The Argonauts at SemEval-2026 Task 9: Multilingual Polarization Detection and Classification Using LLM Prompting and Transformer Fine-Tuning

Sha Newaz Mahmud*, **Sajib Bhattacharjee***, **Md. Refaj Hossan, Kawsar Ahmed and Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
u2004081@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Online polarization, defined as the pronounced division of public opinion into antagonistic groups, poses a significant threat to social cohesion. Automatic detection of polarization across diverse languages and cultures is essential for effective monitoring of online discourse. The challenge extends beyond identifying hate speech to recognizing more nuanced forms, including negative stereotypes, attribution of blame, and dehumanization. This work addresses SemEval-2026 Task 9, which focuses on detecting polarization in multiple languages. Specifically, Subtask 1 involves binary classification of message polarization, while Subtask 2 requires assigning multiple polarization labels in English and Bengali. For Subtask 1, Qwen3-14B is employed with structured few-shot prompting in 4-bit mode, yielding test macro-F1 scores of 0.847 for Bengali (4th place) and 0.808 for English (9th place). For Subtask 2, XLM-RoBERTa-large and RoBERTa-base are fine-tuned using an uneven loss ($\gamma^+ = 1, \gamma^- = 4$) and label-specific thresholds, which increase development macro F1 by up to 24.6 points. The final test macro F1 for English is 0.454 (21th place). Analysis indicates that large language model prompting enhances binary polarization detection, while threshold adjustment is critical for addressing class imbalance in multi-label tasks.

1 Introduction

In the age of social media, polarization has become a pressing societal concern, as opinions increasingly cluster into sharply opposing groups marked by hostility, exclusionary rhetoric, and growing affective animosity. Sunstein (2002) established that group deliberation tends to push individuals toward more extreme positions, a phenomenon amplified by algorithmic content curation on digital platforms. Furthermore, media fragmentation enables selective exposure, reinforcing existing beliefs and deepening societal divides (Prior, 2013). These dynamics manifest across political elections, ethnic conflicts, religious tensions, and gender-based discourse, making automated polarization detection essential for understanding and monitoring online discourse.

From a computational perspective, polarization detection intersects with several established NLP tasks. Stance detection (Mohammad et al., 2016) identifies whether a text favors or opposes a target, while sentiment analysis (Rosenthal et al., 2017) captures affective polarity. However, polarization is broader than any of these: it encompasses not only overt hostility but also subtle mechanisms such as stereotyping, vilification, dehumanization, and deindividuation that create in-group/out-group divisions (Naseem et al., 2026b). This distinction motivates the need for dedicated polarization benchmarks.

SemEval-2026 Task 9 (Naseem et al., 2026a; Ghosh et al., 2026) addresses this gap by introducing the POLAR benchmark (Naseem et al., 2026b), a multilingual dataset spanning 22 languages with annotations for both binary polarization detection (Subtask 1) and multi-label classification of five polarization categories for English and Bengali (Subtask 2), respectively. The main contributions of this work can be summarized as follows:

- We demonstrate the effectiveness of Large Language Models (LLMs) with structured few-shot prompting for binary polarization detection across English and Bengali.
- We propose a multi-label polarization classification framework combining asymmetric loss, back-translation augmentation, and per-label threshold optimization.
- We design a marker-aware data augmentation pipeline for Bengali that preserves

*Authors contributed equally to this work.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2920–2931

domain-specific polarizing terms during back-translation, addressing severe class imbalance in low-resource settings.

2 Background

2.1 Task Description

SemEval-2026 Task 9 (Naseem et al., 2026a; Ghosh et al., 2026) defines polarization as the process in which opinions, beliefs, or behaviors become more extreme or divided, creating greater distance or conflict between groups. The task operationalizes this through six manifestations: stereotyping, vilification, dehumanization, deindividuation, intolerance, and invalidation of others' views. This study focuses on two major Subtasks as defined below.

Subtask 1: Polarization Detection. This Subtask[1] focuses on binary classification of social media text snippets, determining whether each instance exhibits *polarization (1)* or *lacks polarization (0)*.

Subtask 2: Polarization Type Classification. Building on Subtask 1, this Subtask[2] targets polarized texts and requires assigning binary labels (present/absent) across five possible polarization dimensions: (1) political/ideological, (2) racial/ethnic, (3) religious, (4) gender/sexual, and (5) other. As a multi-label task, instances may belong to multiple categories simultaneously.

2.2 Dataset Description

The POLAR benchmark (Naseem et al., 2026b) provides annotated datasets across 22 languages. We focus on English and Bengali. Both subtasks share the same data splits, summarized in Table 1.

Language	Train	Dev	Test
Bengali	3333	166	1501
English	3222	160	1452

Table 1: Dataset sizes per language (both subtasks).

A significant challenge in Subtask 2 is the severe class imbalance. Table 2 shows the imbalance ratios for the English training data.

[1] https://www.codabench.org/competitions/10522/
[2] https://www.codabench.org/competitions/10669/

Polarization Type	Positive %	Ratio
Political/Ideological	49.7%	1:1
Racial/Ethnic	3.5%	27.8:1
Religious	5.2%	18.2:1
Gender/Sexual	2.2%	43.8:1
Other	3.9%	24.6:1

Table 2: Class distribution for Subtask 2 labels in the English training set. Gender/sexual is most underrepresented at 43.8:1.

2.3 Related Work

Polarization research has deep roots in political science. DiMaggio et al. (1996) provided early empirical evidence of increasing attitude divergence in American public opinion, while Sunstein (2002) formalized group polarization theory, showing how deliberation amplifies pre-existing tendencies. Prior (2013) extended this to media consumption, demonstrating how selective exposure deepens partisan divides. Bail et al. (2018) further showed that exposure to opposing views on social media can paradoxically increase political polarization. Computationally, Conover et al. (2011) were among the first to study political polarization on Twitter through retweet and network analysis. Garimella et al. (2018) proposed graph-based methods for quantifying controversy on social media. More recently, Jiang et al. (2022) probed partisan worldviews directly from language model representations. In NLP, polarization detection has been approached through stance classification (Mohammad et al., 2016), sentiment analysis (Rosenthal et al., 2017), and hate speech detection (Basile et al., 2019). However, these tasks capture only fragments of the polarization phenomenon; they focus on individual opinions, affect, or explicit toxicity rather than the broader social dynamics of group division. The POLAR benchmark (Naseem et al., 2026b) addresses this gap by providing the first dedicated multilingual annotation scheme that covers six manifestations of polarization (stereotyping, vilification, dehumanization, deindividuation, intolerance, and invalidation) across 22 languages. Unlike prior studies that address polarization indirectly through stance detection, sentiment analysis, or hate speech identification, our work directly tackles the multi-dimensional polarization framework defined by the POLAR benchmark, jointly handling binary detection and fine-grained multi-

label type classification across languages. Furthermore, while existing computational approaches predominantly rely on network-based or monolingual methods, we combine LLM few-shot prompting with fine-tuned multilingual transformers, leveraging asymmetric loss and per-label threshold optimization to address the extreme class imbalance inherent in multi-label polarization categorization.

3 System Overview

To address the hierarchical polarization detection problem, this section outlines an efficient LLM-based approach that uses structured few-shot prompting for strong binary polarization detection (see Figure 1) and a fine-tuned transformer-based system with asymmetric loss and per-label threshold optimization for accurate multi-label polarization-type classification (see Figure 2). We chose LLM-based prompting for Subtask 1 because binary classification benefits directly from broad world knowledge and zero-shot multilingual reasoning, whereas Subtask 2 requires fine-grained discrimination among five severely imbalanced labels — a setting where task-specific fine-tuning with explicit class-imbalance handling consistently outperforms general-purpose prompting.

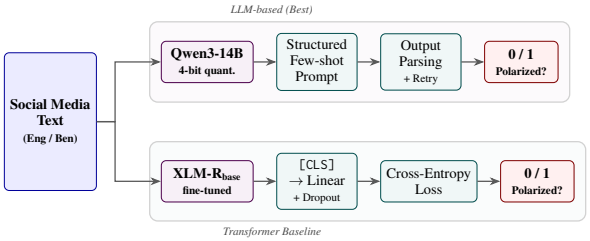

Figure 1: Subtask 1 architecture: binary polarization detection. The LLM branch (top) uses Qwen3-14B with structured few-shot prompting; the transformer branch (bottom) fine-tunes XLM-RoBERTa-base with a linear classification head.

3.1 Subtask 1: Polarization Detection

3.1.1 LLM-Based Approach (Best System)

We employed **Qwen3-14B**[3] with structured instruction prompting. The model is loaded with **4-bit quantization** via bitsandbytes, which reduces peak GPU memory from ∼28 GB to ∼8 GB, enabling inference on a single consumer GPU with negligible quality loss. We use `temperature=0.001`, `do_sample=False`, and `max_new_tokens=3` for deterministic output. The prompt includes: (1) a task definition explaining polarization and its six manifestations, (2) a scope clarification (what the task *is not* about), (3) two to four few-shot examples manually selected from the training set to cover balanced positive/negative cases and diverse polarization manifestations, with language-matched examples for each language, and (4) structured output formatting. We implemented a retry mechanism (up to 3 attempts) with a conservative fallback of 0 (non-polarized). The full prompt templates for both languages are shown in Appendix E.

[3] https://huggingface.co/Qwen/Qwen3-14B

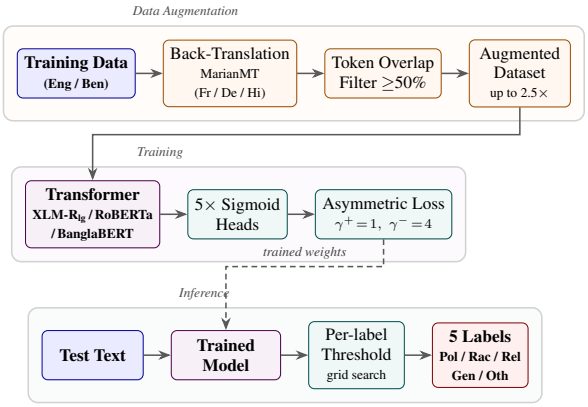

Figure 2: Subtask 2 architecture: multi-label polarization type classification. Training data is augmented via back-translation with token overlap filtering. The model uses asymmetric loss for class imbalance, and per-label thresholds are optimized on the dev set (Pol = Political, Rac = Racial/Ethnic, Rel = Religious, Gen = Gender/Sexual, Oth = Other).

3.1.2 Transformer-Based Models

As the primary transformer baseline, XLM-RoBERTa-base (Conneau et al., 2020) was fine-tuned by adding a linear classification head atop the `[CLS]` token with dropout ($p = 0.1$). The model is trained end-to-end with cross-entropy loss using AdamW (lr = 2e-5, weight decay of 0.01) for 10 epochs on a train (90%)/validation (10%) split. It yields macro F1-scores of 0.780 (accuracy 0.786) on English and 0.853 (accuracy 0.859) on Bengali.

3.2 Subtask 2: Polarization Type Classification

For multi-label classification, we replace softmax with five independent sigmoid outputs, one per polarization type, and train all models with Asymmetric Loss (ASL) (Ridnik et al., 2021).

3.2.1 Approaches

To address the multi-label classification challenge, we evaluated four transformer architectures.

- **XLM-RoBERTa-large** (Conneau et al., 2020): This architecture, trained on 8053 augmented samples for 6 epochs, achieves a dev-set macro F1-score of 0.827 after per-label threshold optimization.
- **RoBERTa-base** (Liu et al., 2019): Trained on 7569 samples (2738 original + 4831 augmented) for 14 epochs, achieving a dev-set macro F1-score of 0.824 after threshold optimization ($\Delta = +4.1$ over default 0.5 threshold).
- **BanglaBERT** (Bhattacharjee et al., 2022): Bengali model trained on 5591 samples (2833 original + 2758 augmented) for 10 epochs and achieved macro F1 of 0.921 in dev set.
- **DeBERTa-v3-small** (He et al., 2021): Trained on 6119 samples (70% augmentation), yields a macro F1 of 0.783 in dev set ($\Delta = +19.0$ from threshold tuning).

We also experimented with Qwen3-14B few-shot prompting for Subtask 2 using a structured multi-label prompt (Appendix E.3). On the English dev set, it achieved per-label macro F1 of 0.769 (micro F1: 0.5), with perfect scores on *religious, gender/sexual, and other* categories but only 0.400 on *political*; indicating that LLMs struggle with the most ambiguous category. However, fine-tuned transformers outperform this approach.

3.2.2 Asymmetric Loss

ASL (Ridnik et al., 2021) has been used across all models with $\gamma^+ = 1$, $\gamma^- = 4$, and clip margin m of 0.05, as shown in Eq. (1).

$$\mathcal{L}_{\text{ASL}} = \begin{cases} (1-p)^{\gamma^+} \log(p) & \text{if } y = 1 \\ p_m^{\gamma^-} \log(1-p_m) & \text{if } y = 0 \end{cases} \quad (1)$$

where $p_m = \max(p - m, 0)$ applies hard thresholding to negatives. The high γ^- aggressively down-weights easy negatives, critical given imbalance ratios up to 43.8:1 (*Gender* category). These hyper-parameters ($\gamma^+ \in \{0, 1, 2\}$, $\gamma^- \in \{2, 4, 6\}$, clip $\in \{0.0, 0.05, 0.1\}$) were selected via grid search on the dev set, with full details in Appendix A (Table A.1).

3.2.3 Data Augmentation

We apply back-translation augmentation using *Helsinki-NLP MarianMT*[4] models to address the severe class imbalance in minority labels (up to 43.8:1 ratio for gender/sexual), generating additional minority-class training examples by translating texts through an intermediate language (French, German, or Hindi) and back, filtering by $\geq$50% token overlap with the original. Table 3 outlines the augmented training sizes.

ST	Lang	Original	Augmented
1	Bengali	3333	5726
	English	3222	5726
2	Bengali	3333	6091
	English	3222	8053

Table 3: Original vs. augmented training set sizes.

3.2.4 Per-Label Threshold Optimization

We optimized per-label thresholds (τ_k) on the dev set by searching $[0.05, 0.95]$ with step 0.02. Optimal thresholds vary widely, i.e., *political* requires 0.35 (balanced class), while *gender/sexual* requires 0.77 (extremely rare). This yields $\Delta = +4.13$ F1 for RoBERTa and $\Delta = +24.6$ for XLM-RoBERTa-large. Full values are detailed in Appendix F. The implementation and source code are publicly available on GitHub[5].

4 Experimental Setup

Training Details: All transformer models use AdamW with learning rate of 2e-5 (encoder) and 5e-4 (head), batch size of 16, gradient accumulation of 4 (effective batch 64), max sequence length of 256, warmup ratio 0.1, and weight decay 0.01. Models are selected by the best dev macro F1 across epochs. Hyperparameter search details are in Appendix A.

Software: PyTorch[6] with HuggingFace Transformers;[7] bitsandbytes[8] for 4-bit quantization; Helsinki-NLP MarianMT for augmentation. All experiments run on a single NVIDIA GPU.

[4] https://huggingface.co/Helsinki-NLP
[5] https://github.com/SM-Shaan/Polarity_SemEval-26
[6] https://pytorch.org/
[7] https://huggingface.co/docs/transformers
[8] https://github.com/TimDettmers/bitsandbytes

Evaluation: Macro F1-score is the official metric for both subtasks (detailed in Appendix B).

5 Results and Discussion

5.1 Subtask 1: Polarization Detection

Table 4 presents dev and test results across all models and subtasks. For Subtask 1, Qwen3-14B achieves the best test performance with **0.808** [English (Eng), rank 9th] and **0.847** [Bengali (Ben), rank 4th], outperforming other approaches.

ST	Model	Lang	Dev	Test	Rank
1	*RemBERT (baseline)*	*Eng*	—	*.694*	—
	RemBERT (baseline)	*Ben*	—	*.754*	—
	XLM-R_{base}	Eng	.780	.762	—
		Ben	.853	.821	—
	$RoBERTa_{base}$	Eng	.751	.738	—
	RoBERTa (no aug)	Eng	.734	.719	—
	DeBERTa-v3	Eng	.726	.711	—
	Qwen3-14B	Eng	.795	**.808**	9th
		Ben	.838	**.847**	4th
2	*RemBERT (baseline)*	*Eng*	—	*.385*	—
	RemBERT (baseline)	*Ben*	—	*.194*	—
	XLM-R_{large}	Eng	.827	**.454**	21st
	$RoBERTa_{base}$	Eng	.824	.438	—
	RoBERTa (no aug)	Eng	.790	.419	—
	DeBERTa-v3	Eng	.783	.352	—
	Qwen3 (few-shot)	Eng	.769	.381	—
	BanglaBERT	Ben	.921	.189	—
	RoBERTa (merged)	Ben	—	**.206**	26th

Table 4: Overview of the performance (macro F1) in both subtasks (ST). Dev scores use optimized thresholds for Subtask 2. Ranks are official leaderboard positions. Organizer baseline (RemBERT, monolingual) scores are from Naseem et al. (2026b).

LLMs Outperform Fine-Tuned Transformers. Despite XLM-RoBERTa-base achieving 0.853 macro F1 on Bengali dev set, Qwen3-14B outperforms it on test (0.847 vs. 0.821 Ben; 0.808 vs. 0.762 Eng). All our systems also exceed the organizer's RemBERT monolingual baseline (0.754 Ben, 0.694 Eng (Naseem et al., 2026b)). Among Subtask 1 transformers, RoBERTa (0.751) and DeBERTa-v3 (0.726) also trail XLM-R on dev. This suggests that LLMs' broad world knowledge and multilingual reasoning are beneficial for binary polarization detection, which requires cultural and political context beyond what small training sets provide.

Asymmetric Error Patterns Across Languages. On English dev ($n = 323$), XLM-RoBERTa produces 49 false positives (FP) vs. 20 false negatives (FN) (21.4% error), over-predicting polarization. On Bengali dev ($n = 334$), this reverses: 17 FP vs. 30 FN (14.1% error), under-predicting polarization. Bengali achieves higher precision (0.861 vs. 0.777) while English achieves higher recall for polarized texts.

5.2 Subtask 2: Polarization Type Classification

Table 5 compares dev and test performance across all Subtask 2 approaches against the organizer's RemBERT baseline (0.385 Eng, 0.194 Ben). XLM-RoBERTa-large achieves the best English test F1 of 0.454, surpassing the baseline by +6.9 points, while RoBERTa (merged) at 0.206 marginally exceeds the Bengali baseline of 0.194.

Model	Lang	Dev	Test	Δ
RemBERT (baseline)	*Eng*	—	*.385*	—
RemBERT (baseline)	*Ben*	—	*.194*	—
XLM-R-large	Eng	.827	**.454**	−.373
RoBERTa	Eng	.824	.438	−.386
RoBERTa (no aug)	Eng	.790	.419	−.371
BanglaBERT	Ben	.921	.189	−.732
RoBERTa (merged)	Ben	—	**.206**	—
DeBERTa-v3	Eng	.783	.352	−.431
Qwen3 (few-shot)	Eng	.769	.381	−.388

Table 5: Subtask 2 macro F1 on dev (optimized thresholds) and test. Δ = test − dev. Organizer baseline (RemBERT, monolingual) from Naseem et al. (2026b) shown at top. All submitted models show large dev-to-test degradation.

Severe Dev-to-Test Drop. XLM-RoBERTa-large achieves 0.827 on dev but only 0.454 on test ($\Delta = -0.373$). BanglaBERT shows the most extreme drop from 0.921 to 0.189 ($\Delta = -0.732$), while DeBERTa has the largest English drop ($\Delta = -0.431$). Qwen3 few-shot also degrades from 0.769 to 0.381 ($\Delta = -0.388$). This consistent, cross-model degradation points to two compounding factors. First, per-label threshold optimization directly maximizes dev macro F1 and is thus prone to overfitting the dev label distribution; when the test set has different class frequencies or boundary cases, the optimized thresholds become miscalibrated. Second, the Bengali dev set contains very few positive samples for minority labels (e.g., gender/sexual, racial/ethnic), making dev F1 artificially high — a single correctly predicted positive inflates

Label	Thresh.	F1	P / R
Political	0.35	.741	.656 / .850
Racial/Ethnic	0.65	.732	.750 / .714
Religious	0.65	.895	.810 / 1.00
Gender/Sexual	0.77	.952	1.00 / .909
Other	0.45	.800	.692 / .947
Macro avg (optimized)		**.824**	
Macro avg (default 0.5)		.783	

Table 6: Per-label dev F1 for RoBERTa (English) with optimized thresholds (Thresh.).

recall to 1.0. The fact that even Qwen3 few-shot (which uses no thresholds) shows a similar Δ of -0.388 indicates that a genuine label-distribution shift between dev and test is also present, independent of threshold overfitting.

Per-Label Performance Is Highly Uneven. Table 6 shows per-label dev F1 for RoBERTa. *Religious* (0.895) and *gender/sexual* (0.952) achieve near-perfect scores due to distinctive markers, while *political* achieves only 0.741 despite being the most frequent class.

LLM Few-Shot Falls Short on Multi-Label. Qwen3-14B achieves 0.769 macro F1 on Subtask 2 dev and 0.381 on test, below all fine-tuned transformers on both splits. Per-label breakdown reveals perfect scores on *religious*, *gender/sexual*, and *other* (F1 = 1.0), but only 0.400 on *political* and 0.444 on *racial/ethnic*. This may appear inconsistent with the LLM's strong Subtask 1 performance: if world knowledge helps binary detection, why does it fail here? The distinction is task structure — Subtask 1 requires a single global judgment (polarized or not) where cultural context is decisive, whereas Subtask 2 requires simultaneously distinguishing five fine-grained, co-occurring labels, including the highly ambiguous *political* category that accounts for nearly half of all positive training instances. For such fine-grained multi-label discrimination, task-specific training signal from fine-tuned transformers proves essential.

Bengali Dev Is Deceptively High. BanglaBERT achieves 0.921 dev F1 with perfect scores on *racial/ethnic*, *religious*, and *gender/sexual* (all 1.000), but these categories have very few positive samples in the Bengali dev set, inflating per-label F1. The test score of 0.189 ($\Delta = -0.732$) confirms dev numbers are unreliable; the submitted RoBERTa (merged) achieves a slightly higher 0.206 on test.

6 Conclusion

This study presented a polarization detection and classification system for SemEval-2026 Task 9 in a hierarchical multilingual setup. In Subtask 1, Qwen3-14B with structured few-shot prompting achieves test macro F1 of 0.847 (Bengali, rank 4$^{\text{th}}$) and 0.808 (English, rank 9$^{\text{th}}$), outperforming fine-tuned transformers (XLM-RoBERTa-base: 0.821 Ben, 0.762 Eng) and the organizer RemBERT baseline (0.754 Ben, 0.694 Eng), demonstrating that LLMs with structured prompts effectively leverage cultural and political world knowledge for binary detection. For Subtask 2, our framework combining asymmetric loss, back-translation augmentation, and per-label threshold optimization achieves test macro F1 of 0.454 (English, rank 21$^{\text{th}}$) and 0.206 (Bengali, rank 26$^{\text{th}}$), surpassing the RemBERT baseline (0.385 Eng, 0.194 Ben). Key findings: (1) LLM few-shot prompting is highly effective for binary polarization detection, especially for Bengali; (2) per-label threshold optimization is powerful yet prone to dev overfitting (Δ up to -0.732 on test); and (3) back-translation augmentation provides modest but consistent gains. Future work includes cross-lingual transfer to additional POLAR languages, distribution-robust threshold calibration, and ensemble strategies combining LLM and transformer predictions.

Limitations

Despite several positive insights, this study has limitations that should be acknowledged. First, our participation was limited to only two of the 22 languages covered by SemEval-2026 Task 9 (POLAR): English and Bengali. Second, LLM experiments are limited to Qwen3-14B; no zero-shot baselines or alternative LLMs (e.g., Gemma, Phi-4, Llama) are evaluated due to compute constraints. Third, back-translation augmentation shows minimal impact; the gap between augmented and non-augmented RoBERTa on Subtask 2 dev is only +3.4 points (0.824 vs. 0.790). Fourth, the severe dev-to-test degradation across Subtask 2 (Δ ranging from -0.371 to -0.732) indicates that per-label threshold optimization overfits to dev and needs regularization. Finally, for Subtask 1, only XLM-RoBERTa-base was fully evaluated per language; other transformers (RoBERTa, DeBERTa-v3) were tested on English only without Bengali breakdown.

Ethical Considerations

Our work focuses on detecting polarized content on social media related to sensitive topics. Automated polarization detection could be misused to censor or suppress legitimate dissent. The models we develop should be viewed as research tools. We follow the shared task's data usage guidelines throughout.

Acknowledgments

We thank the organizers of SemEval-2026 Task 9 for providing the POLAR benchmark and evaluation infrastructure. Experiments were conducted using computing resources at the CUET NLP Lab.

References

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of SemEval-2019*, pages 54–63.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of NAACL*, pages 1318–1327.

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. *Proceedings of ACL*, pages 8440–8451.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have Americans' social attitudes become more polarized? *American Journal of Sociology*, 102(3):690–755.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of ICLR*.

Ruibo Jiang, Xinning Hua, Valerie Berger, Samuel Barkley, Jonathan Bernstein, Lu Xiao, and Jiebo Luo. 2022. CommunityLM: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*.

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*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of SemEval-2016*, pages 31–41.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, 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 Acarürk, 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. *Preprint*, arXiv:2505.20624.

Markus Prior. 2013. Media and political polarization. *Annual Review of Political Science*, 16:101–127.

Tal Ridnik, Emanuel Ben-Baruch, Niv Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of ICCV*, pages 82–91.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of SemEval-2017*, pages 502–518.

Cass R. Sunstein. 2002. The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195.

Appendix

A Hyperparameter Settings

For the LLM-based approach (Subtask 1), Qwen3-14B was loaded with 4-bit quantization via the bitsandbytes framework to ensure memory-efficient inference on a single GPU. We used greedy decoding with `temperature=0.001`, `do_sample=False`, and `max_new_tokens=3` to ensure deterministic label predictions. A retry mechanism (up to 3 attempts) was implemented with a conservative fallback of 0 (non-polarized) for unparseable outputs.

For the transformer fine-tuning approach (Subtask 2), Table A.1 shows the hyperparameter search space for RoBERTa-base, selected by best dev macro F1 across epochs.

Hyperparameter	Range	Best
LR (encoder)	{5e-6, 1e-5, 2e-5}	2e-5
LR (head)	{1e-4, 5e-4, 1e-3}	5e-4
Batch size	{8, 16, 32}	16
Max seq. length	{128, 200, 256}	256
Dropout	{0.1, 0.2, 0.3}	0.2
Weight decay	{0.001, 0.01, 0.1}	0.01
Warmup ratio	{0.0, 0.1, 0.2}	0.1
ASL γ_{neg}	{2, 4, 6}	4
ASL γ_{pos}	{0, 1, 2}	1
ASL clip	{0.0, 0.05, 0.1}	0.05
Grad. accum.	{1, 2, 4}	4

Table A.1: Hyperparameter search for Subtask 2 (RoBERTa-base). Effective batch size: $16 \times 4 = 64$.

Training was conducted with the AdamW 8-bit optimizer using differential learning rates, i.e., 2e-5 for the encoder and 5e-4 for the classification head. A linear learning rate scheduler with a warmup ratio of 0.1 was applied. Models were trained for up to 14 epochs with early stopping based on dev macro F1. Gradient checkpointing was enabled to reduce memory consumption.

B Evaluation Metric

System performance was evaluated separately for the two subtasks using the macro F1-score as the official evaluation metric. For both tasks, evaluation is based on the macro F1-score, which computes the F1-score independently for each class and then averages them equally across all classes, as outlined in Eq. (B.1):

$$\text{Macro-}F_1 = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{B.1}$$

where P_c and R_c denote the precision and recall for class c, and C represents the total number of classes ($C = 2$ for Subtask 1, $C = 5$ for Subtask 2). The use of macro F1-score ensures balanced evaluation across all categories, giving equal importance to both frequent and rare classes.

C Prediction Examples

Table C.1 presents representative prediction examples from our models on the dev set. In Subtask 1, the XLM-RoBERTa model correctly identifies explicit political vilification but occasionally overpredicts polarization for texts that contain political entities without hostile framing. In Subtask 2, the model captures dominant categories well (e.g., political) but struggles with co-occurring labels; for example, texts involving both *racial* and *religious* dimensions are sometimes classified with only one label.

D Augmented Dataset Samples

Table D.1 presents representative examples from our data augmentation pipeline. For Subtask 1 (English), we apply synonym replacement and random word swap, preserving protected domain-specific words (e.g., political party names, slurs, identity terms). For Subtask 2 (English), we use back-translation via *Helsinki-NLP MarianMT* through intermediate languages (French, German, Hindi) with $\geq$50% token overlap filtering. For Bengali Subtask 1, we implement marker-aware augmentation that preserves 40 polarizing marker words, including political party references, derogatory labels, and accusatory terms, during synonym replacement and word reordering.

E Prompts Used for Few-Shot Classification

We designed structured few-shot prompts for both subtasks in English and Bengali. Each prompt includes: (1) a task definition explaining polarization and its six manifestations, (2) a scope clarification (what the task *is not* about), (3) 2–4 few-shot examples drawn from the training set (language-matched), and (4) structured output formatting.

Text	Gold	Pred	Task
Subtask 1: Binary polarization classification (polarized or not-polarized)			
"Musk, for his part, explained how the president is the embodiment of the nation and that resisting his orders is the same as thwarting the will of the people." Ur Fascism property 13, Selective Populism	1	1	ST1
"bad people" I have some conservative values so does that make me one of those Paul? Honest question.	0	1	ST1
"denazification" as an excuse to explain Russias deployment of troops to Ukraine, alleging that Hitler was of Jewish ancestry	1	1	ST1
Jamie Raskin talks impeachment, the future of democracy	0	0	ST1
Subtask 2: multi-label predictions (political, racial, religious, gender, other)			
speaking FACTS. Elites only care about border security when its at their doorstep.	1,0,0,0,0	1,0,0,0,0	ST2
"pro israel, anti genocide" as a political stance is like "pro ignition, anti fire". The existence of israel is based entirely on settler colonial displacement...	1,1,1,0,0	1,1,0,0,0	ST2
Fancy that. A he/him deciding that gender ideology takes priority over victims, womens rights and the law.	1,0,0,1,0	0,0,0,1,0	ST2

Table C.1: Representative prediction examples from dev set evaluation. ST1 uses XLM-RoBERTa-base (binary); ST2 uses RoBERTa-base with optimized thresholds (multi-label). The second ST1 example shows a false positive: conservative values discussion without hostile framing. The ST2 examples show correct dominant-label capture but missed co-occurring labels.

Original Text	Augmented Text	Label	Method
How superstars are silently propagating extreme right wing politics through their movies!	How right propagating silently are extreme superstars wing politics through movies! their	Polarized	Swap
Early voting ends tomorrow. Elections have consequences. GoVoteNow	Early voting GoVoteNow tomorrow. Elections consequences. have ends	Not Pol.	Swap
Congress just passed a bill taking voting rights away from women. Lets hope it doesnt pass in the Senate	congress just passed a bill taking voting rights away from women. lets hope it did roll in the senate	Gender	BT
How surprising. Our govt is still "friends" of this genocidal, apartheid state.	how amazing. our govt is still "wary" of this genocidal, apartheid state.	Racial	BT

Table D.1: Example original and augmented text pairs. Swap = random word swap (Subtask 1); BT = back-translation via MarianMT (Subtask 2). Protected words (political entities, identity terms) are preserved during synonym replacement.

The full prompt templates are described in the following subsections.

E.1 English Subtask 1 Prompt

To enable reliable and deterministic binary polarization detection with Qwen3-14B, we designed a structured few-shot instruction prompt that explicitly defines polarization, clarifies task scope to avoid common misinterpretations, and includes 4 carefully selected in-domain examples. Figure E.1 presents the full English prompt template used.

E.2 Bengali Subtask 1 Prompt

Figure E.2 shows the Bengali binary classification prompt. It uses the same structural template as the English version but includes an extended definition section with Bangla-specific context. The few-shot examples are drawn from the Bengali training set.

E.3 English Subtask 2 Prompt

Figure E.3 shows the English multi-label classification prompt with 4 few-shot examples covering different label combinations.

English Subtask 1 Prompt

```
Below is a binary classification task for English text
polarization detection. Determine whether the text contains
polarized content (1) or is non-polarized (0).

## Definition of Polarization
Polarization denotes: Stereotyping, Vilification,
Dehumanization, Deindividuation, Intolerance of other
people's views, beliefs, and identities.

## Task Scope
This task focuses on: Assessing whether social media messages
reflect attitude polarization. This task is NOT about: The
opinion evoked in the reader, the opinion of someone else
mentioned in the text, or the true opinion of the speaker.

## Instructions
1. Language: English. Output: 0 or 1 only.
2. 0 = Non-polarized (neutral, informative).
   1 = Polarized (division, hatred).

## Examples
Text: Satellite images confirm civilian massacres during
      Russian military occupation
Answer: 1
Text: We were talking about the ukraine invasion and nobody
      genuinely cares.
Answer: 0
Text: These leftist idiots are destroying our country with
      their woke agenda
Answer: 1
Text: The new policy on renewable energy was announced
      yesterday by the government.
Answer: 0

## Your Task
Text: {input_text}
Answer:
```

Figure E.1: English Subtask 1 prompt for Qwen3-14B with 4 few-shot examples covering both polarized and neutral political content.

E.4 Bengali Subtask 2 Prompt

The Bengali Subtask 2 prompt follows the same multi-label structure as the English version. It includes the extended polarization definition with attitude polarization and the six manifestations, task scope clarification, and Bengali few-shot examples drawn from the training set. The output format is identical: 5 comma-separated binary values in the order *political, racial/ethnic, religious, gender/sexual, and other*. Bengali examples emphasize local political contexts (e.g., party-based polarization involving BNP and Awami League) and religious tensions common in Bangladeshi social media discourse.

F Threshold Optimization Details

Table F.1 shows per-label optimized thresholds for the English Subtask 2 models, obtained by grid search over $[0.05, 0.95]$ with step 0.02. XLM-RoBERTa-large benefits most ($\Delta = +24.6$ F1 over default 0.5). For Bengali, BanglaBERT uses thresholds: *political* = 0.63, *racial* = 0.39, *religious* = 0.41, *gender* = 0.21, *other* = 0.87.

Bengali Subtask 1 Prompt

```
Below is a binary classification question in Bangla from a
Polarization Detection dataset. You need to determine whether
a post contains polarized content (1) or is not polarized (0).

## Definition of Polarization
Polarization is the sharp division of opinions into opposing
groups, often with hostility and exclusion.
**Attitude Polarization**: The negative attitude that
individuals display towards out-groups while showing blind
support for in-groups.
**Polarization denotes**: Stereotyping, Vilification,
Dehumanization, Deindividuation, Intolerance of other
people's views, beliefs, and identities.

## Task Scope
This task focuses on: Assessing whether social media messages
reflect attitude polarization. This task is NOT about: The
opinion evoked in the reader, the opinion of someone else
mentioned in the text, or the true opinion of the speaker.

## Instructions
1. Language: Bangla. Answer 0 or 1.
2. Only texts that clearly reflect attitude polarization
   should be labeled 1.
3. Context Matters: Consider the overall meaning, not just
   individual words.

## Few-Shot Examples
Question: [Bengali text: journalist accusing political party
  of arson, blaming Hefazat, calling Awami League supporters
  boot-lickers]
Answer: 1
Question: [Bengali text: neutral statement about waiting and
  observing which editors and intellectuals will stand by them]
Answer: 0

## Your Task
Question: {input_text}
Answer:
```

Figure E.2: Bengali Subtask 1 prompt for Qwen3-14B with 2 few-shot examples. Actual Bengali text examples are drawn from the training set; English glosses shown here for readability.

Label	**RoBERTa**		**XLM-R-large**	
	Thr.	**F1**	**Thr.**	**F1**
Political	0.35	.741	0.47	.712
Racial	0.65	.732	0.51	.674
Religious	0.65	.895	0.65	.872
Gender	0.77	.952	0.73	1.00
Other	0.45	.800	0.69	.878
Macro (opt.)		**.824**		**.827**
Macro (0.5)		.783		.581

Table F.1: Per-label thresholds and F1-scores on English dev set. XLM-R-large benefits most from threshold optimization ($\Delta = +24.6$ over default 0.5).

English Subtask 2 Prompt

```
Below is a multi-label classification task in English
from a Polarization Detection dataset. Classify the
text into 5 polarization categories (each 0 or 1).

## Polarization Categories
1. political: Division based on political parties,
   ideologies, government policies
2. racial/ethnic: Division based on race, ethnicity,
   nationality, cultural background
3. religious: Division based on religious beliefs,
   practices, or groups
4. gender/sexual: Division based on gender identity,
   sexual orientation, gender roles
5. other: Polarization not covered by the above

## Instructions
1. Multi-label: A text can belong to multiple categories.
2. Output: Exactly 5 comma-separated values (0 or 1)
   in order: political,racial/ethnic,religious,
   gender/sexual,other
3. If no polarization: 0,0,0,0,0

## Examples
Text: The Democrats are destroying this country with
  their radical socialist agenda.
Labels: 1,0,0,0,0
Text: These immigrants are taking our jobs and ruining
  our culture. Send them back!
Labels: 0,1,0,0,0
Text: A victory for voting rights in Wisconsin! The
  Supreme Court decided to preserve the ability of
  cities to designate in-person absentee voting sites.
Labels: 0,0,0,0,0
Text: Muslims are all terrorists and should be banned
  from entering the country.
Labels: 0,1,1,0,0

## Your Task
Text: {input_text}
Labels:
```

Figure E.3: English Subtask 2 multi-label prompt for Qwen3-14B with 4 few-shot examples. Examples cover single-label (political), single-label (racial), non-polarized, and multi-label (racial + religious) cases.