Fairness in Language Models Beyond English: Gaps and Challenges

Krithika Ramesh, Sunayana Sitaram, Monojit Choudhury
Microsoft Corporation
{t-kriramesh, sunayana.sitaram, monojitc}@microsoft.com

Abstract

With language models becoming increasingly ubiquitous, it has become essential to address their inequitable treatment of diverse demographic groups and factors. Most research on evaluating and mitigating fairness harms has been concentrated on English, while multilingual models and non-English languages have received comparatively little attention. In this paper, we survey different aspects of fairness in languages beyond English and multilingual contexts. This paper presents a survey of fairness in multilingual and non-English contexts, highlighting the shortcomings of current research and the difficulties faced by methods designed for English. We contend that the multitude of diverse cultures and languages across the world makes it infeasible to achieve comprehensive coverage in terms of constructing fairness datasets. Thus, the measurement and mitigation of biases must evolve beyond the current dataset-driven practices that are narrowly focused on specific dimensions and types of biases and, therefore, impossible to scale across languages and cultures.

1 Introduction

Language models are known to be susceptible to developing spurious correlations and encoding biases that have potentially harmful consequences in downstream tasks. Whilst prior work has documented these harms (Dev et al., 2021) (Bender et al., 2021) (Kumar et al.), there remains much to be studied and criticism for the existing research (or lack thereof) that remains to be addressed.

In the context of language models, fairness can manifest in two forms; *representational* and *allocational* harms. **Representational** harms generally refer to cases where demographic groups end up being misrepresented. This includes stereotypes and negative associations with these groups and even a lack of acknowledgment of certain groups that are underrepresented in the data. **Allocational** harms, on the other hand, refer to the inequitable distribution of resources and opportunities to groups with different demographic attributes associated with them. The nature of allocational harms can vary based on the **sociocultural, economic, and legal** settings where the system has been deployed. However, it can also take shape in terms of the model's functionality across languages with fewer resources (Choudhury and Deshpande, 2021; Liu et al., 2021). While current literature adopts a Euro-American-centric view of fairness, work such as Sambasivan et al. (2021) pushes to recognize algorithmic fairness from a more inclusive lens.

Bias crops up in multiple steps of the pipeline (Hovy and Prabhumoye, 2021) (Sap et al., 2022), including the annotation process, the training data, the input representations, model architecture, and the structure of the research design. Thus, measures to mitigate bias in one of these components alone will likely not suffice as a corrective measure, necessitating human intervention at different stages of the pipeline.

Most work that addresses fairness in NLP addresses it from an Anglo-centric context, with comparatively significantly less work done in grammatically-gendered and low-resource languages. Their inability to capture social and cultural nuances and demographic variations is well-documented (Talat et al., 2022). Despite this, they are ubiquitous, with applications ranging diverse fields, from legal contexts to healthcare. That said, there is insufficient documentation of the harms that could stem from unfair models trained for downstream tasks involving natural language generation, despite Arnold et al. (2018); Bhat et al. (2021); Buschek et al. (2021) indicating the influence of these systems on users. Apart from this, these NLP systems also reinforce and reproduce the social and racial hierarchies observed in society and fail to recognize underrepresented communities that are already marginalized (Dev et al., 2021;

Lauscher et al., 2022b). The ramifications of neglecting these issues are diverse and far-reaching, from minor inconveniences for users in less harmful contexts to compromising their privacy as well as depriving them of opportunities and resources (Cirillo et al., 2020; Köchling and Wehner, 2020).

Finally, while the interplay and tradeoff between privacy, efficiency, and fairness in tabular data has received extensive examination (Hooker et al., 2020; Lyu et al., 2020) comparatively fewer studies have been conducted in NLP (Tal et al., 2022; Ahn et al., 2022; Hessenthaler et al., 2022).

The contributions of this work center around drawing attention to the current state of research on fairness in the context of linguistic and cultural issues in non-English languages and in the context of multilingual models. While thorough survey studies such as Sun et al. (2019); Stanczak and Augenstein (2021); Bhatt et al. (2022) yield valuable insights into some of these aspects, none address the current state of the work in multilingual fairness. Our paper provides insights into the following:

- This work surveys and presents challenges and unanswered questions with respect to fairness in both monolingual and multilingual NLP.

- We analyze bias from both a linguistic and cultural lens for non-English languages and present a comprehensive overview of the literature in bias pertaining to grammatically gendered languages and multilinguality.

- We bring to the forefront challenges in multilingual fairness and begin a dialogue for creating more equitable systems for multilingual NLP.

2 Bias in Monolingual Setups for English

2.1 Metrics for Measurement

Prior to delving into the complexities of fairness in multilingual systems, it is essential to first examine the prevalent biases and challenges in monolingual systems. By prefacing the discussion on bias in multilingual systems with an overview of the current state of fairness evaluation and identifying areas for improvement, we aim to shed light on the potential for similar issues to arise in multilingual systems, as many of the biases present in monolingual systems are likely to persist in multilingual contexts. Some of the initial work on analyzing biases in NLP models (Bolukbasi et al., 2016) propose quantitative measures of evaluating bias in word embeddings. Broadly speaking, bias measures are subcategorized into *i) intrinsic* and *ii) extrinsic* measures. Intrinsic metrics quantify bias in the model's pre-trained representations, whereas extrinsic metrics deal with bias observed in the outputs of the downstream task the model is trained for.

Caliskan et al. (2017); May et al. (2019); Nadeem et al. (2021); Nangia et al. (2020) are commonly used in papers evaluating language models for fairness. Caliskan et al. (2017) proposes the Word Embedding Association Test (WEAT). A fundamental criticism of WEAT is that it can be exploited to overestimate the bias in a model (Ethayarajh et al., 2019). The Sentence Encoder Association Test (SEAT) metric (May et al., 2019) was proposed to address WEAT's limitation of measuring bias only over static word embeddings. SEAT is an adaptation of WEAT that allows us to measure bias over contextualized embeddings.

StereoSet (Nadeem et al., 2021), and CrowS-Pair (Nangia et al., 2020) are crowdsourced datasets specifically geared toward measuring the model's stereotypical proclivity over multiple dimensions, which are inclusive of gender, race, and religion, among others. Blodgett et al. (2021) points out the flaws in the data quality, such as invalid stereotype/anti-stereotype pairs, reliance on indirect group identifiers as a proxy for demographic identification, and logical incongruities in the sentence pairs.

Several other intrinsic measures and adaptations of the aforementioned ones have also been proposed (Kurita et al., 2019; Webster et al., 2020; Kaneko and Bollegala, 2021; Lauscher et al., 2021). Recent studies (Delobelle et al., 2022; Meade et al., 2022) that perform comparative evaluations across these measures provide valuable insights into how and where the metrics can be used, along with their potential drawbacks.

2.2 Intrinsic vs Extrinsic Evaluation

While intrinsic measures are valuable in that they indicate the existence of representational bias in systems, the current literature on fairness evaluation largely concentrates on intrinsic metrics alone. Considerably less work has been done on addressing bias in extrinsic evaluation, with several downstream tasks needing concrete metrics to evaluate bias in their outputs. This is a pressing issue due to the lack of correlation between intrinsic and extrin-

sic measures (Goldfarb-Tarrant et al., 2020; Cao et al., 2022; Delobelle et al., 2022). As emphasized in Orgad and Belinkov (2022), incorporating extrinsic evaluation measures is crucial for several reasons, including the greater relevance of these metrics to bias mitigation objectives. Aside from this, evaluating fairness on the downstream task's outputs allows us to gauge more precisely how a particular demographic may be affected by the biases in the system.

Although work done in fairness evaluation in NLP primarily concentrates on monolingual studies, there remain several unanswered questions and inconclusive results. For instance, although May et al. (2019) claims to use semantically bleached templates, experiments in Delobelle et al. (2022) suggest that they retain some degree of semantic significance. While several bias evaluation methods use template-based data, recent findings (Alnegheimish et al., 2022) suggest that this approach may be unreliable and advocate the use of natural sentence prompts.

2.3 Fairness From the Lens of Multiple Social Dimensions

The focus of much of the existing body of literature is on gender bias, with little that covers other dimensions like race and religion. Evaluation metrics should be able to evaluate harms in language models over the intersectionality of multiple identities, akin to what would realistically be expected in real-world data. While previous research (Talat et al., 2022; Kirk et al., 2021) has emphasized the importance of fairness evaluation and mitigation over intersectional identities, there is relatively sparse work that attempts to address the same (Tan and Celis, 2019; Subramanian et al., 2021; Hassan et al., 2021; Lalor et al., 2022; Câmara et al., 2022). It is also crucial to gauge if reducing bias across one dimension could affect biases in the other dimensions. Most fairness measures do not account for the intersectionality of identities and standards of justice outside the predominantly Western sphere of distributive justice (Sambasivan et al., 2021; Lundgard, 2020).

Whilst there has been an increase in proposing novel methods to mitigate bias in language models, there needs to be more work in benchmarking these debiasing techniques to assess their relative effectiveness. Meade et al. (2022) represents a step forward in this direction. Despite criticism (Etha-

yarajh et al., 2019; Blodgett et al., 2021) of some evaluation metrics, they are still consistently used (and not always in conjunction with other metrics) in bias evaluation studies.

3 Linguistic Aspects

The linguistic variations between languages pose additional problems in the realm of multilingual NLP. Take, for example, the concept of gender, which has multiple definitions in linguistic terms (namely, **grammatical, referential, lexical and bio-social gender**) (Stanczak and Augenstein, 2021). Section 3.1 delves into how the grammatically gendered nature of languages can affect bias in multilingual and monolingual spaces alike. **Referential** gender, on the other hand, deals with terms that referentially address a person's gender, such as pronouns. Terms that non-referentially describe gender fall under the umbrella of **lexical** gender, and the **bio-social** definition of gender involves a mixture of phenotypic traits, gender expression, and identity as well as societal and cultural aspects that influence them (Ackerman, 2019).

Although initial forays into this field investigate bias caused by grammatical gender, problems in these systems can also crop up due to the other definitions of gender. Referential gender terms are not always aligned when used in conjunction with lexically gendered terms, particularly with respect to pronoun-based anaphors for queer-identifying individuals. Several default assumptions regarding the individual's gender identity are made as a consequence (Cao and Daumé III, 2021).

There are multiple varying forms of pronoun complexity (Lindström, 2008; Ballard, 1978). Apart from this, there are instances of substantial variations in their linguistic forms even among languages within a specific region, as highlighted in Nair (2013). Linguistics also involves the presence of constructs like deictic pronouns and honorific pronouns (Goddard, 2005), which in some cases can lead to the pronouns used to reference someone changing based on their social dynamic within the community (Lauscher et al., 2022c). These linguistic aspects represent another line of work that must be addressed for lower-resourced communities that communicate using languages that utilize these.

Lexical gender, while non-referential, finds its own challenges due to the variation of these terms across languages. For example, while certain relationships with individuals in a family may have an

exact mapping in other languages, more often than not (particularly with Southeast Asian languages), there is no precise mapping, and the system ends up making an approximation or ignoring the term altogether. Such issues may also be likely to perforate to other axes such as race, religion, caste, and so forth. In particular, considering that one method of training multilingual embeddings relies on alignment-based approaches, it is imperative that we keep in mind how these design choices could affect the representations of these terms.

Whilst utilizing linguistic features in methods to evaluate and mitigate gender bias is a relatively new field of study, previous work has demonstrated that additional linguistic context can result in performance gains (Volkova et al., 2013; Wallace et al., 2014), thus in alignment with the claim from Hovy and Yang (2021) that LMs must utilize social context to be able to reach human-level performance on tasks. Sun et al. (2021) utilizes linguistic features to capture cross-cultural similarities, and thus, to select languages that are optimal for cross-lingual transfer. However, it is essential to acknowledge that languages are susceptible to cultural and linguistic shifts that occur at both global and local levels over time, as noted in Hamilton et al. (2016). Pretrained models also have the capability to embed sociodemographic information, as evinced by Lauscher et al. (2022a).

It has also been noted that other linguistic forms of gender do not translate well to sociological gender (Cao and Daumé III, 2021). Furthermore, the scarcity of non-binary gender options in different languages can lead to the misgendering of non-binary individuals in these languages, as they may be constricted to fit into a binarized definition of sociological gender.

3.1 Grammatically Gendered Languages

Linguistics recognizes multiple forms of gender (Cao and Daumé III, 2020), as observed in grammatically gendered languages where most or all nouns, including those referring to inanimate objects, possess a syntactic concept of gender. These languages can have anywhere between 2 to 20 forms of grammatical gender divisions. There has been an almost exclusive focus on English for evaluating gender bias, even in the setting of monolingual models and systems. English, however, is not a grammatically-gendered language. This may limit the transferability of techniques used for bias

evaluation and mitigation to other languages that are grammatically gendered.

Zhou et al. (2019) examines bias from the view of grammatically gendered languages by decomposing the gendered information of words in the embedding space into two components; i) semantic and ii) syntactic. For instance, the Spanish word for "man" (*hombre*) is both semantically and syntactically gendered. However, the Spanish word for "water" (*agua*) is not semantically gendered but is considered a feminine noun. The proximity of female occupation words to the feminine side and male occupation words to the masculine side of the semantic gender direction suggests the presence of bias in these Spanish embeddings. Zhou et al. (2019) also demonstrates via experiments on bilingual embeddings that, post-alignment, masculine-gendered words are closer to the English equivalent of the occupation words than feminine-gendered ones. The paper also proposes bias mitigation methods and demonstrates that the quality of the embeddings is preserved via word-translation experiments. Nevertheless, the validity of these mitigation measures would need to be verified by testing them on downstream tasks. Gonen et al. (2019) show that grammatical gender affects the word representations in Italian and German and that inanimate nouns end up being closer to words of the same gender. They propose to address this through the precise use of a language-specific morphological tool and a careful approach to removing all the gender signals from a given text.

The grammatical properties of a language might show some interesting properties to be taken into account when dealing with the fairness of large language models, particularly for gender bias. Studies directed toward them could yield insights into observable trends across language families, with Gonen et al. (2019) demonstrating how the alignment of languages in the embedding space is negatively affected by grammatical gender. They could also prove helpful when analyzing bias in multilingual models, where both grammatically gendered and non-gendered languages are aligned to the same embedding space. The research and datasets available for extrinsic evaluation over other languages remain an area with scope for improvement.

Apart from these grammatical properties that affect the results we observe, the translation of existing bias evaluation datasets into other languages to create parallel corpora does not suffice

when dealing with languages apart from English. This is partly because most languages are inherently rooted in cultural context. Any data curated for these languages must incorporate sociocultural and linguistic aspects unique to the language/region. Depriving NLP systems of cultural context could consequently lead to entire axes over which social biases are measured being ignored. The cultural significance of words and phrases in various languages can vary significantly, as demonstrated in Mohamed et al. (2022), as well as in characteristics such as metaphorical tendencies (Gutiérrez et al., 2016) and communication styles (Miehle et al., 2016; Suszczyńska, 1999). Hovy and Yang (2021) includes an overview and critique of this in the current state of NLP literature, which they claim adopts an oversimplified view and focuses on the information content alone while ignoring the social context of this content. Milios and BehnamGhader (2022); España-Bonet and Barrón-Cedeño (2022) illustrate the inefficiency of direct translation methods, and España-Bonet and Barrón-Cedeño (2022) advocates for the creation of culturally-sensitive datasets for fairness assessment. However, Kaneko et al. (2022) proposes a way to generate parallel corpora for other languages that bears high correlation with human bias annotations.

4 Multilingual Models

Multilingual spaces allow the embeddings of multiple languages to be aligned so that the mappings of every word to its equivalent in other languages are close to each in these embedding spaces. There are numerous ways of training multilingual language models (Hedderich et al., 2021) using monolingual and unlabeled data. Multilingual language models can improve cross-lingual performance on low-resource languages leveraging the data available to higher-resourced languages up to a certain number of languages. Beyond a point, however, the performance across these languages on cross-lingual and monolingual tasks begins to dip as the number of languages increases (Conneau et al., 2020). However, few studies explore the impact of multilingual training on biases. Hovy and Yang (2021) illustrate how language and culture share a strong association, and Khani et al. (2021); Sun et al. (2021) reveal that geographical and cultural proximity among languages could enhance the performance of models.

Languages provide much insight into a society's cultural norms, ideologies, and belief systems (Hershcovich et al., 2022; Wilson et al., 2016). Often, the properties unique to a language are not clearly mapped to other languages or even other dialects within a language, with no direct translations available for several phrases and terminology. Whether or not language models can retain this cultural information and context while utilizing information from higher-resourced languages still requires investigation.

4.1 An Outline of Fairness Evaluation in the Context of Multilinguality

Several datasets have been put forward for the purpose of multilingual evaluation, and Table 1 describes these datasets along with details regarding their utility. These include the languages they cover, whether or not they evaluate bias over pretrained representations or a downstream task, and the downstream tasks and dimensions they cater toward.

Zhao et al. (2020) was among the first papers to quantify biases in multilingual spaces and does so using both extrinsic and intrinsic evaluation techniques. Their findings indicate that some factors that influence bias in multilingual embeddings include the language's linguistic properties, the target language used for the alignment of the embeddings, and transfer learning on these embeddings induces bias. Additionally, there is the possibility that non-Germanic languages do not align well with Germanic ones, and further work would be required to derive conclusions as to how this affects fairness measurements.

Huang et al. (2020) released the first multilingual Twitter corpus for hate speech detection, annotated with the author's demographic attributes (age, country, gender, race/ethnicity), which allows for fairness evaluation across hate speech classifiers. Through experiments, they prove that variations in language, which are highly correlated with demographic attributes (Preoţiuc-Pietro and Ungar, 2018; Osiapem, 2007), can result in biased classifiers. However, there are some promising results from Liang et al. (2020), which proposes a novel debiasing method using Dufter and Schütze (2019). While the multilingual model is originally debiased over English, results show its effectiveness for zero-shot debiasing over Chinese.

Câmara et al. (2022) measures both unisectional and intersectional social biases over gender, race,

Dataset	Languages	Task	Metric	Dimensions
https://github.com/MSR-LIT/MultilingualBias	English, Spanish, German, French	Text Classification	I, E	Gender
https://github.com/xiaoleihuang/DomainFairness	English, Italian, Portuguese, Spanish	Text Classification	E	Gender
https://github.com/kanekomasahiro/bias_eval_in_multiple_mlm	German, Japanese, Arabic, Spanish, Portuguese, Russian, Indonesian, Chinese	Masked Language Modelling	I	Gender
https://github.com/ascamara/ml-intersectionality	English, Arabic, Spanish	Text Classification	E	Gender, Race/Ethnicity, Intersection
https://github.com/liangsheng02/densray-debiasing/	English, Chinese	Masked Language Modelling	I	Gender
https://github.com/xiaoleihuang/Multilingual_Fairness_LREC	English, Italian, Portuguese, Spanish, Polish	Text Classification	E	Age, Country, Gender, Race/Ethnicity
https://github.com/coastalcph/fairlex	English, German, French, Italian and Chinese	Text Classification	E	Gender, Age, Region, Language, Legal Area

Table 1: Datasets for fairness evaluation beyond English. I = Intrinsic, E = Extrinsic

and ethnicity in multilingual language models. This is particularly relevant, as in a practical setting, treating identities as composites of various demographic attributes is a necessity. Kaneko et al. (2022) measures gender bias in masked language models and proposes a method to use parallel corpora to evaluate bias in languages shown to have high correlations with human bias annotations. In cases where manually annotated data doesn't exist, this could prove helpful.

Although there has been research on fairness in multimodal contexts (Wolfe and Caliskan, 2022; Wolfe et al., 2022), in a first-of-its-kind study, Wang et al. (2022) looks at fairness from a multilingual view in multimodal representations. Whilst they find that multimodal representations may be individually fair, i.e., similar text representations across languages translate to similar images, this concept of fairness does not extend across multiple groups.

Talat et al. (2022) expresses criticism over the primary data source for multilingual large language models being English, which they claim is reflective of cultural imperialism. They also advocate for these models to be used only for languages they have been trained for to retain the cultural context unique to a language. The multilingual datasets commonly used tend to be parallel corpora derived directly from English translations, neglecting the socio-cultural nuances specific to a given language, as evidenced by the CommonCrawl corpora (Dodge et al., 2021).

Moreover, recent literature (Al Kuwatly et al., 2020; Parmar et al., 2022; Sap et al., 2022) presents us with yet another potential issue; lack of demographic variation in the annotation of these dataset results could contribute to bias in the pipeline. As of yet, several languages (Aji et al., 2022; Joshi et al., 2020) (such as Hindi, Arabic, and Indonesian, which have tens to hundreds of million of native speakers) have had little to no fairness benchmark-

ing datasets developed for them, an indicator that much remains to be done to develop more equitable language models.

4.2 An Outline of Fairness Mitigation in the Context of Multilinguality

Due to multilingual spaces being a composite of the embeddings of various languages with different linguistic and semantic properties, it would serve mitigation techniques well to consider these differences. Other methods could use these distinctions to reduce bias in downstream tasks. Zhao et al. (2020), for one, show that balancing the corpus and transferring it to a grammatically gendered language's embedding space could reduce bias, and that using debiased embeddings could also aid with bias mitigation.

Huang (2022) takes inspiration from the FEDA domain adaption technique (Daumé III, 2007) to use it to mitigate bias in multilingual text classification and compares this with other mitigation methods. These debiasing baselines involve adversarial training, masking out tokens associated with demographic groups, and instance weighting to reduce the impact of data instances that could lead to more biased classifiers. While Liang et al. (2020) show that zero-shot debiasing can be beneficial for this purpose, further study would be required to ascertain if this is a feasible possibility.

4.3 Problems in Multilingual Evaluation and Mitigation

A major challenge in multilingual fairness is the lack of datasets (including parallel corpora) and literature for evaluation across tasks. Much of the research conducted in monolingual contexts has yet to be replicated in a multilingual setting, which would enable us to determine whether or not bias trends in monolingual spaces are directly transferable to multilingual contexts. Research and data resources also tend to neglect less-represented

demographics, notably those local to a particular region. Further, datasets require thorough documentation, as variations in annotator information can result in different types of biases infiltrating the pipeline (Mohamed et al., 2022; Joshi et al., 2016; Bracewell and Tomlinson, 2012). These could include attitudes towards other cultures and languages, which must be assessed and reported during data collection. Multilingual users speak multiple languages, and there is no work on evaluating bias in language contact settings such as code-switching. Certain axes along which systems may discriminate may be contained to a given region. Due to the underrepresented nature of marginalized identities (such as immigrant communities), models will likely not learn useful representations of these identities.

5 Culture

Language and culture are intrinsically linked with each other. However, NLP research has historically placed a considerable emphasis on the information content of the data, as opposed to the contextual information surrounding the same data. Hovy and Yang (2021) propose a broad taxonomy of 7 social factors that encompasses various aspects of this contextual information. This could be incorporated into models to improve performance and make them aware from a socio-cultural perspective.

The differences between a pair of languages or even a pair of dialects could reflect across multiple attributes; this could lead to variations in language's *phonology, tone, text, and lexical forms*. Some of these attributes are controlled by the speaker and receiver involved. Despite evidence of gains in performance by leveraging these features, systems still retain the potential to discriminate against marginalized communities, as evinced in Sap et al. (2019). This necessitates the proposal of evaluation methods to analyze the potential harms that people from different cultural backgrounds might expose themselves to via the use of such systems.

Multilingualism also entails the need to navigate the nuances of language, including the potential for stereotypes and discriminatory language, which may not have precise equivalents in other languages. Cultural taboos and stereotypes can be highly localized. As an example, pregnant or lactating women are discouraged from consuming nutritious food in certain cultures (Meyer-Rochow,

2009). Such contextual information might be underrepresented or nonexistent in the data that the model is exposed to. While some culture-specific behaviors may be prohibited or frowned upon in some parts of the world, there are yet other places that may encourage or remain indifferent to these very same behaviors.

Additionally, the axes we consider require to be treated differently in different cultural and linguistic settings. Take, for instance, gender. While gender has, for the most part, been treated as a binary variable in these studies, this does not echo what is observed in real-world settings, where several individuals have non-binary gender identities (Devinney et al., 2022). Non-binary gender identities encompass a broad spectrum of gender identities, and the term is generally considered an umbrella term for any identity outside the binary. The inability of models to incorporate this additional information on gender has subsequently led to them developing meaningless representations of non-binary genders in text (Dev et al., 2021). This translates to the systematic erasure of their identities. Baumler and Rudinger (2022) show that much remains to be done concerning addressing non-binary identities outside the Western context. For instance, several non-binary identities, such as the Aravanis and the Māhūs (local to India and Hawaii, respectively) are likely to have little to no meaningful coverage in the training data of the models. These identities can also have unanswered nuances in literature; for example, the Acaults of Myanmar do not consider transsexualism, transvestism, and homosexuality to be distinct categories. This is also applicable to languages such as Arabana-Wangkangurru, which make use of deictic pronouns (previously discussed in Section 3) (Lauscher et al., 2022c; Hercus, 1994).

Further, given that models are highly suscept to the kind of data they are trained on, it is unlikely that our models can recognize that certain forms of prejudice are more frequent in specific socio-cultural environments than others. The targets of this discrimination are also likely to vary from region to region, another nuance that models find difficult to account for. India and Nepal, for instance, are two countries that still suffer from the effects of the hierarchy of a historically caste-based society that (despite sharing similar roots) bear differences in terms of representation of the various castes and how they are referred to (Jodhka et al., 2010; Rao,

2010). It is important to note that the ability of a system to incorporate information from these social factors to mitigate biases is task-dependent. Downstream tasks like machine translation and dialogue/response generation may depend more on cues related to speaker and receiver characteristics from the taxonomy proposed in Hovy and Yang (2021) than other tasks. Extrinsic metrics for machine translation focus primarily on the gender bias of the mappings of nouns and pronouns from one language to another (Cho et al., 2019). On the other hand, more open-ended, subjective tasks like NLG are prone to encoding underlying biases and stereotypes across multiple axes and reproducing these in their outputs (Henderson et al., 2018).

It is critical to consider intersectionality in these studies, as every individual is a composite of multiple identities across multiple axes. When conducting inquiries into the biased nature of these systems, we encourage researchers to use metrics that treat fairness as an intersectional concept and keep in line with the recommendations as suggested in Talat et al. (2022); Blodgett et al. (2020) to document the affected demographics. Testing the validity and reliability of bias measurement and debiasing metrics is essential to ensuring the effectiveness of proposed methods (Blodgett et al., 2020), and it is crucial to report any limitations of the same.

6 Moving Towards Inclusive Systems in All Languages

The issue of fairness in multilingualism presents a number of challenges. Although current practitioners encourage making systems multicultural and developing systems to be used only for specific cultural contexts (Talat et al., 2022), we posit that this may not be a viable solution due to various practical considerations. The vast diversity of cultures and ethnicities across the world presents significant difficulties in terms of creating equitable multilingual systems. Even within languages such as English, several dialectal variants, both of the regional and social kind (Nguyen et al., 2016), still need to be accounted for. Blodgett and O'Connor (2017) is an example of how this could further stigmatize oppressed communities. Language and various social aspects related to language are ever-evolving. Modeling aspects such as lexical variants and the syntactical difference between languages, elements like phonology, and speech inflections in spoken language could contribute to the complexity of these systems.

Several countries have diverse concentrations of people from all regions of the world with unique backgrounds. The intricacies of the social interactions resulting from the population's diverse linguistic backgrounds and issues arising from language contact make the study of the fairness of multilingual systems that would be deployed to cater to these populations essential. It is not possible to make models agnostic to demographic attributes. Even with the omission of certain attributes, models can still exhibit bias based on factors such as linguistic variations in dialect, or the linguistic features employed, as demonstrated by Hovy and Søgaard (2015) who highlight the improved performance of NLP systems on texts written by older individuals. The data that large language models (LLMs) are trained on tends to be biased towards certain demographic strata (Olteanu et al., 2019). Although curating more diverse datasets and following recommendations to mitigate bias in the data pipeline would be a step forward to mitigating this problem (B et al., 2021), various resource constraints could hinder this or make it impractical.

Due to all these challenges and the ubiquity of language technologies that are used by large populations of non-English speaking users, addressing fairness and bias, taking into account diverse linguistic, socio-linguistic, and cultural factors, is of utmost importance. Interdisciplinary and multicultural teams are crucial to identifying, measuring, and mitigating harms caused by bias in multilingual models. Better evaluation benchmarks covering diverse linguistic phenomena and cultures will lead to better fairness evaluation.

Regarding data collection, as discussed in Section 3.1, it would be prudent to avoid directly translating datasets for training or evaluation in applications where fairness is critical. As we have shown in this survey, it is not enough to collect datasets in multiple languages for measuring and mitigating bias, although even these are lacking for most languages worldwide. Zero-shot techniques that ignore the cultural nuances of a language should be used with care in fairness-critical applications, as linguistically similar languages may have different cultural values and vice versa. Finally, multilingual models and systems need to incorporate shared value systems that take into account diverse cultures, although some cultural differences may still go unacknowledged.

Limitations

Our work surveys fairness literature in languages other than English, including bias measurement and mitigation strategies. Although we call out the fact that bias in literature is studied from an Anglo-centric point of view, it is conceivable that we miss many diverse perspectives on linguistic and cultural aspects of bias in different languages and cultures of the world due to the relatively heterogeneous background (in terms of nationality, ethnicity and field of study) of the authors. There may also be other relevant work in the social science literature that we may have missed including in this survey.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa*, 4.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle, United States. Association for Computational Linguistics.

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2018. Sentiment bias in predictive text recommendations results in biased writing. *Proceedings of the 44th Graphics Interface Conference*.

Senthil Kumar B, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An overview of fairness in data – illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv. Association for Computational Linguistics.

William L. Ballard. 1978. More on yuchi pronouns. *International Journal of American Linguistics*, 44(2):103–112.

Connor Baumler and Rachel Rudinger. 2022. Recognition of they/them as singular personal pronouns in coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3426–3432, Seattle, United States. Association for Computational Linguistics.

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.

Advait Bhat, Saaket Agashe, and Anirudha Joshi. 2021. How do people interact with biased text prediction models while writing? In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 116–121, Online. Association for Computational Linguistics.

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)*, pages 727–740, Online only. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english.

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

David Bracewell and Marc Tomlinson. 2012. The language of power and its cultural influence. In *Proceedings of COLING 2012: Posters*, pages 155–164, Mumbai, India. The COLING 2012 Organizing Committee.

Daniel Buschek, Martin Zurn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

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

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12710–12718.

Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, and Nikolaos Mavridis. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine*, 3(1):81.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2019. Analytical methods for interpretable ultradense word embeddings. 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 1185–1191, Hong Kong, China. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages –, Online and Abu Dhabi, UAE. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Cliff Goddard. 2005. The languages of east and southeast asia: An introduction.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

E.D. Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 123–129, New York, NY, USA. Association for Computing Machinery.

Luise Hercus. 1994. A grammar of the arabana-wangkangurru language : Lake eyre basin, south australia.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging fairness and environmental sustainability in natural language processing.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Xiaolei Huang. 2022. Easy adaptation to mitigate gender bias in multilingual text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 717–723, Seattle, United States. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 1440–1448, Marseille, France. European Language Resources Association.

Surinder S. Jodhka, Ghanshyam Shah, Tudor Kalinga, P Silva, Paramsothy Sivapragasam, Thanges, Sri Lanka, Uddin Iftekhar, Chowdhury, Bangladesh Zulfi, Ali Shah, Krishna Bhattachan, Tej Sunar, Yasso Bhattachan, Nepal Senapati, Sobin George, Martin Macwan, S Thorat, Vincent Manoharan, and Gowhar Yakoob. 2010. Comparative contexts of discrimination: Caste and untouchability in south asia. *Economic and Political Weekly*, 45.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of Indian annotators and American text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99, Berlin, Germany. Association for Computational Linguistics.

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*, pages 6282–6293, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Unmasking the mask – evaluating social biases in masked language models.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Nikzad Khani, Isidora Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. Cultural and geographical influences on image translatability of words across languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 198–209, Online. Association for Computational Linguistics.

Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frédéric A. Dreyer, Aleksandar Shtedritski, and Yuki Markus Asano. 2021. How true is gpt-2? an empirical analysis of intersectional occupational biases. *CoRR*, abs/2102.04130.

Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Business Research*, 13(3):795–848.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. Language generation models can cause harm: So what can we do about it? an actionable survey.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022a. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022b. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022c. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Tobias Lüken, and Goran Glavas. 2021. Sustainable modular debiasing of language models. *CoRR*, abs/2109.03646.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Eva Lindström. 2008. *Language complexity and interlinguistic difficulty*, page 217–242.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alan Lundgard. 2020. Measuring justice in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Victor Benno Meyer-Rochow. 2009. Food taboos: their origins and purposes. *Journal of Ethnobiology and Ethnomedicine*, 5(1):18.

Juliana Miehle, Koichiro Yoshino, Louisa Pragst, Stefan Ultes, Satoshi Nakamura, and Wolfgang Minker. 2016. Cultural communication idiosyncrasies in human-computer interaction. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 74–79, Los Angeles. Association for Computational Linguistics.

Aristides Milios and Parishad BehnamGhader. 2022. An analysis of social biases present in bert variants across multiple languages. *ArXiv*, abs/2211.14402.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. 2022. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Ravi Sankar S Nair. 2013. Tribal languages of kerala.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Alexandra Olteanu, Carlos Castillo, Fernando D. Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Iyabo Osiapem. 2007. Florian coulmas, sociolinguistics: The study of speakers' choices -. *Language in Society - LANG SOC*, 36.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don't blame the annotator: Bias already starts in the annotation instructions.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jasmine Rao. 2010. The caste system: Effects on poverty in india, nepal and sri lanka. *Glob. Majority E-J.*, 1.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. *CoRR*, abs/2101.09995.

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, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer

learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Małgorzata Suszczyńska. 1999. Apologizing in english, polish and hungarian: Different languages, different strategies. *Journal of Pragmatics*, 31(8):1053–1065.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *CoRR*, abs/1911.01485.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Jialu Wang, Yang Liu, and Xin Wang. 2022. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *CoRR*, abs/2010.06032.

Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling topic models: A cross-cultural analysis of personal values through words. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 143–152, Austin, Texas. Association for Computational Linguistics.

Robert Wolfe and Aylin Caliskan. 2022. American == white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 800–812, New York, NY, USA. Association for Computing Machinery.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2022. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. 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 5276–5284, Hong Kong, China. Association for Computational Linguistics.