Does Hate Transfer? Cross-Lingual Generalisation of Offensive Content Detection Across Indic Languages

Madduri Purandhar Reddy **Sara Renjit**
Department of Computer Science & Engineering
IIIT Kottayam, India
{2022bcs0179, sara}@iiitkottayam.ac.in

Abstract

A common assumption in low-resource NLP is that cross-lingual transfer from a related language can substitute for target-language annotation when labelled data is scarce. We test this assumption for offensive content detection across five Indic languages by evaluating all twenty directed transfer pairs from a LLaMA-3.1-8B model fine-tuned with Low-Rank Adaptation (LoRA) on the MACD benchmark. Only three of twenty pairs achieve tolerable transfer loss below 15%, all involving Malayalam as the source language. Telugu is the hardest transfer target (average loss 33.8%), while Malayalam is the most transferable source (average loss 16.8%). Confusion-matrix analysis reveals two distinct failure modes: Tamil- and Kannada-trained models are conservative under-flaggers that miss 73–82% of offensive content with near-zero false alarms, while Malayalam-trained models are aggressive flaggers that miss far less (39%) but over-flag at 21%. These patterns do not follow typological structure: a Spearman correlation between URIEL typological similarity and transfer F1 yields $\rho = -0.254$ ($p = 0.281$), failing to confirm the typological hypothesis. Our results indicate that cross-lingual shortcuts are unreliable for this task and that language-specific annotation cannot be avoided by appealing to linguistic family membership.

1 Introduction

Building offensive content detection systems for India's many languages is resource-intensive, requiring native-speaker expertise and substantial annotation effort. A natural response is cross-lingual transfer: train on one language and apply to another. This is intuitive in the Indic context—Tamil, Telugu, Kannada, and Malayalam share Dravidian morphological structure (Krishnamurti, 2003); Hindi shares contact features with Dravidian languages through centuries of convergence (Emeneau, 1956).

We test this expectation systematically using a LLaMA-3.1-8B model fine-tuned on MACD (Gupta et al., 2022) via LoRA (Hu et al., 2022), evaluating all twenty directed zero-shot cross-lingual transfer pairs. We further test whether transfer patterns are explained by typological relatedness using URIEL feature vectors (Littell et al., 2017).

Cross-lingual transfer correlates with typological proximity on structural NLP tasks (Pires et al., 2019; Hu et al., 2020). However, for pragmatically motivated tasks such as offensive language detection, prior work suggests that cultural rather than structural factors play a larger role (Sun et al., 2021; Zhou et al., 2023; Ousidhoum et al., 2019). Existing work on Indic offensive language detection has largely focused on monolingual or code-mixed settings (Chakravarthi et al., 2022; Gupta et al., 2022), and cross-lingual studies typically cover a partial subset of pairs or directions without a quantitative typological control. Our work extends these findings to the Indic setting, providing the first complete directed transfer matrix across five Indic languages on a single benchmark together with a quantitative typological control.

Our contribution is empirical and cautionary: for offensive content detection in Indic languages, language-specific annotation cannot be substituted by appealing to linguistic family membership.

2 Experimental Setup

Dataset. We use MACD (Gupta et al., 2022), real user-generated ShareChat comments annotated for offensive versus non-offensive content. MACD defines offensive content as comments containing abuse, profanity, insults, threats, or content targeting identity attributes such as caste, religion, gender, or community, following the original annotation guidelines of Gupta et al. (2022). Table 1 shows per-language splits and class balance; all subsets are near-balanced (46–52% offensive). The

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

15% tolerable loss threshold follows Ousidhoum et al. (2019), above which degradation is considered significant for deployed moderation systems.

Language	Train	Dev	Test	% Off.
Tamil	18,000	6,000	6,000	46.4
Telugu	18,000	6,000	6,000	52.0
Hindi	20,183	6,728	6,728	52.0
Kannada	19,761	6,587	6,587	48.3
Malayalam	15,508	5,170	5,170	49.5

Table 1: MACD dataset splits and per-language offensive-class percentage. All five subsets are near-balanced (46–52% offensive, imbalance ratio ≤ 1.16 in every case), so class-balance differences are unlikely to explain the cross-lingual transfer asymmetries reported in Section 3.

Model. We fine-tune LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) on each MACD language independently using LoRA (Hu et al., 2022) (r=8, α=16, dropout 0.05, learning rate 1.5×10^{-4}, 3 epochs, bf16 mixed precision on an AMD MI300X GPU). The checkpoint with the highest validation macro-F1 is used. Monolingual test performance is shown in Table 2; these serve as upper-bound references for transfer evaluation.

Language	P	R	F1	MCC
Tamil	0.872	0.870	0.869	0.740
Telugu	0.882	0.881	0.881	0.762
Hindi	0.857	0.856	0.856	0.713
Kannada	0.852	0.847	0.846	0.697
Malayalam	0.840	0.840	0.840	0.677
Average	0.861	0.859	0.858	0.718

Table 2: Monolingual test performance (macro-P, macro-R, macro-F1, MCC).

Transfer protocol. A model fine-tuned on source s is applied zero-shot to target t's test set. Transfer loss is defined as:

$$\Delta F1(s \to t) = \frac{F1_{\text{mono}}(t) - F1_{\text{transfer}}(s \to t)}{F1_{\text{mono}}(t)} \times 100$$

We evaluate all twenty directed pairs. Losses below 15% are tolerable; above 30% are prohibitive. Typological control uses URIEL feature vectors (syntactic, phonological, geographic) via lang2vec, with Spearman correlation computed between cosine typological similarity and transfer F1 over all 20 pairs.

Source	Target	F1	Loss (%)
Tolerable ($\Delta F1 < 15\%$)			
Malayalam	Tamil	0.728	14.1
Kannada	Hindi	0.727	12.9
Malayalam	Hindi	0.716	14.0
Moderate ($15\% \leq \Delta F1 < 30\%$)			
Malayalam	Telugu	0.691	19.0
Malayalam	Kannada	0.647	19.9
Telugu	Hindi	0.638	21.7
Hindi	Kannada	0.638	20.8
Hindi	Malayalam	0.620	22.0
Telugu	Kannada	0.618	22.7
Hindi	Tamil	0.617	25.2
Hindi	Telugu	0.597	28.4
Tamil	Hindi	0.588	26.8
Telugu	Malayalam	0.556	28.4
Prohibitive ($\Delta F1 \geq 30\%$)			
Telugu	Tamil	0.561	30.8
Tamil	Malayalam	0.539	30.1
Kannada	Tamil	0.504	36.6
Kannada	Malayalam	0.502	33.8
Kannada	Telugu	0.491	39.0
Tamil	Kannada	0.475	37.1
Tamil	Telugu	0.390	49.0

Table 3: Complete directed cross-lingual transfer results (all 20 pairs), sorted by F1 within tier. ΔF1 is relative to the monolingual target-language baseline in Table 2.

3 Results and Analysis

Table 3 presents the complete transfer matrix; Figure 1 plots typological similarity against transfer F1.

Malayalam is the strongest transfer source. All four Malayalam-source pairs fall within the tolerable or low-moderate range, yielding an average source loss of 16.8%—the lowest of any source language. This holds across within-family Dravidian targets (Tamil, Kannada) and the cross-family Hindi target. The shared Proto-South-Dravidian I ancestry of Tamil and Malayalam (Krishnamurti, 2003) may partially explain this advantage, though we treat this as a hypothesis consistent with the data rather than a demonstrated mechanism. The error analysis below shows that Malayalam-trained models are also the most aggressive flaggers of offensive content, recovering substantially more of it than other sources but at the cost of a higher false-alarm rate.

Telugu is the hardest transfer target. Transfers to Telugu yield the highest average target loss at 33.8%, including from Tamil and Kannada which share Dravidian morphological structure with Telugu. Even Malayalam achieves only 19.0% loss on Telugu. The monolingual Telugu model achieves

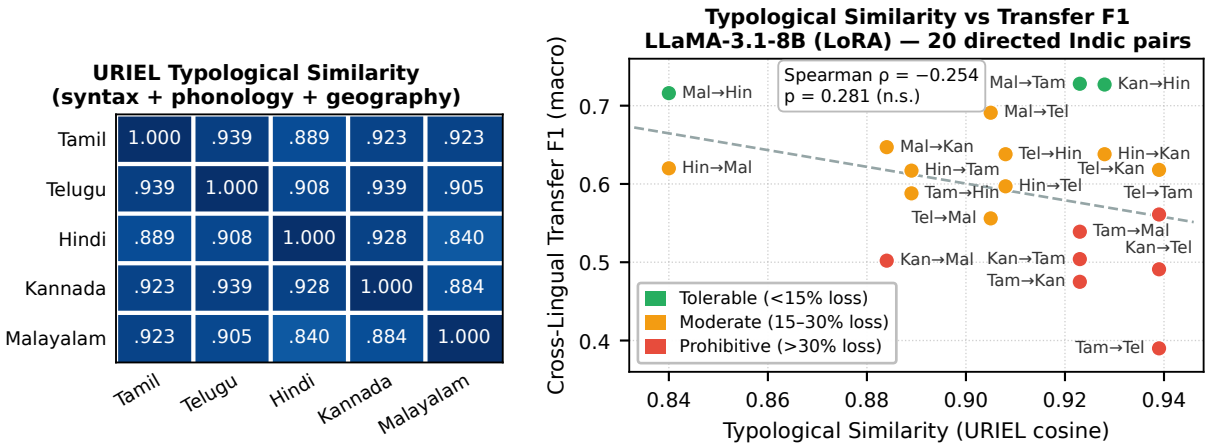

Figure 1: **Left:** URIEL typological similarity matrix (cosine similarity of syntax + phonology + geographic features). All pairwise similarities fall in the narrow range 0.840–0.939. **Right:** Typological similarity versus cross-lingual transfer F1 for all 20 directed pairs. Colour indicates transfer tier: green = tolerable, orange = moderate, red = prohibitive. Despite near-uniform typological similarity, transfer F1 ranges from 0.390 to 0.728. The non-significant Spearman correlation ($\rho = -0.254$, $p = 0.281$) is consistent with cultural rather than structural factors as a candidate explanation, though we do not directly test this.

the highest F1 (0.881), indicating that Telugu offensive content is learnable in-language but that its semantics are highly localised (see the error analysis below).

Transfer is strongly asymmetric. The largest asymmetry is Malayalam→Tamil (F1: 0.728) versus Tamil→Malayalam (F1: 0.539), a gap of 0.189; Malayalam→Kannada similarly outperforms Kannada→Malayalam by 0.145. Asymmetry of this magnitude is not predicted by typological distance, which is symmetric by definition.

Typological similarity does not predict transfer. URIEL pairwise similarities span a compressed range (0.840–0.939). Spearman rank correlation between typological similarity and transfer F1 yields $\rho = -0.254$ ($p = 0.281$), not significant at $\alpha = 0.05$. Transfer F1 varies widely (0.390–0.728) with no coherent typological pattern, consistent with prior work showing cultural rather than typological features predict transfer on pragmatic tasks (Sun et al., 2021; Zhou et al., 2023).

Error analysis. Confusion matrices reveal that transferred models fail by missing offensive content rather than by over-flagging in every pair (false negatives exceed false positives in all 20). The failure mode varies sharply by source. Tamil- and Kannada-trained models are extreme under-flaggers: they miss on average 82% and 73% of offensive content while producing false alarms on just 2–5% of benign content — source-trained "offensive" essentially does not fire on out-of-language text. Malayalam-trained models show the opposite profile, recovering the most offensive content of any source (e.g. 69% recall in Tamil, 60% in Kannada) but at the highest false-alarm rates (21% on average, peaking at 31% on Kannada). Hindi and Telugu sources are intermediate (miss 54–67%, FA 9–12%). Macro-F1 alone does not disclose which regime a deployed moderator inherits, and the two failure modes have very different downstream consequences.

4 Conclusion

We have presented a complete directed cross-lingual transfer analysis of offensive content detection across five Indic languages with a quantitative typological control. Only 3 of 20 directed transfer pairs achieve tolerable loss, all sourced from Malayalam; Telugu is the most resistant target; transfer is strongly asymmetric; and patterns are not explained by URIEL typological similarity ($\rho = -0.254$, $p = 0.281$). Confusion-matrix analysis further shows that sources differ qualitatively in failure mode: Tamil and Kannada sources under-flag almost everything, while Malayalam sources over-flag but recover the most offensive content. These results are consistent with — but do not establish — cultural proximity between speaker communities as a candidate explanation (Sun et al., 2021; Zhou et al., 2023). The practical implication is direct: for Indic offensive content moderation, zero-shot cross-lingual shortcuts are unreliable, and practitioners should treat each language's offensive

semantics as substantially distinct.

Limitations

MACD's binary label taxonomy does not formally distinguish implicit from explicit offensive content; we follow the original authors' characterisation but this cannot be verified from labels alone. The cultural-proximity interpretation is a candidate explanation rather than a tested mechanism: we do not compute a direct cultural similarity measure, and the URIEL control has limited statistical power across only 20 data points, so its non-significant result should be read as failure to confirm rather than disconfirmation. All experiments use a single decoder LLM with LoRA; replication on encoder-only models (e.g. XLM-R, IndicBERT) is the most important open question and the primary direction for follow-up work. Our headline finding is robust to the 15% threshold: at 10% only one pair is tolerable, at 20% five are — the conclusion that most directed transfer is unreliable, with Malayalam the sole consistently transferable source, holds in all three regimes.

Ethics Statement

This work uses the publicly available MACD dataset released with appropriate ethical review. The finding that Telugu moderation via cross-lingual transfer is unreliable is an argument for directing investment toward language-specific annotation, not a reason to delay content moderation investment.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand M. Kumar, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, R. L. Hariharan, John P. McCrae, and Elizabeth Sherly. 2022. DravidianCodeMix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):1 –1060.

Murray B. Emeneau. 1956. India as a linguistic area. *Language*, 32(1):3–16.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Hastagiri P. Vanchinathan, and Animesh Mukherjee. 2022. MACD: Multilingual abusive comment detection at scale for Indic languages. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.