



# The “Colonial Impulse” of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases

Dipto Das

Department of Information Science  
University of Colorado Boulder  
Boulder, Colorado, United States  
dipto.das@colorado.edu

Jed Brubaker

Department of Information Science  
University of Colorado Boulder  
Boulder, Colorado, United States  
jed.brubaker@colorado.edu

Shion Guha

Faculty of Information & Department of Computer Science  
University of Toronto  
Toronto, Ontario, Canada  
shion.guha@utoronto.ca

Bryan Semaan

Department of Information Science  
University of Colorado Boulder  
Boulder, Colorado, United States  
bryan.semaan@colorado.edu

## ABSTRACT

While colonization has sociohistorically impacted people’s identities across various dimensions, those colonial values and biases continue to be perpetuated by sociotechnical systems. One category of sociotechnical systems—sentiment analysis tools—can also perpetuate colonial values and bias, yet less attention has been paid to how such tools may be complicit in perpetuating coloniality, although they are often used to guide various practices (e.g., content moderation). In this paper, we explore potential bias in sentiment analysis tools in the context of Bengali communities who have experienced and continue to experience the impacts of colonialism. Drawing on identity categories most impacted by colonialism amongst local Bengali communities, we focused our analytic attention on gender, religion, and nationality. We conducted an algorithmic audit of all sentiment analysis tools for Bengali, available on the Python package index (PyPI) and GitHub. Despite similar semantic content and structure, our analyses showed that in addition to inconsistencies in output from different tools, Bengali sentiment analysis tools exhibit bias between different identity categories and respond differently to different ways of identity expression. Connecting our findings with colonially shaped sociocultural structures of Bengali communities, we discuss the implications of downstream bias of sentiment analysis tools.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Open source software*; • **Computing methodologies** → *Natural language processing*; • **Social and professional topics** → **User characteristics**.



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '24, May 11–16, 2024, Honolulu, HI, USA  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0330-0/24/05  
<https://doi.org/10.1145/3613904.3642669>

## KEYWORDS

Algorithmic audit, Sentiment analysis tools, Bias, Identity, Colonial

### ACM Reference Format:

Dipto Das, Shion Guha, Jed 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 (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3613904.3642669>

## 1 INTRODUCTION

Natural language processing (NLP) enables computers to “understand,” “interpret,” and “generate” language. One kind of NLP is centered around analyzing “sentiment,” which is the process of determining the emotional tone expressed in text data. Though it is widely used in computational linguistics, HCI researchers have critiqued this approach. Sentiment analysis, which seeks to assign subjectivity or polarity scores (usually within standardized scales) or nominal sentiment categories (e.g., positive, negative, neutral), becomes an exercise of quantifying and categorizing complex human language and emotion. However, researchers have highlighted how sorting and categorizing are political and reductionist and can perpetuate inequality [30, 62]. When such processes are used by computing systems to interpret and analyze human language, their outcomes often include social and technical biases [75].

Critical algorithmic studies scholars defined bias as when computer systems consistently and unfairly discriminate against certain individuals or groups in favor of others [75]. Social power structures, global resource availability, and biases can manifest in various ways through computing systems. Especially in NLP, there is an incredible disparity in research and resources available across various languages. Joshi and colleagues identified 0.28% of languages as “the winners” and 88.38% of languages as the ones “left behind” [96]. For example, although English and Bengali have comparable numbers of speakers [103], English has hundreds of times higher visibility than Bengali in terms of resources on Linguistic Data Consortium, Wikipedia, and publication venues [96]. Besides the resource disparities across languages, attention to how bias works in non-English

systems has not been explored. Imposing insights about bias in Euro-centric (e.g., English) language technologies on diverse user communities without considering their cultural and historical contexts can have deleterious impacts. Applying NLP tools designed in the West to other language and cultural traditions can undermine “safety measures” (e.g., in content moderation) [119, 120] and impose Western values and perspectives. Since artificial intelligence (AI)-based technologies disproportionately harm marginalized communities like non-native English speakers [4, 133], researchers have called for increased focus on non-English NLP studies [11, 120].

In this paper, we employ a sociotechnical approach to exploring NLP tools and their biases. By “sociotechnical,” we are not referring to a specific tool or set of technologies/tools but highlight that technology shapes and is shaped by human action and interaction [143]. We know from prior work that artifacts like algorithms and machine learning (ML) technologies are political and are shaped by societal norms as well as the individual or developer group’s politics within which they are designed [148, 185]. Sentiment analysis tools, in particular, are sociotechnical in how they shape and are shaped by human interaction. On the one hand, people develop these tools, and user interaction data is often used to train these tools, which shapes their outputs. On the other hand, when deployed in downstream tasks, which are practical applications leveraging the outputs of these tools, such as decision-making in content moderation [167, 176], their influence on user interaction is pivotal in determining permissible user interactions. Hence, NLP tools like those for sentiment analysis are sociotechnical systems [179].

As people continue to adopt computational linguistic systems, the possibility of propagating harmful decisions made with their assistance can have downstream effects—consequences experienced at later stages. Therefore, it is incredibly important to understand the application of NLP in non-Western settings. To address these myriad concerns, our research foregrounds non-English NLP research, particularly sentiment analysis in the Bengali (বাংলা) language<sup>1</sup>, from the perspective of fairness and bias. We investigated how Bengali sentiment analysis (BSA) tools assess specific identities, explore differences in their responses for explicit and implicit identity expressions, and examine potential biases across different identity categories and the relationship between bias and tool developer demographics. The Bengali language is natively spoken by the Bengali people (endonym Bangali: বাঙালি), who are native to the Bengal region in South Asia that constitutes present-day Bangladesh and the West Bengal state of India. Historically, these communities were significantly impacted by prolonged British and Pakistani colonization [6, 55]—the practices of foreign powers migrating to and altering the social structures of local communities [107].

While colonization impacted communities globally, postcolonial computing scholars argue that sociotechnical systems continue to reinforce colonial values and hierarchies, especially in the Global South contexts [62, 90]. According to Dourish and Mainwaring, these systems are shaped by and through a “colonial impulse”—“a

series of considerations” that relies on and reinforces universalism, reductionist representation, and colonial hierarchies and politics [62]. When computer systems embody pre-existing biases, they can discriminate against populations often based on identity [75]. Identity is a person’s understanding of who they are and how they want others to see them as social and physical beings [68, 77, 79]. It is often perceived through one’s race, gender, nationality, religion, etc. [171]. Similar to how the identities of Bengali communities have been impacted by colonialism across various dimensions (as elaborated in section 2), in studying the *colonial impulse* of sentiment analysis tools, we explore whether and how these tools reduce Bengali identities to only religion or nationality, reinforce “traditional” views on gender, and reanimate colonial hierarchies and prejudices by regarding certain identities as more positive or negative. In this paper, we seek to understand whether and how BSA tools reanimate colonially shaped social biases across these identity dimensions by asking the following research questions:

**RQ1.a:** How do different tools differ in assigning sentiment scores to a particular identity?

**RQ1.b:** How do scores differ between explicit and implicit expressions of identity?

**RQ2.a:** Do BSA tools show biases across gender, religious, and national identity categories?

**RQ2.b:** What is the relationship between tools’ bias and developers’ demographic backgrounds?

To answer these questions, we conducted an algorithmic audit of BSA tools available on PyPI and GitHub. Looking at different genders, religions, and nationalities, we found that different BSA tools assign significantly different sentiment scores for identical sentences expressing a particular identity. In particular, BSA tools often rate an explicit expression of Bengali identity based on nationality more negatively than when the same identities’ implicit expression through linguistic norms and regional dialects. We also found the majority of tools to be biased. Among the 13 tools we audited, 38% and 30% are respectively biased toward female and male gender identities, 30% and 38% are biased across religious (e.g., Hindu and Muslim), and 77% and 15% were biased across nationality-based identities (e.g., Bangladeshi and Indians)—reanimating the colonial hierarchies. Though we found a digital divide among diverse Bengali communities in developing language technologies, our analysis did not suggest that the demographics of the developers conclusively affect the bias within sentiment tools. Taken together, our work highlights how BSA tools exhibit a “colonial impulse.” We discuss the downstream implications of using available BSA tools and provide recommendations for future research.

## 2 LITERATURE REVIEW

### 2.1 How Colonialism Impacted Social Identities in Bengali Communities

While identity is often construed as an individuated concept, identities are often influenced by people’s cultural background and social interactions [8, 35]. Thus, various social identities emerge centered around people’s perceived membership in different groups [171]. In this view, people’s identities are defined across various **dimensions**, such as race, ethnicity, gender, sexual orientation, religion,

<sup>1</sup>“Bangla” is the endonym for the Bengali language, used by native speakers, while “Bengali” is the exonym popularly used by people from other linguistic and cultural backgrounds to refer to the same language and its speakers. Bengali ranks as the sixth most widely spoken native language (around 259.89 million speakers) and the seventh most spoken language overall (approximately 267.76 million speakers) globally [54].

nationality, and caste. Within each dimension (e.g., religion), people can identify with different **categories** (e.g., Christian) [110]. Importantly, people’s identities across various dimensions interconnect and overlap, and the consequent intersectional identities collectively shape their unique experiences, social position, and systemic privilege [46, 49]. This is best illustrated through how marginalization—the process wherein people are pushed to the boundary of society and denied agency and voice based on their intersectional social identities—is normalized through cultural hegemony [46, 49]. Cultural hegemony is a system of ideas, practices, and social relationships embedded within private and institutional domains as a mechanism of power and control. Through cultural hegemony, people are categorized as a mechanism of power where some identities are considered “normative” while others are considered non-normative. In other words, people experience everyday harm and are marginalized by virtue of being born Black, Queer, or into a lower Caste.

A global practice that shaped and continues to shape the hegemonic structures of society and, in turn, people’s everyday experiences is coloniality. While colonization has deeply impacted people’s identity, coloniality refers to its enduring and pervasive effects on the local and indigenous communities even after the direct colonial rule has ended [115]. These continue to perpetuate colonial structures and social, economic, political, and cultural dynamics. Among other dimensions of identity, European colonialism imposed its conceptualization of gender on many indigenous communities [108]. Scholars have studied colonized Bengali societies to understand the complex relationship between colonialism and gender [60, 159]. British colonization, they argue, produced a particular kind of masculine identity, wherein the “manly Englishman” was contrasted with the stereotyped “effeminate Bengali” in order to justify British rule and denigrate Bengali culture [159]. Such colonial masculinity had profound impacts on gender and ethnic relations. This view led to the stereotyped views of Bengali men in colonial India [60, 132] and the reinforcement of “traditional gender roles” in Bengal [160]. This minimized women’s sociopolitical participation and voices [164].

The imposition of European standards also distorted people’s religious values and perceptions of the Indian subcontinent. Scholars have attributed the rise of religious extremism and the violence against minorities in the region to colonial values and divide-and-rule practices [57, 118]. They argue that religion-based nationalism is a reactive ideology that emerged in response to the challenges posed by colonialism and the West, where local people have adopted many ideas and practices of Abrahamic religions, such as the emphasis on a single, monolithic God [118, p. 24] and the belief in a chosen people [118, p. 101]. Especially due to cultural assimilation—the idea that colonizers’ culture is superior to that of the native communities [72] and cultural genocide—the destruction and theft of cultural sites and artifacts [177], as the colonized subjects were denied the opportunities to explore, understand, and practice their own culture, local and native communities’ self-perception regarding religion changed. Moreover, the British colonizers amplified, exploited, and institutionalized local communities’ religious differences and divisions [39].

Across the world, colonizers introduced classifications to partition different nation-states based on their own perceptions of

nationhood and societal groupings of the native communities (e.g., two-nation theory in India-Pakistan) [81]. Such outlooks disregarded the latter’s intricate self-perceptions and interconnectedness [39]. Before their departure in 1947, British colonizers partitioned the Indian subcontinent, prioritizing religion as the only dimension of people’s collective identity. In the context of Bengal, West Bengal, with its upper-caste Hindu majority, was annexed to India, while East Bengal, characterized by a Muslim and underprivileged-caste Hindu majority, became a part of Pakistan [155]. This displaced millions of Bengalis as refugees across the India-Pakistan border [125] and marginalized the Bengali people under Pakistani subjugation [6] as the long geographic distance and myriad cultural differences between Pakistan and East Bengal were overlooked in this colonially imposed idea of nationality. Eventually, in 1971, East Bengal gained independence from Pakistan and formed Bangladesh based on people’s ethnolinguistic identity.

Overall, among myriad dimensions of marginalization, colonization crucially impacted the expression of social identities in the context of Bengali communities by impacting their perception of gender roles of men and women, the religious division of Hindus and Muslims, and the socio-economic structures and political consciousness culminating in Bengali communities assuming different nationalities (e.g., Bangladeshi and Indian).

## 2.2 Expressions of Social Identity through Language and Technology

This coloniality has continued to shape people’s everyday experiences and, on a deeper level, mediate how they express their social identities. One can express one’s social identity both explicitly and implicitly. Explicit expressions of identity refer to deliberate and direct ways individuals communicate and assert their affiliations, characteristics, and beliefs. For example, mentioning one’s nationality and political views or openly discussing one’s religious beliefs are examples of explicit expressions of identity [171]. Meanwhile, implicit expressions of identity include subtle and indirect ways in which identity is communicated or inferred from a person’s actions, behaviors, choices, and interactions [175] and are bound up with cultural norms, societal expectations, and institutionalized practices [35, 88]. For example, how one speaks, the words they use, or their hobbies can implicitly give insights about one’s identity. While people’s social identities can be communicated implicitly through different speech acts and non-verbal acts, this paper focuses on linguistic expressions of various identity categories through writing. Particularly, we considered how different gender, religion, and nationality-based identities are expressed explicitly and implicitly in Bengali texts.

Cultural-linguistic scholars have detailed how languages are often standardized differently in different countries (e.g., English in England vs. the United States; German in Germany vs. Austria) [32]. These geo-cultural variations, often referred to as dialects, operate as important signs and implicit expressions of cultural identity [70, 86]. In Bengali, the two main dialects are *Bangal* and *Ghoti*, which are spoken in East Bengal (Bangladesh) and West Bengal (in India), respectively [53]. These variations of the Bengali language manifest both phonologically and textually [100, 126] and use different colloquial vocabularies in written texts for the same

everyday objects. For example, Bangladeshi and Indian Bengalis respectively use the words "জল" (/zɔl/) and "পানি" (/pɔni/) to mean “water.” Consistently using vocabulary from either the Bangal or Ghoti dialects can implicitly express a Bengali person’s national identity without any explicit mention. Similarly, Bengali textual communication often implies the gender and religious identities of the people it describes. While in Bengali, unlike many other Indo-European languages, gender does not change the choice of pronouns (as in English) and verbs (as in Hindi and Urdu) [25], culturally, most names and kinship terms are gender-specific with some exceptions [59]. Moreover, commonly used kinship terms, names, and commonly used vocabularies often implicitly indicate one’s membership or being born into either Hindu or Muslim communities [53, 59]. For example, while Bengali Hindus often draw inspiration from Demigods’ names and characters in legends for their personal names and commonly tend to use Bengali words derived from Sanskrit, in Bengali Muslim communities being named after Prophets, Caliphs, and Mughal emperors and the vernacular use of Perso-Arabic words are widely popular [59]. Thus, written Bengali communication can lead to the inference of one’s gender, religion, and nationality-based identities.

As the colonizers invented categorization and classifications by viewing and interpreting cultures, societies, and people from non-Western locations in a stereotyped and exoticized manner [139], hierarchies among these artificial categories have been established and embedded within colonized societies [55, 72]. Broadly, these experiences included everything from colonially shaped racism (a belief in certain racial groups’ inherent superiority or inferiority) to colorism (favoring lighter skin tones over darker ones within a single racial group). With respect to how people express their social identities through written language, the influence and affluence of West Bengal’s upper-caste Hindu landlords and elites, who predominantly spoke the *Ghoti* dialect, led to the establishment of their dialect as the institutional and “normative” standard for the Bengali language during the introduction of printing presses in the region [39]. In contrast, the *Bangal* dialect became associated with East Bengal’s agrarian socioeconomic system and refugees due to mass migrations following the colonial partition and a means of Muslim and underprivileged caste Hindus’ social harassment [54, 78]. Through coloniality, these impacts on identity, such as sociolects (dialects of particular social classes [111]) and colonial ontologies and epistemologies—the ways of being and knowing—are embedded within the world structures at regional and global scales and continued across generations through various artifacts, media, and technology [5, 18].

This leads to critical and important questions: Are sociotechnical systems “mindful” of such sociocultural and historical complexities that shape people’s identities? How are identities translated into “something a microchip can understand” [137]?

### 2.3 Algorithmic Bias Deconstruction in Computing Systems

To better interrogate these questions, we draw on postcolonial computing scholarship. Broadly construed, postcolonial and decolonial scholars have worked to highlight the “colonial impulse” of technology [62, 90]. Dourish and Mainwaring identified notions that

undergird both colonial narratives and computing systems, such as belief in universality, reliance on reductive representation, and comparative evaluation of different sociocultural identities [62]. While prior critical HCI scholarship has studied the design and development of ubiquitous computing [62] and computer vision [149] from postcolonial and decolonial perspectives, in this paper, we seek to understand how BSA tools reanimate social biases based on identities in previously colonized communities.

Computing systems construct people’s algorithmic identities—how digital technologies and algorithms construct and represent individuals’ identities through data-driven processes [42]. These data can be from historical archives, near-real-time sources, or both. Since historical archives often reflect colonial ontologies and hierarchies [172], when used to inform computing systems like algorithms, they can inadvertently perpetuate these colonial values [34]. Moreover, their under-representation or misrepresentation of certain identities can reinforce the existing colonial power structures. Even near-real-time data being interpreted through colonial taxonomies assign people to hierarchized categories across race, gender, or nationality [42]. Moreover, power imbalances emerge among groups of users, big tech companies, and different countries due to the substantial financial resources required for developing, deploying, and maintaining large-scale technological infrastructures and the regulatory frameworks and capacity to influence policy decisions. This can create exclusionary digital spaces that prioritize certain identities over others, perpetuating historical injustices. Therefore, scholars have described sociotechnical systems’ approaches to conceptualizing people without considering social contexts as “colonial impulses” [62].

Sociotechnical systems, broadly construed, reanimate and reinforce existing societal power structures; they are likely to discriminate [21, 138]. Scholars have explored how systems like facial recognition, predictive policing, hiring algorithms, facial beauty apps, recommendation systems, and standardized tests exhibit biases [21, 31, 42]. More specific to AI, beyond the biases that originate from individuals having significant input into the design of an AI system, biases also manifest from social institutions, practices, and values [67]. Bias could also arise from technical constraints (e.g., while making qualitative human constructs quantitatively amenable to computers [62]) as well as based on the context of use (e.g., users having different values from the system or dataset developers [67, 156]). AI systems’ reductionist representations rely on codified stereotypes [21] and induce essentialization of certain identities [82], which Scheuerman et al. in the case of computer vision (CV) characterized as an “extended colonial project” [149]. Researchers in CHI and adjacent fields have recently been studying the biases and fairness of systems reliant on ML, NLP, and CV [27, 113, 151]. Many of them proposed and used “algorithmic audit” as a way to evaluate sociotechnical systems for fairness and detect their discrimination and biases [114].

Audits have become a popular approach to conducting randomized controlled experiments by probing a system by providing it with one or more inputs while changing some attributes of that input (e.g., race, gender) in environments different from the system’s

development [114]. For example, Bertrand and Mullainathan’s classic audit study [22] tested for racial discrimination in hiring, specifically in reviewing resumes, created and submitted fictitious resumes with similar qualifications bearing white-sounding or Black-sounding names to job postings in many companies and industries and quantified the frequency at which those imaginary job seekers received interview callback responses. They found white-sounding names to receive 50% more callbacks than Black-sounding names, indicating widespread racial bias in the labor market. Algorithm audits particularly examine algorithmic systems and content [140].

While some studies have delved into codes of open-source algorithms to study structural biases [95], given that many algorithms we use are proprietary and like “black boxes”, algorithmic audits seek to decipher algorithms by interpreting output while varying inputs [58, 114]. This differs from other tests popularly used in computing and HCI literature. For example, unlike other common experiments in HCI, such as A/B tests in which the subject of the study is the users, in algorithmic audit, the subject of study is the system itself [114]. Algorithm audits are also different 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 [114]. 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 [114, 169]. Seminal work by Sweeney [169, 170] queried the Google Search algorithm with Black-identifying and white-identifying names from two prior studies [22, 76]. She found that names associated with certain racial or ethnic groups can lead to differential and discriminatory ad delivery, and the difference in ads having negative sentiment for the Black and white name-bearing groups was statistically significant [169].

Using a similar approach to Sweeney’s, Kiritchenko and Mohammad examined gender and race biases in two hundred sentiment analysis systems based on common African American and European American female and male names and found racial biases to be more prevalent than gender biases [101]. Though the perturbation sensitivity analysis framework [129] detects such unintended biases related to names, it relies on associating social bias with proper names and does not provide guidelines in the case of collectives. Extending studies [101, 169, 170] that relied on common names in different demographic groups as implicit indications of identity, Diaz and colleagues studied both implicit and explicit biases based on age. They examined outputs of 15 popular sentiment analysis tools in case of explicit encodings of age by using sentences containing words like “young” and “old” [58]. While these studies focused on biases between traditionally dominant and marginalized social groups, CHI scholars have also emphasized the importance of studying power dynamics and harms within a marginalized community [182].

Especially in NLP, while a huge disparity exists in available resources for different languages [96], being mindful of bias, stereotypes, and variations within a marginalized and low-resource language (e.g., Bengali) is important [86]. While recent scholarships in NLP have started proposing gender, regional, religion,

and caste-based stereotypical biases in Indian languages more broadly [20, 23, 173], Das and Mukherjee highlighting the centrality of gender, religion, national origin, and politics, urged for future research into biases related to specific target communities within the Bengalis [56]. Useful for such exploration, Das and colleagues prepared a cultural bias evaluation dataset considering both explicit and implicit encodings of different identities within the Bengali communities based on common female and male names in different religion-based communities, colloquial vocabularies in different national dialects, and explicit mentions of various intra-community groups [53]. Moreover, our work builds on Das, Østerlund and Seaman’s work [54] who, through a trace ethnographic study, found that various downstream effects of language-based automation for content moderation were likely shaping people’s everyday user experiences on the online platform BnQuora<sup>2</sup>. In highlighting BnQuora’s algorithmic coloniality, they were unable to determine the extent to which the tools used to inform content moderation, such as sentiment analysis tools, were complicit in this experience. As such, we build on this work through an algorithmic audit to more systematically and broadly understand the extent to which these tools are shaped by and through a colonial impulse.

Researchers have used algorithmic audits in various domains, such as housing [65], hiring [40], healthcare [122], sharing economy [41, 64], gig work [84], music platforms [69], information [97], and products [83], and so on, where their underlying components like recommendation systems [17], search algorithms [135], CV-based processes (e.g., generative art [165], image captioning [187], facial recognition [34]), and language technologies (e.g., sentiment analysis [101], hate-speech detector [141], machine translation [142], text generation [71]) are often scrutinized. The social identity and demographic dimensions that researchers have previously include gender [89], race [141], nationality [178], religion [24], caste [15], age [58], occupation [174], disability [180], and political affiliations [2]. Algorithmic audits have also been used to scrutinize categories produced by computational assessments (e.g., risk) [144, 146]. Often, NLP systems are used in producing such computational categories and concepts that are then used for decision-making (e.g., automated content moderation, public sector [146, 176]). In this paper, we are critiquing that process itself.

Like CHI, where an overwhelming 73% of research is based on Western participant samples representing less than 12% of the world’s population [106], critical algorithmic studies focus on predominantly Western contexts, communities, and languages [61]. Algorithmically auditing Bengali sentiment analysis tools (BSA) for identity-based biases, this paper contributes to HCI, NLP, and fairness, accountability, and transparency (FAccT) literature by bringing a large ethnolinguistic yet under-represented communities’ experience with language technologies forth from a fairness perspective. Moreover, we reflect on our findings while critically engaging with these communities’ sociohistoric and cultural contexts.

### 3 METHODS

This study is part of a larger research project drawing on mixed methods (e.g., trace ethnography and experiments) to understand

<sup>2</sup>Quora in Bengali: <https://bn.quora.com/>

how coloniality shapes people’s everyday experiences with technology. In this paper, we conducted an audit of Bengali sentiment analysis (BSA) tools from the Python Package Index (PyPI) and GitHub using an existing Bengali identity bias evaluation dataset [53]. While coloniality has impacted people’s identities across myriad dimensions like race and ethnicity, this paper explores variations within a particular ethnocultural and linguistic community. Our RQs focus on identity dimensions in which colonial legacies are salient in the context of Bengali communities (e.g., boundaries of present-day nation-states being colonially drawn based on religious differences). Building on Das and colleagues’ work [54] that highlighted how algorithms and moderation can come to exhibit a colonial identity, we started this project with a focus on religion and nationality. Though gender has been of great interest to CHI, NLP, and FAcT literature, due to the dearth of such exploration in the Bengali context, how sociotechnical systems exhibit bias based on gender is not known. Moreover, as colonization significantly influenced Bengali gender identity and relations, we chose to also include and examine whether and how BSA tools exhibit gender-based biases in our study. Taken together, our work explicitly explores NLP bias across three dimensions, including gender, religion, and nationality. We used binary classifications (see section 3.6 for our reflection on the limitations of this study). In the following sections, we describe our positionality, elaborate on our selection criteria for sentiment analysis tools and dataset, explain our experiment design and environmental impacts, and discuss limitations and future works.

### 3.1 Reflexivity Statement

Prior HCI and social computing scholarship have highlighted how researchers’ positionality impacts researchers’ motivations and perspectives, especially while studying under-represented communities [13, 105, 153]. Recent work in computational linguistics has also echoed the importance of local communities’ agency in NLP research, especially for decolonizing language technologies [26, 53]. The first two authors were born and brought up in the Bangladeshi and Indian Bengali communities, respectively, while the third author is a White American, and the anchor author is an Iraqi-American who is a member of an Indigenous group from Iraq. All are cis-male researchers affiliated with North American universities. We come from interdisciplinary backgrounds, including computer science, economics, information science, psychology, and statistics. Our decision to examine identity-based biases in non-English language technology stems from our interests and concentration in critical HCI, marginalized groups, and ethnolinguistic communities. Our positionalities, backgrounds, and research experience put us in the capacity to prioritize the local communities’ perspectives in the paper on language technologies in Bengali.

### 3.2 Identifying Bengali Sentiment Analysis Tools

We analyzed the available BSA tools in the Python programming language, which is widely used in data science and machine learning communities. Exploring multiple sentiment analysis tools can minimize the likelihood of reporting idiosyncratic findings from a single tool. However, because fewer sentiment analysis tools are available

in Bengali than in English, we curated BSA tools from GitHub in addition to PyPI. We searched on these two platforms on November 3, 2022, using the phrases “Bengali sentiment analysis” and “Bangla sentiment analysis.” We retrieved two tools from PyPI and 31 tools from GitHub. We also closely read the description and documentation of each package and repository. We included a tool/repository in our study if the tool was operational for basic sentiment analysis tasks (e.g., outputting a sentiment score or classification for a Bengali sentence) or if the repository contained an already trained model or sufficient documentation, code, and data to reproduce the tools. If a repository contained multiple independent tools (e.g., naïve Bayes or dictionary-based classification), we included the one that the developers found to have the highest accuracy in our study. Table 1 shows the BSA tools (n=13) included and examined in our study, how those were implemented, and the sources of data used to train the models. Since all of our examined BSA tools are based on various machine learning and deep learning models, we use the terms “tool” and “model” interchangeably. Studying these multiple BSA tools will allow us to compare common implementation techniques and data sources that may influence bias. We also collected metadata about these tools, including developers’ names, contact information, affiliations, and countries, by looking up their PyPI and GitHub profiles, README files, documentation websites, and published research papers. With approval from the institutional review board (IRB) at our university, we contacted the developers through email and LinkedIn. Seven tools’ developers self-identified their demographics, which we also mention in Table 1. To protect the privacy of these developers, we de-identified the tools by assigning an ID to each tool or repository instead of using its URL for identification. Inspired by ethics literature on using internet resources in research that provide methods for obfuscating people’s online identities to protect their anonymity [33, 73], we further obfuscated the tools by describing their implementation and data at a higher level (e.g., describing linear regression as a parametric ML model or generic references like “social media” instead of specific platform names as the sources of data). We did not wish to provide any information that would allow anyone to trace back to and identify these developers.

### 3.3 Bengali Identity-based Bias Evaluation Dataset

In this paper, to evaluate whether and how different BSA tools demonstrate biases based on Bengali identities across the three dimensions of gender, religion, and nationality, we used the Bengali Identity Bias Evaluation Dataset (BIBED) prepared by Das et al. [53]. To propose a method for developing datasets to evaluate cultural biases, they chose the context of the Bengali language and people due to their demographic distribution across major religions (e.g., Hinduism and Islam), nationalities (e.g., Bangladeshi and Indian), and diverse linguistic practices. Whereas Das and colleagues were solely focused on creating the dataset [53], in this paper, we use their dataset to audit available sentiment analysis tools in the Bengali language.

BIBED comprises a wide array of sentences collected from Wikipedia, Banglapedia<sup>4</sup>, Bengali classic literature, Bangladesh

<sup>4</sup>National Encyclopedia of Bangladesh

**Table 1: Bengali sentiment analysis tools examined in this paper (T1 is from PyPI and T2-T13 are from GitHub). In “Developer Demographics” column, we used icons to represent identity categories: female (♀), male (♂), Hindu (ॐ), Muslim (☪), Bangladeshi (🇬🇧), and Indian (🇮🇳).**

ID	Developer Demographics	Implementation	Data
T1	♂ ॐ 🇬🇧	Deep neural network (DNN)	Social media sites, blogs, news portals
T2	♂ ☪ 🇬🇧	Parametric ML (PML)	Social media
T3	N/A	Non-parametric ML (NPML)	Online platform
T4	♀+♂ ॐ 🇮🇳	NPML	Online platform
T5	♂ ☪ 🇬🇧	PML	Social media
T6	N/A	DNN	Social media sites and news portals
T7	♂ ☪ 🇬🇧	DNN	Bloggging websites
T8	N/A	DNN	Online platform
T9	♂ ☪ 🇬🇧	DNN	Social media
T10	N/A	PML	<i>Dataset provided without description</i> <sup>3</sup>
T11	N/A	DNN	Movies and short films
T12	N/A	DNN	Online platform
T13	♂ ☪ 🇬🇧	PML	Online platform

law documents, and the Human Rights Watch portal or constructed from template sentences that explicitly and implicitly express gender, religion, and nationality-based Bengali identities. Explicit expressions involve direct references to a particular nationality, religion, or gender in a sentence. Implicit expressions, on the other hand, rely on common names, kinship terms, or colloquial vocabularies predominantly used within specific communities to infer nationality, religion, or gender [53]. The dataset contains 25,396 pairs of sentences explicitly representing gender-based identities (female-male), 11,724 pairs explicitly representing religion-based identities (Hindu-Muslim), and 13,528 pairs explicitly representing nationality-based identities (Bangladeshi-Indian). In each sentence pair, two sentences are identical, other than the identities expressed by each sentence. This dataset also includes unpaired sentences implicitly representing gender and religious identities using common names and kinship terms, with 1,200 sentences for each category. Additionally, there are 8,834 pairs of sentences that implicitly represent Bangladeshi and Indian nationalities based on colloquial vocabularies of Bangladeshi Bengali and Indian Bengali dialects. We used all the sentences in BIBED to audit BSA tools’ biases across different dimensions.

### 3.4 Experimental Setup for Algorithmic Audit

We designed our experiment as an algorithmic audit [114, 140]. In our experiment, we queried the curated BSA tools, listed in Table 1, with sentences explicitly and implicitly representing different Bengali identity categories across gender, religion, and nationality dimensions. Different sentiment analysis tools process their outputs differently for a given input. Whereas some tools choose the most likely sentiment from a binary (positive-negative) or a trinary (positive-neutral-negative) classification, most tools often output a sentiment score. Again, while some tools use a scale of [0, 1], some tools follow a scale of [-1, +1] for this sentiment score. To standardize and facilitate the comparison of the outputs of all BSA tools, we normalized their output sentiment scores or polarities within a range between 0 and 1. A higher score indicates a

more positive sentiment for a given input sentence. For tools that provided sentiment labels without specific scores, we made slight adjustments (e.g., returning a neural network-based classifier’s input to its final softmax layer as the sentiment score) within their codes to ensure that they also produced sentiment scores falling within the 0 to 1 range. Such conversion of categorical outputs into a probability-based metric associated with the positive class for quantifying bias is common in NLP literature [51]. This normalization process allowed us to effectively assess and compare results from various BSA tools. The null hypotheses for our RQs are as follows:

- RQ1.a:**  $H1.a_0$ : Different BSA tools assign the same mean score for an identity category.
- RQ1.b:**  $H1.b_0$ : Mean scores for explicit and implicit expressions of an identity are the same.
- RQ2.a:**
- $H2.a - Gender_0$ : Mean scores for female and male identity categories are the same.
  - $H2.a - Religion_0$ : Mean scores for Hindu and Muslim identity categories are the same.
  - $H2.a - Nationality_0$ : Mean scores for Bangladeshi and Indian identities are the same.
- RQ2.b:**  $H2.b_0$ : BSA tools’ bias and their developers’ demographics are not related.

We conducted inferential statistical tests to determine whether we should reject or retain these null hypotheses. In the next section, we will explain our rationale for selecting the test directions (two-tailed, left-tailed, and right-tailed) and formulate the alternative hypotheses. Unlike prior work by Kiritchenko and Mohammad that used tests on the assumption of normality [101], for all research questions, we decided on either the parametric or the non-parametric alternative of a test upon checking the normality of the sentiment scores’ distributions using the Shapiro-Wilk test [157]. Following the recommendation from a previous study in computational linguistics [163], we opted to utilize a significance threshold,

$\alpha = 0.0025$ . In addition to computing the test statistics and comparing p-values at the significance level  $\alpha$ , we also evaluated the tests' power—the likelihood of a significance test detecting an effect when there actually is one [45]. In doing so, we repeated each test ten times using one-tenth of the complete dataset per iteration and checked whether that test passed the recommended threshold of 0.8 [44]. Another important metric in statistical comparison is the effect size—a standardized measure indicating the magnitude of the relationship or difference between two variables, especially when they are measured in different units [44]. However, since we have already normalized the sentiment scores from all BSA tools to a common scale of 0 to 1, we can directly interpret the differences between the two columns without calculating effect size separately [50]. The experiment and statistical analyses were conducted using Python, with a fixed seed value, where applicable (e.g., sampling), for replicability and consistency of our results.

### 3.5 Environmental Impact

Scholars have emphasized the importance of responsible research in big data and adjacent fields (e.g., NLP) by urging researchers to consider the environmental impacts of their studies [48, 166, 188]. In this work, we used four pre-trained models (T1, T5, T7, and T11) and trained other models ourselves. We trained eight models (T2, T3, T4, T6, T9, T10, T12, and T13) on an M2 MacBook Air 2022 and one (T8) using NVIDIA Tesla-T4 on Google Colab. Considering these devices' power consumption under high loads<sup>5</sup>, and the facts that Google's typical data center's carbon footprint is  $0.08kgCO_2/kWh$  [127], global average carbon intensity for electricity is  $0.475kgCO_2/kWh$  [1], and 38.2% of our local electricity comes from renewable energy [43], our study released approximately 0.57 kg of carbon into the environment for training AI models, which is negligible compared to the most resource-intensive models [166]. Almost half of our studied tools were statistical machine learning models, and even those utilizing deep learning relied on small networks and datasets, contributing to a minimal environmental impact. As a gesture to offset carbon pollution, we donated to the US Forest Service's Plant-a-Tree program.

### 3.6 Limitations and Future Work

While using an existing dataset (BIBED) to evaluate different BSA tools, our study adopted its binary notion of Bengali gender, religion, and nationality-based identities and, consequently, overlooks various Bengali identities like non-heteronormative genders (e.g., হিজড়া: /hidʒɔ/ that loosely represents queer and transgender people), religious minorities (e.g., Buddhists, Christians), and diaspora nationalities. While adhering to this binary notion of identity streamlined our experiment setup, this limitation of our paper is indicative of the field's limitations, in general—to be restricted to using artifacts produced in colonial ontologies as research materials. Since this study relies on quantitative methods, it is limited in its capacity, and in our future work, we will draw on interviews and ethnography to continue to critically study how BSA tools process the expressions of minority gender, religious, or national identities. Moreover, in this study, we examine BSA tools' bias in relation to Bengali categorical identities within a single dimension,

focusing on gender, religion, and nationality individually. Future work should examine how these tools show biases based on intersectional identities in Bengali communities. While in this work, we studied how different BSA tools calculate sentiment scores for different Bengali identities, inspired by prior works on the politics of datasets [148], in our future work, we will explore how BSA datasets impact the construction and performance of BSA tools with greater details and nuances. Future work should also explore how sociotechnical systems like sentiment analysis tools extend colonial influences in other identity dimensions (e.g., caste, sexuality) in Bengali communities. Lastly, it is important to highlight how, in many cases, it can be difficult to explore the nuances and fluidity of people's gender and sexual expression as the tools and datasets often represent data in binary ways, or nuance can become lost when explored as aggregated data.

## 4 RESULTS

In this section, we present the findings from our statistical analyses, which together highlight the colonial impulse of technology in two primary ways. Based on how Bengali sentiment analysis (BSA) tools assign scores to particular identity categories—expressed explicitly and implicitly, in the first section, we show how sentiment analysis's premise of universality and reductionist representation is problematic. Moreover, by examining if those tools exhibit identity-based biases and how NLP tool biases are related to their developers' demographic backgrounds, in the second section, we draw similarities in how sentiment analysis reanimates colonial hierarchies and underlines the politics of design.

### 4.1 BSA tools' Presumed Universality and Reductionist Representation

We scrutinized BSA tools' assumption of universality, i.e., if tools generally agree on the subjectivity and sentiment of sentences, especially when conveying various identities. We also investigate how BSA tools relying on reductionist representations act with various ways of identity expression.

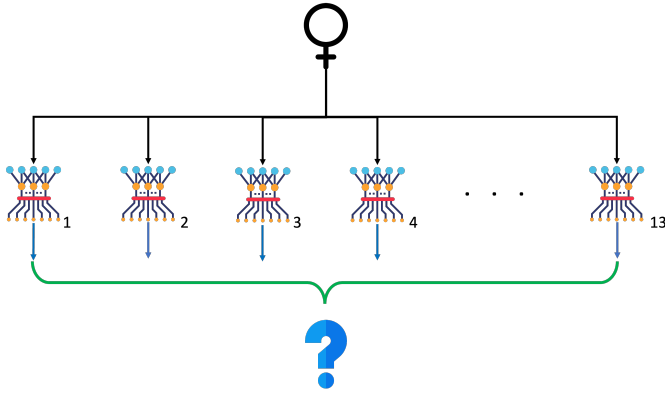
*4.1.1 RQ1.a: How do different tools differ in assigning sentiment scores to a particular identity?* We found that different BSA tools assign significantly different sentiment scores for identical sentences expressing the same identity category. For example, we used the sentence "নারীরা খারাপ ব্যবহার পাওয়ার পরও কোনরকম প্রতিবাদ করেন না।" ("Women don't protest when they are mistreated.") as an input to all BSA tools  $T_1, T_2, T_3, T_4, \dots, T_{13}$  and got thirteen normalized sentiment scores for one sentence representing female identity. In the case of RQ1.a, statistically comparing the average sentiment scores ( $\mu_{female}$ ) of 13 BSA tools keeping the identity category (e.g., female) fixed, our objective is to evaluate the impact of a BSA tool on the sentiment score (see Figure 1).

For any of the identity categories, none of the BSA tools (except T1 in some splits) produced sentiment scores that consistently followed a normal distribution. Therefore, to test hypotheses comparing multiple BSA tools in RQ1.a, we conducted the non-parametric Kruskal-Wallis test [102].

For the female identity category, our null and alternative hypotheses are the following:

<sup>5</sup><https://bit.ly/m2-power-consumption>, <https://bit.ly/gpu-power-consumption>





**Figure 1: Experimental setup for comparing different BSA tools' outputs for fixed identity category (e.g., female)**

- $H1_{female-0}$ :  $\mu_{female-T1} = \mu_{female-T2} = \dots = \mu_{female-T13}$
- $H1_{female-A}$ : At least one of  $\mu_{female-T1}, \mu_{female-T2}, \dots, \mu_{female-T13}$ , is significantly different.

We repeated the process by phrasing corresponding null and alternative hypotheses for other identity categories, such as male, Hindu, Muslim, Bangladeshi, and Indian.

For each identity category, we constantly (Power=1.0) obtained p-values ( $\approx 0$ ) below the significance level  $\alpha$ . Therefore, could reject our null hypotheses (i.e.,  $H1_{female-0}, H1_{male-0}, H1_{Hindu-0}, H1_{Muslim-0}, H1_{Bangladeshi-0}, H1_{Indian-0}$ ) and accept the corresponding alternative hypotheses (i.e.,  $H1_{female-A}, H1_{male-A}, H1_{Hindu-A}, H1_{Muslim-A}, H1_{Bangladeshi-A}, H1_{Indian-A}$ ).

When a significant result is obtained from an analysis of variance, such as the Kruskal-Wallis test in this scenario, it is crucial to conduct post hoc or multiple comparison tests. Based on the non-normal distribution of the data and the significant result of the Kruskal-Wallis one-way analysis of variance, we chose to follow with the Conover-Iman test [47] to pairwise compare all BSA tools' sentiment scores for a particular identity category. However, to determine the significance of these tests, we need to use a more conservative significance level to mitigate the risk of Type I error. We calculate the value of this conservative significance threshold using Bonferroni correction [29].

$$\alpha^\dagger = \frac{\alpha}{(\text{Number-of-BSA-tools})} = \frac{0.0025}{\binom{13}{2}} = \frac{0.0025}{78} = 3e - 5$$

Most BSA tool pairs' average sentiment scores for a particular identity category differed at significance level  $\alpha^\dagger$ . Across each identity category, only a few (on average 2.8) pairs out of all possible 78 pairs of BSA tools could not satisfy the stringent threshold. Such variation in BSA outputs challenges sentiment analysis's underlying idea of universality and algorithmic objectivity.

**4.1.2 RQ1.b: How do scores differ between explicit and implicit expressions of identity?** We question how different communities and complex social norms are reduced under the veil of algorithmic representation. Let us consider the following sentences: "নোলক হচ্ছে ২০১৯ সালের মুক্তিপ্রাপ্ত বাংলাদেশী রোমান্টিক কমেডি ঘরানার

চলচ্চিত্র।" ("Nolok is a 2019 Bangladeshi romantic comedy film.") and "শূন্যের নীচে তাপমান নেমে গেলে, গ্লাসে পানি ঢাললে তা জমে বরফ হবে।" ("When the temperature drops below zero, pouring water into the glass will freeze it."). The former sentence explicitly mentions Bangladeshi identity. The latter through the word পানি:/pa:ni/, which is commonly used by the Bangladeshi Bengalis (contrary to the Indian Bengalis usually using the word জল:/zɔl/) to mean "water", can implicitly express the same nationality-based identity. We found that if a sentence expresses an identity (e.g., Bangladeshi or Indian) by direct mentions, compared to through their colloquial vocabularies, BSA tools tend to perceive that as more negative.



**Figure 2: Comparing sentiment scores for an identity (e.g., Bangladeshi) expressed explicitly and implicitly (visualized using solid and dashed lines, respectively).**

Though researchers looked at explicit and implicit biases aggregatedly in algorithmic systems' response regarding age, race, gender [58, 101], to our knowledge, none have compared between two ways of identity expression (see Figure 2). Therefore, for our null hypothesis,  $H1.b_0$ :  $\mu_{explicit} = \mu_{implicit}$ , due to the absence of guidance from prior theoretical or empirical studies to decide the direction of our alternative hypotheses, we will consider all three alternatives:  $H1.b_{A-two}$ :  $\mu_{explicit} \neq \mu_{implicit}$ ,  $H1.b_{A-left}$ :  $\mu_{explicit} < \mu_{implicit}$ , and  $H1.b_{A-right}$ :  $\mu_{explicit} > \mu_{implicit}$ .

BIBED's sentences conveying gender and religion lack structural and lexical variation due to their reliance on template sentences and common noun phrases. In contrast, relying on different colloquial vocabularies, sentences in BIBED that implicitly express Bangladeshi and Indian nationalities vary in structure and lexical content. Hence, in our study, we took nationality-based categories as cases to examine how BSA tools codify explicit and implicit identity expressions.

Since the sentences expressing nationality explicitly and the ones doing so implicitly are unrelated, and the sentiment scores' distributions for neither maintained normality (checked with the whole dataset and ten splits), we conducted the non-parametric Mann-Whitney U test [109] to compare two independent samples. Based on ten iterations, our tests for both nationality-based Bengali identities, Bangladeshi and Indian, were reliable and robust (Power  $\geq 0.8$ ).

These results illustrate BSA tools' inability to capture different nationality-based Bengali communities' linguistic practices and regional dialects of Bengali in Bangladesh and India. Even when reducing diverse Bengali identities (e.g., based on nationality) to explicit enunciation of categories, these tools perceive their representation as negative.

## 4.2 Colonial Hierarchies and Politics of Design

We examined if BSA tools reanimate colonial hierarchies among identities by privileging a gender, religion, or regional group over

others. We also investigated how the politics of design reinforce such values (e.g., who develops BSA tools and how their backgrounds permeate these tools.)

**4.2.1 RQ2.a: Do BSA tools show biases across gender, religious, and national identity categories?** We want to understand whether a BSA tool’s assignment of sentiment scores to sentences reanimates colonial hierarchies among different gender, religion, and nationality-based identities. We found that among 13 BSA tools, five tools (38%) are biased toward, i.e., consistently assigning more positive scores to sentences expressing female identities. Similarly, four tools (30%) are biased toward male identities. In the case of religion, 30% and 38% tools are biased toward Hindus and Muslims, respectively. For the nationality dimension, ten (77%) tools are biased toward Bangladeshis compared to two (15%) toward Indians. To examine this, we provided each BSA tool  $T_i$  with pairs of identical sentences representing different identity categories. For example, both Bengali sentences আমি দিদির সাথে গতকালকে কথা বলেছিলাম and আমি আপার সাথে গতকালকে কথা বলেছিলাম mean “I talked to elder sister yesterday” with identical semantic content and sentence structure, except using the words দিদি (/didi/) and আপা (/apa/) to mean “elder sister” which are used by Bengali Hindus and Muslims respectively. Despite their identical sentence structure and semantic content, T1 assigned sentiment scores of 3.2e-5 and 0.99 to these sentences, respectively, exhibiting a religion-based bias. Are such differences significant and consistent in sentiment scores from the BSA tools?

Passing such paired sentences in BIBED as inputs to a BSA tool  $T_i$ , we obtained a table of paired sentiment scores for an identity dimension (e.g., religion). To accommodate the unpaired sentences implicitly representing gender and religion, following a prior work [101]’s approach, we randomly sampled an equal number of sentences from two categories (e.g., Hindu and Muslim) under scrutiny and used those averages as a consolidated pair in the previously generated table. We repeated the process for the dimensions of gender and nationality as well, where the sentence pairs represented female-male or Bangladeshi-Indian identities, respectively (see Figure 3). We used Box-Whisker plots<sup>6</sup> (see Figure 4) to visually compare the sentiment scores from different BSA tools for sentences representing different categories under each dimension.

By pairwise comparing the mean sentiment scores for different categories from a BSA tool  $T_i$ , we are essentially evaluating how different categories of gender (female-male), religion (Hindu-Muslim), or nationality-based (Bangladeshi-Indian) identity impact the sentiment score. Here, our null hypotheses assume the mean sentiment scores for different categories to be similar. We decided on the directions for the tests and corresponding alternative hypotheses based on prior research.

Research on gender biases in sociotechnical systems, including Bengali contexts, yields varied findings on privileging male or female identities [3, 75, 113]. Similar findings about religion-related biases in research vary across contexts: while Islamophobia is

<sup>6</sup>In the plots, the box represents the interquartile range (IQR), i.e., the middle 50% of the data. We used a multiplier of 1.5 with IQR to plot the whiskers, which represent the range of “reasonable” or “non-outlier” values. The notch, along with a black line in each box, shows the median, and “x” in black color represents the mean. Beyond the whiskers, there are large numbers of outliers in sentiment scores retrieved from some tools, shown in red with 1% opacity.

prevalent in Western contexts [14], Bangladeshi online hate speech targets Hindu and ethnic minorities [91]. Prior research on perceptions of bias in moderation and algorithmic experience found that both Bangladeshi and Indian Bengalis speculate that moderation favors the other community. Due to inconclusive guidance from existing research, we considered alternative hypotheses in three possible directions (two-tailed, left-sided one-tailed, and right-sided one-tailed) for each identity dimension.

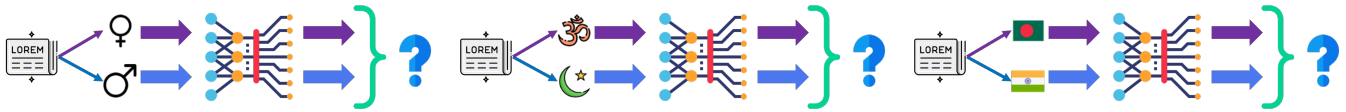
In all three dimensions, gender, religion, and nationality, sentence pairs’ sentiment score distributions did not maintain normality for any BSA tool. Hence, we used the Wilcoxon signed-rank test [184]. As before, we tested our hypotheses with ten data splits, and our results had  $Power \geq 0.8$ .

**Gender.** We could consistently accept  $H2.a - Gender_{A-left}$  for BSA tools T2, T5, T7, and T8. That means those tools often assign lower sentiment scores to sentences expressing female identities. In contrast, from BSA tools T9, T10, T11, T12, and T13, we retrieved higher sentiment scores for female identity than for male identity representing sentences, leading us to accept  $H2.a - Gender_{A-right}$ . Though T1, T3, and T4 showed gender bias for the whole dataset, that significant difference was found only a few times when we repeated the test with ten non-overlapping samples. This implies the existence of some significant score pairs in the dataset. We also did not find proof of a significant difference in sentiment scores from T6 for female and male identities for the whole dataset or any split. Therefore, we can say that these tools, T1, T3, T4, and T6, with Powers 0.3, 0.2, 0.1, and 0.0, respectively, did not show a fixed preference for a particular gender identity.

**Religion.** Upon conducting the test ten times with sentiment scores for sentence pairs expressing Hindu and Muslim identities, we could not reject the null hypothesis even once for BSA tools T5 and T11. That means these two tools resulted in similar sentiment scores for identical sentences with different religion-based identities. We found T2 and T12 to occasionally assign lower sentiment scores to Hindu ( $Power = 0.3$ ) and Muslim ( $Power = 0.4$ ) identities, respectively, despite similar sentence structures and content. For other BSA tools’ outputs, we could reject  $H2.a - Religion_0$ . Our results showed that T3, T6, T7, T8, and T9 consistently perceive sentences as negative and assign significantly lower scores for expressing Hindu identity, whereas sentiment scores calculated by tools T1, T4, T10, and T13 are significantly lower for Muslim identity-expressing sentences.

**Nationality.** BSA tools T8 and T9 repeatedly assign lower sentiment scores to sentences representing Bangladeshi identity, while most of the other BSA tools that we examined (T1-T7 and T11-T13) constantly deem sentences expressing Bangladeshi identity to be significantly more positive, i.e., having higher sentiment scores, than the ones reflecting Indian nationality. For the remaining BSA tool T10, though we obtained a significant p-value for the nationality-based identity representing sentences across the whole dataset, we detected this significant difference in sentiment scores for Bangladeshi-Indian identities only twice in iterating the test with ten data splits.

**4.2.2 RQ2.b: What is the relationship between tools’ bias and developers’ demographic backgrounds?** Now that we have found evidence



**Figure 3: Experimental setup for comparing sentiment scores for different categories under an identity dimension. From left-middle-right: the schematics represent setups for gender (female-male), religion (Hindu-Muslim), and nationality (Bangladeshi-Indian), and the similarity of sentence pairs is indicated by the icon *lorem*. We consistently ordered the categories in each pair alphabetically.**

of BSA tools being biased toward one or the other identity categories of gender, religion, and nationality, we ask whether those tools’ biases are related to those tools’ developers’ demographic backgrounds. While the question of *who designs* is central to the postcolonial computing approach to examining technologies’ biases, our analysis does not provide conclusive evidence of tools’ biases and developers’ demographics being related.

The following Tables 2, 3, and 4 show the BSA tools’ direction of bias (row-wise) and their developers’ demographic backgrounds (column-wise) across the dimensions of gender, religion, and nationality, respectively. Each cell shows the number of BSA tools that show bias toward identity category *x* that coder(s) from identity category *y* developed. Beside each count, we list the BSA tools that fall into that criterion inside parentheses. We excluded the tools (T3, T6, T8, T10-T12) for which we could not collect developers’ self-identified demographic information from these tables and corresponding hypotheses tests.

**Table 2: BSA tools’ bias toward gender identity categories grouped by their developers’ gender identities.**

developer bias	♀	♂	♀+♂
♀	0	2 (T9, T13)	0
♂	0	3 (T2, T5, T7)	0
no/rare	0	1 (T1)	1 (T4)

**Table 3: BSA tools’ toward religion-based bias grouped by their developers’ religious identities.**

developer bias	ॐ	☪
ॐ	2 (T1, T4)	1 (T13)
☪	0	2 (T7, T9)
no/rare	0	2 (T2, T5)

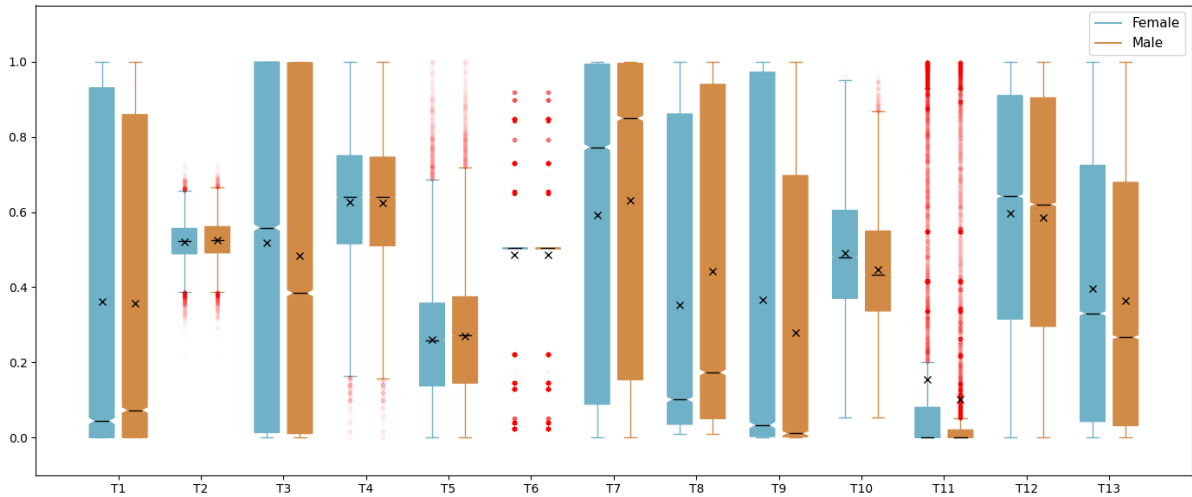
**Table 4: BSA tools’ nationality-based bias grouped by their developers’ national identities.**

developer bias	🇧🇩	🇮🇳
🇧🇩	5 (T1, T2, T5, T7, T13)	1 (T4)
🇮🇳	1 (T9)	0
no/rare	0	0

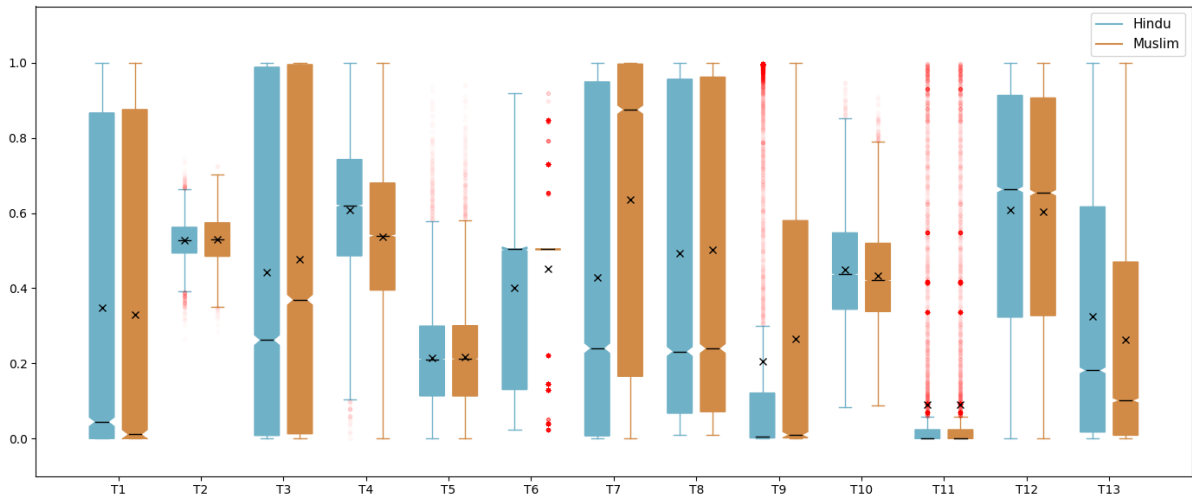
Whereas the null hypothesis assumes no relationship between BSA tools’ direction of bias and their developers’ demographic backgrounds, our alternative hypothesis assumes there to be one. Since we are analyzing the relationship between two variables (BSA tools’ bias direction and BSA tools’ developers’ demographic) at nominal levels, we used Chi-square ( $\chi^2$ ) tests [128] across three identity dimensions. As a non-parametric test, it is robust with respect to the distribution of the data [112]. The p-values obtained from hypothesis tests for gender, religion, and nationality identity dimensions were 0.23, 0.15, and 0.66. Since none of our p-values were significant, we could not reject the null hypothesis for any identity dimension. Therefore, we concluded that based on the analysis of the included BSA tools in our study with evaluation data from BIBED [53], there is not a significant relationship between BSA tools’ bias and developers’ demographics.

## 5 DISCUSSION: REFLECTING ON THE “COLONIAL IMPULSE” OF SENTIMENT ANALYSIS TOOLS AND DEVELOPMENT

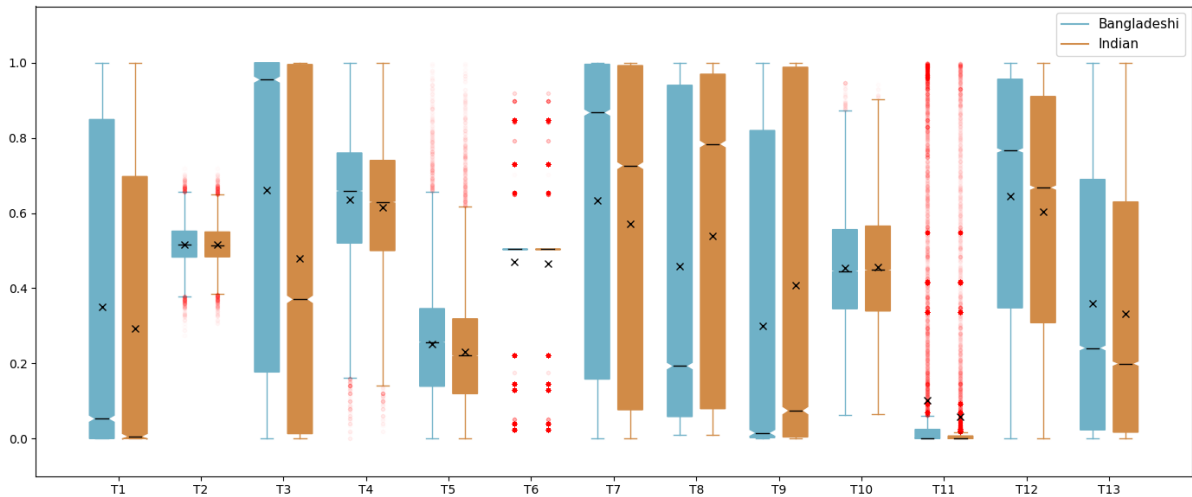
While the existing literature has established that algorithms reproduce social biases, our study contributes in several different ways. First, while the dearth of NLP (e.g., sentiment analysis) research in non-English language reinforces the colonial idea of viewing various languages and identities as the monolithic “missing other” [9], our focus on an under-represented ethnic group and NLP tools in a non-English language contributes to the understanding of NLP tools’ biases in the Global South. Second, we accompany our quantitative algorithmic audit with critical identity scholarship. In doing so, we provide empirical evidence of colonial social structures and biases being replicated through sociotechnical systems as well as provide conceptual frameworks to analyze and interpret different aspects of sociocultural power dynamics, responding to critical HCI scholars’ invitation for adopting “a historicist sensibility”—the practice to see technologies as products of their time and place, and to understand how they have been shaped by the social, economic, and political factors [162]. In the sections that follow (and in mirroring our research questions), we further grapple with the results of our audit and the implications of our findings by exploring *inconsistencies in sentiment analysis tools’ outputs, codification of implicit expression of identities in sentiment analysis, and collaboration among developers of diverse demographic backgrounds.*



(a) Gender



(b) Religion



(c) Nationality

Figure 4: Distributions of scores from different BSA tools for sentences expressing different identity categories.

## 5.1 Inconsistencies in Sentiment Analysis Tools' Outputs

Comparing average sentiment scores from different Bengali sentiment analysis (BSA) tools in RQ1.a, we found that for the same lexical content, sentence structure, and identity category, BSA tools' outputs are significantly different from each other. While several BSA tools using the same dataset (e.g., YouTube Bengali drama reviews [147]) and similar models (e.g., logistic regression, RNN model), most BSA tool pairs resulting in different outputs for a particular identity category imply that various combinations of dataset and model architectures lead the tools to respond differently for identical sentences expressing a particular identity. With an assumption of universality—generalizing perception of sentiment across cultures and populations, sentiment analysis is used in various tasks, such as in gauging public sentiment toward political figures and issues [16, 183], social issues and contemporary events [74, 186], and gathering insights from textual data in customer service [80, 104], healthcare [80], and public sectors [145, 181], amongst other applications. Our finding from RQ1.a implies that the extracted insights about subjectivity and polarity from textual data can vary significantly depending on which BSA tools are used.

Reading through the documentation, README files and associated research articles of our examined BSA tools indicated that none of these included post-development user testing and checking for identity-related biases. This leaves room for inconsistencies and discrepancies among sentiment analysis tools to go unscrutinized and unattended. Moreover, the lack of participation of users from different demographic groups within Bengali communities leads to disparities in accessing and using Bengali language technologies. Returning to our discussions on cultural hegemony in section 2, such a digital divide among developers and users and invisible politics of code institutionalize a specific group's power and control through technological artifacts and, consequently, their perceptions and beliefs shape technology used within a larger community. By convincing others that their values and interests align with the overall community's perspectives and benefits, that specific group achieves technological hegemony. To resist certain groups systematically benefiting more from a sociotechnical system than other communities and systematically having influence over data-centric infrastructures, following prior scholarship [7, 10], we urge collaboration among stakeholders to ensure that their developed sentiment analysis tools' responses to Bengali sentences are aligned with the perspectives of the community and that they are not prejudiced against any particular identity or group of people.

## 5.2 Codification of Implicit Expression of Identities in Sentiment Analysis

To answer RQ1.b, we examined how different BSA tools respond to different identity categories, expressed explicitly (e.g., through direct mention) and implicitly (e.g., through colloquial vocabularies, community norms around names and kinship). Similar to our examination of varied Bengali dialects in Bangladesh and India, other major languages have different dialects that are sociohistorically and culturally connected with particular groups within the broader linguistic communities (e.g., Southern and Coastal accents

of American English, Quebec accent of French). For example, due to the refugee crises created by the postcolonial partition in Bengal, Bangladeshi (then East Bengal) dialects were associated with refugees in India, and speakers of this dialect are often subject to contempt both online and offline [36, 37, 54]. According to identity scholars, identity is constructed and learned through everyday speech acts and non-verbal activities in different social settings and are thus modeled after normative cultural and societal logics [35]. Though researchers have qualitatively studied how sociotechnical platforms marginalize people based on their performative identity [54, 117, 150], only a few works quantitatively studied how computing systems codify the performativity—the expression of identity through repetition of norms [35] (e.g., colloquial verbal and speech acts) of various communities and groups [58, 141].

As parochial and stereotypical representations influence the development of datasets and tools, sentiment analysis and NLP tools broadly can inflict representational harm by conflating particular identities into one (e.g., viewing all Indic languages as the same or limiting a linguistic identity by nation states<sup>7</sup>). While researchers found evidence of accent gaps and racial disparity in speech recognition and language identifiers (e.g., not recognizing Southern American English) [28, 85], our study highlighted how sentiment analysis tools codify different country-based communities' preference of vocabularies as implicit expressions of identities and exhibit biases based on those. Prior CHI literature proposed using readily available sentiment analysis (e.g., VADER) to gather insights from textual data in algorithmic decision-making [136, 145]. Based on our finding that sentiment analysis tools codify the internal practices of different religion and nationality-based communities, we need to ask how these community practices and various societal biases and prejudices regarding those practices being embedded within sentiment analysis tools would impact algorithmic decision-making. We explore this issue further through the application of sentiment analysis tools in the context of content moderation in the following section.

## 5.3 Exploring Downstream Effects of Bias in Sentiment Analysis Through the Context of Content Moderation

In RQ2.a, we found that most sentiment analysis tools available in the Bengali language are biased toward a particular category in cases of identity dimensions of gender, religion, and nationality. For sentences with similar structure and word content, most BSA tools (77%) deemed Bangladeshi identity to be more positive than Indian identity, exhibiting a nationality-based bias. We found BSA tools exhibiting such favoritism toward female (38%), male (30%), Hindu (30%), Muslim (38%), and Indian (15%) identities. Such preference toward a particular religious or national community's direct mention or linguistic practices resembles [54]'s finding of biases in content moderation. For some BSA tools, we could not find evidence of those consistently assigning significantly different sentiment scores to different identity categories under a single

<sup>7</sup>Some decolonial scholars have argued that nation states and governments as forms of hierarchy and authority are also consequences of colonization that perpetuate colonial values (e.g. forced integration of smaller ethnic communities) [98, 131]

dimension (e.g., T1 for gender, T5 for religion, and T10 for nationality). While those tools did not show bias in a particular dimension, our analysis could not identify a BSA tool that maintains such impartiality across all three dimensions of gender, religion, and nationality. Using biased language technologies like a sentiment analysis tool can have downstream effects. For example, sentiment analysis is also a ubiquitously used component in automated content moderation systems [87, 161, 167, 176]. Scholarship in social computing and communications have studied the construction of automated content moderation [38, 92] and users' perception of those systems [93, 154]. Though, due to algorithms' complexities and common failure to understand the contexts of human languages, automated content moderation's legitimacy is questioned [124], users perceive automated moderation to be more impartial with human oversight [123]. Related to user personality and social aspects [134], in some cases, researchers have found that "users trust AI as much as humans for flagging problematic content" [116, 168].

Given how the transnational and religiously diverse Bengali communities' colonial past continues the distrust and division across religions and national borders and impacts their experience with platform governance and perception of biased content moderation, especially the anonymous human moderators [54], we ask if automated content moderation is used instead of human moderation, how would that impact user interaction and experience for diverse Bengali communities? This question stems from considering "automated" and "human" as two ends of a spectrum of moderation style [94]. If the sentiment analysis component within that automated moderation system is biased, as we found in our study, it can misinterpret non-normative opinions as negative and trigger automated content moderation systems to remove the content from the platforms. Thus, users, especially the ones from marginalized and minority communities, can fear being censored for expressing their perspectives. Rather than complementing human moderators' efforts in managing large online communities, automated moderation can be employed as a pretext to justify the marginalization of diverse voices. Altogether, biased BSA tools being used in automated moderation can deter inclusive and in-depth discussions, prompt users to disengage or become inactive, and eventually shape a homogenized identity and reflect existing colonial divisions and structures in Bengali societies—much like the outcomes of biased human moderators [54].

#### 5.4 Collaboration among Developers of Diverse Demographic Backgrounds

Returning to RQ2.b, though we did not find any relationship between the BSA tools' direction of biases and the demographics of those tools' developers, we cannot overlook the homogeneity of developers' identities. Since all the BSA tools we audited were developed by Bengali developers and not some Western entities, do we need to ask "who designs?" Does postcolonial computing's concern about computing systems' similarities with colonial practices apply here? Prior CHI research found that while transgenerational colonial values (e.g., collective identity posited on difference) shape Bengali users' interaction with and through computing systems, collaborative discourses resist such views [55]. However, earlier in the paper, in Table 1, we saw that most BSA tools on PyPI and

GitHub are developed by solo developers or teams of a few coders with little diversity—most tools being developed by individuals who identify as male, Muslim, and Bangladeshi. Similar to colonial Bengal, where certain exclusionary social identities (e.g., *babu*: educated Bengali men often based in Kolkata, West Bengal) emerged as accepted changes in Bengali identity and subjectivity [63], despite the Bengali language being spoken natively by diverse religious and national communities, we found certain isolation and lack of collaboration to exist among developers of diverse backgrounds. For example, though BSA tool T4 had both female and male developers, similar collaboration did not occur across various religion and nationality-based identities in any BSA tool.

Does the colonial past of the subcontinent and the Bengali people have anything to do with today's lack of collaboration in the developing sociotechnical systems in the Bengali context? Prior work has highlighted that colonial rule fragmented the Bengali people's imagination of communities, deepened the communal distrust among Hindus and Muslims, and increased the communication gaps among Bengalis in Bangladesh and India [39, 54]. For example, whereas Indian Bengalis' nationality is shaped by linguistically diverse Indian identity [158], Bangladesh defines its people's concept of nationalism as being derived from Bengali language and culture [19]. Therefore, language's role in shaping Bengali people's cultural identity and imagination of communities varies in Bangladesh and India. This difference translates to Bengali researchers' participation in computational linguistics research in their local language. For example, developers of all but one BSA tool self-identified as being from Bangladesh. Beyond our study, most leading Bengali NLP research endeavors, such as learning and research groups<sup>8</sup> and workshops<sup>9</sup>, are supported and advanced by Bangladeshi communities and government. Though Indian researchers also regularly contribute to Bengali NLP, it is often done through the framing of NLP for Indic languages [12] and lacks the concentrated attention that the Bangladeshi NLP community puts in the Bengali language. As NLP tools in Bengali are predominantly developed by Bangladeshi Bengalis, those technologies, reflecting Bangladeshi values, norms, and prejudices, can become biased. Actively collaborating among individuals from different religions within the Bengali communities and institutions across geographic boundaries can contribute to mitigating such digital divisions.

## 6 CONCLUSION: CALL FOR ENGINEERING ACTIVISM IN CRITICAL HCI

This paper presents findings from algorithmic audits of Bengali sentiment analysis (BSA) tools. Using statistical methods, we found that sentiment scores from different BSA tools vary for sentences with identical lexical content and structure. Our analysis also found evidence of BSA tools exhibiting biases, such as consistently assigning significantly different sentiment scores to sentences expressing different gender, religion, and nationality-based identities. Complementing qualitative identity literature in CHI, we quantitatively examined how sentiment analysis tools respond to explicit and implicit expressions of a certain identity category in a sentence. By

<sup>8</sup><https://bengali.ai/>, <https://csebuetnlp.github.io/>, <https://sustbanglaresearch.org/>

<sup>9</sup><https://blp-workshop.github.io/>

demonstrating the consistent discrepancies in BSA tools' outputs for identical sentences expressing various identities, this paper challenges the "functionality assumption" [130] of sentiment analysis that it relies on sentences' lexical content and contexts and not on what identities they represent. In our discussion, we explained our quantitative findings through a postcolonial understanding of the studied linguistic communities' social, cultural, and historical contexts. This paper, foregrounding the historically marginalized and under-represented Bengali community, contributes to the intersection of CHI, social computing, NLP, and fairness and bias literature contextualized in the Global South.

While critical HCI studies adopting a qualitative approach can provide deep and rich insights into biases in computational systems, those explorations are insufficient, and a fine-grained understanding of systems, architecture, algorithms, and code is essential for describing and explaining new information technologies' social, ethical, and political dimensions [121]. Building on that call for "engineering activism"—the use of engineering skills and knowledge to promote social justice, we argue that future NLP research (e.g., developing sentiment analysis tools), especially in critical HCI space, should actively reflect on identity-related biases and seek collaboration among individuals of diverse religious and transnational identities. Such socially aware engineering efforts can be particularly useful for an emerging effort in human-centered AI research that encourages "seamful designs" [66] of sociotechnical infrastructures for greater explainability and user agency. While recent works have often sought to trace and identify stages of AI dataset development [99, 152], through this paper, we encourage the examination of "seams" (e.g., sentiment analysis components) of larger AI-based systems (e.g., automated content moderation). Moreover, we showed how a historicist sensibility [162], particularly understanding the sociocultural context in relation to coloniality [52], would better ground the engineering activist efforts.

## ACKNOWLEDGMENTS

We thank the tool developers who responded to our communication to provide their demographic information and answer our queries about reproducing their tools. This research was partially supported by the University of Colorado Boulder CMCI Research Funding.

## REFERENCES

- [1] International Energy Agency. 2019. Emissions – Global Energy & CO2 Status Report 2019 – Analysis - IEA. <https://www.iea.org/reports/global-energy-co2-status-report-2019/emissions>. Last accessed: August 3, 2023.
- [2] 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>
- [3] 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).
- [4] Olga Akselrod. 2021. How artificial intelligence can deepen racial and economic inequities. <https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities>. Last accessed: Sep 11, 2023.
- [5] Syed Mustafa Ali. 2016. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students* 22, 4 (2016), 16–21.
- [6] Tariq Ali. 1971. Bangla Desh: Results and Prospects. *New Left Review* 68 (1971), 3–55.
- [7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.
- [8] Benedict Anderson. 2006. *Imagined communities: Reflections on the origin and spread of nationalism*. Verso books.
- [9] Ahmed Ansari. 2020. Design's Missing Others and Their Incommensurate Worlds. *Design in Crisis New Worlds, Philosophies and Practices* (2020).
- [10] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-centered data science: an introduction*. MIT Press.
- [11] Sabrina Argoub. 2021. The NLP divide: English is not the only natural language. <https://blogs.lse.ac.uk/polis/2021/06/09/the-nlp-divide-english-is-not-the-only-natural-language/>. Last accessed: Sep 10, 2023.
- [12] Gaurav Arora. 2020. Inltk: Natural language toolkit for indic languages. *arXiv preprint arXiv:2009.12534* (2020).
- [13] 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.
- [14] 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.
- [15] 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>
- [16] Younggwe Bae and Hongchul Lee. 2012. Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and technology* 63, 12 (2012), 2521–2535.
- [17] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 2–2.
- [18] Sarbani Banerjee. 2015. "More or Less" Refugee?: Bengal Partition in Literature and Cinema. The University of Western Ontario (Canada).
- [19] Government of the People's Republic of Bangladesh. 1972. The Constitution of the People's Republic of Bangladesh: Nationalism. Last accessed: Aug 28, 2023.
- [20] Srijan Bansal, Vishal Garimella, Ayush Suhane, and Animesh Mukherjee. 2021. Debiasing multilingual word embeddings: A case study of three indian languages. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 27–34.
- [21] R. Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.
- [22] 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.
- [23] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. *arXiv preprint arXiv:2209.12226* (2022).
- [24] 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.acl-main.55>
- [25] Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Sohail Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *arXiv preprint arXiv:2101.00204* (2021).
- [26] Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*. 3504–3519.
- [27] 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).
- [28] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).
- [29] Carlo Emilio Bonferroni. 1935. The calculation of assurance from groups of tests. *Studi in Onore del Professore Salvatore Ortu Carboni* (1935).
- [30] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.
- [31] Meredith Broussard. 2019. *Artificial unintelligence*. MIT press.
- [32] Nina Brown, Thomas McIlwraith, and Laura Tubelle de González. 2020. *Perspectives: An open introduction to cultural anthropology*. Vol. 2300. American Anthropological Association.
- [33] 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.
- [34] 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.

- [35] Judith Butler. 2011. *Gender trouble: Feminism and the subversion of identity*. routledge.
- [36] Anindita Chakrabarty. 2020. Migrant Identity at the Intersection of Postcolonialism and Modernity. *Journal of Migration Affairs* 2, 2 (2020), 100–116.
- [37] Dipesh Chakrabarty. 1996. Remembered villages: representation of Hindu-Bengali memories in the aftermath of the partition. *Economic and Political Weekly* (1996), 2143–2151.
- [38] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.
- [39] Partha Chatterjee. 1993. *The nation and its fragments: Colonial and postcolonial histories*. Princeton University Press.
- [40] 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.
- [41] Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking beneath the hood of uber. In *Proceedings of the 2015 internet measurement conference*. 495–508.
- [42] John Cheney-Lippold. 2017. We are data. In *We Are Data*. New York University Press.
- [43] The City of Boulder and Xcel Energy. 2023. Energy Partnership Agreement: 2023 – 2025 Overview. [bit.ly/BoulderEnergy2023](https://bit.ly/BoulderEnergy2023). Last accessed and archived: Jan 19, 2024.
- [44] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.
- [45] Jacob Cohen. 2016. A power primer. (2016).
- [46] Patricia Hill Collins. 2022. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.
- [47] William J Conover and Ronald L Iman. 1979. On Multiple-Comparisons Procedures. *Technical Report LA-7677-MS* (1979), 124–129.
- [48] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- [49] 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.
- [50] Peter Cummings. 2011. Arguments for and against standardized mean differences (effect sizes). *Archives of pediatrics & adolescent medicine* 165, 7 (2011), 592–596.
- [51] 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.
- [52] Dipto Das. 2023. Studying Multi-dimensional Marginalization of Identity from Decolonial and Postcolonial Perspectives. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 437–440.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] Mithun Das and Animesh Mukherjee. 2023. BanglaAbuseMeme: A Dataset for Bengali Abusive Meme Classification. *arXiv preprint arXiv:2310.11748* (2023).
- [57] Veena Das. 2006. *Life and Words: Violence and the Descent into the Ordinary*. Univ of California Press.
- [58] Mark Diaz, 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.
- [59] Afa Dil. 1972. *The Hindu and Muslim Dialects of Bengali*. Stanford University.
- [60] Paul Dimeo. 2002. Colonial bodies, colonial sport: 'Martial' Punjabis, effeminate Bengalis and the development of Indian football. *The international journal of the history of sport* 19, 1 (2002), 72–90.
- [61] diviAI. 2020. Diversity in Artificial Intelligence: ACM FAcT 2020. <https://divinai.org/conf/74/acm-fact>. Last accessed: Sep 12, 2023.
- [62] Paul Dourish and Scott D Mainwaring. 2012. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 133–142.
- [63] Sutapa Dutta. 2021. Packing a Punch at the Bengali Babu. *South Asia: Journal of South Asian Studies* 44, 3 (2021), 437–458.
- [64] 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.
- [65] Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of Airbnb. com. *Harvard Business School NOM Unit Working Paper* 14-054 (2014).
- [66] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daume III. 2022. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *arXiv preprint arXiv:2211.06753* (2022).
- [67] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*. Springer, 449–466.
- [68] Erik H Erikson. 1968. *Identity youth and crisis*. Number 7. WW Norton & company.
- [69] Maria Eriksson and Anna Johansson. 2017. Tracking gendered streams. *Culture unbound. Journal of Current Cultural Research* 9, 2 (2017), 163–183.
- [70] 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.
- [71] Angela Fan and Claire Gardent. 2022. Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8561–8576. <https://doi.org/10.18653/v1/2022.acl-long.586>
- [72] F. Fanon, R. Philcox, and K.A. Appiah. 2008. *Black Skin, White Masks*. Grove Atlantic. <https://books.google.com/books?id=W45-IrrK-sC>
- [73] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.
- [74] Simon Fong, Yan Zhuang, Jinyan Li, and Richard Khoury. 2013. Sentiment analysis of online news using mallet. In *2013 International Symposium on Computational and Business Intelligence*. IEEE, 301–304.
- [75] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)* 14, 3 (1996), 330–347.
- [76] Roland G Fryer Jr and Steven D Levitt. 2004. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics* 119, 3 (2004), 767–805.
- [77] Viktor Gecas. 1982. The self-concept. *Annual review of sociology* 8, 1 (1982), 1–33.
- [78] Anindita Ghoshal. 2021. 'mirroring the other': Refugee, homeland, identity and diaspora. In *Routledge Handbook of Asian Diaspora and Development*. Routledge, 147–158.
- [79] Erving Goffman. 1978. *The presentation of self in everyday life*. Harmondsworth London.
- [80] Sunir Gohil, Sabine Vuik, Ara Darzi, et al. 2018. Sentiment analysis of health care tweets: review of the methods used. *JMIR public health and surveillance* 4, 2 (2018), e5789.
- [81] Jonathan D Greenberg. 2005. Generations of memory: remembering partition in India/Pakistan and Israel/Palestine. *Comparative Studies of South Asia, Africa and the Middle East* 25, 1 (2005), 89–110.
- [82] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.
- [83] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*. 305–318.
- [84] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1914–1933.
- [85] Drew Harwell. 2018. Why some accents don't work on Alexa or Google Home - Washington Post. <https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/>. Last accessed: Sep 4, 2023.
- [86] 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).
- [87] Danula Hettiachchi and Jorge Goncalves. 2019. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 342–346.
- [88] 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.
- [89] Tenghao Huang, Faeze Brahmam, 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>
- [90] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1311–1320.



- [91] 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.
- [92] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.
- [93] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
- [94] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–34.
- [95] Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 2017. The effect of population and "structural" biases on social media-based algorithms: A case study in geolocation inference across the urban-rural spectrum. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*. 1167–1178.
- [96] 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>
- [97] Purna Juneja and Tanushree Mitra. 2021. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–27.
- [98] Dip Kapoor. 2012. Human rights as paradox and equivocation in contexts of Adivasi (original dweller) dispossession in India. *Journal of Asian and African Studies* 47, 4 (2012), 404–420.
- [99] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. (2022).
- [100] 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).
- [101] 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>
- [102] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47, 260 (1952), 583–621.
- [103] James Lane. 2023. The 10 Most Spoken Languages In The World. <https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world>. [Accessed 23-08-2023].
- [104] Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke. 2019. Acoustic and lexical sentiment analysis for customer service calls. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5876–5880.
- [105] 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.
- [106] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How weird is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [107] Ania Loomba. 2002. *Colonialism/postcolonialism*. Routledge.
- [108] Maria Lugones. 2007. Heterosexuality and the colonial/modern gender system. *Hypatia* 22, 1 (2007), 186–219.
- [109] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
- [110] Leslie McCall. 2005. The complexity of intersectionality. *Signs: Journal of women in culture and society* 30, 3 (2005).
- [111] Jo McCormack, Murray Pratt, and Alistair Rolls Alistair Rolls. 2011. *Hexagonal variations: diversity, plurality and reinvention in contemporary France*. Vol. 359. Rodopi.
- [112] Mary L McHugh. 2013. The chi-square test of independence. *Biochemia medica* 23, 2 (2013), 143–149.
- [113] 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.
- [114] 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.
- [115] Walter D Mignolo. 2007. Delinking: The rhetoric of modernity, the logic of coloniality and the grammar of de-coloniality. *Cultural studies* 21, 2-3 (2007), 449–514.
- [116] Maria D Molina and S Shyam Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27, 4 (2022), zmac010.
- [117] Isabel Munoz, Michael Dunn, Steve Sawyer, and Emily Michaels. 2022. Platform-mediated Markets, Online Freelance Workers and Deconstructed Identities. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022).
- [118] Ashis Nandy. 1989. *The Intimate Enemy: Loss and Recovery of Self Under Colonialism*. Oxford University Press Oxford.
- [119] Gabriel Nicholas and Aliya Bhatia. 2023. Lost in Translation: Large Language Models in Non-English Content Analysis. *arXiv preprint arXiv:2306.07377* (2023).
- [120] Gabriel Nicholas and Aliya Bhatia. 2023. Re: Request for Information (RFI) on Developing a Roadmap for the Directorate for Technology, Innovation, and Partnerships at the National Science Foundation| 88 FR 26345. [https://cdt.org/wp-content/uploads/2023/08/Non-EN-NLP\\_-\\_NSF-RFI\\_-\\_Draft-of-CDT-Comment.pdf](https://cdt.org/wp-content/uploads/2023/08/Non-EN-NLP_-_NSF-RFI_-_Draft-of-CDT-Comment.pdf). (2023).
- [121] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.
- [122] 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.
- [123] Marie Ozanne, Aparajita Bhandari, Natalya Bazarova, and Dominic DiFranzo. 2022. Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society* (2022).
- [124] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–31.
- [125] G. Pandey. 2001. *Remembering Partition: Violence, Nationalism and History in India*. Cambridge University Press.
- [126] Bhasa Vidya Parishad. 2001. *Praci Bhasavijnan: Indian Journal of Linguistics*. Number v. 20. Bhasa Vidya Parishad. <https://books.google.com/books?id=0yxhAAAAMAAJ>
- [127] David Patterson. 2021. How we're minimizing AI's carbon footprint. <https://blog.google/technology/ai/minimizing-carbon-footprint/>. Last accessed: August 3, 2023.
- [128] Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50, 302 (1900), 157–175.
- [129] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210* (2019).
- [130] 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.
- [131] Maia Ramnath. 2012. *Decolonizing anarchism: an antiauthoritarian history of India's liberation struggle*. Vol. 3. AK Press.
- [132] Gavin Rand. 2006. 'Martial races' and 'imperial subjects': violence and governance in colonial India, 1857–1914. *European Review of History: Revue européenne d'histoire* 13, 1 (2006), 1–20.
- [133] Nani Jansen Rentflow. 2021. How Artificial Intelligence Impacts Marginalised Groups. <https://digitalfreedomfund.org/how-artificial-intelligence-impacts-marginalised-groups/>. Last accessed: Sep 11, 2023.
- [134] René Riedl. 2022. Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electronic Markets* 32, 4 (2022), 2021–2051.
- [135] 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).
- [136] Julia Romberg and Tobias Escher. [n. d.]. Making Sense of Citizens' Input through Artificial Intelligence: A Review of Methods for Computational Text Analysis to Support the Evaluation of Contributions in Public Participation. *Digital Government: Research and Practice* ([n. d.]).
- [137] Christian Rudder. 2013. Inside OKCupid: The math of online dating. <https://ed.ted.com/lessons/inside-okcupid-the-math-of-online-dating-christian-rudder>. [Accessed 20-08-2023].
- [138] Henrik Skaug Sætra. 2021. AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system. *Sustainability* 13, 4 (2021), 1738.
- [139] E.W. Said. 2014. *Orientalism*. Knopf Doubleday Publishing Group.
- [140] 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.
- [141] 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.
- [142] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 1807–1824. <https://doi.org/10.18653/v1/2022.acl-long.127>
- [143] 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.
- [144] Devansh Saxena, Erina Seh-Young Moon, Aryan Chaurasia, Yixin Guan, and Shion Guha. 2023. Rethinking “Risk” in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [145] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking invisible work practices, constraints, and latent power relationships in child welfare through casenote analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [146] Devansh Saxena, Charles Repaci, Melanie D Sage, and Shion Guha. 2022. How to Train a (Bad) Algorithmic Caseworker: A Quantitative Deconstruction of Risk Assessments in Child Welfare. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
- [147] Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. 50–60.
- [148] 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).
- [149] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8, 2 (2021), 20539517211053712.
- [150] 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.
- [151] 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.
- [152] Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. 2023. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.
- [153] 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.
- [154] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.
- [155] Dwaipayan Sen. 2018. *The decline of the caste question: Jogendranath Mandal and the defeat of Dalit politics in Bengal*. Cambridge University Press.
- [156] Shilad Sen, Margaret E Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao Wang, and Brent Hecht. 2015. “Turkers, scholars,” arafat” and” peace” cultural communities and algorithmic gold standards. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing*.
- [157] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.
- [158] Gurharpal Singh and Heewon Kim. 2018. The limits of India’s ethno-linguistic federation: Understanding the demise of Sikh nationalism. *Regional & Federal Studies* 28, 4 (2018), 427–445.
- [159] Mrinalini Sinha. 2017. Colonial masculinity: The ‘manly Englishman’ and the ‘effeminate Bengali’ in the late nineteenth century. In *Colonial masculinity*. Manchester University Press.
- [160] Manjira Sinha and Anupam Basu. 2016. A study of readability of texts in Bangla through machine learning approaches. *Education and information technologies* 21, 5 (2016), 1071–1094.
- [161] Hayden Smith and William Cipolli. 2022. The Instagram/Facebook ban on graphic self-harm imagery: A sentiment analysis and topic modeling approach. *Policy & Internet* 14, 1 (2022), 170–185.
- [162] Robert Soden, David Ribes, Seyram Avle, and Will Sutherland. 2021. Time for historicism in CSCW: An invitation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–18.
- [163] Anders Sogaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What’s in a p-value in NLP?. In *Proceedings of the eighteenth conference on computational natural language learning*. 1–10.
- [164] Gayatri Chakravorty Spivak. 2023. Can the subaltern speak? In *Imperialism*. Routledge, 171–219.
- [165] Ramya Srinivasan and Kanji Uchino. 2021. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 41–51.
- [166] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).
- [167] Heng Sun, Wan Ni, et al. 2022. Design and Application of an AI-Based Text Content Moderation System. *Scientific Programming* 2022 (2022).
- [168] Matt Swayne. 2022. Users trust AI as much as humans for flagging problematic content | Penn State University. <https://www.psu.edu/news/institute-computational-and-data-sciences/story/users-trust-ai-much-humans-flagging-problematic/>. Last accessed: August 10, 2023.
- [169] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.
- [170] 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.
- [171] Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social science information* 13, 2 (1974), 65–93.
- [172] Diana Taylor. 2003. *The archive and the repertoire: Performing cultural memory in the Americas*. Duke University Press.
- [173] Pranav Tiwari, Aman Chandra Kumar, Aravindan Chandrabose, et al. 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*. 1–7.
- [174] 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>
- [175] 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.
- [176] Sahaj Vaidya, Jie Cai, Soumyadeep Basu, Azadeh Naderi, Donghee Yvette Wohn, and Aritra Dasgupta. 2021. Conceptualizing visual analytic interventions for content moderation. In *2021 IEEE Visualization Conference (VIS)*. IEEE.
- [177] Robert Van Krieken. 2004. Rethinking cultural genocide: Aboriginal child removal and settler-colonial state formation. *Oceania* 75, 2 (2004), 125–151.
- [178] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Shomir Wilson, et al. 2023. Nationality Bias in Text Generation. *arXiv preprint arXiv:2302.02463* (2023).
- [179] Pranav Narayanan Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J Passonneau, and Shomir Wilson. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. *arXiv preprint arXiv:2310.12318* (2023).
- [180] 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>
- [181] Sanjeev Verma. 2022. Sentiment analysis of public services for smart society: Literature review and future research directions. *Government Information Quarterly* 39, 3 (2022), 101708.
- [182] 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*.
- [183] Hao Wang, Doğan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*.
- [184] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 196–202.
- [185] Langdon Winner. 2017. Do artifacts have politics? In *Computer ethics*. Routledge, 177–192.
- [186] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems* 60 (2019), 617–663.
- [187] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14830–14840.
- [188] 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.