MSqrd at SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization

Syeda Samah Daniyal[1], Manal Hasan[1], Muneeba Badar[1], Shifa Shah[1], Sandesh Kumar[1], Dr. Abdul Samad[1]
[1]Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan
Correspondence: sd07838@st.habib.edu.pk, mh08438@st.habib.edu.pk, mb08091@st.habib.edu.pk, ss08336@st.habib.edu.pk

Abstract

Online polarization, the critical division between social, political, or identity groups, often leads to hate speech and social fragmentation. Detecting polarization, especially across diverse linguistic and cultural contexts, is a critical challenge. This paper presents our submission for SemEval-2026 Task 9, which focuses on detecting online polarization of multilingual, multicultural and multievents (Naseem et al., 2026b,a). The task is divided into three subtasks: (1) binary polarization detection, (2) multi-label classification of polarization type (e.g., political, racial, religious), and (3) multi-label identification of its manifestation (e.g., stereotype, vilification, dehumanization). For each subtask, we employ fine tune BERT-based transformer models. Model configurations are described in Section 4. The results are evaluated using F1 macro score. We have achieved scores of 78.6, 55.8, 44.6 on the development test set for subtasks 1, 2, and 3, respectively. Overall, the results demonstrate the effectiveness of BERT-based models for multilingual polarization detection.

1 Introduction

Online polarization is characterized by sharp division and hostility between social, political, or identity groups in digital spaces, often preceding offensive discourse, hate speech, and broader social fragmentation (POL, 2026). This project addresses SemEval-2026 Task 9: "Detecting Multilingual, Multicultural, and Multievent Online Polarization". The aim is to capture the polarization in online discourse across diverse contexts, and structured into three distinct subtasks (Naseem et al., 2026a):

- Subtask 1- Polarization Detection: This subtask involves binary classification to determine whether a given social media text is polar or not.
- Subtask 2- Polarization Type Classification: This subtask is a multi-label classification problem aiming to identify the types of polarization in the given text. This includes political, racial/ethnic, religious, gender/sexual, and other types of polarization.
- Subtask 3- Manifestation Identification: This subtask is also a multi-label classification problem to classify how polarization is expressed by detecting manifestation characteristics such as stereotyping, vilification, dehumanization, and etc.

2 Literature Review

In the recent years, multilabel and multilingal text classification has become a popular after transformer based model like BERT proved successful. This review will look at different methods of polarization detection and its types by focusing on the contextual meaning of the words used in a text instead of solely looking at the words. We explore several studies that solve similar problems using varied techniques and models.

The binary detection of polarization in a message is similar to earlier research on recognizing emotions and detecting harmful or offensive speech. Lepekhin and Sharoff (Lepekhin and Sharoff, 2025) explored emotion recognition across multiple languages including Amharic and Hausa which are low resource languages also present in POLAR's dataset. Their work on detection of emotions like anger and disgust aligns with POLAR's ST1 goal of detecting polarization in a text. The SemEval-2024 task on Persuasion Techniques in Memes (Dimitrov et al., 2024) also includes a binary subtask to identify if a meme utilizes any persuasion technique, similar to the binary detection of ST1.

The classification of specific social dimensions such as political, religious, or racial parallels research in identity-based hate speech and narrative classification. Toraman et al. (Toraman et al., 2022) worked on large-scale hate speech detection in En-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1664–1671

glish and Turkish (a low-resource language also in the POLAR dataset), grouping speech into domains like religion, gender, politics, and racism. These categories are exactly similar to POLAR's polarization types (religious, gender/sexual identity, political, racial/ethnic). Kiousis (Kiousis, 2025) and Faye et al. (Faye et al., 2025) worked on multilingual narrative classification across 93 categories related to geopolitical events like climate change and war. This is similar to ST2's goal of identifying what social or political dimension is being targeted across different global events. Prasad et al. (Prasad et al., 2022) studied abusive comment detection in Tamil which is a low resource language. The paper mentioned categories like Misandry, Misogyny, and Xenophobia to identify identity-based hate which is similar to ST2's classes targeting ethnicity and gender.

The SemEval-2024 task on Persuasion Techniques in Memes (Dimitrov et al., 2024) identifies 22 techniques, including Smears and Name Calling which are similar to the manifestation techniques in POLAR ST3. The EDOS task (Kirk et al., 2023) identifies 11 sexism categories including dehumanizing attacks and condescending explanations. These categories, alongside PCL categories (Perez-Almendros et al., 2022) like Unbalanced Power Relations, are related to manifestation techniques Invalidation, Dehumanization, and Vilification found in POLAR ST3.

The most common approach to solving these problems was the use of fine-tuned Transformer-based models, like XLM-RoBERTa, mDeBERTa, and RemBERT. These models work well with single language as well as multilingual problems. They can be further improved with techniques such as ensemble learning, adversarial training, and continued pre-training to handles issues such as limited data and class imbalance. Research for detecting specific issues, like hate speech, or certain types of harmful language, like sexism and condescension exists but the POLAR task is more complex because it needs to detect a wide range of polarization types (ST2) and detailed manifestations techniques (ST3) across multiple events at the same time. Our project uses advanced multilingual transformers and data-balancing techniques to tackle this broader polarization challenge.

Compared to prior work, which typically focuses on single-task or monolingual settings, the POLAR task introduces a significantly more complex setup requiring simultaneous handling of multilinguality, multi-label classification, and cross-event generalization. This increases both modeling complexity and the need for robust imbalance handling strategies.

Table 1: Dataset Statistics for ST1, ST2, and ST3

Feature	ST1	ST2	ST3
Type of Classification	Binary	Multi-label	Multi-label
Samples	77,368	77,368	64,810
Columns	1	5	6
Languages	22	22	18
Imbalance	1:1.13	1:3.23	1:2.92
Label Balance	Non-polarized (46.89%), Polarized (53.11%)	Political (27.40%), Gender/Sexual (8.48%)	Stereotype (33.60%), Dehumanization (11.51%)

3 Dataset Description and Understanding

The dataset provided by the organizers of SemEval-2026 Task 9 consists of short, multilingual social media texts spanning three distinct sub-tasks (ST). While ST1 focuses on binary polarization detection, ST2 and ST3 involve complex multi-label classifications across diverse categories such as political, gender, and dehumanization labels. With 22 languages covered, including low-resource languages like Hausa and Amharic, and a "multi-event" design, the data is remarkably diverse. To address the significant class imbalances identified in Table 1, specifically in ST2 and ST3, the study utilizes specialized loss functions and deep semantic modeling to ensure robust performance across minority classes and cross-lingual transfers (Naseem et al., 2026b; Polar-SemEval, 2026).

4 Models and Experiments

4.1 Subtask1

For multilingual binary polarization detection, we experimented with multiple transformer architectures, data augmentation techniques and imbalance-handling strategies before arriving at the final configuration.

Baseline Architectures We first evaluated roBERTa, LaBSE, XLM-R, mBERT, and mDeBERTa under standard cross-entropy loss. This configuration achieved a Macro-F1 score of approximately 73, but there was strong overfitting behavior and substantial variance across languages.

In particular, German, Urdu, and Spanish exhibited noticeably lower scores.

We further experimented with mDeBERTa + data augmentation to overcome overfitting. Several augmentation techniques were tried such as, back translation, synonym replacement, random insertion, and contextual word embeddings. However, the results showed no improvement or reduced validation F1 during training. We believe that the augmentated texts were not able to retain the contextual meaning of the sentences.

XLM-R Large without Focal Loss We then evaluated XLM-R-Large-Polarization-Classifier under standard cross-entropy loss. This improved the average Macro-F1 to 74 and produced more consistent multilingual accuracy compared to mDeBERTa. However, validation loss exhibited an increasing trend despite stable F1, suggesting residual overfitting to dominant language patterns.

Imbalance-Aware Optimization via Focal Loss. To address class imbalance and hard-instance sensitivity, we replaced cross-entropy with Focal Loss. This modification increased the average Macro-F1 to 76 and significantly reduced overfitting behavior observed in prior configurations and stabilized training dynamics.

Data Augmentation Revisited. We additionally evaluated XLM-R Large + Focal Loss + augmentation. Contrary to expectation, augmentation yielded negligible improvement. Analysis suggests that binary polarization detection is highly sensitive to contextual nuance, where the augmentation may have distorted signals critical for classification, thereby weakening decision boundaries.

Targeted Language-Specific Experiments. Across all prior experiments, Spanish and Italian consistently remained among the lower-performing languages. Therefore, we conducted language specific fine-tuning using monolingual BERT variants, including SaBERT (spanish BERT), alBERTo (Italian BERT), bert-base-italian-cased (Italian BERT). The hypothesis was that monolingual pretraining might better capture nuanced polarization cues. However, these targeted experiments did not yield significant improvements in Macro-F1.

LLM-Based Experimentation. We also experimented with meta-llama/Llama-3.1-8B-Instruct in a classification setting. While the overall Macro-F1 reached approximately 75, the behavior was uneven across languages, performing strongly for some while underperforming for others. Although promising, it lacked the consistency and stability achieved earlier.

Hyperparameter Sensitivity Analysis. Across all architectural configurations, we tried different hyperparameters to ensure that performance differences were not attributable to suboptimal optimization settings. These experiments included variations in maximum sequence length, learning rate, batch size and gradient accumulation steps, weight decay and regularization strength, warmup ratio and scheduler configurations.

We observed that learning rates above 1×10^{-5} led to unstable validation curves and degraded Macro-F1, particularly for low-resource languages. Increasing sequence length beyond 256 tokens produced negligible gains while substantially increasing computational cost. Other parameters were considered according to the GPU limitations.

Model Selection Rationale. Overall, experimentation indicates that architectural scaling or augmentation alone did not guarantee improved multilingual robustness. Instead, handling class imbalance at the loss level proved decisive. The final configuration XLM-R-Large-Polarizationm Classifier with Focal Loss offered the best balance between cross-lingual stability, overfitting control, and Macro-F1 performance, and was therefore selected for submission.

Final Training Configuration. The final model was trained with a learning rate of 1×10^{-5}, batch size of 16 (effective via gradient accumulation), and 3 epochs. Maximum sequence length was set to 256. Focal Loss was used with $\gamma = 2$. Model selection was based on validation Macro-F1.

4.2 Subtask2

This task involved multi-label type classification for the five categories (political, racial/ethnic, religious, gender/sexual, and other) in polarized text. Our study involved rigorous testing of multilingual BERT-based models including LaBSE, mDeBERTa, XLM-RoBERTa, and the XLM-T sentiment model cardiffnlp/twitter-xlm-roberta-base-sentiment (Barbieri et al., 2022; CardiffNLP, 2022).

Baseline Models. During initial experiments, we used LaBSE with standard Binary Cross-Entropy

(BCE) loss achieved a Macro-F1 score of approximately $39.0 - 45.0$. However, we found low scores for low-resource languages and for languages with higher class imbalances, such as English, indicating that while LaBSE provided strong multilingual alignment, it struggled to generalize across skewed label distributions. Subsequent experiments with mDeBERTa and XLM-RoBERTa underperformed relative to LaBSE performance. cardiffnlp/twitter-xlm-roberta-base-sentiment achieved the best results, likely benefiting from domain-specific pretraining on Twitter data.

Data Augmentation. We explored multiple data augmentation strategies to address the issue of class imbalances and less data for low-resource language, including translation (Spanish $\rightarrow$ English), back-translation (English $\rightarrow$ Spanish $\rightarrow$ English), and synonym replacement. However, augmentation techniques decreased the accuracy of our models, likely due to the change in semantic meaning. We eventually discarded these strategies and focused more on loss-level adjustments.

Class Imbalance Handling. To address severe label skew across categories, we experimented Binary Cross Entropy (BCE), Weighted BCE, Focal Loss, and Focal Tversky Loss. The final adopted solution was Asymmetric Loss (ASL), which uses separate parameters for positive and negative samples. Predicted probabilities are clamped to the interval $[\delta, 1-\delta]$ where $\delta = 0.005$ before computing the weighted BCE. We set $\gamma_{\text{pos}} = 1.5$, $\gamma_{\text{neg}} = 1.5$, which formulated best results by balancing hard positives and easy negatives.

Training Strategy and Regularization. We employed the following techniques to prevent overfitting and improve generalization. **Early stopping** with a patience of 2 epochs and an improvement threshold of 0.002 was used alongside weight decay of 0.01 for L_2 regularization, with best-model selection based on validation Macro-F1. To help the model converge smoothly, we employed a Cosine Learning Rate Scheduler with a warmup ratio of 0.15. Training was further stabilized using gradient accumulation over 2 steps with gradient clipping at $\|\nabla\|_{\max} = 1.0$, preventing exploding gradients under large effective batch sizes. Lastly, per-label threshold tuning was performed on the validation set after training, sweeping thresholds $t \in [0.05, 0.90]$ for each label independently to maximize per-label F1 before final prediction. Threshold tuning was performed per-label on the validation set, and applied globally across languages for final inference.

4.3 Subtask3

For multi-label, cross-lingual manifestation detection in polarized text, we experimented with multilingual transformer architectures including mDeBERTa, LaBSE, and XLM-R-Large-Polarization-Classifier (Ashraf et al., 2024). The task predicts six non-mutually exclusive labels (stereotype, vilification, dehumanization, extreme language, lack of empathy, invalidation) under a substantial cross-lingual distribution shift.

Baselines. Using mDeBERTa and LaBSE with Binary Cross-Entropy (BCE) yielded Macro-F1 scores of 0.33–0.35. Errors showed a clear precision–recall imbalance: the models favored majority manifestations (stereotype, vilification) and under-detected minority labels such as dehumanization and lack of empathy.

Imbalance handling. We added positive class weighting and multi-label Focal Loss with per-label α_t and p_t. Single-model performance improved but stabilized around Macro-F1 ≈ 37.0 on validation, suggesting remaining calibration instability.

Language-aware optimization and training. Given 18 languages and uneven per-language label distributions, we used a language-balanced + label-aware WeightedRandomSampler, global + per-language pos_weight matrices, and language–label-aware loss scaling, which improved both recall and precision. To reduce overfitting and stabilize training, we applied LoRA (via PEFT), gradient checkpointing, early stopping with best-model selection by validation Macro-F1, per-label threshold tuning on validation, and language-specific threshold tuning at inference.

Few-shot LLM. We also tried few-shot prompting with Llama 3.2 3B using label definitions and 4–5 language-aware examples, with low temperatures (0.1 for classification, 0.4 for generation; $top_p = 0.95$) to enforce deterministic JSON. This approach performed worse (Macro-F1 ≈ 30.0), indicating fine-tuned discriminative models are better suited for dense multi-label classification.

Ensemble. The submitted system ensembled XLM-R-Large-Polarization-Classifier and mDeBERTa to combine XLM-R's calibration strength

with mDeBERTa's stable multilingual discrimination. We averaged per-label probabilities and applied thresholds tuned on validation. This reduced extreme fluctuations, especially for minority labels and low-resource languages, and produced more balanced precision–recall trade-offs across languages.

5 Results and Discussions

The following sections present our results for the POLAR task, discussing the performance achieved for each sub-task and outlining the advances achieved over previous research. We also acknowledge the limitations encountered and propose reasonable future directions.

5.1 Subtask1

The Average F1 Macro Score obtained for this subtask across all 22 languages is 78.65, with an **Average Accuracy** of 82.1 (as detailed in Table 2).

Compared to the benchmark scores shown in (Figure A1) for the 7 languages, i.e. Amharic, Arabic, English, German, Hausa, Spanish, and Urdu, the majority of these languages showed significant improvement using fine-tuned XLM-Roberta-Large. The best performance was observed in Nepali (89.59), Chinese (88.06), and Telugu (87.39), indicating effective cross-lingual learning achieved by the trained model for both high and low-resource languages.

Across languages, the results of this subtask show stable convergence of the binary decision. The use of Focal Loss reflected in good Macro-F1 scores across all diverse languages.

While overall stability was maintained, a few low scoring languages. Italian (63.49) and German (71.7) show a noticeable difference in comparison to the other languages, which may reflect higher lexical ambiguity in polarized expressions

5.2 Subtask2

The model was evaluated across 22 languages for subtask 2, obtaining an overall F1 Macro Score of 55.8 (as detailed in Table 2).

The model demonstrates significant increases from the crosslingual baselines that were mentioned in Figure A1. The best performing languages for our model include Urdu (76.68), Hindi (77.11), Nepali (75.07) and Chinese (73.35). For the low resources languages such as Hausa (30.68), Odia (53.01), Swahili (47.33), Telugu (40.64), Khmer (65.55) we noticed that our model did not perform very well, likely due to limited training data.

The results indicate that the model struggles with low-resource languages. While we also notice that even with some high-resource languages, such as English and Italian we are unable to reach higher scores due to lower number of positive samples in the dataset.

5.3 Subtask3

The final submitted ensemble was evaluated on eighteen languages and improved over the baseline languages (Table 2, Figure A1), achieving an overall Macro-F1 of 0.446.

Performance varied significantly by language. Urdu (77.68) and Hindi (71.81) performed best, while Arabic (54.85) and Chinese (52.48) remained relatively stable. In contrast, very low-resource languages such as Hausa (13.27) and Odia (23.28) were still difficult, likely due to sparse positive examples for minority manifestations.

Relative to single-model runs (Macro-F1 $\approx$ 37.0), the ensemble improved minority-label recall and reduced cross-language volatility, with the clearest gains in mid-resource languages. Subtask 3 remains the most challenging due to dense multi-label structure, cross-lingual distribution shift, and severe per-language imbalance.

5.4 Error Analysis

We observe that most errors occur in minority labels such as dehumanization and lack of empathy, where models tend to confuse subtle contextual cues with general negative sentiment. Additionally, low-resource languages exhibit higher false negatives, suggesting insufficient representation during training. These findings highlight the importance of better calibration and more balanced multilingual data.

6 Conclusion

In conclusion, this paper presents our contributions for the cross-lingual and multilabel polarization detection problem of SemEval-2026 Task 9: "Detecting Multilingual, Multicultural and Multievent Online Polarization" (Naseem et al., 2026b,a). We see significant improvements in our model from the baselines that were given to us for all three subtasks and in every language.

Our findings suggest that BERT based models have outperformed in all three subtasks. In addition

Table 2: Consolidated F1 Macro Scores x 100 by Language across All Subtasks (ST1, ST2, ST3)

Language	ST1	ST2	ST3
Amharic (amh)	74.95	59.54	49.68
Arabic (arb)	80.62	61.08	54.85
Bengali (Ben)	82.77	36.08	23.27
German (deu)	71.7	54.97	45.52
English (eng)	77.43	47.34	46.48
Persian (fas)	81.07	60.88	38.85
Hausa (hau)	73.41	30.68	13.27
Hindi (hin)	79.08	77.11	71.81
Italian (ita)	63.49	27.33	-
Khmer (khm)	73.08	65.55	37.33
Bermese (mya)	85.68	68.14	-
Nepali (nep)	89.59	75.07	58.21
Odia (ori)	78.11	53.01	23.28
Punjabi (pan)	77.32	46.27	42.72
Polish (pol)	79.68	51.57	-
Russian (rus)	77.92	49.88	-
Spanish (spa)	74.08	64.77	45.20
Swahili (swa)	77.27	47.33	47.42
Telugu (tel)	87.34	40.64	29.06
Turkish (tur)	77.34	59.68	45.99
Urdu (urd)	78.74	76.68	77.68
Chinese (zho)	88.06	73.35	52.48
Average F1 Macro Score	**78.6**	**55.8**	**44.6**

to that our approaches to fine-tuning the model excelled in detecting polarization in the text while also being able to detect the type and manifestation of polarization.

In future works we will look into applying advanced fine-tuning techniques to the model to further increase the accuracy. We will be utilizing gen-AI models like llama and BERT based architectures followed by advanced fine-tuning techniques such as few-shot prompting, ensemble learning, s and improve accuracy for low resource languages, which requires further exploration.

References

2026. POLAR@SemEval-2026: Detecting multilingual, multicultural and multievent online polarization. Official Website.

Shaina Ashraf, Isabel Bezzaoui, Ionut Andone, Alexander Markowetz, Jonas Fegert, and Lucie Flek. 2024. DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

CardiffNLP. 2022. cardiffnlp/twitter-xlm-roberta-base-sentiment. `https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment`. Hugging Face model repository.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Geraud Faye, Guillaume Gadek, Wassila Ouerdane, Celine Hudelot, and Sylvain Gatepaille. 2025. NotMyNarrative at SemEval-2025 task 10: Do narrative features share across languages in multilingual encoder models? In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 58–66, Vienna, Austria. Association for Computational Linguistics.

Panagiotis Kiousis. 2025. IRNLP at SemEval-2025 task 10: Multilingual narrative characterization and classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 54–57, Vienna, Austria. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. *Preprint*, arXiv:2303.04222.

Mikhail Lepekhin and Serge Sharoff. 2025. Domain_adaptation at SemEval-2025 task 11: Adversarial domain adaptation for text-based emotion recognition. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 49–53, Vienna, Austria. 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.

Carla Perez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States. Association for Computational Linguistics.

Polar-SemEval. 2026. Polar @ semeval-2026 task 9: Official task dataset. `https://github.com/Polar-SemEval/data-public/`. Data provided by task organizers.

Gaurang Prasad, Janvi Prasad, and Gunavathi C. 2022. GJG@TamilNLP-ACL2022: Using transformers for abusive comment classification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 93–99, Dublin, Ireland. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

A Appendix

		Monolingual						Cosslingual					
Task	**Lang.**	**InfoXLM**	**LaBSE**	**RemBERT**	**XLM-R**	**mBERT**	**mDeBERTa**	**InfoXLM**	**LaBSE**	**RemBERT**	**XLM-R**	**mBERT**	**mDeBERTa**
1. Polarization	Amharic	77.96	78.92	**81.93**	80.64	53.71	74.98	23.26	49.44	5.63	0.00	0.00	**83.21**
	Arabic	58.39	61.94	67.19	**70.42**	52.48	54.68	29.68	35.87	27.76	2.78	**37.79**	25.63
	English	72.09	69.84	75.43	**76.08**	72.77	74.94	1.90	51.68	53.41	46.89	33.44	**53.45**
	German	29.75	63.46	67.50	**67.78**	58.45	64.38	0.62	51.97	**64.59**	41.87	46.05	25.94
	Hausa	59.57	66.41	**67.43**	66.41	59.32	65.62	8.21	22.92	6.47	1.50	6.25	**25.13**
	Spanish	38.69	64.04	**70.98**	57.00	59.31	54.16	45.70	**68.66**	67.26	21.09	68.34	62.17
	Urdu	1.07	66.35	**79.95**	43.91	65.63	60.68	1.60	30.36	0.00	0.00	3.68	**68.67**
2. Types	Amharic	24.22	38.13	**43.65**	25.71	17.93	25.00	11.56	20.65	2.49	0.00	0.00	**26.76**
	Arabic	22.97	40.23	**42.12**	37.58	27.53	35.22	11.49	**23.73**	8.78	2.11	12.47	7.93
	English	16.04	23.25	**31.38**	24.10	21.07	17.70	10.84	**19.69**	16.42	12.22	11.70	3.93
	German	13.52	58.58	**61.19**	58.56	54.13	40.64	12.03	**35.86**	29.59	9.72	23.98	11.88
	Hausa	18.56	19.14	17.38	18.09	**19.61**	18.87	3.20	**9.97**	3.91	1.39	5.78	5.90
	Spanish	43.26	66.07	**67.76**	58.07	57.94	43.40	7.89	**47.06**	15.38	0.43	32.02	14.60
	Urdu	27.14	**51.94**	51.60	45.38	38.02	33.13	3.77	13.62	6.33	0.00	3.94	**20.30**
3. Manifestations	Amharic	43.52	47.56	**47.63**	43.17	33.18	43.29	15.57	27.07	8.64	0.00	0.00	**43.58**
	Arabic	40.05	51.55	52.52	**55.61**	42.18	47.73	16.56	**30.68**	22.14	0.00	19.57	17.24
	English	14.40	15.01	**19.39**	18.61	18.60	15.15	7.16	**10.16**	10.05	5.62	10.69	8.85
	German	38.49	49.88	**52.74**	51.70	46.85	51.91	2.38	**36.12**	27.05	0.00	23.38	12.93
	Hausa	19.23	**20.04**	19.18	18.89	18.74	18.93	5.74	5.86	3.77	3.12	6.19	**6.43**
	Spanish	38.94	50.00	**51.04**	45.02	45.17	35.09	2.34	**40.63**	11.83	0.40	35.99	23.56
	Urdu	34.26	52.20	**53.64**	41.32	45.90	47.01	2.10	19.16	11.54	0.00	1.88	**48.58**

Figure A1: Baseline Results for Monolingual and Cross-lingual Polarization Detection across Subtasks from (Naseem et al., 2026a). Highest F1 scores x 100 in each block are highlighted in color (blue for Monolingual, orange for Cross-lingual).