PolaFusion at SemEval-2026 Task 9: Ensemble Transformers with Targeted Augmentation for Multilingual Polarization Detection

Abdullah Mohammad
Delhi Skill and Entrepreneurship University (DSEU)
Okhla, New Delhi, India
abdulla90t@gmail.com

Abstract

We present **PolaFusion**, our system for SemEval-2026 Task 9, which requires detecting polarization in social media posts across 22 languages, classifying its type (Subtask 2), and identifying its rhetorical manifestation (Subtask 3). The task is characterized by severe and pervasive class imbalance across all three subtasks and all 22 languages. We address this through a combination of three strategies: a hierarchical gating architecture where a binary gatekeeper model gates two specialist classifiers trained exclusively on polarized content; an eight-model mega-ensemble combining five-fold mDeBERTa-v3-base (He et al., 2021) and three-fold XLM-RoBERTa-large (Conneau et al., 2020) with soft-vote probability aggregation; and a Macro-F1-aware augmentation strategy using Qwen3-235B (Yang et al., 2025) that generates synthetic minority-class examples only for language-label pairs that are both scarce and poorly learned. Throughout training, inverse-frequency class weighting within `BCEWithLogitsLoss` forces the model to attend proportionally to rare labels. Our system achieves official Macro-F1 scores of **0.800**, **0.576**, and **0.502** on Subtasks 1–3 respectively, outperforming the POLAR baseline (Naseem et al., 2026b) by +0.040, +0.089, and +0.082 average Macro-F1 across languages. Our code is publicly available at https://github.com/Abdullah4152/PolaFuse.

1 Introduction

Polarization in online discourse operates through many distinct mechanisms: political sloganeering, ethnic dehumanization, religious vilification, gendered invalidation, and subtle forms of empathy denial that are no less corrosive for being harder to identify. Characterizing these phenomena computationally at scale—and across language boundaries—is both socially urgent and technically demanding (Sunstein, 2018; Bail et al., 2018).

SemEval-2026 Task 9 (Naseem et al., 2026a,b) structures this challenge as a three-layer prediction problem over 22 languages, evaluated with Macro-F1—a choice that forces systems to confront extreme label imbalance rather than ignore it.

That imbalance is the central difficulty of this task. In Subtask 1, the proportion of polarized posts ranges from under 11% (Hausa) to over 90% (Khmer). In Subtasks 2 and 3, certain labels—*Dehumanization* in Odia, *Lack of Empathy* in Hausa, *Racial/Ethnic* in Bengali—appear in well under 2% of training instances. Standard cross-entropy training on such distributions learns to ignore these labels entirely, collapsing to high accuracy but near-zero Macro-F1.

We respond to this with **PolaFusion**, a system designed from the ground up around the imbalance problem. We describe a hierarchical gating architecture, an eight-model cross-lingual ensemble, a targeted LLM-based augmentation strategy, and a cost-sensitive training objective that encodes class imbalance directly into the loss function. We show through careful ablation that each component contributes, and that some commonly used techniques (per-label threshold tuning) actively hurt generalization in this data regime.

2 Related Work

Online polarization detection builds on hate speech and toxic language classification. Shared tasks such as OffensEval (Zampieri et al., 2019), HaSpeeDe (Sanguinetti et al., 2020), and SemEval tasks on misogyny identification (Fersini et al., 2022) established paradigms for content-level hostility detection, while HateXplain (Mathew et al., 2021) added explainability. Multilingual extensions (Ousidhoum et al., 2019; Leite et al., 2020) showed that cross-lingual transfer partially bridges annotation gaps but struggles with culturally situated bias.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2877–2885

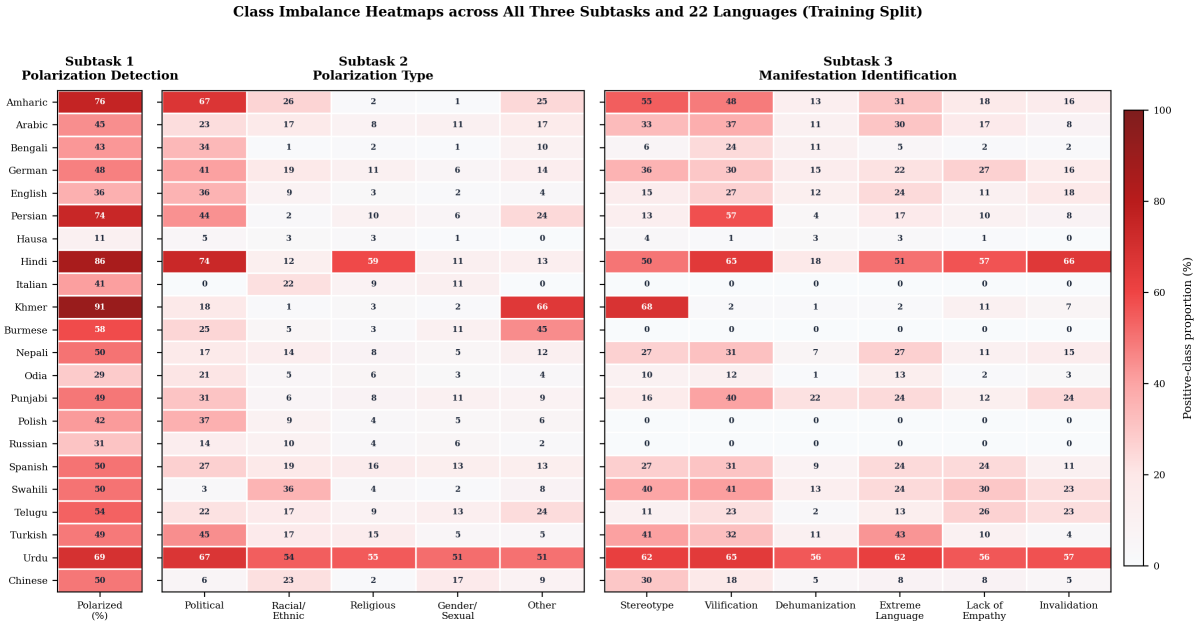

Figure 1: Class imbalance heatmaps for all three subtasks and 22 languages. Each cell shows the positive-class proportion (%). Darker red = more positive instances; near-white = severe label scarcity. Rows are aligned across panels so imbalance can be compared directly.

On the modeling side, cross-validated ensembles of multilingual encoders consistently outperform single models by reducing correlated errors (Dietterich, 2000). Data augmentation has evolved from EDA (Wei and Zou, 2019) and SMOTE (Chawla et al., 2002) toward LLM-based generation conditioned on label semantics. Our work integrates these strategies, targeting augmentation only where model performance warrants it.

3 Task and Data

Subtask 1 is binary classification over all posts: polarized (1) or non-polarized (0). Subtask 2 is multi-label classification into five types—*Political*, *Racial/Ethnic*, *Religious*, *Gender/Sexual*, and *Other*—applied only to polarized posts. Subtask 3 is multi-label classification into six rhetorical manifestations—*Stereotype*, *Vilification*, *Dehumanization*, *Extreme Language*, *Lack of Empathy*, and *Invalidation*—likewise applied only to polarized posts. The underlying data is drawn from the POLAR benchmark (Naseem et al., 2026b), a multilingual, multicultural corpus covering 22 languages.

Class imbalance. Figure 1 presents a unified heatmap of the positive-class proportion across all three subtasks, all 22 languages, and all labels in the training split. For Subtask 1, the spread is extreme: Hausa has only 10.7% polarized posts while Khmer reaches 90.8%. In Subtask 2, *Political* dominates most languages—Hindi reaches 74%, Urdu 67%—while *Gender/Sexual* and *Religious* sit below 15% for the majority, and Italian has zero *Political* training examples entirely. Subtask 3 exposes the most severe imbalance: *Dehumanization*, *Lack of Empathy*, and *Invalidation* are near-absent in Hausa (3%, 1%, and 0% respectively) and Odia (1%, 2%, 3%). Italian, Polish, Russian, and Burmese show all-zero Subtask 3 cells, as these languages were not annotated for manifestations in the training split. This heatmap directly motivates our cost-sensitive loss and targeted augmentation strategy.

4 System

Figure 2 provides an overview of the PolaFusion architecture.

4.1 Hierarchical Gating

The Subtask 2 and 3 specialists are trained exclusively on the *polarized* portion of the training data, removing the diluting effect of all-zero label vectors from non-polarized instances. At inference, posts predicted as non-polarized by the gatekeeper receive all-zero Subtask 2 and 3 labels without invoking the specialists. A forced-argmax correction handles the residual edge case: if a post is predicted as polarized but all specialist probabilities

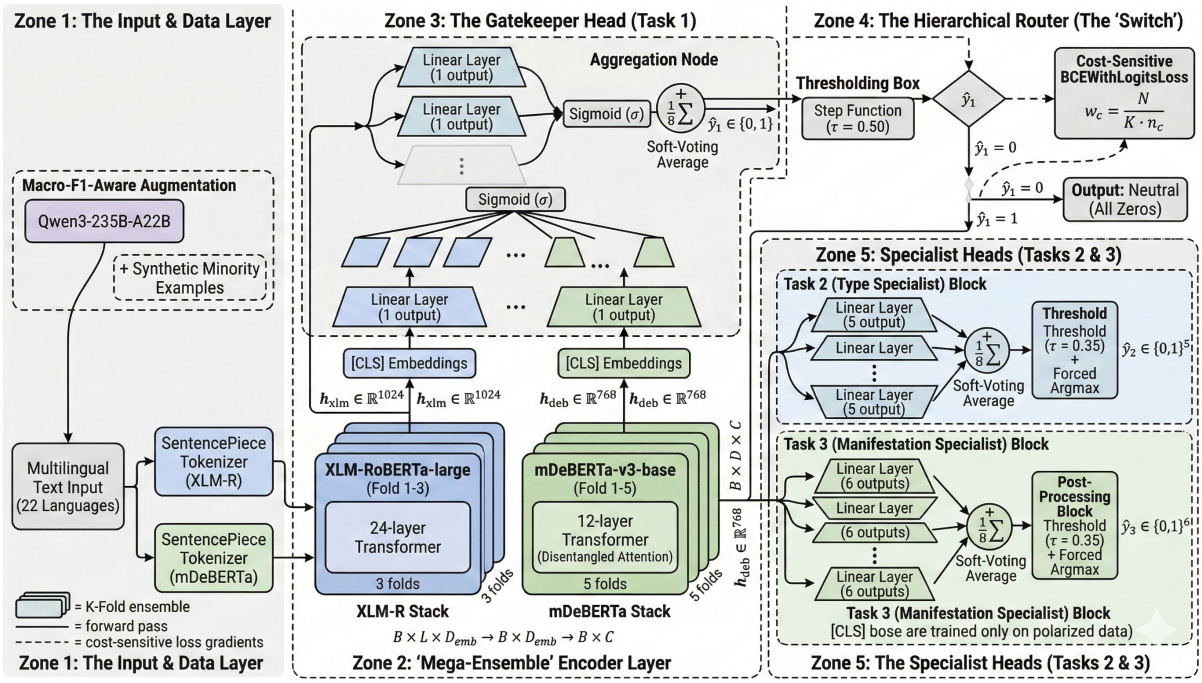

Figure 2: PolaFusion architecture. The Gatekeeper (Subtask 1) routes each post: non-polarized posts receive all-zero outputs immediately; polarized posts are passed to the Type and Manifestation specialists, which are trained exclusively on polarized instances.

fall below threshold, the label with the highest predicted probability is forced to 1, ensuring label consistency. We note that training the Subtask 2 and 3 specialists on the full dataset (including non-polarized instances) degraded performance in preliminary experiments, as the all-zero vectors dominated gradient updates and suppressed learning of minority labels.

4.2 Mega-Ensemble

PolaFusion combines two cross-lingual transformer families. mDeBERTa-v3-base (He et al., 2021) employs disentangled attention over token content and relative position, benefiting morphologically complex languages. XLM-RoBERTa-large (Conneau et al., 2020) offers larger model capacity trained on a significantly larger multilingual corpus, helping low-resource languages where mDeBERTa may lack coverage. We train five cross-validated folds of mDeBERTa and three of XLM-R-large, yielding eight models per subtask. The choice of five folds for mDeBERTa and three for XLM-R reflects a compute–diversity trade-off: mDeBERTa is lighter and benefits from additional fold diversity, while XLM-R-large is computationally expensive and already provides substantial capacity per fold. Final predictions are produced by averaging raw sigmoid probabilities across all eight models before thresholding—soft voting that smooths calibration without requiring learned weighting.

4.3 Handling Class Imbalance

We address imbalance through two complementary mechanisms applied at training time.

Cost-sensitive BCE loss. For each label c, we compute an inverse-frequency class weight:

$$w_c = \frac{N}{K \cdot n_c} \tag{1}$$

where N is the total training instances, $K = 2$ for binary BCE labels, and n_c is the positive count for class c. These weights are passed to `BCEWithLogitsLoss` via `pos_weight`. For Hausa *Invalidation* ($n_c = 9$ out of 3,651 instances), $w_c \approx 202$; well-balanced labels receive $w_c \approx 1.0$. This handles the full imbalance spectrum without resampling or the interpolation artifacts of SMOTE (Chawla et al., 2002), and avoids the tunable γ of focal loss (Lin et al., 2017).

Macro-F1-aware augmentation. We generate synthetic training examples using Qwen3-235B-A22B (Yang et al., 2025), but only when two conditions hold simultaneously: the class is underrepresented (minority/majority ratio below 0.35 for Subtask 1, or label/max-label ratio below 0.25

for Subtasks 2 and 3) *and* the observed per-label Macro-F1 on the development set falls below a threshold (0.60 for Subtask 1, 0.40 for Subtasks 2 and 3). The ratio thresholds correspond to the point below which inverse-frequency weighting alone fails to produce non-zero recall; the F1 thresholds mark the level below which a label is clearly being missed. These values were not extensively tuned and serve as conservative gates.

The joint condition concentrates augmentation where it can make a difference—a class that is rare but easy to learn does not benefit, while one that is rare *and* being missed does. The number of synthetic examples is:

$$s = \min\big(0.8 \times n_{\max},\ 2 \times n_c\big) - n_c \qquad (2)$$

where $n_{\max}$ is the most frequent label's count, capping synthetic output to prevent overrepresentation. For example, Hausa Subtask 1 (F1 = 0.760, ratio = 0.120) triggered generation of 412 synthetic polarized examples from few-shot seeds. To illustrate, consider the following English *Gender/Sexual* training instance and its Qwen3-generated synthetic variation:

> **Original:** *"She does look like a Fox News Stepford Wife."* **[Political, Gender/Sexual]**
>
> **Synthetic:** *"She really does resemble a brainwashed Fox News Barbie doll."* **[Political, Gender/Sexual]**

The synthetic post preserves the stance, target, and gendered register while substantially varying surface form, consistent with our prompt constraints.

Unlike template-based methods such as EDA (Wei and Zou, 2019), augmentation is conditioned on observed development-set F1, ensuring synthetic effort is concentrated only where the model is actively failing. The LLM is prompted with the target language, label definition, and few-shot seeds from the training set, producing contextually appropriate synthetic posts. We note that the type and manifestation labels available for Subtasks 2 and 3 could in principle also inform Subtask 1 augmentation by conditioning synthetic examples on specific polarization types; we leave this as future work.

4.4 Training Details

All models are fine-tuned with AdamW (Loshchilov and Hutter, 2017) at 2×10^{-5} learning rate, batch size 16, maximum sequence length 128 tokens, up to 10 epochs with early stopping (patience 3). Language-stratified splits guarantee at least 10% of each language's data in every validation fold, preventing empty folds for low-resource languages. Fixed global binarization thresholds: 0.50 (Subtask 1), 0.35 (Subtask 2), 0.30 (Subtask 3). All experiments run on NVIDIA Tesla T4 x2 GPUs via Kaggle Notebooks.

5 Ablation Study

Figure 3 summarizes the incremental Macro-F1 gains as each component is added to the system; points that degrade performance are marked ↓.

Ensembling dominates. The jump from a single mDeBERTa fold (0.350) to the full mega-ensemble (0.740) on Subtask 1 accounts for 0.390 Macro-F1 points—the largest single contribution. Combining one fold of each architecture already reaches 0.690, confirming that architectural diversity drives most of the gain. A single XLM-R fold (0.540) outperforms a single mDeBERTa fold (0.350), but the five-fold mDeBERTa ensemble (0.690) matches the three-fold XLM-R (0.680), illustrating how fold diversity compensates for per-model capacity.

Augmentation provides targeted gains. Macro-F1-aware augmentation adds 0.030 on Subtask 1, 0.009 on Subtask 2, and 0.037 on Subtask 3. Subtask 3 benefits most because *Dehumanization* and *Lack of Empathy* are near-absent in many languages; even a small number of synthetic examples provides enough positive signal to form a decision boundary. These gains are modest but consistent across subtasks and concentrated in long-tail labels.

Threshold tuning overfits. Label-level and joint language+label threshold tuning both degrade performance (↓0.015 and ↓0.008 respectively). The per-language-label validation counts are too small to estimate reliable thresholds; tuned values overfit to development-set noise. Fixed global thresholds generalize better and are used in the final system.

Forced argmax is necessary for Subtask 3. Without it, roughly 8% of polarized posts receive all-zero manifestation predictions (F1 = 0 per instance). The correction recovers 0.015–0.023 Macro-F1 points and is especially important for low-resource languages where ensemble probabilities are poorly calibrated.

	Subtask 1					Subtask 2			Subtask 3			
Lang	**Acc**	**Prec**	**Rec**	**F1**	**BL**	**F1-Mic**	**F1-Mac**	**BL**	**F1-Mic**	**F1-Mac**	**BL**	**Exact**
AM	0.843	0.858	0.942	0.776	.715	0.740	0.670	.372	0.587	0.544	.443	0.171
AR	0.834	0.813	0.818	0.833	.796	0.636	0.620	.486	0.646	0.602	.390	0.508
BE	0.844	0.836	0.784	0.839	.853	0.609	0.349	.289	0.389	0.249	.087	0.515
DE	0.720	0.718	0.685	0.719	.671	0.567	0.562	.408	0.505	0.490	.349	0.406
EN	0.807	0.744	0.724	0.791	.780	0.631	0.504	.333	0.506	0.491	.410	0.557
FA	0.846	0.874	0.925	0.787	.842	0.720	0.605	.463	0.586	0.458	.200	0.363
HA	0.924	0.652	0.611	0.794	.775	0.419	0.394	.204	0.246	0.204	.746	0.853
HI	0.913	0.935	0.965	0.813	.738	0.874	**0.790**	.791	0.782	**0.759**	.235	0.303
IT	0.689	0.771	0.486	0.671	.677	0.212	0.256	.376	—	—	—	—
KH	0.920	0.937	0.978	0.703	.659	0.842	0.699	.627	0.721	0.400	.610	0.619
MY	0.877	0.893	0.893	0.874	.821	0.776	0.699	.477	—	—	—	—
NE	0.909	0.913	0.905	**0.909**	.880	0.766	0.774	.722	0.683	0.645	.131	0.539
OR	0.842	0.776	0.627	0.794	.777	0.611	0.515	.560	0.385	0.286	.384	0.690
PA	0.768	0.737	0.812	0.768	.790	0.553	0.472	.365	0.519	0.504	.456	0.386
PO	0.819	0.796	0.763	0.813	.724	0.635	0.535	.449	—	—	—	—
RU	0.835	0.730	0.709	0.801	.746	0.638	0.565	.590	—	—	—	—
SP	0.788	0.779	0.796	0.788	.727	0.653	0.650	.594	0.534	0.507	.509	0.405
SW	0.771	0.791	0.738	0.771	.757	0.664	0.411	.442	0.568	0.547	.221	0.422
TE	0.873	0.896	0.855	0.873	.644	0.473	0.446	.315	0.506	0.429	.674	0.447
TU	0.794	0.792	0.821	0.793	.696	0.658	0.586	.471	0.609	0.515	.769	0.412
UR	0.823	0.860	0.889	0.787	.789	0.780	0.778	.713	0.808	**0.808**	.532	0.651
ZH	0.902	0.915	0.890	**0.902**	.869	0.793	**0.791**	.670	0.626	0.607	.000	0.665
Avg	**0.834**	0.819	0.801	**0.800**	.760	**0.648**	**0.576**	.487	**0.567**	**0.502**	.420	**0.495**

Table 1: Per-language results with POLAR baseline (BL) Macro-F1 from the official leaderboard (Naseem et al., 2026b). Subtask 1: Prec/Rec are for the polarized class. Subtask 3: dashes = languages not evaluated. Bold = highest F1-Mac per subtask.

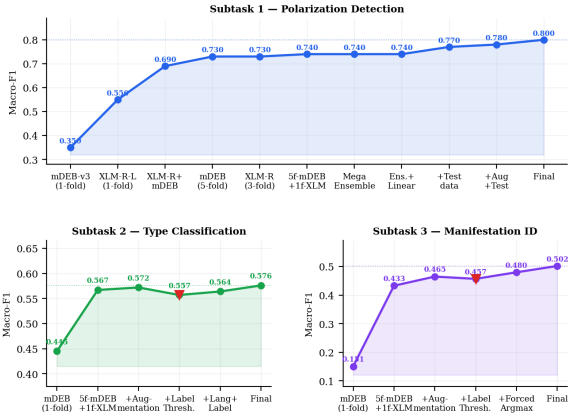

Figure 3: Ablation study across all three subtasks. Each point shows the incremental Macro-F1 score as components are added. Red triangles (↓) mark configurations where a component degraded generalization (threshold tuning); these are excluded from the final submission.

6 Results

Subtask 1. The system achieves 0.800 average Macro-F1 and 0.834 average accuracy across 22 languages. Nepali (0.909), Chinese (0.902), and Burmese (0.874) are the strongest. Italian (0.671) is the weakest, with polarized-class recall of only 0.486; Italian polarization tends to be expressed through irony and understatement, which cross-lingual models trained on more direct expressions fail to transfer.

Subtask 2. The system averages 0.576 Macro-F1 and 0.648 Micro-F1 across 22 languages. Chinese (0.791) and Hindi (0.790) are the strongest. Italian (0.256), Bengali (0.349), and Hausa (0.394) are the hardest cases. The *Other* and *Racial/Ethnic* labels score lowest across virtually all languages, mirroring the near-empty cells visible in Figure 1.

Subtask 3. Across 18 languages, the system achieves 0.502 Macro-F1 and 0.495 exact-match ratio. Urdu (0.808) and Hindi (0.759) are the strongest, reflecting their large and relatively balanced training sets. Hausa (0.204), Bengali (0.249), and Odia (0.286) are the hardest, each suffering from both data scarcity and near-absent minority labels. *Dehumanization* and *Lack of Empathy* remain the most difficult manifestations globally, with per-language F1 frequently below 0.30.

Comparison with POLAR baseline. Table 1 compares PolaFusion against the official POLAR baseline (Naseem et al., 2026b) (BL columns). On Subtask 1, PolaFusion outperforms the baseline in 17 of 22 languages (+0.040 average), with the largest gains on Telugu (+0.229) and Turkish (+0.098). On Subtask 2, we exceed the baseline in 17 of 22 languages (+0.089 average), led by Amharic (+0.298) and Burmese (+0.222). On Subtask 3, our system wins in 12 of 18 languages

(+0.082 average), with notable gains on Hindi (+0.524) and Nepali (+0.513); however, the baseline substantially outperforms PolaFusion on Hausa (-0.541) and Turkish (-0.254).

7 Discussion

The results reveal a consistent hierarchy of difficulty: binary detection (F1 = 0.800) is substantially easier than type classification (0.576), which in turn is easier than manifestation identification (0.502). The persistent gap between Micro-F1 and Macro-F1 in Subtasks 2 and 3 confirms that common labels (*Political*, *Stereotype*) are learned well while rare ones (*Racial/Ethnic*, *Dehumanization*) remain systematically underpredicted. Our BCE weighting is effective because it operates during optimization rather than through data manipulation, scaling naturally to any degree of imbalance, though it cannot substitute for real signal when positive counts approach zero.

Italian and Hausa represent two structurally different failure modes. Italian's problem is linguistic register—sarcasm and politically coded vocabulary that cross-lingual models cannot transfer—suggesting culturally grounded few-shot augmentation as a remedy. Hausa's problem is representational: the language is underrepresented in pretraining corpora of both encoders, pointing toward continued pre-training or retrieval-augmented generation from native sources. The POLAR baseline comparison (Table 1, BL columns) further reveals that PolaFusion's gains are not uniform: the baseline outperforms our system on Hausa and Turkish in Subtask 3, suggesting its approach better handles certain manifestation distributions.

8 Conclusion

We presented PolaFusion, a system that treats class imbalance as the primary design constraint. Hierarchical gating, mega-ensemble soft voting, inverse-frequency BCE weighting, and Macro-F1-aware LLM augmentation together achieve 0.800, 0.576, and 0.502 Macro-F1 on the three subtasks, outperforming the POLAR baseline by +0.040, +0.089, and +0.082 average Macro-F1 respectively. Ensembling is the dominant gain; augmentation provides consistent but moderate improvements concentrated in data-scarce conditions; threshold tuning overfits at this data scale and should be avoided. Persistent failure on rare labels in low-resource languages remains the key open challenge.

Limitations

Our final submission incorporates unlabeled test data into training via pseudo-labeling: the mega-ensemble's predictions on the official test set are treated as silver labels, and models are retrained on the union of training and pseudo-labeled data. This prevents computing unbiased held-out estimates, though official leaderboard scores remain unaffected. LLM augmentation may introduce distributional biases where few-shot seeds are limited. Fixed global thresholds are suboptimal for languages with extreme imbalance. The forced-argmax correction guarantees label consistency but can assign incorrect labels when confidence is uniformly low. We did not isolate augmentation and gating effects without the ensemble; a single-model ablation would clarify whether these components provide independent gains.

Acknowledgments

We thank the SemEval-2026 Task 9 organizers for the dataset and evaluation platform. Experiments were run on NVIDIA Tesla T4 x2 GPUs via Kaggle Notebooks.

References

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB 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.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

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 association for computational linguistics*, pages 8440–8451.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop*

on Semantic Evaluation (SemEval-2022), pages 533–549.

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

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

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. *Preprint*, arXiv:2505.20624.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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 4675–4684.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2 @ EVALITA 2020: Overview of the EVALITA 2020 hate speech detection task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*.

Cass R Sunstein. 2018. Republic: Divided democracy in the age of social media.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

A Augmentation Prompts

We use three task-specific prompt templates for Qwen3-235B-A22B, one per subtask. All prompts share the same structural skeleton but differ in their label context block. Each prompt is called with temperature 0.7, top-p 0.9, and a maximum of 128–150 tokens via the NVIDIA NIM API.

Subtask 1 — Polarization Detection

Prompt Template

```
[Task]
Act as a linguistic expert in social
    media discourse.
Generate one synthetic variation of
    the text below
for a machine learning dataset.

[Constraints]
- Label Context: The text is {
    LABEL_DESCRIPTION}.
- Meaning: Preserve the same opinion
    , stance, and target.
- Linguistic Style: Maintain the
    original level of
  slang, intensity, and informal
    grammar.
- Content: Do NOT soften the
    language. Do NOT add
  apologies, disclaimers, or new
    entities.
- Output: Return ONLY the new
    sentence. No preamble.

[Text to Transform]
{TEXT}

[Augmented Result]
```

where {LABEL_DESCRIPTION} takes one of two values depending on the ground-truth label:

- **Label 1** (polarized): *"HIGHLY POLARIZED/BIASED. Keep the aggressive or divisive tone."*
- **Label 0** (neutral): *"NEUTRAL/NON-POLARIZED. Keep the objective or factual tone."*

Subtask 2 — Polarization Type

Prompt Template

```
[Task]
Act as a linguistic expert in social
    media discourse
specialized in {LANGUAGE}.
Generate one synthetic variation of
    the {LANGUAGE}
text below.

[Context]
This text is polarized based on: {
    LABELS}.

Requirements for labels:
- POLITICAL: Maintain political
    ideology, party bias,
  or political hostility.
- RACIAL/ETHNIC: Keep the bias/
    hostility toward the
  specific race, ethnicity, or
    nationality.
- RELIGIOUS: Preserve the hostility
    or blame directed
  at the religion or religious group.

- GENDER/SEXUAL: Maintain the bias
    or insult regarding
  gender or sexual orientation.
- OTHER: Preserve targets like media
    , economy, or
  technology not covered by identity
    labels.

[Constraints]
- Preserve the same meaning, stance,
    and target.
- Maintain the same level of
    aggression and tone.
- Do NOT sanitize, soften, or
    neutralize the language.
- Do NOT add or remove entities or
    targets.
- Do NOT change the polarization
    category.
- Do NOT use words like "please", "
    respectfully",
  or "balanced".
- Output ONLY the new sentence. No
    explanation.

[Text to Transform]
{TEXT}

[Augmented Result]
```

Subtask 3 — Manifestation Identification

Prompt Template

```
[Task]
Act as a linguistic expert in social
    media discourse
specialized in {LANGUAGE}.
Generate one synthetic variation of
    the {LANGUAGE}
text below.

[Context]
This text expresses: {LABELS}.
Requirement: {LABEL_HINTS}

[Constraints]
- Preserve the same meaning, stance,
    and target.
- Maintain the same level of
```

```
    aggression and intensity.
- Do NOT sanitize, soften, or
    neutralize the language.
- Do NOT add or remove entities or
    targets.
- Do NOT switch to another
    manifestation type.
- Do NOT use words like "please", "
    respectfully",
  or "balanced".
- Do NOT replace slurs with generic
    insults.
- Keep slang and informal grammar if
     present.
- Output ONLY the new sentence. No
    explanations.

[Text to Transform]
{TEXT}

[Augmented Result]
```

The `{LABEL_HINTS}` placeholder is instantiated with a definition for each manifestation class present in the sample:

- **Stereotype** — use generalized claims about a group as inherent traits.
- **Vilification** — depict the target as immoral or corrupt.
- **Dehumanization** — portray the target as animals, objects, or less than human.
- **Extreme Language** — use highly aggressive or exaggerated expressions.
- **Lack of Empathy** — show indifference toward the target's suffering.
- **Invalidation** — dismiss the target's experiences or identity.

All three prompts use temperature 0.7, top-p 0.9, and a maximum of 128–150 tokens. The model is called via the NVIDIA NIM API.

B Augmentation Plans

Tables 2–4 list only the language–label pairs where both trigger conditions were met (ratio below threshold *and* F1 below threshold) and synthetic examples were generated. Pairs not shown were not augmented. The full augmented datasets are available in our code repository.

Table 2: Subtask 1 augmentation plan (triggered pairs only). Minority label: 1 = polarized, 0 = non-polarized.

Lang	F1	Ratio	Label	Gen.
Hausa	0.760	0.120	1	412
Khmer	0.614	0.101	0	642

Table 3: Subtask 2 augmentation plan (triggered pairs only).

Lang	Label	Count	F1	Ratio	Gen.
Amharic	Religious	69	0.400	0.030	69
Amharic	Gender/Sexual	20	0.000	0.009	20
English	Racial/Ethnic	295	0.526	0.244	295
English	Religious	117	0.429	0.097	117
English	Gender/Sexual	75	0.286	0.062	75
English	Other	132	0.000	0.109	132
Bengali	Racial/Ethnic	26	0.000	0.022	26
Bengali	Religious	68	0.250	0.057	68
Bengali	Gender/Sexual	18	0.500	0.015	18
Hausa	Gender/Sexual	30	0.000	0.160	30
Hausa	Other	15	0.000	0.080	15
Khmer	Racial/Ethnic	103	0.267	0.022	103
Punjabi	Racial/Ethnic	106	0.125	0.191	106
Polish	Gender/Sexual	115	0.308	0.125	115
Polish	Other	163	0.375	0.177	163
Chinese	Religious	89	0.500	0.087	89
Russian	Other	83	0.000	0.170	83
Swahili	Gender/Sexual	164	0.286	0.063	164
Swahili	Other	583	0.357	0.224	583
Odia	Racial/Ethnic	125	0.400	0.240	125
Odia	Gender/Sexual	83	0.000	0.159	83
Odia	Other	91	0.000	0.175	91
German	Gender/Sexual	196	0.533	0.143	196
Turkish	Other	120	0.500	0.108	120
Persian	Racial/Ethnic	84	0.364	0.055	84
Persian	Gender/Sexual	207	0.435	0.136	207

Table 4: Subtask 3 augmentation plan (triggered pairs only).

Lang	Label	Count	F1	Ratio	Gen.
Telugu	Dehumanization	62	0.000	0.095	62
Amharic	Dehumanization	460	0.508	0.241	460
Bengali	Lack of Empathy	66	0.000	0.078	66
Bengali	Invalidation	62	0.000	0.073	62
Hausa	Lack of Empathy	34	0.000	0.207	34
Hausa	Invalidation	9	0.000	0.055	9
Khmer	Dehumanization	85	0.333	0.018	85
Khmer	Lack of Empathy	765	0.482	0.161	765
Khmer	Invalidation	456	0.329	0.096	456
Arabic	Invalidation	288	0.375	0.218	288
Chinese	Dehumanization	226	0.581	0.167	226
Odia	Dehumanization	17	0.000	0.051	17
Odia	Lack of Empathy	39	0.000	0.117	39
Turkish	Lack of Empathy	237	0.318	0.221	237
Turkish	Invalidation	100	0.133	0.093	100
Persian	Dehumanization	149	0.316	0.075	149
Persian	Lack of Empathy	341	0.233	0.172	341
Persian	Invalidation	276	0.263	0.139	276