U4RASD at StanceNakba Shared Task: Data Augmentation and Auxiliary Objectives for Arabic Stance Detection

Nancy Hamdan[1], **Aya Jouni**[1], **Aya Saïd**[2], **Fadi A. Zaraket**[1,2]
[1]Arab Center for Research and Policy Studies, [2]American University of Beirut
{nhamdan, ajouni, fzaraket}@dohainstitute.edu.qa; aas165@mail.aub.edu

Abstract

This paper describes a submission to Track B of the StanceNakba Shared Task on Arabic cross-topic stance detection in the political domain. We investigate LLM-based data augmentation, auxiliary training objectives including contrastive and multi-task learning, zero-shot prompting, and a preliminary terminology-based clustering approach. Our final system, based on MARBERTv2 with dialect-aware LLM-based augmentation, achieved 86% macro-F1 on the blind test set and ranked 3rd out of 10 teams. Our results show that dialect-aware augmentation substantially improved performance in a low-resource Arabic stance detection setting, while not all auxiliary objectives or clustering-based strategies yielded consistent gains. We release our code at https://acr.ps/1L9B9Tw.

Keywords: Arabic Stance Detection, Data Augmentation, Auxiliary Objectives

1. Introduction

Stance detection concerns the automated identification of the position expressed in an input text towards a specified target idea, object, or proposition (Somasundaran and Wiebe, 2010; Mohammad et al., 2016). Stance *favor*, *against*, and *neutral* labeling differs from sentiment analysis as it also concerns the relationship between the text and the specified target (Mohammad et al., 2016; Alturayeif et al., 2024). This makes stance nuanced and sensitive to *targets* where opinions may be implicit or sarcastic.

Stance detection plays an important role in analyzing public opinion and political polarization and can complement traditional surveys and polls (Küçük and Can, 2020). It can be formulated as either single-target, where one separate model is trained per target, or multi-target, where a unified model takes the target as part of its input and learns jointly across several targets (Sobhani et al., 2017; Xu et al., 2018).

This paper describes our submission to Track B of the StanceNakba Shared Task (Aldous et al., 2026), which focuses on Arabic cross-topic stance detection for the normalization with Israel (Topic 1) and refugee presence in Jordan (Topic 2) topics. We build upon the shared task baseline system and make the following contributions.

- We explore data augmentation strategies and identify an effective approach.
- We investigate auxiliary training objectives, including contrastive and multi-task learning.
- We conduct a preliminary study of terminology-based clustering using stance-specific embedding centroids.
- We evaluate zero-shot prompting as a training-free alternative.

Our best performing system, based on MARBERTv2 (Abdul-Mageed et al., 2021) with dialect-aware LLM-based data augmentation, achieved 88.7% macro-F1 on the validation set and 86% on the blind test set, ranking 3^{rd} out of 10 teams. Our results show that dialect-aware augmentation substantially improved performance, while not all augmentation, auxiliary objectives, or clustering-based strategies yielded consistent gains. We further analyze the interaction between contrastive learning and data augmentation using effective rank analysis.

2. Background: Setup and Dataset

Track B focuses on Arabic cross-topic stance detection. Given a social media post and an associated topic, the model predicts a *pro*, *against*, or *neutral* stance label.

The dataset, subset of MARASTA (Charfi et al., 2024), provides 1,205 Arabic posts covering **Topic 1:** (*Normalization with Israel*) and **Topic 2:** (*Refugee/Immigrant Presence in Jordan*). Table 1 presents example samples. Table 2 summarizes the distribution across the training and development subsets. The relatively small size of the dataset, combined with Arabic dialectal variation and rich morphology, increases model generalization difficulty.

3. Related Work

Stance detection was formalized in the target-specific setting in SemEval-2016 (Mohammad et al., 2016), and later extended to multi-target and multi-task settings (Sobhani et al., 2017; Xu et al., 2018; Li and Caragea, 2021).

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 206–211

ID	Sentence	Topic	Label
1	انا من فلسطين واتمنى أنهم العراقيين يجو لا عنا *I am from Palestine, and I hope Iraqis come to us.*	Topic 2	Pro
6	والله لو تصيحون الى ان تنشف حلوقكم مافيه تطبيع معكم *Even if you shout until your throats dry, there will be no normalization with you.*	Topic 1	Against
26	سفير الدولة في الأردن يزور المخيم الأردني للاجئين السوريين *The country's ambassador to Jordan visits the Jordanian camp for Syrian refugees.*	Topic 2	Neutral

Table 1: Sample excerpts of instances from the Track B training dataset.

Topic	Label	Train	Dev	Topic total
Norm. w/ Israel	Pro	120	26	491
	Against	139	29	
	Neutral	145	32	
Refugees in Jordan	Pro	166	36	533
	Against	159	34	
	Neutral	114	24	
Total		**843**	**181**	**1,024**

Table 2: Track B label distribution by topic for the training and development datasets.

Arabic stance detection remains relatively under-resourced, though recent efforts have expanded datasets and modeling approaches. The MAWQIF dataset (Alturayeif et al., 2022), used in StanceEval 2024 (Alturayeif et al., 2024), includes stance along with sentiment and sarcasm labels. The winning system incorporated multi-task learning and contrastive loss objectives (Badran et al., 2024). ArabicStanceX later introduced a larger multi-topic dataset for Arabic stance detection (Alkhathlan et al., 2025).

4. System Overview

Our system is based on fine-tuning an encoder LLM for stance detection. We explored four directions to improve performance: (1) data augmentation, (2) auxiliary training objectives, (3) terminology-based clustering, and (4) zero-shot prompting. We provide a concrete algorithmic walkthrough for each component.

4.1. Data Augmentation

We explored LLM-based augmentation strategies to increase data diversity, improve generalization and compensate for the limited size of the data.

Counterfactual Stance Augmentation We used Gemini 2.5 Flash Lite (Google, 2025b) to generate counterfactual variants of each training sample. For every instance labeled as *pro*, *against*, or *neutral*, the model generated two additional variants corresponding to the remaining stance labels while preserving the topic and applying minimal edits.

External Data Augmentation We incorporated data from an internal benchmark on politically sensitive topics.[1] For Topic 1, we used Palestine/Israel-related data, and for Topic 2, refugee/immigration-related data, then expanded them with Gemini 2.5 Flash Lite while preserving topic and stance labels.

Dialect-Aware Augmentation We used Gemini 3 Flash Preview (Google, 2025a) to generate paraphrased versions of each training sample. For each training instance, the model generated three paraphrases while strictly preserving both stance and topic. The prompt enforced dialectal consistency: dialectal samples remain dialectal, MSA samples remain MSA, and mixed samples remain mixed. This increased lexical diversity while preserving the original data distribution.

4.2. Auxiliary Training Objectives

In addition to data augmentation, we experimented with modifying the training objective.

Contrastive Learning We combined supervised contrastive loss using L2 distance (Hadsell et al., 2006) with cross-entropy loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{contrastive}}, \quad \lambda = 0.1.$$

The contrastive term encourages samples with the same stance to have closer embeddings while separating different stance classes.

Multi-Task Learning We explored multi-task learning (Ruder, 2017) using LLM-generated sarcasm and sentiment labels. We used Gemini 2.5 Flash Lite to produce the auxiliary annotations, and the model was trained with three classification heads using uncertainty-based loss weighting (Cipolla et al., 2018).

[1] https://acr.ps/1L9F2Dn

Hyper-parameter	Value
Maximum sequence length	512 tokens
Train batch size	32
Learning rate	2.0×10^{-5}
Number of epochs	10 (30 with augmented data)
Warm-up ratio	10%
Weight decay	0.01
Optimizer	AdamW

Table 3: Training hyper-parameters used for model fine-tuning.

Preliminary Terminology-Based Clustering aims to characterize the stance towards a topic with specific terms and entities. A domain expert could define stance-indicative terminology and use it to infer stance. We investigated an automatic construction alternative.

For each topic and stance label, we collected 10 representative articles and embedded them using Qwen3 8B embeddings (Zhang et al., 2025). We computed a centroid embedding for each stance group. Due to limited experimentation we excluded this promising method from our submission, and we plan to further explore it in future work.

4.3. Zero-Shot Prompting

We evaluated zero-shot prompting with Gemini 2.5 Flash Lite (Google, 2025b), prompting it to predict *pro*, *against*, or *neutral* given the topic while accounting for sarcasm and dialect.

5. Experimental Setup

In this section, we describe the data preprocessing, training configuration, and evaluation protocol we followed.

Data Preprocessing We applied standard Arabic normalization: removing diacritics, tatweel (ـ), URLs, mentions, and hashtag symbols; normalizing whitespace; and unifying common character variants such as alef (ا آ إ أ), alef maqsura (ى), and taa marbuta (ة).

Training Configuration All models were trained for 10 epochs, or 30 with augmented data, using a maximum sequence length of 512 and batch size 32. Full hyperparameters are listed in Table 3.

Models We trained *AraBERT* v0.2-base (Antoun et al., 2020) as the baseline provided by the organizers, and evaluated *MARBERT* v2 (Abdul-Mageed et al., 2021). We adopted *MARBERT*, as it outperformed *AraBERT*, in our subsequent experiments.

Models were trained using HuggingFace Transformers (Wolf et al., 2020) and we used PyTorch Metric Learning (Musgrave et al., 2020) for the contrastive loss function. For data augmentation and zero-shot prompting, we used the OpenRouter[2] platform. Appendix A shows the version numbers and URLs for all tools and libraries used.

Evaluation Models were trained on the official training set, with augmented samples when applicable, and evaluated using stratified 5-fold cross-validation on the training data and the full development set. We report macro-F1 as the primary metric, along with macro precision, macro recall, and accuracy. The model with the best development set performance was selected for our final submission.

6. Results

Table 4 reports performance on the training set using 5-fold cross-validation and on the development set. We report macro precision, recall, F1, and accuracy. Our submitted system achieved a macro-F1 score of 86.0 on the blind test set as shown in Table 5.

Backbone Model Comparison Replacing *AraBERT* with *MARBERT* improved macro-F1 by approximately 5.5 points on the development set. This is likely due to better domain alignment, as *MARBERT* was exclusively pretrained on Arabic tweets, similar to the shared task data.

Effect of Contrastive and Multi-Task Learning Adding contrastive loss improved *MARBERT* by about 1.1 macro-F1 points, while multi-task learning with LLM-generated sarcasm and sentiment labels reduced performance by about 1 point, likely due to noise in the auxiliary annotations.

Effect of Data Augmentation Counterfactual and external augmentation reduced performance by about 1 macro-F1 point relative to *MARBERT*. In contrast, dialect-aware augmentation yielded the best results, improving macro-F1 by about 5.4 points and corresponding to our submitted system. Combining it with contrastive loss reduced performance by 2.8 points.

Zero Shot Prompting Gemini 2.5 Flash Lite achieved a macro-F1 of 77.4 in the zero-shot setting, approaching the baseline fine-tuned model without task-specific training.

[2]https://openrouter.ai/

Table 4: Performance comparison across different stance detection configurations on the training set using 5-fold stratified cross-validation (CV), and on the development set (Dev). All metrics are Macro-averaged except for Accuracy. **Bold** indicates best performance; underline indicates second-best performance per metric and category. The final submitted system is highlighted.

ID	Description	Precision		Recall		Macro F1		Accuracy	
		CV	Dev	CV	Dev	CV	Dev	CV	Dev
Overall Performance									
1	Baseline (AraBERTv0.2)	.795	.783	.791	.779	.791	.780	.791	.779
2	MARBERTv2	.819	.833	.812	.834	.812	.833	.814	.834
3	MARBERTv2 + Contrastive Loss	.809	.848	.804	.843	.804	.844	.805	.845
4	MARBERTv2 + Multitask Learning	.781	.830	.770	.821	.772	.823	.773	.823
5	MARBERTv2 + Counterfactual + External Aug.	.977	.822	.976	.823	.977	.822	.976	.823
6	MARBERTv2 + Dialect-Aware Aug.	.949	.888	.948	.887	.948	**.887**	.948	.889
7	MARBERTv2 + Dialect-Aware Aug. + Contrastive Loss	.943	.861	.943	.859	.943	.859	.943	.861
8	Gemeni 2.5 Flash Lite + Zero Shot Prompting	-	.819	-	.769	-	.774	-	.773
9	Qwen 3 8B + Clustering	-	.340	-	.357	-	.307	-	.364
Topic 1: Normalization with Israel									
6	MARBERTv2 + Dialect-Aware Aug.	.927	.839	.926	.846	.926	**.841**	.925	.839
7	MARBERTv2 + Dialect-Aware Aug. + Contrastive Loss	.917	.795	.920	.798	.918	.797	.917	.793
Topic 2: Refugees and Immigrants in Jordan									
6	MARBERTv2 + Dialect-Aware Aug.	.970	.933	.968	.933	.969	**.933**	.969	.936
7	MARBERTv2 + Dialect-Aware Aug. + Contrastive Loss	.967	.930	.967	.920	.967	.924	.967	.925

System	Macro-F1 (Blind Test)
MARBERTv2 + Dialect-Aware Augmentation	.860

Table 5: Official blind test performance of our submitted system to track B in the StanceNakba shared task.

Terminology-Based Clustering The preliminary clustering approach performed substantially worse than supervised models, achieving 30.7 macro-F1 on the development set. This indicates that simple centroid-based clustering with limited seed documents is insufficient for reliable stance classification.

Topic Level Analysis Across configurations, performance was consistently lower on Topic 1 than Topic 2. The highest macro-F1 on Topic 1 was 84.1, compared to 93.3 on Topic 2.

Error Analysis Most errors for the submitted system on the development set occur between the *neutral* and *against* classes (7 neutral samples were predicted as against and 6 against samples as neutral), while *pro* is more reliably detected. For example, sarcastic or implicit posts such as "الاردن حيث البطاله المستفحله للاردنيين الحكومه تعد بتوفير 200الف فرصه عمل للاجئين السوريين؟!" ("Jordan, where unemployment is rampant among Jordanians, while the government promises 200,000 job opportunities for Syrian refugees?!"; neutral classified as against) are misclassified due to implicit stance expression. These cases are inherently ambiguous, and some *neutral*/*against* confusions may reflect plausible alternative interpretations of the stance.

Combining dialect-aware augmentation with contrastive loss degraded performance compared to augmentation alone. We hypothesize that this is related to dimensional collapse in contrastive learning (Jing et al., 2022). To examine this, we computed the effective rank of the learned representations using RankMe (Garrido et al., 2023). The submitted system achieved a RankMe score of 3.18, while the model trained with augmentation and contrastive loss scored 1.70, a 46.5% reduction. This suggests reduced diversity of learned representations when contrastive learning was combined with data augmentation.

7. Conclusion

We presented our submission to Track B of the StanceNakba Shared Task (Aldous et al., 2026). Our system based on MARBERTv2 with dialect-aware LLM-based augmentation ranked 3rd on the blind test set, showing that controlled dialect-aware augmentation improves performance, while auxiliary objectives did not consistently help. This work is limited by the small dataset size and its restriction to two topics, which may affect generalizability. In future work, we will evaluate the approach on broader datasets and further explore clustering-based methods.

8. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 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, Online. Association for Computational Linguistics.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

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

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, Bangkok, Thailand. Association for Computational Linguistics.

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

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, Marseille, France. European Language Resources Association.

Mohamed Badran, Mo'men Hamdy, Marwan Torki, and Nagwa El-Makky. 2024. AlexUNLP-BH at StanceEval2024: Multiple contrastive losses ensemble strategy with multi-task learning for stance detection in Arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 823–827, Bangkok, Thailand. Association for Computational Linguistics.

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

Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. 2023. Rankme: assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Google. 2025a. Gemini 3 flash: frontier intelligence built for speed. Google Blog.

Google. 2025b. We're expanding our gemini 2.5 family of models. Google Blog.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning.

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

Yingjie Li and Cornelia Caragea. 2021. A multi-task learning framework for multi-target stance detection. In *Findings of ACL-IJCNLP*, pages 2320–2326.

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.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. Pytorch metric learning. *ArXiv*, abs/2008.09164.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks.

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, Valencia, Spain. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models.

A. Tools and Libraries

Table 6 summarizes the tools and libraries we used in our experiments and their corresponding versions and URLs.

Tool / Library	Version	Usage
PyTorch	`2.10.0`	Deep learning framework
HuggingFace Transformers	`4.49.0`	Model training and fine-tuning
PyTorch Metric Learning	`2.9.0`	Contrastive loss implementation
OpenAI Python SDK	`2.16.0`	Client library used to access OpenRouter-hosted LLMs
OpenRouter API	`v1`	LLM provider endpoint used for data augmentation and zero-shot prompting

Table 6: Tools, libraries, and APIs used in the experiments, with versions and URLs.