When Does Hate Transfer?
A Large-Scale Study of Cross-Lingual Hate Speech Detection

Anonymous ACL submission

Abstract

Effective detection of hate content is essential for maintaining healthy online communities, yet performance can vary substantially across languages. This work investigates a core practical question: Can training data from one language support hate-speech detection in another language with comparable performance? To answer this question, we conduct a large-scale study across 14 languages. Our analysis shows that transfer effectiveness could depend on several factors—including, but not limited to, dataset size, resources and temporal overlap. At the same time, we observe consistent patterns in which certain language pairs exhibit stronger transfer performance. Upon further analysis, we find that these patterns can be attributed to cultural cues or shared societal characteristics. Our analysis includes low-resource languages, offering practical insights into when cross-lingual transfer can support them and where its limitations emerge. These findings provide guidance for designing more reliable and generalizable hate detection systems.

Warning: This paper includes examples of hate speech for research purposes.

1 Introduction

Hate language detection aims to identify harmful expressions in online text, such as hate speech, harassment, and threats (Hoang et al., 2024). As digital communication expands globally, exposure to such content has increased, motivating efforts to build automated systems capable of supporting safer online communities. However, what counts as hate speech often depends on several criteria and even human annotators vary in how they perceive hate across contexts (Sap et al., 2022a). These differences pose a challenge for developing models that generalize reliably across languages.

Large Language Models (LLMs) offer strong multilingual representations and are increasingly used in hate detection (Khondaker et al., 2023; Kumar et al., 2024; Abaskohi et al., 2024). However, the cross-lingual performance is not uniform, and the factors shaping transfer effectiveness should be studied (Meyer and Buys, 2024). This raises a practical and underexplored question: Can training data from one language support hate-speech detection in another language with comparable performance? Several factors could influence transfer success, including but not limited to dataset size, label definitions, temporal alignment, and linguistic similarity. Preliminary work suggests that broader social or regional contexts may also shape how hate expressions are formed and recognized (Bokaei et al., 2025; Pawar et al., 2024).

In this study, we conduct a large-scale analysis across 14 languages from diverse regions and domains. We construct a comprehensive cross-lingual transfer matrix by training hate-speech classifiers in each language and evaluating them on all others to assess the effectiveness of cross-lingual transfer for hate-speech detection. We then perform two complementary experiments: (1) integrating all languages into a single multilingual training set, (2) constructing a balanced multilingual dataset with equal representation per language. These setups allow us to examine how data volume and representational balance influence cross-lingual generalization. The current study addresses the following research questions (RQs):

RQ1. How does the performance of cross-lingual training compare to in-language training for hate-speech detection?

RQ2. Would using training data from multiple languages improve performance compared to relying on a single source language?

RQ3. What could be the potential reasons for the success or failure of transfer learning in hate-speech detection?

These questions aim to clarify whether cross-

lingual transfer is effective for hate-speech detection, and to identify the conditions under which it succeeds or fails, providing insights for developing more reliable and generalizable hate detection systems. We use LLaMA 3 as our model, based on preliminary experiments indicating that it achieves the best average performance across our 14 languages: Turkish, Urdu, Persian, Arabic, Hindi, Bengali, Indonesian, Malay, English, Spanish, French, German, Korean, and Chinese.

2 Related Work

Hate Speech Detection. Early work on hate language detection focused primarily on English, using Machine Learning and Deep Learning approaches to classify hate speech (Asogwa et al., 2022; Davidson et al., 2017; Mullah and Zainon, 2021; Malik et al., 2024; Zimmerman et al., 2018; Zhou et al., 2020; Roy et al., 2020; Zhang et al., 2018; Bade et al., 2024; Aiyanyo et al., 2020; Cao et al., 2020; Risch et al., 2020; Wang et al., 2020; Pamungkas and Patti, 2019; Van Hee et al., 2015; Guo and Gauch, 2024; Cano Basave et al., 2013). Subsequent research expanded to a wider range of languages, including Indonesian (Ibrohim and Budi, 2019), Danish (Sigurbergsson and Derczynski, 2020), Arabic (Mubarak et al., 2021; Bensalem et al., 2023), Korean (Jeong et al., 2022), Chinese (Deng et al., 2022), Greek (Pitenis et al., 2020), Persian (Delbari et al., 2024), and Hindi (Gupta et al., 2022; Kapoor et al., 2019).

Recent work highlights key challenges in multilingual hate detection, including comparisons of translation-based and multilingual pipelines (Bell et al., 2025), limitations in zero-shot transfer for taboo expressions (Nozza, 2021), and the influence of annotator identity and beliefs on hate judgments (Sap et al., 2022b). Multimodal datasets such as Multi3Hate reveal substantial cross-cultural variation in hate-speech perception (Bui et al., 2025), while surveys document strong geographic and cultural imbalances in existing hate-speech resources (Tonneau et al., 2024). Additional studies explore the feasibility of parallel multilingual hate-speech data via machine translation (Korre et al., 2024). Shared workshops such as OffensEval, HASOC, OSACT5, and GermEval continue to support benchmarking across languages (Zampieri et al., 2019; Mandl et al., 2019; Mubarak et al., 2022; Wiegand et al., 2018). **Cross-lingual Transfer Learning for Hate Speech.** Transfer learning has been widely applied across NLP tasks, including sentiment analysis (Fsih et al., 2022; Husain et al., 2022), irony detection (Golazizian et al., 2020), machine translation (Zoph et al., 2016; Kim et al., 2019; Adebara and Abdul-Mageed, 2021), fake news detection (Cruz et al., 2020), and offensive language classification (Zhou et al., 2023). In multilingual hate detection, prior work examines cross-lingual transfer from English to low-resource languages using contextual embeddings (Ranasinghe and Zampieri, 2021), compares multilingual and translation-based pipelines for Arabic–English (El-Alami et al., 2022), and improves low-resource performance through nearest-neighbor retrieval (Ghorbanpour et al., 2025). A recent survey reviewing 67 studies provides a comprehensive overview of cross-lingual offensive language detection methods and datasets (Jiang and Zubiaga, 2024). Research explicitly examining cultural or social influences on transfer remains limited. Zero-shot studies show that language-specific taboo or colloquial expressions often fail to transfer across languages (Nozza, 2021). Cross-lingual analyses of English, Chinese, and Korean further reveal culture-specific biases that affect transfer, although few-shot adaptation can mitigate some of these effects (Zhou et al., 2023). Multimodal work also highlights cross-cultural variation in hate-speech judgments (Bui et al., 2025). For example, Zhou et al. (2023) find that Korean-to-Chinese transfer can outperform English-to-Chinese, suggesting that contextual or social alignment may support better generalization. Similarly, Bokaei et al. (2025) report that transfer into Persian is stronger from Arabic and Indonesian than from English, indicating that cross-lingual effectiveness varies substantially across source languages. Our study complements the literature by analyzing transfer behavior across 14 languages and examining correlations between transfer effectiveness and different factors. Additional discussion of related work is provided in the Appendix.

3 Dataset

We use publicly available datasets from peer-reviewed publications and shared evaluation tasks to ensure data quality and comparability across languages. Our study includes fourteen languages drawn from diverse linguistic families and geographical regions. Most datasets originate from Twitter, though some languages include additional sources such as Facebook and YouTube. Their collection periods span roughly 2016–2023. Table 1

Language	Data Size	Train	Test	Train$_{Hate}$	Test$_{Hate}$	Source	Period	Ref.
Bengali (BN)	30000	9007	1002	4504	501	YT, FB	2017–2020	(Romim et al., 2020)
Urdu (UR)	10041	9000	1000	4500	300	X	–	(Rizwan et al., 2020)
Hindi (HI)	8192	4171	467	2222	241	X, FB, WA	–	(Bhardwaj et al., 2020)
Persian (FA)	7000	2060	230	1002	117	X	2020–2022	(Delbari et al., 2024)
Indonesian (IN)	13169	9005	1003	4502	502	X	2018	(Ibrohim and Budi, 2019)
Malay (MS)	5014	3403	378	1701	189	X	2022–2023	(Maity et al., 2023)
Arabic (AR)	15014	9005	1002	4503	502	X	2016–2018	(Albadi et al., 2018)
Turkish (TR)	52672	9010	1003	4505	502	X, IG	2016–2023	(Toraman et al., 2022)
Korean (KD)	9341	3065	300	1520	142	X	2018–2020	(Moon et al., 2020; Lee et al., 2022)
Chinese (ZH)	8065	4950	318	2470	151	Web	–	(Moosa and Najiba, 2022)
Spanish (ES)	29856	9000	1000	4731	625	X	2016–2023	(Tonneau et al., 2024)
German (DE)	58025	9008	1003	4503	502	X, FB	2016–2023	(Tonneau et al., 2024)
French (FR)	18072	7483	829	3739	413	X	2016–2023	(Tonneau et al., 2024)
English (EN)	360494	9000	1038	4500	502	FB, YT, RD, X	2016–2023	(Tonneau et al., 2024)

Table 1: Hate language detection datasets. FB = Facebook, YT = YouTube, X = Twitter, WA: WhatsApp, IG = Instagram, RD = Reddit. The definition of Hate and more details for each dataset are shown in Table 4

summarizes the dataset sizes, the training set sizes, the number of hate samples used for fine-tuning, and their sources and collection periods.

All datasets exhibit class imbalance, with varying proportions of hate versus non-hate content. To conduct controlled comparisons across languages, we standardize the amount of training data wherever possible. Specifically, among the languages with sufficiently large datasets, we identify the smallest available number of hate examples (approximately 4,500) and sample an equal number of non-hate instances, producing balanced datasets of roughly 10,000 examples per language. We use an 80/10/10 split for train, validation, and test sets. Balancing all training sets to a similar size helps isolate the effect of language rather than dataset volume, and avoids situations where high-resource languages (such as English or German) would appear to transfer better simply because they have much larger datasets. It also keeps the computational requirements reasonable for training fourteen models. All balancing steps were applied only when the underlying dataset contained enough annotated hate examples to support this procedure. However, this was not the case for all languages. For Hindi, Malay, Korean, Chinese, and Persian, the available datasets contain fewer hate instances.

To enable consistent evaluation across datasets with fine-grained annotation schemes (e.g., hostile, hate, offensive, abusive), we focus only on hate-speech samples and ignore other labels. The original label schemes and hate-speech definitions are summarized in Appendix Table 4. Most datasets adopt a broadly consistent definition of hate speech as language attacking individuals or groups based on protected identity attributes.

4 Experimental Setup

We investigate cross-lingual transfer in hate detection using the datasets summarized in Table 1. Our experimental design includes (1) single-language fine-tuning with cross-lingual evaluation, (2) zero-shot baselines, and (3) two complementary multilingual fine tuning setups to investigate alternative explanations for transfer behavior.

Model Selection and Baselines: Before selecting a primary model to be fine-tuned, we initially experimented multiple baselines: 1) Zero-shot baselines, including GPT-4o, Gemma-3, and LLaMA-3-Instruct, and 2) Fine-tuning baselines, including Gemma-3 and LLaMA-3-8B. Fine-tuned baselines using both Gemma and LLaMA outperformed all zero-shot baselines across all the 14 languages. Further, fine-tuning LLaMA-3-8B yields higher average macro-F1 compared to Gemma 3, which aligns with recent work (Guo and Sarker, 2025; Bokaei et al., 2025) highlighting LLaMA-3's multilingual strength. Thus, we adopt LLaMA-3-8B for all our upcoming fine-tuning experiments. Table 5 in the Appendix reports the baseline results.

Fine-tuning Configuration: All fine-tuning uses the official LLaMA-3-8B pretrained checkpoint. We apply AdamW with a learning rate of 1e-5, batch size 8, and a maximum sequence length of 512 tokens. Training proceeds for up to 10 epochs with early stopping (patience = 3) based on validation F1 score. All experiments are run with five random seeds, and we report the mean and standard deviation of F1 across runs to assess result stability. All experiments are conducted on a single NVIDIA Quadro RTX 8000 (48 GB). Performance is reported using macro-F1.

Single-Language Fine-Tuning: For each language, we fine-tune a separate LLaMA-3-8B model

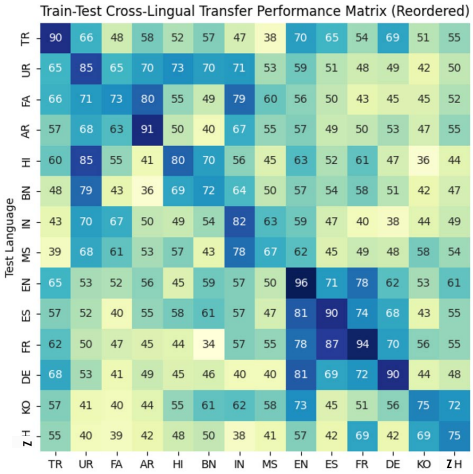

Figure 1: Cross-lingual F1 scores across 14 languages.

on its own training set and evaluate it on all 14 test sets. For example, a model fine-tuned on English is evaluated on English and on the remaining 13 languages. This procedure produces a 14 × 14 cross-lingual transfer matrix, allowing us to examine empirical transfer patterns. Because LLaMA is pretrained as a multilingual model, fine-tuning on one language does not remove its multilingual knowledge; instead, it specializes the model toward the target language's hate characteristics.

The 14×14 single-language fine-tuning experiment directly addresses RQ1 and reveals if any language pairs transfer well or poorly. To answer RQ2, we design two complementary experiments. The first examines whether increased data volume and multilingual diversity improve performance. We refer to this experiment as Global Integration (Imbalanced). The other evaluates whether ensuring equal representation across languages affects transfer behavior by removing dataset-size differences as a confounding variable. We refer to this experiment as Global Integration (balanced). Together, these additional setups allow us to investigate whether transfer effectiveness stems from data size, multilingual exposure, or any structural similarities observed in the main 14×14 matrix.

Global Integration (Imbalanced) Experiment: We merge the full training sets from all 14 languages into one multilingual dataset and fine-tune a single LLaMA-3-8B model. The model is then evaluated separately on each language's test set. This experiment tests whether greater data volume and diversity produce better generalization.

Global Integration (Balanced) Experiment: To control for dataset-size differences, we sample an equal number of hate (n = 1000) and non-hate (n = 1000) instances from each language and merge them into a balanced multilingual dataset. A single model is fine-tuned and evaluated on each language's test set, isolating the effect of balanced representation by controlling for dataset size.

Cluster Tuning: Finally, from the main 14×14 transfer matrix, we identify clusters of languages that show strong mutual transfer from the initial experiment. For each cluster, we merge the training sets and fine-tune a single model, which is then evaluated separately on each language within the cluster. In all attempts, we use the same train, validation, and test splits for each language to ensure comparability.

5 Results

Figure 1 presents the 14 × 14 cross-lingual transfer matrix for hate content detection. Each cell reports the F1 score obtained when a model is fine-tuned on the language shown on the x-axis and evaluated on the test language on the y-axis. The matrix contains 196 train–test combinations, providing a comprehensive view of both in-language and cross-lingual performance. Across languages, the highest F1 scores mostly occur on the diagonal, where the model is fine-tuned and tested on the same language. These in-language results are consistently strong: Turkish (90), Arabic (91), English (96), French (94), Spanish (90), and German (90) achieve F1 scores at or above 90. In a few cases, cross-lingual models outperform the in-language baseline; for example, models fine-tuned on Arabic or Indonesian yield higher scores for Persian than models fine-tuned on Persian itself.

Off-diagonal values show substantial variation in the quality of cross-lingual transfer. Based on these off-diagonal patterns, several cluster of languages emerge in which transfer performance is consistently higher within the cluster than across the cluster. These include: (1) Persian–Arabic–Urdu–Indonesian, (2) Urdu–Hindi–Bengali, (3) Indonesian–Malay, (4) Korean–Chinese, and (5) French–Spanish–English–German. As Figure1 represents, models fine-tuned on Urdu perform strongly on Hindi (85), and English-trained models yield relatively high scores on French (78), Spanish (74), and German (69). There are also cases of strong mutual transfer, such as $French \leftrightarrow English$ (78) and $Bengali \leftrightarrow Urdu$ (BN→UR: 70; UR→BN: 79). Lower scores appear for languages like Bengali, whose models perform less well on European test sets (BN→FR:

	TR	UR	FA	AR	HI	BN	IN	MS	EN	ES	FR	DE	KO	ZH
Solo-Source TL	90 ±0.6	85 ±0.6	80 ±0.8	91 ±0.6	85 ±0.7	79 ±0.7	82 ±0.6	78 ±0.7	96 ±0.5	90 ±0.6	94 ±0.5	90 ±0.6	**75** ±0.7	75 ±0.7
Global-Imb	**92** ±0.3 [+2]	88 ±0.3 [+3]	64 ±0.4 [-16]	87 ±0.3 [-4]	73 ±0.4 [-12]	66 ±0.4 [-13]	84 ±0.3 [+2]	62 ±0.4 [-16]	**98** ±0.2 [+2]	95 ±0.3 [+5]	**96** ±0.3 [+2]	94 ±0.3 [+4]	71 ±0.4 [-4]	72 ±0.4 [-3]
Global-bal	91 ±0.5 [+1]	87 ±0.5 [+2]	67 ±0.6 [-13]	89 ±0.5 [-2]	81 ±0.6 [-4]	69 ±0.6 [-10]	83 ±0.5 [+1]	63 ±0.6 [-15]	96 ±0.4 [0]	92 ±0.5 [+2]	94 ±0.4 [0]	93 ±0.5 [+3]	73 ±0.6 [-2]	76 ±0.6 [+1]
Observed Cluster	90 ±0.4 [0]	**90** ±0.4 [+5]	**83** ±0.5 [+3]	**93** ±0.4 [+2]	**90** ±0.5 [+5]	**83** ±0.5 [+4]	**85** ±0.4 [+3]	**80** ±0.5 [+2]	97 ±0.3 [+1]	**96** ±0.4 [+6]	**96** ±0.4 [+2]	**95** ±0.4 [+5]	**75** ±0.5 [0]	**78** ±0.5 [+2]

Table 2: Performance comparison across different fine-tuning strategies. Results are reported as mean ± standard deviation over five random seeds. Bracketed numbers indicate improvement or decline relative to Solo-Source TL.

51; BN→DE: 42), and Chinese models show weaker transfer to several languages (ZH→ES: 42; ZH→AR: 38). Languages such as Korean, Chinese, and Malay generally show lower transfer performance both as train and test languages.

Table 2 summarizes results from the three complementary experiments. In the Global-Imbalanced setup, combining the full training data from all languages leads to performance gains for several high-resource languages, including English, French, and Spanish (average gain: +2.86), as well as improvements for languages such as Urdu. In contrast, most lower-resource languages—such as Bengali, Persian, and Malay—show declines relative to their single-language baselines (average: -5.43).

In the observed cluster tuning experiment, fine tuning in language-selective clustes from the 14×14 matrix yields consistent performance improvements for nearly all cluster members (average gain: +5.62). Gains are observed in groups such as Hindi–Bengali–Urdu, Arabic–Persian–Indonesian, and English–French–Spanish–German, though the magnitude of improvement varies across languages. In the Global-Balanced setup, equal sampling from all languages reduces the effect of data imbalance. High-resource languages still improve over their single-language baselines, though gains are smaller than in the imbalanced condition (average: +1.57). For lower-resource languages, the balanced dataset mitigates some of the declines seen in the Global-Imbalanced experiment, with most showing small gains and some remaining close to baseline.

6 Analysis

The 14×14 transfer matrix in Figure 1 shows several structural patterns. We have already mentioned the five clusters of languages with relatively strong mutual transfer. While most languages achieve their highest performance through in-language learning, we also observe groups of language pairs where transfer learning produces better—or at least comparable—results. Table 6 in the Appendix presents some hate samples correctly predicted only via transfer learning. We first examine whether dataset size or temporal overlap can account for these patterns. To understand whether dataset-related factors can account for the transfer patterns observed in Figure 1, we evaluate each cluster in terms of three properties: dataset size, resource level, and temporal overlap.

Dataset Size: Across the clusters, dataset size shows partial but inconsistent alignment. The French–Spanish–English–German group comprises some of the largest datasets (7k–9k hate samples), which could make within-group transfer more stable. The Urdu–Hindi–Bengali cluster also falls within a relatively similar medium range (3k–5k hate samples), which may support some internal consistency. In contrast, the Persian–Arabic–Urdu–Indonesian cluster includes languages with markedly different dataset sizes: Persian (∼ 1k hate) is much smaller than Arabic or Indonesian (∼4.5k hate), yet it clusters with them despite this imbalance. Similarly, Indonesian–Malay shows large variation, as Malay contains substantially fewer hate examples. The Korean–Chinese group also exhibits different dataset scales, with Chinese containing roughly twice as many hate examples as Korean. Taken together, dataset size sometimes aligns with cluster boundaries (notably in English- German- French - Spanish and South Asian groups), but not in clusters involving Persian, Malay, Korean, and Chinese, indicating that dataset size alone cannot explain the overall structure.

Dataset Resource: Many clusters, such as Urdu–Hindi–Bengali, Persian–Arabic–Urdu–Indonesian, and Indonesian–Malay, consist entirely of Twitter-based datasets. However, this pattern does not distinguish them meaningfully, because almost all languages

in the study—including those that do not cluster together—are also drawn primarily from Twitter. Conversely, the Korean–Chinese cluster does not share a unified source either: Korean is drawn from Twitter, while Chinese originates from general web data. Thus, data source alignment does not systematically track the cluster structure.

Temporal Overlap: A similar mixed pattern appears when comparing dataset collection periods: English- German- French - Spanish- share nearly identical time spans (2016–2023), which could plausibly support more consistent transfer. The Urdu–Hindi–Bengali datasets likewise cover broadly similar years. However, the Persian–Arabic–Urdu–Indonesian cluster shows substantial variation: Persian (2020–22) does not overlap with Arabic (2016–18) or Indonesian ($\leq$2018), yet all appear in the same cluster. In addition , the Indonesian–Malay pair has almost no temporal overlap ($\leq$2018 vs. 2022–23), yet they form a clear transfer cluster. The analysis for Korean–Chinese is not possible since the Chinese time window is unspecified. These comparisons show that temporal alignment may support transfer in some clusters (English - German - French - Spanish and South Asian) but fails to predict clusters in others.

6.1 Quantitative Analysis

We examined several measurable dataset properties—dataset size difference, temporal overlap, and linguistic similarity using typological feature vectors derived from URIEL/lang2vec (Bjerva et al., 2019)—to assess whether they explain the transfer structure observed in the matrix. For each language pair, we compute Spearman correlations (Spearman, 1961) between these factors and the off-diagonal F1 scores from the transfer matrix. A brief summary of the analysis is provided here, while the full details are reported in the Appendix. As represented in Table 3, Linguistic similarity shows the strongest association with transfer performance, followed by temporal overlap, while dataset size difference has only a weak effect. However, several cases suggest that these factors alone cannot fully account for the observed patterns. For instance, the transfer occurs from Indonesian to Persian (F1=79), or Arabic to Urdu (F1= 70), despite the dataset size difference or linguistic dissimilarity. Conversely, some pairs with moderate similarity or overlap still show limited transfer, such as Turkish to Arabic (F1=58). Further detail is represented in Table 8

Factor	Spearman ρ	p-value
Dataset size difference	-0.19	0.011
Temporal overlap	0.35	1.6×10^{-4}
Linguistic similarity	0.46	$< 10^{-9}$

Table 3: Spearman correlations between transfer performance and dataset properties.

6.2 Qualitative Analysis

To better understand when transferability occurs, we performed a cross-lingual qualitative error analysis by inspecting examples correctly classified only after applying transfer learning and comparing them with those correctly predicted by the in-language model. Further analysis of the cluster comprising Arabic–Indonesian–Urdu–Persian shows that Arabic and Urdu particularly support the detection of religious and political hate, especially sociopolitical hate prevalent in the Middle Eastern context. The Arabic dataset provides contextual cues that align well with Persian discourse in these domains, while Urdu contributes strongly to detecting political hate, often even more than Arabic. Indonesian contributes more to detecting role-based hate (e.g., targeted specific profession) and performs well on longer tweets that mix neutral and hateful sentiments. Table 6 presents Persian examples correctly predicted only after transfer learning, alongside those captured only by the Persian fine-tuned model.

We also observe strong transfer between Indonesian and Malay. Both datasets contain comparable expressions of political frustration, religious commentary, and identity-based accusations. Religious references are frequently used to judge or shame others, including accusations related to religious observance. Similar metaphorical and idiomatic expressions also appear in both languages, including insults referencing moral faults or entities, allowing Indonesian fine-tuned models to generalize well to Malay. Both languages also employ similar proverbs, insults, and metaphorical language, particularly the use of animals such as "dog" and "devil" as slurs. Social issues such as traffic caused by bazaars, public frustration over commodity scarcity, and emotional sharing on social media are likewise reflected in both contexts. However, the reverse transfer direction is weaker, likely due to the limited amount of hate data in the Malay dataset, which restricts its effectiveness as a source language. A similar pattern appears

for Persian, where limited hate training data reduces its effectiveness as a target language. Another cluster where transfer learning outperforms in-language learning includes Urdu, Hindi, and Bengali. Models fine-tuned on Urdu often perform better on Bengali and Hindi than the respective in-language models. Transfer from Urdu is particularly effective for highly polarized, overt hate expressions involving explicit slurs, religious antagonism, or politicized attacks. In contrast, in-language models more accurately classify tweets containing idiomatic expressions or contextually grounded criticism. For instance, socially embedded sarcasm in Bengali was frequently missed by the Urdu-fine-tuned model but correctly identified by the Bengali-fine-tuned counterpart. This difference appears related to dataset composition: the Urdu dataset contains a high density of explicit hate markers—religious slurs, profane language, and direct references to identifiable targets—which exhibit lexical consistency and limited idiomatic variation, enabling reliable pattern learning. By comparison, Bengali and Hindi datasets contain more ambiguous, sarcastic, or culturally embedded forms of expression; specifically, the Hindi dataset includes implicit hate, which is harder to generalize across languages. We also observe that although topics such as political or religious hate appear across multiple datasets, transfer strength does not depend on topical similarity alone. For example, although English and French contain religious or political hate, transfer learning from Arabic or Urdu yields larger gains when detecting such hate in Persian. A similar pattern appears when transferring from Arabic to English or French.

To broaden the analysis, we also inspected examples from language pairs outside the same empirical cluster, examining transfer when alignment is weaker. Our results show that even when languages do not belong to the same cluster (e.g., Spanish–Chinese, Persian–French), transfer learning can still detect certain types of hate speech, though less consistently. In particular, explicit, emotionally extreme, or lexically offensive expressions tend to transfer well across languages. For example, in our experiments, the Spanish→Chinese transfer achieved moderate performance in detecting overtly hateful Chinese instances such as *You're a brainless feminist bitch* or (*So many feminist bitches on Weibo!*). However, the Spanish-trained model struggled with more context-dependent expressions, such as: *Emperor Tang Minghuang's lifetime reputation was ruined at the hands of the Yang family - A historical reference used to delegitimize women*, or *I also saw someone called 'Pepsi256' who claims to be a feminist. Lately, she's been supporting the Sichuan Fine Arts Institute's pro-invasion-of-Taiwan military otakus and happily interacting with the nationalists (a.k.a. little pinks).*

In contrast, training the model on Korean resulted in significantly better performance not only on overt hate expressions but also on more subtle, context-dependent instances like those discussed above. In other cases, experiments with French and Persian show that a Persian fine-tuned model can still correctly identify highly explicit hate content in French. For example: "Vous pouvez crever sale race" (You can die, filthy race). This phrase combines an aggressive death wish with a racial slur, making it emotionally extreme and lexically explicit, which enables the Persian fine-tuned model to detect it despite belonging to a different cluster. These results suggest that even across dissimilar clusters, transfer learning succeeds when hate content is lexically explicit or emotionally extreme, but consistently struggles with expressions relying on background social or historical context.

Overall, the Global-Imb experiment demonstrates that integrating data from all languages in an imbalanced manner primarily benefits high-resource languages. While performance improves for high-resource settings, this approach weakens language-specific cues for low-resource languages, whose contextually grounded patterns may be overshadowed by dominant high-resource signals. The Cluster experiment mitigates this wash-out effect by grouping languages with similar characteristics, producing the most consistent gains, particularly for low-resource languages. Finally, the Global-Bal experiment provides a middle ground: balancing representation across languages reduces disadvantages for low-resource languages, although it slightly decreases gains for high-resource ones. (More samples are in Table 6 in the Appendix.)

7 Discussion

In addressing RQ1—whether cross-lingual training can match in-language hate speech detection—our results show that cross-lingual models can, in several cases, achieve performance comparable to or even exceeding in-language fine-tuning. Across the 14 languages studied, this effect is most evident within the observed clusters, where languages such as Persian, Indone-

sian, Urdu, Malay, Hindi, and Bengali benefit substantially from training on related languages. Regarding RQ2—whether multilingual training improves performance beyond single-source transfer—our findings show that only specific multilingual setups yield consistent gains. Globally imbalanced training primarily benefits high-resource languages while obscuring cues needed for low-resource ones. In contrast, cluster-based training provides the most reliable improvements across languages. RQ3 concerns potential reasons underlying successful or failed transfer. A qualitative analysis explores whether conceptual similarities in hate discourse could shed light on transferability. Our analysis shows that, within several clusters, the transferable expressions are not random: they often reflect recurring cultural cues and communicative norms shared across languages. Since the clearest cases where cross-lingual transfer outperforms in-language training arise in the Arabic–Persian–Indonesian setting, we examine this cluster more closely to understand the underlying factors. Our analysis of correctly transferred cases suggests that Arabic data helps models detect hate expressions related to religious and political discourse, particularly forms of sociopolitical hostility common in Middle Eastern contexts. These expressions frequently involve condemnatory statements about religious authority, ideological positions, or political actors, which appear in comparable forms across Arabic and Persian discourse. Such patterns provide contextual signals that align well with Persian hate expressions and therefore enrich the training signal for the target language. In contrast, Indonesian transfer contributes differently. The Indonesian dataset contains a higher proportion of tweets in which neutral narrative and hateful expressions appear within the same message, often combining background explanation with localized insults. Models trained on Indonesian data therefore show stronger performance on Persian tweets that follow a similar structure, where hateful remarks are embedded within longer narrative contexts rather than expressed as isolated insults. Indonesian-trained models are also particularly effective in detecting attacks directed toward individuals rather than groups, especially when hostility is framed through professional roles or references to public figures. These complementary patterns are reflected in the transfer experiments, where combining Arabic and Indonesian training data produces the strongest results. Table 7 in the Appendix shows some relevant instances. In addition, in the Indonesian–Malay pair, transferable cases similarly reflect shared idiomatic insults, moral accusations, and socially embedded metaphors. In the Urdu–Hindi–Bengali cluster, transfer succeeds when expressions rely on explicit slurs or polarized antagonism, echoing shared sociopolitical tensions, suggesting that some of the empirical clusters reflect overlapping cultural cues or communicative patterns (Zhou et al., 2023).

Our observations are consistent with prior work highlighting the importance of cultural context in hate speech detection. Lee et al. (2023) show that models trained within a single cultural context often perform poorly across cultures because they fail to capture culturally specific targets and discourse norms. Similarly, Lee et al. (2024) demonstrate that even within the same language, annotations vary across countries due to differences in cultural background and social norms. Our findings extend this by showing that such cultural and discourse differences also influence cross-lingual transfer, where languages with similar sociopolitical contexts enable stronger transfer, while culturally specific expressions reduce performance.

Conclusion

This study analyzes cross-lingual transfer learning for hate speech detection across fourteen languages, finding that while in-language fine-tuning remains the most robust baseline, cross-lingual models can achieve comparable or superior performance by leveraging clusters of languages with similar hate discourse patterns. Results indicate that multilingual training is most effective when utilizing these empirical clusters, as globally imbalanced integration provides limited support for low-resource languages. Crucially, successful transfer is not purely a function of dataset size, resource availability, or temporal overlap; rather, it reflects shared communicative norms and sociopolitical framing. While explicit or highly polarized hate expressions transfer effectively across distant languages, culturally embedded and context-dependent forms remain challenging. These findings suggest that the efficacy of cross-lingual transfer is tied to broader social dynamics, necessitating multilingual systems that account for the diverse ways online hate is expressed.

Limitations

This study has several limitations. First, the datasets used across the fourteen languages vary in source platforms, annotation practices, and topical distributions. Although we standardize dataset size where possible and map labels to a binary hate/non-hate scheme, differences in domain, explicitness, and annotation criteria may still influence transfer behavior. In particular, implicit hate appears only in a subset of languages, limiting our ability to evaluate how implicit expressions transfer across languages. Second, while our analysis considers dataset size and collection period, we do not control for all factors that may shape transfer performance, such as differences in sampling strategies or demographic characteristics of annotators. As such, the empirical clusters we identify reflect observed transfer patterns.

Third, the thematic observations drawn from our qualitative error analysis rely on manual inspection of representative examples. Although these analyses provide useful insights, they may not capture the full variability of hate discourse across languages. A more extensive human evaluation—including multiple annotators would offer a stronger foundation for interpreting the nature of transferable expressions. Finally, although we discuss broad social or discourse-related similarities as one possible contributor to transfer effectiveness, we do not measure such factors directly. We rely on transfer patterns themselves as empirical signals, and therefore cannot isolate the influence of linguistic, social, or dataset-driven properties. Future work can build on these insights by developing culturally informed models, expanding coverage of low-resource languages, and better capturing the social contexts underlying hateful discourse.

Ethics Statement

This study analyzes publicly available hate datasets and does not involve collecting new user data. All datasets were obtained from prior peer-reviewed work or shared tasks that follow established ethical guidelines. Because hate datasets may contain harmful or offensive language, examples shown in this paper are minimized and presented only when necessary for scientific transparency. No personally identifiable information is included in our datasets or model outputs. All experiments were conducted using anonymized text. Models trained in this work are not intended for deployment without further evaluation, fairness auditing, and context-specific calibration.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203, Torino, Italia. ELRA and ICCL.

Ife Adebara and Muhammad Abdul-Mageed. 2021. Improving similar language translation with transfer learning. In *Proceedings of the Sixth Conference on Machine Translation*, pages 273–278, Online. Association for Computational Linguistics.

Imatitikua D Aiyanyo, Hamman Samuel, and Heuiseok Lim. 2020. A systematic review of defensive and offensive cybersecurity with machine learning. *Applied Sciences*, 10(17):5811.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification usingSVM and Naive Bayes. *arXiv preprint arXiv:2204.07057*.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244, St. Julian's, Malta. Association for Computational Linguistics.

Samuel Bell, Eduardo Sánchez, David Dale, Pontus Stenetorp, Mikel Artetxe, and Marta R Costa-jussà. 2025. Translate, then detect: Leveraging machine translation for cross-lingual toxicity classification. In *Proceedings of the Tenth Conference on Machine Translation*, pages 253–268.

Imene Bensalem, Meryem Mout, and Paolo Rosso. 2023. Offensive language detection in Arabizi. In *Proceedings of ArabicNLP 2023*, pages 423–434.

Mohit Bhardwaj, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *CoRR*, abs/2011.03588.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9290–9304, Vienna, Austria. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^3hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731.

Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 109–117, Nagoya, Japan. Asian Federation of Natural Language Processing.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, pages 11–20.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of fake news detection via multitask transfer learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2596–2604, Marseille, France. European Language Resources Association.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: a Persian multi-label hate speech dataset with annotator rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17889–17897.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6048–6056.

Emna Fsih, Sameh Kchaou, Rahma Boujelbane, and Lamia Hadrich-Belguith. 2022. Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 431–435, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. 2025. Data-efficient hate speech detection via cross-lingual nearest neighbor retrieval with limited labeled data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29662–29680, Suzhou, China. Association for Computational Linguistics.

Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. Irony detection in Persian language: A transfer learning approach using emoji prediction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2839–2845, Marseille, France. European Language Resources Association.

Xiaoyu Guo and Susan Gauch. 2024. Using sarcasm to improve cyberbullying detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 52–59, Torino, Italia. ELRA and ICCL.

Yuting Guo and Abeed Sarker. 2025. Benchmarking open-source large language models on healthcare text classification tasks. *arXiv preprint arXiv:2503.15169*.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual abusive comment detection at scale for Indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Nhat Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Anh Tuan Luu. 2024. ToXCL: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472, Mexico City, Mexico. Association for Computational Linguistics.

Fatemah Husain, Hana Al-Ostad, and Halima Omar. 2022. A weak supervised transfer learning approach for sentiment analysis to the Kuwaiti dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 161–173, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aiqi Jiang and Arkaitz Zubiaga. 2024. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *arXiv preprint arXiv:2401.09244*.

Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9951–9952.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2024. The challenges of creating a parallel multilingual hate speech corpus: An exploration. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15842–15853, Torino, Italia. ELRA and ICCL.

Ankit Kumar, Richa Sharma, and Punam Bedi. 2024. Towards optimal NLP solutions: Analyzing GPT and LLaMA-2 models across model scale, dataset size, and task diversity. *Engineering, Technology & Applied Science Research*, 14(3):14219–14224.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 35–46.

Krishanu Maity, Shaubhik Bhattacharya, Sriparna Saha, and Manjeevan Seera. 2023. A deep learning framework for the detection of malay hate speech. *IEEE Access*, 11:79542–79552.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2024. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, pages 1–16.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Francois Meyer and Jan Buys. 2024. A systematic analysis of subwords and cross-lingual transfer in multilingual translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2194–2200, Mexico City, Mexico. Association for Computational Linguistics.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Wajid Hassan Moosa and Najiba. 2022. Multilingual hate speech dataset. `https://www.kaggle.com/datasets/wajidhassanmoosa/multilingual-hatespeech-dataset`. Open Data Commons Attribution License (ODC-By) v1.0.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 9:88364–88376.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Faiz Ghifari Haznitrama, Junho Myung, Alice Oh, Isabelle Augenstein, et al. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–13.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *Preprint*, arXiv:2012.09686.

Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. 2020. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204951–204962.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022a. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022b. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Charles Spearman. 1961. The proof and measurement of association between two things.

Manuel Tonneau. 2024. Spanish hate speech superset. Accessed: 2025-05-17.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. 2020. Detect all abuse! toward universal abusive language detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, and Nick Savage. 2020. Deep learning based fusion approach for hate speech detection. *IEEE Access*, 8:128923–128929.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix

Lang	Reference	Labels	Hate definition as stated in reference
BN	(Romim et al., 2020)	Hate, Not hate	A sentence that dehumanizes one or multiple persons or a community (e.g., comparing them to insects, objects, criminals, or targeting race, gender, disability).
UR	(Rizwan et al., 2020)	Hate-Offensive, Normal, Abusive, Offensive, Sexism, Religious Hate, Profane	Language that expresses hatred, abuse, or offensive content such as profanity, vulgar or insulting language targeting individuals or groups; includes gender-based hatred and religion-based hatred.
HI	(Bhardwaj et al., 2020)	Hostile, Non-hostile, Fake News, Hate, Offensive, Defamation, Normal	A post targeting a specific group of people based on ethnicity, religious beliefs, geographical belonging, race, etc., with malicious intentions of spreading hate or encouraging violence.
FA	(Delbari et al., 2024)	Normal, Violence, Hate, Vulgar	Communication targeting individuals or groups with the intention of attacking, dehumanizing, inciting violence, promoting discrimination, or fostering hostility based on characteristics such as race, ethnicity, religion, gender, sexual orientation, political beliefs, social class, disability, or appearance.
IN	(Ibrohim and Budi, 2019)	Hate, Not hate	Communication that disparages a person or a group based on characteristics such as race, ethnicity, gender, sexual orientation, nationality, religion, or other attributes.
MS	(Maity et al., 2023)	Hate, Not hate	Communication intended to attack the dignity of a group on the basis of characteristics such as race, gender, ethnicity, sexual orientation, nationality, religion, or other characteristics.
AR	(Mubarak et al., 2022)	Normal, Abusive, Hate	Any communication that disparages a person or a group based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.
TR	(Toraman et al., 2022)	Hate, Offensive, Normal	Tweets that target, incite violence against, threaten, or call for physical damage toward an individual or a group because of some identifying trait or characteristic.
KO	(Moon et al., 2020)	Hate, Offensive, None	Comments that explicitly express hatred against an individual or group based on attributes such as sex, gender, sexual orientation, gender identity, age, appearance, social status, religion, disability, ethnicity, or national origin, or that severely insult or attack them.
ES	(Tonneau, 2024)	Hate, Not hate	Speech that denigrates a person or multiple persons based on their membership to a group, usually defined by race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or views.
DE	(Tonneau et al., 2024)	Hate, Not hate, Offensive, Profanity	Describing dehumanization or deficiencies to groups or individuals because of being members of a group (e.g., race, political opinion, sexual orientation, gender, social status, health condition).
FR	(Tonneau et al., 2024)	Abusive, Hate, Offensive, Disrespectful, Fearful, Normal	Speech promotes the dehumanization of people who are typically from minority groups and may incite hate crimes.
EN	(Tonneau et al., 2024)	Abusive, Hate, Offensive, Disrespectful, Fearful, Normal	Speech promotes the dehumanization of people who are typically from minority groups and may incite hate crimes.

Table 4: Comparison of hate speech definitions and annotation schemes across languages. As shown, most datasets adopt a broadly consistent definition of hate speech as language attacking individuals or groups based on protected identity attributes.

method	model	En	Ar	Fa	BN	Hi	ZH	Kr	SP	Fr	De	IN	MS	Tu	UR
Zero-shot	GPT - 4o	82	64	55	53	67	67	60	75	77	75	65	56	71	45
	Llama 3	82	72	45	58	70	66	49	79	76	74	64	51	70	52
	Gemma 3	85	59	58	54	56	52	43	67	69	70	40	51	46	38
Fine-tuning	Llama 3	**96**	**91**	73	**72**	**80**	75	75	**90**	**94**	90	**82**	67	**90**	**85**
	Gemma 3	94	90	**75**	70	75	**82**	**76**	82	80	**94**	75	**71**	88	**85**

Table 5: F1 scores of different models for multilingual hate-speech detection.

Lang	TL From	Tweet	Explanation
FA	IN	رندی به محضر فقیهی رسیدو حرکات رقص را جداجدا انجام می داد و می پرسید آیا حرام است؟ فقیه میگفت نه. پس رند شروع به رقصیدن کرد.فقیه گفت تجزیه اش خوب بود ولی مرده شور ترکیش رو بردن حالا حکایت این عدالت خواراست بعضیشون عیبیی بچه های خوبین ولی مرده شور ترکیشون رو برده من برم به کارای خودم برسم خدافظ #عدالتخواران#انتخابات_مجلس A trickster went to a cleric and performed dance moves separately, asking if they were forbidden. The cleric said no. Then the trickster started dancing, and the cleric said, 'Breaking it down was fine, but damn the combination!' This is exactly the case with these so-called justice-seekers—some of them are actually good kids, but damn their combination! Anyway, I'll get back to my own business. Goodbye. #JusticeSeekers #ParliamentElection	Long tweet containing a mixture of neutral and toxic sentiments, with shared religious content.
FA	AR	وقتی یه جوان بدبخت میخاد استخدام بشه ریشه و بیخ مسائل خانوادگیش رو درمیارن اما زیباکلام با دختربازیش و شراب خواریش هنوز تو دانشگاه اسلامی!!! تدریس میکنه When a poor young man wants to get hired, they dig up every root of his family issues, but Zibakalam, with his womanizing and wine-drinking, is still teaching at an Islamic university!	Religious hypocrisy framed with moral outrage, a sentiment common to both Persian and Arabic contexts.
MS	IN	Oh dan kereta mmg suka menghimpit ke motor walaupun jelas2 ade moto kt situ smpi terpaksa mengelak n kereta tu buat bodoh je. Ramai ye perangai mcm ni. Oh and cars really like to squeeze in on motorcycles even when it's obvious there's a bike there, to the point the rider has to avoid them, and the car just acts clueless. There are many people with this kind of attitude.	Everyday frustration or complaint expressed in the form of Malay-language hate.
BN	UR	বেয়াদবের বাচ্চা বেয়াদব হ্যাকার গ্রুপ বর্ণনা হ্যাকার গ্রুপ বার করে দিব লুচ্চামি কি করস শরীরটা দেখছিস নিজের ছবিটা দেখ Impudent child's impudent hacker group description hacker group will be kicked out, what lewdness are you doing, looking at the body, look at your own picture.	Accusing a group of hackers of lewd behaviour and threatening expulsion.
HI	UR	हम तेरे लिये कुछ नही कर सकते ऐ दोस्त...!!! हमें अभी और मंदिर और मस्जिद बनानी है...!! We can't do anything for you my friend... We still have to build more temples and mosques...	Politicians prioritize religious projects over people's real needs.
ZH	KO	唐明皇一世英名就毁在杨家人手上了 Emperor Tang Minghuang's lifetime reputation was ruined at the hands of the Yang family - A historical reference used to delegitimize women	subtle, context-dependent instances.
ZH	KO	我还看到一个百事256的说自己是女权的，最近在力挺川美侵台军宅，和粉红互动得可欢了呢。 I also saw someone called 'Pepsi256' who claims to be a feminist. Lately, she's been enthusiastically supporting the Sichuan Fine Arts Institute's pro-invasion-of-Taiwan military otakus and happily interacting with the nationalists (a.k.a. little pinks	subtle, context-dependent instances

Table 6: Hate samples correctly predicted via transfer learning across languages.

tweet	Ar	Ind	Type
تف تو مملکتی که دیه الناز رکابی از مهدی ترابی کمتره 🤔 😳 !... Shame on a country where Elnaz Rekabi's blood money is worth less than Mehdi Torabi's...! 😳 🤔	1	0	Politics
رندی به محضر فقیهی رسیدو حرکات رقص را جداجدا انجام می داد و می پرسید آیا حرام است؟ فقیه میگفت نه. پس رند شروع به رقصیدن کرد.فقیه گفت تجزیه اش خوب بود ولی مرده شور ترکیش رو بردن حالا حکایت این عدالت خواراست بعضیشون عیبیی بچه های خوبین ولی مرده شور ترکیشون رو برده من برم به کارای خودم برسم خدافظ #عدالتخواران#انتخابات_مجلس A trickster went to a cleric and performed dance moves separately, asking if they were forbidden. The cleric said no. Then the trickster started dancing, and the cleric said, 'Breaking it down was fine, but damn the combination!' This is exactly the case with these so-called justice-seekers—some of them are actually good kids, but damn their combination! Anyway, I'll get back to my own business. Goodbye. #JusticeSeekers #ParliamentElections	0	1	Mix of Neutral and Hate (bold)
رفتم دماغمو عمل کنم دکتره یه نگاهی کرد گفت داداش شما صبر کن یکم دیگه علم پیشرفت کنه کلا سرتو عوض کن 😂 حالا شده حکایت ما مملکتی که رقابت خوشگلترین مسئولش بین جهانگیری و احمد خاتمی و احمد نژاده باید رید توش البته از نظر عملکرد شان هم باید رید توش #مهسا_امینی #ترور_بیولوژیکی I went to get my nose done, and the doctor took one look at me and said, 'Bro, just wait a little longer until science advances enough to replace your whole head. 😂' That's exactly our situation—when the competition for the most handsome official in the country is between Jahangiri, Ahmad Khatami, and Ahmadinejad, you know it's doomed. And in terms of their performance and dignity, well, it's even worse. #Mahsa_Amini #Biological_Assassination	0	1	Mix of Neutral and Hate (bold)
از ماست که بر ماست تا به این دین و باورهای بیابان گرد ملخ خوار باور داریم همین آش و همین کاسه It is up to us to believe in this religion and beliefs of the desert, the locust-eating locusts, the same soup and the same bowl.	0	0	idiomatic

Table 7: Case study of Persian hate speech instances correctly predicted through transfer learning from cluster languages, or misclassified in all experimental settings.

8 Quantitative Analysis of Transfer Factors

To better understand the factors influencing cross-lingual transfer performance, we conducted a quantitative analysis examining the relationship between several dataset properties and transfer performance. For each pair of languages, the transfer performance was defined as the off-diagonal F1 score from the 14×14 transfer matrix presented in Figure 1. This results in 182 source–target language pairs. For each pair we computed three measurable factors: dataset size difference, temporal overlap, and linguistic similarity. The resulting dataset consists of one row per source–target pair containing the transfer F1 score and the corresponding factor values 8

8.1 Dataset Size Difference

Dataset size difference measures the disparity between the number of hate instances used for training in the source and target datasets. For each language pair we computed the absolute log-ratio of the hate training sizes:

$$size_diff = |\log(N_{source}) - \log(N_{target})|$$

where N_{source} and N_{target} denote the number of hate training instances reported in Table 1. This produced 182 values corresponding to the off-diagonal entries of the transfer matrix.

8.2 Temporal Overlap

Temporal overlap measures how much the data collection periods of two datasets intersect. For languages where collection periods were available, we computed the overlap between the time ranges using intersection over union:

$$temporal_overlap = \frac{|T_{source} \cap T_{target}|}{|T_{source} \cup T_{target}|}$$

where T_{source} and T_{target} represent the years covered by each dataset. Language pairs with missing collection dates were excluded from this calculation.

8.3 Linguistic Similarity

To measure linguistic proximity between languages, we used typological feature vectors derived from the URIEL/lang2vec database (Bjerva et al., 2019). Each language is represented by a vector of typological and phylogenetic features. Pairwise similarity between languages was computed using cosine similarity between these vectors:

$$ling_similarity = cosine(v_{source}, v_{target})$$

where v_{source} and v_{target} denote the feature vectors for the source and target languages.

8.4 Correlation Analysis

To assess the relationship between these factors and transfer performance, we computed Spearman rank correlations (Spearman, 1961) between each factor and the transfer F1 scores across all language pairs. The correlation results reported in Table 3 indicate that linguistic similarity has the strongest association with transfer performance, followed by temporal overlap, while dataset size difference shows only a weak relationship. This trend is also visible in several high-performing language pairs with strong typological proximity. For example, transfer from Urdu to Hindi achieves the highest score in the matrix with very high linguistic similarity , while Urdu to Bengali and English to German show similarly strong performance. High transfer is also observed among closely related Romance languages, such as Spanish to French . Likewise, Indonesian transfers effectively to Malay , reflecting their near-identical linguistic structure. However, not all strong transfers align perfectly with linguistic proximity. For instance, Turkish transfers relatively well to English and German. Similarly, Arabic transfers effectively to Urdu, Indonesian to Persian despite belonging to different language families. These cases suggest that while typological similarity contributes to transfer effectiveness, it does not fully account for the patterns observed in the transfer matrix, and that broader discourse or sociocultural similarities may also facilitate cross-lingual transfer.

source	target	F1	size_diff	ling_similarity	temporal_overlap
TR	FA	66	1.5	0.7	0.29
TR	MS	39	0.97	0.46	0.14
TR	ES	57	0.05	0.47	1
TR	FR	62	0.19	0.48	1
FA	TR	48	1.5	0.7	0.29
FA	AR	63	1.5	0.78	0
FA	BN	43	1.5	0.7	0
FA	IN	67	1.5	0.58	0
FA	KO	40	0.42	0.4	0
AR	FA	80	1.5	0.78	0
AR	MS	53	0.97	0.56	0
AR	ES	55	0.05	0.58	0.29
AR	KO	44	1.09	0.38	0
BN	MS	43	0.97	0.61	0
BN	ES	61	0.05	0.61	0.43
IN	MS	78	0.97	0.92	0
IN	ES	57	0.05	0.63	0.29
IN	FR	57	0.19	0.64	0.29
IN	KO	62	1.09	0.42	0
MS	IN	63	0.97	0.92	0
MS	EN	50	0.97	0.66	0.14
MS	ES	47	1.02	0.63	0.14
MS	FR	55	0.79	0.64	0.14
EN	FA	56	1.5	0.63	0.29
EN	FR	78	0.19	0.88	1
EN	KO	73	1.09	0.45	0.29
ES	TR	65	0.05	0.47	1
ES	FA	50	1.55	0.61	0.29
ES	BN	54	0.05	0.61	0.43
ES	IN	47	0.05	0.63	0.29
FR	TR	54	0.19	0.48	1
FR	BN	58	0.19	0.62	0.43
FR	IN	40	0.19	0.64	0.29
FR	MS	49	0.79	0.64	0.14
DE	FA	45	1.5	0.64	0.29
DE	ES	68	0.05	0.88	1
KO	TR	51	1.09	0.35	0.29
KO	IN	44	1.09	0.42	0
KO	MS	58	0.11	0.43	0
KO	ES	43	1.14	0.44	0.29

Table 8: Cross-lingual transfer cases with corresponding dataset properties for a representative sample.