KvochurHegel at StanceNakba: Robust Stance Detection with Regularized Natural Language Inference

Minh-Hoang Le
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
24520542@gm.uit.edu.vn

Abstract

Actor-level stance detection over noisy, politically sensitive data can present challenges that standard training procedures fail to handle reliably. This paper presents KvochurHegel, our submission to the StanceNakba 2026 Shared Task, which addresses these challenges by framing stance classification as Natural Language Inference (NLI) to capture actor-level granularity. The official StanceNakba dataset contains high label noise and topic-correlated spurious features, such as texts discussing unrelated global conflicts using in-domain political vocabulary. To handle these conditions within a three-class schema, we construct templates encoding stance hypotheses for specific actors (e.g., "The author expresses support for Palestine") and introduce a broadened neutral class designed to absorb spurious out-of-domain inputs. A DeBERTa-v3 Cross-Encoder independently evaluates the entailment between the input text and each class-specific hypothesis. Because standard cross-entropy training tends to memorize contradictory annotations under these conditions, we regularize the training procedure with R-Drop and label smoothing. This regularized setup likely contributed to robustness against distribution shifts between the competition's evaluation phases (the public leaderboard and private test set), allowing our model to improve from a Macro-F1 of 0.9094 to 0.9384 without requiring large generative models, cross-validation, or inference-time ensembling.

Keywords: Actor-Level Stance Detection, Natural Language Inference (NLI), Label Noise, R-Drop, DeBERTa, StanceNakba

1. Introduction

Actor-level stance detection in polarized social media discourse often requires models to interpret implicit rhetoric and complex ideological framing (ALDayel and Magdy, 2021). The Palestinian-Israeli conflict represents one of the most contested and heavily discussed topics in online political discourse, making automated stance detection over such data both practically relevant and technically challenging. To address this and provide a standardized benchmark, Aldous et al. (2026) introduced the StanceNakba 2026 Shared Task. This paper describes our system submitted to this competition, specifically targeting Subtask A: Actor-Level Stance Detection. The objective is to classify an author's general political alignment into one of three categories: Pro-Palestine, Pro-Israel, or Neutral.

Analysis of the training and development sets identifies severe label noise, contradictory annotations, and topic-correlated spurious features. Texts discussing unrelated geopolitical topics often use in-domain vocabulary (e.g., "military", "colonization"), creating spurious out-of-domain inputs. Under these conditions, standard cross-entropy training tends to memorize contradictory annotations rather than learning generalizable patterns.

To mitigate this behavior, we frame stance detection as an NLI task (Yin et al., 2019) using a DeBERTa-v3 Cross-Encoder. We construct actor-specific hypothesis templates and a broadened neutral class, regularizing the training procedure with R-Drop and label smoothing to prevent overfitting on the noisy annotations. This approach likely contributed to robustness against distribution shifts between the competition's two evaluation phases, improving from a Macro-F1 of 0.9094 on the public leaderboard to 0.9384 on the private test set, without cross-validation or synthetic data.

2. Background

2.1. Related Work

Stance detection has been approached primarily as target-specific text classification, with early systems relying on lexical features and more recent work adopting pre-trained transformer fine-tuning (Mohammad et al., 2016; Küçük and Can, 2020). These approaches, however, rely on surface-level associations that fail under implicit rhetoric and topic-correlated spurious features. Yin et al. (Yin et al., 2019) demonstrate that reformulating text classification as Natural Language Inference improves robustness on implicit reasoning tasks by leveraging pre-trained entailment representations, motivating our hypothesis-based formulation.

A separate challenge in stance detection is annotation noise, particularly in politically sensitive domains where rater disagreement is high (ALDayel and Magdy, 2021). Label smoothing (Szegedy et al., 2016) and consistency regularization via R-Drop (Liang et al., 2021) are established techniques

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 113–117

for reducing model overconfidence under noisy supervision. Our system combines NLI formulation with both techniques, addressing implicit rhetoric and label noise simultaneously without requiring generative model inference.

2.2. Task Definition

Subtask A of the StanceNakba Shared Task focuses on predicting general political alignments at the actor level. Formally, given an input social media text X, the system must predict a stance label $y \in \{\text{Pro-Palestine}, \text{Pro-Israel}, \text{Neutral}\}$. The official dataset consists of 1,401 English samples, partitioned into a 70/15/15 split for training, development, and testing.

2.3. Data Characteristics and Challenges

Analysis of the training and development sets identified three primary phenomena that disrupt standard classification pipelines, detailed in Table 1.

Implicit Rhetoric Stances are often expressed through indirect ideological markers rather than direct declarations of support (Table 1). Models relying on lexical overlap fail to map terms like "colonization" to the correct geopolitical actor without an explicit inference mechanism.

Spurious Out-of-Domain Inputs The dataset includes texts that use the vocabulary of conflict and politics but express no stance on the Palestinian-Israeli conflict. If a model associates words like "military" strictly with the Pro-Israel or Pro-Palestine classes, it will misclassify these inputs.

Contradictory Annotations A particularly disruptive challenge is the presence of severe label noise. Polarized texts are occasionally assigned the "Neutral" label. Training on such data using standard cross-entropy severely degrades generalization, as the network is forced to map explicitly partisan phrasing to neutrality—a failure mode we address through regularization during fine-tuning.

3. Methodology

3.1. Hypothesis Design and NLI Formulation

To preserve the original discourse features of the social media data—including capitalization, punctuation, and special characters—we apply no text preprocessing or normalization. The raw texts are passed directly to the tokenizer. We formulate actor-level stance detection as an independent Natural Language Inference (NLI) task. We design three class-specific hypotheses (H_c), each articulating the defining characteristics of one stance label. This approach allows the model to jointly encode the input text and hypothesis, computing attention across both sequences at every layer.

By including in-domain ideological markers (e.g., "IDF", "Zionism", "resistance") and defining the neutral scope ("unrelated topics"), these hypotheses directly address the implicit rhetoric and spurious inputs identified in Section 2. Evaluating each hypothesis independently, rather than applying a joint three-class classification head, prevents the model from learning inter-class shortcuts and preserves the binary entailment semantics of each hypothesis. During inference, the model evaluates the premise X against all three hypotheses. The final predicted class $\hat{y}$ is the hypothesis that yields the maximum entailment probability (obtained via softmax over the raw NLI logits):

$$\hat{y} = \text{argmax}_c P(\text{Entailment} \mid X, H_c)$$

Although independent evaluation requires three forward passes per example, inference remains efficient given the dataset scale and the absence of ensemble mechanics.

3.2. Architecture and Training Configuration

Our system uses a DeBERTa-v3 Cross-Encoder (He et al., 2023) initialized with the `MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli` checkpoint[1], which is pre-trained on MNLI, FEVER-NLI, and ANLI, offering robust NLI priors for the entailment scoring task. For each forward pass, the tokenizer concatenates the input sequence as `[CLS] X [SEP] H_c [SEP]`. The pooled `[CLS]` representation is passed through a linear classification head to produce the standard three NLI logits (entailment, neutral, contradiction), from which only the entailment logit is used.

The model is optimized using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5}, a weight decay of 0.1, and a warmup ratio of 0.1. We train for a maximum of 10 epochs using a training batch size of 16 (and an evaluation batch size of 32) with `bfloat16` mixed precision. To prevent overfitting to the noisy training annotations, we implement early stopping on the validation loss with a patience of 2 epochs, which triggered at epoch 9 in our final training run.

3.3. Regularization Strategy

Because standard cross-entropy training tends to memorize the severe label noise present in the

[1] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli

Phenomenon	Gold Label	Example Text	Modeling Challenge
Implicit Stance	Pro-Palestine	*"I'm a Hindu,! And you don't speak in behalf of others! Zionists colonization..."* (ID: 807)	Lacks explicit entity mentions; relies entirely on ideological terminology.
Spurious Inputs	Neutral	*"Did you ever serve in the military? ."* (ID: 247)	Contains conflict-related vocabulary but lacks geopolitical stance.
Severe Label Noise	Neutral (Mislabeled)	*"Allah stands with Israel. He has blessed them with many victories over the vile rats..."* (ID: 70)	Contradictory annotation. The text is explicitly partisan but labeled Neutral.

Table 1: Examples of dataset ambiguity, spurious out-of-domain inputs, and label noise in the StanceNakba dataset.

Class	Hypothesis Text (H_c)
Israel	*"The author expresses support for Israel, including its government, military (IDF), security, Zionism, or the Jewish people's right to self-defense."*
Palestine	*"The author expresses support for Palestine, Gaza, or the resistance, including calls for liberation, human rights, or stopping the violence."*
Neutral	*"The text provides a factual, non-judgmental report on the conflict, or discusses completely unrelated topics (like other global wars or daily life)."*

Table 2: Class-specific hypotheses encoding ideological markers for stance classes and a broadened neutral scope to absorb spurious inputs.

StanceNakba dataset, we apply two regularization techniques during fine-tuning: label smoothing and R-Drop.

First, we apply label smoothing ($\epsilon = 0.2$) (Szegedy et al., 2016; Müller et al., 2019) to the cross-entropy objective ($\mathcal{L}_{CE}$). This softens the one-hot target distributions, discouraging the network from assigning maximum confidence to any training label, which is particularly beneficial under contradictory annotations.

Second, we implement R-Drop consistency regularization (Liang et al., 2021) via a custom training loop, as the standard Hugging Face Trainer API does not natively support the requisite double forward pass per step. For each training step, the input sequence is passed through the model twice with independently sampled dropout masks, generating two distinct logit distributions, P_1 and P_2. We enforce predictive consistency between these sub-models by minimizing their bidirectional Kullback-Leibler (KL) divergence ($\mathcal{L}_{KL}$):

$$\mathcal{L}_{KL} = \frac{1}{2}\Big(D_{KL}(P_1 \parallel P_2) + D_{KL}(P_2 \parallel P_1)\Big)$$

The final objective function $\mathcal{L}$ is the sum of the averaged smoothed cross-entropy loss from both forward passes and the KL divergence loss, scaled by $\alpha = 4.0$ (selected via evaluation on the development set):

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{CE}^{(1)} + \mathcal{L}_{CE}^{(2)}) + \alpha\mathcal{L}_{KL}$$

The impact of this regularized setup on model robustness is evaluated in Section 4.

3.4. Implementation Details

Our system is implemented in Python 3.13.2 using PyTorch 2.7.1 (Paszke et al., 2019)[2] and the Hugging Face Transformers library 4.55.4 (Wolf et al., 2020)[3]. Dataset loading and preprocessing use the Hugging Face Datasets library 4.5.0 (Lhoest et al., 2021)[4]. Evaluation metrics are computed using scikit-learn 1.7.1 (Pedregosa et al., 2011)[5]. The complete training notebook is publicly available in our repository[6].

4. Results and Discussion

4.1. Official Evaluation

Submitted under the team name KvochurHegel, our system achieved a Macro-F1 score of 0.9384, securing Rank 4 on the official leaderboard.

As shown in Table 3, precision, recall, and accuracy align closely with the Macro-F1 score, suggesting broadly balanced per-class performance.

[2] https://pytorch.org/
[3] https://huggingface.co/transformers
[4] https://huggingface.co/docs/datasets/
[5] https://scikit-learn.org/
[6] https://github.com/lmhoang06/StanceNakba2026_KvochurHegel

Phase	Macro-F1	Acc.	Prec.	Rec.
Public	0.9094	0.9095	0.9096	0.9095
Private	0.9384	0.9384	0.9387	0.9384

Table 3: System performance across the public and private evaluation phases.

For context, the top three systems on the leaderboard achieved Macro-F1 scores of 0.9620, 0.9525, and 0.9426, respectively. Our single-run submission performs within 0.024 points of the winning system.

4.2. Performance and Generalization

The 0.029-point Macro-F1 improvement between evaluation phases likely reflects two contributing factors. First, the regularization strategy (R-Drop and label smoothing) reduced overfitting to training noise, consistent with the improved private test performance. Second, the private test distribution may have naturally contained fewer contradictory annotations or spurious out-of-domain inputs than the public split, constituting an objectively easier evaluation set.

4.3. System Limitations

While the system avoids complex inference-time ensembles, evaluating three independent NLI hypotheses per sample triples the inference cost relative to a standard classification head. However, this overhead remains acceptable given the dataset scale. Furthermore, the current formulation evaluates each post in isolation, without access to conversational context such as parent posts or reply threads. This restricts the model's ability to resolve implicit stances that depend entirely on discourse context (e.g., agreeing with an unincluded parent post).

4.4. Per-Class Performance and Error Analysis

To clarify the evaluation framework, the terms "public" and "private" test sets refer strictly to the two sequential phases of the shared task evaluation, rather than standard static development and test splits. Because per-class breakdowns were not provided by the organizers for the final private phase, Table 4 reports the detailed metrics and corresponding confusion matrix evaluated on the public test set.

The system achieves its highest F1 on Pro-Israel (0.936), with Pro-Palestine (0.899) and Neutral (0.894) showing comparable but lower performance. Pro-Palestine exhibits the highest misclassification count (8 instances), with the dominant

Per-Class Metrics

Class	P	R	F1
PI	0.930	0.943	0.936
PP	0.912	0.886	0.899
Neu	0.887	0.900	0.894

Confusion Matrix

	PI	PP	Neu
PI	66	2	2
PP	2	62	6
Neu	3	4	63

Table 4: Public test set performance breakdown. Matrix rows represent the true labels, and columns represent the predicted labels (PI: Pro-Israel, PP: Pro-Palestine, Neu: Neutral). P: Precision, R: Recall.

error being confusion with the Neutral class (6 instances), consistent with the spurious input challenge identified in Section 2.

Inspection of the 19 misclassified public test samples reveals two recurring failure modes. First, the absence of conversational context forces errors on texts whose stance depends on an unincluded parent post. For example, *", and I say no to war. No I don't stand with ..."* (ID: 1008) is predicted Pro-Palestine but labeled Neutral, likely because the implied stance requires the unincluded parent post for resolution. Second, residual label noise that regularization cannot fully override produces irreducible errors. The text *"I stand against jihad. Merci Israel."* (ID: 1088) contains explicitly Pro-Israel rhetoric but carries a Pro-Palestine gold label; the model predicts Neutral, a more semantically coherent output than the gold annotation. These cases illustrate that performance is bounded by both annotation quality and the absence of conversational context, consistent with the limitations noted in Section 4.3.

5. Conclusion and Future Work

This paper formulated actor-level stance detection as an independent Natural Language Inference task. We addressed spurious out-of-domain inputs through class-specific hypothesis design, and mitigated the memorization of severe label noise via R-Drop and label smoothing. Achieving Rank 4 with a single-run DeBERTa-base model demonstrates that combining NLI formulation with engineered hypotheses and regularization achieves competitive performance without requiring large generative models, cross-validation, or inference-time ensembling. Future work may extend this framework to incorporate discourse context, enabling the resolution of implicit stances within reply threads while maintaining the efficiency of the single-run architecture.

6. Bibliographical References

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

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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, San Diego, California. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.