Taien at SemEval-2026 Task 9: Multilingual Polarization Detection Using Transformer-based Models

Saida Islam Taien
Dept. of CSE
BGC Trust University Bangladesh
Chittagong-4381, Bangladesh
saidataien@gmail.com

Palash Hossen
Dept. of CSE
University of Chittagong
Chittagong-4331, Bangladesh
palash.cu.hossen@gmail.com

Abstract

The rapid growth of online communication platforms has intensified ideological polarization in political and socioeconomic discourse, making multilingual polarization detection a critical yet challenging task in natural language processing. Unlike traditional sentiment or stance analysis, polarization detection requires modeling ideological extremity and nuanced opinion divergence across diverse linguistic settings. In this work, we describe our system for Subtask 1 of SemEval 2026 Task 9, where we develop a multilingual transformer-based approach covering 22 languages. Our system leverages two pretrained multilingual models, XLM-RoBERTa-base and mDeBERTa-v3-base, which are fine-tuned on the shared task dataset. To improve predictive robustness and cross-lingual stability, we employ a probability-level ensemble strategy that combines the outputs of both models. Experimental results show that transformer-based models significantly outperform traditional machine learning baselines, while the ensemble method consistently improves performance stability across languages. These findings demonstrate the effectiveness of multilingual transformer ensembles for large-scale polarization detection.

1 Introduction

Polarization detection identifies whether a text expresses ideologically polarized viewpoints rather than simple sentiment. Unlike sentiment analysis, it focuses on extremity, oppositional framing, and implicit bias. We participate in Subtask 1 of SemEval 2026 Task 9, addressing multilingual polarization across 22 languages. The task is critical due to online discourse growth, where polarized narratives shape opinion, spread misinformation, and deepen divisions. Multilingual settings add complexity as polarization varies across linguistic and cultural contexts. This work builds on SemEval-2026 Task 9 and the POLAR dataset (Naseem et al., 2026a,b).

Our system uses a multilingual transformer-based approach, fine-tuning XLM-RoBERTa-base and mDeBERTa-v3-base on the shared task dataset. To enhance robustness, we ensemble the models' softmax outputs via soft voting, capturing complementary contextual representations and improving stability in ambiguous or implicit cases. The same preprocessing and fine-tuning pipeline is applied across all languages to ensure fairness and reproducibility.

Transformer-based multilingual representations outperform classical baselines in modeling implicit polarization, and the ensemble consistently stabilizes predictions, particularly for low-resource languages. Nonetheless, class imbalance and subtle linguistic differences affect performance, with implicit or context-dependent polarization remaining challenging. Overall, our approach achieves competitive results across languages.

The implementation is fully reproducible; source code and model configurations are released to support future research: [1].

2 Background

2.1 Dataset

We use the POLAR dataset introduced by (Naseem et al., 2026b), which is provided as part of the SemEval-2026 Task 9 shared task. It contains short texts labeled as polarized (1) or not (0). Example:

> "I completely disagree with the policy; it is unfair and biased." → Polarization = 1

2.1.1 Language Coverage

The dataset covers 22 languages: Amharic, Arabic, Bengali, German, English, Persian, Hausa, Hindi, Italian, Khmer, Burmese, Nepali, Odia, Punjabi, Polish, Russian, Spanish, Swahili, Telugu, Turkish,

[1] https://github.com/saida-taien/Multilingual-Polarization-22lang/tree/main

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2070–2077

Urdu, and Chinese. It includes both high- and low-resource languages across diverse scripts. Predefined train, validation, and test splits are provided (Naseem et al., 2026a). The texts are collected from social media and online news sources on topics such as elections, conflicts, gender rights, and migration.

2.1.2 Dataset Statistics

Language	Train	Dev	Test
Amharic (amh)	3332	166	1501
Arabic (arb)	3380	169	1521
Bengali (ben)	3333	166	1501
German (deu)	3180	159	1432
English (eng)	3222	160	1452
Persian (fas)	3295	164	1484
Hausa (hau)	3651	182	1664
Hindi (hin)	2744	137	1236
Italian (ita)	3334	166	1539
Khmer (khm)	6640	332	2988
Burmese (mya)	2889	144	1301
Nepali (nep)	2205	100	903
Odia (ori)	2368	118	1066
Punjabi (pan)	1700	100	809
Polish (pol)	2391	119	1077
Russian (rus)	3348	167	1508
Spanish (spa)	3305	165	1488
Swahili (swa)	6991	349	3147
Telugu (tel)	2366	118	1066
Turkish (tur)	2364	115	1093
Urdu (urd)	3563	177	1606
Chinese (zho)	4280	214	1927

Table 1: Train, Dev, and Test sizes for the POLAR@2026 multilingual polarization dataset.

This dataset provides a robust benchmark for multilingual polarization detection, emphasizing cross-lingual transfer, low-resource performance, and implicit ideological cues.

2.2 Related Work

Social media and news platforms often exhibit polarized viewpoints affecting political stability, social cohesion, and policy. Detecting polarization automatically is a key NLP challenge, especially in multilingual low-resource settings where linguistic diversity and data scarcity are hurdles. Pre-trained multilingual transformers learn shared representations across languages, mitigating these issues (Pikuliak et al., 2021).

Classical ML methods such as SVM, Naive Bayes, and Logistic Regression using bag-of-words, TF-IDF, or n-grams have been strong baselines in low-resource multilingual tasks (Das et al., 2023). Though effective with limited data, they cannot capture deep contextual semantics needed for polarization detection. Deep learning models, including CNNs, LSTMs, and hybrid CNN-LSTM architectures, improved hierarchical representation learning and performed well in bilingual and code-mixed sentiment tasks (Roy, 2024), but require large labeled datasets and careful tuning.

Pre-trained transformers like BERT, XLM-RoBERTa, and mDeBERTa leverage bidirectional contextual embeddings, offering strong cross-lingual transfer for limited-data scenarios (Conneau et al., 2020b). Fine-tuned multilingual transformers outperform traditional and shallow neural baselines in sentiment and text classification (Hu et al., 2023; Kumar et al., 2024; Conneau et al., 2020a; Dai et al., 2023), and ensemble strategies further improve robustness and accuracy (Wang et al., 2021; Ali and Eshete, 2020).

Polarization detection goes beyond sentiment, capturing ideological extremity and stance divergence. Early work analyzed network structures and clustering to show sustained polarized discourse (Conover et al., 2021). NLP approaches treat polarization as text classification, integrating stance detection and contextual embeddings to detect implicit cues (Garimella et al., 2018). Shared tasks like SemEval advance multilingual research, addressing class imbalance and cross-domain robustness (Mohammad et al., 2016; Schlechtweg et al., 2020; Wang et al., 2023).

Recent studies show polarization's societal impact: (Ali et al., 2025) examines social media's role in conflict dynamics and digital peacebuilding, and (Banim, 2025) shows how algorithm-driven information bubbles amplify divisive narratives.

Building on these works, this study provides a strong baseline for multilingual polarization detection across 22 languages in SemEval-2026 Task 9. We fine-tune XLM-RoBERTa and mDeBERTa individually and in an ensemble, addressing class imbalance, low-resource issues, and cross-lingual generalization, offering a replicable benchmark for future research.

3 System Overview

We propose a multilingual polarization detection system designed to operate consistently across 22 languages under data imbalance and morphological diversity. The final submission is a probability-level ensemble of two pretrained multilingual transformer encoders: **XLM-RoBERTa-base** (Conneau et al., 2020c) and **mDeBERTa-v3-base** (He et al., 2021). The system is fully language-independent and uses no external task-specific resources beyond the provided training data and publicly available pretrained models.

3.1 Task Formulation

We model multilingual polarization detection as a supervised binary classification task. Given an input sentence x, the system predicts a label:

$$y \in \{0, 1\} \tag{1}$$

where 0 denotes non-polarized content and 1 denotes polarized content. The model is trained using cross-entropy loss, while Macro-F1 is used as the primary validation metric for model selection.

3.2 Input Processing and Encoding

All languages are processed using a uniform preprocessing pipeline consisting of Unicode normalization, whitespace normalization, and removal of encoding artifacts. No language-specific stemming, stop-word removal, or filtering is applied.

Input sentences are tokenized using subword segmentation (SentencePiece/BPE). For a sentence x, tokenization produces:

$$x \rightarrow (input\_ids, attention\_mask) \tag{2}$$

$$input\_ids \in R^L, \quad attention\_mask \in \{0, 1\}^L \tag{3}$$

$$L = 128 \tag{4}$$

3.3 Parallel Transformer Modeling

The encoded input is forwarded in parallel to two pretrained multilingual transformer encoders. Each encoder produces a contextual sentence representation, which is passed to a task-specific linear classification head. Given a representation H, logits are computed as:

$$z = WH + b \tag{5}$$

Here, H denotes the contextual embedding of the input sentence, while W and b are the trainable weight matrix and bias vector of the linear classifier. This transformation maps the sentence representation to class-specific scores.

Training is performed using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{2} y_i \log(p_i) \tag{6}$$

The loss measures the discrepancy between the predicted class probabilities and the true label, enabling end-to-end fine-tuning of both the encoder and classification head.

3.4 Probability-Level Ensemble

Each fine-tuned encoder produces a probability distribution over the two classes. The final ensemble probability is computed via soft voting:

$$P_{final} = \frac{1}{2}\left(P_{XLM} + P_{DeBERTa}\right) \tag{7}$$

This averaging combines the confidence scores of both models, allowing complementary representational strengths to contribute to the final decision. The predicted label is obtained as:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} P_{final}(c) \tag{8}$$

The class with the highest averaged probability is selected as the final prediction. Probability-level averaging reduces prediction variance while preserving complementary information from both encoders.

3.5 System Configurations and Challenges

Three configurations are evaluated: fine-tuned XLM-RoBERTa-base, fine-tuned mDeBERTa-v3-base, and their probability-level soft-voting ensemble, which is used as the final model. All models share the same preprocessing and classification setup to ensure fair comparison. The ensemble improves prediction stability by combining the strengths of individual models.

Three main challenges are present: multilingual data imbalance, morphological variation, and implicit contextual polarization. These are addressed through multilingual pretraining, subword tokenization, and contextual transformer representations. However, no explicit class imbalance handling is applied, which remains a limitation. The approach improves generalization and stability across languages, especially in low-resource settings.

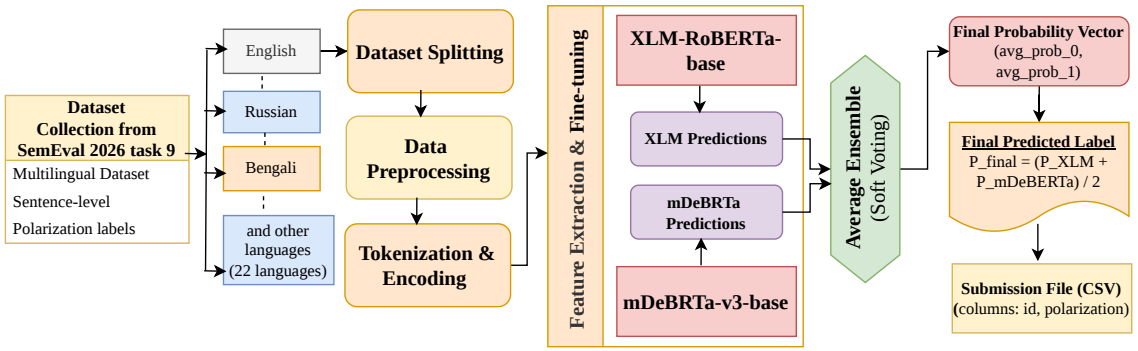

Figure 1: Proposed Multilingual Classification Architecture

Figure 1 illustrates the overall architecture of the proposed system. Additional training details, implementation specifications, and extended descriptions are provided in Appendix A.

4 Experimental Setup

We describe dataset usage, preprocessing, training configuration, implementation, and evaluation protocol for reproducibility.

4.1 Dataset Splits

We use the official SemEval-2026 Task 9 dataset (22 languages). For each language, the provided training data is split via stratified sampling into 90% train and 10% validation sets, preserving class distribution. The validation set is used for hyperparameter tuning and checkpoint selection. Official development and test sets are used strictly for final evaluation; no test data is used during training or selection.

4.2 Preprocessing and Encoding

All languages follow identical preprocessing: newline removal, whitespace normalization, Unicode normalization, and encoding cleanup. No stopword removal or language-specific processing is applied.

Models use pretrained subword tokenizers with maximum sequence length $L = 128$, using truncation and padding. Attention masks prevent padded tokens from influencing self-attention. Max length is 128 tokens for efficiency, which may truncate context.

4.3 Training Configuration

All models are trained under identical hyperparameters selected based on validation performance. Evaluation is performed after each epoch, retaining the checkpoint with highest validation weighted F1.

Optimizer	AdamW
Learning Rate	2e-5
Epochs	3
Train Batch Size	16
Eval Batch Size	32
Weight Decay	0.01
Max Length	128
Model Selection	Best validation weighted F1

Table 2: Hyperparameter configuration for all models.

4.4 Implementation Details

Experiments are implemented using PyTorch and HuggingFace Transformers (see Appendix B.4) on a single NVIDIA GPU. Public pretrained checkpoints of XLM-RoBERTa-base and mDeBERTa-v3-base are used. No additional corpora, lexicons, or external resources are incorporated.

4.5 Evaluation Metrics

Following the official SemEval-2026 Task 9 protocol, weighted F1-score is the primary metric due to class imbalance. Accuracy, Precision, Recall, and F1-score are also reported. Weighted F1 is used for model comparison and checkpoint selection.

See Appendix B for more details.

5 Results

5.1 Official Submission Performance

We report the performance of our final submitted system on the official SemEval-2026 Task 9 test

sets across 22 languages. The primary evaluation metric is weighted F1-score.

Our final submission (soft-voting ensemble of XLM-RoBERTa-base and mDeBERTa-v3-base) achieves an average weighted F1-score of **0.788** and an average accuracy of **0.822** across all languages.

The system achieves 1st rank in Burmese (F1 = 0.8913), and top-5 performance in Persian (rank 2), Hausa (rank 5) and Swahili (rank 3). Performance varies across languages, reflecting differences in class distribution, linguistic structure, and implicit polarization patterns. Lower scores are observed in languages such as Italian (F1 = 0.5798) and Odia (F1 = 0.7309), suggesting challenges related to subtle stance expression and data sparsity.

Complete per-language results and leaderboard positions are provided in Appendix C.

5.2 Comparison with Individual Models

On validation data, XLM-RoBERTa-base achieves an average weighted F1 of 0.781, while mDeBERTa-v3-base achieves 0.774. The soft-voting ensemble improves performance to 0.788, indicating complementary strengths between multilingual pretraining and disentangled attention mechanisms. The ensemble consistently improves robustness across both high-resource and low-resource languages.

5.3 Analysis of Design Decisions

We conduct controlled ablation experiments on the **official development set** to quantify the contribution of key design choices. These analyses are separate from the official test submission; all test results reported in Section 5.1 correspond to the submitted system without modification.

Fusion Strategy. Probability-level soft voting improves Macro-F1 by approximately **+0.9** compared to decision-level majority voting. Averaging confidence distributions enables more reliable resolution of borderline and implicitly polarized instances.

Checkpoint Selection. Selecting checkpoints based on best validation Macro-F1 consistently outperforms using final-epoch weights, yielding higher average performance and reduced cross-language variance, indicating improved generalization.

Language Variation. Performance differences reflect linguistic factors such as morphological complexity, flexible syntax, and implicit discourse structure. The ensemble consistently reduces cross-lingual variance relative to individual models, demonstrating improved robustness.

5.4 Error Analysis

Validation-set analysis reveals conservative behavior, with more false negatives than false positives, indicating missed polarization is more common than over-prediction.

Errors fall into three categories. **Non-literal Expressions:** divergence between surface meaning and intended stance requires pragmatic inference, leading to frequent misclassification. **Implicit Polarization:** indirect framing or subtle evaluative cues are often undetected without explicit sentiment markers. **Context Dependence:** sentence-level inputs lack broader discourse context, a limitation amplified in low-resource languages.

Overall, errors cluster in implicit and discourse-sensitive cases. While probability-level ensembling improves robustness, modeling subtle polarization remains challenging. Implicit and context-dependent polarization remains challenging.

6 Ethical Considerations

Polarization detection affects sensitive socio-political domains and misclassification may misrepresent viewpoints. The model could be misused for automated content filtering or suppressing dissent. Bias arises from class imbalance, low-resource languages, and pretrained transformer limitations. Deployment requires transparency and human oversight. Minority opinions may be disproportionately impacted. Future work should address fairness and bias mitigation for responsible multilingual use.

7 Conclusion

We propose a multilingual polarization detection system using XLM-RoBERTa and mDeBERTa with probability-level ensembling, achieving strong Macro-F1 across languages. Ablation results show soft voting and checkpointing improve robustness. Limitations include class imbalance, 128-token input length, sentence-level scope, and higher computational cost. Future work explores adapters, LoRA, improved context modeling, adaptive ensembles, and large language models, with further evaluation on fairness, bias, and comparisons with mmBERT and Glot500. Despite its simplicity, the ensemble shows strong multilingual performance.

References

Abdullah Ali and Birhanu Eshete. 2020. Best-effort adversarial approximation of black-box malware classifiers. *Preprint*, arXiv:2006.15725.

Adem Chanie Ali, Seid Muhie Yimam, Abinew Ali Ayele, Chris Biemann, and Martin Semmann. 2025. Silenced voices: social media polarization and women's marginalization in peacebuilding during the northern ethiopia war. *i-com*, 24(2):407–432.

G. Banim. 2025. Very large online platforms—how big is your polarization footprint? towards a metric to give eu citizens transparency around an online systemic risk driving conflict in our societies. *Build Up*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020c. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2021. Political polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):89–96.

Yi Dai, Hao Lang, Yinhe Zheng, Fei Huang, and Yongbin Li. 2023. Long-tailed question answering in an open world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6362–6382, Toronto, Canada. Association for Computational Linguistics.

Rajesh Kumar Das, Mirajul Islam, Md Mahmudul Hasan, Sultana Razia, Mocksidul Hassan, and Sharun Akter Khushbu. 2023. Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. *Heliyon*, 9(9):e20281.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 913–922, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

Dou Hu, Lingwei Wei, Yaxin Liu, Wei Zhou, and Songlin Hu. 2023. Ucas-iie-nlp at semeval-2023 task 12: Enhancing generalization of multilingual bert for low-resource sentiment analysis. *Preprint*, arXiv:2306.01093.

Ankit Kumar, Shivam Mishra, and Anil Singh. 2024. Multilingual hate speech and offensive language detection using transformer-based models. *Scientific Reports*, 14(1):60210.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, 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)*, San Diego, CA, USA. 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.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Pradeep Kumar Roy. 2024. Deep ensemble network for sentiment analysis in bi-lingual low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(1).

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi.

2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023. KnowComp at SemEval-2023 task 7: Fine-tuning pre-trained language models for clinical trial entailment identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1–9, Toronto, Canada. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Ivan Titov, Shay B. Cohen, Zhilin Lei, and Ting Liu. 2021. A closer look into the robustness of neural dependency parsers using better adversarial examples. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2344–2354, Online. Association for Computational Linguistics.

A Extended Implementation Details

A.1 End-to-End Inference Algorithm

The inference pipeline proceeds as follows:

1. Receive input sentence x.
2. Apply Unicode and whitespace normalization.
3. Tokenize using subword segmentation (maximum length 128).
4. Generate contextual representations using:
   - XLM-RoBERTa-base,
   - mDeBERTa-v3-base.
5. Compute logits using linear classification heads.
6. Apply Softmax to obtain class probabilities.
7. Average probabilities across models.
8. Output $\arg\max$ prediction.

A.2 Implementation Specifications

The system is implemented using PyTorch and the HuggingFace Transformers library. Both encoders consist of 12 transformer layers with hidden dimension 768 and multi-head self-attention. Task-specific linear classification heads are added during fine-tuning. No additional corpora, lexicons, translation systems, or language-specific heuristics are used.

B Additional Training Details

B.1 Hardware

Experiments are conducted on a single NVIDIA GPU with mixed-precision training enabled to reduce memory usage and improve computational efficiency.

B.2 Reproducibility Settings

A fixed random seed is applied for Python, NumPy, and PyTorch to ensure deterministic behavior where possible. Data shuffling and stratified splitting are performed using the same seed across experiments.

B.3 Training Stability

Gradient clipping is applied to prevent exploding gradients. No language-specific reweighting or sampling strategies are used. All languages follow identical preprocessing and optimization settings to ensure controlled comparison.

B.4 Evaluation Metrics

The evaluation metrics are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Weighted F1-score is used as the primary metric for model comparison.

B.5 Software Versions

The implementation is developed using PyTorch (v2.x), HuggingFace Transformers (v4.x), and Python (v3.x). Exact version numbers and environment configuration files will be released with the code repository to ensure full reproducibility.

C Detailed Scores and Leaderboard Positions

This section reports the official test results of our final submission across all 22 languages. Weighted F1-score (F1 Macro) is the primary ranking metric used in the competition. The system achieves an

Language	Acc	Prec	Rec	$\mathbf{F1}_{Bin}$	$\mathbf{F1}_{Macro}$	$\mathbf{F1}_{Micro}$	Rank
Amharic	0.8095	0.8319	0.9295	0.8780	0.7217	0.8095	35
Arabic	0.8297	0.8226	0.7900	0.8060	0.8271	0.8297	15
Bengali	0.8301	0.8225	0.7615	0.7908	0.8239	0.8301	35
German	0.7193	0.7034	0.7157	0.7095	0.7190	0.7193	15
English	0.8023	0.7412	0.7092	0.7248	0.7853	0.8023	35
Persian	0.8720	0.9084	0.9199	0.9141	0.8314	0.8720	2
Hausa	0.9380	0.7417	0.6400	0.6871	0.8263	0.9380	5
Hindi	0.9094	0.9368	0.9582	0.9474	0.8109	0.9094	14
Italian	0.5988	0.6149	0.4080	0.4905	0.5798	0.5988	32
Khmer	0.9321	0.9416	0.9864	0.9634	0.7418	0.9321	7
Burmese	0.8932	0.9185	0.8926	0.9054	0.8913	0.8932	1
Nepali	0.8948	0.8904	0.9002	0.8953	0.8948	0.8948	31
Odia	0.8068	0.7462	0.4851	0.5880	0.7309	0.8068	36
Punjabi	0.7515	0.7008	0.8524	0.7692	0.7501	0.7515	34
Polish	0.7892	0.7383	0.7694	0.7535	0.7847	0.7892	30
Russian	0.8103	0.7770	0.5111	0.6166	0.7453	0.8103	39
Spanish	0.7628	0.7553	0.7687	0.7620	0.7628	0.7628	31
Swahili	0.7985	0.7893	0.8163	0.8026	0.7985	0.7985	3
Telugu	0.8837	0.8919	0.8822	0.8871	0.8836	0.8837	11
Turkish	0.7768	0.7846	0.7873	0.7860	0.7763	0.7768	23
Urdu	0.8064	0.8582	0.8636	0.8609	0.7713	0.8064	25
Chinese	0.8760	0.8853	0.8681	0.8766	0.8760	0.8760	36

Table 3: Language-wise performance and leaderboard rank of the proposed system on the official test set.

overall average weighted F1-score of 0.788 and an average accuracy of 0.822.

Table 3 presents the language-wise performance of our official submission on the test set, including accuracy, precision, recall, binary F1, macro-F1, micro-F1, and leaderboard rank.

C.1 Performance Distribution Analysis

Across languages, performance ranges from 0.5798 (Italian) to 0.8948 (Nepali) in weighted F1-score. The system achieves first rank in Burmese, and top-5 ranking in Persian, Hausa and Swahili. Performance variation reflects differences in class imbalance, morphological complexity, and implicit polarization patterns.

Languages with higher recall (e.g., Hindi, Khmer, Persian) indicate effective detection of polarized content, whereas languages such as Odia and Russian show lower recall, contributing to higher false negative rates.

Overall, the ensemble demonstrates stable multilingual generalization while highlighting persistent challenges in detecting implicit and context-dependent polarization.

C.2 Model-wise Comparative Analysis and Final Discussion

A language-wise comparison highlights complementary strengths between the two transformer architectures. mDeBERTa-v3-base shows comparatively stronger performance in morphologically rich and structurally flexible languages such as Persian, Hindi, and Khmer, indicating that its disentangled attention mechanism effectively models positional variations and long-range dependencies.

XLM-RoBERTa-base demonstrates consistent behavior across high-resource languages including English, Spanish, and German, benefiting from large-scale multilingual pretraining and strong cross-lingual semantic alignment.

In lower-resource settings such as Amharic, Odia, and Hausa, individual models exhibit higher prediction variance, while the probability-level ensemble mitigates this instability by averaging confidence scores and reducing model-specific bias.

Overall, the results confirm that although each model has distinct representational advantages, their combination provides the most stable and generalizable performance across diverse languages.