On Evaluating and Mitigating Gender Biases in Multilingual Settings

Aniket Vashishtha[*] **Kabir Ahuja**[*] **Sunayana Sitaram**
Microsoft Research India
{t-aniketva,t-kabirahuja,sunayana.sitaram}@microsoft.com

Abstract

While understanding and removing gender biases in language models has been a longstanding problem in Natural Language Processing, prior research work has primarily been limited to English. In this work, we investigate some of the challenges with evaluating and mitigating biases in multilingual settings which stem from a lack of existing benchmarks and resources for bias evaluation beyond English especially for non-western context. In this paper, we first create a benchmark for evaluating gender biases in pre-trained masked language models by extending DisCo to different Indian languages using human annotations. We extend various debiasing methods to work beyond English and evaluate their effectiveness for SOTA massively multilingual models on our proposed metric. Overall, our work highlights the challenges that arise while studying social biases in multilingual settings and provides resources as well as mitigation techniques to take a step toward scaling to more languages.

1 Introduction

Large Language Models (LLMs) (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020) have obtained impressive performance on a wide range of NLP tasks showing great potential in several downstream applications for real world impact. However, these models have shown to be prone to picking up unwanted correlations and stereotypes from the pre-training data (Sheng et al., 2019; Kurita et al., 2019; Hutchinson et al., 2020) which, can perpetuate harmful biases for people belonging to marginalized groups. While there has been a great deal of interest in understanding and mitigating such biases in LLMs (Nadeem et al., 2021; Schick et al., 2021; Meade et al., 2022), the focus of such studies has primarily been on English.

While Massively Multilingual Language Models (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021), have shown impressive performances across a wide range of languages, especially with their surprising effectiveness at zero-shot cross-lingual transfer, there still exists a lack of focused research to evaluate and mitigate the biases that exist in these models. This can lead to a lack of inclusive and responsible technologies for groups whose native language is not English and can also lead to the dissemination of stereotypes and the widening of existing cultural gaps.

Past work on evaluating and mitigating biases in multilingual models has mostly been concerned with gender bias in cross-lingual word embeddings (Zhao et al., 2020; Bansal et al., 2021) which fails to account for contextual information (Kurita et al., 2019; Delobelle et al., 2022), making them unreliable for LLMs. Other methods for estimating biases in contextualized representations involve Multilingual Bias Evaluation (Kaneko et al., 2022, MBE), which utilizes parallel translation corpora in different languages that might lack non-western cultural contexts (Talat et al., 2022). For debiasing LLMs, Lauscher et al. (2021) proposed an adapter (Houlsby et al., 2019) based approach. However, the biases are measured in the word representations and only English data was used for debiasing, missing out on cultural context for other languages.

To address these concerns, we make the following key contributions in our work. *First*, we extend the DisCo metric (Webster et al., 2020) by creating human-corrected templates for 6 Indian languages. DisCo takes sentence-level context while measuring bias and our templates are largely culturally agnostic making them more generally applicable. *Second*, we extend existing debiasing strategies like Counterfactual Data Augmentation (Zhao et al., 2018) and Self-Debiasing (Schick et al., 2021) to mitigate gender biases across languages in Masked Language Models (MLMs).

Finally, we also evaluate the transferability of debiasing MLMs from one source language to other

[*] Equal contribution

Findings of the Association for Computational Linguistics: ACL 2023, pages 307–318
July 9-14, 2023 ©2023 Association for Computational Linguistics

target languages and observe limited transfer from English to languages lacking western context. However, we do observe that typologically and culturally similar languages aid each other in reducing gender bias. While there have been multiple studies on measuring biases in multilingual models, previous work has not explored mitigating gender biases from these models on multiple languages and studying the transferability of debiasing across different languages. This is especially true while using non-embedding based approaches for evaluation and debiasing. To the best of our knowledge, ours is the first work to debias multilingual LLMs for different languages and measure the cross-lingual transfer for gender bias mitigation. To encourage future research in this area, we will release our code and datasets publically[1].

2 Measuring Bias in Multilingual Models

In this section, we describe the benchmarks to evaluate biases in MLMs across different languages. Since most existing benchmarks for bias evaluation in contextualized representations are designed for English, we discuss our multilingual variant of DisCo and the recently proposed MBE metric.

2.1 Multilingual DisCo

Discovery of Correlations (DisCo) is a template-based metric that measures unfair or biased associations of predictions of an MLM to a particular gender. It follows a slot-filling procedure where for each template, predictions are made for a masked token, which are evaluated to assess whether there is a statistically significant difference in the top predictions across male and female genders. For calculating the bias score using DisCo, a χ^2 test is performed to reject the null hypothesis (with a p-value of 0.05) that the model has the same prediction rate with both male and female context. We use the modified version of the metric from (Delobelle et al., 2022) that measures the fraction of slot-fills containing predictions with gendered associations (fully biased model gets a score of 1, and fully unbiased gets a score of 0).

We extend the **Names** variant of DisCo, as personal names can act as representatives for various socio-demographic attributes to capture cultural context (Sambasivan et al., 2021). Especially for India, surnames are a strong cultural identifier. Majority Indian surnames are typically an identifier of belonging to a particular caste, religion and culture. We use surnames from specific cultures which speak the languages for which we prepare the name pairs for. We further use these surnames to filter out personal first names for both male and female from an open-source Indian names list containing a large number of popular Indian names (details in Appendix A.1) and word-translated the names from English to the corresponding languages, to be used for slot-filling. Further, unlike nouns and pronouns which might be gender-neutral in some languages, names are indicative of gender to a large extent across cultures.

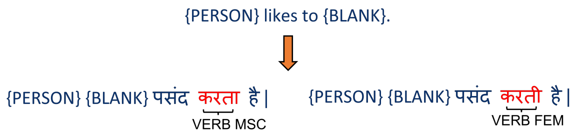

{PERSON} likes to {BLANK}.
{PERSON} {BLANK} पसंद करता है |
VERB MSC
{PERSON} {BLANK} पसंद करती है |
VERB FEM

Figure 1: Example template translation for "*{PERSON} likes to {BLANK}*" in Hindi for creation of our multilingual dataset.

Dataset Construction: We start with the 14 templates provided in Webster et al. (2020) and translate them using Bing translation API [2] to 6 Indian languages of varying resources. We use the Class taxonomy from (Joshi et al., 2020) to characterize language resources, where Class 5 represent high resource and Class-0 for lowest resource languages. Our set of Indian Languages contain Class 4 language Hindi (*hi*); Class 3 language Bengali (*bn*); Class 2 languages Marathi (*mr*) and Punjabi (*pa*); and Class 1 language Gujarati (*gu*). A challenge while transferring templates from English to these languages is that, unlike English, a common template might not be applicable to both genders. For eg. the template "'*{PERSON} likes to {BLANK}*"', will have different translations in Hindi, depending upon the gender of the slot fill for {PERSON}, as Hindi has gendered verbs. Hence, during translation we first filled the *{PERSON}* slot with a male and a female name to obtain two templates corresponding to each gender (see Figure 1). All the translated templates in our dataset were then thoroughly reviewed and corrected by human annotators who are native speakers of the languages (details in Appendix A.1).

[1] https://aka.ms/multilingual-bias

[2] https://www.microsoft.com/en-us/translator/

2.2 Multilingual Bias Evaluation (MBE)

We also evaluate MLMs with the MBE score proposed in (Kaneko et al., 2022) containing datasets for bias evaluation in 8 high resource languages: German (de), Japanese (ja), Arabic (ar), Spanish (es), and Mandarin (zh) belonging to Class 5; Portuguese (pt) and Russian (ru) in Class 4; and Indonesian (id) in Class 3. For evaluation, it first considers parallel corpora from English to different languages and extracts the set of sentences containing male and female words. Next, the likelihood for each sentence is evaluated with the MLM, and the bias score is measured as the percentage of total pairs for which a male sentence gets a higher likelihood than a female sentence. Hence a value close to 50 for an MLM indicates no bias towards both groups while greater or smaller values indicate a bias towards females and males respectively. For better interpretability of metrics, we report $|50 - \mathbf{MBE}|$ in our results.

3 Mitigating Bias in Multilingual Models

We next discuss how we extend bias mitigation techniques to work beyond English along with different fine-tuning and prompting strategies that we deploy in our experiments.

3.1 Counterfactual Data Augmentation (CDA)

CDA (Zhao et al., 2018) is an effective method for reducing biases picked up by the language models during pre-training. It operates by augmenting an unlabeled text corpus with counterfactuals generated for each sentence based on a specific dimension like gender. As an example, the counterfactual for a sentence s = *"The doctor went to* ***his*** *home"* will be $\hat{s}$ = *"The doctor went to* ***her*** *home"*. The model is then fine-tuned on the augmented data, which helps balance out any spurious correlations that would have existed in the pre-training dataset.

To generate counterfactuals in English, we do word replacements on Wikipedia data using 193 gendered term pairs (eg. {he, she}, {actor, actress}, etc.) following Lauscher et al. (2021). However, generating counterfactuals for languages other than English can be challenging as acquiring term pairs need recruiting annotators which can be expensive for low-resource languages. Further, word replacement can prove unreliable for languages that mark gender case to objects (like Hindi), producing ungrammatical sentences (Zmigrod et al., 2019).

Generating Multilingual Counterfactuals: We use a translation-based approach to obtain counterfactually augmented examples in different languages. We first select the sentences in the Wikipedia English corpus containing India-related keywords which were extracted using ConceptNet (Speer et al., 2017) which include keywords related to Indian food, location, languages, religions, etc. Using these keywords we select a set of 20K sentences to avoid under-representation of Indian culture specific context. Also, generating counterfactuals for the whole corpus and fine-tuning MLMs for each of the languages will require substantial energy consumption (Strubell et al., 2019), so we decided to use the set of filtered 20k sentences for debiasing the MLMs. Further, we augment the 193 term pairs list to contain pairs of Indian personal names as well. We align the male and female names through a greedy search for selecting pairs with minimum edit distance. Finally, using the augmented term pairs list and the filtered data with Indian context, we generate counterfactuals using word replacements and translate the obtained data to the 6 Indian languages.

Once we have obtained CDA data in different languages, we can utilize it to debias the model. We define CDA-$\mathcal{S}$ as a fine-tuning setup where the MLM is debiased using CDA data for languages belonging to the set $\mathcal{S} \subset \mathcal{L}$, where $\mathcal{L} = \{\mathrm{en}, \mathrm{hi}, \mathrm{pa}, \mathrm{bn}, \mathrm{ta}, \mathrm{gu}, \mathrm{mr}\}$. In particular, we explore the following classes of fine-tuning setups:

1. CDA-$\{\mathrm{en}\}$: Fine-tune the model with English CDA data only (zero-shot debiasing).

2. CDA-$\{l\}$: Fine-tune the model with language l specific CDA data (monolingual-debiasing).

3. CDA-$\{l, \mathrm{en}\}$: Fine-tune the model with English and language l's CDA data (few-shot debiasing).

4. CDA-$\mathcal{L} \setminus \{\mathrm{en}\}$: Fine-tune the model with CDA data in all non-English languages (multilingual-debiasing).

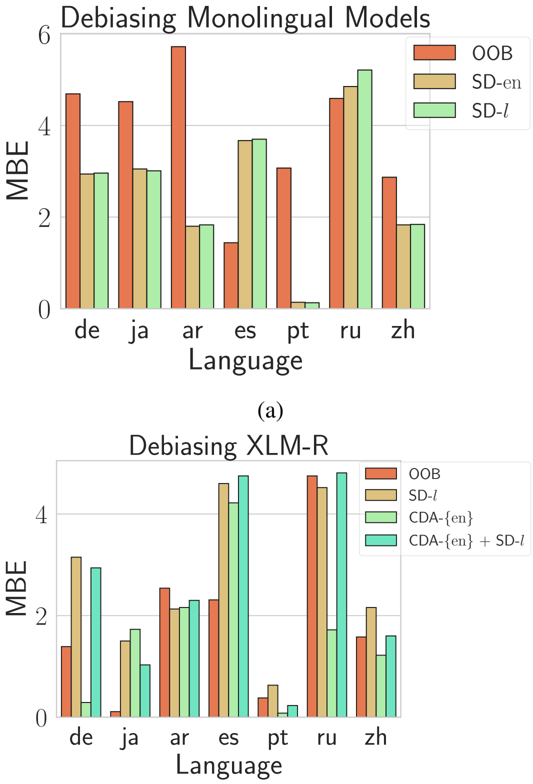
3.2 Self-Debiasing

Self-Debiasing (Schick et al., 2021) is a post-hoc method to reduce corpus-based biases in language models. It is based on the observation that pretrained language models can recognize biases in text data fairly well and prepends the input text with prompts encouraging the model to exhibit undesired behavior. Using this, it recognizes the undesirable predictions of the model as the ones with an increase in likelihood when the prompt is provided and suppresses them in the final predictions. We translate the English prompt *"The following text discriminates against people because of their gender"* in different languages and use them for bias mitigation (**SD-**l). We also experiment with using English prompt for other languages (**SD-**en).

4 Results

We evaluate the Out Of Box (OOB) biases as well the effect of applying aforementioned debiasing techniques in multilingual MLMs like XLMR-base (Conneau et al., 2020), IndicBERT (Kakwani et al., 2020), and mBERT (cased) (Devlin et al., 2019) using our multilingual DisCo metric. Additionally, we also evaluate language-specific monolingual models (refer Table 3 in appendix) and XLMR on the MBE score.

Comparison Between Different Fine-tuning Setups for CDA: We first compare the results of bias mitigation across all 4 classes of finetuning setups for CDA to understand the effect each had on the final bias reduction. As can be seen in Table 1 even though zero-shot transfer from English (CDA-$\{\text{en}\}$) results in some reduction in biases when compared to the models without any debiasing (OOB), most of the other fine-tuning setups that use language-specific counterfactuals incur better drops in the DisCo score. Specifically, few-shot debiasing (CDA-$\{l, \text{en}\}$) and multilingual-debiasing (CDA-$\mathcal{L} \setminus \{\text{en}\}$) perform consistently the best for both models with CDA-$\mathcal{L} \setminus \{\text{en}\}$ performing slightly better for XLMR and substantially so for Indic-BERT. This shows that even though language-specific counterfactuals were translated, using them for the debiasing of models helped in considerable bias reduction. We also observe that the monolingual debiasing (CDA-$\{l\}$) leads to a drop similar to CDA-$\{\text{en}\}$, and we conjecture that it might be attributed to the low amount of data we have in languages other than English for debiasing. Further, the dominant performance of CDA-$\mathcal{L} \setminus \{\text{en}\}$ highlights that languages from a similar culture can collectively help improve biases in such models. We also observe similar results for mBERT which are provided in Table 4 in the appendix.



(a)

(b)

Figure 2: MBE scores for monolingual and multilingual models and the impact of debiasing across languages

Comparison Between CDA and Self-Debiasing: Counter to CDA, Self-Debiasing shows different bias mitigation trends for Indian languages. Table 1 shows that for both multilingual MLMs, the overall bias ends up increasing when Self-Debiasing is applied, and that too by a considerable amount for IndicBERT. This seems to be in contrast to the past work (Meade et al., 2022) that shows Self-Debiasing to be the strongest debiasing technique. However, we will see next the cases where it can indeed be effective in reducing biases.

Evaluation on MBE Metric: We first investigate the effect of Self-Debiasing on monolingual models when evaluated for the MBE metric. As can be observed in Figure 2a, for most languages (except Russian and Spanish), both variants of Self-Debiasing manage to reduce the biases substantially. However, when we compare the results on a multilingual model i.e. XLMR in Figure 2b, we again observe the same phenomenon as for multilingual DisCo, where the biases tend to increase upon applying Self-Debiasing. Figure 2a shows that SD-en and SD-l have similar debiasing performance for monolingual models. It is intriguing that monolingual models are able to debias so well based on English prompts. This similarity in results with non-English and English prompts could possibly

MLM	Method	Languages	en	hi	pa	bn	ta	gu	mr	$\mathcal{L} \setminus \{\text{en}\}$
XLM-R	OOB	{}	0.78	0.83	0.92	0.94	0.94	0.86	0.86	0.89
	Self-Debiasing	{en}	0.82	0.88	0.92	0.93	0.94	0.86	0.87	0.90
		$\{l\}$	0.82	0.89	0.93	0.94	0.92	0.89	0.88	0.91
	CDA	{en}	**0.61**	0.83	0.83	0.89	0.90	0.82	0.83	0.85
		$\{l\}$	**0.61**	0.81	0.84	0.90	0.92	0.78	0.83	0.85
		$\{l, \text{en}\}$	-	**0.74**	0.79	0.88	0.87	**0.70**	**0.69**	0.78
		$\mathcal{L} \setminus \text{en}$	0.73	0.75	**0.61**	**0.87**	**0.87**	0.78	0.76	**0.77**
IndicBERT	OOB	{}	**0.70**	0.79	0.84	0.93	0.86	0.82	0.76	0.83
	Self-Debiasing	{en}	0.78	0.86	0.93	0.98	0.93	0.86	0.87	0.90
		$\{l\}$	0.78	0.86	0.89	0.96	0.91	0.84	0.87	0.89
	CDA	{en}	0.70	0.76	**0.72**	0.95	0.89	0.83	0.85	0.83
		$\{l\}$	0.70	0.80	0.80	0.82	0.90	0.79	0.78	0.82
		$\{l, \text{en}\}$	-	0.75	0.80	0.83	0.80	0.86	0.75	0.80
		$\mathcal{L} \setminus \text{en}$	0.72	**0.66**	0.75	**0.80**	**0.79**	**0.66**	**0.73**	**0.73**

Table 1: Multilingual DisCo metric results (score of 1 being fully biased and 0 being fully unbiased) of debiasing using CDA and Self-Debiasing using various fine-tuning settings on different languages. Refer to Table 4 for the full version of the results.

be explained by contamination in the pretraining monolingual data (Blevins and Zettlemoyer, 2022). We also compare the effect of CDA-{en}on reducing the biases and we observed it does obtain more success in most languages (except Spanish and Japanese). Even though MBE and Multilingual DisCo have different experimental setups, obtaining consistent results while using the two different metrics like English-only debiasing being insufficient to reduce biases in other languages. Self-debiasing being ineffective for mitigating biases in multilingual models strenghtens the applicability of our results. Our results indicate that Self-Debiasing might be limited for multilingual models and we leave the investigation of this phenomenon to future work.

5 Conclusion

In this work, we investigated gender biases in multilingual settings by proposing a bias evaluation dataset in 6 Indian languages. We further extended debiasing approaches like CDA and Self-Debiasing to work for languages beyond English and evaluated their effectiveness in removing biases across languages in MLMs. One of our key findings is that debiasing with English data might only provide a limited bias reduction in other languages and even collecting a limited amount of counterfactual data through translation can lead to substantial improvements when jointly trained with such data from similar languages. Finally, we showed that despite being effective on monolingual models, Self-Debiasing is limited in reducing biases in multilingual models with often resulting in an increase in overall bias. We hope that our work will act as a useful resource for the community to build more inclusive technologies for all cultures.

6 Limitations

The present study is limited to exploring biases in MLMs for the gender dimension only. For future work, important dimensionalities can be explored, especially for non-western contexts like Caste, Ethnicity, etc (Ahn and Oh, 2021; Bhatt et al., 2022). We also used Machine Translation on English counterfactuals to obtain CDA data in each language in our dataset. Translations are prone to errors and issues like *Translaionese* (Gellerstam, 1986), especially for the lower resource languages, and therefore can lead to the unreliability of the quality of generated counterfactuals were generated. In the future, we would like to explore learning generative (Wu et al., 2021) or editing models (Malmi et al., 2022) for automatically generating gender counterfactuals given text data in different languages. This can help us scale our counterfactual generation process to a much higher number of samples while also avoiding any losses in quality that may arise due to machine translation. Our multilingual DisCo metric is currently limited to 6 Indian languages and we hope our work will inspire further extension to cover different language families for improving the focus on multilingual biases evaluation.

7 Ethical Considerations

Our work dealt with evaluating biases in MLMs and different methods for bias mitigation in multilingual settings. While most of the current work is disproportionately in favor of high-resource languages like English, it is extremely important to improve this linguistic disparity for building inclusive and responsible language technology. Through our work, we provided a dataset to evaluate gender biases in languages of varying resources as well as methods to reduce such biases.

Acknowledgements

We would like to thank the following people who helped in evaluating and improving the Multilingual DisCo templates: Ranajoy Sadhukhan, Atharv Sonwane, Abhinav Rao, Krut Patel and Mirza Baig.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Srijan Bansal, Vishal Garimella, Ayush Suhane, and Animesh Mukherjee. 2021. Debiasing multilingual word embeddings: A case study of three indian languages.

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.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. Translation studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. 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.

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 the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

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.

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.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, and Aliaksei Severyn. 2022. Text generation with text-editing models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 1–7, Seattle, United States. 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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. 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.

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. *arXiv preprint arXiv:2010.06032*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving 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 6707–6723, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Dataset Construction Details

Scraping Langauge-Specific Personal Names: We curated a list of personal names corresponding to the cultures for each language by scraping the popular surnames associated with each culture from Wikipedia[3]. We then obtain the open source list of Indian male[4] and female[5] names, and we segment the names to different languages by referring to our culture-specific surnames list. The names obtained this way our in Latin script, so we transliterate them to the corresponding languages using the Bing Translator API.

Annotator Details: For verifying the templates obtained using machine translation we asked human annotators to correct them. Our annotators were colleagues working at our research lab and all of them were of South Asian (Indian) descent, native to different parts of India, and each having one of the six Indian languages that we consider as their L1. They all identify as males and are in their mid-20s. The annotators were provided original English templates along with the translated ones in their native language and were asked to verify that they were grammatically correct and conveyed the exact same meaning as the original base template. Further, they were asked to make corrections to ensure

[3] https://en.wikipedia.org/wiki/Category:Indian_surnames

[4] https://gist.github.com/mbejda/7f86ca901fe41bc14a63

[5] https://gist.github.com/mbejda/9b93c7545c9dd93060bd

Language	Number of Name Pairs
Hindi	164
Punjabi	50
Bengali	33
Gujarati	51
Tamil	19
Marathi	49

Table 2: Total number of gendered name pairs for each language used in Multilingual DisCo

that a template pair was as close to each other as possible except for modifications in the gendered terms, like verbs in the case of Hindi (Figure 2).

Dataset Statistics: Our dataset consists of 14 templates in each language and for each language the number of name pairs are given in Table 2.

A.2 Experimental Setup

We performed all our experiments on a single A100 GPU. For the fine-tuning setup CDA-$\{\text{en}\}$, we trained for 50K steps using a batch size of 32, a learning rate of 2e-5, and a weight decay of 0.01. We follow the same hyperparameters for other fine-tuning setups as well, but instead of fine-tuning for 50K steps, we train for 1 epoch following (Lauscher et al., 2020) as the amount of data is limited in other languages. For Self-Debiasing, we used the default hyperparameters i.e. the decay constant $\lambda = 50$ and $\epsilon = 0.01$. For all of our experiments, we used the pre-trained models provided with HuggingFace's transformers library (Wolf et al., 2020). The details of all the pre-trained models that we use in the paper are provided in Table 3

Model Name	**Variant**	**Supported Languages**	**Number of Parameters**
Multilingual Masked Language Models			
XLM-R	xlm-roberta-base	100 languages from (Conneau et al., 2020)	270M
IndicBERT	indic-bert	12 Indian Languages	12M
mBERT	bert-base-multilingual-cased	Top 104 Wikipedia Languages [6]	110M
Monolingual Masked Language Models			
GBERT (Chan et al., 2020)	gbert-base	de	110M
BERT Japanese[7]	bert-base-japanese-whole-word-masking	ja	110M
AraBERT (Antoun et al., 2020)	bert-base-arabertv02	ar	110M
Spanish Pre-trained BERT (Cañete et al., 2020)	bert-base-spanish-wwm-uncased	es	110M
BERTimbau(Souza et al., 2020)	bert-base-portuguese-cased	pt	110M
RoBERTa-base for Russian [8]	roberta-base-russian-v0	ru	110M
Chinese BERT (Cui et al., 2020)	chinese-bert-wwm-ext	zh	100M

Table 3: Description of MLMs that we use in our experiments

MMLM	Debiasing Method	Languages Used	en	hi	pa	bn	ta	gu	mr	$\mathcal{L} \setminus \{\text{en}\}$
XLM-R	OOB	$\{\}$	0.78	0.83	0.92	0.94	0.94	0.86	0.86	0.89
	Self-Debiasing	$\{\text{en}\}$	0.82	0.88	0.92	0.93	0.94	0.86	0.87	0.90
		$\{l\}$	0.82	0.89	0.93	0.94	0.92	0.89	0.88	0.91
	CDA	$\{\text{en}\}$	**0.61**	0.83	0.83	0.89	0.90	0.82	0.83	0.85
		$\{l\}$	**0.61**	0.81	0.84	0.90	0.92	0.78	0.83	0.85
		$\{\text{en}, l\}$	-	**0.74**	0.79	0.88	0.87	**0.70**	**0.69**	0.78
		$\mathcal{L} \setminus \text{en}$	0.73	0.75	**0.61**	**0.87**	0.87	0.78	0.76	**0.77**
		$\mathcal{L}$	0.72	0.78	0.74	0.89	**0.85**	0.75	0.79	0.80
mBERT	OOB	$\{\}$	0.88	0.87	0.72	0.93	**0.79**	0.84	**0.71**	0.81
	Self-Debiasing	$\{\text{en}\}$	0.88	0.90	0.87	0.98	0.94	0.91	0.89	0.91
		$\{l\}$	0.88	0.86	0.81	0.98	0.92	0.91	0.82	0.88
	CDA	$\{\text{en}\}$	**0.68**	0.90	0.73	0.94	0.85	0.79	0.75	0.83
		$\{l\}$	**0.68**	**0.76**	0.72	0.89	0.86	0.77	0.79	0.80
		$\{\text{en}, l\}$	-	0.84	**0.67**	0.86	0.80	**0.73**	0.76	**0.78**
		$\mathcal{L} \setminus \text{en}$	0.88	0.82	0.73	**0.80**	**0.79**	0.79	0.88	0.80
		$\mathcal{L}$	0.88	0.83	0.79	0.81	0.82	0.75	0.92	0.82
IndicBERT	OOB	$\{\}$	0.70	0.79	0.84	0.93	0.86	0.82	0.76	0.83
	Self-Debiasing	$\{\text{en}\}$	0.78	0.86	0.93	0.98	0.93	0.86	0.87	0.90
		$\{l\}$	0.78	0.86	0.89	0.96	0.91	0.84	0.87	0.89
	CDA	$\{\text{en}\}$	0.70	0.76	**0.72**	0.95	0.89	0.83	0.85	0.83
		$\{l\}$	0.70	0.80	0.80	0.82	0.90	0.79	0.78	0.82
		$\{\text{en}, l\}$	-	0.75	0.80	0.83	0.80	0.86	0.75	0.80
		$\mathcal{L} \setminus \text{en}$	0.72	**0.66**	0.75	**0.80**	**0.79**	**0.66**	**0.73**	**0.73**
		$\mathcal{L}$	**0.62**	0.73	0.82	0.85	0.85	0.79	0.76	0.80

Table 4: Complete version of results of debiasing using CDA and Self-Debiasing using various fine-tuning settings on different languages and MMLMs.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will release the datasets, code and pretrained models we created as open source

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix section A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix section A.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix section A.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2 and Appendix section A1.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix section A.1

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix section A.1

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The data was collected from in-house researchers and interns

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix section A.1