Detecting Multilingual, Multicultural and Multievent Online Polarization Final Report

SemEval-2026 Task 9 - Subtask 1

Zeynep Şahin (30553), Suat Emre Karabıçak (30649), Sıla Horozoğlu (30916), Korhan Erdoğdu (30838)

1. Introduction

Over the last decade, social media has become one of the main places where people argue about politics, identity, religion, migration and other sensitive topics. These discussions often shift from simple disagreement to a much sharper pattern. People strongly align with an “us” and speak about “them” in hostile, stereotyping or even dehumanizing ways. Automatically identifying such polarized content is important for understanding how online discussions become fragmented and how attitudes differ across languages and communities.

In this project, we address SemEval-2026 Task 9: Multilingual Polarization Detection, specifically focusing on Subtask 1, which frames the problem as a binary classification task. Given a short social media post, the goal is to determine whether it contains polarized language. The dataset provided for this task supports 22 languages, covering a wide geographic and linguistic area. Polarization detection differs from related tasks such as hate speech or offensive language detection. While those tasks identify harmful content, polarization detection focuses on identifying language that divides audiences into opposing groups, often through us-versus-them framing, moral absolutism, or emotional intensity.

Our methodology spans a broad range of architectures, progressing from traditional machine-learning baselines to state-of-the-art multilingual transformers. We first establish a strong baseline using TF-IDF features with Logistic Regression, achieving 68.80% macro F1. We then evaluate neural sequence models, including CLSTM and BiLSTM-based architectures, where incorporating attention and explicit language embeddings improves performance up to 76.50% macro F1. Finally, we fine-tune several multilingual transformer models, with XLM-RoBERTa-large and mDeBERTa-v3-base outperforming other single models. Our best performance is obtained with a weighted ensemble of five transformer models, reaching 79.69% macro F1, which highlights the effectiveness of combining models whose performance varies across languages to reduce language-specific weaknesses.

2. Related Work

Our study builds upon key research in multilingual NLP, stance detection and cross-lingual text classification. As noted by AlDayel and Magdy (2021), stance detection and polarization detection tasks share similar characteristics as both require understanding not only the sentiment but also the positioning of text with respect to opposing groups. This overlap allows us to leverage stance detection methodologies. Mohammad et al. (2016) established

important baselines for stance detection in their SemEval-2016 shared task, demonstrating that TF-IDF representations combined with linear classifiers provide strong baselines for social media text classification. We adopt a similar baseline approach to quantify improvements from more sophisticated methods.

Before the development of transformer-based architectures, recurrent and hybrid neural models represented the state of the art for sentence-level text classification. Zhou et al. (2015) proposed the C-LSTM model, which combines convolutional layers for local n-gram feature extraction with LSTM layers for modeling long-range dependencies, achieving better performances than both separate models on sentiment classification tasks. Subsequently, attention-based bidirectional LSTM models were developed to allow the network to focus on the most informative tokens in a sequence without relying on external linguistic features and other NLP systems, outperforming other methods (Zhou et al., 2016). These architectures serve as strong pre-transformer baselines and motivate our exploration of CLSTM and BiLSTM+Attention models alongside modern decoder-only transformer approaches.

AlDayel and Magdy (2021) further provide a comprehensive survey of stance detection methods on social media, demonstrating that transformer-based models substantially outperform traditional machine learning approaches. Conneau et al. (2020) introduced XLM-RoBERTa, which achieves strong cross-lingual transfer by training on 100 languages with 2.5TB of CommonCrawl data. Similarly, Lai et al. (2020) demonstrate in their MultiTACOS system that joint training across multiple languages improves performance compared to language-specific models, particularly for low-resource languages. Together, these findings motivate our choice to adopt unified multilingual transformer models rather than separate per-language classifiers.

Finally, Naseem et al. (2025) introduced the POLAR benchmark, which underlies the SemEval-2026 Task 9 dataset. Their benchmark study comparing various multilingual models informed our selection of models for the ensemble, particularly their finding that RemBERT consistently achieved strong scores across multiple languages.

3. Methodology

3.1 Dataset

We conduct our experiments using the dataset provided by the organizers of SemEval-2026 Task 9, based on the POLAR benchmark. The dataset provides labeled social media posts for 22 languages, each annotated as either polarized (1) or non-polarized (0). The training set contains approximately 73,681 instances across all languages, with a held-out 20% validation split for internal evaluation.

Our analysis reveals significant class imbalance across languages. For example, Hindi has 85.5% polarized posts while Hausa has only 10.7%. Languages also vary substantially in text length, with Italian averaging 143 characters while Chinese averages only 34 characters. This heterogeneity motivated our use of focal loss and weighted sampling strategies.

3.2 Model Selection

Based on prior work, we implement a wide range of models from traditional baselines, neural sequence models and multilingual transformers. Transformer-based architectures are adopted following AlDayel and Magdy (2021), who show their superiority over traditional models for stance-related tasks. XLM-RoBERTa (Conneau et al., 2020) serves as a core multilingual encoder due to its strong cross-lingual transfer capabilities, while RemBERT is included based on its robust performance in the POLAR benchmark (Naseem et al., 2025). In addition, we follow Lai et al. (2020) and train unified multilingual models rather than language-specific classifiers. To address severe class imbalance, we apply focal loss as proposed by Lin et al. (2017).

3.3 Preprocessing

We apply minimal preprocessing to preserve meaningful linguistic cues. For the TF-IDF baseline, we remove URLs, anonymize user mentions, remove emojis, and convert text to lowercase. For transformer models, we preserve case and emojis (which carry sentiment information), only removing URLs and user mentions. Texts are tokenized with model-specific tokenizers and padded/truncated to a maximum sequence length of 128 tokens. We split each language's training data into 80% training and 20% validation using stratified sampling to maintain class balance. All splits use random seed 42 for reproducibility.

3.4 Baseline: TF-IDF with Logistic Regression

We establish a baseline using TF-IDF features with logistic regression, following Mohammad et al. (2016). We extract unigrams and bigrams with a maximum of 5,000 features, minimum document frequency of 2, and maximum document frequency of 95%. The logistic regression classifier uses balanced class weights to address class imbalance. This baseline achieves 68.8% macro F1-score across all languages.

3.5 Neural Sequence Models

To compare modern transformer fine-tuning with strong pre-transformer neural baselines, we implemented several sequence-based architectures. We first implemented a CLSTM model, which applies a one-dimensional CNN over token embeddings to extract local n-gram features, followed by an LSTM to model long-range dependencies. We then implemented a BiLSTM model and a BiLSTM with token-level attention, allowing the model to focus on the most informative tokens in a sequence. Finally, we evaluated a language-conditioned variant by learning a small language embedding and concatenating it with token embeddings, enabling the model to exploit language identity during multilingual training. All neural baselines were trained on a stratified 90/10 split of the training data into training and validation sets. Text was tokenized using a multilingual subword tokenizer (“bert-base-multilingual-cased”).

3.6 Multilingual Transformer Models

Our primary approach leverages pre-trained multilingual transformers. We experiment with five models:

1. XLM-RoBERTa-base: Pre-trained on 100 languages with 270M parameters
2. XLM-RoBERTa-large: Larger variant with 550M parameters
3. mDeBERTa-v3-base: Microsoft's multilingual DeBERTa with disentangled attention
4. InfoXLM-base: Microsoft's cross-lingual model with improved cross-lingual alignment
5. RemBERT: Google's multilingual model with 580M parameters, shown to perform well in POLAR benchmark

All encoder transformer models are fully fine-tuned with the following hyperparameters: batch size of 64, learning rate of 2e-5, warmup ratio of 0.06, weight decay of 0.01, and maximum 10 epochs with early stopping (patience=3) based on validation F1 macro. We use mixed-precision (FP16) training for efficiency. Training is conducted on the concatenated data from all 22 languages to enable cross-lingual transfer.

In addition to encoder-based models, we evaluated a decoder-only multilingual transformer, mGPT, to assess how generative architectures perform on classification tasks, motivated by prior coursework which showed decoders' competitive classification performance. We adopted parameter-efficient fine-tuning using QLoRA, enabling 4-bit quantization with low rank adaptation for this model.

3.7 Handling Class Imbalance

To address the severe class imbalance in several languages, we employ two strategies for encoder models during training. First we use focal loss (Lin et al., 2017) with $\alpha=0.25$ and $\gamma=2.0$. This loss function down-weights well-classified, easy examples and focuses training on hard and minority-class instances, which is particularly effective for imbalanced datasets. Second, we compute sample weights inversely proportional to class frequency within each language, ensuring that minority classes receive adequate representation during training.

3.8 Ensemble Method

Our final system uses a weighted soft-voting ensemble of all five transformer models. Weights are computed based on per-language validation performance, with better-performing models receiving slightly higher weights. The ensemble predictions are generated by averaging probability distributions across models, weighted by their performance:

$$P(y|x) = \Sigma_i \, w_i \cdot P_i(y|x)$$

where w_i is the normalized weight for model i based on its average F1 score across languages. The final weights are XLM-R-large (0.204), mDeBERTa (0.200), InfoXLM (0.198), RemBERT (0.199), XLM-R-base (0.199).

4. Results

4.1 Overall Performance

Table 1 summarizes the overall validation performance of our approaches. The ensemble method achieves the best performance with 79.69% macro F1, representing a 12.49 percentage point improvement over the TF-IDF baseline. For individual transformer models, XLM-RoBERTa-large gives the highest result and per-language scores are given in Figure D.1 in the Appendix. Among non-transformer models, BiLSTM with attention and language embedding achieved the best performance with 0.7650 macro F1 score. The confusion matrix (Figure A.1), per-language analysis (Figure A.2), training loss curve (Figure A.3) and precision-recall curve (Figure A.4) for this model is given in the Appendix.

Table 1: Overall Model Performance

Model	F1 Macro	Macro Precision	Macro Recall
Traditional baseline			
TF-IDF + Logistic Regression	68.80	69.59	69.74
Neural Sequence Models			
CLSTM	0.7471	0.7471	0.7470
BiLSTM + Attention	0.7578	0.7577	0.7587
BiLSTM + Attention + Language Emb.	0.7650	0.7648	0.7655
Decoder-based transformers			
mGPT	0.7527	0.7562	0.7518
Encoder-based transformers			
XLM-RoBERTa-base (Baseline Tra)	76.07		
XLM-RoBERTa-base + Focal Loss	76.50		
mDeBERTa-v3-base	77.91		
InfoXLM-base	77.43		
XLM-RoBERTa-large	78.14		
RemBERT	76.98		
Ensemble (5 models)	79.69		

Performance varies substantially across languages, reflecting differences in data size, class distribution, and linguistic characteristics. mDeBERTa-v3-base achieves the strongest individual performance on Amharic (74.88%) and Italian (61.77%), while XLM-RoBERTa-large performs best on Persian (83.22%) and Turkish (80.84%). RemBERT yields the highest score among individual models for Chinese (89.01%). In other words, different architectures excel on

different subsets of languages, a trend also reported in the POLAR benchmark study. This complementary behavior directly motivates our ensemble approach, which aggregates model predictions to reduce language-specific weaknesses. Detailed per-language ensemble results are reported in Table F.1 in the Appendix.

Our models have significantly better results than the models reported in the POLAR benchmark study. Out of seven languages reported in crosslingual setting in the benchmark paper, all of our transformer models perform better. There is only one case that they reported a higher result, which is mDeBERTa for Amharic (83.21% F1-Macro).

5. Discussion

5.1 Dataset Impact

The POLAR dataset's characteristics significantly influence system performance. The severe class imbalance ranging from 10.7% to 90.7% polarization makes training challenging, particularly for languages like Hausa where polarized examples are scarce. Our focal loss and weighted sampling strategies partially address this issue, but performance gaps remain.

The multilingual nature of the dataset enables cross-lingual transfer learning, which benefits low-resource languages like Nepali and Amharic. However, languages with unique scripts (Chinese, Arabic) or limited representation in pre-training data may not benefit equally from this transfer.

5.2 Approach Advantages and Disadvantages

Our ensemble approach uses the complementary strengths of different architectures. mDeBERTa's disentangled attention mechanism captures fine-grained semantic relationships, while RemBERT's larger capacity enables better representation of diverse languages. We combine their predictions using weighted averaging to reduce reliance on a single model and stabilize performance across languages.

The ensemble requires training and storing five separate models, increasing computational costs and inference time by approximately 5x. Memory constraints prevented us from completing full ensemble evaluation on GPU (CUDA out of memory during ensemble inference). Additionally, the equal weighting across languages may not be optimal for languages where certain models significantly outperform others.

5.3 Comparison with Polar Benchmark

Compared with the POLAR benchmark study (Naseem et al., 2025), our ensemble approach demonstrates strong performance. On the validation set, our 5-model ensemble achieves 79.69% macro F1, while our Codabench submissions show 80.3% F1 on 22 test

languages (Figure G.1). The benchmark reports RemBERT as achieving the highest performance in their monolingual evaluation setup. Our multilingual joint training approach with XLM-RoBERTa-base alone reaches 75.6% F1 on the test set (Figure G.2), demonstrating that cross-lingual transfer within a single unified model per architecture can achieve competitive results. The ensemble method further improves performance by combining complementary strengths of different transformer architectures.

5.4 Limitations

This project has several limitations. GPU memory constraints prevented running full ensemble inference on the GPU, requiring CPU-based evaluation, which increased inference time and limited further experimentation. Performance also varies across languages. Results for lower-resource languages such as Italian (64.25%) and German (70.17%) remain below those of high-resource languages, indicating that multilingual transfer does not fully address data imbalance. Interpretability is another limitation, as the ensemble provides limited insight into which linguistic features drive polarization predictions, restricting qualitative analysis. Finally, the models are trained on a fixed set of events and topics, which may limit generalization to unseen polarization contexts.

5.5 Potential Improvements

Given more time and resources, there are several ways we would improve this work further. Our first priority would be targeting the languages that struggled, specifically Italian and German, using more specialized fine-tuning strategies. We also see potential in explicitly teaching the model to recognize linguistic cues like emotional intensity or 'us-vs-them' rhetoric rather than relying solely on raw text embeddings. To address data gaps, we would use back-translation to generate synthetic examples for underrepresented classes. Finally, to make the system practical for real-world use, we would distill our heavy ensemble into a single, lighter model to cut down on inference costs.

6. Conclusion

We presented multiple approaches to multilingual polarization detection for SemEval-2026 Task 9. Our system progresses from a TF-IDF baseline (68.8% F1-Macro) through neural sequence models (76.50% F1-Macro) and multilingual transformers (76.07% F1 with XLM-RoBERTa-base) to a weighted ensemble of five models achieving 79.69% macro F1, corresponding to an improvement of 12.5 percentage points over the baseline. These results outperform the results shared by POLAR benchmark study. (Naseem et al., 2025)

Key findings include: (1) multilingual joint training enables effective cross-lingual transfer, benefiting low-resource languages; (2) focal loss and weighted sampling are essential for handling severe class imbalance across languages; (3) ensemble methods combining diverse architectures provide consistent improvements over individual models; and (4) significant performance variation across languages persists, with Italian and German remaining challenging.

Our work demonstrates that modern multilingual transformers can effectively detect polarization across diverse languages, though substantial room for improvement remains, particularly for languages with limited training data or unique linguistic characteristics.

7. Individual Contributions

- **Korhan Erdoğdu:** Implemented focal loss and weighted sampling strategies for handling class imbalance. Contributed to cross-lingual transfer experiments and addressed weak language performance (Amharic, Italian).

- **Zeynep Şahin:** Performed the exploratory data analysis and built different preprocessing pipelines.Developed the baseline TF-IDF model and initial transformerXLM-RoBERTa implementation. Trained Rembert model. Contributed to weighted loss implementation and report writing.

- **Suat Emre Karabıçak:** Implemented ensemble methods and soft voting mechanism. Developed checkpoint management system for training resilience. Conducted model comparison experiments. Trained XLM-RoBERTa-large and InfoXLM models. Developed the weighted ensemble optimization. Created submission generation pipeline for CodaBench.

- **Sıla Horozoğlu:** Implemented and evaluated parameter-efficient fine-tuning of ai-forever/mGPT using QLoRA (4-bit) + LoRA for polarization classification. Implemented neural baselines (CLSTM, BiLSTM, BiLSTM+attention), FastText-initialized word-level baseline, and a language-conditioned BiLSTM+attention model.

8. References

- AlDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management, 58*(4), Article 102597. https://doi.org/10.1016/j.ipm.2021.102597
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747
- Lai, M., Cignarella, A. T., Hernández Farías, D. I., Bosco, C., Patti, V., & Rosso, P. (2020). Multilingual stance detection in social media political debates. *Expert Systems with Applications, 143*, 113045. https://doi.org/10.1016/j.eswa.2019.113045
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324
- Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 31–41). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1003
- Naseem, U., et al. (2025). *POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization*. arXiv. https://arxiv.org/abs/2505.20624
- Zhou, C., Sun, C., Liu, Z., & Lau, F. C. M. (2015). A C-LSTM neural network for text classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 39–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1013
- Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 207–212). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1022

APPENDIX A

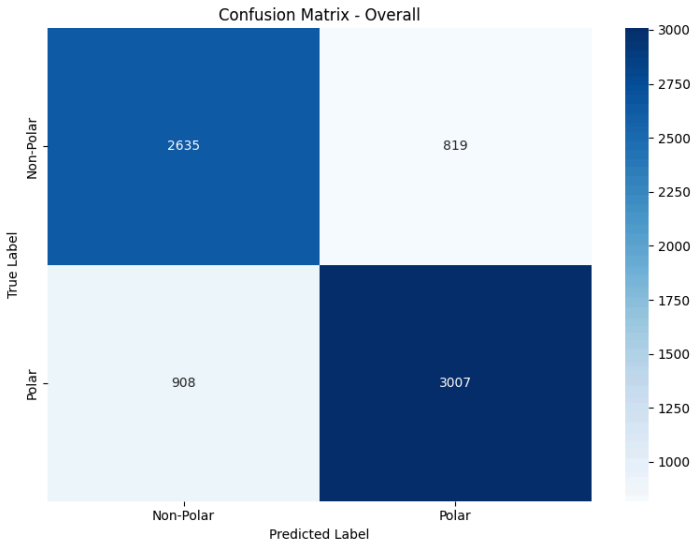

Figure A.1: Confusion matrix of the BiLSTM with token-level attention and language embeddings on the validation set. The model shows balanced performance across polar and non-polar classes, with comparable false positive and false negative rates.

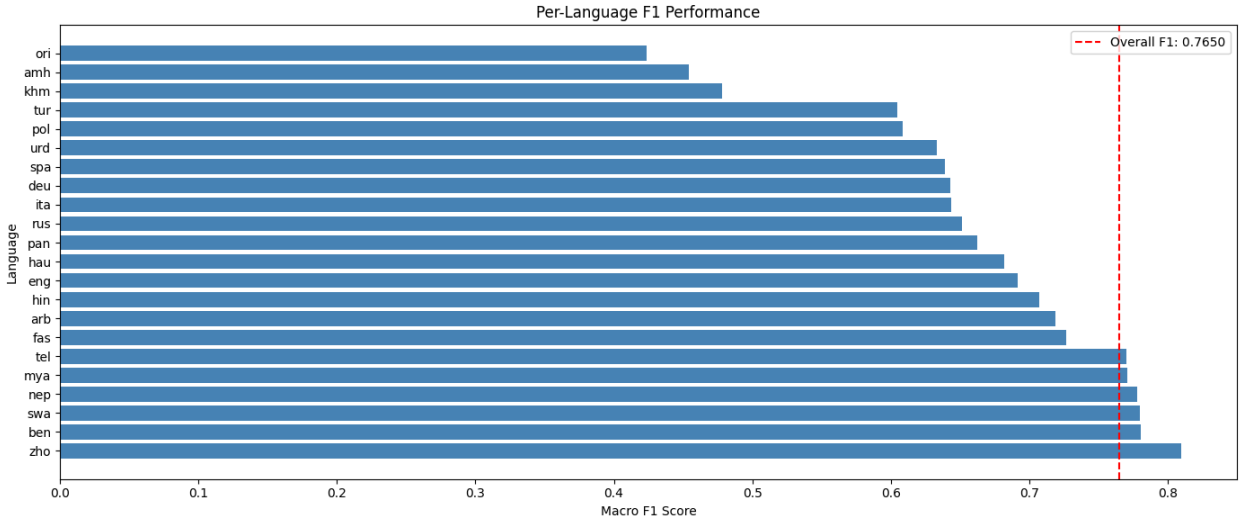

Figure A.2: Per-language macro F1 scores for the BiLSTM with attention and language embeddings. The dashed line indicates the overall macro F1 score (0.765), highlighting performance variation across languages.

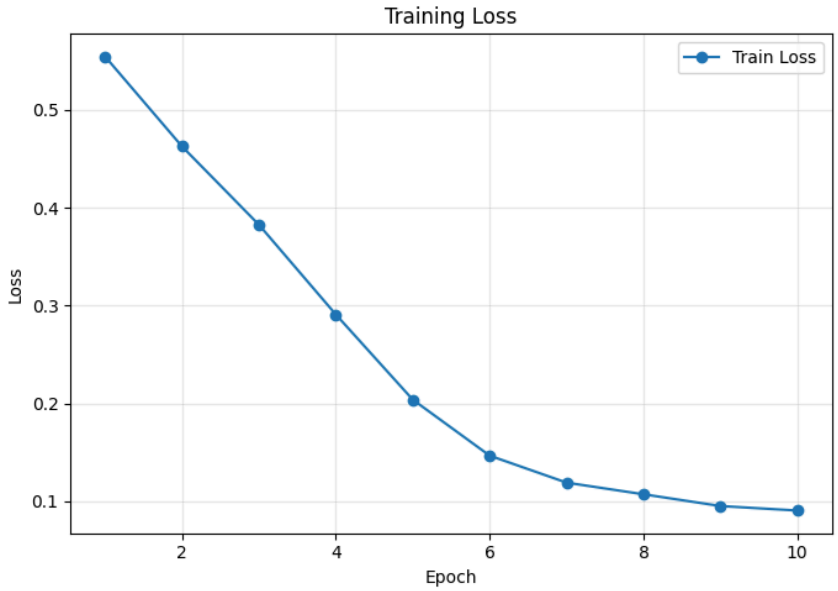

Figure A.3: Training loss across epochs for the BiLSTM with attention and language embeddings, showing stable convergence without signs of optimization instability.

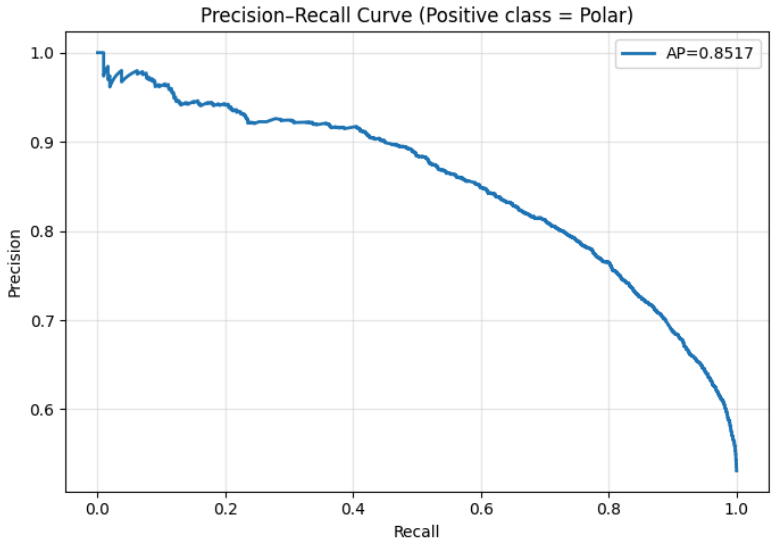

Figure A.4: Precision-recall curve for the BiLSTM with attention and language embeddings.

APPENDIX B

Exploratory Data Analysis and Dataset Statistics

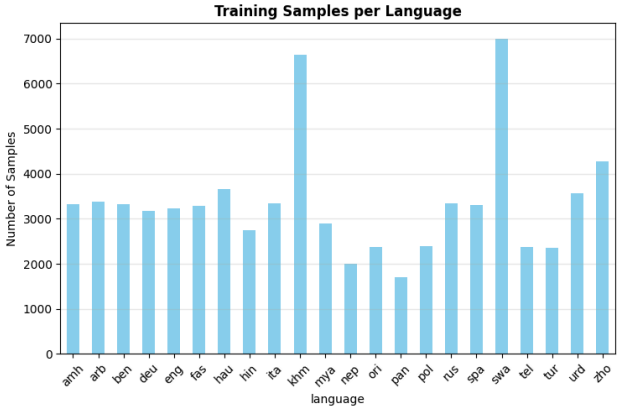

Figure B.1: Training samples per language in the SemEval-2026 Task 9 dataset. Swahili (7,340) and Khmer (6,972) have the most samples, while Punjabi (1,800) has the fewest.

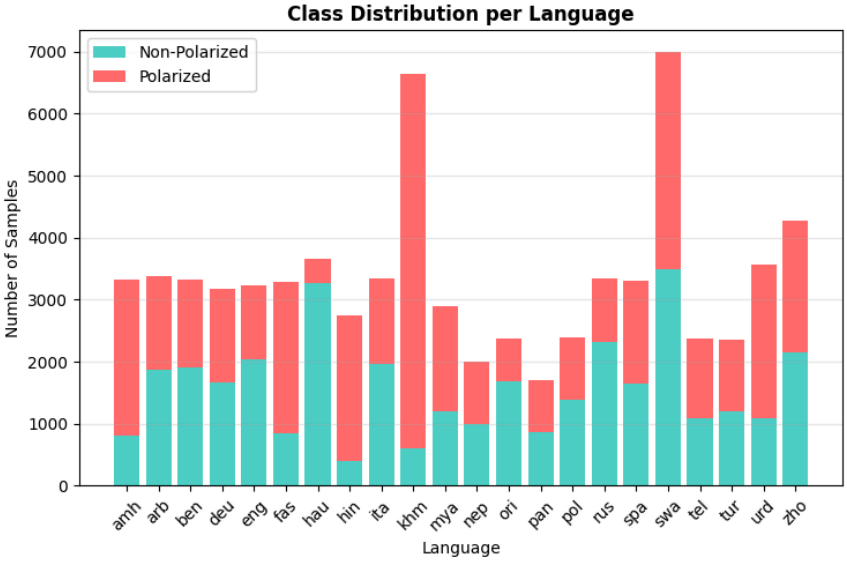

Figure B.2: Class distribution per language showing polarized (red) vs. non-polarized (teal) instances.

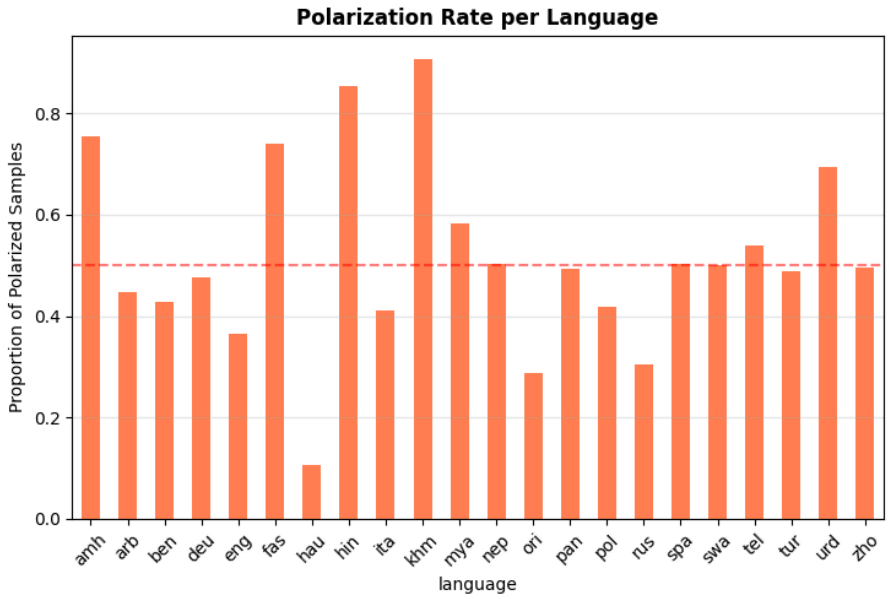

Figure B.3: Polarization rates across 22 languages. Eight languages show severe imbalance (>70% or <30% polarization), while 14 languages are relatively balanced (40-60%).

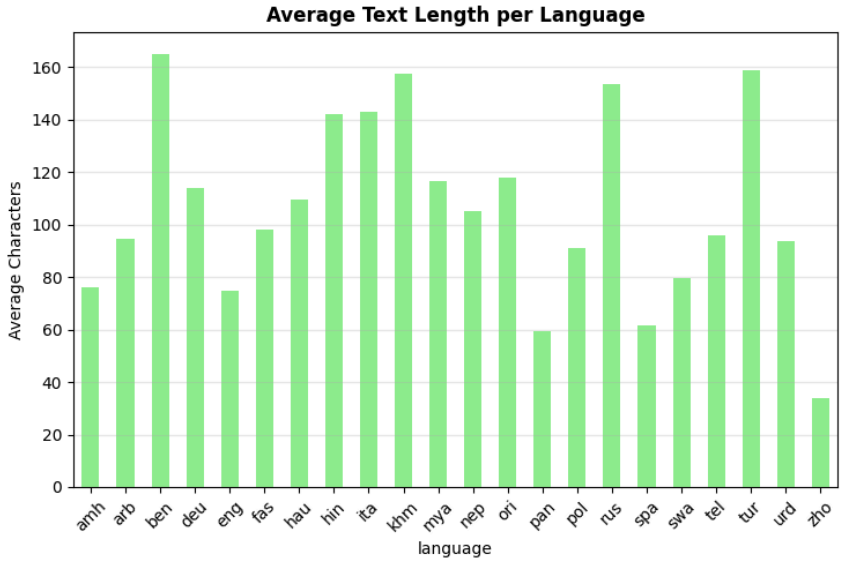

Figure B.4: Average text length per language in characters. Chinese averages 31 characters (logographic compression), while Bengali, Khmer, Turkish, and Russian exceed 140 characters.

Per Language Summary

Language	Train_Samples	Dev_Samples	Total_Samples	Train_Polarized	Train_NonPolarized	Polarization_Rate	Avg_Text_Length
swa	6991	349	7340	3504	3487	50.12%	80
khm	6640	332	6972	6029	611	90.80%	157
zho	4280	214	4494	2121	2159	49.56%	31
ben	3651	182	3833	392	3259	10.74%	110
urd	3563	177	3740	2476	1087	69.49%	94
arb	3380	169	3549	1512	1868	44.73%	95
rus	3348	167	3515	1023	2325	30.56%	154
ita	3334	166	3500	1368	1966	41.03%	143
ben	3333	166	3499	1424	1909	42.72%	165
amh	3332	166	3498	2518	814	75.57%	76
spa	3305	165	3470	1660	1645	50.23%	62
fas	3295	164	3459	2440	855	74.05%	98
eng	3222	160	3382	1175	2047	36.47%	75
deu	3180	159	3339	1512	1668	47.55%	114
mya	2889	144	3033	1682	1207	58.22%	117
hin	2744	137	2881	2346	398	85.50%	142
pol	2391	119	2510	1003	1388	41.95%	91
ori	2368	118	2486	683	1685	28.84%	118
tel	2366	118	2484	1274	1092	53.85%	96
tur	2364	115	2479	1155	1209	48.86%	159
nep	2005	100	2105	1008	997	50.27%	105
pan	1700	100	1800	840	860	49.41%	60

Figure B.5: Comprehensive dataset statistics for all 22 languages including sample counts, class distribution, polarization rates, and average text lengths.

APPENDIX C

TF-IDF Baseline Results

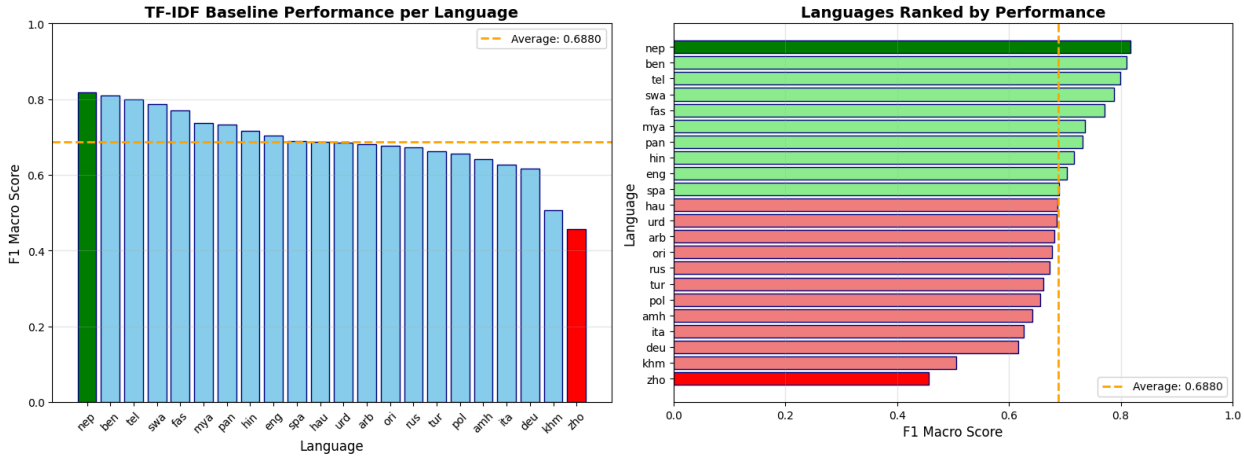

Figure C.1: TF-IDF baseline performance per language on the left. Nepali achieves the highest F1 score (0.88), while Amharic performs worst (0.48). The average macro F1 is 0.688. On the right languages ranked by TF-IDF baseline performance. Green bars indicate above-average performance, red bars indicate below-average performance.

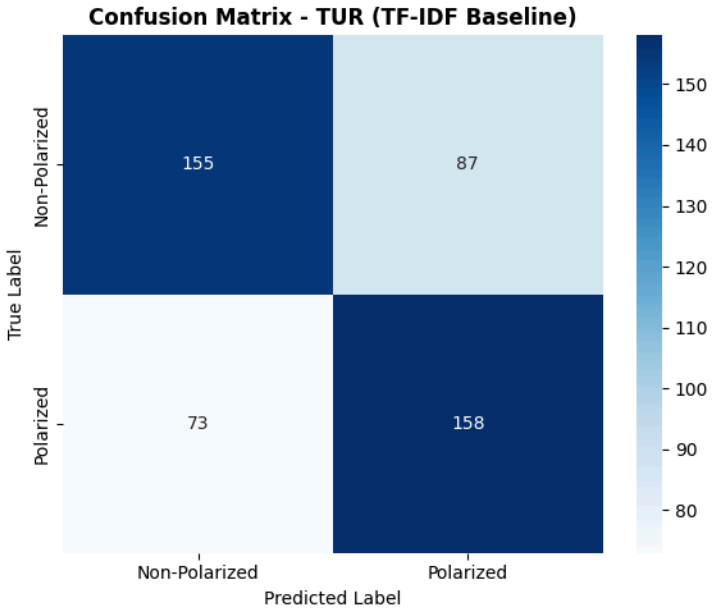

Figure C.2: Confusion matrix for Turkish (TUR) using TF-IDF baseline. The model correctly classifies 158 polarized and 155 non-polarized posts, with 87 false positives and 73 false negatives.

APPENDIX D

XLM-RoBERTa-Base Results

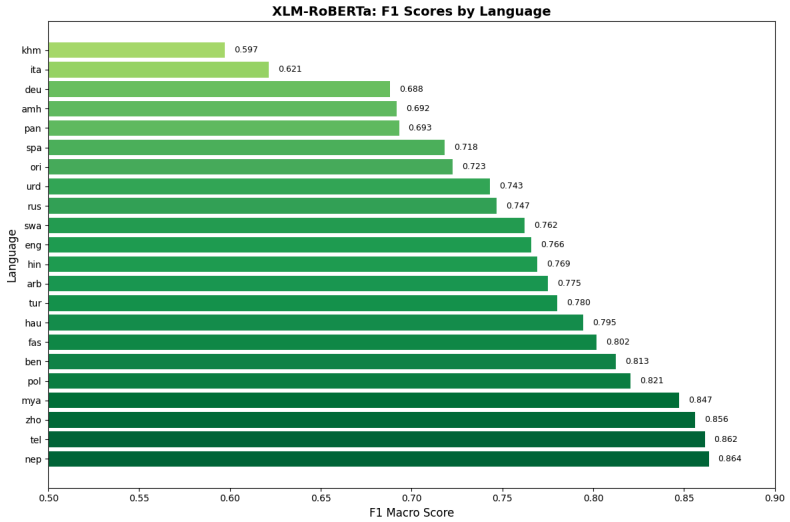

Figure D.1: XLM-RoBERTa-base F1 scores by language. Nepali (0.864), Telugu (0.862), and Chinese (0.856) achieve the best scores, while Khmer (0.597), Persian (0.621), and German (0.688) perform worst. Average F1: 0.761.

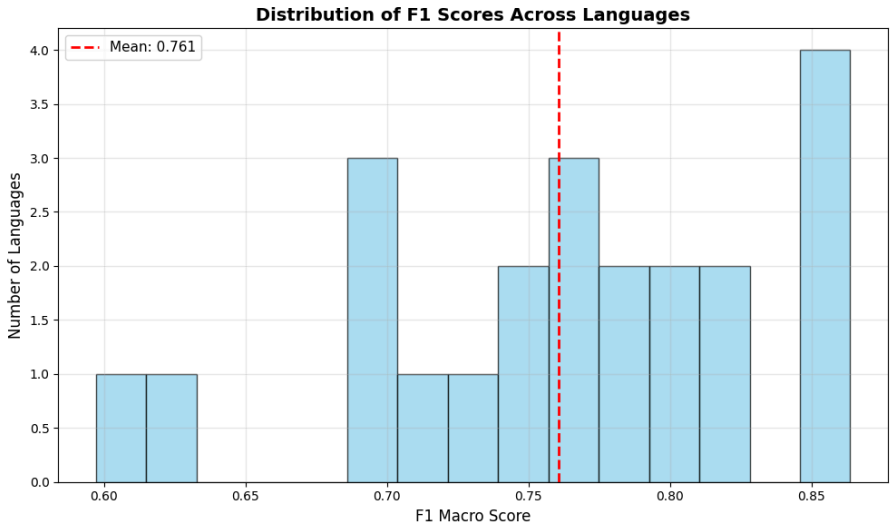

Figure D.2: Distribution of F1 scores across 22 languages for XLM-RoBERTa-base. Most languages cluster around 0.75-0.85, with a few outliers below 0.65.

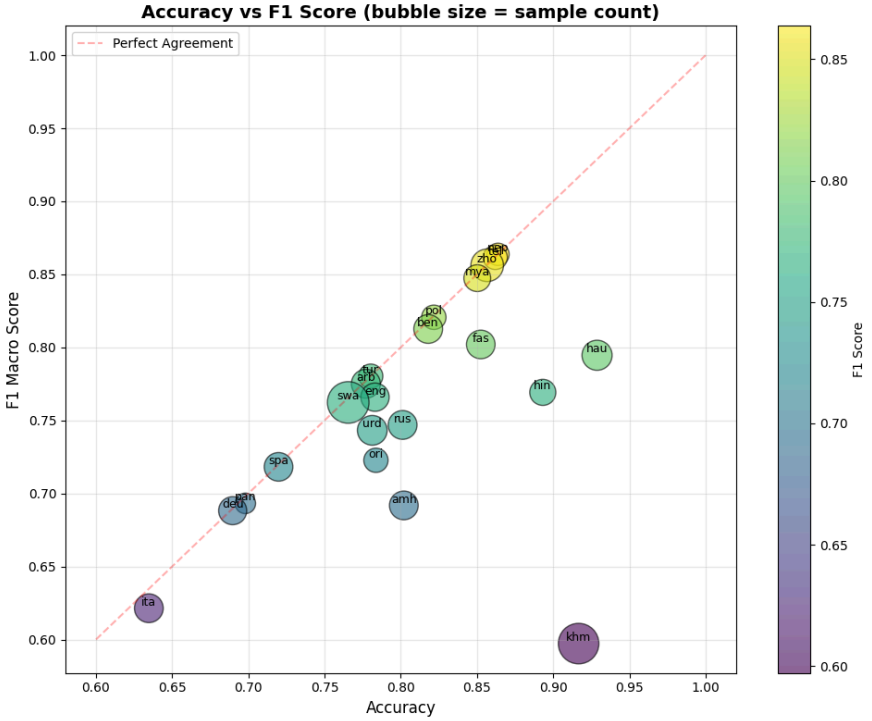

Figure D.3: Accuracy vs. F1 score scatter plot for XLM-RoBERTa-base. Bubble size represents sample count. Languages near the diagonal (perfect agreement line) have balanced performance. Khmer shows high accuracy (0.93) but lower F1 (0.60) due to class imbalance.

APPENDIX E

Individual Transformer Model Results

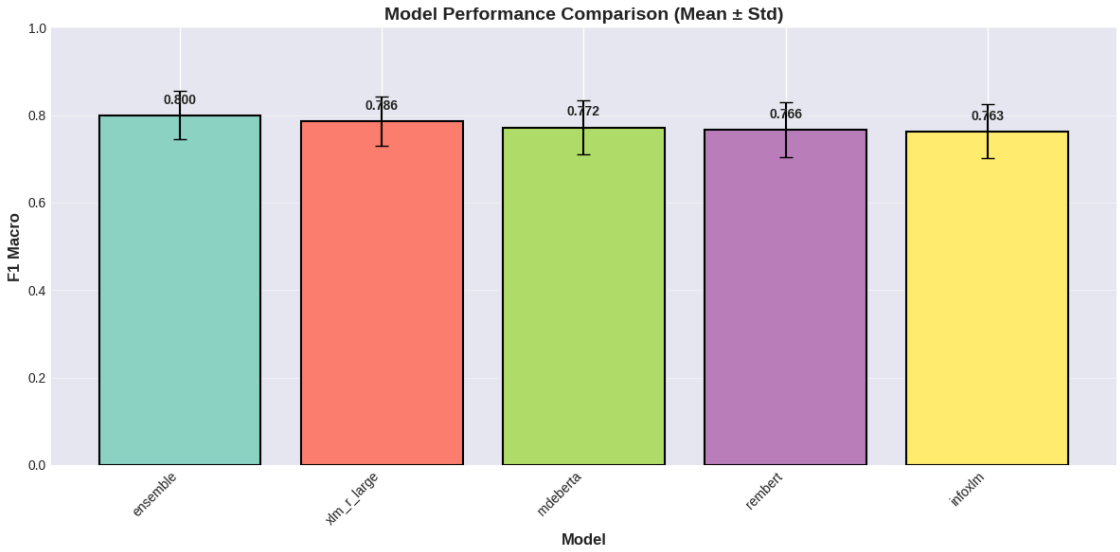

Figure E.1: Model performance comparison showing mean F1 scores with standard deviation. The ensemble (0.797) outperforms all individual models: XLM-R-large (0.781), mDeBERTa (0.779), InfoXLM (0.774), RemBERT (0.770), and XLM-R-base (0.769).

APPENDIX F

Ensemble Model Results

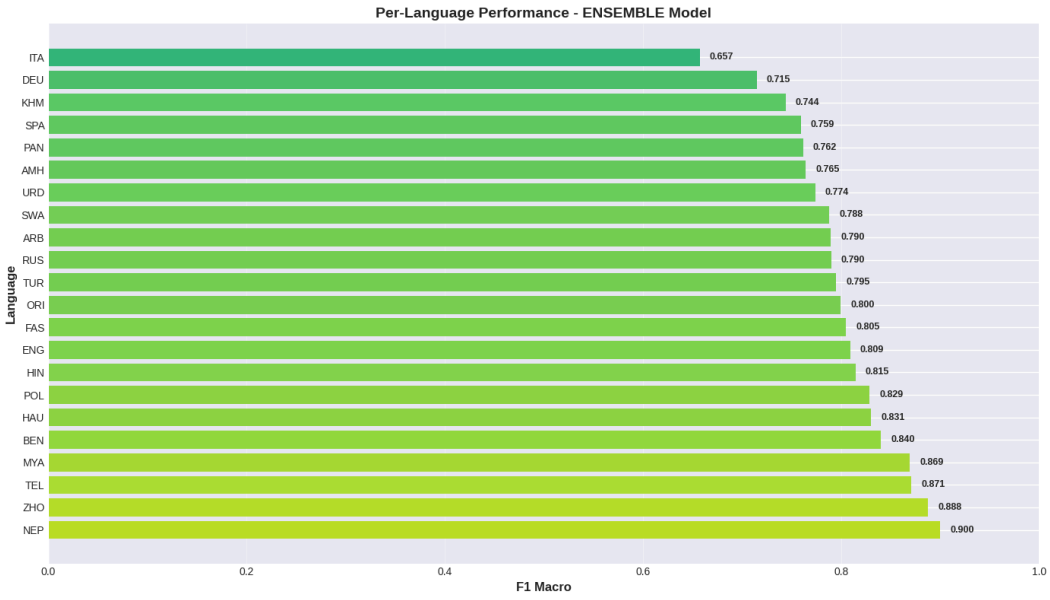

Figure F.1: Per-language macro F1 scores for the ensemble model. Nepali achieves the highest score (0.900), followed by Chinese (0.888) and Telugu (0.871). Italian (0.657), German (0.715), and Khmer (0.744) remain challenging despite ensemble methods.

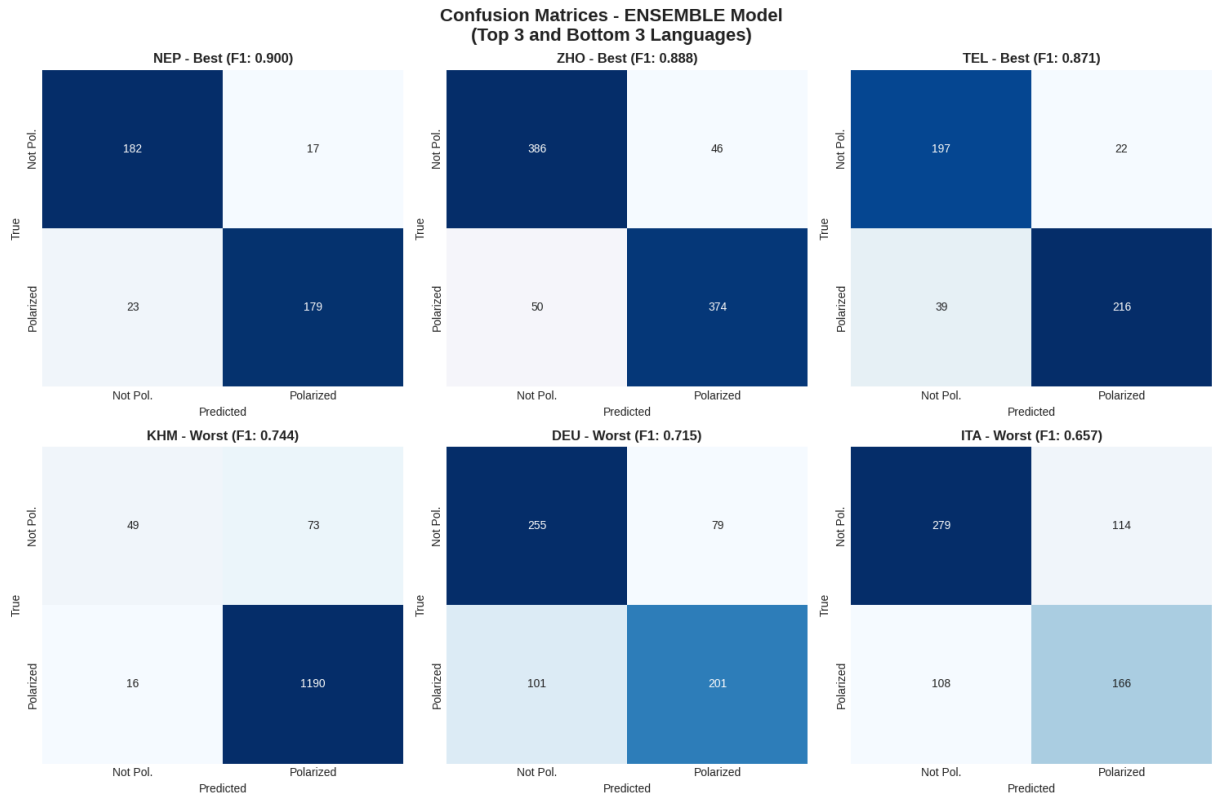

Figure F.2: Confusion matrices for the top 3 (Nepali, Chinese, Telugu) and bottom 3 (Khmer, German, Italian) performing languages using the ensemble model. Top-performing languages show strong diagonal dominance, while bottom-performing languages exhibit higher confusion rates.

Figure F.3: Performance comparison across three approaches: TF-IDF + Logistic Regression (0.672), Random Baseline (0.500), and Ensemble (0.800). The ensemble achieves a 19.0% relative improvement over the TF-IDF baseline.

Evaluation Scores per Language Table

Language	F1-Macro	Precision	Recall	Accuracy	Samples
NEP	0.9002	0.9005	0.9004	0.9002	401
ZHO	0.8878	0.8879	0.8878	0.8879	856
TEL	0.8711	0.8712	0.8733	0.8713	474
MYA	0.8695	0.8673	0.8731	0.8720	578
BEN	0.8403	0.8413	0.8394	0.8441	667
HAU	0.8305	0.8577	0.8083	0.9398	731
POL	0.8288	0.8338	0.8255	0.8351	479
HIN	0.8147	0.8019	0.8295	0.9035	549
ENG	0.8093	0.8122	0.8068	0.8248	645
FAS	0.8049	0.8043	0.8055	0.8498	659
ORI	0.7998	0.8142	0.7891	0.8418	474
TUR	0.7949	0.7949	0.7951	0.7949	473
RUS	0.7902	0.8002	0.7825	0.8269	670
ARB	0.7897	0.7891	0.7908	0.7914	676
SWA	0.7881	0.7903	0.7885	0.7884	1399
URD	0.7740	0.7736	0.7743	0.8079	713
AMH	0.7645	0.7645	0.7645	0.8261	667
PAN	0.7617	0.7625	0.7620	0.7618	340
SPA	0.7592	0.7607	0.7596	0.7595	661
KHM	0.7440	0.8480	0.6942	0.9330	1328
DEU	0.7149	0.7171	0.7145	0.7170	636
ITA	0.6573	0.6569	0.6579	0.6672	667

Table F.1: Detailed evaluation scores per language for the ensemble model, including F1 macro, precision, recall, accuracy, and sample counts.

APPENDIX G

Submission Results to Codabench

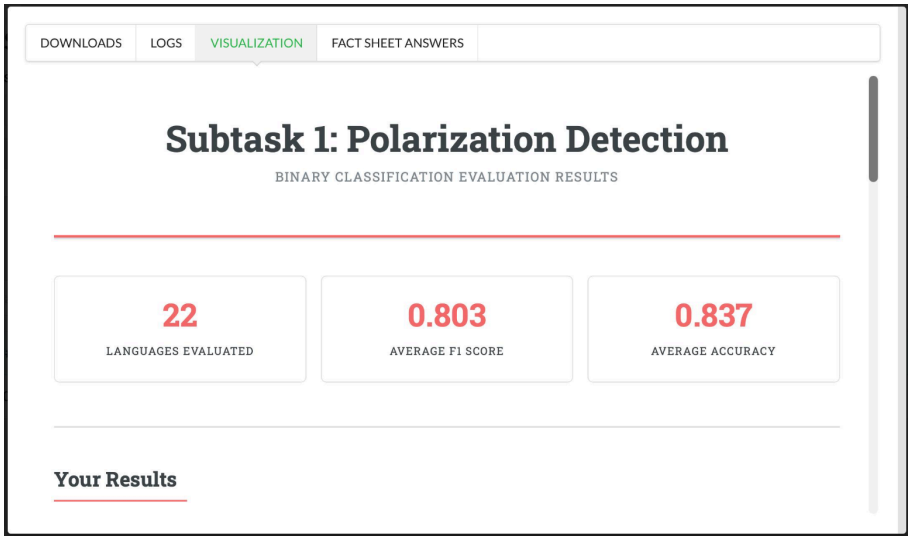

Figure G.1: Submission results of Transformers ensemble of 5 models with focal loss and weighted sampling.

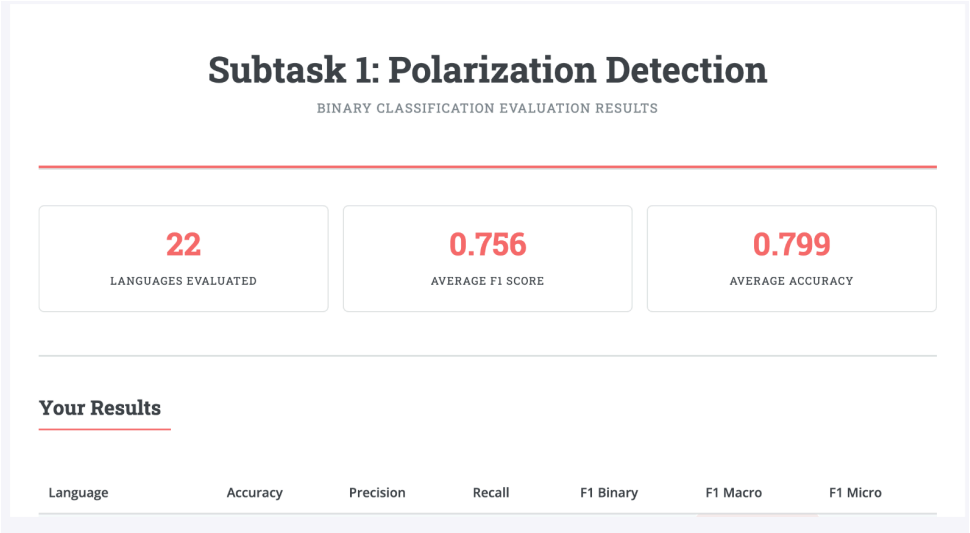

Figure G.2: Submission results of XLM-Roberta-Base

Figure G.3: Submission results of ensemble of 3 models (miniLM, mbert, distilmbert)