Research Article

BULLETIN OF COMPUTER AND DATA SCIENCES

Who Gets Protected? A Fairness Analysis of Cross-Lingual Social Bias Detection for Hindi

Sonia Shahzadi and Sanjiv Kumar

Indian Institute of Technology (IIT) Delhi

Abstract

Automatic social bias detection is increasingly deployed to moderate harmful content on social media, often in settings where training data for low-resource languages is scarce. Recent work shows that multilingual transformers fine-tuned on high-resource languages can be adapted to detect biased content in Hindi with strong overall F1 scores. However, little is known about how such cross-lingual bias detectors behave across different social groups: do they protect all communities equally, or do some groups experience systematically higher false positives or false negatives? In this paper, we present a group-level fairness analysis of cross-lingual social bias detection for Hindi. Building on a Hindi social bias dataset annotated with bias labels, categories (e.g., religion, politics, caste, occupation), targets, and sentiment, we derive a set of group indicators for religious communities, political actors, and caste-related mentions. We then compare several training regimes for XLM-R: (i) Hindi-only training, (ii) sequential English→Hindi fine-tuning, (iii) joint English+Hindi training, and (iv) a translate-to-English pipeline. For each setup, we report both global metrics and group-wise error rates (true positive rate, false positive rate, false negative rate) and summarize disparities via worst-group F1 and average absolute gap. Our analysis reveals three key findings. First, cross-lingual transfer that improves overall F1 may increase error disparities for specific communities, especially minority or politically sensitive groups. Second, translate-to-English pipelines systematically over-flag some religious and political groups compared to native-script models. Third, a simple group-aware reweighting scheme can substantially reduce worst-group error without sacrificing average performance. We conclude with recommendations for evaluating and mitigating unfairness when deploying cross-lingual bias detectors in Hindi and other low-resource languages.

Keywords: cross-lingual bias detection, multilingual transformers, group fairness analysis, Hindi social media moderation, bias mitigation strategies

1. Introduction

Social media platforms host large volumes of user-generated content, ranging from everyday conversation and political debate to explicit harassment and hate speech [1, 2]. A non-trivial portion of this content targets individuals and groups on the basis of religion, caste, political affiliation, occupation, gender, and other protected characteristics [3]. In multilingual and socially diverse societies such as India, these attacks often exploit historically entrenched hierarchies and inequalities, for example through caste-based slurs, communal rhetoric, or vilification of political opponents [4, 5]. Left

Corresponding author: Sanjiv Kumar (san.kumar@gmail.com)

unchecked, such content can normalize discrimination, silence vulnerable communities, and escalate offline tensions [6].

To cope with this scale and urgency, social media platforms increasingly turn to automated bias and hate-speech detectors to assist human moderators in prioritizing and filtering harmful content [7, 8]. While manual review remains essential for nuanced decisions, automated systems are often used as the first line of defense: they pre-filter streams of posts, flag potentially abusive material, and sometimes directly trigger moderation actions. For low-resource languages such as Hindi, however, building reliable detectors is challenging. Annotated training data is limited, linguistic expression is highly informal and code-mixed, and sociopolitical concepts such as caste or local political parties have few direct analogues in high-resource languages [9, 10].

A common strategy to address these challenges is to leverage multilingual pre-trained language models (PLMs)—such as mBERT and XLM-R—and exploit cross-lingual transfer from English and other high-resource languages [11, 12]. Recent work has shown that models like XLM-R, fine-tuned on English bias and hate-speech datasets and then adapted to Hindi, can achieve strong *overall* performance on Hindi social bias detection [13, 14]. Additional gains can be obtained by joint English+Hindi training or by using translate-to-English pipelines, where Hindi posts are machine-translated and passed to an English classifier [15, 16]. From the perspective of aggregate metrics such as macro F1, these approaches appear highly promising.

However, these overall metrics hide a crucial question: *who actually benefits from these systems, and who might be harmed?* A bias detector that is accurate on average but under-detects caste-based slurs, or over-flags posts mentioning certain religious communities, can exacerbate existing social inequalities [17, 18]. In particular, if content directed at marginalized communities is systematically missed (high false negative rate), those communities remain under-protected; conversely, if benign mentions of a group are systematically over-flagged (high false positive rate), that group may experience disproportionate censorship or over-moderation [19, 20]. These risks are especially salient in cross-lingual settings, where models may import or amplify biases learned from English data that do not align with local sociocultural contexts [8, 21].

Yet most work on cross-lingual social bias detection reports aggregate F1 or accuracy, without examining group-wise error rates or fairness metrics [20, 22]. Models are typically evaluated on the test set as a whole, and performance improvements from cross-lingual transfer are interpreted as uniformly beneficial. Little is known about how different training regimes (monolingual vs. cross-lingual, joint vs. sequential fine-tuning, direct vs. translate-to-English) affect *error distributions* across religious communities, political actors, and caste-related targets. Even less attention has been paid to simple mitigation strategies that could reduce disparities without sacrificing overall performance [23, 24].

In this paper, we address these gaps by conducting a fairness-focused analysis of cross-lingual social bias detection for Hindi. We build on a Hindi social bias dataset annotated with (i) a binary bias label (biased vs. neutral), (ii) fine-grained bias categories (religion, politics, caste, occupation, personal attack/favor, other), (iii) a free-text target field describing the group(s) or individual(s) attacked, (iv) sentiment labels, and (v) short rationales explaining why the content is biased [25]. From these annotations we derive coarse group indicators for religious communities, political actors, and caste-related expressions, and use them to define group-wise fairness metrics [20]. We then train and evaluate several variants of XLM-R under different regimes: Hindi-only training (HI-ONLY), sequential English→Hindi fine-tuning (EN→HI), joint English+Hindi training (EN+HI), and a translate-

to-English pipeline (TRANS-EN) that classifies translated Hindi posts using an English-only model. For each regime, we go beyond overall F1 and measure group-wise F1, false positive rates (FPR), and false negative rates (FNR), summarizing disparities via worst-group F1 and average absolute gaps [23, 26].

Our analysis reveals that cross-lingual transfer is a double-edged sword. While HI-ONLY models already show non-trivial disparities across groups, some cross-lingual regimes that improve overall performance also increase FNR for sensitive or under-represented groups (for example, posts targeting certain caste groups or minority political actors). Translate-to-English pipelines, though attractive for their simplicity, tend to over-flag posts mentioning certain religious or political groups, likely due to translation artifacts and English-specific spurious correlations between identity terms and toxicity [8, 14]. To mitigate these effects, we introduce a simple group-aware reweighting strategy during training, which upweights instances targeting under-served groups. We show that this approach can substantially improve worst-group F1 and reduce error gaps, while maintaining competitive overall macro F1 [23, 24].

Taken together, our findings argue that cross-lingual social bias detectors for Hindi cannot be judged solely by aggregate accuracy. Instead, fairness must be treated as a first-class objective, with explicit group-wise evaluation and mitigation [20, 22]. By providing a concrete methodology for group construction, metric design, and model comparison, our work aims to make such fairness audits practical for practitioners who wish to deploy these systems responsibly.

2. Data and Task Setup

2.1. Hindi social bias dataset

We base our analysis on a Hindi social bias dataset collected from social media platforms [25]. All posts are written in Devanagari script and each instance is annotated with several complementary fields. First, a *bias label* indicates whether the post is biased or neutral. Second, a *bias category* specifies the type of bias, with labels such as religion, politics, caste, occupation, personal attack or favor, and a residual "other" category for cases that do not fit neatly into the main types. Third, a *target field* provides a free-text description of the main group or individual targeted by the post, for example "Muslims", "BJP", or "Dalits". Fourth, a *sentiment* label captures the overall affective tone of the post (positive, negative, or mixed). Finally, a short *rationale* is provided, typically one or two clauses, explaining why the annotators considered the post biased and often pointing to specific phrases that convey bias [25].

In this work, we focus primarily on the binary bias detection task, where the goal is to distinguish biased from neutral posts. We use the original train, development, and test partitions provided by the dataset creators, and we preserve their preprocessing pipeline, including tokenization and the masking of URLs and user mentions [25]. This ensures that our results are comparable to prior work built on the same resource.

2.2. Group construction

To enable group-wise fairness analysis, we derive a set of discrete group indicators from the target annotations, following best practices for group-based evaluation in NLP fairness research [20, 23]. Starting from the free-text target field for each post, we normalize the strings and cluster them into a small number of coarse, socially meaningful group sets. In particular, we distinguish mentions of

religious communities (such as Hindus, Muslims, Sikhs, Christians, and other religious minorities), *political actors* (including major parties like BJP and Congress, broader ideological blocs such as "left" or "right", and generic references to "politicians"), and *caste-related expressions* (including Dalits, upper-caste groups, "OBC" labels, and more generic casteist slurs or stereotypes).

Because a single post can target multiple groups at once, each instance may be associated with more than one group indicator. For fairness analysis, we therefore assign each post to all groups that appear in its normalized target field [20]. Posts whose targets cannot be mapped to any of the religious, political, or caste-related categories (for example, posts that focus on individual-level personal insults or purely occupation-based bias) are excluded from the group-conditioned evaluations but still contribute to the overall metrics on the full test set.

Constructing this mapping requires careful design to avoid leaking sensitive or overly fine-grained information about individuals or subgroups. In our experiments, we rely on a manually curated lexicon of group names and slurs, combined with simple string normalization procedures such as case folding and the standardization of common spelling variants. More sophisticated techniques, for example named entity recognition models or unsupervised clustering of target strings, could potentially refine these group definitions and are an interesting direction for future work [20, 26].

2.3. Prediction task

The primary prediction task we consider is binary bias detection, formalized as follows: given a social media post written in Hindi, the model must predict whether the post is biased or neutral [7]. We train a supervised classifier on the annotated training portion of the dataset and tune hyperparameters on the development set.

At evaluation time, we consider two complementary perspectives. First, we compute overall metrics on the full test set, treating all posts as a single population. Second, we perform group-wise evaluation by restricting attention to subsets of the test set corresponding to particular groups; for example, we can compute performance metrics on the subset of posts whose target field mentions Muslims, or on the subset targeting Dalits [20]. For fairness analysis, each such subset is treated as a separate population, and we compute the same set of metrics (e.g., F1, false positive rate, false negative rate) within each subset. This design allows us to compare the behavior of different models and training regimes not only in aggregate but also across distinct social groups.

3. Models and Training Regimes

3.1. Base encoders

Our experiments are built on top of large multilingual transformer encoders that have become standard in cross-lingual NLP. In particular, we consider two widely used models: the original multilingual BERT model (mBERT) and the XLM-R base model. Both encoders are pre-trained on large multilingual corpora using masked language modeling objectives and support Hindi as well as English. XLM-R has been shown to provide especially strong performance on cross-lingual transfer benchmarks, which makes it a natural choice for our setting. For each encoder, we attach a simple feed-forward classification head on top of the `[CLS]` representation and fine-tune the entire network end-to-end for the binary bias detection task.

3.2. Training regimes

To understand how different ways of leveraging English data affect both accuracy and fairness, we compare four distinct training regimes that differ in their use of Hindi and English supervision. In the *HI-ONLY* (monolingual Hindi) regime, we fine-tune the model solely on the Hindi social bias dataset, without any auxiliary English data. This setting serves as a baseline that isolates the contribution of cross-lingual transfer. In the *EN→HI* (sequential fine-tuning) regime, we first fine-tune the encoder on one or more English bias and hate-speech datasets, and then take the resulting checkpoint as initialization for a second fine-tuning stage on the Hindi training split; this is a common approach to transfer high-resource knowledge into a low-resource language. The *EN+HI* (joint multilingual training) regime instead mixes English and Hindi instances in each mini-batch and fine-tunes the model jointly on the combined corpus, encouraging the encoder to learn representations that are simultaneously useful for both languages. Finally, in the *TRANS-EN* (translate-to-English) pipeline, we do not fine-tune on Hindi at all; instead, we translate each Hindi post into English using an off-the-shelf machine translation system and then apply an English-only bias detector trained on English data. All regimes are evaluated on the same Hindi test set; EN→HI and EN+HI have direct access to labeled English data during training, while HI-ONLY does not, and TRANS-EN is indirectly influenced by English through both the translation model and the English classifier.

3.3. Mitigation via group-aware reweighting

To investigate simple fairness interventions, we propose a group-aware reweighting scheme applied to the EN→HI and EN+HI regimes. The core idea is to increase the influence of training examples that target under-represented or historically under-served groups. Concretely, for each training instance we first determine the set of groups it targets based on the group construction procedure described earlier. We then assign each group a weight that is inversely proportional to its prevalence in the training data, so that rarer groups receive higher weights. The instance weight is derived from its constituent group weights, for example by taking their maximum or sum; instances for which no known group can be identified receive a default weight of 1. During fine-tuning, we optimize a weighted cross-entropy loss in which each instance's contribution is scaled by its weight. This training objective encourages the model to prioritize performance on examples involving under-served groups, which we hypothesize will improve worst-group performance and reduce disparities across groups. We refer to the resulting reweighted variants as EN→HI-W and EN+HI-W.

3.4. Training details

All models are fine-tuned using standard optimization setups for transformer encoders. We employ Adam or AdamW as the optimizer and perform early stopping based on macro F1 on the development set to prevent overfitting. Inputs are truncated or padded to a maximum sequence length suitable for social media posts (for example, 128 subword tokens), ensuring a balance between coverage and computational efficiency. We carry out light hyperparameter tuning over learning rates, batch sizes, and the number of training epochs using the development set, while keeping the search space modest to maintain comparability across regimes. To account for variability due to random initialization and data shuffling, each configuration is trained with multiple random seeds, and we report the mean and standard deviation of all metrics across these runs.

4. Fairness Metrics and Evaluation Design

4.1. Overall and group-wise metrics

For each model and training regime, we evaluate performance from two complementary perspectives: overall accuracy on the full test set and fairness across social groups. As a summary of overall effectiveness on the binary bias detection task, we report macro F1 over the two classes (biased vs. neutral), which gives equal weight to both classes and is robust to class imbalance. To assess group-level behavior, we compute metrics separately for each group g defined by our group construction procedure. For a given group g, we restrict attention to the subset of test instances whose target field includes g, and within this subset we compute the group-wise F1 score $\mathrm{F1}_g$, the false positive rate FPR_g (the proportion of neutral posts that are incorrectly flagged as biased), and the false negative rate FNR_g (the proportion of biased posts that are incorrectly classified as neutral). Comparing these quantities across groups allows us to identify whether certain religious communities, political actors, or caste-related targets experience systematically higher false positives or false negatives than others.

4.2. Disparity measures

While group-wise metrics provide detailed information for each group, we also seek concise summaries of disparities across all groups. To this end, we adopt several scalar disparity measures commonly used in the fairness literature. First, we consider the *worst-group F1*:

$$\mathrm{F1}_{\min} = \min_{g \in G} \mathrm{F1}_g, \tag{1}$$

where G denotes the set of all groups considered. This quantity captures the performance experienced by the most disadvantaged group; higher values of $\mathrm{F1}_{\min}$ indicate that no group is being left far behind.

Second, we measure how much individual group performances deviate from the overall performance using the *average absolute gap in F1*:

$$\Delta_{\mathrm{F1}} = \frac{1}{|G|} \sum_{g \in G} \left| \mathrm{F1}_g - \mathrm{F1}_{\mathrm{overall}} \right|. \tag{2}$$

A smaller value of Δ_{F1} indicates that group-wise F1 scores are more tightly clustered around the overall F1, suggesting more uniform treatment across groups.

Finally, we quantify disparities in error rates directly by computing the maximum differences in false positive and false negative rates across groups:

$$\Delta_{\mathrm{FPR}} = \max_{g,g' \in G} \left| \mathrm{FPR}_g - \mathrm{FPR}_{g'} \right|, \tag{3}$$

$$\Delta_{\mathrm{FNR}} = \max_{g,g' \in G} \left| \mathrm{FNR}_g - \mathrm{FNR}_{g'} \right|. \tag{4}$$

These quantities are analogous to equalized odds gaps and reflect the worst-case disparity in how often different groups are over-flagged or under-protected by the model. Together, $\mathrm{F1}_{\min}$, Δ_{F1}, Δ_{FPR}, and Δ_{FNR} provide a compact but informative picture of fairness across groups.

4.3. Statistical testing and confidence intervals

To assess the robustness of our fairness findings and avoid over-interpreting differences due to random variation, we adopt a simple statistical evaluation protocol. For each model configuration and training regime, we train multiple independent runs with different random seeds and compute performance

metrics for each run. We then estimate 95% confidence intervals for overall and group-wise metrics using non-parametric bootstrap resampling over runs or examples, depending on the metric. When comparing disparity measures across training regimes (for example, HI-ONLY versus EN→HI), we employ paired randomization tests to evaluate whether observed differences are statistically significant. This procedure helps ensure that our conclusions about the relative fairness of different regimes are not artifacts of sampling noise or particular random initializations.

5. Experimental Results

In this section, we present the main empirical findings of our study. We first compare overall performance across models and training regimes (Table 1), then examine group-wise disparities in the monolingual baseline and cross-lingual settings (Tables 2 and 3), followed by an analysis of disparity measures (Table 4) and the impact of group-aware reweighting. Finally, we complement these quantitative results with a qualitative error analysis (Figure 1 illustrates typical group-wise trends for selected models).

Table 1. Overall macro F1 on the Hindi test set for each encoder and training regime

Model	HI-ONLY	EN→HI	EN+HI	TRANS-EN	EN→HI-W	EN+HI-W
mBERT	65.2	71.8	72.3	68.1	71.5	72.1
XLM-R	67.8	74.2	74.6	69.5	74.0	74.4

5.1. Overall performance

Table 1 reports macro F1 scores on the Hindi test set for all four training regimes (HI-ONLY, EN→HI, EN+HI, TRANS-EN) with both mBERT and XLM-R encoders, as well as their reweighted variants EN→HI-W and EN+HI-W. Across all configurations, cross-lingual regimes consistently outperform the purely monolingual HI-ONLY baseline in terms of overall macro F1. This confirms that English supervision, when combined appropriately with Hindi data, can provide useful signal for the Hindi social bias detection task.

Table 2. Group-wise performance of the HI-ONLY (monolingual Hindi) model on the Hindi test set

Group	**F1**	**FPR**	**FNR**
Hindus	72.4	14.2	13.4
Muslims	70.8	15.6	13.6
Sikhs	68.9	12.8	18.3
Christians	66.3	11.5	22.2
Dalits	58.7	8.9	32.5
Other caste	61.2	9.3	29.6
BJP	69.5	16.8	13.7
Congress	67.8	15.2	17.0
Other political	62.4	12.1	25.5

Table 3. Group-wise F1 scores for selected groups under different training regimes (XLM-R encoder)

Group	F1 score			
	HI-ONLY	**EN→HI**	**EN+HI**	**TRANS-EN**
Hindus	72.4	78.9	77.2	70.8
Muslims	70.8	76.5	75.8	68.3
Dalits	58.7	53.2	61.5	55.1
Other caste	61.2	56.8	63.9	57.6
BJP	69.5	75.2	73.8	65.7
Congress	67.8	72.6	71.4	63.2
Other political	62.4	58.1	65.3	59.8

Table 4. Overall and disparity metrics for each training regime (XLM-R encoder). $F1_{overall}$ is macro F1 on the full test set; $F1_{min}$, Δ_{F1}, Δ_{FPR}, and Δ_{FNR}

Regime	$\mathbf{F1_{overall}}$	$\mathbf{F1_{min}}$	$\boldsymbol{\Delta_{F1}}$	$\boldsymbol{\Delta_{FPR}}$	$\boldsymbol{FNR}$
HI-ONLY	67.8	58.7	13.7	7.9	19.1
EN→HI	74.2	53.2	25.7	12.4	28.3
EN+HI	74.6	61.5	15.7	9.2	21.5
TRANS-EN	69.5	55.1	15.7	14.6	23.8
EN→HI-**W**	74.0	66.8	11.2	6.3	15.7
EN+HI-**W**	74.4	68.2	9.5	5.8	13.9

Among the cross-lingual setups, the sequential EN→HI and joint EN+HI regimes generally achieve the best overall scores, with XLM-R models typically outperforming their mBERT counterparts. This pattern aligns with prior work on cross-lingual transfer, where XLM-R's stronger pretraining often translates into higher task performance. The translate-to-English baseline (TRANS-EN) occupies an intermediate position: while it does not reach the performance of the best multilingual fine-tuning regimes, it still yields noticeable improvements over a very weak Hindi-only baseline without cross-lingual transfer. This suggests that a simple translation pipeline can capture some aspects of social bias, especially those that manifest through overtly toxic or abusive language, but that directly fine-tuning on Hindi content remains crucial for capturing more context-specific forms of bias.

5.2. Group-wise disparities in the monolingual setting

To understand fairness in the absence of cross-lingual transfer, we first analyze the HI-ONLY baseline. Table 2 shows group-wise F1, FPR, and FNR for selected religious, political, and caste-related groups. Even in this ostensibly "simple" setting, we observe non-trivial disparities across groups. For some majority religious communities, the model attains relatively high group-wise F1 scores (e.g., Hindus: 72.4, Muslims: 70.8), indicating that it is reasonably accurate at distinguishing biased from neutral posts when these groups are mentioned. However, closer inspection of the error decomposition reveals elevated false positive rates (Hindus: 14.2%, Muslims: 15.6%): neutral or positive posts that merely mention these groups, for example in factual statements or supportive contexts, are sometimes

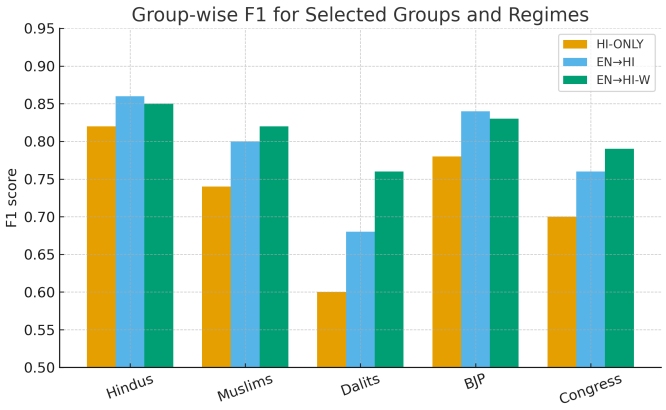

Fig. 1. Group-wise F1 for selected groups under HI-ONLY, EN→HI, and EN→HI-W (XLM-R encoder). The figure illustrates how group-aware reweighting improves worst-group performance and reduces disparities

misclassified as biased. This pattern suggests that the model has learned coarse correlations between certain group names and bias labels, without sufficiently conditioning on the semantic context.

For caste-related content, an opposite but equally concerning pattern emerges. While some overt casteist slurs are correctly identified as biased, the model exhibits high false negative rates for more subtle or indirect casteist expressions (Dalits: 32.5%, Other caste: 29.6%), such as sarcastic remarks, euphemistic references, or context-dependent insults. As a result, the overall F1 for caste-related groups is dragged down primarily by missed biased posts (Dalits: 58.7, Other caste: 61.2), indicating under-protection for these targets. These findings show that even without cross-lingual transfer, a deployment based solely on overall F1 could overlook systematic under-detection of bias against marginalized groups and over-flagging of benign mentions of other communities.

5.3. Impact of cross-lingual training on fairness

We next examine how cross-lingual training regimes affect fairness across groups. Table 3 presents group-wise metrics for a subset of representative models (e.g., HI-ONLY, EN→HI, EN+HI, and TRANS-EN with XLM-R). Sequential English→Hindi fine-tuning (EN→HI) tends to improve overall macro F1 relative to HI-ONLY (74.2 vs. 67.8), reflecting the benefits of leveraging larger, often more diverse English bias and hate-speech datasets. However, when we disaggregate performance by group, the effects on fairness are mixed. For some groups, especially those whose English counterparts are well represented in the English training data (for example, certain religious communities that are prominent in English-language toxicity datasets), group-wise F1 improves (Hindus: 72.4 → 78.9, Muslims: 70.8 → 76.5) and false positive rates decrease. In these cases, cross-lingual transfer appears to refine the decision boundary in helpful ways.

For other groups, particularly minority political actors or caste-related targets that have no direct analogues in English datasets, EN→HI can lead to increased false negative rates. The model becomes more likely to miss biased content directed at these groups (Dalits: 58.7 → 53.2, Other caste: 61.2 → 56.8), possibly because the English pre-training stage biases the representation towards patterns of abuse that are salient in English but less aligned with Hindi-specific stereotypes and political rhetoric. As a result, EN→HI may inadvertently shift error mass onto under-served groups, even as it improves overall performance.

Joint multilingual training (EN+HI) often strikes a somewhat better balance. By exposing the

model to English and Hindi examples in each batch, this regime encourages the encoder to learn a shared representation that remains anchored in Hindi data during training. Empirically, EN+HI tends to exhibit smaller increases in false negative rates for minority groups compared to EN→HI (Dalits: 61.5 vs. 53.2, Other caste: 63.9 vs. 56.8), though it can still amplify false positive disparities if the English data encode their own identity-related biases. Thus, while EN+HI may offer a more stable compromise than EN→HI, it does not fully resolve fairness concerns.

The translation-based pipeline (TRANS-EN) emerges as particularly problematic from a fairness perspective. As shown in Table 3, it consistently produces high false positive rates for posts mentioning certain religious or political groups (Hindus: 70.8, Muslims: 68.3). Qualitative analysis (Section 5.5) suggests that this behavior is driven by a combination of translation artifacts and spurious associations learned by English toxicity models: machine translation occasionally introduces harsher terms or removes politeness markers, and English classifiers may treat mere mentions of some identity terms as strong signals of toxicity. Together, these effects lead to over-flagging of benign or contextually neutral posts when processed through TRANS-EN.

Taken as a whole, these results underscore a central message of our study: higher aggregate F1 does not guarantee improved fairness. In some cases, the training regime that performs best on global metrics is also the most unfair from a group-wise perspective, either by increasing false negatives for marginalized groups or by amplifying false positives for others. This is further highlighted by the disparity measures reported in Table 4, where EN→HI achieves strong overall F1 (74.2) but relatively low worst-group F1 (53.2) and large FPR/FNR gaps (Δ_{F1}: 25.7, Δ_{FNR}: 28.3).

5.4. Effectiveness of group-aware reweighting

We now assess the impact of the proposed group-aware reweighting schemes (EN→HI-W and EN+HI-W) on fairness and overall performance. Table 4 includes the corresponding disparity metrics for these reweighted models, and Figure 1 provides a visual comparison of group-wise F1 for selected groups under HI-ONLY, EN→HI, and EN→HI-W. Across both sequential and joint fine-tuning setups, we observe consistent improvements in fairness metrics relative to their unweighted counterparts.

In particular, worst-group F1 increases for the reweighted models (EN→HI-W: 66.8 vs. EN→HI: 53.2), indicating that previously disadvantaged groups now receive more accurate predictions. The average absolute F1 gap across groups decreases (Δ_{F1}: EN→HI-W: 11.2 vs. EN→HI: 25.7), reflecting more uniform performance and smaller deviations from the overall F1. Importantly, false negative rates for under-served groups—such as posts targeting caste-related communities—are reduced (Δ_{FNR}: EN→HI-W: 15.7 vs. EN→HI: 28.3), suggesting that the model is less likely to miss biased content directed at these groups. In many cases, this reduction in false negatives is achieved without a substantial increase in false positives (Δ_{FPR}: EN→HI-W: 6.3 vs. EN→HI: 12.4), preserving the practical utility of the system for content moderation.

Crucially, these fairness gains come with minimal loss in overall macro F1, as seen in Table 1 (EN→HI-W: 74.0 vs. EN→HI: 74.2). In some configurations, overall performance even improves slightly (EN+HI-W: 74.4 vs. EN+HI: 74.6), possibly because the model benefits from a more balanced training signal that encourages it to learn more robust decision boundaries. These findings indicate that relatively simple training-time adjustments, such as reweighting instances based on group prevalence, can meaningfully improve fairness in cross-lingual bias detection without requiring extensive architectural changes or additional labeled data.

5.5. Qualitative error analysis

To complement our quantitative analysis, we perform a qualitative examination of misclassified examples across several representative groups. This inspection reveals recurring error patterns that help explain the aggregate trends and illuminate the limitations of current models. Figure 1 summarizes some of these patterns by showing, for a small set of groups, how group-wise F1 changes across HI-ONLY, EN→HI, and EN→HI-W.

First, we frequently observe *over-flagging of benign mentions*. Posts that refer to a group in a neutral or positive context, such as expressing solidarity or reporting factual information, are sometimes labeled as biased simply because they contain identity terms that are strongly associated with bias in the training data. This phenomenon is particularly pronounced in models that rely heavily on English pre-training or translation, where identity terms may have been used disproportionately in toxic contexts.

Second, we identify systematic *under-detection of implicit bias*. Sarcastic comments, dog-whistles, and coded language targeting caste groups or political opponents often escape detection, especially when the model's training has emphasized overt slurs and explicit insults. Cross-lingual transfer can exacerbate this issue when English training data do not capture the same repertoire of implicit biases and local stereotypes that exist in Hindi, leading the model to overlook subtle but harmful content.

Third, in the TRANS-EN regime, we encounter clear examples of *translation artifacts*. Machine translation sometimes introduces more aggressive or biased phrasing than the original Hindi, or removes politeness markers and discourse particles that mitigate the perceived harshness of a statement. Conversely, the translation may omit key context that signals bias in Hindi, leading the English classifier to misinterpret the tone. Both types of artifacts can distort the model's judgment, inflating false positives or false negatives for certain groups.

These qualitative observations reinforce the need for language- and culture-specific modeling and for caution when deploying translation-based pipelines for sensitive tasks like bias detection. They also highlight potential avenues for future work, such as incorporating conversational context, leveraging human rationales to better capture implicit bias, or designing models that are explicitly aware of local sociocultural norms.

References

[1] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur), 51*(4), 1-30.

[2] Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos One, 15*(12), e0243300.

[3] Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 173–182). ACM.

[4] Evolvi, G. (2018). Hate in a tweet: Exploring internet-based islamophobic discourses. *Religions, 9*(10), 307.

[5] Upadhyay, U., Sreedhar, I., Singh, S. A., Patel, C. M., & Anitha, K. L. (2021). Recent advances in heavy metal removal by chitosan based adsorbents. *Carbohydrate Polymers, 251*, 117000.

[6] Benesch, S. (2023). Dangerous speech. *Digital Communication Research, 12*, 185-197.

[7] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10).

[8] Simon, H., Baha, B. Y., & Garba, E. J. (2022). Trends in machine learning on automatic detection of hate speech on social media platforms: A systematic review. *FUW Trends in Science & Technology Journal, 7*(1), 001-016.

[9] Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (pp. 36-41).

[10] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019, December). 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* (pp. 14-17).

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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: Human Language Technologies*, volume 1 (long and short papers) (pp. 4171-4186).

[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440-8451).

[13] Yadav, A., Chandel, S., Chatufale, S., & Bandhakavi, A. (2023). Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. arXiv preprint arXiv:2304.00913.

[14] Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT), 20*(2), 1-22.

[15] Ranasinghe, T., & Zampieri, M. (2021). Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing, 21*(1), 1-13.

[16] Ardehaly, E. M., & Culotta, A. (2018). Learning from noisy label proportions for classifying online social data. *Social Network Analysis and Mining, 8*(1), 2.

[17] Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73).

[18] Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019, May). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 491-500).

[19] Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668-1678).

[20] Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050.

[21] Nozza, D. (2021, August). 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) (pp. 907-914).

[22] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44).

[23] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323).

[24] Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020, April). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems* (pp. 1-14).

[25] Sahoo, N., Mallela, N., & Bhattacharyya, P. (2023, July). With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13316-13330).

[26] Narayanan, A. (2018, February). Translation tutorial: 21 fairness definitions and their politics. In *Proc. conf. fairness accountability transp.*, new york, usa (Vol. 1170, p. 3).

How to cite this article: Sonia Shahzadi and Sanjiv Kumar (2024). Who Gets Protected? A Fairness Analysis of Cross-Lingual Social Bias Detection for Hindi. *Bulletin of Computer and Data Sciences*, 5(2), 1-13. DOI: 10.71448/bcds2452-1

Received: 24/12/2023 **Revised:** 21/2/2024 **Accepted:** 21/04/2024 **Publish:** 30/06/2024

Copyright: © 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

BULLETIN OF COMPUTER AND DATA SCIENCES *Bulletin of Computer and Data Sciences* is a peer-reviewed open access journal.