CoPol at SemEval-2026 Task 9: Modeling Polarization Type Co-occurrence with Label Correlation Networks

Pushkar Arora
Delhi Skill and Entrepreneurship University (DSEU)
Okhla, New Delhi, India
Pushkararora88@gmail.com

Abstract

We present **POLAR-LDA** for SemEval-2026 Task 9 Subtask 2: multi-label classification of five polarization types across 22 languages. Two underexplored challenges define this task: absolute positive scarcity certain language-label pairs contain zero or one positive training instance, below the recoverable threshold for any supervised classifier and culturally variable label co-occurrence patterns that independent classifiers cannot capture. POLAR-LDA extends mDeBERTa-v3-base with a Label Correlation Network applying Graph Attention over language-specific co-occurrence matrices, Asymmetric Loss for positive-scarcity regimes, and language-family grouped ensemble training. The system achieves 0.567 macro F1 on the official evaluation (range: 0.784 Hindi to 0.256 Italian; per-label: 0.685 political to 0.444 other) ; per-label data volume dominates linguistic family as a predictor of performance ($\sigma_{\text{intra-family}}$=0.119 **vs.** $\sigma_{\text{inter-family}}$=0.086). Diagnostic analysis identifies data voids as the hard floor of multilingual polarization classification and shows that per-label training volume dominates linguistic family as a predictor of cross-lingual performance variance.

1 Introduction

Polarization detection targets societal division along ideological, ethnic, religious, and gender lines distinct from hate speech, which targets individual-level toxicity (Naseem et al., 2026). Prior shared tasks treat related phenomena as binary or single-label problems (Zampieri et al., 2019, 2020; Basile et al., 2019). SemEval-2026 Task 9 (POLAR) is the first multilingual benchmark formalizing polarization *type* classification as a multi-label problem across 22 languages and 5 types (Naseem et al., 2026).

Two properties expose fundamental limits of standard multi-label classifiers. First, **absolute positive scarcity**: certain language-label pairs contain zero or one positive training instance—a data void no loss function can overcome (F1 = 0.0 regardless of model capacity), with a broader scarcity gradient degrading BCE into all-negative predictions across additional pairs. Second, **culturally structured co-occurrence**: political–religious dominates South Asian contexts while political–racial dominates Western contexts (section 7), yet independent sigmoid classifiers discard this signal entirely.

POLAR-LDA (**L**abel-**D**ependency **A**ware) addresses the recoverable portion of these challenges via: (i) a *Label Correlation Network* (LCN) applying Graph Attention (Veličković et al., 2018) over language-specific co-occurrence matrices, unlike prior global-matrix methods (Chen et al., 2019); (ii) *Asymmetric Loss* (Ridnik et al., 2021) (γ^-=4, γ^+=0) preventing collapse where minimal signal exists; and (iii) language-family *grouped ensemble training* mitigating gradient dominance by high-resource languages.
Our contributions are:

- Identification of data voids zero or one positive training instance as an irreducible boundary condition of multilingual polarization classification, distinct from modeling failures.
- Language-specific label dependency modeling via graph attention capturing culturally structured co-occurrence across 22 languages.
- Empirical evidence that per-label training volume, not linguistic family, is the primary predictor of cross-lingual performance variance.

2 Background

Label Dependency Modeling. Standard multi-label classification treats labels independently; graph-based methods recover co-occurrence signal. ML-GCN (Chen et al., 2019) applies GCNs over a *global* co-occurrence matrix for image classification; in NLP, label-specific attention (Ma et al.,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3621–3627

2021) and dual graph networks model dependencies for document categorization. All prior approaches assume a single shared co-occurrence structure none account for language-specific variation, which is critical when label patterns differ across cultural contexts the gap POLAR-LDA's language-specific $\mathbf{A}_L$ matrices directly address.

Loss Functions for Extreme Imbalance. Focal Loss (Lin et al., 2017) down-weights well-classified negatives but applies symmetric focusing. Asymmetric Loss (ASL; Ridnik et al., 2021) decouples positive and negative focusing parameters (γ^+, γ^-), outperforming both BCE and Focal Loss when positive rates fall below 5% the regime that characterizes multiple language-label pairs in this task.

3 Task and Data

SemEval-2026 Task 9 Subtask 2 (Naseem et al., 2026) requires multi-label assignment of five polarization types *political*, *racial/ethnic*, *religious*, *gender/sexual*, *other* to texts already identified as polarized, across 22 languages. Evaluation uses macro F1 averaged over all languages.

The training data exhibits extreme distributional skew: language-label pairs range from thousands of positive instances (Hindi political) to absolute data voids—zero instances for Hausa other, one for Bengali racial/ethnic—where F1 = 0.0 is the only possible supervised outcome. Table 1 summarizes the language groupings and dominant co-occurrence patterns that motivate our architecture.

Group	Languages	Dominant Pair
Indic	Hin, Nep, Urd, Ben, Pan, Odi, Tel	Pol. ↔ Rel.
Global	Eng, Deu, Spa, Ita, Pol, Rus, Tur, Zho, Arb, Fas	Pol. ↔ Rac.
AfroAs.	Amh, Hau, Swa, Mya, Khm	Rac. ↔ Rel.

Table 1: Language groups with dominant label co-occurrence pairs observed in training data, motivating the LCN (4.2) and grouped ensemble (4.4).

4 System Overview

POLAR-LDA comprises three components integrated sequentially: (1) a multilingual encoder, (2) a Label Correlation Network (LCN) modeling language-specific co-occurrence, and (3) Asymmetric Loss, trained within a grouped ensemble.

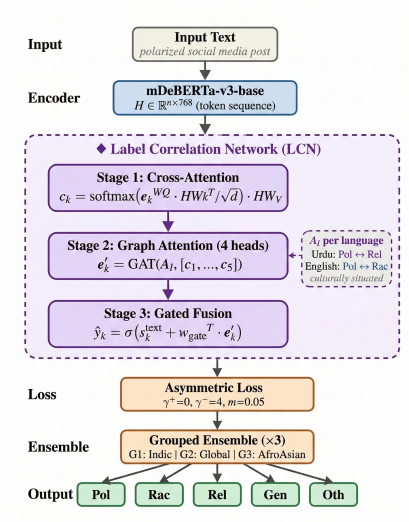

Figure 1: CoPol Architecture

Each addresses a specific failure mode of the mDeBERTa + BCE baseline identified during development: classifier collapse under positive scarcity, discarded label co-occurrence signal, and gradient dominance by high-resource languages. Figure 1 provides an overview.

4.1 Encoder

We use `mdeberta-v3-base` He et al., 2023 (a multilingual transformer (Devlin et al., 2019) with 86M parameters) over XLM-R-large (560M) for two reasons: (i) mDeBERTa-v3 achieves comparable or superior cross-lingual transfer on XNLI (Conneau et al., 2018) and XTREME benchmarks despite 6.5× fewer parameters (He et al., 2023), and (ii) the reduced memory footprint permits training three group-specific models simultaneously on a single GPU within the competition's compute constraints. SentencePiece tokenization covers 100+ languages. Given input text, we extract the full token-level representation $\mathbf{H} \in \mathbb{R}^{n\times 768}$ (where $n \leq 256$ tokens) and the [CLS] pooled vector $\mathbf{h} \in \mathbb{R}^{768}$.

4.2 Label Correlation Network

Independent sigmoid classifiers assume label independence an assumption violated by the culturally structured co-occurrence documented in Table 1. The LCN recovers this signal through three stages.

(1) Language-specific adjacency. For each language L, we construct an asymmetric co-occurrence matrix $\mathbf{A}_L \in \mathbb{R}^{5\times 5}$:

$$\mathbf{A}_L[i,j] = P(\text{type}_j{=}1 \mid \text{type}_i{=}1) \quad (1)$$

estimated from training-set label counts with Laplace smoothing (α=1) to regularize low-count cells. Edges with $\mathbf{A}_L[i,j] < 0.1$ are pruned; this threshold was selected on the development set from $\{0.05, 0.1, 0.15, 0.2\}$. $\mathbf{A}_L$ is *not* symmetrized: the directed graph preserves asymmetric dependencies (e.g., $P(\text{rel} \mid \text{pol}) \neq P(\text{pol} \mid \text{rel})$).

(2) Cross-attention over token sequence. Each label $k \in \{1, \ldots, 5\}$ has a learnable embedding $\mathbf{e}_k \in \mathbb{R}^{768}$, initialized from the mDeBERTa encoding of its natural-language name (e.g., "racial or ethnic polarization"). Each embedding attends over the *full token sequence* $\mathbf{H}$, not only [CLS]:

$$\mathbf{c}_k = \text{softmax}\left(\frac{\mathbf{e}_k \mathbf{W}_Q (\mathbf{H}\mathbf{W}_K)^\top}{\sqrt{d}}\right)\mathbf{H}\mathbf{W}_V \quad (2)$$

This produces a label-specific text representation $\mathbf{c}_k \in \mathbb{R}^{768}$, allowing each label to focus on different spans (e.g., *religious* attending to theological terms while *political* attends to institutional references). A preliminary score is computed as $s_k^{\text{text}} = \text{MLP}([\mathbf{e}_k; \mathbf{c}_k])$.

(3) Graph attention over label graph. The concatenated representations $\{\mathbf{e}_k \| \mathbf{c}_k\}_{k=1}^5$ are passed through a 4-head GAT layer (Veličković et al., 2018) with $\mathbf{A}_L$ as the directed adjacency, producing refined embeddings $\mathbf{e}'_k$. The final prediction fuses text-only and graph-refined signals via a learned gate:

$$\hat{y}_k = \sigma\big(s_k^{\text{text}} + \mathbf{w}_g^\top \mathbf{e}'_k\big) \quad (3)$$

Unlike ML-GCN (Chen et al., 2019), which applies a single global matrix across all inputs, our LCN uses *per-language directed graphs*, capturing asymmetric cultural co-occurrence structure.

Reliability caveat. $\mathbf{A}_L$ quality depends on sufficient co-occurrence counts. For languages where any label has fewer than 20 training instances (Hausa, Bengali), conditional probability estimates have high variance and the GAT may propagate noise. This creates a structural asymmetry: the LCN is most beneficial where label statistics are reliable, and least reliable where cross-label signal is most needed. A Bayesian estimator with cross-lingual priors would address this; we leave it to future work.

4.3 Asymmetric Loss

BCE assigns equal focusing to positives and negatives, collapsing to all-negative predictions when positive rates drop below $\sim$5%—observed for Italian political (F1 = 0.037) and Hausa other (F1 = 0.0). ASL (Ridnik et al., 2021) decouples focusing:

$$\mathcal{L}_{\text{ASL}} = -\frac{1}{N}\sum_i \Big[y_i(1-p_i)^{\gamma^+}\log p_i + (1-y_i)\,\hat{p}_i^{\gamma^-}\log(1-\hat{p}_i)\Big] \quad (4)$$

where $\hat{p}_i = \max(p_i - m, 0)$ applies hard thresholding. Hyperparameter choices: $\gamma^+ = 0$ (all positive samples contribute equally—critical when some labels have $\leq$5 instances), $\gamma^- = 4$ (aggressive suppression of easy negatives, following Ridnik et al., 2021), $m = 0.05$ (probability margin clipping negative-class predictions that are already confident).

4.4 Grouped Ensemble

Training a single model on all 22 languages causes high-resource languages (Hindi: $\sim$2k instances; Chinese: $\sim$1.5k) to dominate gradient updates, starving low-resource languages (Hausa: $\sim$200) of learning signal a known failure mode in massively multilingual training (Conneau et al., 2020). We partition languages into three groups (Table 1) based on script family, geographic–cultural proximity, and pretraining data volume, training an independent mDeBERTa + LCN + ASL model per group.

At inference, we ensemble across all three models with per-language weights optimized on the development set:

$$\hat{\mathbf{y}}_L = w_1^L f_{G_1}(\mathbf{x}) + w_2^L f_{G_2}(\mathbf{x}) + w_3^L f_{G_3}(\mathbf{x}) \quad (5)$$

where $\sum_i w_i^L = 1$ and weights are constrained to $[0, 1]$. The primary group typically receives $w \geq 0.7$; secondary models contribute marginal cross-lingual regularization. We note this grouping is typologically motivated; as shown in §7, within-group performance variance ($\bar{\sigma}_{\text{intra}}$=0.161) substantially exceeds between-group variance (σ_{inter}=0.011), indicating data-driven clustering may outperform linguistic heuristics.

5 Experimental Setup

We train exclusively on instances labeled as polarized in the official POLAR training split (Naseem et al., 2026). Co-occurrence matrices $\mathbf{A}_L$ are computed per-language from training labels with Laplace smoothing. All experiments use 5-fold

stratified cross-validation; fold probabilities are averaged for final predictions. The official evaluation metric is macro F1 averaged across all 22 languages. All training runs use a single NVIDIA T4x2 (32 GB) via Kaggle; grouped training completes in 8 hours total.

Parameter	G1	G2	G3
Learning rate	2×10^{-5}	2×10^{-5}	3×10^{-5}
Batch size	16	32	8
Max epochs	10	5	15

Shared across groups: AdamW (ϵ=1e-8, wd=0.01), seq. len=256, dropout=0.1, ASL $\gamma^+/\gamma^-/m = 0/4/0.05$, GAT heads=4, LCN edge threshold=0.1

Table 2: Hyperparameters by language group.

Table 2 reports group-specific hyperparameters. G3 (low-resource) uses a higher learning rate and more epochs to compensate for smaller training sets; G2 uses larger batches enabled by its data volume. All groups share AdamW (Loshchilov and Hutter, 2019) optimization with linear warmup (10% of steps) and decay, early stopping with patience 3 on dev macro F1.

6 Results

6.1 Official Performance

The submitted system achieves **0.567** macro F1 on the official CodaBench test set, using a single mDeBERTa-v3-base backbone per group without data augmentation or external resources. Performance spans a 0.528-point range: Hindi (0.784) to Italian (0.256).

Per-label difficulty varies substantially: political polarization is detected most reliably (avg. F1 = 0.685), followed by religious (0.613), racial/ethnic (0.553), gender/sexual (0.539), and other (0.444). The 0.241-point gap between political and other indicates that label-level ambiguity, not model capacity, is the binding constraint for overall performance.

Table 3 reports per-language results grouped by *linguistic family* (distinct from the training groups in Table 1, which partition by script and data regime). East/SE Asian languages achieve the highest family average (0.709), while European languages underperform (0.508) despite high pretraining coverage—driven primarily by Italian's anomalous collapse. Within-family variance dominates between-family variance: Hindi (0.784) vs. Bengali (0.340) within Indic alone spans a wider range than any two family averages.

Family	Language	F1
E/SE Asian	Chinese	.783
	Khmer	.675
	Burmese	.670
	Avg.	*.709*
Indic	Hindi	.784
	Urdu	.777
	Nepali	.774
	Odia	.515
	Telugu	.438
	Punjabi	.426
	Bengali	.340
	Avg.	*.579*
Turkic/Iran.	Persian	.605
	Turkish	.573
	Avg.	*.589*
Afro-Asiatic	Amharic	.658
	Arabic	.618
	Hausa	.325
	Avg.	*.534*
European	Spanish	.646
	Russian	.563
	German	.553
	Polish	.526
	English	.504
	Italian	.256
	Avg.	*.508*
Bantu	Swahili	.460

Table 3: Macro F1 per language, grouped by family.

6.2 Ablation

Table 4 reports the incremental contribution of each component on the test set. ASL provides the largest single gain (+.041 F1), recovering labels where BCE produces all-negative predictions. The LCN contributes +.030, confirming that structured label dependencies carry signal beyond independent classification. The grouped ensemble adds +.018, primarily reducing per-language variance rather than shifting the mean.

7 Analysis

Data voids as boundary conditions. Four language-label pairs yield near-zero F1: Bengali racial/ethnic (0.0), Hausa other (0.0), Italian political (0.037), and Italian other (0.035). These reflect absolute positive scarcity, not modeling failures—

Configuration	F1	Δ
mDeBERTa + BCE (baseline)	.478	—
+ Asymmetric Loss	.519	+.041
+ Label Correlation Net	.549	+.030
+ Grouped ensemble	.567	+.018

Table 4: Incremental ablation on test set (macro F1, 22-language average). Each row adds one component to the configuration above.

no loss function generates signal from absent data. Italian religious achieves F1 = 0.610 in the *same language*, isolating per-label instance count as the causal factor over language-level representation quality. The *other* label produces the lowest mean F1 (0.444) and highest cross-language variance (σ=0.231), consistent with its definition by exclusion.

Intra-family variance dominates inter-family variance. σ_{inter}=0.086 across six family means versus $\bar{\sigma}_{\text{intra}}$=0.119 within families. The Indic family alone spans 0.444 F1 points (Hindi 0.784 vs. Bengali 0.340; σ=0.193), exceeding the full range of family means. Burmese (0.670) and Khmer (0.675) outperform English (0.504) despite lower mDeBERTa pretraining coverage, indicating pretraining–domain alignment matters more than raw token coverage.

Language-specific co-occurrence structure. Per-language $\mathbf{A}_L$ matrices confirm culturally variable co-occurrence: Pol↔Rac dominates in English and German; Pol↔Rel in Hindi, Urdu, Nepali, Persian, and Turkish; Rac↔Rel in Swahili and Italian (Figure 2). The LCN exploits this—detecting political polarization in Urdu activates the political→religious edge in $\mathbf{A}_{\text{urd}}$, propagating signal that independent sigmoids discard.

8 Conclusion

POLAR-LDA achieves 0.567 macro F1 across 22 languages. Three findings emerge: (i) failures trace to per-label data scarcity, not representation quality; (ii) polarization co-occurrence is culturally situated, requiring language-specific dependency modeling; (iii) within-family variance ($\bar{\sigma}_{\text{intra}}$=0.119) exceeds between-family variance (σ_{inter}=0.086), indicating data volume dominates linguistic typology as a transfer predictor. Cross-lingual retrieval augmentation is the most promising direction for addressing data voids in low-resource pairs.

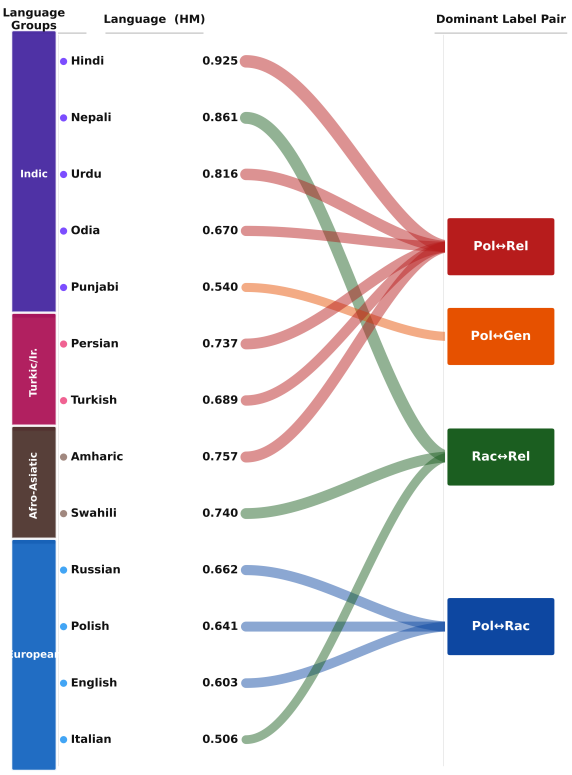

Figure 2: Dominant label pair per language by harmonic mean (HM) of per-label F1. Ribbon width ∝ HM score. Pol↔Rel, Rac↔Rel, Pol↔Rac, Pol↔Gen. Bengali and Hausa excluded (data-void labels).

Ethical Considerations

Polarization detection risks misuse for censorship. Our system classifies *types* without assessing harm—a judgment requiring cultural context unavailable at classification time—and a single decision boundary across 22 languages risks bias against culturally specific forms of political expression.

Limitations

Co-occurrence estimation under scarcity. For Hausa and Bengali (<20 co-occurrences), $\mathbf{A}_L$ estimates are high-variance; the GAT may propagate noise rather than signal.

Residual category incoherence. *Other* (σ=0.231, range 0.0–0.889) lacks coherent semantic boundaries; named subcategories would yield more learnable targets.

Domain mismatch. English (0.504) and German (0.553) underperform Burmese (0.670) and Khmer (0.675) despite higher pretraining coverage, indicating stylistic distance from polarization discourse outweighs vocabulary coverage as a bottleneck.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

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

Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3855–3864.

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. 2026. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

A Full Per-Language Per-Label Results

Family	Language	Pol.	Rac./Eth.	Rel.	Gen./Sex.	Other	Macro
E/SE Asian	Chinese	.747	.812	.921	.851	.584	.783
	Khmer	.776	.439	.685	.587	.889	.675
	Burmese	.826	.609	.532	.571	.810	.670
Indic	Hindi	.920	.812	.930	.763	.496	.784
	Urdu	.861	.768	.775	.743	.736	.777
	Nepali	.725	.844	.879	.771	.649	.774
	Odia	.703	.447	.640	.483	.302	.515
	Telugu	.556	.455	.271	.384	.524	.438
	Punjabi	.663	.284	.447	.455	.281	.426
	Bengali	.764	**.000**†	.366	.174	.395	.340
Turkic/Iran.	Persian	.813	.342	.674	.531	.665	.605
	Turkish	.763	.577	.628	.537	.361	.573
Afro-Asiatic	Amharic	.859	.641	.676	.571	.542	.658
	Arabic	.761	.642	.606	.548	.533	.618
	Hausa	.423	.467	.254	.483	**.000**†	.325
European	Spanish	.703	.586	.605	.785	.548	.645
	Russian	.686	.640	.595	.629	.262	.563
	German	.643	.573	.620	.639	.289	.553
	Polish	.764	.552	.517	.460	.336	.526
	English	.725	.516	.496	.500	.280	.503
	Italian	**.037**†	.432	.610	.166	**.035**†	.256
Bantu	Swahili	.344	.730	.750	.220	.255	.460
Average		**.685**	**.553**	**.613**	**.539**	**.444**	**.567**

† Data void: F1 < 0.05 due to ≤ 1 positive training instance for this language-label pair.

Table 5: Per-label F1 on the official test set for all 22 languages, grouped by linguistic family and sorted by macro F1 within each family. Four language-label pairs (†) fall below F1 $= 0.05$, corresponding to data voids identified in §7.