How do datasets, developers, and models affect biases in a low-resourced language?

Anonymous Author(s)

ABSTRACT

Sociotechnical systems, such as language technologies, frequently exhibit identity-based biases. These biases exacerbate the experiences of historically marginalized communities and remain understudied in low-resource contexts. While models and datasets specific to a language or with multilingual support are commonly recommended to address these biases, this paper empirically tests the effectiveness of such approaches in the context of gender, religion, and nationality-based identities in Bengali, a widely spoken but low-resourced language. We conducted an algorithmic audit of sentiment analysis models built on `mBERT` and `BanglaBERT`, which were fine-tuned using all Bengali sentiment analysis (BSA) datasets from Google Dataset Search. Our analyses showed that BSA models exhibit biases across different identity categories despite having similar semantic content and structure. We also examined the inconsistencies and uncertainties arising from combining pre-trained models and datasets created by individuals from diverse demographic backgrounds. We connected these findings to the broader discussions on epistemic injustice, AI alignment, and methodological decisions in algorithmic audits.

1 INTRODUCTION

Sociotechnical systems reinforce and perpetuate the systematic privileging of certain social identities and marginalization of others [58]. Marginalization refers to pushing individuals or groups to the fringes of society due to one or more intersecting aspects of their identities [28, 120]. When computer systems (e.g., algorithms) systematically marginalize and unfairly discriminate against certain individuals or groups in favor of others on unreasonable or inappropriate grounds, Friedman and Nissenbaum defined such incidents as bias [58]. While algorithmic audits seek to identify such biases in computing systems [88], these studies often focus on predominantly Western contexts and languages [47].

Given the resource disparity in natural language processing (NLP) [77], there is a dearth of critical studies in many major languages [38]. In this paper, we focus on the sentiment analysis task–the computational process of identifying, extracting, and categorizing the subjective information/tone expressed in text to determine whether the attitude toward something is positive, negative, or neutral, in the Bengali language (বাংলা: /baŋla/, endonym: Bangla), which is spoken by more than 260 million people [39]. Bengali people's colonial past profoundly shaped gender relations, exacerbated religious divisions between Hindus and Muslims [22], and fractured their nationality-based identities [36] in the Bengali (বাঙালি: /baŋali/, endonym: Bangali) ethnolinguistic communities [121]. Given their demographic distribution among different genders, Hindu (28%), Muslim (70%), Bangladeshi (57%), and Indian (34%) identities [18, 74], and their strong cultural presence online [40, 77], it is time the algorithmic fairness, accountability, and transparency (FAccT) researchers focused on the NLP datasets and models for this widely spoken language.

In the absence of well-rounded efforts in representing major global languages in state-of-the-art NLP research, language models pre-trained using multilingual data are often proposed as alternatives [43], though all languages are not represented equally in those models [138]. In some instances, researchers have prepared a few comparatively large datasets in local languages and pre-trained popular language models on those [13, 66]. While the fallacy of AI functionality–the mistaken belief that an AI system functions as intended simply because it performs well in evaluation settings [102]–prevents users from seeing where systems do not function as expected and obscures these points where different components make connection, contrast, or transition [52], these can lead to algorithmic biases that disproportionately impact marginalized communities [37, 58]. Therefore, we need to examine the usefulness of such approaches and components through critical studies and audits.

Prior scholarship has found gender, religion, and nationality-based biases in off-the-shelf Bengali sentiment analysis (BSA) tools [37], but falls short in tracing the origins of these identity-based biases. In this paper, we algorithmically audited 19 BSA datasets identified from the Google Dataset Search and two language models, `mBERT` and `BanglaBERT`, to identify their biases in terms of gender, religion, and nationality-based identities. We aim to investigate their connections to the BSA training datasets, the demographic backgrounds of their developers, and the underlying pre-trained language models. Here, our study is guided by the following three research questions:

- **RQ1:** Do language models fine-tuned with BSA datasets show biases based on gender, religion, and nationality?
- **RQ2:** Are the biases of the fine-tuned BSA models related with the dataset developers' demographic backgrounds?
- **RQ3:** How do the combinations of different language models and datasets influence the fine-tuned models' biases?

We fine-tuned 38 models based on two pre-trained models and 19 BSA datasets. This paper focuses on systematically auditing identity-based biases in sentiment analysis models, not on evaluating the sentiment of the text itself. In auditing those, we found that 61% are biased toward, i.e., assign significantly higher sentiment scores to male identity, while 24% are biased toward female identities. In the case of religion-based Bengali identities, we found that 24% and 61% are respectively biased toward the direct mentions of Hindus and Muslims or the resemblance of these communities' linguistic norms. Among the fine-tuned BSA models, half (50%) were biased toward Bangladeshi identities, compared to 26% being biased toward the explicit or implicit expression of Indian nationality. Though we found that predominantly male, Muslim, and Bangladeshi Bengalis were involved in the curation and development of BSA datasets, our analysis did not suggest a relationship

between their demographics and the biases of the BSA models. By scrutinizing the combinations of different language models and datasets rather than concealing them, we found that the language-specific `BanglaBERT` pre-trained model yields comparatively less biased fine-tuned models than `mBERT` does for Bengali sentiment analysis, highlighting the importance of language-specific models over multilingual ones. We also quantified BSA datasets demonstrating varying degrees of fairness, where no single dataset was free from bias–those with less bias in one identity dimension (e.g., gender) often exhibited significant biases in other identity dimensions (e.g., religion and nationality). This observation underscores the complexity of achieving comprehensive fairness in algorithmic systems. We connect these findings to the broader discussion on epistemic injustice in NLP, decolonizing NLP for AI alignment, and making methodological decisions for algorithmic audits.

2 LITERATURE REVIEW

2.1 Marginalization of Social Identities and Linguistic Expression in Bengali

While identity is often understood as an individual construct rooted in self-perception [60], it is also shaped by one's sense of belonging to various social groups [131]. These social identities, which are often interconnected, are defined across various **dimensions**, such as race, ethnicity, gender, sexual orientation, religion, nationality, and caste. Within each dimension (e.g., religion), people can identify with different **categories** (e.g., Christian, Muslim, Hindu) [84]. We view these categories as shaped by long-standing societal norms and practices, driven by a myriad of cultural, institutional, and political forces [20, 39]. Someone can express their social identities both explicitly and implicitly. Explicit identity expressions are deliberate and direct ways individuals communicate their affiliations, characteristics, and beliefs [131]. In contrast, implicit expressions involve subtle, indirect cues implied by actions, behaviors, and choices shaped by cultural norms, societal expectations, and institutional practices [20, 70, 133]. For example, a person may directly mention their nationality or political views, while they can also implicitly communicate and enact such identities by conforming to societal norms and certain practices through language and appearances [20]. Let's examine the cultural and linguistic norms in the Bengali language.

Bengali people's geo-cultural variations manifest in the forms of two major dialects: *Bangal* and *Ghoti* and bear important signifiers of cultural identity [55, 67]. The first one is spoken in Bangladesh, whereas the second one is commonly spoken in the Indian state of West Bengal [38]. These two dialects are different both phonologically and in their use of different colloquial vocabularies for written texts and verbal communication [78, 96]. For example, to mean the word "water," Bangladeshi and Indian Bengalis respectively use the words "জল" (/ʤɔl/) and "পানি" (/ˈpɑːniː/). Thus, a Bengali person's consistent use of words normative in the Bangal or Ghoti dialect would *implicitly* indicate their national identity. Though, unlike many other Indo-European languages, gender in Bengali does not affect pronouns (as in English) and verbs (as in Hindi and Urdu) [37], the common names and kinship terms used to describe people in Bengali textual communication can often imply their gender as well as their membership or birth into either Hindu or Muslim communities [38, 46]. For example, Bengali Hindus culturally tend to use Bengali words derived from Sanskrit, whereas the vernacular use of Perso-Arabic words is widely popular among Bengali Muslims. Both religious groups draw inspiration from their respective sacred texts for personal names (e.g., demigods, legendary characters, prophets, caliphs, and emperors) [46]. Thus, linguistic styles in Bengali texts can express one's gender, religion, and nationality.

While long-standing norms shape such expressions of social identities, historical events can significantly alter these identity norms. As identity dimensions often interconnect and overlap, the consequent intersectional identities collectively shape their unique experiences, social positions, and systemic privileges [27, 31]. For example, the Bengali communities' history with colonization impacted different gender, religion, and nationality-based identity categories. British colonial masculinity reinforced gender stereotypes, limiting women's sociopolitical roles and deepening ethnic and gender divides in Bengali societies [124]. It reshaped religious values in the Indian subcontinent, fueled religious extremism and violence through divide-and-rule tactics among Hindus and Muslims [41, 90]. Exploiting that religious division, Bengal was used as a site of partition, causing massive displacement [95]. Consequently, it annexed West Bengal with Hindu-majority India and marginalized the Muslims and underprivileged caste Hindus in East Bengal under Pakistani subjugation until gaining independence as Bangladesh [36, 118].

Similarly, as certain identities are perpetuated as normative in global and regional structures through media and technology [3, 8], other identities and practices are rendered non-normative and become marginalized. For instance, the normative use of English has marginalized non-native speakers and eroded linguistic diversity [97]. In the context of the Bengal region and the Bengali language, during the introduction of the printing press in Bengal, the influential upper-caste Hindu landlords' *Ghoti* dialect from West Bengal became the de facto standard [22], while the *Bangal* dialect, was associated with the agrarian system of and refugees from East Bengal (now Bangladesh) and marginalized [39, 61]. This dialect also became associated with Muslims and lower-caste Hindus, reflecting social biases that have come to shape people's everyday experiences [39, 61]. In standardizing the dialects of particular social classes or sociolects [86], different speech and non-verbal acts can serve as the vehicle for marginalizing certain identities [16], and this marginalization continues to be perpetuated by and through technology, such as NLP models and datasets. In this paper, we are particularly interested in understanding the marginalization of different gender, religion, and nationality-based identities by NLP models and datasets based on their explicit and implicit expression in Bengali texts.

2.2 Social and Algorithmic Identities' Relationship with Sociotechnical Systems' Biases

We employ a sociotechnical approach to exploring NLP technologies and their biases. Instead of referring to a specific technology, a sociotechnical perspective is guided by the idea that technology, broadly construed, is interconnected with people across contexts.

Underlying this view is the perspective that technology shapes and is shaped by human action and interaction [110]. In sociotechnical systems, people's identities are algorithmically constructed through a dynamic interplay between pre-existing social categories (e.g., gender, race) and social norms, cultural contexts, and historical understandings. As algorithms become increasingly integral to sociotechnical systems, users' data and interactions are analyzed to construct these algorithmic identities [24]. For example, people are assigned algorithmic identities through various interpretations of their preferred languages of interaction, search histories, social connections on social media, and more. As a result, while identities in sociotechnical systems are continuously shaped and reshaped by human-defined categories, technology and its underlying algorithmic and data-driven processes rely on reductionist and stereotyped representations of social relationships and identities [48].

This dynamic of technology perpetuating reductionism and stereotyping results in sociotechnical systems that reinforce existing societal biases while generating new intersectional biases through algorithmic extrapolations, interpolations, and decisions [37, 48]. For example, studies have found that NLP tools are often unable to understand racial, ethnic, and religious minorities' dialects [81] or classify their linguistic practices as negative and abusive [37, 42, 109]. Researchers previously examined the biases of computational systems across different social identity dimensions [15, 87], such as gender [72], race [109], nationality [134], religion [12], caste [6], age [44], occupation [132], disability [135], and political affiliations [1]. Such biases can be put into three categories [58]: preexisting, technical, and emergent.

Preexisting bias has its roots in social institutions, practices, and prejudicial attitudes, which can be reinforced in sociotechnical systems through various means. For example, researchers studied how online interaction among Bengali users is shaped by and reflects their historical religious and national divisions [36, 39]. Studying how governance shapes users' everyday experiences on online platforms, Das and colleagues explain how moderators enforce dialects used by certain groups as the standard form of language, protect selective identity groups from hate speech, and how users' collective surveillance and reporting foster a majoritarian privilege. These adversarial experiences of and biases against marginalized groups on computing platforms originate from and are perpetuated through deeply ingrained pre-existing social attitudes (e.g., toward different religions) and norms (e.g., dialects). Hence, contemporary critical scholarship in fields such as FAccT [9], human-computer interaction (HCI) [65], and NLP [14] have urged interrogating the positionality and investigating the issues around power among technology users, designers, and developers.

Technical bias arises from technical constraints or considerations [58]. When developers attempt to replicate fuzzy and qualitative social heuristics through quantitative measurements in algorithmic systems, they encounter inherent technical constraints. Exacerbating this issue, many technical artifacts rarely contain underlying source material for how different identities (e.g., race, gender) are defined and deem classifications of identities as insignificant, indisputable, and apolitical [113–115]. This leads to frequent misclassification, biased decisions, and disproportionate resource allocation in various domains, including online community moderation [37], child welfare [111], higher education [85], and policing [64]. Algorithmic systems' failure to capture complex social understanding around identities leads to users facing technical biases. Although FAccT studies on algorithmic systems identify and address such biases, existing scholarship has predominantly has focused on and been guided by Western and US-centric contexts, communities, and languages [47, 82], which Laufer et al. characterized as "narrow inquiry." Similarly, in NLP, only 0.28% of languages are considered 'winners,' while 88.38% are 'left behind' in research and technical resources [77].

While it is possible to identify pre-existing and technical biases during system design, emergent bias arises only in the context of use, especially when new societal knowledge and mismatches between users and system design emerge [58]. It is often a consequence of a technology being used in a different use case than for which it was originally intended. For example, Eubanks explored how algorithms designed for surveillance and policing can lead to bias and inequality when applied in different contexts, such as welfare or social services [54]. While such practices of leveraging models or datasets from one use case for other related tasks, especially for low-resourced contexts [140], are quite common, algorithmic fairness scholars urge for accountable and transparent approaches to developing and deploying AI systems [100, 117, 122].

2.3 Algorithmic Audits for Bias Detection in Computing Systems

Prior scholarship on algorithmic fairness, accountability, and transparency proposed "algorithmic audit" as a way for evaluating sociotechnical systems and content for fairness and detecting their biases [88, 108]. In this process, researchers conduct randomized controlled experiments by probing a system with one or more inputs while changing some attributes of that input (e.g., identity category) in a setting different from the system's development [88]. Unlike other common experiments, such as A/B tests that consider the users as the subjects, in algorithmic audits, the system itself is the subject of study [88]. Audits differ from other types of system testing due to their broader scope, resulting in systematic evaluations rather than binary pass/fail conclusions for individual test cases. Moreover, audits are purposefully intended to be external evaluations based only on outputs, without insider knowledge of the system or algorithm being studied [88]. Traditionally, querying an algorithm with a wide range of inputs and statistically comparing the corresponding results has been one of the most effective ways for algorithmic audits [88, 129].

While audit has been widely adopted in algorithmic fairness research, its origin is credited to Bertrand and Mullainathan [11], who examined racial discrimination in hiring by submitting fictitious resumes with white-sounding or Black-sounding names to job postings and found that otherwise similar resumes with white-sounding names received 50% more callbacks. Building on this approach, computing researchers have queried algorithmic systems like Google Ad delivery [129, 130] and sentiment analysis tools [37, 80] with common names associated with particular gender and racial groups and found that names associated with certain identities can lead to significantly different outputs. Recent studies examined biases in computing systems in response to explicit references to certain

demographic groups and have also considered other implicit indicators of identity, such as community-specific colloquial vocabularies, kinship terms, and distinct writing styles [37, 44]. Researchers have employed algorithmic audits across various domains, including housing [51], hiring [23], healthcare [93], policing [64], the sharing economy [50], and gig work [63], to examine fairness and biases of complex and often proprietary sociotechnical systems such as recommendation systems [7], search algorithms [106], music platforms [53], facial recognition [19], and large language models [89].

While most algorithmic fairness research studies the biases between traditionally dominant and marginalized social groups (e.g., the racial majority and minorities in the US), scholars have also urged to study the power dynamics and harm within marginalized communities [104, 136] (e.g., different economic classes among racial minorities). For example, within the underserved Bengali ethnolinguistic group, Das and colleagues [37, 38] examined biases toward different Bengali social groups defined by gender, religion, and nationality. They prepared a cultural bias evaluation dataset of sentences that explicitly and implicitly express gender, religion, and nationality-based identities within the Bengali communities [38]. Using that dataset, they audited off-the-shelf Bengali sentiment analysis (BSA) tools and identified the colonial impulses in their identity-based biases [37]. Their study is closest to the focus of this paper. However, their investigation of existing BSA tools falls short of explaining how those tools' biases relate to the pre-trained models, fine-tuning datasets, and the demographics of dataset developers–a gap that we seek to examine in this paper.

Our study focusing on the colonially marginalized Bengali communities also responds to Laufer and colleagues' call to foreground non-Western and Indigenous values and politics [82]. Despite some recent focus on South Asian contexts and languages (e.g., Hindi) [10, 62, 101], there is a dearth of literature on algorithmic fairness in Bengali language technologies. Given the reliance of pre-trained models and transfer learning in such low-resourced contexts, we build on prior FAccT scholarship that examined their adoption, use, and impacts [21, 59, 128]. Many researchers identified inappropriate blaming and unclear choice of pre-trained models as a barrier to transparency [29, 92], while others foregrounded the issues of datasets and their politics [73, 99]. Notably, existing research focused on accounting for individual and collective identities in crowdsourced dataset annotation [45] and meaning making of categories [105, 112].

3 METHODS

In this paper, we conducted an audit of sentiment analysis in Bengali, a low-resource language in NLP, given the scarcity of dataset availability and model support in this language. Considering how colonization has and continues to impact Bengali communities and their identities, we focused on biases across three identity dimensions and corresponding major binary categories: gender (female: ♀ and male: ♂), religion (Hindu: 🕉 and Muslim: ☪), and nationality (Bangladeshi: 🇧🇩 and India 🇮🇳). Here, we describe our approach to identifying Bengali sentiment analysis (BSA) datasets, conducting a survey with their developers to collect their demographic information, identifying language models pre-trained with Bengali data, and setting up the experiment for algorithmic audit, including details about fine-tuning, the bias evaluation data set, the statistical approach for comparison, and metrics for quantifying group bias.

3.1 Identifying Bengali Sentiment Analysis Datasets and Contacting Their Developers

To streamline the search for datasets, we utilized Google Dataset Search[1], which enables the discovery of datasets hosted on popular repositories (e.g., Kaggle and Mendeley Data)–platforms frequently used by NLP researchers and dataset developers. Given the wide variance in how sentiment datasets are often described (e.g., sentiment analysis/classification/categorization), we searched for Bengali sentiment analysis (BSA) datasets on Google Dataset Search using the phrases ``Bengali sentiment'' and ``Bangla sentiment'' on January 10, 2024. We excluded duplicates and datasets for other tasks (e.g., fake news detection) from the search results by reading through their descriptions. Similar to prior work [34, 123], in cases of datasets for related tasks (e.g., multi-class emotion classification), we compressed the multiple fine-grained positive/negative classes into a single positive/negative class following the instructions provided in the corresponding dataset's documentation, if available. Finally, we included 19 BSA datasets in this study, each with an average of 16,415 labeled data instances. We also collected metadata about these datasets, including developers' names, contact information, affiliations, and countries, by reviewing their data repository profiles (e.g., Kaggle, GitHub), README files, and published research papers. With the approval of the institutional review board (IRB), we invited the developers to participate in an online survey to know their demographic information. We received responses from developers of 12 BSA datasets, whom we compensated with $20 for their time. Since our study also involves examining the links between BSA models trained on these datasets and their developers, we did not intend to associate our critique with the developers personally or provide any information that would allow anyone to trace back and identify them. Hence, we obfuscated the datasets to protect the developers' anonymity following methods from ethics literature on using internet resources in research [17, 56]. In doing so, we de-identified the datasets (see Table 1) by using random identifiers.

Table 1: Examined BSA datasets and their developers' demographic backgrounds.

Dataset IDs	Developer Demographics
D1, D8, D10, D12-D14, D17	N/A
D2, D3, D5, D6, D9, D15, D16, D18, D19	♂ ☪ 🇧🇩
D4, D11	♀ ☪ 🇧🇩
D7	♂ ☪+Agnostic 🇧🇩

3.2 Identifying Language Models for Bengali

We fine-tuned pre-trained language models for sentiment analysis tasks using a specific BSA dataset to identify biases that are unique

[1] https://datasetsearch.research.google.com/

to that dataset. Doing so can provide insights into how the biases in both the pre-trained model and the BSA dataset influence the model's sentiment analysis. We considered some variants of Bidirectional Encoder Representations from Transformers (BERT) [43], which were pre-trained using Bengali data. For example, BERT's multilingual variant (henceforth, `mBERT`) is pre-trained and *"generalizes"* in 104 languages [98], and Bengali is one of those languages. There exists the `BanglaBERT` model, which was pre-trained *"specifically"* with Bengali corpora with both Bengali and Romanized scripts and reportedly outperformed other similar models for sentiment classification tasks in Bengali [13]. Given their pre-training data's linguistic diversity, we refer to `mBERT` and `BanglaBERT` as generalized and specialized language models, respectively. Though the Bengali alphabet doesn't have case variation, considering that a few BSA datasets (e.g., D9) contain Romanized Bengali, where case variation is used to indicate different sentiments by Bengali speakers online [35], we used the case-sensitive `mBERT` but `BanglaBERT` has no case-sensitive version.

3.3 Experiment Setup for Algorithmic Audit

We designed our experiment as an algorithmic audit [88, 108]. First, we fine-tuned `mBERT` and `BanglaBERT` models using the BSA datasets, D1-D19, as shown in Figure 1 (a). We audited gender, religion, and nationality-based biases in the resulting $\binom{2}{1} * \binom{19}{1} = 38$ fine-tuned BSA models. We queried each fine-tuned BSA model $Di - x$ (where $i \in [1 - 19]$ and $x \in \{\texttt{mBERT}, \texttt{BanglaBERT}\}$) with pairs of identical sentences from the Bengali identity bias evaluation dataset (BIBED) [38] that explicitly (through direct mentions) and implicitly (through linguistic norms) represent different Bengali gender, religion, and nationality-based identity categories (see Figure 1 (b)).

3.3.1 Bengali Identity Bias Evaluation Dataset. During this study, BIBED [38] is the only identity-based bias evaluation dataset in Bengali, which has been used by several audits as a benchmark dataset [37, 107]. The sentences in BIBED were sourced from Wikipedia, Banglapedia, Bengali classic literature, Bangladesh law documents, and the Human Rights Watch portal. These sentences either explicitly or implicitly express female-male, Hindu-Muslim, and Bangladeshi-Indian Bengali identities. In the case of explicit expression, the sentence pairs directly mention different gender-based (25,396), religion-based (11,724), and nationality-based (13,528) identities. Each pair contains two identical sentences, differing only in the mentioned identities. The implicit expressions of these identities rely on linguistic norms, including common names, kinship terms, and community-specific colloquial vocabularies, which are different in various cultural groups defined by major religions and nationalities among the Bengali people. There are 1,200 unpaired sentences implicitly representing gender and religion and 8,834 pairs implicitly representing Bangladeshi and Indian nationalities.

3.3.2 Comparison Approaches and Metrics. For an input sentence, a fine-tuned BSA model predicts both nominal class and sentiment score. The sentiment scores, normalized on a scale of 0 to 1, indicate "the probability associated with the positive" class [34]. For each sentence pair in BIBED, we will obtain pairs of sentiment classes and scores from a fine-tuned model. In the case of unpaired sentences, following [37, 80], we sampled an equal number (10%) of sentences from different identity categories under scrutiny (e.g., female-male) and aggregated the outputs (mode for nominal classes and average for numeric scores) into consolidated pairs. We quantified and statistically compared biases based on how the fine-tuned BSA models assigned sentiment classes and scores for different identities.

Statistical Comparison of Groups. Algorithmic audits often use statistical comparisons, such as Wilcoxon signed rank [37], t-test [80], or regression [44] to compare numerical scores assigned to different identity groups by some algorithmic entity, and χ^2 analysis [129] to examine the relationship between identity groups and nominal classification.

To answer **RQ1**, we statistically compared fine-tuned BSA models' outputs–both nominal categories and numeric scores. From an algorithmic fairness angle, there would be no relationship between the identity a sentence represents and the sentiment category it is assigned to (null hypothesis $H1cat_0$). We used the χ^2 test to assess the relationship between two nominal variables: identity category and sentiment classification. To examine whether and how different gender (female-male), religion (Hindu-Muslim), or nationality-based (Bangladeshi-Indian) identity categories impact the numeric sentiment scores, we pairwise compared the mean sentiment scores for different categories from a fine-tuned BSA model. Here, the null hypothesis ($H1num_0$) assumes the mean sentiment scores for different categories in an identity dimension to be similar (i.e., $\mu_{female} = \mu_{male}$, $\mu_{Hindu} = \mu_{Muslim}$, and $\mu_{Bangladeshi} = \mu_{Indian}$). Given the differing findings of prior studies on the direction of biases toward different gender [2, 58, 87], religion [5, 75], and nationality [39, 91]-based identities, especially in the context of the Bengali communities [37, 39], we tested two-tailed, left-tailed, and right-tailed alternative hypotheses to identify the direction of biases–the identity categories to which it assigns higher sentiment scores. To consider the tests' results significant and consistent enough to declare the outputs as biased, we used threshold values, $\alpha = 0.01$ and $power \geq 0.8$ following recommendations of [25, 26]. Since sentiment scores from all models are normalized on a common scale (0 to 1), we can interpret differences between the two columns directly without separately calculating the effect size–a standardized measure indicating the magnitude of the relationship or difference [33]. Similar to [37, 80], for an identity dimension and a fine-tuned BSA model, if the sentence pairs' sentiment score distributions maintained normality [119], we used a parametric test like the pairwise t-test [126], otherwise a non-parametric equivalent, such as the Wilcoxon signed-rank test [137] for statistical inference.

For answering **RQ2**, we examined whether the directions of a model's bias are related to the identity categories of the developers of the corresponding BSA datasets. Following [37, 129], we used χ^2 test for checking the null hypothesis ($H2_0$): "Bias of language models trained with BSA datasets are not related with their developers' demographic backgrounds."

Quantifying Group Bias. To answer how different combinations of pre-trained models and training datasets influence the biases in fine-tuned models (**RQ3**), we need to quantify those resulting models' group biases. To compare nominal classifications, we followed [34, 49]'s guidelines of demographic parity that looks for

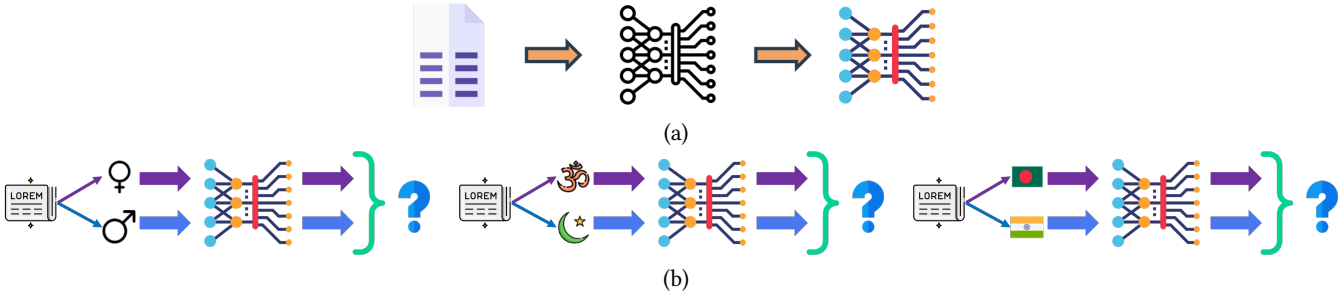

Figure 1: (a) Fine-tuning `mBERT` or `BanglaBERT` (B/W diagram in middle) with BSA datasets, Dx (icon on left) to get fine-tuned language models (color diagram on right) (b) Auditing the fine-tuned `Dx-mBERT` or `Dx-BanglaBERT` models' gender, religion, and nationality biases (First paragraph of this section lists the icons used for indicating different categories).

an equal positive classification rate (PCR) across different groups. Let T be the set of all identity categories under a particular dimension. In case of gender, $T = \{female, male\}$, for religion, $T = \{Hindu, Muslim\}$, and for nationality, $T = \{Bangladeshi, Indian\}$. S^{t_i} denotes a subset of examples associated with an identity group t_i, and $\Phi(S^{t_i})$ be the number of sentences in the set S^{t_i} that were predicted as positive by a fine-tuned BSA model, and $|S^{t_i}|$ be the size of that set. We calculate the PCRs for protected groups t_i and t_j in T and identify the identity category toward which a model's output is biased using Equation 1:

$$argmax(\frac{\Phi(S^{t_i})}{|S^{t_i}|}, \frac{\Phi(S^{t_j})}{|S^{t_j}|}) \tag{1}$$

In the case of comparing two fine-tuned models having similar PCR, we used a secondary quantifying metric of group bias, which is called pairwise comparison metric (PCM). For a sample of sentence pairs expressing different identities, PCM calculates the average difference of sentiment scores [34]. Using the aforementioned notations for PCR, let $|T|$ be the set T's size. $\phi(A)$ is the sentiment score for some set of examples A, and $d(x, y)$ means the difference between two scalar values x and y. We adopted the PCM metric defined by [34] for our experiment (see Equation 2) to compare paired sentiment scores from a fine-tuned BSA model for a set of evaluation sentence pairs, as follows:

$$\frac{1}{n} \sum_{t_i, t_j \in \binom{T}{2}} d(\phi(S^{t_i}), \phi(S^{t_j})), \quad n = \binom{|T|}{2} \tag{2}$$

3.4 Setup for Fine-tuning Models

Hooker argued that given the advent of domain specialized hardware (e.g., graphics processing unit or GPU in machine learning) we need to make it easier to quantify the opportunity cost of experiments in terms of hardware accessibility and specialized software expertise [69]. The experiment and statistical analyses were conducted using Python. We used pre-trained `mBERT`[2] and `BanglaBERT`[3] models from Hugging Face. While fine-tuning these pre-trained BERT variants, we followed [43]'s recommendations for choosing the values for hyperparameters, batch size: 16 (training) and 32 (evaluation), learning rate (Adam): 5e-5, and number of epochs: 3. We used the NVIDIA A100 (40GB PCIe) GPU on Google Colab. Wherever applicable (e.g., sampling data splits on a MacBook Air M2), we used a fixed seed value for the replicability and consistency of our results.

3.5 Researcher Positionality

Researchers' identities reflexively bring certain affinities into perspective while studying underserved communities [4, 83, 116]. In particular, our work follows Bird's call for decolonizing language technologies [14, 38] by focusing on a low-resource language spoken by colonially marginalized transnational communities from the Global South. The first two authors were born and raised in the [anonymized nationality] and [anonymized nationality] Bengali communities, respectively, and the anchor author is an [anonymized nationality] who is a member of an Indigenous group from [anonymized country]. All authors identify as [anonymized gender] [anonymized sexual orientation] and are affiliated with [anonymized region] universities. Besides our positionalities, our interdisciplinary backgrounds, including computer science, economics, information science, and statistics, and our research experience in critical studies, algorithmic bias and fairness, cross-cultural NLP, and marginalized ethnolinguistic groups contribute to our motivation and capacities and this study's mindfulness and care toward under-represented Bengali communities.

3.6 Environmental Impacts

Mindful of the concerns of environmental colonialism and injustice–pollution from activities, like the development of large AI models disproportionately and adversely affecting marginalized communities who do not even benefit from those models, researchers have previously encouraged considering environmental impacts in responsible research in big data and related fields like NLP [30, 125, 139]. In this work, we fine-tuned 38 models using the NVIDIA A100 (40GB PCIe) GPU on Google Colab. Considering that this device's power consumption under high loads is 250W[4], and Google's typical data center's carbon footprint is 0.082 $kgCO_2/kWh$, training models in our study released approximately 0.2 kg CO_2, which is

[2]https://huggingface.co/bert-base-multilingual-cased
[3]https://huggingface.co/csebuetnlp/banglabert
[4]https://bit.ly/a100-power-consumption

negligible compared to the most resource-intensive models [125]. As a gesture to offset this carbon pollution, we donated to the United States Forest Service's Plant-a-Tree program. Moreover, our study advocates for historically marginalized Bengali communities by highlighting language models' and datasets' biases and identifying fairness considerations for their deployment in downstream tasks, like content moderation [127].

3.7 Limitations and Future Work

Using BIBED [38], which highlighted two major genders, religions, and nationalities, our study overlooked non-binary genders, smaller religious minorities, diaspora nationalities, and smaller regional linguistic norms. It was the only Bengali dataset to identify bias during our study, which was the primary reason for adopting the binary identity classification. Such common practice of binarification in NLP datasets and artifacts that shape and restrict algorithmic audits is indicative of the field's limitations. Despite our intention and efforts (e.g., connecting with developers of different religious beliefs) to go beyond binaries, we were limited by the ontologies of available resources. Beyond examining the biases in each dimension of fine-tuned models individually, future work should investigate their intersectional biases and other vital identity dimensions, such as caste and sexual orientation. However, relying on quantitative methods, this paper is limited in its capacity. In our future work, we will draw on interviews and ethnography to understand how developers prepare datasets and choose pre-trained models in low-resource contexts.

4 RESULTS

In this section, we first explain whether and how language models fine-tuned with BSA datasets exhibit biases. Second, by examining the relationship between the identities fine-tuned models are biased toward and the identities of the dataset developers, we underline the politics of design. Third, we foreground the influences on the fine-tuned models that stem from different combinations of language models and BSA datasets.

4.1 RQ1: Do language models fine-tuned with BSA datasets show biases based on gender, religion, and nationality?

In this study, we audited 38 fine-tuned BSA models using pairs of sentences with identical semantic content, structure, and meaning that differ only in the identity the sentences represent. Consider the following two sentences: "পানি পরিবেশের একটি গুরুত্বপূর্ণ উপাদান।" and "জল পরিবেশের একটি গুরুত্বপূর্ণ উপাদান।", both of which mean "Water is an important element of the environment." In addition to their exact same meaning, these two sentences have identical semantic content and sentence structures, except using the underlined words পানি (/ˈpɑːniː/) and জল (/zɔl/) to mean the word "water." Between these two synonymous words, Bangladeshi Bengalis commonly use the first word, while Indian Bengalis typically use the second. Despite the same structure and similar semantic content, while D1-mBERT categorized the first sentence as positive (sentiment score 0.9758), the second was categorized as negative (sentiment score 0.1062). This discrepancy of sentiment categories and scores for sentences in the pair exhibits a nationality bias based on linguistic norms. For **RQ1**, the question is whether these output discrepancies in sentiment analysis tasks are significant and consistent across language models fine-tuned with BSA datasets.

The results of our χ^2 suggest that the nominal sentiment classifications of nine fine-tuned models, including (D2, D4, D5, D6, D7, D10, D11, D18)-mBERT and (D15)-BanglaBERT, consistently (e.g., with a power ≥ 0.8), relate to the gender represented in a sentence. For 12 fine-tuned models: (D1, D2, D4, D5, D9, D10, D11, D17, D19)-mBERT and (D1, D10, D17)-BanglaBERT, sentiment classifications were often related to the religion-based identities expressed by the Bengali sentences. In the case of nationality-based identity, outputs of nine fine-tuned BSA models, including (D1, D11, D14, D16, D17, D19)-mBERT and (D1, D3, D13)-BanglaBERT, were related to whether the sentences explicitly mentioned or followed the linguistic norms of Bangladeshi or Indian Bengalis. Among the 38 fine-tuned models audited in our study, this approach identifies less than half of these as biased in each identity dimension.

Table 2 presents the results of pairwise comparisons of the numeric sentiment scores for different categories in each identity dimension. Details about χ^2 and Wilcoxon signed rank or paired t-tests are in Table 3 in the Appendix.

Comparing sentiment score pairs, we found that among these models, 9 fine-tuned models (24%) are biased toward female identity (e.g., consistently assign more positive sentiment scores to sentences that explicitly or implicitly express female identities). Similarly, 23 models (61%) are biased toward male identities. In the case of religion-based identities, fine-tuned models that are biased toward Hindu and Muslim identities amount to 24% and 61%, respectively. For the nationality dimension, 50% of the fine-tuned models were biased toward, i.e., perceived Bangladeshi identity more positively, compared to 26% models being biased toward Indian identity.

4.2 RQ2: Are the biases of the fine-tuned BSA models related to the dataset developers' demographic backgrounds?

In answering the previous RQ, we found how mBERT and BanglaBERT, being fine-tuned with different BSA datasets, exhibit biases toward one or the other identity categories of gender, religion, and nationality. Given that most BSA dataset developers share similar identities, could the biases of the models fine-tuned using those datasets be surfacing the lack of representation from other identities and the potential misalignment among the diversities within Bengali communities? In **RQ2**, we investigate whether the demographic backgrounds of the developers of these datasets are related to how these datasets influence the direction of the biases in the fine-tuned models. This question is particularly important given the emphasis on the positionality of designers in critical scholarship in HCI, as discussed in section 2. However, our analysis did not provide conclusive evidence that the biases of mBERT and BanglaBERT models fine-tuned with BSA datasets are related to the demographic background of the dataset developers. Tables 4, 5, and 6 in the Appendix present the direction of bias in the fine-tuned BSA models and the demographic backgrounds of their developers across the dimensions of gender, religion, and nationality, respectively. We excluded the fine-tuned models trained with datasets for which

Table 2: Results of statistical tests pairwise comparing numerical sentiment scores.

	H_a/Directions of bias	mBERT	BanglaBERT
Gender	$\mu_{female} < \mu_{male}$	D2, D5, D7, D9-D11, D13-D18 **(n=12)**	D1, D2, D5-D9, D11, D14, D16, D17 **(n=11)**
	$\mu_{female} > \mu_{male}$	D1, D3, D4, D6, D19 **(n=5)**	D12, D15, D18, D19 **(n=4)**
	no/rare	D8, D12 **(n=2)**	D3, D4, D10, D13 **(n=4)**
Religion	$\mu_{Hindu} < \mu_{Muslim}$	D1, D2, D5, D7-D11, D13, D15, D17 **(n=11)**	D1, D2, D4-D6, D8-D11, D14, D16, D17 **(n=12)**
	$\mu_{Hindu} > \mu_{Muslim}$	D3, D4, D12, D14, D16, D18, D19 **(n=7)**	D12, D15 **(n=2)**
	no/rare	D6 **(n=1)**	D3, D7, D13, D18, D19 **(n=4)**
Nationality	$\mu_{Bangladeshi} < \mu_{Indian}$	D10, D12, D18, D19 **(n=4)**	D2, D6, D8, D10, D13, D18 **(n=6)**
	$\mu_{Bangladeshi} > \mu_{Indian}$	D1, D2, D4, D5, D7-D9, D11, D13, D14, D16, D17 **(n=12)**	D1, D3, D7, D9, D14, D16, D19 **(n=7)**
	no/rare	D3, D6, D15 **(n=3)**	D4, D5, D11, D12, D15, D17 **(n=6)**

we could not collect the corresponding developers' self-identified demographic information from the corresponding hypothesis tests.

For this RQ, our null hypothesis assumes no relationship between the direction of bias in BSA tools and their developers' demographic backgrounds, whereas our alternative hypothesis assumes one exists. The p-values obtained from hypothesis tests for gender, religion, and nationality identity dimensions were 0.77, 0.27, and 1.0. Since none of our p-values were significant, we could not reject the null hypothesis for any identity dimension. Hence, based on our statistical tests, we concluded that there is no significant evidence to suggest that the biases in these fine-tuned BSA models are related to the demographic identities of the dataset developers. Then, we asked whether and how the combinations of two key components of downstream NLP systems–pre-trained language models and fine-tuning datasets–influence these biases.

4.3 RQ3: How do the combinations of different language models and datasets influence the fine-tuned models' biases?

In **RQ3**, we explore how the combinations of different pre-trained models and datasets influence the biases of the fine-tuned models. Beyond determining whether the fine-tuned models are biased, we quantified the group biases of those models using the positive classification rate (PCR) and the pairwise comparison metric (PCM).

We identified the identity toward which a fine-tuned model was biased based on PCR across ten splits of the evaluation dataset. Figure 2 shows that most of the combinations of the pre-trained models (e.g., mBERT or BanglaBERT) and fine-tuning BSA datasets exhibited a positive classification bias toward one or the other category (seen in dark blue or dark red in the heatmap) ten out of ten times we calculated those models' PCRs. Let's refer to such cases of fine-tuned models being biased toward an identity category across all data splits as "constant bias."

Figure 2 also shows how certain BSA datasets, irrespective of the pre-trained base model, always lead to identity bias toward a specific gender, religion, or nationality (e.g., models fine-tuned with D2 and D18 being biased toward Bangladeshis and Indians, respectively). This raises a question about the role of these datasets in leading to such biased models. In contrast, when we fine-tuned mBERT using D1, D3, D4, and D6, the resulting models consistently categorized female identity-expressing sentences as positive in all data splits. However, the same base model, when fine-tuned using the BSA datasets D2, D5, D7-11, and D13-18, exhibited a similarly constant positive classification bias toward male identities explicitly or implicitly expressed in Bengali sentences. Such shifts in the direction of gender bias in fine-tuned models, depending on the BSA dataset used for fine-tuning a pre-trained model, align with the common argument that critiques the problematic nature of data.

However, we observed cases where fine-tuned models challenge the notion that biases in algorithmic systems stem solely from biased training datasets. For example, though the BSA datasets D1 and D6 shaped the mBERT model to show constant bias toward female identity, the same datasets when being used in conjunction with BanglaBERT, resulted in fine-tuned models that favored male identity-representing sentences. Similarly, for religion and nationality-based identities, we saw instances of different BSA datasets shifting the same pre-trained models' direction of bias through fine-tuning (e.g., D14 and D15 leading to constant bias toward different religious identities) as well as of the same BSA dataset affecting different base models' biases to move in different directions (e.g., mBERT and BanglaBERT fine-tuned with D19 showing constant biased toward Indian and Bangladeshi identities, respectively).

Unlike the fine-tuned models we described as showing constant bias, there exist models that exhibit biases toward different genders, religions, and nationalities in different splits of evaluation data. Examining these combinations and considering instances where bias directions were less consistent than in the cases above can help identify the pre-trained model and fine-tuning dataset pairings that result in reduced bias. For example, when we used the dataset D19 to fine-tune mBERT, it resulted in a BSA model that showed a positive classification bias toward male identity seven out of ten times. When we calculated PCR for the D19-BanglaBERT fine-tuned model, we found it to be biased toward female identity-representing sentences six times out of ten. These datasets fine-tune models to favor one identity (e.g., Bangladeshi) occasionally and at other times favor the opposite (e.g., Indian). In other words, depending on the pre-trained model, these datasets slightly shift the bias direction of the BSA model but are not consistently biased, unlike the others. Models in Figure 2 with mid-spectrum colors, like off-white (e.g.,

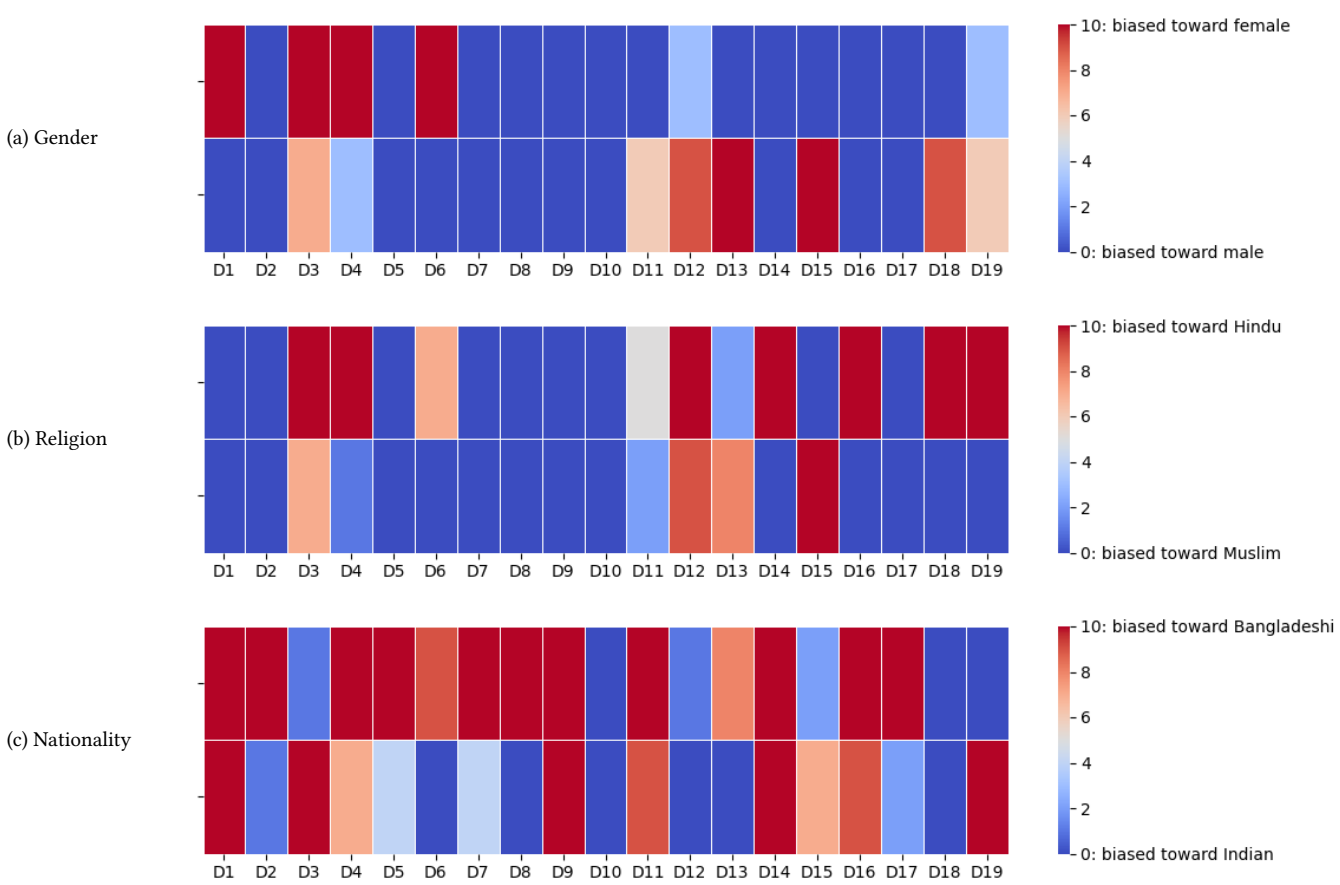

Figure 2: Heatmap showing the directions of biases of the fine-tuned models based on PCR, i.e., in how many iterations a particular combination of `mBERT` (top) or `BanglaBERT` (bottom) with different BSA datasets more frequently classified a category as positive.

D11-`mBERT`), indicate being biased toward different categories (e.g., Hindu and Muslim) an equal number of times (e.g., 5 and 5) across all data splits. All fine-tuned models' PCR are presented in Table 7 in the Appendix.

However, excluding the models that show constant bias (colored with dark blue or dark red in Figure 2), most models with inconsistent bias directions in different iterations do not have exactly equal PCRs. Therefore, to decide between two fine-tuned models that have somewhat similar PCRs, we can consider the values of PCM (see Table 7 in the Appendix) that compares the average pairwise differences of normalized sentiment scores for different categories in paired inputs. The higher this score is for a fine-tuned model, on average the more different sentiment scores that the model assigns to different categories (e.g., Bangladeshi and Indian) in a particular identity dimension (e.g., nationality). Hence, for models with equal PCRs, a lower PCM pinpoints the model that assigns less different scores to different identities.

Considering these arguments, we found that fine-tuning `BanglaBERT` with different BSA datasets resulted in fewer models with a consistent bias toward certain gender, religious, and national identities. This implies that while most fine-tuned models are likely to exhibit algorithmic bias, the pre-trained model specializing in the language of the downstream task, in this case, Bengali, is more malleable than the generalized `mBERT` model during fine-tuning.

5 DISCUSSION

Our study provides empirical evidence of language models and datasets exhibiting biases across different gender, religion, and nationality-based identities in the low-resource Bengali language. We also examine how the demographic background of the dataset developers relates to these biases and the effectiveness of multilingual and language-specific pre-training in mitigating those. Here, we reflect on our findings and their implications by connecting them to the concept of *epistemic injustice* for NLP broadly, *decolonizing NLP* to resist the dominance of certain social values in AI alignment, and *choosing among various metrics and methods* for algorithmic audits.

5.1 Epistemic Injustice in Natural Language Processing

Natural Language Processing (NLP) can be viewed as a form of epistemology [79], given its application in understanding, categorizing, and generating human language. NLP-based technologies can prioritize certain ways of interpreting information through various datasets, models, and tools [37, 44, 80]. We found that fine-tuned BSA models associate specific gender-, religion-, and nationality-categories with positive sentiments and others with negative connotations. We can conceptualize such biases while interacting with language technologies through the lens of epistemic injustice.

Epistemic injustice is unfairly discrediting someone's testimony, prejudicially undermining their ability to participate, and misrepresenting their views in knowledge practices [57]. It can manifest in two forms. First, testimonial injustice occurs when prejudice causes a hearer to give the speaker less credibility based on the latter's identity. When language models assign less favorable scores to sentences that mention a specific gender, religion, or nationality, or reflect the linguistic norms of those identity groups, this highlights the models' testimonial injustice. Second, hermeneutical injustice occurs at a prior stage, where the social experiences of members of marginalized groups are left inadequately conceptualized and ill-understood due to gaps in their respective hermeneutics. Despite English and Bengali having comparable numbers of native speakers, the latter has fewer resources available than the former by a factor of thousands [77]. Moreover, as our study found, there are serious concerns regarding bias in the limited number of labeled Bengali datasets. Since Bengali communities have a strong online presence [77], their interactions can enable NLP tools to effectively understand diverse Bengali hermeneutics. While prior work has shown that models trained on specific language families tend to outperform those trained on diverse but unrelated languages [94], our study complements this critique by demonstrating that the language-agnostic model `mBERT` systematically dismisses, conflates, or distorts dialects and linguistic styles, thereby exacerbating disadvantages for low-resourced languages. Consequently, language technologies can be unjust toward users and render their interactions with sociotechnical systems in terms of content and style structurally prejudicial [37, 81].

5.2 Decolonizing NLP as Addressing Cultural Differences in AI Alignment

AI alignment aims to ensure that AI systems align with widely shared values [71, 76]. In historically marginalized communities, participatory methods help resist cultural imposition, decolonize language technologies, and develop community-driven resources and artifacts through the negotiation of local values [14]. We found a clear under-representation of BSA dataset developers who identify as female, Hindu, and Indian, which can risk inadequately conceptualizing their experiences, cultural appropriation, and exploitation resulting from data sourced about underserved and colonially marginalized people, such as the Bengalis, without informed consent.

Contributing factors to this underrepresentation may include various social elements, such as a lack of financial incentives and insufficient political will. For example, while Bengali is India's second most spoken language, the recent government-sponsored promotion of Hindi disadvantages it in the multilingual country [103]. Considering decolonial scholarship, which views governments as continuations of colonial hierarchies, such dominance over local languages can be seen as a colonial legacy. On the contrary, Bengali being Bangladesh's national language, NLP research in that language benefits from community endeavors and governmental initiatives [37].

Let's consider ways to align AI models with the values of diverse nationalities, genders, and religious communities speaking the Bengali language. Forward alignment aims to align AI systems via alignment training, whereas backward alignment assesses the systems' alignment and governs them appropriately to avoid exacerbating misalignment risks [76]. Given the scarcity of labeled datasets in Bengali, especially those that consider fairness and equity, the feasibility of alignment training might be limited, and backward AI alignment could be a more pragmatic approach. Here, the goal is to develop robust models that do not perpetuate existing societal biases, such as predicting negative sentiment solely based on unrelated factors.

Considering the technological and infrastructural challenges in the Global South, where many low-resource languages are spoken, reflecting on sustainable and accessible NLP approaches becomes essential. Even with data availability, large models' computational demands can make them impractical. In such cases, knowledge distillation, where a smaller model is trained to replicate the behavior of a larger, more complex model, can be a viable alternative [32, 68] to reduce computational needs and support community-driven and decolonized language technology research.

5.3 Decisions around Methods and Quantification in Algorithmic Audit

We used multiple statistical tests and evaluation metrics in our audit. For example, to identify identity-based biases in fine-tuned BSA models, we compared nominal sentiment categories using the χ^2 test and numerical sentiment scores using paired t-tests or Wilcoxon signed-rank tests. Although both approaches revealed biases, more fine-tuned models were identified as biased by comparing numerical sentiment scores (summed across different identity categories, gender: 85%, religion: 85%, and nationality: 76%) than through the nominal category comparison (gender: 24%, religion: 32%, and nationality: 24%). These differences could be due to the fine-tuned models missing subtle nuances when classifying data into discrete categories rather than using continuous scores. Therefore, while some prior studies have focused on nominal categories [129], we recommend using numerical scores for a more vigilant assessment of biases.

Similarly, to examine different combinations of pre-trained models and BSA datasets, we used two metrics to quantify group bias: positive classification rate (PCR) and pairwise comparison metric (PCM), which rely on nominal categories and numerical scores, respectively. In our experiment, we found several fine-tuned models where the PCR values for different identity categories were

significantly different, indicating strong biases, but the same models had low PCM values, suggesting less bias. For example, the D11-BanglaBERT model classified Muslim identity expressing sentences as positive more frequently in more splits than in data splits where the explicit or implicit expression of Hindu identities was categorized as positive with a higher rate (see Table 7 in Appendix for details and a few more other examples). Despite such religion-based bias in this model's outputs, which leads us to expect a higher PCM based on pairwise differences in sentiment scores of sentence pairs, this model has a low PCM value. How do we interpret the inconsistencies between our expectations and observations about a particular metric? The aggregation of the differences in pairwise sentiment scores across all sentence pairs, as per the formula by [34], might have minimized the PCM value. While summing absolute differences instead of numerical differences may better capture the overall differences in sentiment scores across large datasets, its effectiveness should be confirmed through future empirical validation.

6 CONCLUSION

We presented findings from algorithmic audits of fine-tuned Bengali sentiment analysis (BSA) models based on existing BSA datasets and two BERT models: one multilingual and one specifically pre-trained for the Bengali language. Using statistical comparison and quantifying group biases, we found that BSA models exhibit biases by consistently assigning significantly different sentiment scores to sentences expressing different gender, religion, and nationality-based identities. Our study foregrounded the downstream biases of pre-trained models, examined their possible relationship to the training dataset developers' identities, and inconsistencies stemming from different combinations of pre-trained models and datasets. As algorithms become more prevalent in global sociotechnical infrastructure, we call for more audits in low-resource and cross-cultural contexts, focusing on datasets, pre-trained models, and developers. Transparency fostered through such practices in selecting datasets, models, and fairness metrics for audits can address misalignments of values and exclusion, promote social justice, and foster more inclusive and accountable AI regulations.

REFERENCES

[1] Samyak Agrawal, Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2022. Towards Detecting Political Bias in Hindi News Articles. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Dublin, Ireland, 239–244. https://doi.org/10.18653/v1/2022.acl-srw.17

[2] Sibbir Ahmad, Songqing Jin, Veronique Theriault, and Klaus Deininger. 2023. Labor market discrimination in Bangladesh: Experimental evidence from the job market of college graduates. (2023).

[3] Syed Mustafa Ali. 2016. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students* 22, 4 (2016), 16–21.

[4] Mariam Attia and Julian Edge. 2017. Be (com) ing a reflexive researcher: a developmental approach to research methodology. *Open review of educational research* 4, 1 (2017), 33–45.

[5] Imran Awan. 2016. Islamophobia on social media: A qualitative analysis of the facebook's walls of hate. *International Journal of Cyber Criminology* 10, 1 (2016), 1.

[6] Senthil Kumar B, Pranav Tiwari, Aman Chandra Kumar, and Aravindan Chandrabose. 2022. Casteism in India, but Not Racism - a Study of Bias in Word Embeddings of Indian Languages. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1–7. https://aclanthology.org/2022.lateraisse-1.1

[7] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 2–2.

[8] Sarbani Banerjee. 2015. *"More or Less" Refugee?: Bengal Partition in Literature and Cinema*. The University of Western Ontario (Canada).

[9] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.

[10] Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2024. Tackling Language Modelling Bias in Support of Linguistic Diversity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 562–572.

[11] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[12] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing Fairness in NLP: The Case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online only, 727–740. https://aclanthology.org/2022.aacl-main.55

[13] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1318–1327. https://doi.org/10.18653/v1/2022.findings-naacl.98

[14] Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*. 3504–3519.

[15] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[16] Nina Brown, Thomas McIlwraith, and Laura Tubelle de González. 2020. *Perspectives: An open introduction to cultural anthropology*. Vol. 2300. American Anthropological Association.

[17] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4 (2002), 217–231.

[18] Bangladesh Statistics Bureau BSB. 2022. Preliminary Report on Population and Housing Census 2022 : English Version. https://drive.google.com/file/d/1Vhn2t_PbEzo5-NDGBeoFJq4XCoSzOVKg/view. [Accessed: Feb 28, 2023].

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[20] Judith Butler. 2011. *Gender trouble: Feminism and the subversion of identity*. routledge.

[21] Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 370–378.

[22] Partha Chatterjee. 1993. *The nation and its fragments: Colonial and postcolonial histories*. Princeton University Press.

[23] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[24] John Cheney-Lippold. 2017. *We are data: Algorithms and the making of our digital selves*. New York University Press.

[25] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

[26] Jacob Cohen. 2016. A power primer. (2016).

[27] Patricia Hill Collins. 2022. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.

[28] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality*. John Wiley & Sons.

[29] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 864–876.

[30] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[31] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*. Routledge, 23–51.

[32] Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, Tom Sercu, Kartik Audhkhasi, Abhinav Sethy, Markus Nussbaum-Thom, and Andrew Rosenberg. 2017. Knowledge distillation across ensembles of multilingual models for low-resource languages. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4825–4829.

[33] Peter Cummings. 2011. Arguments for and against standardized mean differences (effect sizes). *Archives of pediatrics & adolescent medicine* 165, 7 (2011), 592–596.

[34] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics* 9 (2021), 1249–1267.

[35] Dipto Das and Anthony J Clark. 2019. Construct of Sarcasm on social media platform. In *2019 IEEE international conference on humanized computing and communication (HCC)*. IEEE, 106–113.

[36] Dipto Das, Dhwani Gandhi, and Bryan Semaan. 2024. Reimagining Communities through Transnational Bengali Decolonial Discourse with YouTube Content Creators. *arXiv preprint arXiv:2407.13131* (2024).

[37] Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024. The"Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[38] Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward Cultural Bias Evaluation Datasets: The Case of Bengali Gender, Religious, and National Identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. 68–83.

[39] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. " Jol" or" Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[40] Dipto Das and Bryan Semaan. 2022. Collaborative identity decolonization as reclaiming narrative agency: Identity work of Bengali communities on Quora. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[41] Veena Das. 2006. *Life and Words: Violence and the Descent into the Ordinary*. Univ of California Press.

[42] 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*, Vol. 11. 512–515.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[44] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[45] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2342–2351.

[46] Afia Dil. 1972. *The Hindu and Muslim Dialects of Bengali*. Stanford University.

[47] divinAI. 2020. Diversity in Artificial Intelligence: ACM FAccT 2020. https://divinai.org/conf/74/acm-facct. Last accessed: Sep 12, 2023.

[48] Paul Dourish and Scott D Mainwaring. 2012. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 133–142.

[49] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[50] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics* 9, 2 (2017), 1–22.

[51] Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of Airbnb. com. *Harvard Business School NOM Unit Working Paper* 14-054 (2014).

[52] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.

[53] Maria Eriksson and Anna Johansson. 2017. Tracking gendered streams. *Culture unbound. Journal of Current Cultural Research* 9, 2 (2017), 163–183.

[54] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[55] Oliver Falck, Stephan Heblich, Alfred Lameli, and Jens Südekum. 2012. Dialects, cultural identity, and economic exchange. *Journal of urban economics* 72, 2-3 (2012), 225–239.

[56] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.

[57] Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

[58] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)* 14, 3 (1996), 330–347.

[59] Joshua Gardner, Renzhe Yu, Quan Nguyen, Christopher Brooks, and Rene Kizilcec. 2023. Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1664–1684.

[60] Viktor Gecas. 1982. The self-concept. *Annual review of sociology* 8 (1982), 1–33.

[61] Anindita Ghoshal. 2021. 'mirroring the other': Refugee, homeland, identity and diaspora. In *Routledge Handbook of Asian Diaspora and Development*. Routledge, 147–158.

[62] Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, et al. 2024. Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1926–1939.

[63] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1914–1933.

[64] MD Romael Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. 2024. Are We Asking the Right Questions?: Designing for Community Stakeholders' Interactions with AI in Policing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[65] Christina N Harrington, Shamika Klassen, and Yolanda A Rankin. 2022. "All that You Touch, You Change": Expanding the Canon of Speculative Design Towards Black Futuring. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–10.

[66] Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2612–2623. https://doi.org/10.18653/v1/2020.emnlp-main.207

[67] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural NLP. *arXiv preprint arXiv:2203.10020* (2022).

[68] Geoffrey Hinton. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[69] Sara Hooker. 2021. The hardware lottery. *Commun. ACM* 64, 12 (2021), 58–65.

[70] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1686–1690.

[71] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1395–1417.

[72] Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering Implicit Gender Bias in Narratives through Commonsense Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3866–3873. https://doi.org/10.18653/v1/2021.findings-emnlp.326

[73] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[74] Office of the Registrar General India. 2011. Census of India: Comparative speaker's strength of Scheduled Languages. https://www.censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf. Last accessed: September 16, 2020.

[75] Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 555–560.

[76] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2023).

[77] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[78] Shafkat Kibria, Ahnaf Mozib Samin, M Humayon Kobir, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal. 2022. Bangladeshi Bangla speech corpus for automatic speech recognition research. *Speech Communication* 136 (2022).

[79] Minsu Kim and James Thorne. 2024. Epistemology of Language Models: Do Language Models Have Holistic Knowledge? *arXiv preprint arXiv:2403.12862* (2024).

[80] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, 43–53. https:

//doi.org/10.18653/v1/S18-2005

[81] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences* 117, 14 (2020), 7684–7689.

[82] Benjamin Laufer, Sameer Jain, A Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four years of FAccT: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 401–426.

[83] Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–47.

[84] Leslie McCall. 2005. The complexity of intersectionality. *Signs: Journal of women in culture and society* 30, 3 (2005).

[85] Kelly McConvey and Shion Guha. 2024. "This is not a data problem": Algorithms and Power in Public Higher Education in Canada. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

[86] Jo McCormack, Murray Pratt, and Alistair Rolls Alistair Rolls. 2011. *Hexagonal variations: diversity, plurality and reinvention in contemporary France*. Vol. 359. Rodopi.

[87] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[88] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[89] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics* (2023), 1–31.

[90] Ashis Nandy. 1988. *The intimate enemy: Loss and recovery of self under colonialism*. Oxford University Press.

[91] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 116–122. https://doi.org/10.18653/v1/2023.eacl-main.9

[92] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2 (1996), 25–42.

[93] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[94] Tolulope Ogunremi, Dan Jurafsky, and Christopher D Manning. 2023. Mini but mighty: Efficient multilingual pretraining with linguistically-informed data selection. In *Findings of the Association for Computational Linguistics: EACL 2023*. 1251–1266.

[95] G. Pandey. 2001. *Remembering Partition: Violence, Nationalism and History in India*. Cambridge University Press.

[96] Bhasa Vidya Parishad. 2001. *Praci Bhasavijnan: Indian Journal of Linguistics*. Number v. 20. Bhasa Vidya Parishad. https://books.google.com/books?id=0yxhAAAAMAAJ

[97] Robert Phillipson and Tove Skutnabb-Kangas. 2017. English, language dominance, and ecolinguistic diversity maintenance. *The Oxford handbook of world Englishes* (2017), 312–322.

[98] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4996–5001. https://doi.org/10.18653/v1/P19-1493

[99] Lindsay Poirier. 2022. Accountable Data: The Politics and Pragmatics of Disclosure Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1446–1456.

[100] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.

[101] Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. 2023. AI's regimes of representation: A community-centered study of text-to-image models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 506–517.

[102] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[103] Amit Ranjan. 2021. Language as an Identity: Hindi–Non-Hindi Debates in India. *Society and Culture in South Asia* 7, 2 (2021), 314–337.

[104] Mohammad Rashidujjaman Rifat, Dipto Das, Arpon Poddar, Mahiratul Jannat, Robert Soden, Bryan Semaan, and Syed Ishtiaque Ahmed. 2024. The Politics of Fear and the Experience of Bangladeshi Religious Minority Communities Using Social Media Platforms. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–32.

[105] Mohammad Rashidujjaman Rifat, Abdullah Hasan Safir, Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohammad Ruhul Amin, and Syed Ishtiaque Ahmed. 2024. Data, Annotation, and Meaning-Making: The Politics of Categorization in Annotating a Dataset of Faith-based Communal Violence. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2148–2156.

[106] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018).

[107] Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2024. Social bias in large language models for bangla: An empirical study on gender and religious bias. *arXiv preprint arXiv:2407.03536* (2024).

[108] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[109] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[110] Steve Sawyer and Mohammad Hossein Jarrahi. 2014. Sociotechnical approaches to the study of information systems. In *Computing handbook, third edition: Information systems and information technology*. CRC Press, 5–1.

[111] Devansh Saxena and Shion Guha. 2024. Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–32.

[112] Morgan Klaus Scheuerman and Jed R Brubaker. 2024. Products of positionality: How tech workers shape identity concepts in computer vision. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[113] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[114] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[115] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW1 (2020), 1–35.

[116] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 5412–5427.

[117] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[118] Dwaipayan Sen. 2018. *The decline of the caste question: Jogendranath Mandal and the defeat of Dalit politics in Bengal*. Cambridge University Press.

[119] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.

[120] David Sibley. 2002. *Geographies of exclusion: Society and difference in the West*. Routledge.

[121] Mrinalini Sinha. 2017. Colonial masculinity: The 'manly Englishman'and the 'effeminate Bengali'in the late nineteenth century. In *Colonial masculinity*. Manchester University Press.

[122] Dylan Slack, Sorelle A Friedler, and Emile Givental. 2020. Fairness warnings and Fair-MAML: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 200–209.

[123] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[124] Gayatri Chakravorty Spivak. 2023. Can the subaltern speak? In *Imperialism*. Routledge, 171–219.

[125] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[126] Student. 1908. The probable error of a mean. *Biometrika* 6, 1 (1908), 1–25.

[127] Heng Sun and Wan Ni. 2022. Design and Application of an AI-Based Text Content Moderation System. *Scientific Programming* (2022).

[128] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1609–1621.

[129] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[130] Latanya Sweeney. 2013. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue* 11, 3 (2013), 10–29.

[131] Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social science information* 13, 2 (1974), 65–93.

[132] Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational Biases in Norwegian and Multilingual Language Models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Seattle, Washington, 200–211. https://doi.org/10.18653/v1/2022.gebnlp-1.21

[133] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. 1987. *Rediscovering the social group: A self-categorization theory*. Oxford: Blackwell.

[134] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Shomir Wilson, et al. 2023. Nationality Bias in Text Generation. *arXiv preprint arXiv:2302.02463* (2023).

[135] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1324–1332. https://aclanthology.org/2022.coling-1.113

[136] Ashley Marie Walker and Michael A DeVito. 2020. "'More gay'fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

[137] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 196–202.

[138] Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? *arXiv preprint arXiv:2005.09093* (2020).

[139] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A Koenig, Jacob Metcalf, et al. 2017. Ten simple rules for responsible big data research.

[140] 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*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 1568–1575. https://doi.org/10.18653/v1/D16-1163

A APPENDIX

A.1 RQ1 Tables

Table 3: Power of χ^2 and Wilcoxon/t-tests comparing sentiment labels and scores assigned for different identity categories by fine-tuned models using different combinations of datasets and language models.

Identity Dimension		Gender				Religion				Nationality			
Statistical Test		χ^2	Wilcoxon/t-test			χ^2	Wilcoxon/t-test			χ^2	Wilcoxon/t-test		
ID	**Language Model**		two	left	right		two	left	right		two	left	right
D1	mBERT	0.5	**1.0**	-	**1.0**	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	-	**1.0**
D2	mBERT	**1.0**	**1.0**	**1.0**	-	**0.8**	**1.0**	**1.0**	-	0.1	**1.0**	-	**1.0**
	BanglaBERT	0.7	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-
D3	mBERT	0.2	**1.0**	-	**1.0**	0.1	**1.0**	-	**1.0**	-	0.6	0.7	-
	BanglaBERT	-	0.5	-	0.5	-	-	-	-	**1.0**	**1.0**	-	**1.0**
D4	mBERT	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	**1.0**	0.1	**1.0**	-	**1.0**
	BanglaBERT	-	0.5	-	0.7	-	**1.0**	**1.0**	-	-	0.1	0.1	-
D5	mBERT	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	0.1	-	0.2
D6	mBERT	**1.0**	**1.0**	-	**1.0**	-	0.2	0.3	-	-	0.1	-	0.1
	BanglaBERT	0.2	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	0.1	**1.0**	**1.0**	-
D7	mBERT	**0.9**	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	-	0.5	0.5	-	-	**1.0**	-	**1.0**
D8	mBERT	0.2	0.5	-	0.6	0.2	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-
D9	mBERT	-	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	0.6	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
D10	mBERT	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-
	BanglaBERT	-	0.5	0.6	-	**1.0**	**1.0**	**1.0**	-	0.2	**1.0**	**1.0**	-
D11	mBERT	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	-	-	-
D12	mBERT	-	0.3	-	0.4	-	**1.0**	-	**1.0**	-	**1.0**	**1.0**	-
	BanglaBERT	-	**0.8**	-	**0.8**	-	**1.0**	-	**1.0**	-	0.5	0.7	-
D13	mBERT	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
	BanglaBERT	-	0.2	-	0.3	-	-	-	-	**1.0**	**1.0**	**1.0**	-
D14	mBERT	-	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**	**0.9**	**1.0**	-	**1.0**
	BanglaBERT	0.1	**1.0**	**1.0**	-	0.3	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**
D15	mBERT	-	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	-	-	-
	BanglaBERT	**0.9**	**1.0**	-	**1.0**	-	**1.0**	-	**1.0**	-	0.3	-	0.3
D16	mBERT	-	**1.0**	**1.0**	-	-	**1.0**	-	**1.0**	**1.0**	**1.0**	-	**1.0**
	BanglaBERT	0.1	**1.0**	**1.0**	-	-	**1.0**	**1.0**	-	-	0.7	-	**0.8**
D17	mBERT	-	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	-	**1.0**
	BanglaBERT	-	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	-	-	-	0.1	-
D18	mBERT	**0.8**	**1.0**	**1.0**	-	0.1	**1.0**	-	**1.0**	0.5	**1.0**	**1.0**	-
	BanglaBERT	-	**1.0**	-	**1.0**	-	0.5	0.5	-	-	**1.0**	**1.0**	-
D19	mBERT	-	**1.0**	-	**1.0**	**1.0**	**1.0**	-	**1.0**	**1.0**	**1.0**	**1.0**	-
	BanglaBERT	-	**1.0**	-	**1.0**	-	-	-	-	-	**1.0**	-	**1.0**

A.2 RQ2 Tables

Each cell of these tables shows the number of fine-tuned BSA models that show bias toward identity category x that developer(s) from identity category y developed. Beside each count, we list the fine-tuned BSA models that fall into that criterion inside parentheses. To avoid repeating the base BERT models' names in the tables' cells, we used `Dxm` and `DxB`, respectively, to indicate the fine-tuned models resulting from training `mBERT` and `BanglaBERT` using the BSA dataset `Dx`.

A.3 RQ3 Tables

Table 4: Fine-tuned BSA models' bias toward gender identity categories grouped by the BSA datasets' developers' gender identities.

bias \ developer	♀	♂	♀+♂
♀	2 (D4m, D6m)	4 (D3m, D15B, D19m, D19B)	0
♂	3 (D6B, D11m, D11B)	12 (D2m, D2B, D5m, D5B, D7m, D7B, D9m, D9B, D15m, D16m, D16B, D18m)	0
no/rare	1 (D4B)	2 (D3B, D18B)	0

Table 5: Fine-tuned BSA models' bias toward religion-based identity categories grouped by the BSA datasets' developers' religious identities.

bias \ developer	🕉	☪	☪+Agnostic
🕉	0	5 (D4m, D15B, D16m, D18m, D19m)	1 (D7m)
☪	0	13 (D2m, D2B, D3B, D4B, D5m, D6B, D5B, D9m, D9B, D11m, D11B, D15m, D16B)	0
no/rare	0	4 (D3m, D6m, D18B, D19B)	1 (D7B)

Table 6: Fine-tuned BSA models' bias toward nationality-based identity categories grouped by the BSA datasets' developers' national identities.

bias \ developer	🇧🇩	🇮🇳
🇧🇩	12 (D2m, D3B, D4m, D5m, D7m, D7B, D9m, D9B, D11m, D16m, D16B, D19B)	0
🇮🇳	5 (D2B, D6B, D18m, D18B, D19m)	0
no/rare	7 (D3m, D4B, D5B, D6m, D11B, D15m, D15B)	0

Table 7: Quantified Bias Metrics (average PCM and PCR) in ten data splits.

Identity Dimension		Gender		Religion		Nationality	
ID	**Language Model**	PCM	PCR (♀, ♂)	PCM	PCR (🕉, ☪)	PCM	PCR (🇧🇩, 🇮🇳)
D1	mBERT	146.98	10, 0	104.7	0, 10	76.34	10, 0
	BanglaBERT	79.97	0, 10	180.25	0, 10	62.61	10, 0
D2	mBERT	54.12	0, 10	31.57	0, 10	38.44	10, 0
	BanglaBERT	71.82	0, 10	31.1	0, 10	37.66	1, 9
D3	mBERT	55.46	10, 0	32.92	10, 0	45.89	1, 9
	BanglaBERT	67.62	7, 3	33.23	7, 3	55.21	10, 0
D4	mBERT	92.04	10, 0	49.51	10, 0	50.44	10, 0
	BanglaBERT	33.16	3, 7	11.14	1, 9	22.15	7, 3
D5	mBERT	87.18	0, 10	47.46	0, 10	52.73	10, 0
	BanglaBERT	58.48	0, 10	39.47	0, 10	24.9	4, 6
D6	mBERT	66.12	10, 0	24.69	7, 3	58.21	9, 1
	BanglaBERT	110.49	0, 10	52.99	0, 10	81.21	0, 10
D7	mBERT	76.23	0, 10	19.66	0, 10	46.34	10, 0
	BanglaBERT	42.18	0, 10	22.43	0, 10	29.84	4, 6
D8	mBERT	42.35	0, 10	35.27	0, 10	46.51	10, 0
	BanglaBERT	54.4	0, 10	29.04	0, 10	39.76	0, 10
D9	mBERT	49.23	0, 10	64.55	0, 10	70.98	10, 0
	BanglaBERT	75.62	0, 10	44.73	0, 10	31.36	10, 0
D10	mBERT	93.7	0, 10	62.63	0, 10	60.07	0, 10
	BanglaBERT	48.51	0, 10	67.38	0, 10	67.21	0, 10
D11	mBERT	7.28	0, 10	3.8	5, 5	6.26	10, 0
	BanglaBERT	5.81	6, 4	2.62	2, 8	5.17	9, 1
D12	mBERT	26.52	3, 7	15.9	10, 1	25.9	1, 9
	BanglaBERT	37.81	9, 1	14.41	9, 1	35.1	0, 10
D13	mBERT	17.34	0, 10	9.94	2, 8	13.75	8, 2

Table 7 continued: Quantified Bias Metrics (average PCM and PCR) in ten data splits.

Identity Dimension		Gender		Religion		Nationality	
ID	**Language Model**	PCM	PCR (♀, ♂)	PCM	PCR (🕉, ☪)	PCM	PCR (🇧🇩, 🇮🇳)
	BanglaBERT	4.46	10, 0	1.59	8, 2	7.05	0, 10
D14	mBERT	118.66	0, 10	26.31	10, 0	70.33	10, 0
	BanglaBERT	108.36	0, 10	52.63	0, 10	50.43	10, 0
D15	mBERT	58.25	0, 10	28.56	0, 10	46.22	2, 8
	BanglaBERT	111.41	10, 0	38.55	10, 0	64.18	7, 3
D16	mBERT	29.79	0, 10	16.08	10, 0	67.04	10, 0
	BanglaBERT	60.6	0, 10	20.86	0, 10	36.58	9, 1
D17	mBERT	36.71	0, 10	90.19	0, 10	77.79	10, 0
	BanglaBERT	96.24	0, 10	121.86	0, 10	48.57	2, 8
D18	mBERT	36.49	0, 10	10.19	10, 0	52.87	0, 10
	BanglaBERT	59.48	9, 1	39.9	0, 10	32.27	0, 10
D19	mBERT	39.28	3, 7	30.91	10, 0	51.45	0, 10
	BanglaBERT	73.11	6, 4	30.6	0, 10	53.64	10, 0