Tralaleros at SemEval-2026 Task 9: Multilingual Polarization Detection with Transformer-based Models

Adrian Dahl
Kiel University
stu228183@mail.uni-kiel.de

Adam Mierzwa
University of Hamburg
adam.mierzwa@studium.uni-hamburg.de

Bado Völckers
University of Hamburg
bado.voelckers@studium.uni-hamburg.de

Abstract

Online polarization, the sharp division of public opinion into opposing and often hostile groups, has become a multilingual phenomenon, but most existing corpora and detectors are restricted to a single language or political event. SemEval 2026 Task 9 addresses this gap by releasing a 22-language polarization corpus drawn from political discourse on social media. This paper describes our submission to Subtask 1, the binary polarization-detection track. We systematically investigate four training strategies for transformer-based classifiers, namely data rebalancing through undersampling and back-translation, hyperparameter optimisation with focal loss, model scaling from mBERT to larger multilingual encoders, and ensemble aggregation across architectures. A per-language analysis shows that no single model dominates and that the gain from ensembling is small once individual models are well tuned. Our final weighted-average ensemble reaches 0.7936 Macro F1 on the official leaderboard, with top-3 placements on Hindi and Telugu. All four fine-tuned models are released on HuggingFace.

1 Introduction

Polarization, the sharp division of public opinion into opposing and often hostile groups, has long been studied in political science but has only recently become a focus of multilingual natural language processing. The phenomenon is inherently cross-lingual, yet most automatic detectors and corpora to date have been restricted to a single language and a single political event (Naseem et al., 2026b), and it is not clear how the design choices that work for English transfer to typologically distant low-resource languages. SemEval 2026 Task 9 (POLAR @ SemEval-2026, 2026; Naseem et al., 2026a) addresses this question by requiring binary classification of social-media texts across 22 typologically diverse languages[1] as either *Polarized* or *Non-Polarized*, covering high-resource languages such as English and German alongside low-resource languages such as Amharic, Hausa, and Khmer.

We systematically explore four training strategies for transformer-based classifiers on this corpus. We first investigate whether the severe per-language class imbalance in the training data should be corrected by undersampling or by back-translation augmentation. We then conduct a random-search hyperparameter study on mBERT (Devlin et al., 2019), comparing focal loss against cross entropy under a range of learning-rate, freezing, and weight-decay configurations. With the resulting configuration as a starting point, we scale to three larger multilingual encoders, namely XLM-RoBERTa Large (Conneau et al., 2020), RemBERT (Chung et al., 2020), and mDeBERTa-v3 (He et al., 2023), each chosen to vary one architectural axis relative to mBERT. Finally we aggregate the four fine-tuned classifiers into a weighted-average ensemble.

The empirical picture that emerges is one of architectural specialisation rather than universal superiority. Undersampling consistently degrades performance, suggesting that in shared multilingual embedding spaces the additional majority-class signal outweighs class parity. XLM-RoBERTa Large achieves the highest single-model dev Macro F1 of 0.7929, but a per-language breakdown shows that RemBERT is consistently strongest on Indic languages, mDeBERTa on Semitic and morphologically rich languages, and mBERT remains competitive on Bengali, Burmese, Spanish, and Swahili. Despite this complementarity, our ensemble improves over the best single model by only 0.0008

[1] Amharic (amh), Arabic (arb), Bengali (ben), Burmese (mya), English (eng), German (deu), Hausa (hau), Hindi (hin), Italian (ita), Khmer (khm), Nepali (nep), Odia (ori), Persian (fas), Polish (pol), Punjabi (pan), Russian (rus), Spanish (spa), Swahili (swa), Telugu (tel), Turkish (tur), Urdu (urd), and Chinese (zho).

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 343–353

on dev, which we read as evidence that universal aggregation is not the right way to exploit per-language strengths, and that future work should focus on language-specific routing.

Our final test-set submission, the weighted-average ensemble, reaches 0.7936 Macro F1 on the official Subtask 1 leaderboard,[2] with top-3 placements on Hindi (3/35) and Telugu (3/33). All four fine-tuned models are released on HuggingFace: mBERT,[3] XLM-RoBERTa Large,[4] RemBERT,[5] and mDeBERTa-v3.[6] A privacy-preserving Firefox browser extension that runs the quantised mBERT locally via ONNX Runtime is released alongside.[7]

2 Background

The data for this shared task comes from POLAR (Naseem et al., 2026b), a multilingual, multicultural, and multi-event corpus of over 110K instances across 22 languages, annotated along three axes: polarization detection, type, and manifestation. Unlike prior corpora restricted to a single language or political event, POLAR enables systematic comparison of models and training strategies across diverse linguistic families. We participate in Subtask 1 (binary polarization detection) using the subset released for SemEval-2026 Task 9 (Naseem et al., 2026a), and do not compare against the benchmark results reported in the original POLAR paper; our contribution is orthogonal, focusing on training-strategy and model-family choices.

The SemEval training split contains 73,681 samples across 22 languages, ranging from 1,700 (Punjabi) to 6,991 (Swahili). While globally balanced (53.13% polarized), per-language polarization rates vary drastically, from $< 10\%$ (Hausa) to $> 90\%$ (Khmer); full sample-count and polarization-rate distributions are reported in Appendix E. Texts are social-media posts and online comments related to political events. Token length analysis showed 128 tokens cover most samples; we set this as our maximum sequence length. We use the official train/dev/test splits and evaluate with Macro F1.

[2] https://github.com/Polar-SemEval/Leaderboards

[3] https://huggingface.co/DavidGetter/mbert-polarized-model

[4] https://huggingface.co/DavidGetter/xlmr-large_normal_dataset-polar

[5] https://huggingface.co/DavidGetter/rembert-polarized

[6] https://huggingface.co/DavidGetter/mdeberta-v3-polarized

[7] https://github.com/adriandahl/PolarHighlighter

Related Work: Online polarization detection is closely related to stance detection and hate-speech classification, tasks for which fine-tuned pre-trained transformers (Vaswani et al., 2017) have become the standard approach. In the multilingual setting, mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) exploit shared sub-word vocabularies and large unlabeled corpora to enable cross-lingual transfer, and have been shown to generalize to low-resource languages when fine-tuned on task data. More recent models refine these foundations along different axes: RemBERT (Chung et al., 2020) decouples input and output embeddings to obtain stronger cross-lingual representations, while mDeBERTa-v3 (He et al., 2023) combines disentangled attention with ELECTRA-style replaced-token detection. Because each of these models encodes different inductive biases, recent work on multilingual classification has increasingly treated them as complementary rather than interchangeable, which directly motivates the cross-model analysis we present in Section 5.

A recurring obstacle in multilingual text classification is severe class imbalance, both globally and at the per-language level. Two complementary families of methods have been explored: loss reweighting and data-level resampling or augmentation. Focal Loss (Lin et al., 2018), originally proposed for dense object detection, downweights well-classified majority examples and has been widely adopted for imbalanced NLP tasks. On the data side, back-translation through a strong pivot language, typically using large multilingual translation models such as NLLB-200 (Team et al., 2022), provides a label-preserving way to synthesize additional minority-class samples while largely retaining source-language semantics. Both families inform our rebalancing strategy (Section 5), where we compare Focal Loss, undersampling, and NLLB-based back-translation against an mBERT reference trained on the unmodified full dataset, with a per-language majority-class baseline reported as a non-learning floor.

3 System Overview

Our pipeline explores four complementary strategies for binary text classification as either *Polarized* or *Non-Polarized*, organised so that each strategy addresses a distinct aspect of multilingual classifier design: data composition (Strategy A),

fine-tuning hyperparameters (Strategy B), encoder choice (Strategy C), and aggregation across encoders (Strategy D). The implementation uses HuggingFace Transformers throughout.

3.1 Strategy A: Data Rebalancing

Per-language polarization rates in the training split range from below 10% (Hausa) to above 90% (Khmer), and prior work on multilingual classification has shown that severe imbalance can encourage the classifier to collapse onto the majority class during fine-tuning. We therefore compare two complementary corrections against a baseline that uses the unmodified training data. The first, which we refer to as undersampling, restores a 60/40 class balance for any language whose majority class exceeds 70%, removing 12,652 samples in total from Amharic, Hausa, Khmer, and similar skewed languages. The second is back-translation augmentation: we synthesise additional minority-class examples by round-tripping through English with NLLB-200 (Team et al., 2022), chosen for its strong coverage of the low-resource languages in the corpus. English-to-English back-translation is skipped to avoid duplication, and the augmentation for each language is capped at a 1:1 ratio relative to the minority class. The resulting augmented corpus contains 9,542 additional samples, expanding the training data by approximately 13%.

3.2 Strategy B: Hyperparameter Optimization

Beyond data rebalancing, the choice of fine-tuning hyperparameters has a large effect on the behaviour of multilingual encoders, particularly under class imbalance, where loss reweighting can shift the decision boundary independently of the data distribution. Focal Loss (Lin et al., 2018), originally proposed for dense object detection, downweights well-classified majority examples through the term $L_{FL} = -\alpha_t(1-p_t)^{\gamma}\log p_t$ and has since been adopted widely for imbalanced NLP tasks; it is therefore a natural complement to the data-side rebalancing of Strategy A. We conducted a random search over mBERT (Devlin et al., 2019) fine-tuning hyperparameters, sampling 18 configurations that varied the loss (cross entropy vs. focal loss), the backbone learning rate ($5\mathrm{e}{-6}$–$5\mathrm{e}{-5}$), the classifier-head learning rate ($1\mathrm{e}{-5}$–$1\mathrm{e}{-4}$), and the freeze strategy (no freezing, embeddings only, or the first six layers). The full table of runs is reported in Appendix A, and configurations were ranked by dev-set Macro F1.

The winning run used focal loss with frozen embeddings, decoupled learning rates ($LR_{back} = 2\mathrm{e}{-5}$, $LR_{head} = 1\mathrm{e}{-4}$), batch size 16, and three epochs, reaching Macro F1 of 0.744, well above both the majority-class baseline of 0.381 and our initial mBERT reference run of 0.730. Three patterns emerged from the search. Focal loss was more robust than cross entropy: the two lowest-scoring runs (F1 0.340 and 0.561) both used cross entropy. Freezing the input embeddings helped consistently, with four of the five best runs adopting this setting; we attribute this to the embeddings already being well aligned across languages from mBERT pre-training. An aggressive backbone learning rate of $5\mathrm{e}{-5}$ was uniformly destructive, with all four runs using it scoring at or below 0.699. We carry this best configuration forward as the starting point for the per-model search in Strategy C.

3.3 Strategy C: Model Scaling

Although Strategy B identified a strong fine-tuning configuration for mBERT, it leaves open the question of whether the remaining performance gap is a property of the loss and freezing choices or of the mBERT encoder itself. We therefore selected three larger multilingual encoders, each chosen to vary one architectural axis relative to mBERT and otherwise to share the same fine-tuning recipe. XLM-RoBERTa Large (Conneau et al., 2020) controls for raw scale and pre-training corpus, with 560M parameters trained on the CC-100 multilingual web crawl. RemBERT (Chung et al., 2020) controls for the input/output embedding design through its decoupled, rebalanced embeddings, which earlier work has shown to improve cross-lingual transfer in low-resource settings. mDeBERTa-v3 (He et al., 2023) controls for the attention mechanism and pre-training objective through its disentangled attention and ELECTRA-style replaced-token detection. Any gain over mBERT can therefore be attributed to a specific architectural change rather than to a confounded combination of factors.

Starting from the mBERT-optimal configuration of Strategy B, we ran a smaller per-model random search of four to six configurations sampled around that optimum, varying the backbone and head learning rates, batch size, freeze strategy, weight decay, and dropout, while holding focal loss and the warmup ratio fixed. The mBERT optimum transferred almost directly to XLM-R Large and mDeBERTa-v3, whereas RemBERT required

the largest deviation, namely no freezing, higher learning rates of $3e{-}5$ and $2e{-}4$, and dropout 0.15. Per-model best configurations are reported in Appendix B.

3.4 Strategy D: Ensemble Learning

The cross-model results in Section 5 make clear that each of the four fine-tuned models is the strongest on a different subset of languages, which is the canonical setting in which ensembling can yield gains. We therefore combine the four checkpoints from Strategies B and C (mBERT, XLM-R Large, RemBERT, and mDeBERTa-v3), using each model's best dev-set configuration (Appendix B). We compare four aggregation methods, namely weighted average, majority voting, simple average, and max confidence. For weighted average, each model receives a weight proportional to its dev Macro F1, normalised to sum to one ($w_i = \mathrm{F1}_i / \sum_j \mathrm{F1}_j$); the narrow F1 band of 0.74–0.79 produced near-uniform weights of approximately 0.24–0.26, which is why we did not pursue learned weighting such as stacking.

4 Experimental Setup

We use the official SemEval splits, training on the full 73,681-sample training set, validating on the dev set, and submitting predictions on the held-out test set. Texts are tokenised with the HuggingFace AutoTokenizer for each backbone and padded or truncated to a maximum length of 128 tokens, which earlier token-length analysis showed to cover the bulk of the corpus. For all four models we use HuggingFace's `AutoModelForSequenceClassification` head with `num_labels=2`, which adds a single linear projection on top of the pooled `[CLS]` representation; predictions are obtained as the argmax of the softmax over the two logits, i.e. a fixed 0.5 decision threshold on the polarized-class probability. The same softmax probabilities are also the inputs to the ensemble aggregation methods of Strategy D.

Optimisation uses AdamW with gradient clipping at 1.0, three epochs of training, and early stopping on dev-set Macro F1. Batch size, backbone and head learning rates, dropout, weight decay, warmup ratio, and freeze strategy are tuned per model and reported in Appendix B. The primary evaluation metric is Macro F1 averaged across the 22 per-language F1 scores, with binary F1 and accuracy as secondary metrics. All experiments were implemented with HuggingFace Transformers and PyTorch and trained on NVIDIA L4 and A100 GPUs; the deployed browser extension uses ONNX Runtime for on-device inference.

5 Results

5.1 Overall Performance and Ranking

On the dev set, XLM-RoBERTa Large achieved our best single-model performance with Macro F1 of 0.7929, a 5.4-point improvement over our tuned mBERT at 0.7440 (configuration r07; Appendix A). Our final test-set submission used the weighted-average ensemble described in Section 5.4, which reached an average Macro F1 of 0.7936 across the 22 languages on the official Subtask 1 leaderboard,[8] closely matching the dev-set ensemble result of 0.7937 and confirming that the system generalises from dev to test.

The full per-language leaderboard ranks are reported in Appendix D, and we summarise the headline positions here. Our team (*Tralaleros*) achieved a top-3 finish on Hindi (rank 3 of 35, F1 0.8178) and on Telugu (rank 3 of 33, F1 0.8968), and a top-10 finish on three additional languages: Punjabi (6/33), Spanish (7/37), and Hausa (10/31). Across the remaining languages we ranked between 11 and 21 of typically 30–37 teams, with the notable outlier of Khmer (25/31). The Khmer ranking is consistent with the persistent difficulty observed for that language across all four models on the dev set, where the Macro F1 stays around 0.53 (see Table 6); the combination of a 90.8% polarized class rate and a script with limited representation in the pre-training corpora of all four encoders likely accounts for this.

5.2 Impact of Data Rebalancing

A per-language majority-class baseline reaches only 0.381 Macro F1 (Table 1), which all mBERT conditions clear by a wide margin and which we use as a non-learning floor. Against the mBERT-on-full-data reference at 0.730, undersampling degraded every metric, reducing Macro F1 to 0.690 (a 5.5-point drop) with degradation in 18 of the 22 languages (Appendix C). We read this as evidence that, in shared multilingual embeddings, data volume outweighs class parity: even majority-class samples provide useful cross-lingual signal that is lost when those samples are removed.

[8] https://github.com/Polar-SemEval/Leaderboards

Back-translation augmentation showed a more mixed pattern (Macro F1 0.735, narrowly above the reference). The largest gain was on Hindi (+8 points), but most other languages either showed no change or a small degradation, which we attribute to cultural context being lost or distorted during the round trip through English. This is consistent with the well-known limitation that pivot-based back-translation tends to neutralise stylistic and culturally specific cues, which are precisely the cues that polarization detection depends on.

Table 1: Aggregate performance comparison. The majority-class baseline predicts the most frequent training-set label per language. Green/Red shading marks improvement or degradation of the rebalancing conditions relative to the mBERT (full data) reference.

Metric	Maj. Class	mBERT (full)	Undersampled	Augmented
Macro F1 (avg)	0.381	0.730	0.690	0.735
Binary F1 (avg)	0.353	0.781	0.744	0.747
Accuracy (avg)	0.624	0.804	0.760	0.785

Table 2: Ensemble Strategy Comparison (Avg F1-Macro)

Strategy	Avg F1-Macro	Improvement
Weighted Average	**0.7937**	+0.0008
Majority Voting	0.7937	+0.0008
Simple Average	0.7928	-0.0001
Max Confidence	0.7904	-0.0025

Detailed per-language results before and after undersampling are reported in Appendix C.

5.3 Cross-Model Analysis: No Universal Solver

Although XLM-RoBERTa Large achieves the highest average dev Macro F1 (0.7929), it does not dominate across the 22 languages of the corpus, and the per-language breakdown in Table 6 (Appendix C) shows a clear pattern of architectural specialisation. XLM-R Large performs best on European languages, including German (0.778) and Russian (0.819), and on a subset of Indic languages such as Punjabi (0.850) and Odia (0.785), which is consistent with its scale and the composition of CC-100. RemBERT leads on Persian (0.911), Hindi (0.843), Telugu (0.907), and Urdu (0.755); we attribute its strength on these complex scripts to its rebalanced output embeddings, which give relatively more capacity to languages with large vocabularies. mDeBERTa is strongest on the Semitic and morphologically rich languages of the corpus, including Amharic (0.716), Arabic (0.846), and Turkish (0.835), which is consistent with its disentangled-attention mechanism being particularly useful for languages with rich agglutinative or templatic morphology. Despite being the smallest of the four models, mBERT retains top performance on Bengali (0.839), Burmese (0.879), Spanish (0.733), and Swahili (0.802); we suspect this reflects the relatively higher quality of the Wikipedia data on which mBERT was pre-trained for these languages.

This distribution suggests that each architecture captures distinct linguistic features, and the error-complementarity analysis in Appendix F supports the same reading: XLM-R and mBERT show the lowest error overlap (Jaccard 0.420, complementarity 0.408), and Cohen's κ across model pairs ranges from 0.696 to 0.745, indicating that the models make systematically different mistakes.

5.4 Ensemble Learning

Despite the strong error complementarity reported in the previous subsection, the four aggregation methods yielded only minimal gains on the dev set (Table 2). Weighted average and majority voting both reached 0.7937, an improvement of 0.0008 over the XLM-R single-model baseline, while simple average (0.7928) and max confidence (0.7904) underperformed it. This marginal improvement suggests that the weaker models, in particular mBERT at 0.7440, dilute the ensemble more than their complementary errors compensate for, and is also consistent with the observation that the strongest model differs across languages: a single global aggregation cannot exploit per-language specialisation. We therefore read the result as evidence that language-specific routing, in which a gating mechanism selects the best model for each language, is a more promising direction than universal aggregation. We selected the weighted-average ensemble for our final test-set submission; its per-language leaderboard performance is reported in Section 5.1 and Appendix D.

5.5 Error Analysis

To understand where the system fails, we manually inspected 91 misclassifications produced by the ensemble across Polish, German, and English. We restricted the analysis to these three languages because the authors are fluent in them, and translation could mask the linguistic and cultural nuances that determine whether a text is polarized. The error distribution is roughly balanced, with 46.2% false positives (non-polarized classified as polar-

ized) and 53.8% false negatives (polarized missed by the model), and four recurring patterns emerge.

The largest category, accounting for 38% of analysed errors, involves implicit political stance, in which the model struggles with strong political opinions that lack overt hostile language. The German sentence "*Finds aber interessant, dass dann Grüne sehr schockiert sind, dass man sie Linksgrün nennt*" is annotated as non-polarized but predicted as polarized; it expresses political commentary without any explicit polarization marker, and the model appears to rely on surface lexical cues that do not generalise to this register.

A second category, 29% of errors, is sarcasm and irony, in which the predicted label corresponds to the literal reading of the text rather than the intended one. The English sentence "*Yes if we all post the Ukrainian flag on social media it'll stop the Russian aggression*" is in fact polarized, criticising performative activism, but the model labels it non-polarized. This is a well-documented limitation of fine-tuned transformer classifiers, which lack explicit pragmatic reasoning.

A third category, 23% of errors, is context-dependent hostility, in which short statements require cultural or political knowledge to be classified correctly. The German sentence "*Türken wählen größtenteils SPD*" ("Turks mostly vote SPD") is annotated as polarized because it implies an ethnic-political stereotype, but the model treats it as a neutral statement of fact.

The remaining 10% of errors are borderline cases where the polarization judgement is itself subjective; these likely reflect annotation inconsistencies in the training data rather than a model failure mode.

The language breakdown of the analysed errors (German 45.1%, English 34.1%, Polish 20.9%) shows that all three languages contribute, with German over-represented; this is consistent with the relatively low per-language Macro F1 on German across all four models on the dev set.

6 Conclusion

Our participation in SemEval 2026 Task 9 yields three main observations. First, on this corpus, data volume outweighs class parity: undersampling degraded performance across 18 of the 22 languages despite severe per-language imbalance, which suggests that in shared multilingual embedding spaces the additional majority-class signal contributes useful cross-lingual information that should not be discarded. Second, no single architecture dominates the 22 languages of the corpus; while XLM-RoBERTa Large achieves the highest dev-set average at 0.7929, RemBERT is consistently strongest on Indic languages and mDeBERTa on Semitic languages, indicating that the four models we studied have complementary inductive biases. Third, despite this complementarity, the gain from a weighted-average ensemble is only 0.0008 on the dev set, which we read as evidence that aggregating uniformly across languages is the wrong way to exploit per-language strengths and that a routing or gating mechanism that selects the best model for each language is a more promising direction for future work, alongside higher-quality cross-cultural augmentation and pragmatic features for sarcasm and implicit stance.

We release the four fine-tuned models on HuggingFace, together with *Polar Highlighter*,[9] a Firefox browser extension that runs the quantised mBERT locally via ONNX Runtime WebAssembly (approximately 173 MB) for fully on-device, privacy-preserving inference.

Acknowledgments

We thank the SemEval 2026 Task 9 organizers for providing the dataset and evaluation framework.

References

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *Preprint*, arXiv:2010.12821.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

[9] https://github.com/adriandahl/PolarHighlighter

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

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, Juan Ren, Saba Anwar, Sarah Kohail, Rudy Alexandro Garrido Veliz, Robert Geislinger, Aisha Jabr, Idris Abdulmumin, Laiba Qureshi, Aarushi Ajay Borkar, Maryam Ibrahim Mukhtar, Abinew Ali Ayele, Ibrahim Said Ahmad, Adem Ali, Martin Semmann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

POLAR @ SemEval-2026. 2026. Detecting multilingual, multicultural and multievent online polarization (task 9). Accessed: 2026-01-29.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Hyperparameter Search Results

Run	LR_{back}	LR_{head}	Batch	Ep	W.Dec	Drop	Warm	Freeze	Loss	F1
r07	2e-5	1e-4	16	3	0.00	0.1	0.06	Emb	Focal	**0.744**
r06	5e-6	1e-4	16	4	0.00	0.1	0.06	Emb	Focal	0.731
r04	1e-5	1e-5	16	3	0.01	0.2	0.10	Emb	CE	0.729
r05	3e-5	5e-5	16	3	0.05	0.1	0.10	None	Focal	0.729
r03	2e-5	5e-5	32	2	0.00	0.1	0.00	Emb	Focal	0.723
r08	2e-5	1e-4	16	4	0.05	0.1	0.00	None	CE	0.723
r09	1e-5	1e-5	8	2	0.00	0.2	0.06	None	CE	0.715
r10	2e-5	1e-4	8	3	0.01	0.3	0.10	None	CE	0.714
r13	5e-6	5e-5	16	2	0.05	0.1	0.10	None	Focal	0.711
r12	5e-6	1e-5	16	3	0.01	0.2	0.10	None	CE	0.701
r11	5e-5	5e-5	16	2	0.00	0.2	0.00	First 6	Focal	0.699
r02	1e-5	1e-4	8	3	0.00	0.3	0.00	None	CE	0.699
r01	2e-5	1e-5	8	4	0.00	0.2	0.00	None	CE	0.697
r14	1e-5	1e-5	16	3	0.01	0.2	0.00	First 6	CE	0.695
r15	2e-5	5e-5	32	2	0.01	0.3	0.00	Emb	Focal	0.686
r16	5e-5	1e-5	32	2	0.01	0.3	0.06	First 6	Focal	0.663
r17	5e-5	1e-5	16	2	0.01	0.1	0.00	None	CE	0.561
r18	5e-5	1e-4	16	4	0.05	0.3	0.00	Emb	CE	0.340

Table 3: Hyperparameter search results, sorted by dev-set Macro F1. Abbreviations: Ep = Epochs, W.Dec = Weight Decay, Drop = Dropout, Warm = Warmup Ratio, Emb = Embeddings Frozen, First 6 = First 6 Layers Frozen.

B Per-Model Best Hyperparameter Configurations

Table 4 reports the best configuration for each of the four fine-tuned models, used to produce the dev-set results in Table 6. The starting point for each per-model refinement random search (Strategy C) was the mBERT-optimal run `r07` from Table 3. All four configurations share the focal loss formulation ($\alpha = 1$, $\gamma = 2$), the AdamW optimizer, a maximum sequence length of 128 tokens, gradient clipping at 1.0, and 3 training epochs with early stopping on dev Macro F1. All four fine-tuned checkpoints are released on HuggingFace (Section 1) for direct reproduction of the reported numbers.

Hyperparameter	mBERT	XLM-R	RemBERT	mDeBERTa
LR_{back}	2e-5	2e-5	3e-5	2e-5
LR_{head}	1e-4	1e-4	2e-4	1e-4
Batch size	16	32	24	32
Weight decay	0.01	0.0	0.0	0.0
Dropout	0.1	0.1	0.15	0.1
Warmup	0.06	0.06	0.05	0.06
Freeze	Emb	Emb	None	Emb
Loss	Focal	Focal	Focal	Focal

Table 4: Best hyperparameter configuration per model after the per-model refinement random search. *Emb* denotes that input embeddings are frozen during fine-tuning; *None* denotes no freezing. All four models trained for 3 epochs with early stopping on dev Macro F1.

C Per-language Results for Data Rebalancing

Table 5: Per-language Macro F1 of the majority-class baseline alongside mBERT trained on the full dataset, undersampled data (Cut), and back-translation augmented data. Green/Red shading marks improvement or degradation of the rebalancing conditions relative to the mBERT (full data) reference.

Language	Code	Maj. Class (F1)	mBERT full (F1)	Undersampling (F1)	Augmentation (F1)
Amharic	AMH	0.4236	0.4820	0.4601	0.5251
Arabic	ARB	0.3574	0.7179	0.7447	0.7152

Continued on next page...

Language	Code	Maj. Class (F1)	mBERT full (F1)	Undersampling (F1)	Augmentation (F1)
Bengali	BEN	0.3664	0.8373	0.8277	0.8190
German	DEU	0.3430	0.6774	0.6967	0.6894
English	ENG	0.3870	0.7803	0.7642	0.7625
Persian	FAS	0.4164	0.8724	0.8282	0.8577
Hausa	HAU	0.4709	0.7080	0.6389	0.7495
Hindi	HIN	0.4498	0.7469	0.7053	0.8246
Italian	ITA	0.3688	0.6193	0.6367	0.6142
Khmer	KHM	0.4755	0.4755	0.5525	0.4747
Burmese	MYA	0.3600	0.8670	0.8189	0.8652
Nepali	NEP	0.3377	0.8700	0.8300	0.8699
Odia	ORI	0.4129	0.5178	0.4982	0.5603
Punjabi	PAN	0.3464	0.7591	0.7552	0.7469
Polish	POL	0.3670	0.7850	0.7188	0.6683
Russian	RUS	0.4078	0.7081	0.6821	0.7069
Spanish	SPA	0.3373	0.6907	0.6969	0.7384
Swahili	SWA	0.3327	0.8080	0.7994	0.7993
Telugu	TEL	0.3333	0.8896	0.8559	0.8808
Turkish	TUR	0.3391	0.7358	0.7123	0.7301
Urdu	URD	0.4120	0.7319	0.7248	0.6890
Chinese	ZHO	0.3292	0.8832	0.8831	0.8738
Average	**AVG**	**0.3807**	**0.7300**	**0.7153**	**0.7346**

Table 6: Model performance comparison. Green indicates the best performing model for each language.

Language	Code	mBERT	XLM-R Large	RemBERT	mDeBERTa
Amharic	AMH	0.5356	0.6927	0.7041	0.7162
Arabic	ARB	0.7384	0.7893	0.7439	0.8455
Bengali	BEN	0.8388	0.8307	0.8147	0.8074
German	DEU	0.6954	0.7782	0.7093	0.7289
English	ENG	0.7519	0.7983	0.7684	0.7956
Persian	FAS	0.8527	0.8827	0.9105	0.8082
Hausa	HAU	0.7744	0.8044	0.7943	0.7842
Hindi	HIN	0.8074	0.8171	0.8432	0.8232
Italian	ITA	0.6442	0.6595	0.6174	0.6545
Khmer	KHM	0.4755	0.5850	0.6442	0.5594
Burmese	MYA	0.8794	0.8657	0.8436	0.8714
Nepali	NEP	0.8800	0.8900	0.8197	0.8499
Odia	ORI	0.5438	0.7853	0.6228	0.7510
Punjabi	PAN	0.7596	0.8499	0.7500	0.7999
Polish	POL	0.7477	0.7780	0.7207	0.7966
Russian	RUS	0.6868	0.8185	0.7207	0.7487
Spanish	SPA	0.7331	0.7141	0.6654	0.6513
Swahili	SWA	0.8023	0.7937	0.7844	0.7934
Telugu	TEL	0.8644	0.8713	0.9067	0.7869
Turkish	TUR	0.7550	0.8086	0.7652	0.8347
Urdu	URD	0.7194	0.7283	0.7549	0.7120
Chinese	ZHO	0.8878	0.9018	0.9018	0.8877
Average	**AVG**	**0.7440**	**0.7929**	**0.7639**	**0.7730**

D Per-Language Test-Set Leaderboard Results

Table 7 reports our team's (*Tralaleros*) official test-set Macro F1 and per-language rank on the POLAR @ SemEval-2026 Subtask 1 leaderboard,[10] alongside the top score and top team for each language. Top-3 placements are highlighted in green; ranks of top-10 placements are bolded.

[10] https://github.com/Polar-SemEval/Leaderboards

Table 7: Per-language test-set Macro F1 from the official Subtask 1 leaderboard for our team *Tralaleros*. Green highlights top-3 placements; bolded ranks indicate top-10 placements. "Total" is the number of submitting teams for that language.

Language	Code	Rank	Our F1	Top F1	Top Team	Total
Amharic	AMH	20	0.7619	0.8002	PSK	30
Arabic	ARB	19	0.8199	0.8488	UTokyo Tsuruoka Lab	33
Bengali	BEN	13	0.8359	0.8625	UTokyo Tsuruoka Lab	37
German	DEU	16	0.7183	0.7608	NYCU-NLP	33
English	ENG	16	0.7973	0.8252	UTokyo Tsuruoka Lab	44
Persian	FAS	16	0.8049	0.8424	POLAR Baseline	32
Hausa	HAU	**10**	0.8070	0.8336	PhatThachDau	31
Hindi	HIN	**3**	0.8178	0.8236	PSK	35
Italian	ITA	16	0.6129	0.7303	mdok-style	32
Khmer	KHM	25	0.6242	0.7744	SMASH	31
Burmese	MYA	14	0.8750	0.8913	taien	30
Nepali	NEP	14	0.9014	0.9236	NYCU-NLP	33
Odia	ORI	21	0.7736	0.8255	UTokyo Tsuruoka Lab	33
Punjabi	PAN	**6**	0.7960	0.8257	UTokyo Tsuruoka Lab	33
Polish	POL	16	0.8042	0.8431	Lingo Research Group	32
Russian	RUS	11	0.7939	0.8303	UTokyo Tsuruoka Lab	31
Spanish	SPA	**7**	0.7869	0.8030	UTokyo Tsuruoka Lab	37
Swahili	SWA	12	0.7890	0.8113	PSK	31
Telugu	TEL	**3**	0.8968	0.9053	Sagarmatha	33
Turkish	TUR	19	0.7754	0.8329	NYCU-NLP	31
Urdu	URD	16	0.7760	0.8196	UTokyo Tsuruoka Lab	35
Chinese	ZHO	17	0.8910	0.9315	yunkuang0329	33
Average	**AVG**	—	**0.7936**	—	—	**33.2**

E Dataset Distributions

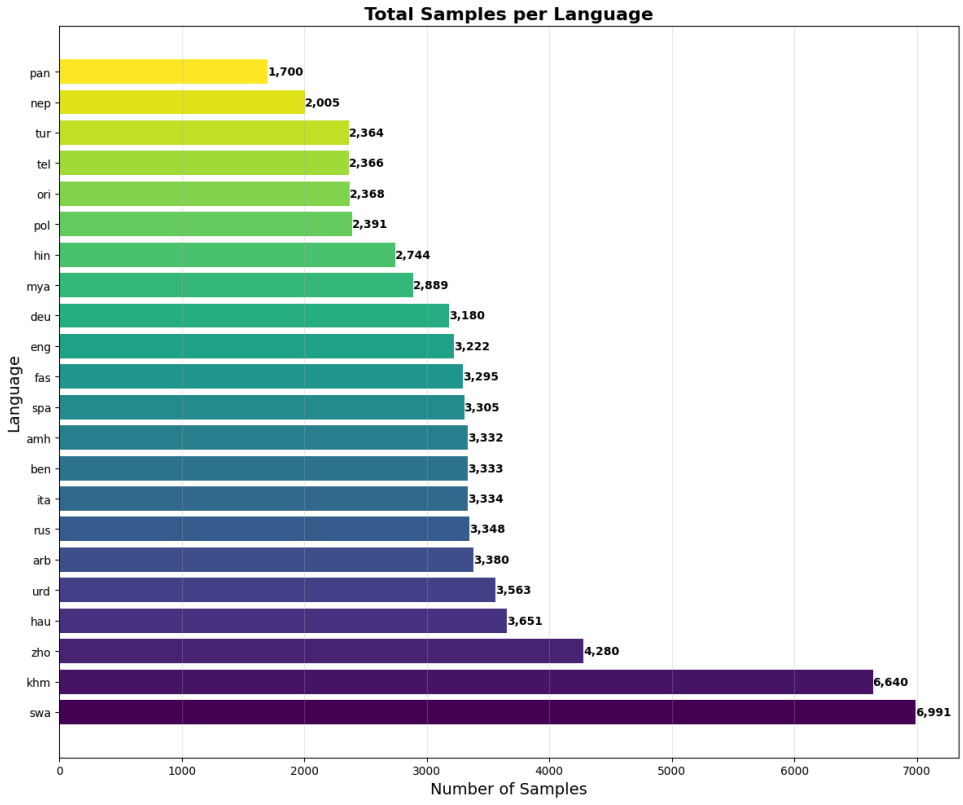

Figure 1: Total training samples per language. High-resource languages such as Swahili contain ∼4× the data of low-resource ones such as Punjabi.

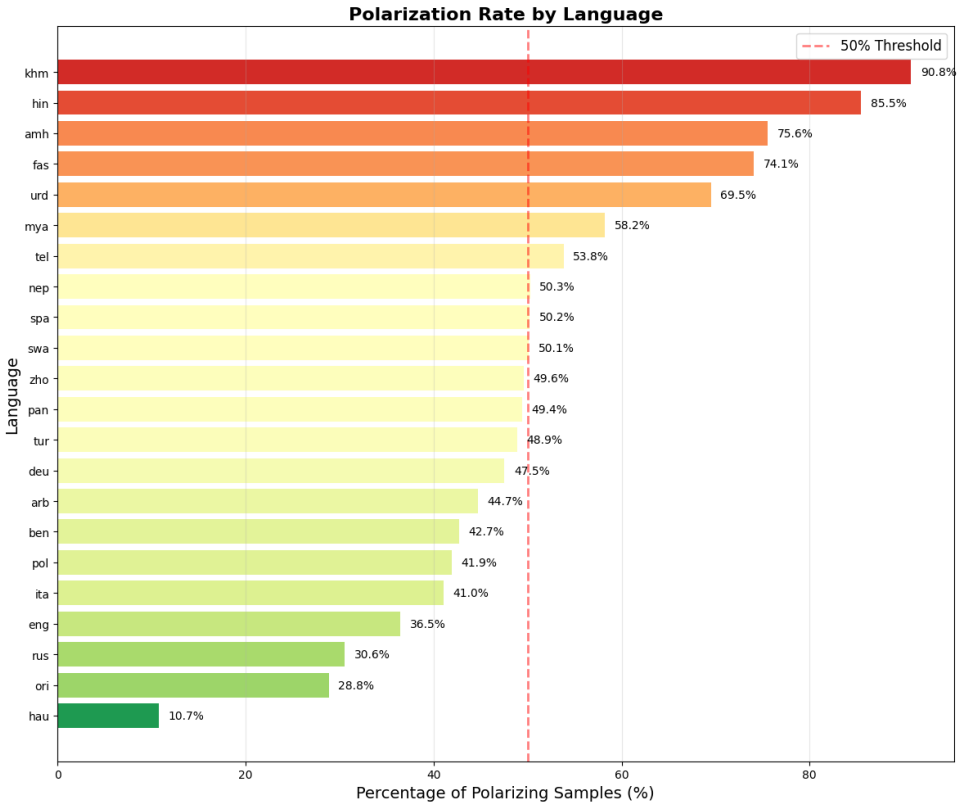

Figure 2: Polarization rate by language. Khmer and Hindi exceed 85% polarized, while Hausa is dominated by non-polarized examples.

F Error Complementarity Across Models

Figure 3 reports the pairwise error agreement (Jaccard similarity over the set of misclassified dev examples) and the complementarity score (1− Jaccard) for the four fine-tuned models from Strategy C. Lower agreement and higher complementarity indicate that the models disagree on which examples they get wrong, which is a precondition for ensemble gains.

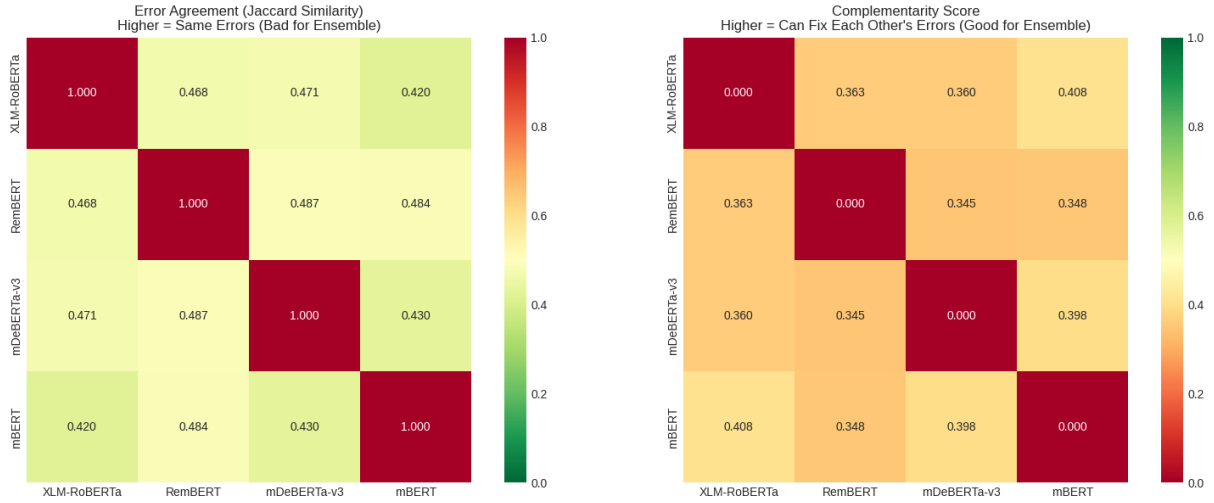

Figure 3: Pairwise error agreement (Jaccard similarity) and complementarity scores across the four fine-tuned models. Lower agreement and higher complementarity indicate stronger ensemble potential.