Re-contextualizing Fairness in NLP: The Case of 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 focus on social disparities in West, and are not directly portable to other geo-cultural contexts. In this paper, we focus on NLP fairness in the context of India. We start with a brief account of the prominent axes of social disparities in India. We build resources for fairness evaluation in the Indian context and use them to demonstrate prediction biases along some of the axes. We then delve deeper into social stereotypes for Region and Religion, demonstrating its prevalence in corpora and models. Finally, 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 NLP capabilities and resources, and adapting to Indian cultural *values*. While we focus on India, this framework can be generalized to other geo-cultural contexts.

1 Introduction

While Natural Language Processing (NLP) has seen impressive advancements recently (Devlin et al., 2018a; Raffel et al., 2019; Brown et al., 2020; Chowdhery et al., 2022), it has also been demonstrated that language technologies may capture, propagate, and amplify societal biases (Blodgett et al., 2020). Although NLP is adopted globally, most studies on assessing and mitigating biases are in the Western context,[1] focusing on axes of disparities in the West, relying on Western data and justice norms, and are not directly portable to non-Western contexts (Sambasivan et al., 2021).

This is especially troubling for India, a pluralistic nation of 1.4 billion people, with fast-growing investments in NLP from the government and the private sector.[2] There is commendable recent work on fairness in NLP models for Indian languages such as Hindi, Bengali, and Telugu (Pujari et al., 2019; Malik et al., 2021; Gupta et al., 2021). But, for a nation with many religions, ethnicities, and cultures, re-contextualizing NLP fairness needs to account for the various axes of social disparities in the Indian society, their proxies in language data, the disparate NLP capabilities in Indian languages, and the (lack of) resources for bias evaluation.

Sambasivan et al. (2021) proposed a research agenda for AI fairness for India based on interviews of 36 experts on Indian society and technology. In this paper, we build on their work with a focus on NLP. We start with a brief discussion on the major axes of social disparities in India (§3). We then discuss the proxies of some of these axes in language and empirically demonstrate prediction biases around these proxies in NLP models (§4). We then delve deeper into stereotypes along the axes of *Region* and *Religion*, demonstrating their prevalence in data and models (§5). Finally, we build on these empirical demonstrations to propose an overarching research agenda along the *societal*, *technological*, and *value alignment* aspects important to formulating fairness research for the Indian context (§6). While we focus on India in this paper, our framework can be adapted to re-contextualize fairness research for other geo-cultural contexts.

To summarize, our main contributions are: (1) an overarching research agenda for NLP fairness in the Indian context accounting for societal, technological, and value aspects; (2) resources (curated and created) for enabling fairness evaluations in the Indian context available;[3] and (3) empirical demon-

[1] We use *Western* or *the West* to refer to the regions, nations & states consisting of Europe, the U.S., Canada, and Australasia, and their shared norms, values, customs, religious beliefs, & political systems (Kurth, 2003).

[2] In government (https://bhashini.gov.in) and private sector (https://tinyurl.com/indiaai-top-nlp-startups, https://tinyurl.com/google-idf-language).

[3] https://www.github.com/google-research-datasets/nlp-fairness-for-india

strations of prediction biases and over-prevalence of social stereotypes in data and models.

2 Related Work

Research on undesirable biases has been a growing priority in NLP (Caliskan et al., 2017; Blodgett et al., 2020; Sheng et al., 2021; Ghosh et al., 2021). Social biases are shown to be baked into pretrained language models (Bender et al., 2021) and models for downstream tasks such as sentiment analysis (Kiritchenko and Mohammad, 2018) and toxicity detection (Sap et al., 2019). While the majority of NLP fairness research focuses on gender (Bolukbasi et al., 2016; Sun et al., 2019; Zhao et al., 2017) and racial biases (Sap et al., 2019; Davidson et al., 2019; Manzini et al., 2019), other axes of disparities such as ability (Hutchinson et al., 2020), age (Diaz et al., 2018), and sexual orientation (Garg et al., 2019) have also gotten some attention. However, the majority of this research is framed in and for the Western context, relying on data and values reflecting the West (Sambasivan et al., 2021).

Recently, fairness research in NLP has also been expanded to non-English languages such as Arabic (Lauscher et al., 2020), Japanese (Takeshita et al., 2020), Hindi, Bengali, and Telugu (Pujari et al., 2019; Malik et al., 2021). Evidence of cultural biases for different countries have also been recorded (Ghosh et al., 2021) in LMs. Our work adds to this line of research. Building on Sambasivan et al. (2021), we take a more holistic approach towards NLP fairness in the specific geo-cultural context of India. More specifically, we re-frame the agenda they proposed along "re-contextualising data and model fairness; empowering communities by participatory action; and enabling an ecosystem for meaningful fairness" with an NLP-centric lens.

3 Axes of Disparities

Identifying prominent axes of disparities is the first step in laying out a holistic NLP fairness research agenda for the Indian context. We follow Sambasivan et al. (2021) who identify the major axes of potential ML (un)fairness (Table 1 of Sambasivan et al. (2021)), and include *Region*, *Caste*, *Gender*, *Religion*, *Ability*, and *Gender Identity and Sexual Orientation*.[4] We further group them into globally salient axes (such as *Gender* and *Religion*) with local manifestations (such as different religions - for example, *Jainism*) and axes that are unique and/or specific to India (such as *Region* and *Caste*).

Further, amplified social biases may be faced by those with overlapping categories of marginalized groups. We do not focus on this *Intersectionality* here and leave discussion about it to Section 6.

3.1 India-specific axes

Region: Region as an axis can manifest globally (for example as nationality), but here we predominantly focus on the ethnicity associated with geographic regions of India and hence categorize it as India-specific. While the census does not recognise racial or ethnic groups,[5] India is home to many ethno-lingusitic groups with diverse cultures and traditions.[6] Most states in India comprise a dominant ethno-lingusitic group (such as *Haryanvis* in *Haryana*, *Goans* in *Goa*). Early research has documented various stereotypes for regional subgroups (Borude, 1966; de Souza, 1977). de Souza (1977) reported that students from a college in Mumbai ascribed traits such as crooked to Andhraites, cunning to Kannadigas, and brave to Punjabis, observing that South Indians were ascribed "unfavorable" traits more frequently. Disparities and stereotypes also exist in India at broader regional levels (for example, negative stereotypes and rampant discrimination has been documented against North-East Indians (McDuie-Ra, 2012; Haokip, 2021)), and groups belonging to smaller regions within or across states (like Konkani in parts of Goa, Maharashtra, and Karnataka).

Caste: Caste is an inherited hierachical social identity, that has been basis of historical marginalization. Despite the intended eradication of caste-based discrimination envisioned decades ago (Ambedkar, 2014), lower rungs of the caste hierarchy continue to have low literacy rates, misrepresentation, poverty, low technology access, and exclusion in language data (Deshpande, 2011; Kamath, 2018; Krishna et al., 2019).[7] Caste-based prejudices have been documented in matrimonial ads (Rajadesingan et al., 2019) and social media (Vaghela et al., 2021). Fonseca et al. (2019) found that news coverage of "lower caste" groups were focus excessively on prejudice, violence, and conflict, and ignore other aspects of their life and identity.

[4] Sambasivan et al. (2021) include *Class* as an axis, however we see class as an attribute that cuts across multiple axes, rather than as an immutable characteristic.

[5] https://www.censusindia.gov.in/

[6] https://tinyurl.com/SA-ethnic-groups

[7] https://tinyurl.com/oxfamindia-caste

3.2 Global axes in the Indian context

Gender: Although gender is a prominent axis of disparity across the globe, the specifics of how gender manifests in society (and hence, in data) varies greatly across geo-cultural contexts (Kurian, 2020). Re-contextualization of the gender axis needs to account for India-specific gender stereotypes and the structural disparities in engagement of women in society. For example women in India are 58% less likely to connect to mobile Internet then men (Sambasivan et al., 2019), have literacy rate of 65% compared to 85% for men, and 21% labor force participation compared to 76% for men.[8] Gender roles and stereotypes in India vary from the West (Sethi and Allen, 1984; Leingpibul and Mehta, 2006) and so do their potrayal in media (Griffin et al., 1994; Khairullah and Khairullah, 2009; Das, 2011).

Religion: Religious biases have been studied in NLP (Dev et al., 2020; Nadeem et al., 2020; Abid et al., 2021), however the social disparities and stereotypes about various religious groups differ significantly in India from the West, (Malik et al., 2021). For example, Christianity (typically a majority religion in the West) is a minority religion (2.3% of the population) in India, along with Sikhism (1.9%), Buddhism (0.8%), and Jainism (0.4%).

Ability: Awareness about (dis)ability is relatively recent in India (Ghosh, 2016; Ghai, 2019). Representation of disability in social discourse and the barriers it poses are significantly different for India than the West (Chaudhry and Shipp, 2005; Johnstone et al., 2017). For example people with disabilities are often abandoned at birth or socially segregated (Kumar et al., 2012) due to being seen as deceitful, unable to progress to adulthood, and dependent on charity and pity (Ghai, 2002). Disability is often mocked, portrayed as a punishment, and heteronormative narratives of 'fixing' disability are prevalent in Indian cinema (Sawhnet).

Gender Identity and Sexual Orientation: Discourse around gender identity and sexual orientation has historically been largely absent from the Indian public discourse (Abraham and Abraham, 1998). While India reflects the growing positive attitude towards LGBTQ+ issues (Anand, 2016) along with the recent decriminalisation of homosexuality (Tamang, 2020), there still exist challenges to acceptance and visibility. Furthermore, understanding LGBTQ+ related biases in the Indian context needs engagement with the social situatedness of groups like the *hijra* community, a socially outcast intersex and transgender community.

[8] https://tiny.cc/labor-gender-in

4 Proxies of Axes and Predictive Disparities

Bias evaluation in NLP relies on proxies of subgroups in language, such as identity terms and personal names, to reveal the undesirable associations present in models and data (Caliskan et al., 2017; Maudslay et al., 2019). In the Indian context, we identify three major kinds of proxies: *identity terms*, *personal names*, and *dialectal features*.

Using such proxies however poses unique challenges in the Indian context. For example, there are thousands of caste identities and hundreds of ethnolinguistic regional identities that are not codified in any authoritative sources. Similarly, there do not exist any large resources that provide subgroup associations for personal names, such as the US Census data (for race) or SSA data (for gender) in the West. Building exhaustive resources to capture such fine-grained social groups is outside the scope of this paper. However, in this section we curate identity terms and personal names with prototypical identity associations. We adopt a black-box evaluation strategy to demonstrate predictive biases in standard NLP pipelines/models and also demonstrate the utility of India-specific resources. Finally we note that these resources and studies are meant to be demonstrative, not exhaustive.

4.1 Identity Terms

We curated lists of India-specific identity terms along three different axes:

- *Region*: demonyms for states & union territories like *Kashmiri*, *Andamanese*.[9]
- *Caste*: frequently used terms-[10] *Brahmin*, *Kshatriya*, *Vaishya*, *Shudra*, *Dalit*, *SC/ST* (Scheduled Castes/Scheduled Tribes), *OBC* (Other Backward Classes).
- *Religion*: terms for populous religions- *Hindu*, *Muslim*, *Christian*, *Sikh*, *Buddhist*, *Jain*.

We now demonstrate biases in the default HuggingFace sentiment pipeline which is DistilBERT-base-uncased (Sanh et al., 2019) fine-tuned on the SST-2 (Socher et al., 2013).[11] We perform per-

[9] https://tinyurl.com/wiki-in-regions

[10] Broad (and overlapping) categories, not caste names.

[11] https://tinyurl.com/hf-sentiment.

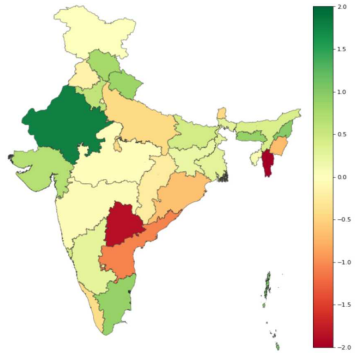

Figure 1: Relative sentiment score shift when regional identity terms are perturbed showing negative (e.g., *Mizoram*) and positive (e.g., *Rajasthan*) associations.

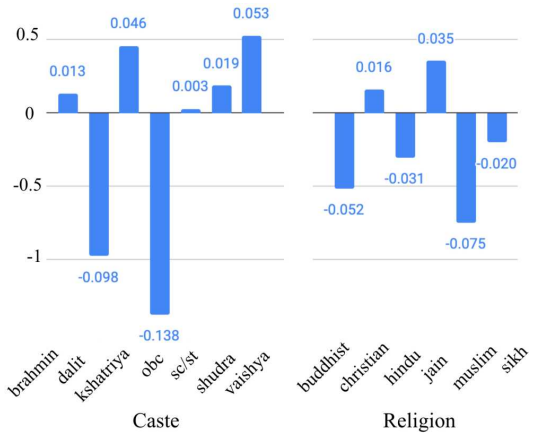

Figure 2: Relative sentiment score shift when caste and religious identities are perturbed showing negative associations with marginalized groups (e.g. *obc*, *muslim*).

turbation sensitivity analysis (Prabhakaran et al., 2019) that reveals biases by counterfactual replacement of terms of same semantic category in natural sentences. For example, the sentence "Gujarati people love food." is perturbed with regional identity terms leading to sentences like "Kashmiri people love food", "Andamanese people love food" etc. We report the normalized shift in sentiment scores for these perturbed sentences, essentially demonstrating the degree to which the scores are affected by the identity term present in the sentence.

For this analysis, we extract sentences in which an identity term occurs from IndicCorp-en (Kunchukuttan et al., 2020), and randomly select equal number of sentences for every identity term to prevent the topical content from being biased towards any subgroup. We extract 10, 150, & 200 sentences, totalling in 357, 1050, and 1200 sentences along region (some region terms had less than 10 sentences), caste, and religion respectively.

Figure 1 shows the shift in scores for regional identities. We find *Mizoram* and *Telangana* have among the most negative score shifts, while *Rajasthan* and *Gujarat* had among the most positive association. Figure 2 shows the relative shift for caste and religion. For caste, the model had significant negative association towards the terms *obc* and *dalit*, both of which represent historically marginalized groups; and for religion, we find negative association towards the terms *muslim* and *hindu*, while *jain* and *christian* have positive associations.

4.2 Personal Names

Personal names *can be* strong proxies for various socio-demographic identity groups in India, including gender, religion, caste, and regional ethnolinguistic identities (Sambasivan et al., 2021). We curate a list of Indian first names with prototypical binary gender association . We build this list by querying the MediaWiki API using a seed list of Wikipedia category pages listing Indian names.[12]

We now perform analysis of gendered correlation in pretrained models using the DisCo metric (Webster et al., 2020) which measures if the predictions of a language model have disproportionate association to a particular gender. Following Webster et al. (2020), we perform slot filling using a set of templates and names, and record the number of candidate words generated by the language model having statistically significant association with a gender, averaged over the number of templates. A higher value for DisCo metric means more associations. We analyze two language models: MuRIL (Khanuja et al., 2021) and multilingual BERT (mBERT) (Devlin et al., 2018a). MuRIL uses the same architecture as mBERT, but is trained on more data derived from the Indian context, and significantly outperforms mBERT on multiple benchmark tasks for Indian languages, including 20% improvement in NER.

We calculate DisCo metric in two ways: (1) using a list of 300 American male and female names (such as, *Mary*, *John*) and (2) using 300 Indian male and female names (such as, *Rahul*, *Pooja*).

Results in figure 3 leads to 2 observations. First, in line Webster et al. (2020), gender bias is encoded for personal names in the Indian context. Second, India-specific resources are critical to bias evaluation. This is because, using American names, it appears that MuRIL has a lesser amount of bias than mBERT. However, using Indian names reveals that

[12] https://tinyurl.com/wiki-indian-names

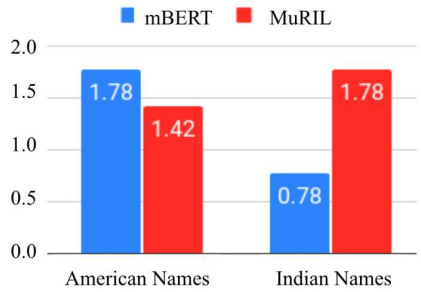

Figure 3: DiSCO metric (higher value means more gendered correlations) mBERT and MuRIL

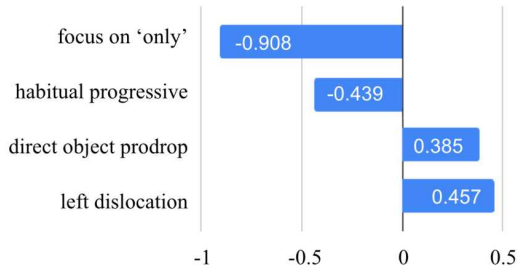

Figure 4: Relative sentiment score shift showing model sensitivity to dialectal features of Indian English

while MuRIL learned to detect names better (i.e., improved NER performance), it also learned more stereotypical associations around those names.

4.3 Dialectal Features

Presence of dialectal features is often associated with demographic subgroups (like socio-economic class (Bernstein, 1960; Kroch, 1978)), and hence can act as a proxy for many axes. Dialects are not monolithic; distinctions are often captured by the presence, absence, and frequency of many features (such as, *article omission*) (Demszky et al., 2021). For this study, we use the minimal pairs dataset built by (Demszky et al., 2021) with 266 sentences annotated with presence of 22 morpho-syntactic dialectal features prevalent in Indian English. For each sentence with a dialect feature, the dataset also contains an equivalent sentence without the feature; effectively functioning as a counterfactual dataset for dialect features. We run this dataset through the sentiment model described earlier, and assess its sensitivity to the presence of dialect features.

We find the sentiment model is sensitive to the presence/absence of dialect features. However, there was no overall trend in any one direction. Figure 4 shows the top 2 features in terms of score shift in either direction; refer to Appendix A for full results. The presence of certain dialect features like *left dislocation* (e.g., "*my father*, he works for a solar company") causes a positive shift in sentiment score while other dialect features like the use of *only* to signify focus (e.g., "I was there yesterday *only*") shifts the score in the negative direction. Although it is difficult to infer systematic patterns of model behaviour due to the small number of sentences in this analysis, the high sensitivity to dialectal features prevalent in the Indian context is concerning in a fairness perspective. Finally, we note that this analysis is w.r.t to dialects of Indian vs western English. However, within India, dialects are not monolithic and resources to map dialectal features to social identities are needed to perform similar analysis for dialectal features within India.

5 Stereotypes in Indian Context

We now turn our attention to the prevalence of social stereotypes from the Indian society in NLP data and models. There is limited literature and resources on social stereotypes in the Indian context, as outlined in Section 2. Notably, de Souza (1977) reported stereotypes around region and religion subgroups in India. They report the top 5 and bottom 5 traits that participants associate with 11 regional and 4 religious identities. But, the study is narrowly scoped to limited adjectives and is from decades ago thus may not reflect the current Indian society. Recent research within NLP has built large stereotype datasets such as Stereoset (Nadeem et al., 2020) and CrowS-P (Nangia et al., 2020) to evaluate models, but they may not capture the stereotypes relevant to India.

We build a set of stereotypical associations based on prior work but employing Indian annotators. Like (de Souza, 1977), we focus on the *Region* and *Religion*. This choice is motivated by the availability of resources and the challenges in studying the other axes (outlined in Section 6). We then use the stereotypes reported by de Souza (1977) and our created dataset to analyse NLP corpora and models for the prevalence of these stereotypes.

5.1 Dataset Creation

We build a dataset of tuples (i, t) where i is an identity term, and t is a word token that represents a concept that is stereotypically associated (or not) with i, for instance, (*Bihari*, *labourer*).

Generating Candidate Associations: We build the set of candidate association tuples (i, t) using identity terms described in Section 4 for re-

ligion and region. We then create a list of tokens based on prior work (Malik et al., 2021; Nangia et al., 2020; Nadeem et al., 2020); including lists of professions, subjects of study (*history*, *science*, etc.), action-verbs, and adjectives for behaviour, socio-economic status, food habits, and clothing preferences. Tuples are formed by a cross product between tokens and identity terms. Since this cross product gives a prohibitively large number of tuples, we further prune this list by including only those tuples that co-occur (are present in the same sentence) in IndicCorp-en (Kunchukuttan et al., 2020) which contains 54M sentences from Indian news and magazine articles and hence likely to reflect the stereotypes prevalent in the Indian public discourse. Tuples with tokens appearing with all identity terms of a given axis are removed.

Obtaining stereotype annotations: We now obtain annotations for each tuple (i, t), where an annotator chooses if the association is *Stereotypical* or *Non-Stereotypical*. The question to the annotator was "Do you think this is a Stereotype widely held by the society?", and thus their annotations reflect community-held opinion, rather than their personal beliefs. They could also mark a tuple as *Unsure*.

We recruited six annotators with diverse gender and region identities: 3 male, 3 female, 2 each from the North east and Central India, and 1 each from West and South India. Virtual training sessions were held to explain the task with examples. We first conducted a pilot where each annotation required a justification which were reviewed by the authors, and any misconceptions were clarified. The annotators were paid 1$ per 3 tuples.

We are interested in building a "high precision" dataset that captures associations that are highly likely to be stereotypes held by a large portion of the society. Hence, we performed the annotation in two phases. First, each tuple is annotated by 3 annotators. The second phase is performed only for the tuples that are labeled stereotypical by at least 2 annotators in phase 1. We retain individual annotations in the dataset to capture potential differences in annotator behavior owing to their socio-cultural background and lived experiences (Prabhakaran et al., 2021). For the analysis presented in this paper, we report results at different levels: S>=1, S>=2, & S>=3, where S denote the number of annotators who marked the tuple as stereotypical.[13] Our resource is both larger in size (See table 1), and captures more diverse perspectives as compared to de Souza (1977). There is only a minimal overlap (10 tuples) between the set of tuples. Table 2 shows some example tuples from our data and the number of annotators who labeled it Stereotypical.

[13]Too few tuples had S>= 4,5,6 to gain reliable insights.

	S=0	S>=1	S>=2	S>=3	Total
Region	2083	473	86	15	2556
Religion	692	604	229	52	1296

Table 1: Number of tuples in our dataset marked as stereotypical by 0, >=1, >=2, >=3 annotators.

Tuple (identity term, attribute token)	Num. S
Region	
(tamilian, mathematician)	6
(marwari, business)	6
(bengali, poet)	5
(punjabi, farmer)	4
(bihari, labourer)	4
(bihari, farmer)	3
(punjabi, army)	3
(rajasthani, dance)	3
Religion	
(christian, missionary)	6
(hindu, pandit)	6
(jain, vegetarian)	5
(muslim, butcher)	5
buddhist, calm)	3
(buddhist, kind)	3
(muslim, terrorist)	3
(sikh, angry)	3

Table 2: Example tuples from our dataset with number of annotators who labeled them as Stereotypical (S).

5.2 Corpus Analysis

Data can be a primary source of biases in LMs (Bender et al., 2021), so we analyze prevalence of stereotypical tuples in large corpora used to train LMs. We analyze the Wikipedia corpus used to train LMs like BERT (Devlin et al., 2018b), and the IndicCorp-en corpus used in training multilingual models like IndicBERT (Kakwani et al., 2020). We measure co-occurrence counts (CC), where a tuple is considered co-occurring if both the identity term (or its plural form) and the token (or one of its inflections) occur in the same sentence.[14]

In the analysis using tuples from de Souza (1977) (Figure 5 - top row) we find co-occurrence counts

[14]We obtain similar trends for nPMI (Aka et al., 2021) metric, and a window size of 2, i.e., co-occurrence within the two tokens before/after the identity term .

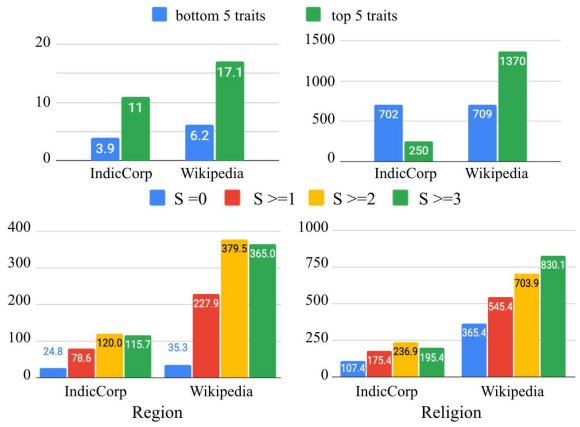

Figure 5: Average co-occurrence of tuples from de Souza (1977) (top row) and our dataset (bottom row) in IndicCorp-en and Wikipedia

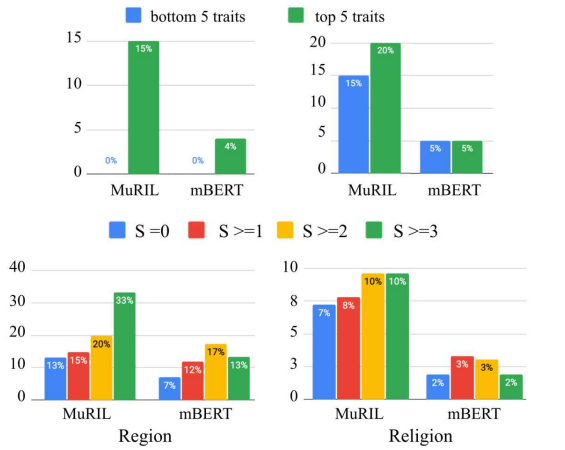

Figure 6: Percentage of tuples from de Souza (1977) (top row) and our data (bottom row) in top 5 predictions of mBERT and MuRIL

are higher for tuples representing top 5 traits compared to bottom 5 traits,[15] We observe similar trend for our dataset (Figure 5 - bottom row). Tuples that all annotators agreed to be not stereotypes (i.e., S=0) have the lowest co-occurrence counts. The average co-occurrence counts increase as more number of annotators mark the tuple as stereotype. The co-occurrence counts in Wikipedia are consistently higher, likely due its larger size as compared to IndicCorp-en (174M vs 54M sentences). In summary, we find that stereotypical associations are preferentially encoded in both corpora.

5.3 Model Analysis

Following previous work (Webster et al., 2020; Hutchinson et al., 2020), we probe MuRIL and mBERT with the task of predicting the masked token in a sentence. We hand-craft templates for each category of tokens in our list. For e.g, a template for the profession category of tokens is: "[i_t] are most likely to work as <MASK>." [16] For each tuple (i, t), we replace i_t in the template with identity term i and record if the token t, or its inflections occur in the top K (K=5)[17] predictions of the model.

Figure 6 show the percentage of tuples occurring in top 5 predictions for the de Souza (1977) and our dataset. Similar to corpus analysis, for tuples from de Souza (1977), we find that the top 5 associated traits are more likely to appear in model predictions as compared bottom 5 traits for both MuRIL and mBERT. For the dataset we built, the percentage of tuples appearing in top 5 model predictions increase as more annotators label the tuple as Stereotype.[18] We also find that MuRIL shows consistently higher percentage of Stereotypical tuples in top 5 predictions suggesting that it has learned more stereotypes in the Indian context due to data sourced from India.

5.4 Limitations

While our dataset can serve as a starting point in evaluation and development of more such datasets, it is not meant as an exhaustive resource for this purpose. First of all, we capture only two axes of disparities: region and religion, and in English. We attempted to collect data for gender identity and caste, but these efforts did not yield reliable results, possibly because of the annotator pool not having the necessary familiarity with those marginalized groups and their lived experiences. Our approach towards filtering the set of tuples for annotation based on co-occurrence limit our data to only capture those stereotypes that are explicitly mentioned in text, but there might exist stereotypes in society that are not captured in corpora and hence will not be captured by our dataset. Additionally, our methods may not capture Stereotypes that are implicit or beyond our token categories.

6 Re-contextualizing Fairness

Given the empirical demonstration of biases in the Indian context in data and models, we now return to the broader agenda for re-contextualizing NLP

[15] One tuple for religion had very high co-occurrence in the IndicCorp-en corpus, resulting in the flipped trend.

[16] Complete list of templates is available with the resources.

[17] We saw similar trends for K=3, 10, 25, 50

[18] S>=3 for mBERT is an exception, with a slight dip, we leave a detailed analysis of this to future work.

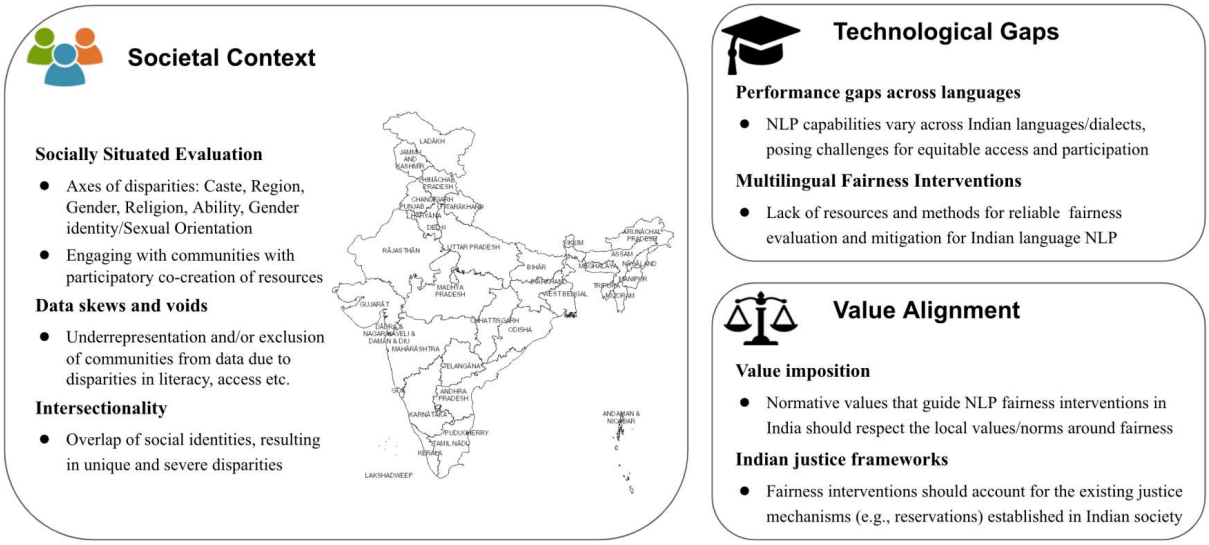

Figure 7: A holistic research agenda for NLP Fairness in the Indian context: accounting for societal disparities in India (Section 3-5), bridging technological gaps in NLP capabilities/resources, and adapting fairness interventions to align with local values and norms (Section 6). (Map source: https://indiamaps.gov.in/soiapp/)

fairness. We re-frame the agenda of Sambasivan et al. (2019) along three aspects: accounting for *Social Disparities*, bridging *Technological gaps*, and adapting to *Values & Norms*.

6.1 Accounting for Indian *Societal* context

We provided a comprehensive account of prominent axes of disparities in Indian society (Section 3), and demonstrated biases around them encoded in NLP data and models (Section 4-5). Our work is just the first step and is far from over.

Socially Situated Evaluation: Most of our analysis is focused on region and religion. A major hurdle in expanding axis coverage is the (lack of easy) access to diverse annotator pools who have familiarity and/or lived experiences of the marginalized groups especially as the public discourse around (dis)ability, gender identity and sexual orientation is relatively new and limited. We believe that participatory approaches (Lee et al., 2020) to create resources for fairness evaluation will be crucial for meaningfully addressing this gap.

Data Voids: Social disparities in literacy and internet access might cause entire communities to be excluded from language data (Sambasivan et al., 2021). Further, the risk of unintentionally excluding marginalized communities based on dialect or other linguistic features while filtering data to ensure quality (Dodge et al., 2021; Gururangan et al., 2022) is even higher in the Indian context because of very limited computational representation of marginalized communities. Accounting for data voids and intentional data curation (such as by collecting language data specifically from marginalized communities (Abraham et al., 2020; Nekoto et al., 2020)) can significantly help bridge this gap.

Intersectionality: Due to the interplay of all the diverse axes in the Indian context, intersectional biases (Collins and Bilge, 2020) experienced by different marginalized groups are often more severe (Sabharwal and Sonalkar, 2015). 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 are also disparate. Bias evaluation and mitigation interventions should account for these intersectional biases.

6.2 Bridging cross-lingual *Technological* gaps

While we focus on English language data and models in this paper, it is crucial to mitigate the gaps in NLP capabilities and resources across Indian languages, both in general and for fairness research.

Performance gaps across languages: India is a vastly multilingual country with hundreds of languages and thousands of dialects. But there are wide disparities in NLP capabilities across these languages and dialects. These disparities pose a major challenge for 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

addressing this gap in recent years (Khanuja et al., 2021), 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 bias evaluation resources and while we present such resources for the Indian context, we limited our focus to only English. It is crucial to expand this effort into Indian languages, along the lines of recent work on Hindi, Bengali, and Telugu (Malik et al., 2021; Pujari et al., 2019). This is especially important since biases may manifest differently in data and models for different languages. Additionally, 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 dimension of language brings.

6.3 Adapting to Indian *Values and Norms*

Fairness interventions essentially impart a normative value system on model behaviour. It is crucial to ensure that these interventions are not at odd with Indian values, norms, and legal frameworks.

Accounting for Indian justice models: India has established legal restorative justice measures for resource allocation, colloquially known as the "reservation system" (Ambedkar, 2014), where historically marginalized communities (like Dalits, backward castes, tribals, and religious minorities) are afforded fixed quotas in educational and government institutions to counter historical deprivation. NLP fairness interventions should conform to these established measures that are otherwise non-existent, and hence not thought for in the West.

Avoiding value imposition: Fairness inquiries answer questions such as: what fairness means, and how fair is fair enough? These questions, and their answers risk value imposition. While, implicitly these answers draw largely from Western values rooted in egalitarianism, consequentialism, deontic justice, and Rawls' distributive justice (Sambasivan et al., 2021), the philosophy of fairness in India is rooted in social restorative justice. More work should look into such value alignment challenges for fairness interventions (Gabriel, 2020).

7 Conclusion

In this paper, we holistically re-contextualize fairness research for the Indian context taking an NLP-centric lens to Sambasivan et al. (2021). We lay out a research agenda advocating to account for the societal context in India, bridge technological gaps in capability and resources, and align with local values and norms (Section 6). Our focus here is on India, but the broader framework of this work can be used to recontextualize fairness for any geo-cultural context. We outline the prominent axes of disparities in India (Section 3), and demonstrate biases around them in NLP models and corpora. To summarize: First, our perturbation analysis reveals that sentiment model predictions are significantly sensitive to regional, religious, and caste identities (Section 4.1), and dialectal features (Section 4.3). Second, our DisCo analysis shows the necessity of India-specific resources for revealing biases in the Indian context (Section 4.2). Third, we build a stereotype dataset for the Indian context and demonstrate preferential encoding of stereotypical associations in both NLP data and models (Section 5). While there is more work to be done, we believe this is an essential first step towards a meaningful NLP fairness research agenda for India.

8 Ethical considerations

We build resources to demonstrate biases in models, these resources alone are insufficient to capture all the undesirable biases in the Indian society. As described in Section 5.4, our dataset lacks coverage across the various Indian axes of disparities, languages, and reflects the judgements of a small number of annotators. Hence, they should be used only for diagnostic and research purposes, and not as benchmarks to prove lack of bias. We also urge that the list of names with prototypical binary gender associations from Wikipedia (used in Section 4.2) not be used to train gender prediction models.

Acknowledgements

We thank Nithya Sambasivan for her groundbreaking research and early guidance on this project. We thank Ben Hutchinson, Kellie Webster, Ding Wang, Molly FitzMorris, and Reena Jana for their critical insights on earlier drafts. We are grateful to the anonymous reviewers for their helpful feedback. We thank Dinesh Tewari for his work on facilitating the project. We thank the annotation team for facilitating our data collection.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models.

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. 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. European Language Resources Association.

Kuruvilla C Abraham and Ajit K Abraham. 1998. Homosexuality: some reflections from india. *The Ecumenical Review*, 50(1):22.

Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. 2021. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–335.

BR Ambedkar. 2014. *Annihilation of Caste: The Annotated Critical Edition*. Verso Books.

Pooja V Anand. 2016. Attitude towards homosexuality: A survey based study. *Journal of Psychosocial Research*, 11(1):157.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Basil Bernstein. 1960. Language and social class. *The British journal of sociology*, 11(3):271–276.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. 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.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Ramdas Borude. 1966. Linguistic stereotypes and social distance. *Indian Journal of Social Work*, 27(1):75–82.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Vandana Chaudhry and Tom Shipp. 2005. Rethinking the digital divide in relation to visual disability in india and the united states: towards a paradigm of information inequity. *Disability Studies Quarterly*, 25(2):2.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

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

Mallika Das. 2011. Gender role portrayals in indian television ads. *Sex Roles*, 64(3):208–222.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Thomas A. de Souza. 1977. Regional and communal stereotypes of bombay university students. *Indian Journal of Social Work*, 38(1):37–44.

Dorottya Demszky, Devyani Sharma, Jonathan H Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338.

Ashwini Deshpande. 2011. *The grammar of caste: Economic discrimination in contemporary India*. Oxford University Press.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding.

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*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. 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.

António Filipe Fonseca, Sohhom Bandyopadhyay, Jorge Louçã, and Jaison A Manjaly. 2019. Caste in the news: a computational analysis of indian newspapers. *Social Media+ Society*, 5(4):2056305119896057.

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

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Anita Ghai. 2002. Disabled women: An excluded agenda of indian feminism.

Anita Ghai. 2019. *Rethinking disability in India*. Routledge India.

Nandini Ghosh. 2016. *Interrogating Disability in India*. Springer.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media.

Michael Griffin, K Viswanath, and Dona Schwartz. 1994. Gender advertising in the us and india: Exporting cultural stereotypes. *Media, Culture & Society*, 16(3):487–507.

Gauri Gupta, Krithika Ramesh, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation.

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

Thongkholal Haokip. 2021. From 'chinky' to 'coronavirus': racism against northeast indians during the covid-19 pandemic. *Asian Ethnicity*, 22(2):353–373.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities.

Christopher J Johnstone, Sandhya Limaye, and Misa Kayama. 2017. Disability, culture, and identity in india and usa. In *Inclusion, Disability and Culture*, pages 15–29. Springer.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Anant Kamath. 2018. "untouchable" cellphones? old caste exclusions and new digital divides in peri-urban bangalore. *Critical Asian Studies*, 50(3):375–394.

Durriya HZ Khairullah and Zahid Y Khairullah. 2009. Cross-cultural analysis of gender roles: Indian and us advertisements. *Asia Pacific Journal of Marketing and Logistics*.

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

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*, pages 43–53.

Vijesh V Krishna, Lagesh M Aravalath, and Surjit Vikraman. 2019. Does caste determine farmer access to quality information? *PloS one*, 14(1):e0210721.

Anthony S Kroch. 1978. Toward a theory of social dialect variation. *Language in society*, 7(1):17–36.

S. Ganesh Kumar, Gautam Roy, and Sitanshu Sekhar Kar. 2012. Disability and rehabilitation services in india: Issues and challenges.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for Indic languages. *arXiv preprint arXiv:2005.00085*.

Abhishek Kurian. 2020. Sex, laws and inequality : comparison between India and the U.S.A. https://blog.ipleaders.in/sex-laws-inequality-comparison-india-u-s Accessed: 2022-04-29.

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

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. 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.

Kittipong; Mehta Nikhil; Leingpibul, Thaweephan; Laosethakul and Anju Mehta. 2006. The cross cultural study concerning gender stereotyping in computing: Comparison between the us and india. *AMCIS 2006 Proceedings*.

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

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275.

Duncan McDuie-Ra. 2012. *Northeast migrants in Delhi: Race, refuge and retail*. Amsterdam University Press.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

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. 2020. 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. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. 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.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. 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. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ashwin Rajadesingan, Ramaswami Mahalingam, and David Jurgens. 2019. Smart, responsible, and upper caste only: measuring caste attitudes through large-scale analysis of matrimonial profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 393–404.

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

Nithya Sambasivan, Nova Ahmed, Amna Batool, Elie Bursztein, Elizabeth Churchill, Laura Sanely Gaytan-Lugo, Tara Matthews, David Nemar, Kurt Thomas, and Sunny Consolvo. 2019. Toward gender-equitable privacy and security in south asia. *IEEE Security & Privacy*, 17(4):71–77.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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*, pages 1668–1678.

Kartik Sawhnet. Tracing the portrayal of disability in indian cinema.

Renuka R Sethi and Mary J Allen. 1984. Sex-role stereotypes in northern india and the united states. *Sex roles*, 11(7):615–626.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *CoRR*, abs/2105.04054.

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

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can existing methods debias languages other than english? first attempt to analyze and mitigate japanese word embeddings. In *GEBNLP*.

Nisha Tamang. 2020. Section 377: Challenges and changing perspectives in the indian society. *Changing Trends in Human Thoughts and Perspectives: Science, Humanities and Culture Part I*, page 68.

Palashi Vaghela, Ramaravind K Mothilal, and Joyojeet Pal. 2021. Birds of a caste-how caste hierarchies manifest in retweet behavior of indian politicians. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–24.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

A Perturbation Sensitivity Analysis with dialectal features: full results

In §4.3 we perform perturbation sensitivity analysis with sentences from Demszky et al. (2021). Here we provide the complete results for this analysis, where in-text we provided only the top-2 most positively shifted and negatively shifted features.

Dialectal Feature	**Relative sentiment score shift**
focus 'only'	-0.908
habitual progressive	-0.439
inversion in embedded clause	-0.412
topicalized non-argument constituent	-0.205
lack of copula	-0.029
stative progressive	-0.019
invariant tag ('isn't it', 'no', 'na')	-0.010
focus 'itself'	-0.007
resumptive object pronoun	0.000
non-initial existential 'X is / are there'	0.004
resumptive subject pronoun	0.009
mass nouns as count nouns	0.009
article omission	0.023
preposition drop	0.025
lack of inversion in wh-questions	0.036
extraneous 'the' (often generic) or 'a'	0.084
prepositional phrase fronting	0.186
object fronting	0.192
use of 'and all'	0.208
lack of agreement	0.274
direct object prodrop	0.385
left dislocation	0.457

Table 3: Relative sentiment score shift due to presence or absence of dialectal features