Phatthachdau at SemEval-2026 Task 9: A Multi-Stage Augment-Judge-Train Pipeline for Multilingual Online Polarization Detection

Thanh Phat Phan Duc
VNUHCM - University of Information Technology
23521149@gm.uit.edu.vn

Abstract

For SemEval-2026 Task 9 (Detecting Online Polarization), we address two contrasting challenges: the extreme label imbalance in low-resource languages (e.g., Hausa, with only 11% polarized instances) and the sophistication of implicit hostility in high-resource languages (English). We propose the Augment-Judge-Train (AJT) pipeline to overcome these barriers. By leveraging Gemini 2.0 Flash for taxonomy-driven synthetic data generation and Gemini 2.5 Pro as an LLM-as-a-Judge layer for quality control, we effectively expand the minority class while preserving cultural nuances. Our final system utilizes a weighted soft-voting ensemble, combining specialized Encoders (for local linguistic patterns) with LLM-LoRA (for deep contextual reasoning). Evaluated on the official test set, our system achieved 1st Place in Hausa (0.8336 Macro-F1) and ranked in the top 10 for English (0.8092 Macro-F1), demonstrating the efficacy of culture-aware synthetic data in enhancing social NLP.

1 Introduction

Detecting online polarization is fundamentally distinct from standard toxicity detection; it relies on capturing implicit "us-versus-them" rhetoric, group identity conflicts, and culture-specific nuances rather than simply matching explicit offensive keywords. This complexity poses a significant challenge for standard language models, which often default to surface-level heuristics and fail to capture implicit group conflict.

The SemEval-2026 Task 9 (Naseem et al., 2026a) addresses these gaps through the POLAR dataset (Naseem et al., 2026b)—a multilingual, multicultural, and multi-event benchmark containing over 110,000 instances across 22 languages. In this paper, we describe the Phatthachdau system for the POLARDETECT task (Naseem et al., 2026a) (binary polarization detection). We focus on two languages that represent distinct computational challenges: English (high-resource) and Hausa (low-resource). Participating in this task revealed two contrasting obstacles: extreme data scarcity in Hausa—where only 11% of instances are labeled as polarized—and the sophistication of implicit hostility in English discourse.

To overcome these barriers, we propose the Augment-Judge-Train (AJT) pipeline. Our approach mitigates extreme label imbalance while preserving the cultural DNA of social media texts, leveraging the following key contributions:

- **Taxonomy-Guided Augmentation:** Using Gemini 2.0 Flash to generate targeted synthetic data. By prompting the model with a predefined taxonomy of polarization types (e.g., Political, Racial/Ethnic, Religious, Gender/Sexual), we synthesize culture-aware minority class samples for Hausa and difficult "hard-negative" samples for English.
- **Data Quality Assurance:** Implementing an LLM-as-a-Judge layer using Gemini 2.5 Pro to filter and ensure the integrity of augmented samples before model training, preventing the introduction of noisy or weakly aligned data.
- **Ensemble Architecture:** A weighted soft-voting ensemble that combines specialized Encoder models (optimized for local linguistic scripts) with the deep contextual reasoning of LLM-LoRA.

Experimental results demonstrate that the AJT method provides consistent improvements, particularly in addressing severe label imbalance. Our system achieved 1st Place for Hausa with a Macro-F1 of 0.8336 and ranked within the top 10 for English with a score of 0.8092. These results suggest that integrating human-expert knowledge through a structured taxonomy into the data augmentation

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1616–1620

process is useful for building robust, culture-aware social NLP models.

Our code and augmented datasets are publicly available at: https://github.com/kelvin2250/SemEval2026_Task9.

2 Background and Related Work

2.1 Task Definition and Affective Polarization

The SemEval-2026 Task 9 focuses on identifying attitude polarization in multilingual online discourse (Naseem et al., 2026a,b). We address Subtask 1 (binary classification) for English and Hausa. A major challenge is the extreme label imbalance in the Hausa dataset, where the polarized class represents only 11% of instances. To effectively capture polarization, our system builds upon the *Polarization Footprint* framework (Build Up, 2025), which defines affective polarization as identity-based animosity. We specifically target rhetorical markers like dehumanization and vilification during our synthetic data generation process.

2.2 Data Augmentation and Robust Ensembling

Traditional data augmentation techniques, such as back-translation, often fail to preserve the nuanced cultural DNA of social media texts. While recent approaches successfully employ Large Language Models (LLMs) for paraphrase augmentation, generating high-volume synthetic data can introduce noisy or weakly aligned samples. To address label scarcity and ensure data integrity, our *Augment-Judge-Train* (AJT) pipeline advances standard augmentation by combining a taxonomy-driven generation approach with an LLM-as-a-Judge (Zheng et al., 2023) filtration layer. Finally, to handle the complex and imbalanced nature of social NLP, we employ an ensemble strategy that merges the local linguistic pattern recognition of specialized encoders with the deep contextual reasoning of LLM-LoRA.

3 System Overview

Our proposed system, **Phatthachdau**, centers on a multi-stage pipeline named **Augment-Judge-Train (AJT)**. This framework is designed to tackle the inherent "long-tail" distribution of the POLAR dataset (Naseem et al., 2026b) by synthesizing additional data, filtering for quality, and optimizing through an ensemble of heterogeneous architectures (Dietterich, 2000).

- **Stage 1: Taxonomy-Driven Augmentation.** To address data scarcity (e.g., the 11% polarization minority in the Hausa dataset), we utilize Gemini 2.0 Flash (Gemini Team et al., 2023) to generate targeted synthetic data. For Hausa, we apply a *Multilabel-Conditioned* strategy, leveraging labels to anchor the generation of authentic *Hausa baka* (colloquial) samples. For English, we employ *Hard-Negative Mining* to create sensitive-but-neutral texts, alongside *DNA-Mimicry* to simulate partisan slang and harsh rhetorical structures found in the seed data (Build Up, 2025).

- **Stage 2: Quality Control via LLM-as-a-Judge.** To prevent the introduction of noisy or weakly aligned data, all synthetic samples undergo an automated validation step using Gemini 2.5 Pro (Gemini Team et al., 2023) based on the LLM-as-a-Judge paradigm (Zheng et al., 2023). We evaluate each sample's relevance on a 0.0–1.0 scale; only high-quality instances achieving a score of $s \geq 0.70$ are integrated into the final training set, ensuring contextual reliability for downstream fine-tuning.

- **Stage 3: Hybrid Modeling and Soft-Voting Ensemble.** We capture both local linguistic nuances and deep contextual reasoning by combining specialized Encoders, such as XLM-RoBERTa (Conneau et al., 2020) and RoBERTa (Liu et al., 2019), with LLM-LoRA architectures (Hu et al., 2022). The final classification relies on a weighted soft-voting ensemble (Dietterich, 2000): $P_{ens} = \frac{\sum w_i P_i}{\sum w_i}$, where the weight w_i prioritizes the contribution of each individual model based on its validation performance.

3.1 Model Architectures

We utilize a hybrid modeling approach to capture both local linguistic nuances and global contextual reasoning:

- **Specialized Encoders:** We finetune `xlm-roberta-base-fine-tuned-hausa` (Conneau et al., 2020) for low-resource scripts and `RoBERTa-base` (Liu et al., 2019) for English, focusing on efficient token representation of social media signals (emojis, hashtags).

- **LLM-LoRA:** For English, we employ Large Language Models fine-tuned through Low-Rank Adaptation (LoRA) (Hu et al., 2022). This allows the system to detect "latent polarization"—where hostility is implied through sarcasm or cultural metaphors rather than explicit keywords.

3.2 Weighted Soft Voting Ensemble

The final system output is derived from an ensemble that aggregates class probabilities from N models trained on varying synthetic-to-original data ratios. Unlike simple majority voting, our weighted soft voting approach prioritizes models with higher validation performance.

The accumulated weighted probability P_{ens} for the polarized class is calculated as:

$$P_{ens} = \frac{\sum_{i=1}^{n} w_i \cdot P_i(\text{class} = 1)}{\sum_{i=1}^{n} w_i} \quad (1)$$

where w_i is the weight assigned to the i-th model and P_i is the predicted probability for the polarized class. The final classification decision Y is determined by the following decision rule:

$$Y = \begin{cases} 1 & \text{if } P_{ens} \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

This strategy reduces sensitivity to individual model errors and provides a consensus-based decision boundary that is more robust to the noise inherent in social media data.

4 Experimental Setup

This section details the empirical configuration of the **Phatthachdau** system, including the data partitioning strategy, the specific parameters of the AJT pipeline, and the hardware environment used for training.

4.1 Data Processing and Balancing

To prepare the noisy social media text from the POLAR dataset for model training, we apply a series of normalization steps:

- **Text Normalization:** We standardized user mentions to `@USER` and URLs to `HTTPURL` to reduce vocabulary sparsity.
- **Demojizing:** Emojis were converted into their textual descriptions using the `emoji` library, preserving the emotional signals critical for polarization detection.
- **Handling Imbalance:** For the Hausa subtask, which exhibits an extreme 11% polarization ratio, we utilized the AJT pipeline to oversample the minority class until a balanced distribution was achieved. For English, we focused on generating "Hard-Negative" samples to refine the decision boundary between sensitive-but-neutral and polarized discourse.

4.2 Augmentation and Validation Settings

Data generation and quality control used separate models and stage-specific settings (summarized in Table 1). Augmentation used Gemini 2.0 Flash, while validation used Gemini 2.5 Pro. We only explicitly set temperature in code; other decoding parameters (e.g., top-p, max output tokens) used API defaults.

Phase	Hyperparameter	Value
Augmentation (Gemini)	Backbone	Gemini 2.0 Flash
	Temperature	0.75 (English), 0.85 (Hausa)
Validation (Judge)	Backbone	Gemini 2.5 Pro
	Temperature	0.30
	Scoring Threshold (G_{val})	≥ 0.70

Table 1: Hyperparameters for the AJT pipeline stages.

4.3 Training and Fine-Tuning Configuration

All experiments were conducted in a Python 3.13 environment using standard libraries (PyTorch, Transformers, PEFT) and trained on a single 16GB VRAM NVIDIA P100 GPU. To ensure reproducibility, we used an 80/20 stratified train-validation split with a fixed random seed of 42, selecting the best checkpoints based on the validation Macro-F1 score.

For the ST1 Encoders, we utilized `RoBERTa-base` for English and `xlm-roberta-base-finetuned-hausa` for Hausa. These models were trained for 10 epochs using a learning rate of 2×10^{-5}, a batch size of 16 for training and 32 for evaluation, a maximum sequence length of 128, a weight decay of 0.01, and a warm-up ratio of 0.06.

For the LLM-LoRA architecture, we fine-tuned `Meta-Llama-3.1-8B` using 4-bit NF4 quantization (bfloat16 compute). The LoRA adapters were configured with $r = 16$, $\alpha = 32$, and a dropout of 0.05. The training was conducted for 2 epochs with a learning rate of 2×10^{-4}, a batch size of 4 with 4 gradient accumulation steps, and a maximum sequence length of 256. Evaluation occurred every 80 steps, coupled with an early stopping patience of 4.

5 Results and Analysis

5.1 Official Leaderboard Results

Language	Team Rank	Username	Score (F1)
Hausa	**1st Place**	**phatthachdau**	**0.8336**
	2nd Place	pfr812	0.8324
English	1st Place	howard	0.8252
	10th Place	**phatthachdau**	**0.8092**

Table 2: Official leaderboard results for the POLARDETECT task.

5.2 Dataset Statistics and Augmentation Impact

Lang.	Label	Orig.	Aug.	Final
Hausa	Polarized (1)	392	2,481	2,873
	Non-Pol (0)	3,259	0	3,259
	Total	**5,082**	**2,481**	**7,563**
English	Non-Pol (0)	2,047	1,827	3,874
	Polarized (1)	1,175	1,096	2,271
	Total	**3,222**	**2,923**	**6,145**

Table 3: Statistics of the training datasets after AJT expansion.

5.3 Detailed Ensemble Configuration

Language	Model Component	Dataset Split	Weight
Hausa	XLM-R (Hau-1)	Partial Aug (df_1)	0.20
	XLM-R (Hau-2)	Medium Aug (df_2)	0.20
	XLM-R (Hau-3)	Full Aug (df_3)	0.30
	XLM-R (Hau-Orig)	Original Set	0.30
English	**LLM-LoRA (1)**	**LLM 500 Full**	**0.40**
	LLM-LoRA (2)	LLM 850 Aug	0.25
	LLM-LoRA (3)	LLM Full Data	0.10
	RoBERTa (1)	RBase 300 Full	0.15
	RoBERTa (2)	RBase Full Data	0.10

Table 4: Detailed ensemble configuration weights.

All weights in Table 4 are fixed constants in the ensemble script, and final labels are derived with a global threshold $P_{ens} \geq 0.5$.

5.4 Analysis and Discussion

The experimental results support the effectiveness of our *Augment-Judge-Train* (AJT) pipeline. For Hausa, dividing the 2,481 augmented samples into different training splits (df_1 to df_3) allowed the ensemble to capture diverse polarization patterns. In English, adding 1,827 non-polarized samples through Gemini 2.0 Flash augmentation helped the model distinguish between intense debate and actual attitude polarization.

6 Conclusion

In this paper, we described the **Phatthachdau** system for SemEval-2026 Task 9: POLAR. Our experiments show that the *Augment-Judge-Train* (AJT) pipeline is effective in mitigating extreme label imbalance in low-resource settings. By leveraging LLM-LoRA architectures and a multi-split ensemble of specialized encoders, we successfully addressed the 11% polarization minority in Hausa through the targeted generation of 2,481 high-quality samples.

Our system achieved **1st Place** in the Hausa subtask with a Macro-F1 of 0.8336 and a top 10 ranking in English with 0.8092. These results suggest that combining taxonomy-driven generation with a rigorous LLM-as-a-Judge filtration layer supports data integrity and enhances the model's ability to detect both explicit and latent polarization markers. Future work will explore more efficient ways to capture culture-specific nuances and improve the scalability of the ensemble architecture.

7 Limitations

Despite its strong performance, our system faces several limitations. First, the ensemble strategy—combining up to five models per language—and the reliance on external Gemini APIs for data augmentation introduce considerable computational overhead, latency, and potential cost. Furthermore, hardware constraints (16GB VRAM) necessitated 4-bit quantization (Dettmers et al., 2023) and LoRA, which may slightly restrict model expressivity compared to full-parameter fine-tuning. Regarding data preparation, our strict LLM-as-a-Judge filtration threshold (Score ≥ 0.70) risks excluding borderline samples that could otherwise offer valuable linguistic diversity. Finally, the system still struggles with deep socio-linguistic nuances, occasionally misclassifying deep irony in English discourse and complex code-switching or overlapping religious terminology in Hausa.

References

Build Up. 2025. The polarization footprint in europe. Technical report, Build Up.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Asso-*

ciation for Computational Linguistics, pages 8440–8451.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Hao, Zhanghao Wu, Sylvain Ba, Eugene Zhuang, Zihao Lin, Zhuohan Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.