Cultural Re-contextualization of Fairness Research in Language Technologies in India

Shaily Bhatt
Google Research
shailybhatt@google.com

Sunipa Dev
Google Research
sunipadev@google.com

Partha Talukdar
Google Research
partha@google.com

Shachi Dave*
Google Research
shachi@google.com

Vinodkumar Prabhakaran*
Google Research
vinodkpg@google.com

Abstract

Recent research has revealed undesirable biases in NLP data and models. However, these efforts largely focus on social disparities in the West, and are not directly portable to other geo-cultural contexts. In this position paper, we outline a holistic research agenda to re-contextualize NLP fairness research for the Indian context, accounting for Indian *societal context*, bridging *technological* gaps in capability and resources, and adapting to Indian cultural *values*. We also summarize findings from an empirical study on various social biases along different axes of disparities relevant to India, demonstrating their prevalence in corpora and models.

1 Introduction

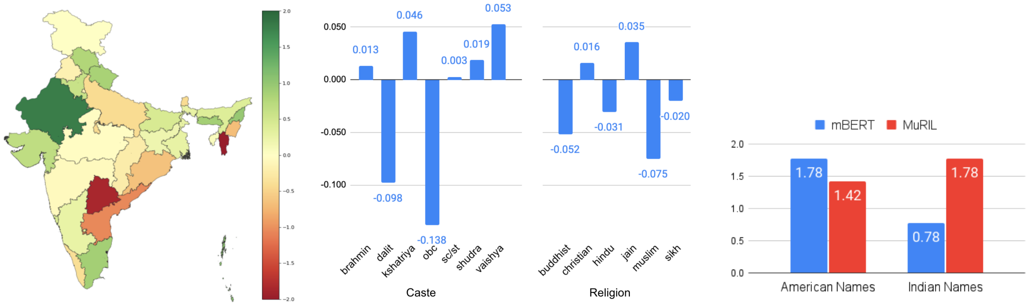
Recent research has demonstrated that language technologies may capture, propagate, and amplify societal biases [4]. While Natural Language Processing (NLP) has seen global adoption, most studies on assessing and mitigating such biases are situated in the Western context,[1] focusing primarily on axes of disparities prevalent in the Western public discourse, and hence not directly portable to non-Western contexts [19]. This is especially troubling in the case of India, a pluralistic nation of 1.4 billion people, with fast-growing investments in NLP research, development, and deployments from the government, the industry, and the startup ecosystem. While there is some recent work on NLP fairness in Indian languages like Hindi, Bengali, and Telugu [17, 13], re-contextualizing NLP fairness for the Indian context requires a holistic approach that accounts for the various relevant axes of social disparities in the Indian society, their proxies in language data, the disparate NLP capabilities across Indian languages and dialects, and the (lack of) availability of resources that enable fairness evaluations and mitigation [19]. In this paper, we summarize takeaways from an empirical analysis of biases in NLP models along various axes of disparities relevant in the Indian context, and then propose a holistic roadmap for re-contextualizing data and model fairness in NLP.

2 Summary of Empirical Results

We first report some highlights from our extensive empirical analysis of social biases in NLP models in the Indian context [3]. The axes of disparities we consider include two India-specific axes: a) *Caste*, which is an inherited hierarchical social identity, that has been the basis of historical marginalization; and b) *Region*, or ethnicity associated with geographic regions of India, as well as four globally-salient

[1] We use *Western* or *the West* to refer to the regions, nations and states consisting of Europe, the U.S., Canada, and Australasia, and their shared norms, values, customs, religious beliefs, and political systems [11].

Preprint. Under review.



(a) Region (b) Caste and Religion (c) Gender

Figure 1: Highlights from an empirical analysis of biases along axes of disparities in the Indian context. Fig (a) and (b) show perturbation analysis results [15] using identity terms for *Region*, *Caste*, and *Religion* on the HuggingFace default sentiment model. Fig (c) shows the DisCo metrics [20] using Indian and American names to measure *Gender* bias in language models mBERT and MuRIL.

axes that have unique manifestations in the Indian context: a) *Gender*, where there are different structural disparities in engagement of women in society as compared to the West; b) *Religion*, wherein the majority and minority religious groups differ compared to the west; c) (dis)*Ability* and 4) *Gender Identity and Sexual Orientation*, around which the social discourse and awareness in India is fairly recent. We analyzed various proxies in language data for these social groups such as identity terms, personal names, and dialectal features to study biases in NLP models.

Figures 1a and 1b shows shifts in sentiment scores in response to perturbation analysis [15] of identity terms for *Region*, *Caste*, and *Religion* on the HuggingFace default sentiment model (a DistillBERT fine-tuned with SST-2). In particular, we see that the model has learnt to associate higher negative sentiment towards marginalized sub-groups, such as 'Dalit' and 'OBC' (other backward castes) in caste, and 'Muslim' in religion. For state identities, the model has learnt to associate more negative sentiment with southern states like Andhra Pradesh and Telangana, and North-Eastern states like Mizoram and Manipur. Figure 1c shows that DisCo metric [20] that measures whether the predictions of a model have statistically significant associations to (binary) gender in language models require Indian names with gender association in order to correctly detect encoded biases. In addition, we also built a human-curated dataset of stereotypes around Region and Religion axes to demonstrate that such stereotypes are preferentially encoded in models and corpora (not shown in the figures above).

3 Towards Cultural Re-contextualization of NLP Fairness in India

The above results demonstrate that NLP models reflect societal biases around socio-demographic subgroups in the Indian context. To effectively address these issues we need a holistic perspective that accounts for the various factors in the ecosystem. Building on [19], we propose a holistic research agenda (Figure 2) for re-contextualizing fairness in NLP along three dimensions: accounting for the *societal context*, bridging the *technological gaps*, and adapting to the local *values and norms*.

3.1 Accounting for Indian *Societal* context

Socially Situated Evaluation: A major hurdle in accounting for different axes is the access to diverse annotator pools who have familiarity and lived experiences of the marginalized groups. This is important for fairness work in general [6], but especially in India where public discourse around (dis)ability, gender identity and sexual orientation is relatively limited. Participatory approaches [12] to co-create resources for fairness evaluation will be crucial for meaningfully addressing this gap.

Data Voids: Entire communities may be excluded from language data due to disparities in literacy and internet access [19]. Not accounting for such data voids might result in biases being baked into the language models that has become base infrastructure for NLP [5]. Further, the risk of unintentionally excluding marginalized communities based on dialect or other linguistic features while filtering data to ensure quality [7, 9] is even higher in the Indian context because of very limited computational representation of marginalized communities. Participatory data curation (e.g., collecting language data specifically from marginalized communities [1, 14] can significantly help bridge such data voids.

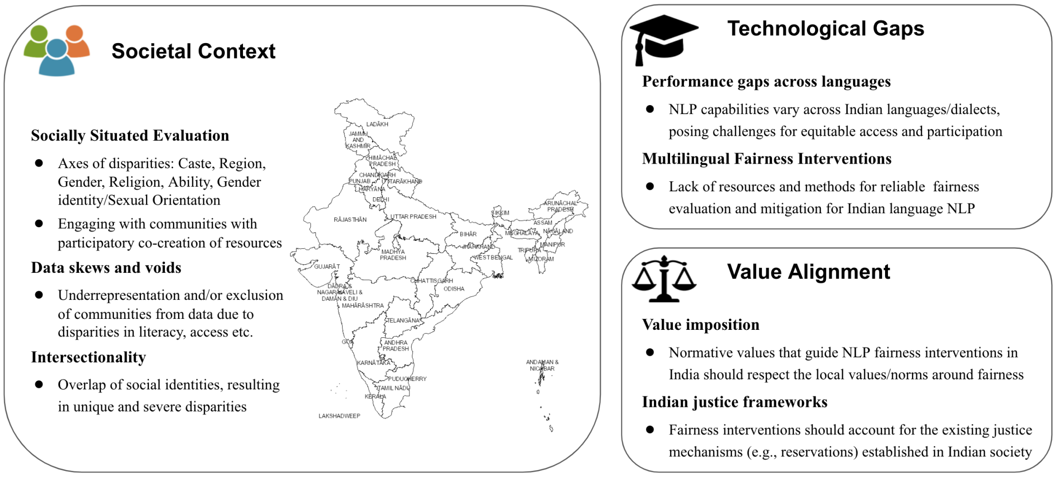

Societal Context
Socially Situated Evaluation
Axes of disparities: Caste, Region, Gender, Religion, Ability, Gender identity/Sexual Orientation
Engaging with communities with participatory co-creation of resources
Data skews and voids
Underrepresentation and/or exclusion of communities from data due to disparities in literacy, access etc.
Intersectionality
Overlap of social identities, resulting in unique and severe disparities
LADĀKH
JAMMU AND KASHMĪR
HIMĀCHAL PRADESH
CHANDIGARH
PUNJAB
UTTARĀKHAND
HARYĀNA
DELHI
RĀJASTHĀN
UTTAR PRADESH
SIKKIM
ARUNĀCHAL PRADESH
ASSAM
BIHĀR
MEGHALAYA
NĀGĀLAND
MANIPUR
JHĀRKHAND
TRIPURA
WEST BENGAL
MIZORAM
MADHYA PRADESH
GUJARĀT
CHHATTISGARH
ODISHA
DĀDRA & NAGAR HAVELI & DAMĀN & DIU
MAHĀRĀSHTRA
TELANGĀNA
GOA
ANDHRA PRADESH
KARNĀTAKA
ANDAMAN & NICOBAR
PUDUCHERRY
TAMIL NĀDU
KERALA
LAKSHADWEEP
Technological Gaps
Performance gaps across languages
NLP capabilities vary across Indian languages/dialects, posing challenges for equitable access and participation
Multilingual Fairness Interventions
Lack of resources and methods for reliable fairness evaluation and mitigation for Indian language NLP
Value Alignment
Value imposition
Normative values that guide NLP fairness interventions in India should respect the local values/norms around fairness
Indian justice frameworks
Fairness interventions should account for the existing justice mechanisms (e.g., reservations) established in Indian society

Figure 2: A holistic research agenda for NLP Fairness in the Indian context: accounting for societal disparities in India, bridging technological gaps in NLP capabilities/resources, and adapting fairness interventions to align with local values and norms. (Map: `https://indiamaps.gov.in/soiapp/`)

Intersectionality: Due to the interplay of all the diverse axes in the Indian context, intersectional biases experienced by different marginalized groups are further exacerbated [18]. With notable differences in literacy, economic stability, technology access, and healthcare access across geographical, caste, religious, and gender divides, representation in and access to language technologies is also disparate. Bias evaluation and mitigation interventions should account for these intersectional biases.

3.2 Bridging cross-lingual *Technological* gaps

Performance gaps across languages: Although India is a vastly multilingual country with hundreds of languages, and thousands of dialects, there are wide disparities in NLP capabilities across these languages and dialects. These disparities hinder equitable access, creating barriers to internet participation, information access, and in turn, representation in data and models. While the Indian NLP community has made major strides in bridging this gap (e.g., [10]), more work is needed in building and improving NLP technologies for marginalized and endangered languages and dialects.

Multilingual fairness research: NLP Fairness research relies on evaluation resources that are currently largely built in and for the Western context. It is crucial to build these resources in Indian languages, along the lines of recent work on Hindi, Bengali, and Telugu [13, 17], since biases may manifest differently in data and models for different languages, and how bias transfers in transfer-learning paradigms for multilingual NLP is unknown. Finally, bias mitigation in one (or a few) language(s) may have counter-productive effects on other languages. Hence, a research agenda for fair NLP in India should address these various unknowns that the multilingual setting brings.

3.3 Adapting to Indian *Values and Norms*

Avoiding value imposition: Fairness inquiries answer questions such as: what does it mean to be fair or unfair, and how fair is fair enough? These questions, and their answers, are rarely made explicit; rather a shared understanding is implicitly assumed, risking value imposition. For instance, these answers often draw largely from Western values of fairness that are rooted in egalitarianism, consequentialism, deontic justice, and Rawls' distributive justice [19]. However, the philosophy of fairness in India is rooted in social restorative justice. More work should look into such value alignment challenges, which is not trivial when it comes to deploying fairness interventions [8, 16].

Accounting for Indian justice models: India has established restorative justice measures in various resource allocation contexts, colloquially known as the "reservations" [2], where historically marginalized communities (such as Dalits, other backward castes, Adivasis (tribals), and religious minorities) are afforded fixed quotas in educational institutes and government jobs to counter historical deprivation. NLP fairness research in these domains should consider how fairness interventions work in the context of such established measures.

References

[1] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France, May 2020. European Language Resources Association.

[2] BR Ambedkar. *Annihilation of Caste: The Annotated Critical Edition*. Verso Books, 2014.

[3] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. Re-contextualizing Fairness in NLP: The Case of India. In *Asia-Pacific Chapter of the Association for Computational Linguistics*. ACL, 2022.

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv e-prints*, pages arXiv–2112, 2021.

[7] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

[8] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[9] Suchin Gururangan, Dallas Card, Sarah K Drier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*, 2022.

[10] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*, 2021.

[11] James Kurth. Western civilization, our tradition. *The Intercollegiate Review*, 39(1-2):5–13, 2003.

[12] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. Human-centered approaches to fair and responsible ai. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[13] Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially aware bias measurements for hindi language representations. *CoRR*, abs/2110.07871, 2021.

[14] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November 2020. Association for Computational Linguistics.

[15] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, 2019.

[16] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A Human Rights-Based Approach to Responsible AI. *arXiv preprint arXiv:2210.02667*, 2022.

[17] Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2019, page 450–456, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Nidhi Sabharwal and Wandana Sonalkar. Dalit women in india: At the crossroads of gender, class, and caste. *Global justice: Theory, Practice, Rhetoric*, 8, 07 2015.

[19] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Reimagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, 2021.

[20] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models, 2020.