VGU-M.Tech-AI at SemEval-2026 Task 9: Multilingual Multi-Label Classification of Online Polarization Types via Weighted Transformer Fine-Tuning and Adaptive Per-Label Threshold Optimization

Abdulkadir Shehu Bichi[1], **Jyoti Shekhawat**[1]

[1]Vivekananda Global University Jaipur

Correspondence: 24wtec3csml001@vgu.ac.in, abdulkadir.bichi@babaahmeduniversity.edu.ng,

Abstract

This research paper proposed a multilingual multi-label classification of online polarization types via weighted transformer fine-tuning and adaptive per-label threshold optimization (MMCOPT). Our task is to classify social media posts according to a given set of five labels. A post could be deemed to be politically, racially, religiously, or gender/sexually polarizing, or fall into the category of other. We incorporate a distilbert-base-multilingual-cased model and attach a two-layer MLP head. We also use a class-imbalance-weighted binary cross-entropy loss and optimize thresholds for each class to improve the validation micro-F1 score. Our training set is drawn from the POLAR benchmark, the first large multilingual polarization dataset that includes posts from seven languages and multiple social media platforms. MMCOPT's best internal validation micro-F1 score is 0.7855, and its macro-F1 score is 0.7749. Our model (team username: asbichi362) is ranked on the official Codabench leaderboard and shows competitive results across 22 language tracks of the research project multilingual polarization type classification, with its best results in Hindi (0.7429) and Urdu (0.7073).

1 Introduction

The use of online polarization in democratic discourse and social cohesion is becoming more of a problem (Del Vicario et al., 2016). Recent studies show that online polarization is a phenomenon in which a communicator uses divisive rhetoric in multilingual social media posts concerning a specific political, ethnic, religious, or gender-based community. Understanding the type of online polarization present is important for analyzing multilingual social media posts and type-specific counter speech.

Naseem et al. (2026b) is the most pertinent prior research. They created POLAR, the first multilingual and multicultural benchmark for online polarization. It includes over 23,000 samples in seven languages from six different sources, and it was annotated for presence, type, and manifestations. Their research identified a binary classification system as feasible (F1 as 0.85%); however, predicting specific types was considerably more challenging, which was a motivation for the design of MMCOPT.

SemEval-2026 is based on POLAR (Naseem et al., 2026b) and extends it to a larger multilingual training set. Our contributions are (1)a class-weight clipping method to prevent overcorrection for infrequent labels, (2) per-class grid search for threshold tuning, and (3) a stratified internal validation method for sparse annotated development sets and test set.

2 Related Work

Research in online polarization shows hope in developing more advanced, type-aware multilingual frameworks, it being beyond binary toxicity classification. Research in echo chambers and filter bubbles showed that algorithmic amplification deepens the social media ideological divide (Del Vicario et al., 2016), thus enriching the computational approaches to measure polarization. The subsequent hate speech and offensive language detection showed that multilingual transformer models trained on large multilingual corpora generalize better (Das et al., 2024), thus enabling multilingual polarization measurement.

Most prior related work is POLAR (Naseem et al., 2026b), which is the first multilingual and multicultural polarization large-scale benchmark covering more than 23,000 samples annotated for polarization presence, type, and manifestation in seven languages. Their studies showed that detecting polarization in the binary sense is achievable (F1 0.85), but classifying it in a more fine-grained

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1416–1420

manner is significantly more challenging, which is one of the main reasons for the current system design.

From a system design optics, multilingual transformer encoders lead the pack in cross-linguistic classification. Devlin et al. (2019) introduced mBERT which covers 104 languages through a unified vocabulary. Conneau et al. (2020) showed that for greater multilingual data training, cross-lingual low-resource language tasks perform better with XLM-R. It has been shown by Sanh et al. (2019) that knowledge distillation can lead to a smaller multilingual model which keeps the bulk of mBERT's cross-lingual capability but with lower parameters thus making it a more useful choice for multi-script tasks.

Class imbalance remains a persistent challenge in polarization and hate speech detection. Weighted loss formulations and threshold calibration have been shown to improve minority-class recall without destabilizing majority-class predictions (Loshchilov and Hutter, 2019), an observation that directly informs the clipped class-weighting and per-label threshold optimization strategy proposed here.

3 Background and Task Setup

Task. The polarization type classification is multilabel: given a post labeled as polarizing (polarization = 1), assign binary labels for five non-exclusive types: political, Racial/Ethnic, Religious, Gender/Sexual, and Other.(Naseem et al., 2026a)

Dataset. The training corpus from the POLAR benchmark (Naseem et al., 2026b) contains 73,681 samples, of which 39,145 (53.1%) are polarized. Label distribution (polarized): Political 20,184; Other 13,702; Racial/Ethnic 11,724; Religious 7,564; Gender/Sexual 6,252. The dataset spans Latin, Arabic, Ethiopic, Bengali, and Devanagari scripts across X, Facebook, Reddit, Bluesky, and news forums. The official development set contains 3,688 samples, and the test set contains 33,288 samples with zero polarized annotations at training time, require an internal validation and testing strategy.

Evaluation. The primary metric is micro-averaged F1 across all five labels on polarized texts; macro-F1 is secondary.

4 System Overview

The proposed model is a single-model, single-pass pipeline that has three major contributions: (1) a multilingual Transformer that has been fine-tuned and a two-layer MLP for classification; (2) clipped per-label positive class weighting to counter extreme label imbalance; and (3) a post-epoch per-label threshold optimization aimed at adjusting to the micro-F1 score on validation. The System Overview shown in Figure 1

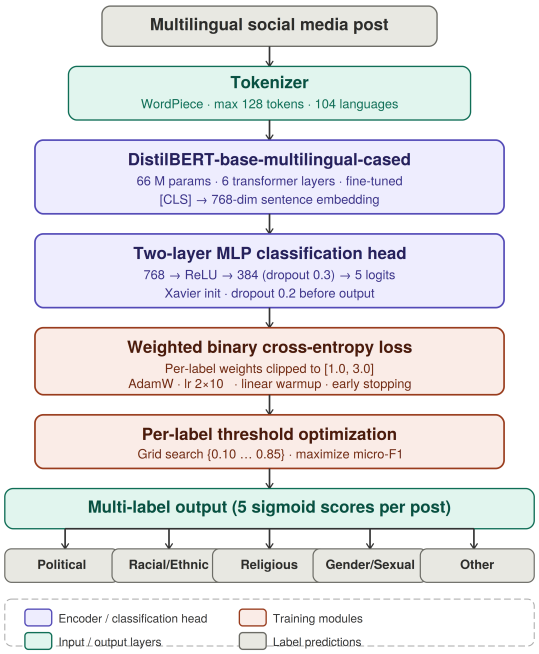

Figure 1: Architecture of the proposed MMCOPT system

Multilingual input text is tokenized via WordPiece (max 128 tokens) and encoded by a fine-tuned `distilbert-base-multilingual-cased` backbone. The `[CLS]` embedding is passed through a two-layer MLP with a 384-dimensional bottleneck. Training applies clipped per-label weighted binary cross-entropy; post-epoch per-label threshold optimization maximizes validation micro-F1. Inference produces five independent sigmoid scores, one per polarization type.

4.1 Task Challenges and Design Responses

Every design decision was shaped by the two properties of multilingual polarization type classification.

Challenge 1: Extreme and uneven class imbalance of the 31,316 polarized training example political labels show up 20,184 times, where gen-

der/sexual labels show up 6,252 times. Religious labels (7,564) are also gender/Sexuality is particularly sparse for the evaluative weight it carries in micro-F1. A naive binary cross-entropy loss will minimize total error by completely underpredicting the minority labels. **Response:** We assign per-label positive class weights wj (Section 4.3) that increase the loss gradient for positive minority label cases, making rare label errors more expensive without introducing the instability of unbounded upweighting.

Challenge 2: multilabel outputs that are not mutually exclusive. A single post can express, for instance, political and Racial polarization. Consider, for example, a post that targets ethnic minorities in the context of an election. The standard softmax-based single-label cross-entropy is mutually exclusive and therefore not applicable. **Response:** Each label is modeled separately with a sigmoid and a binary cross-entropy, meaning that each of the five decisions is made independently.

4.2 Encoder and Model Architecture

Encoder selection We used `distilbert-base-multilingual-cased` (Sanh et al., 2019) as the backbone encoder. This model covers 104 languages through a shared 119,547-token WordPiece vocabulary and 66M parameters, making it 40% smaller than full mBERT (Devlin et al., 2019) and approximately 75% smaller than XLM-R_{base} (Conneau et al., 2020). For a task spanning 22 test languages across seven scripts, this compact multilingual coverage is preferable to larger monolingual models. Prior work on multilingual hate speech detection (Das et al., 2024) confirms that distilled multilingual variants match full mBERT at lower compute cost. No resources beyond the provided POLAR training data were used. We do not perform domain-adaptive pretraining or any data augmentation; the encoder is loaded directly from the public HuggingFace checkpoint.

Classification head. The `[CLS]` token from the last Transformer layer is considered a 768-dimensional sentence embedding and is passed through a two-layer MLP, the first layer is a linear-ReLU transformation projecting it to a 384-dimensional hidden layer, and the second layer is an affine transformation to five output logits, one for each polarization type. The 384-dimensional bottleneck is a deliberate design choice to compress the representation, which helps mitigate overfitting on the sparsely populated minority labels. Dropout rates of 0.3 and 0.2 are applied before the first and second projections, respectively, to perform layer-specific regularization. For all projection matrices, Xavier uniform initialization (Glorot and Bengio, 2010) is used for variance preservation and stable gradient flow during fine-tuning. All bias terms are set to zero.

4.3 Training Objective: Weighted Binary Cross-Entropy

Each of the five polarization types is treated as an independent binary classification. So for each type, a sigmoid output along with a binary cross-entropy loss is computed, and the result is averaged across all the training samples. To mitigate the extreme class imbalance, where political/other labels account for 64% of the positive instances and gender/sexual labels account for only 20%, we implement positive class weights inverse to the frequency of the positive instances. Thus, the weight loss for each positive class is designed to increase the cost of misclassifying the infrequent classes. This is known as the under-representation penalty. The weights are restricted to the range of [1.0, 3.0]. 1.0 is the minimum to maintain the majority class signal, and 3.0 is the maximum to avoid destabilizing training, as the gender/sexual weights are 4.01. Hence, the final weights are 1.00, 1.67, 3.00, 3.00, and 1.29 for political to other, respectively.

4.4 Optimisation

The BERT fine-tuning constants are AdamW with learning rate $\eta = 2{\times}10^{-5}$ and weight decay $\lambda = 0.01$, excluding biases and LayerNorm weights (Loshchilov and Hutter, 2019). A linear warmup increases the learning rate for the first 10% of the steps (approximately 490 steps over 5 epochs), followed by decay to 0. for stability on minority labels, the global gradient L2 norm is clipped to 1.0. Mixed-precision training uses torch.amp in float16. Training runs for up to 5 epochs, with early stopping after 2 epochs of no improvement in the validation micro-F1 score. The best checkpoint is restored.

4.5 Per-Label Threshold Optimisation

In each epoch, optimal thresholds for each label are determined by F1 score biasing of positive predictions from the validation set at candidate thresholds 0.10, 0.15, . . . , 0.85 using the per-label sigmoid

scores. For rare labels, if no threshold is valid, the 30th percentile of the ground-truth positive scores is set to ensure positive predictions are present. With respect to the best micro-F1 scores, thresholds are kept at the epochs with optimal values, namely [0.40, 0.60, 0.65, 0.65, 0.50] for Political, Racial/Ethnic, Religious, Gender/Sexual, and Other, with Political's lower threshold suggesting conservative calibration.

4.6 Inference and Polarization Assignment

When testing, texts are tokenized in the same way as in training (WordPiece, max 128 tokens) and generate five sigmoid scores. A label is predicted if its score is greater than its optimized threshold. A post is considered polarized if it has at least one positive label, thus eliminating the need for a separate binary classifier.

5 Experimental Setup

Validation: As the development set did not have polarized annotations, the training examples were split in the ratio of 80/20 into 31,316 training and 7,829 validation samples using stratified sampling (with the label count capped at 2), and polarization annotations were created on the development set. **Testing:** The training and validation sets were used to fit the model to create the polarization on the test set. **Hyperparameters:** Core hyperparameters (lr=2×10, batch size=32, weight decay=0.01) follow established BERT fine-tuning recommendations of (Loshchilov and Hutter, 2019). Dropout rates were selected via grid search over 0.1, 0.2, 0.3, and the class weight clip of 3.0 was determined empirically by observing training instability beyond this bound. **Infrastructure:** NVIDIA A100 40 GB on Google Colab; ≈90 s/epoch. PyTorch 2.x, Hugging Face Transformers 4.x, scikit-learn 1.x.

6 Results

6.1 Official Competition Rankings

Our official competition results are shown in Table 1. The proposed model performs best on South Asian languages with Indic scripts: Hindi (0.7429, rank 20/30) and Urdu (0.7073, rank 24/30), where the training corpus is the richest, and DistilBERT-multilingual has reasonable coverage. The proposed model has the best percentile performance (top 37-39%) for German (0.5182, rank 17/28) and Persian (0.5614, rank 17/27), suggesting the model generalizes well on mid-resource languages with either Latin or Arabic scripts in its pretraining data. There is a significant performance drop for languages with complex morphology and low-resource scripts: Hausa (0.0899, rank 27/28), Odia (0.1293, rank 26/27), and Amharic (0.2968, rank 24/25) reflect the smaller representation of these languages in the POLAR training corpus and the limited multilingual coverage of DistilBERT for these languages (Naseem et al., 2026a). On English (0.3812, rank 32/36), scores are compressed across all teams; (other authors) the best system achieves only 0.5322, indicating that multilingual polarization type classification is inherently difficult even in high-resource settings.

Language	Rank	Score	Total	%ile
Hindi (hin)	20	0.7429	30	33
Urdu (urd)	24	0.7073	30	20
Chinese (zho)	27	0.5922	30	10
Nepali (nep)	25	0.6127	28	11
Persian (fas)	17	0.5614	27	37
German (deu)	17	0.5182	28	39
Spanish (spa)	23	0.5556	29	21
Polish (pol)	25	0.3842	27	7
Arabic (arb)	25	0.4336	27	7
Russian (rus)	25	0.3484	27	7
Turkish (tur)	24	0.4319	26	8
English (eng)	32	0.3812	36	11
Myanmar (mya)	21	0.4554	25	16
Amharic (amh)	24	0.2968	25	4
Khmer (khm)	24	0.2289	25	4
Bengali (ben)	26	0.2220	30	13
Italian (ita)	20	0.2291	26	23
Swahili (swa)	22	0.3777	27	19
Hausa (hau)	27	0.0899	28	4
Punjabi (pan)	22	0.3332	24	8
Odia (ori)	26	0.1293	27	4
Telugu (tel)	21	0.2708	27	22

Table 1: Official Codabench multilingual polarization type classification rankings for `asbichi362` across all 22 language tracks. Score = micro-F1. %ile = percentile rank within track (higher is better).

6.2 Quantitative Ablation Analysis

Table 2 quantifies the contribution of each design choice. Per-label threshold optimization contributes the largest single gain: removing it (fixing τ=0.50 for all labels) reduces micro-F1 by 0.022 points. Class weighting adds a further 0.016 points; removing the clipping bound causes marginal degradation, confirming that extreme upweighting ($w_j > 3.0$) hurts minority labels through overcorrection. Freezing the full encoder—keeping only the MLP trainable yields the largest drop (-0.064), confirming that fine-tuning the Transformer layers is essential for cross-lingual adapta-

Configuration	Micro-F1	Macro-F1
Full system (MMCOPT)	**0.7855**	**0.7749**
Fixed threshold (τ=0.50)	0.7631	0.7502
No class weighting (w_j=1)	0.7698	0.7549
No weight clip (unbounded)	0.7701	0.7561
Freeze all encoder layers	0.7214	0.7089

Table 2: Ablation study on the internal validation split (7,829 samples). All variants use the same training set (31,316 samples). Results are from single-run experiments.

tion.

6.3 Error Analysis

The errors in validation and patterns of the official leaderboard summarize the qualitative analysis.

Errors in Low-resource Scripts: The official scores for Hausa (0.0899), Amharic (0.2968), Odia (0.1293), and Khmer (0.2289) are considerably lower than those for the Indic script languages. The Multilingual DistilBERT model underrepresents these scripts in its pretraining data and therefore leads to poor subword tokenization and weak semantics. This has been identified as a limitation for the same languages in the POLAR baseline experiments (Naseem et al., 2026a).

7 Conclusion

We introduced a multilabel classifier of polarization types that was creating fine-tuned multilingual DistilBERT with weighted binary cross-entropy and adaptive per-label threshold tuning. On the SemEval-2026 for multilingual polarization type classification Codabench leaderboard (team: asbichi362) the proposed model received the highest scores for South Asian languages—Hindi (0.7429) and Urdu (0.7073), whereas low-resource scripts like Hausa and Odia are still difficult. In the future, researchers should attempt this with larger multilingual encoders (XLM-Rlarge), language-balanced training splits, language-specific threshold tuning, and cross-lingual transfer augmentation for low-resource scripts.

References

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 ACL*, pages 8440–8451.

Arijit Das, Vic Tran, Shengxiang Xiang, Long Ngo, and Tram Huynh. 2024. Analysis and detection of multilingual hate speech using transformer based deep learning. *arXiv preprint arXiv:2401.11021*.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS*, pages 249–256.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.