StanceLab at SemEval-2026 Task 9: Addressing Class Imbalance in Multilingual Polarization Detection

Teodor Ivănușcă and Dan Dodun-Des-Perrieres and Ștefana Gheorghiță
Faculty of Computer Science
"Alexandru Ioan Cuza" University of Iași
Iași, Romania
{teodor.ivanusca, dan.dodun, stefana.gheorghita}@student.uaic.ro

Abstract

Polarization in online discourse poses significant challenges for natural language processing, particularly in multilingual and culturally diverse environments. In this paper, we address the SemEval-2026 POLAR shared task on multilingual polarization detection across 22 languages. We adopt a staged experimental strategy that first investigates the problem in a controlled monolingual English setting before extending the approach to multilingual modeling. Our system evaluates several transformer-based architectures, including RoBERTa, XLM-RoBERTa, MPNet, and mDeBERTa-v3, combined with techniques designed to mitigate class imbalance such as weighted loss functions, focal loss, and data augmentation using back-translation and large language models. Experimental results show that no single configuration consistently dominates across all languages. However, focal loss and augmentation frequently improve performance in languages with skewed label distributions. Our findings highlight the importance of contextual representations, imbalance-aware training strategies, and language-specific considerations for robust multilingual polarization detection.

1 Introduction

Social media platforms have significantly transformed interpersonal communication. While they facilitate rapid information exchange, they also contribute to the spread and reinforcement of polarized narratives, which can intensify social divisions and hinder constructive dialogue. Consequently, detecting and analyzing polarization in online messages has become an important challenge in natural language processing (NLP).

Recent advances in transformer-based language models have substantially improved performance across many NLP tasks, including text classification, sentiment analysis, and stance detection. These models provide strong contextual representations that can capture subtle linguistic signals associated with polarization. However, polarization is strongly shaped by linguistic, cultural, and socio-political factors that vary across languages and regions, making multilingual polarization detection particularly challenging.

To support research in this area, (Naseem et al., 2026a) introduces *POLAR*, a multilingual dataset designed to capture diverse forms of polarized content across multiple languages and cultural contexts. The task (Naseem et al., 2026b) encourages the development of robust models capable of generalizing beyond language-specific patterns and addressing real-world multilingual scenarios.

In this work, we evaluate several transformer-based approaches and investigate strategies for handling class imbalance in both monolingual and multilingual settings, with particular attention to underrepresented and highly imbalanced languages.

2 Related Work

Early work on polarization and related phenomena in online discourse has focused on tasks such as hate speech, stance, and offensive language detection, often within monolingual and event-specific settings. These studies typically relied on supervised learning methods and targeted explicit signals of hostility or strong opinion polarity in social media texts.

More recently, transformer-based language models have become the dominant approach for treating these tasks. Multilingual pretrained models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mDeBERTa (He et al., 2021) have demonstrated strong performance in cross-lingual text classification tasks (Conneau et al., 2020; Arango et al., 2021), enabling transfer learning across languages with limited annotated data.

Experimental results reported by the POLAR

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3608–3620

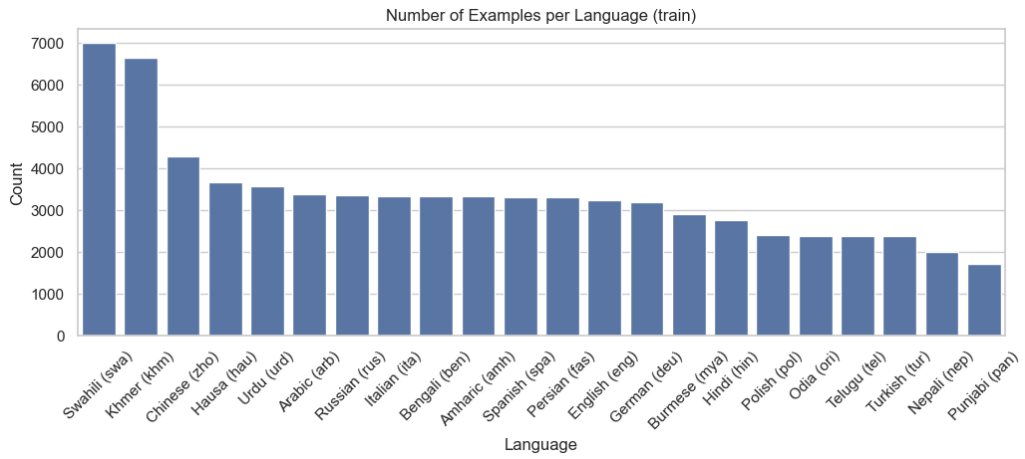

Figure 1: Train Samples per Language

benchmark (Naseem et al., 2026a) dataset authors show that multilingual transformer models achieve reasonable performance on binary polarization detection but exhibit substantial performance degradation on fine-grained tasks such as type and manifestation classification, especially in cross-lingual evaluation scenarios. These findings highlight the inherent complexity of polarization as a socially and culturally grounded phenomenon.

Related research has also explored multi-task learning as a means of improving performance on closely related objectives, such as jointly modeling hate speech, stance, and target detection (Rodriguez-Garcia and Centeno, 2024; Kang et al., 2025). Multi-task transformer architectures have been shown to benefit from shared representations across tasks, particularly when annotated data is limited. Additionally, domain-specific multilingual representations have been proposed to better capture culturally sensitive language patterns, outperforming general-purpose multilingual embeddings in zero-shot transfer settings (Arango et al., 2021).

Overall, existing work suggests that while large multilingual models provide a strong foundation for polarization detection, effectively modeling the linguistic diversity, cultural specificity, and rhetorical strategies involved in polarized discourse remains an open challenge.

Motivated by these observations, we explore transformer-based architectures combined with data augmentation techniques and imbalance-aware training strategies in order to improve robustness in multilingual polarization detection.

3 POLAR Task

This work addresses the SemEval-2026 POLAR shared task (Naseem et al., 2026b), which aims to automatically identify polarized content in online messages across multiple languages and cultural contexts. Given a short text extracted from social media platforms, systems are required to determine whether the text contains polarization and, depending on the subtask, further characterize its nature.

3.1 Task Definition

The POLAR dataset (Naseem et al., 2026a) defines three complementary subtasks: binary polarization detection, polarization type classification, and polarization manifestation identification.

In this work, we focus on the *binary polarization detection* subtask. For each input message, the system predicts whether the content is *Polarized* or *Not Polarized*.

3.2 Dataset

The dataset includes 22 languages spanning diverse linguistic and regional contexts, with data sourced from multiple platforms such as Twitter/X, Facebook, BlueSky, Reddit, and local news outlets (Naseem et al., 2026a). Each instance consists of a unique identifier, the input text, and a binary polarization label.

The languages are represented in different proportions across the dataset, as shown in Figure 1.

3.3 Exploratory Analysis

To better characterize the challenges of multilingual polarization detection, we conduct an exploratory analysis of the training data, focusing on

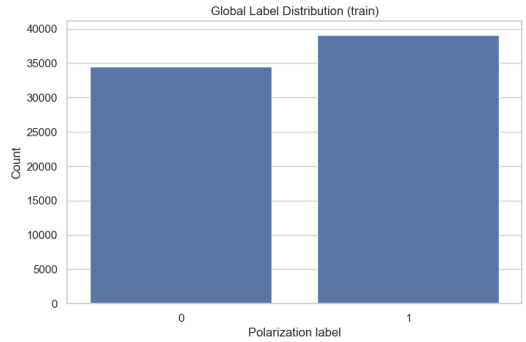

Figure 2: Label Distribution

cross-lingual label distributions, text length properties, and their relationship with the polarization label. Due to space constraints, we focus here on the cross-lingual label distribution and defer the analysis of text length properties to Appendix A.

Cross-lingual label imbalance. Although the full dataset, comprising 73,681 entries, is approximately balanced with respect to polarization labels (Figure 2), the label distribution varies substantially between languages, which strongly affects model performance and results.

Figure 3 illustrates this variability, with polarization rates ranging from around 10% in Hausa to over 90% in Khmer. Several languages exhibit near-balanced distributions (e.g., Chinese, Spanish, Swahili), whereas others are strongly skewed toward one class. This pronounced heterogeneity suggests that polarization manifests differently across linguistic and cultural contexts and motivates the use of macro-averaged evaluation metrics to avoid bias toward languages with extreme class imbalance.

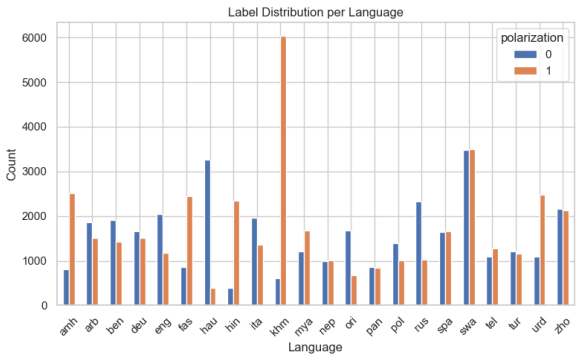

Figure 3: Label Distribution per Language

We adopt a staged strategy combining monolingual and multilingual modeling for polarization detection. We first conduct experiments in English to establish baselines and validate architectural choices in a controlled setting, and then extend the approach to the full multilingual scenario.

3.4 Data Preprocessing

The dataset requires minimal preprocessing. All instances contain valid text and labels, with no missing values. No additional normalization or filtering is applied; social media artifacts such as URLs, mentions, hashtags, and emojis are retained. Tokenization is performed using the tokenizer associated with each model, and input sequences are truncated to a fixed maximum length during training.

3.5 Data Augmentation

To improve robustness and increase minority-class diversity, we experiment with several augmentation strategies. First, we use back-translation through an English–Spanish–English pipeline based on Helsinki-NLP models (Tiedemann and Thottingal, 2020). Second, we generate paraphrases with the OpenAI GPT-4.1 API (Achiam et al., 2023) while preserving the original semantic content and label. Third, we use the `gemini-2.5-flash-lite` API (Comanici et al., 2025) in a few-shot setting to generate additional minority-class examples. Augmented samples are used only in training and are excluded from validation and test splits.

For GPT-4.1 augmentation, we employ a constrained paraphrasing prompt designed to preserve the original stance, target group(s), sentiment polarity, and overall polarization intensity of each example, while explicitly prohibiting the introduction of hedging, neutrality, or new semantic content. The prompt also instructs the model to retain stylistic markers such as hashtags, mentions, emojis, and informal language when present, ensuring that generated paraphrases remain close to the original discourse style.

For Gemini augmentation, we use a few-shot generation prompt conditioned on the target language and class label. The model is provided with a small set of in-language minority-class examples and instructed to generate new social media comments that match their tone, length, and linguistic style, while remaining consistent with the desired polarization label (polarized or neutral/objective). This approach is intended to enrich minority-class diversity while preserving language-specific discourse patterns.

To verify sample quality, we apply both automatic and manual filtering. For GPT-4.1 paraphrases, we enforce length-ratio constraints relative to the source text and reject generations containing banned hedging expressions (e.g., "some people say", "it seems", "allegedly") that could weaken polarization cues or alter label semantics. For Gemini-generated examples, we monitor generation outputs for formatting correctness, linguistic plausibility, and label consistency. In addition, manual inspection of random subsets of generated samples is performed to confirm semantic coherence, preservation of intended label characteristics, and the absence of obvious artifacts, duplicates, or off-topic generations.

3.6 Monolingual Models

We first study English-only polarization detection to validate modeling choices before moving to multilingual classification. As lightweight baselines, we train Naive Bayes and Logistic Regression models using Bag-of-Words features. We then evaluate contextual encoder models, including RoBERTa (Liu et al., 2019) and DeBERTa-v3 (He et al., 2021), with RoBERTa-large providing the strongest validation performance.

Training is initially performed using a single train–validation split, followed by a 5-fold stratified cross-validation setup with soft-voting over fold predictions. Optimization uses AdamW (Loshchilov and Hutter, 2019) with cross-entropy loss, and class weights are incorporated to mitigate moderate label imbalance in the English subset.

We also explore MPNet-based sentence representations (Song et al., 2020) (Appendix C), prompt-based classification with GPT-4.1, and layer-wise learning rate decay; however, these alternatives do not consistently outperform RoBERTa-large. Motivated by the potential usefulness of sentence embedding models such as MPNet, we design a hybrid architecture that aims to combine the strengths of RoBERTa-large with MPNet representations. Specifically, the model integrates RoBERTa `[CLS]` representations with MPNet sentence embeddings through feature concatenation, followed by a lightweight classifier. The architecture of the model is illustrated in Figure 4. However, this system was not evaluated in the multilingual setting because of time and computational constraints.

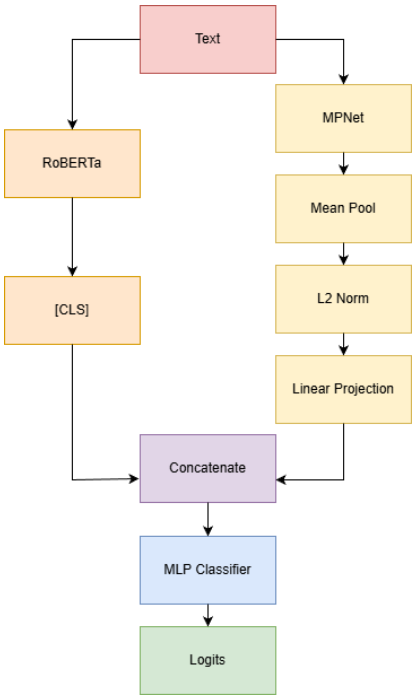

Figure 4: Hybrid Model

3.7 Multilingual Models

For multilingual classification, we evaluate multilingual transformer models including XLM-RoBERTa (Conneau et al., 2020) and mDeBERTa-v3 (He et al., 2021) under different training configurations. In addition, we explore a sentence-embedding–based approach based on MPNet (Song et al., 2020).

To address class imbalance, we experiment with cost-sensitive learning using weighted cross-entropy:

$$L_{\text{WCE}}(y, \hat{y}) = -\sum_{c=1}^{C} w_c y_c \log(\hat{y}_c) \qquad (1)$$

where y_c denotes the ground-truth indicator, $\hat{y}_c$ the predicted probability, and w_c the class-specific weight.

We additionally evaluate focal loss (Lin et al., 2020), which emphasizes harder examples during training:

$$L_{\text{Focal}}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \qquad (2)$$

where p_t is the probability assigned to the gold label and γ controls the focusing strength.

In the multilingual setting, we also assess LLM-based semantic augmentation using the `gemini-2.5-flash-lite` API to increase linguistic diversity while preserving label semantics (Ding et al., 2020; Dai et al., 2025).

4 Experimental Setup

Dataset Splits. For monolingual experiments, we train models on the English portion of the training data. Model selection is performed using a stratified validation split, and final performance is estimated using a 5-fold stratified cross-validation protocol. For multilingual experiments, models are trained on the full multilingual training set provided by the task organizers, in both cases with a (90% / 10%) train–validation split.

Evaluation Metric. Following the official task guidelines, system performance is evaluated using macro-averaged F1-score across classes. This metric equally weights both polarized and non-polarized categories and is therefore suitable for imbalanced datasets.

Training Details. All transformer models are fine-tuned using the AdamW optimizer with cross-entropy loss. We evaluate two main configurations: a baseline setup corresponding to the default training configuration provided by the task organizers, and a modified configuration obtained through hyperparameter tuning. Hyperparameters are selected through Bayesian optimization on the validation set. The search space includes learning rates in the range $[10^{-6}, 3 \times 10^{-5}]$, weight decay values $\{0.0, 0.01, 0.05\}$, label smoothing factors $\{0.0, 0.05, 0.1\}$, and per-device batch sizes of $\{4, 8\}$. Training uses gradient accumulation when necessary to accommodate GPU memory constraints. More details on the parameters are presented in Appendix B.

Implementation. Experiments are implemented using the HuggingFace Transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). Hyperparameter optimization and experiment tracking are conducted using Weights & Biases (Lukas et al., 2020).

5 Results

In Table 1 we report the main test-set results across all languages. The scores present on the official leaderboard are marked with "*". Table 2 presents our official leaderboard results together with the POLAR baseline.

For XLM-R and MPNet we evaluate two parameter configurations: the default baseline setup and a modified configuration. The modified configuration corresponds to the training setup that yielded the best performance for RoBERTa-large in the monolingual English experiments.

Overall, no single model or configuration consistently outperforms the others across all languages. Instead, performance varies depending on both the architecture and the target language, highlighting the inherent difficulty of multilingual polarization detection.

Nevertheless, several trends can be observed. The modified configurations generally improve performance for a subset of languages, suggesting that training settings optimized on English can transfer partially to multilingual models. In addition, the mDeBERTa-based models demonstrate strong performance across many languages, particularly when combined with focal loss or ensemble strategies.

Notably, configurations that incorporate focal loss and data augmentation frequently achieve the best results for individual languages. Focal loss helps emphasize harder or underrepresented examples during training, which appears particularly beneficial in languages with skewed class distributions. Similarly, augmentation-based approaches improve performance in several languages by increasing training diversity and reducing overfitting.

However, these techniques do not yield universal improvements, and their effectiveness remains language-dependent. Taken together, the results suggest that while focal loss and augmentation can provide meaningful gains, no single configuration dominates across the entire multilingual setting.

Overall, the test-set results are largely consistent with those observed on the validation dataset (Appendix D). The main discrepancy occurs for English: although the hybrid architecture with 5-fold cross-validation achieved the best validation performance, the RoBERTa-large model with 5-fold cross-validation proved to be the most stable on the test set.

6 Conclusions

In this work, we investigated multilingual polarization detection in the context of the SemEval-2026 POLAR shared task. The dataset presents several challenges, including strong cross-lingual variation in label distributions, heterogeneous text lengths, and culturally grounded manifestations of polarization. Our exploratory analysis highlighted the limited predictive power of surface-level features and motivated the use of contextual representations

Experiment	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
XLM-R - Baseline	0.7505	0.807	0.8205	0.6664	0.7618	0.7966	0.7278	0.804	0.5848	0.6231	0.8476	0.8837	0.7354	0.7499	0.7721	0.7570	0.7530	0.7629	0.8655	0.7635	0.7593	0.8614
XLM-R - Modified Config	0.727	0.7822	0.84	0.6834	0.7824	0.7979	0.7842	0.7562	0.5931	0.5778	0.8098	0.8399	0.7037	0.7290	**0.8092**	0.7483	0.6776	0.7593	0.8559	0.7561	0.7415	0.8785
XLM-R - Modified Config (Subset)	–	–	–	**0.7095**	0.7929	–	–	–	0.6153	–	–	–	–	–	0.8065	0.7768	0.7584	–	–	–	–	–
MPNet - Baseline	0.706	0.7710	0.8187	0.6784	0.7664	**0.8102***	0.7318	0.7878	0.5796	0.5419	0.8567	0.8781	**0.7651**	0.7428	0.7545	0.7055	0.7563	0.7667	0.806	0.7368	0.7392	0.8416
MPNet - Modified Config	0.7425	0.7681	**0.8329**	0.6961	0.7832	0.7904	0.7406	0.7985	0.6214	0.6065	0.8409	0.8670	0.7576*	0.7414	0.7473	0.7474	0.7424	0.7664	0.8534	0.7636	0.7356	0.8712
mDeBERTa-v3-base weighted training	0.7491	0.8144	0.8233	0.7065	0.7945	0.7905	0.7696	0.7794	**0.672***	**0.761***	0.86	**0.9092***	0.7336	**0.7761***	0.7864	0.7578	0.7581	0.7733	0.8799	0.7613	0.7564	0.8707
mDeBERTa-v3-base ensemble + Focal Loss	**0.7664***	**0.8247***	0.8261	0.7053*	0.7892	0.798	0.7407	0.814	0.6411	0.7245	**0.8756***	0.8992	0.7325	0.7737	0.7837	**0.7771***	**0.7722***	**0.7919***	**0.8818***	**0.7723***	0.7624	0.8791
mDeBERTa-v3-base ensemble + Augmentation	0.7653	0.8139	0.8318*	0.7032	0.7825	0.8004	**0.7922***	**0.8167***	0.6457	0.685	0.8605	0.8992	0.7504	0.7651	0.7908*	0.7754	0.768	0.7828	0.8618	0.7644	**0.7703***	0.8827
RoBERTa-large + 5 fold cross-validation	–	–	–	–	**0.8112**	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Hybrid Architecture	–	–	–	–	0.8008*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Chinese RoBERTa	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**0.9019***

Table 1: Per-language performance comparison across different training configurations on the test set.

	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
Our results (rank)	0.7664 (17)	0.8247 (17)	0.8318 (18)	0.7053 (25)	0.8008 (13)	0.8102 (13)	0.7922 (16)	0.8167 (5)	0.6720 (2)	0.7610 (2)	0.8756 (12)	0.9092 (8)	0.7577 (28)	0.7761 (11)	0.7908 (23)	0.7771 (18)	0.7722 (16)	0.7919 (10)	0.8818 (15)	0.7723 (23)	0.7703 (24)	0.9019 (11)
POLAR baseline (rank)	0.7151 (27)	0.7957 (30)	0.8528 (4)	0.6714 (30)	0.7802 (35)	0.8424 (1)	0.7753 (19)	0.7379 (32)	0.6773 (2)	0.6592 (22)	0.8210 (29)	0.8798 (28)	0.7765 (21)	0.7898 (8)	0.7241 (30)	0.7457 (29)	0.7266 (33)	0.7571 (29)	0.6440 (33)	0.6957 (29)	0.7890 (9)	0.8691 (28)

Table 2: Our results compared with the POLAR baseline on the test set.

and imbalance-aware evaluation.

We adopted a staged experimental strategy, starting with a controlled monolingual setting in English before extending to multilingual modeling. In the monolingual experiments, transformer-based models substantially outperformed Bag-of-Words baselines, highlighting the importance of contextual semantics for detecting polarized discourse. Increasing model capacity and incorporating data augmentation further improved performance, particularly for the minority class. Experiments with prompt-based large language models showed that, without task-specific fine-tuning, these approaches remain unstable and underperform encoder-based models.

In the multilingual setting, we evaluated several transformer architectures and training strategies, including weighted losses, focal loss, data augmentation, and ensembling. Results indicate that no single configuration consistently outperforms others across all languages; instead, performance varies depending on language characteristics and data distribution.

Overall, our findings highlight the importance of contextual modeling, careful handling of class imbalance, and macro-averaged evaluation for multilingual polarization detection. Future work will focus on reducing language-specific performance variability in multilingual settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2021. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *Preprint*, arXiv:2104.14728.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2025. AugGPT: Leveraging ChatGPT for Text Data Augmentation. *IEEE Transactions on Big Data*, 11(3):907–918.

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 4171–4186.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using

ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *ArXiv*, abs/2111.09543.

Long Kang, Jiaqi Yao, Ruoshuang Du, Lu Ren, Haifeng Liu, and Bo Xu. 2025. A stance detection model based on sentiment analysis and toxic language detection. *Electronics*, 14(11).

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Biewald Lukas and 1 others. 2020. Experiment tracking with weights and biases. *Software available from wandb. com*.

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, 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

Raquel Rodriguez-Garcia and Roberto Centeno. 2024. HAMiSoN-MTL at ClimateActivism 2024: Detection of hate speech, targets, and stance using multi-task learning. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 89–95, St. Julians, Malta. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Text Length and Polarization

Text length by polarization label. We analyze text length in terms of word counts across the English data. Figure 5 shows that polarized posts tend to be slightly longer on average than non-polarized ones; however, the overlap between the two distributions is substantial. Although a small number of long outliers are present in both classes, most texts remain relatively short. This indicates that while length may weakly correlate with polarization, it is insufficient as a standalone predictive signal.

Length-feature correlations. To further assess the informativeness of surface-level features, we compute correlations between polarization, character length, and word length. As shown in Figure 6, character and word counts are strongly correlated with each other, while their correlation with the polarization label is weak (absolute correlation below 0.1). This confirms that surface-level length features alone do not capture the complexity of polarization and highlights the need for models that leverage semantic and contextual information.

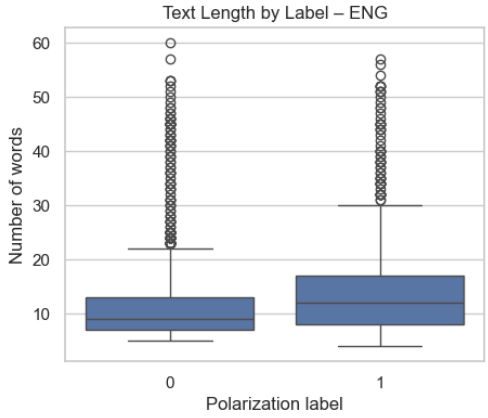

Figure 5: Text length (in words) by polarization label.

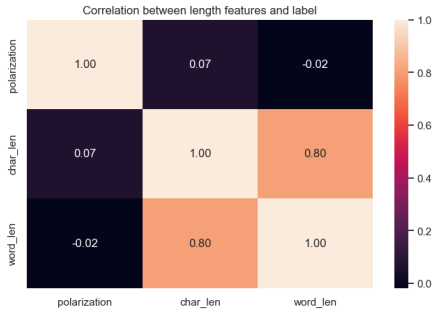

Figure 6: Correlation between polarization and length-based features.

Cross-lingual variation in text length. Text length distributions differ markedly across languages, as illustrated in Figure 7. Languages such as Bengali, Hindi, and Russian exhibit longer median word counts, while others, including Chinese and Khmer, contain substantially shorter texts due to differences in writing systems and tokenization granularity. These discrepancies underline the linguistic diversity of the dataset and pose an additional challenge for multilingual modeling, as a single representation must accommodate highly heterogeneous input characteristics.

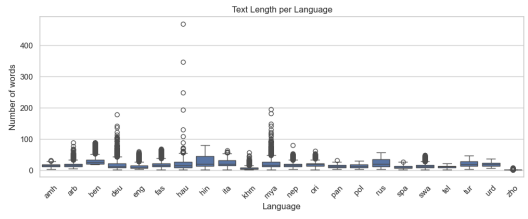

Figure 7: Word length by language

Language-specific length distributions. Figure 8 presents the word-length distribution for English as an illustrative example. The distribution is heavily right-skewed, with most texts containing fewer than 20 words and a long tail of rare but very long posts. Similar patterns are observed across languages, reflecting the social-media-driven nature of the data. The prevalence of short texts increases ambiguity, while the presence of long outliers motivates the use of input truncation during training to ensure computational efficiency and robustness.

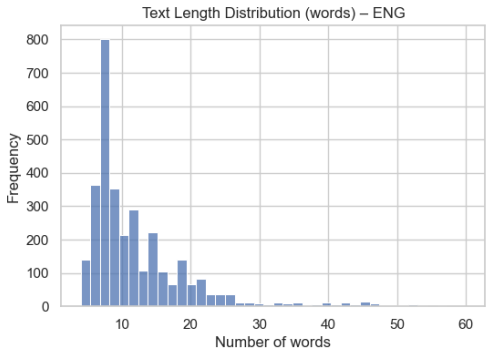

Figure 8: Word-length distribution for English.

Overall, this exploratory analysis highlights three key aspects of the dataset: strong cross-lingual variation in label distributions, substantial heterogeneity in text length across languages, and the limited predictive power of surface-level length features. These observations motivate the adoption of multilingual, contextual models capable of capturing semantic cues beyond simple lexical or structural statistics.

B Experimental Setup

B.1 Monolingual Settings (English)

Monolingual experiments are conducted on the English subset of the dataset and evaluated using macro-averaged F1.

Initial Training Configuration. Initial experiments are performed using a single train–validation split (90% / 10%) to guide model selection. The training setup is as follows:

- Optimizer: AdamW
- Learning rate: 5×10^{-6}
- Scheduler: cosine with warmup ratio 0.10
- Weight decay: 0.01
- Training epochs: 3
- Batch size: 4 (train), 8 (evaluation)

- Gradient accumulation: enabled
- Maximum sequence length: 256
- Label smoothing: 0.05

Cross-Validation Setup. After identifying a stable configuration, we adopt stratified k-fold cross-validation with $k = 5$. Each fold is trained independently using the same hyperparameters, and predictions are aggregated via soft-voting by averaging class probabilities. All reported monolingual results are obtained using this cross-validation ensemble.

B.2 Multilingual Settings

As an initial experiment, we fine-tuned XLM-RoBERTa on a multilingual dataset obtained by combining instances from all available languages. This experiment followed the default configuration provided in the official baseline, using a learning rate of 2×10^{-5} and a per-device training batch size of 64. Training was conducted for 5 epochs.

Subsequently, we explored a more refined training configuration to improve performance. In this setup, we applied weight decay (0.01), a warmup ratio of 0.10, and label smoothing with a factor of 0.05. Training was again performed for 5 epochs, using a reduced per-device batch size of 4 combined with gradient accumulation over 4 steps. A lower learning rate of 5×10^{-6} was adopted to stabilize optimization.

We further evaluated this refined configuration on a subset of languages, namely English, Italian, Spanish, German, Russian, and Polish.

Building on these configurations, we also experimented with the sentence-embedding–based approach described in the previous section, using the `sentence-transformers/paraphrase-multilingual-mpnet-base-v2` model. This model was evaluated under the first two training configurations described above.

C Sentence-Level Representations

We evaluated a sentence-embedding–based classifier using `sentence-transformers/all-mpnet-base-v2`, where token representations are aggregated via mean pooling before classification (Figure 9). This approach simplifies the representation to a single sentence vector, but consistently underperformed token-level fine-tuning, indicating that polarization cues benefit from richer token-level interactions.

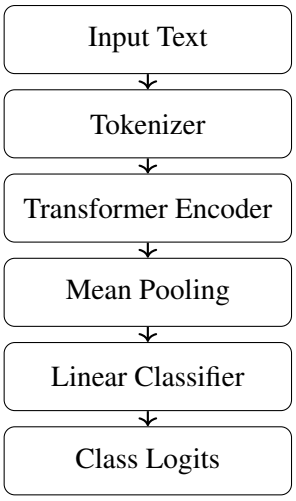

Figure 9: Sentence-Embedding–Based Classification Architecture with Mean Pooling.

D Experimental Results on Validation Dataset

D.1 Monolingual Results (English)

Baseline Results We first evaluate lightweight baseline models on the English subset to establish a lower bound for polarization detection. As shown in Table 3, Naive Bayes achieves a macro-F1 score of 0.686, while Logistic Regression attains higher accuracy but a lower macro-F1 of 0.652, reflecting a strong bias toward the majority class. Overall, both models struggle to reliably capture polarized content, highlighting the limitations of bag-of-words representations.

Model	Acc.	Macro-F1
Naive Bayes (Count)	0.712	0.686
Logistic Regression (TF–IDF)	0.719	0.652

Table 3: Baseline performance on the English subset.

Hyperparameter Tuning. We performed Bayesian hyperparameter optimization to identify effective training configurations for our models. Hyperparameters were selected using Bayesian search, with the objective of maximizing macro-F1 on a held-out validation split derived from the training data. Each trial consisted of a full training run under a sampled hyperparameter configuration, followed by evaluation on the validation split. The resulting macro-F1 score was then used to guide subsequent hyperparameter selection.

The search space covered both optimization-related and regularization-related hyperparameters. The learning rate was sampled from a log-uniform

Figure 10: Parameter effect

distribution in the range $[10^{-6}, 3 \times 10^{-5}]$, enabling exploration across multiple orders of magnitude. Regularization was controlled via weight decay, with values selected from $\{0.0, 0.01, 0.05\}$, and label smoothing factors from $\{0.0, 0.05, 0.1\}$.

We additionally varied the training dynamics by exploring per-device batch sizes of 4 and 8, combined with gradient accumulation steps of 1, 2, or 4, which allowed effective batch size scaling under GPU memory constraints. Learning rate warmup was applied using warmup ratios of 0.05 and 0.10. Finally, we evaluated strategies for handling class imbalance by toggling the use of class-weighted loss functions and optionally balancing the evaluation split.

This Bayesian optimization setup enabled the search procedure to efficiently focus on promising regions of the hyperparameter space, leading to stable high-performing configurations without requiring an exhaustive grid search.

In addition, the influence of individual hyperparameters on the validation macro-F1 score is illustrated in Figure 10. Based on the Bayesian hyperparameter optimization, the best-performing configuration was obtained with a learning rate of 7.5×10^{-6}, weight decay set to 0.01, and a label smoothing factor of 0.1. Training was performed using a per-device batch size of 8, a warmup ratio of 0.05, and gradient accumulation set to 1. In addition, both class balancing on the evaluation set and the use of class weights during training were activated. This configuration consistently achieved the highest validation macro-F1 score within the sweep and was therefore selected for subsequent experiments.

Additional Experiments We report intermediate monolingual experiments that guided the development of the final system, focusing on progressively stronger RoBERTa-based configurations.

RoBERTa-base without augmentation. We first fine-tune RoBERTa-base on the English subset without data augmentation. This configuration yields a clear improvement over bag-of-words baselines, achieving an accuracy of **0.79** and a macro-F1 score of **0.76**. Performance remains stronger for the non-polarized class (F1 = **0.84**) than for the polarized class (F1 = **0.69**), indicating that contextual representations capture meaningful polarization cues but still struggle with minority-class detection in the absence of explicit imbalance handling.

RoBERTa-large with data augmentation. Building on these results, we adopt RoBERTa-large and augment the training data using GPT-based paraphrasing. This setup further improves performance, reaching an accuracy of **0.83** and a validation macro-F1 score of **0.81**. The observed gains suggest that increased model capacity combined with data augmentation leads to better generalization, particularly for polarized content.

Training dynamics for this configuration are shown in Figures 11, 12, and 13. The training loss (Figure 11) decreases smoothly, indicating stable optimization. Validation macro-F1 (Figure 12) improves rapidly during early training and remains stable thereafter, suggesting effective convergence. The validation loss curve (Figure 13) exhibits moderate fluctuations, which are consistent with the increased variability introduced by augmented samples rather than overfitting.

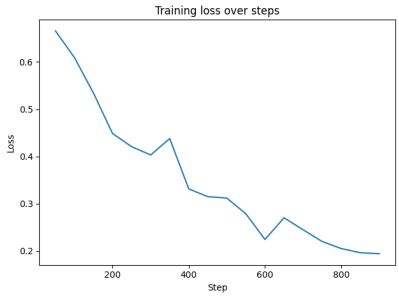

Figure 11: Training loss over steps for RoBERTa-large with data augmentation.

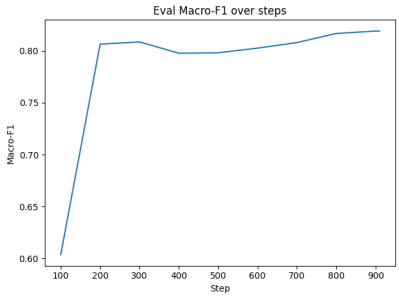

Figure 12: Validation macro-F1 over steps for RoBERTa-large with data augmentation.

GPT-4.1 Prompt-Based Classification. We additionally evaluate a prompt-based classification approach using GPT-4.1 in a zero-/few-shot setting on the English subset. This setup achieves an average accuracy of **0.61** and a macro-F1 score of **0.38**. Despite reasonable accuracy, the low macro-F1 indicates poor balance between classes, suggesting that the model struggles to reliably distinguish polarized from non-polarized content.

These results highlight the limitations of prompt-based large language models for this task when used without task-specific fine-tuning. In particular, predictions are sensitive to prompt formulation and tend to favor the majority class, leading to unstable and inconsistent performance compared to fine-tuned encoder models.

Sentence-Embedding Models. We evaluated a sentence-level approach based on the `sentence-transformers/all-mpnet-base-v2` model using the 5-fold cross-validation protocol. Across the five folds, the model achieved macro-F1 scores of 0.803, 0.793, 0.792, 0.782, and 0.826, respectively. This resulted in a mean macro-F1 of 0.799 with a standard deviation of 0.015, indicating relatively stable performance across folds. While this sentence-embedding-based approach does not outperform the best token-level models in the English setting, its consistent results suggest that it captures relevant global semantic information and serves as a strong baseline for further multilingual and hybrid modeling experiments.

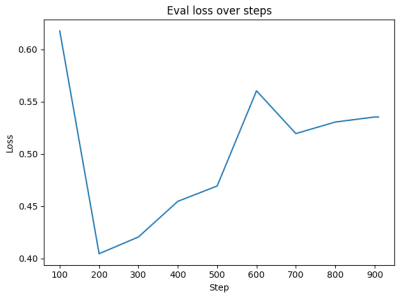

Figure 13: Validation loss over steps for RoBERTa-large with data augmentation.

Best Model on Validation Dataset. Based on the results of our exploratory fusion experiments, the hybrid architecture combining token-level representations from RoBERTa with sentence-level embeddings from MPNet emerged as our best-performing model. Using a 5-fold cross-validation protocol, the hybrid model achieved fold-level macro-F1 scores of 0.884, 0.904, 0.909, 0.920, and 0.878, resulting in a mean macro-F1 of 0.899 with a standard deviation of 0.016 on the training data. This indicates both strong overall performance and stable behavior across folds.

When evaluated on the validation set, the hybrid model achieved a macro-F1 score of approximately 0.83, confirming that the gains observed during cross-validation generalize beyond the training data. In addition to macro-F1, the model attained an accuracy of approximately 0.88, with balanced precision and recall across classes. Specifically, precision values exceeded 0.88 for both classes, while recall remained consistently high, reaching over 0.91 for the non-polarized class and approximately 0.81–0.90 for the polarized class.

The confusion matrices further indicated that the hybrid model substantially reduces false negatives for the polarized class compared to sentence-level or token-level models alone, while maintaining strong performance on the majority class. These results suggest that integrating global sentence-level semantics from MPNet with fine-grained token-level contextual representations from RoBERTa enables a more robust decision boundary, particularly for detecting minority-class instances.

Result Comparison. Table 4 summarizes the performance of a wide range of models and configurations evaluated on the validation set. Traditional machine learning baselines, such as Naive Bayes with count features and logistic regression with TF–IDF, provide reasonable but clearly inferior performance compared to transformer-based approaches, with macro-F1 scores below 0.70. Prompt-based inference using GPT-4.1 performs substantially worse, indicating that zero-shot prompting is insufficient for this task without task-specific fine-tuning.

Among transformer models, we observe a consistent improvement when scaling model capacity and incorporating data augmentation. RoBERTa-based models exhibit strong and stable performance, with RoBERTa-large outperforming its base variant and further gains obtained through augmentation strategies. In particular, RoBERTa-large with back-translation achieves the highest validation performance among non-cross-validation settings, reaching a macro-F1 of 0.844.

Cross-validation experiments further highlight the relative strengths of different architectures. While DeBERTa V3 models achieve competitive macro-F1 scores on the validation folds during training, with DeBERTa V3 large reaching 0.821, these improvements do not consistently translate to better generalization on the validation set. In fact, despite slightly stronger validation performance in some configurations, DeBERTa models were consistently outperformed by RoBERTa-large by over one percentage point on the validation set across all comparable settings.

A similar discrepancy between the performance observed during training and that observed on the validation set is observed for augmentation-based models. Although RoBERTa-large with back-translation achieves a macro-F1 of 0.844 during training, this configuration attains only 0.812 macro-F1 on the validation set, indicating a degree of overfitting to the training data. Notably, the *RoBERTa-large + LLM augmentation* configuration achieves a macro-F1 score of 0.935 under cross-validation on the training data; however, this strong validation performance does not translate to comparable generalization on the validation set, where the same model attains a macro-F1 of approximately 0.79. This further emphasizes the importance of evaluating generalization performance beyond validation metrics alone.

In general, these results suggest that strong performance on a single validation split does not necessarily guarantee superior generalization. RoBERTa-large-based models demonstrate more robust and consistent performance compared to alternative architectures, while the hybrid model achieves the highest macro-F1 under cross-validation. Consequently, our final model selection prioritizes architectures that exhibit stable performance across validation and cross-validation evaluations, rather than those that optimize a single validation metric in isolation.

Additional Monolingual Setting Motivated by the strong performance of RoBERTa-large on English, we evaluated a comparable large-scale Chinese model, hfl/chinese-roberta-wwm-ext-large (Cui et al., 2020). The model obtained an F1 score of 0.9164 on the validation set, suggesting that RoBERTa-style pretraining with whole-word masking is highly effective for Chinese as well.

Setup	Accuracy	Macro-F1
Naive Bayes (Count)	0.712	0.686
Logistic Regression (TF–IDF)	0.719	0.652
GPT-4.1 Prompt-Based	0.61	0.38
RoBERTa-base	0.79	0.76
RoBERTa-large	0.83	0.81
RoBERTa-large + LLM aug	0.85	0.85
RoBERTa-large + backtrans	0.88	0.844
MPNet + fold		0.79
DeBERTa V3 base + fold		0.804
DeBERTa V3 large + fold		0.821
RoBERTa-large + fold		0.8159
RoBERTa-large + LLM aug + fold		0.935
Hybrid Model + fold		0.89

Table 4: Results of Different Configurations on Validation Data

D.2 Multilingual Results

Table 5 presents per-language macro-F1 scores on the validation set for the multilingual models under the different experimental configurations. Overall, the XLM-RoBERTa baseline demonstrates competitive performance across a wide range of languages, with particularly strong results for high-resource languages such as Persian, Nepali, Punjabi, and Chinese. However, performance remains substantially lower for several low-resource or morphologically complex languages, including Khmer and Italian, highlighting the inherent difficulty of multilingual polarization detection.

The refined XLM-RoBERTa configuration yields improvements for a subset of languages, most notably Arabic, English, Persian, Polish, and Turkish. At the same time, these gains are not uniform: for some languages, such as German and Italian, performance slightly decreases compared to the baseline. This variability suggests that

Experiment	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
XLM-R - Baseline	0.7099	0.7361	0.8239	0.7184	0.7541	0.874	0.7842	0.8171	0.6442	0.5505	0.8453	0.85	0.7594	0.8699	0.7715	0.7374	0.7088	0.7907	0.8643	0.7123	0.7532	0.8738
XLM-R - Modified Config	0.727	0.7822	0.84	0.6834	0.7824	0.8843	0.7842	0.7562	0.5931	0.5778	0.8098	0.8399	0.7037	0.8599	0.8092	0.7483	0.6776	0.7593	0.8559	0.7561	0.7415	0.8785
XLM-R - Modified Config (Subset)	–	–	–	0.6869	0.7721	–	–	–	0.6528	–	–	–	–	–	0.7918	0.8002	0.7272	–	–	–	–	–
MPNet - Baseline	0.734	0.7718	0.8373	0.6997	0.7884	0.8874	0.7405	0.8032	0.6193	0.5627	0.8598	0.89	0.7375	0.8397	0.7519	0.7091	0.6787	0.7987	0.8273	0.72	0.7388	0.883
MPNet - Modified Config	0.7425	0.7681	0.8329	0.6961	0.8122	0.8602	0.7443	0.8006	0.6458	0.5594	0.8516	0.8599	0.7485	0.8298	0.7146	0.7374	0.6463	0.8104	0.8557	0.7477	0.6856	0.8831
mDeBERTa-v3-base weighted training	0.762	0.8091	0.816	0.6971	0.8133	0.8274	0.7606	0.8293	0.6324	0.6846	0.8254	0.8292	0.7602	0.8091	0.8091	0.7214	0.7015	0.7707	0.805	0.8173	0.71	0.8551
mDeBERTa-v3-base ensemble Focal Loss	0.7347	0.8071	0.8336	0.6997	0.7803	0.8424	0.7534	0.8803	0.6418	0.673	0.8734	0.8296	0.7642	0.79	0.7918	0.7312	0.6776	0.7994	0.8559	0.8261	0.7224	0.9019
Ensemble + Augmentation	0.7158	0.7848	0.8277	0.7014	0.7922	0.8739	0.8034	0.8375	0.6793	0.6394	0.8725	0.8496	0.8086	0.7698	0.804	0.7481	0.6822	0.7904	0.8133	0.8172	0.7277	0.883

Table 5: Per-language performance comparison across different multilingual training configurations.

global optimization strategies may benefit certain language groups while being less effective for others, likely due to differences in data size, linguistic structure, and label distribution.

Results obtained on the language subset largely follow the same trends observed in the full multilingual setting. For English, Italian, and Polish, performance remains comparable to that of the corresponding full multilingual configuration, indicating that the refined model retains reasonable robustness when trained on a smaller set of languages. Furthermore, performance improvements are particularly pronounced for Russian and Spanish, with Russian showing the most substantial gains among the subset languages. Nevertheless, the absence of consistent improvements across all subset languages further underscores the challenges of balancing multilingual generalization with language-specific optimization.

The sentence-embedding–based MPNet models exhibit competitive performance across several languages, particularly English and Chinese, where they match or exceed the XLM-RoBERTa baseline. However, their performance is more uneven across languages, with notable drops for lower-resource settings. This behavior suggests that while sentence-level semantic representations can be effective in multilingual contexts, they may be more sensitive to data scarcity and language-specific characteristics than token-level multilingual encoders.

Taken together, the results in Table 1 indicate that no single multilingual configuration consistently outperforms the others across all languages. Instead, performance varies substantially depending on both the modeling approach and the target language, highlighting the importance of considering language-specific behavior when designing and evaluating multilingual polarization detection systems.

In addition to the XLM-RoBERTa and MPNet configurations, Table 1 also reports results for more advanced training strategies based on mDeBERTa-v3 and ensemble models. The mDeBERTa-v3-base model trained with class-weighted loss achieves competitive performance across several languages, particularly Amhranic, Arabic, English, Hindi, Polish, and Turkish, indicating that explicit re-weighting can be beneficial for handling class imbalance in multilingual settings. However, its performance remains uneven across languages, with comparatively lower scores for languages such as Italian and German.

The ensemble variant of mDeBERTa-v3-base trained with focal loss further improves performance for a subset of languages, most notably Hindi, Chinese, Burmese , and Turkish, suggesting that focal loss can help emphasize harder or underrepresented examples. Nevertheless, these gains are not consistent across all languages, and in some cases performance slightly degrades compared to the weighted-loss variant. This variability indicates that while focal loss can be effective in multilingual contexts, its benefits are highly language-dependent and sensitive to data distribution.

Finally, the ensemble model combined with data augmentation yields strong results for several high-resource languages, including Persian, Hindi, Burmese, and Chinese, and achieves more balanced performance across a broader set of languages. At the same time, the augmented ensemble does not consistently outperform the best single-model configurations for all languages, and for some lower-resource languages the improvements remain limited. This observation aligns with earlier findings that augmentation-heavy and ensemble-based approaches may improve robustness for certain languages, but do not uniformly translate into superior generalization across the entire multilingual spectrum.

Overall, the results of the final configurations reinforce the conclusion that increasingly complex training strategies, such as loss re-weighting, focal loss, and ensembling, can yield localized improvements, but do not guarantee consistent gains across all languages. These findings highlight the trade-off between model complexity and cross-lingual robustness, and motivate the need for language-aware or adaptive multilingual modeling strategies.