Bias Beyond English: Evaluating Social Bias and Debiasing Methods in a Low-Resource Setting

Anonymous authors
Paper under double-blind review

Abstract

This paper addresses the identification of social harms—including stereotypes and bias—which may be upsetting to some readers.

Social bias in language models can potentially exacerbate social inequalities. Despite it having garnered wide attention, most research focuses on English data. In a low-resource scenario, the models often perform worse due to insufficient training data. This study aims to leverage high-resource language corpora to evaluate bias and experiment with debiasing methods in low-resource languages. We evaluated the performance of recent multilingual models in five languages: English (ENG), Chinese (ZHO), Russian (RUS), Indonesian (IND) and Thai (THA), and analyzed four bias dimensions: *gender*, *religion*, *nationality*, and *race-color*. By constructing multilingual bias evaluation datasets, this study allows fair comparisons between models across languages. We have further investigated three debiasing methods-`CDA`, `Dropout`, `SenDeb`-and demonstrated that debiasing methods from high-resource languages can be effectively transferred to low-resource ones, providing actionable insights for fairness research in multilingual NLP.

1 Introduction

Machine learning frameworks are fundamentally designed as a function that generalizes past data (Chun, 2021). As a result, pretrained language models inevitably learn the inherent social biases embedded in raw real-world text. Studies have shown that these models and their learned representations **retain** and **propagate** biases from their training data. For example, Bolukbasi et al. (2016) found that word embeddings such as `Word2Vec` retain measurable gender bias, with *male*-associated terms being linked to professions like *programmer* and *scientist*, while *female*-associated terms are more commonly linked to *nurse* and *homemaker*. Meanwhile, bias is propagated in downstream applications, thus potentially reinforcing stereotypes and exacerbating information inequality (Bender & Friedman, 2018).

Social bias has been extensively studied in natural language processing (Vanmassenhove et al., 2019; Kiritchenko & Mohammad, 2018; Sap et al., 2019; Nangia et al., 2020; Nadeem et al., 2020). However, most of the relevant works are limited to English and reflect an Anglo-centric social context. The "*subaltern*", a key figure in postcolonial discourse, is historically marginalized and voiceless—"*has no history and thus no voice*" Morris (2009). In computational linguistics, this **voiceless**ness of the subaltern is also pronounced: Most NLP research overlooks thousands of languages spoken by billions of people (Bender & Friedman, 2018; Eberhard et al., 2019). Although multilingual language models are trained on language data rather than cultural data, all languages inherently reflect cultural stereotypes. However, models are primarily trained, evaluated, and aligned using Western data and culture, and therefore debiasing techniques often fail to account for culture-specific discrimination (Khandelwal et al., 2023). The systemic bias present in these models not only affects the fairness and accuracy of multilingual systems, but also negatively impacts social equity and equal access to information. Developing more equitable and inclusive multilingual systems is an urgent challenge in NLP today.

This study aims to systematically evaluate model bias in a multilingual setting. We adopt the masked language model prediction probability method proposed by Nadeem et al. (2020) to measure the bias of language models toward specific social attributes. To compare biases across different models and languages, we propose a new evaluation metric that standardizes model evaluation indicators, enabling more precise cross-model bias comparison.

Based on the CrowS-Pairs dataset Nangia et al. (2020), we created a new multilingual bias evaluation dataset with English, Chinese, Russian, Thai, and Indonesian; each with four representative bias types: gender, race-color, nationality and religion. These languages were selected based on their global use and the distribution of online resources, with Thai and Indonesian considered low-resource languages. In our experiments, we evaluated a series of widely used multilingual language models. We noticed significant differences in the type of model bias across languages. Our findings emphasize the importance of incorporating diverse cultural and linguistic backgrounds into research to ensure fairness on a global scale.

Little research has been done on exploring application in debiasing methods across languages. In this work, we further fine-tuned multilingual models on English Wikipedia datasets with `CDA`, `DO`, computed the bias subspace with `SenDeb`, and measured bias shifts across English and four other languages.

We summarize our contribution as follows.

- we proposed $\mathbb{NBS}$ (§3.1 & §A) as a method to evaluate model bias in a multilingual setting by measuring the normalized probability of a masked-word prediction in a biased context.
- we curated a new dataset (§3.2) for bias evaluation with five languages, representatively selected based on the language resource conditions, with Thai and Indonesian considered low-resource languages.
- our results show that the impact of different bias types (*e.g.*, gender, religion) varies across languages and cultures in the six multilingual models we tested (§3.3.2).
- we demonstrated that debiasing strategies can be effectively transferred to other languages through cross-lingual knowledge sharing (§4.3).

2 Related Works

2.1 Low-Resource Languages

Low-resource language communities face barriers accessing information. Hedderich et al. (2021) discussed how low-resource languages can benefit from annotated resources in high-resource languages. Hu et al. (2020) and Wu & Dredze (2020) noted that there remains a significant performance gap between high-resource and low-resource settings. Despite advancements in multilingual NLP, existing models do not yet serve as truly universal language models, and many languages with over a million speakers remain underrepresented (Lauscher et al., 2020).

2.2 Social Bias in the Multilingual Setting

A significant body of research has emerged on bias in NLP systems. The attempts to measure bias in word embeddings date back Bolukbasi et al. (2016), followed by research exposing bias in contextualized language representations trained for various NLP tasks. For example, Vanmassenhove et al. (2019) identified bias in machine translation, Kiritchenko & Mohammad (2018) found gender and racial bias in sentiment analysis , and Sap et al. (2019) discovered racial bias in hate speech and toxicity detection. However, these are all limited to the English language.

Some research explored social biases in non-English settings. Lauscher et al. (2020) conducted an extensive analysis of bias in Arabic word embeddings, identifying gender and

Language	Language Family	Writing System	Availability	2023-50	2024-10	2024-18
English	Indo-European (Germanic)	Latin	High	44.43%	46.45%	45.51%
Chinese	Sino-Tibetan	Hanzi	Medium	5.08%	4.17%	4.42%
Russian	Indo-European (Slavic)	Cyrillic	Medium	6.03%	5.81%	5.95%
Thai	Kra-Dai	Thai	Low	0.43%	0.41%	0.41%
Indonesian	Austronesian	Latin	Low	0.86%	0.86%	0.92%

Table 1: Linguistic characteristics of selected languages from the Common Crawl archives Crawl (2024). Each entry corresponds to the dataset prefix `CC-MAIN`.

racial biases in Arabic news corpora. Sahoo et al. (2023) created a Hindi social bias detection dataset. Névéol et al. (2022) extended CrowS-Pairs to investigate various biases in French. Zhou et al. (2019) evaluated gender bias in grammatically gendered languages, with experiments on French and Spanish text. B et al. (2022) examined gender and caste bias in monolingual word embeddings for Hindi and Tamil. Their research demonstrated that bias evaluation becomes significantly more complex in a multilingual context due to (1) varying cultural frameworks that influence the definition of bias and (2) differences in grammatical structures, which render some existing evaluation methods highly challenging to apply.

Recently, several studies have begun to examine multilingual bias at a more holistic level. Ahn & Oh (2021) analyzed ethnicity bias in six languages and attempted to mitigate biases seen in monolingual models by using mBERT. Levy et al. (2023) analyzed biases in sentiment analysis in five languages in mBERT and XLM-RoBERTa. Câmara et al. (2022) analyzed gender, race, and ethnicity bias in English, Spanish, and Arabic for the sentiment analysis task. Cabello Piqueras & Søgaard (2022) created parallel cloze test sets in English, Spanish, German and French with mBERT, XLM-R and m-T5. However, all the aforementioned studies have exclusively utilized comparatively small-scale pre-trained language models and have not examined bias behavior in large language models, nor have they focused on low-resource languages.

2.3 Debiasing in Multilingual Systems

Existing debias methods have made progress in reducing biases in models, but still face numerous challenges. Meade et al. (2022) summarized several recent debias methods for pretrained language models including Counterfactual Data Augmentation (**CDA**) (Zmigrod et al., 2019; Webster et al., 2020), **Self-Debias** (Schick et al., 2021), Dropout Regularization (**DO**), Sentence-Level Debiasing (**SenDeb**) (Liang et al., 2020) and Iterative Nullspace Projection (**INLP**) (Ravfogel et al., 2020).

Most debias methods are designed and evaluated for English-only environments, and research on multilingual transfer learning for debiasing is still limited. Since multilingual models share linguistic knowledge across languages, they have the potential to transfer debiased knowledge from English to other languages (Wang et al., 2019). Reusens et al. (2023b) explored the cross-lingual transferability of debiasing techniques in multilingual models using mBERT, demonstrating that debiasing methods effectively reduce bias with SentenceDebias achieving the best results. Nozza (2021) explored cross-lingual debiasing in English, Italian, and Spanish for stereotype detection tasks. However, no current research focuses on debiasing specifically for low-resource languages, highlighting an urgent need for multilingual debiasing transfer learning.

3 Multilingual Bias Evaluation

3.1 Methodology

Adopting the methodology proposed by Nangia et al. (2020), we also assess bias in masked language models (MLM) by predicting the probability of masked words with pseudo-log-likelihood estimation (Wang & Cho, 2019). The probability score $\mathbb{PS}(s_i)$ of sentence s is defined as:

$$\mathbb{PS}(s_i) = \sum_{j=0}^{|U|} \log P(u_j \in U | U_{\setminus u_j}; m_i; \theta) \tag{1}$$

where $M = \{m_0, \ldots, m_n\}$ represent the stereotypical words we modify in this case, $U = \{u_0, \ldots, u_l\}$ represent the unchanged words, $s_i = U \cup m_i$, and θ is the language model parameter. Most recent multilingual models are causal language models (CLM) which are not finetuned to predict masked tokens, therefore, instead of masking a token, we remove it from the input and use the model to generate a probability distribution for that position. We take the prediction scores of the language modeling head (scores for each vocabulary token before SoftMax) as an approximation.

While Nangia et al. (2020)'s method allows for intra-model comparisons, it does not facilitate bias comparisons across different models. Therefore, we define Normalized Bias Score $\mathbb{NBS}$ (3) for comparison across models and provide a benchmark framework:

$$W_{\text{avg}} = \frac{1}{n} \sum_{l \in \text{lang}} \frac{1}{N} \sum_{k=1}^{N} \frac{\left|\mathbb{PS}(s_{l,k}) + \mathbb{PS}(\bar{s}_{l,k})\right|}{2} \tag{2}$$

$$\mathbb{NBS}(\theta) = \frac{1}{W_{avg}} \cdot \frac{1}{N} \sum_{k=1}^{N} \left|\mathbb{PS}(s_{l,k}) - \mathbb{PS}(\bar{s}_{l,k})\right| = 2n \cdot \frac{\sum_{k=1}^{N} \left|\mathbb{PS}(s_{l,k}) - \mathbb{PS}(\bar{s}_{l,k})\right|}{\sum_{l \in lang} \sum_{k=1}^{N} \left|\mathbb{PS}(s_{l,k}) + \mathbb{PS}(\bar{s}_{l,k})\right|} \tag{3}$$

where $n = |\text{lang}|$, s the original sentence and $\bar{s}$ the modified. In our analysis, bias evaluations are conducted using $\mathbb{NBS}$ metric. The closer $\mathbb{NBS}$ is to 0, the lower the bias in the model. If $\mathbb{NBS}$ = 0, it indicates that the model treats the two terms equally and exhibits no intrinsic social bias. For a complete mathematical illustration of this section, please refer Appendix A.

3.2 Dataset Construction

To evaluate bias in low-resource languages, we build our dataset based on CrowS-Pairs (Nangia et al., 2020).

The dataset contains 1,508 examples, covering various bias types and measures model bias by comparing the likelihood of stereotypical *vs.* non-stereotypical sentences. Previous research shows that many widely used language models favor stereotypical sentences in English, revealing internal biases. However, CrowS-Pairs is based on American sociocultural contexts, which limits its applicability to other languages and cultures. To make bias evaluation more representative, this study selects four primary bias types for in-depth analysis: *Gender, Race-Color, Nationality, Religion*, which are relatively generalizable across different cultural and linguistic contexts. After filtering the dataset, we obtained 1,042 sentences, adapted for evaluating bias in non-English languages.

We translated CrowS-Pairs into four languages: Chinese (ZHO), Russian (RUS), Indonesian (IND), and Thai (THA) using the Google Translate API. The languages were chosen based on Common Crawl Crawl (2024), a large-scale web dataset that reflects the availability of online resources in different languages. Table 1 shows the proportion of web content in these languages. According to linguistic resource classifications in NLP (Joshi et al., 2020), Chinese and Russian are considered high-resource, while Thai and Indonesian are low-resource languages, despite having millions of native speakers. The translation tool was chosen considerting both the effectiveness and budget-resource constraints. Despite the resulting corpus being artificial and translation-based, we will see in the later section that the experimental results still reveal meaningful patterns, demonstrating the value of the evaluation albeit limitations.

Language	Bias Type	Bias Score per Model					
		mBERT	**XLM-R**	**XGLM**	**Gemma 3**	**Qwen 2.5**	**LLaMA 3**
English	Gender	52.57	41.31	20.42	18.98	11.99	7.53
	Nationality	43.37	36.87	16.24	14.23	9.90	7.45
	Race-color	44.51	33.54	16.24	17.86	11.67	7.50
	Religion	48.49	44.07	18.97	16.80	11.16	9.28
	Average	**46.76**	**37.06**	**17.57**	**17.48**	**11.43**	**7.68**
Chinese	Gender	58.19	43.24	25.00	26.67	25.76	11.46
	Nationality	59.35	47.46	20.49	28.96	29.24	10.94
	Race-color	56.79	55.71	21.01	25.84	25.25	11.04
	Religion	60.49	50.54	18.81	25.71	29.35	10.16
	Average	**57.91**	**50.79**	**21.71**	**26.51**	**26.40**	**11.04**
Russian	Gender	53.27	37.11	26.42	40.91	14.77	11.24
	Nationality	49.59	37.13	19.44	25.86	10.89	9.45
	Race-color	48.79	39.42	25.48	29.49	13.07	10.51
	Religion	46.65	37.61	31.16	25.09	12.53	10.76
	Average	**49.82**	**38.30**	**25.37**	**31.36**	**13.11**	**10.56**
Indonesian	Gender	59.42	38.72	17.58	14.91	24.95	8.62
	Nationality	62.83	50.43	12.94	12.73	21.23	7.61
	Race-color	57.07	55.38	19.64	14.90	23.69	10.02
	Religion	60.84	55.08	24.93	20.68	33.23	8.37
	Average	**58.92**	**50.40**	**18.63**	**15.16**	**24.59**	**9.13**
Thai	Gender	74.84	55.15	20.92	33.34	13.01	11.51
	Nationality	103.26	65.37	22.04	25.94	11.67	7.52
	Race-color	87.08	64.93	21.41	30.68	12.75	10.21
	Religion	128.99	73.60	25.57	40.94	15.88	11.02
	Average	**90.69**	**63.41**	**21.80**	**31.66**	**12.96**	**10.21**

Table 2: Bias Score Comparison of Different Models Across Languages

3.3 Experiment

3.3.1 Setting

In this study we selected six pre-trained models: mBERT, XLM-RoBERTa, XGLM, Gemma 3, Qwen 2.5 and LLaMA 3. These models are widely used particularly in multilingual tasks. For the details of the models, refer Appendix B. Note that all language models in this list support all evaluated languages, except Qwen 2.5, which has reported to support the other four language but not Thai, as indicated in its technical documentation (Yang et al., 2025). We still report the bias scores calculated for Qwen 2.5 in Thai as per our evaluation methodology; however, the reader should be advised to keep the information in mind that Qwen 2.5 model lacks proper support for Thai language, as will be discussed in later sections.

All experiments were conducted on NVIDIA A100 PCIE GPUs. To ensure fair comparison, all models were tested under identical hardware and software conditions.

3.3.2 Results & Analysis

We first ran inference on 1,042 dataset samples to obtain key bias evaluation metrics: $\mathbb{PS}(s)$ and $\mathbb{PS}(\bar{s})$, which represent the model's bias toward specific social attributes. We then applied Equation 3 to compute bias scores across different scenarios. The results are detailed in Table 2, which provides an overview of the bias scores for different models and languages. Figure 1 visualizes the results of the Gemma model, illustrating the bias distribution across different languages and social attributes. To view all graphical results, please refer to Appendix D.

Through our classification of bias types, we observe notable differences across different experimental settings. Some potential insights are as follows:

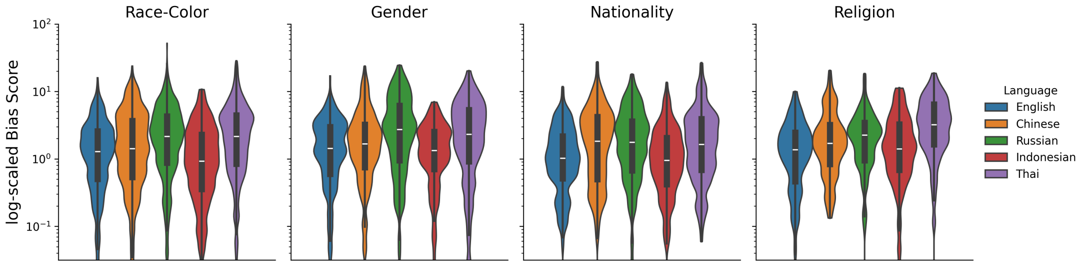

Figure 1: Bias Score across different languages with different bias categories in Gemma.

Model-wise Smaller models like XLM-RoBERTa and mBERT exhibit the highest bias scores among the tested models across multiple languages, particularly in Thai (90.69). XGLM, Gemma and Qwen 2.5 demonstrate moderate bias scores, often lower than XLM-RoBERTa and mBERT but still displaying noticeable biases across languages. LLaMA 3 consistently displays the lowest bias scores across all categories in English, suggesting it incorporates stronger alignment strategies to mitigate social biases. The lower bias scores observed in LLaMA 3 align with recent advancements in alignment-focused training methodologies. Meta's development of LLaMA models emphasizes reinforcement learning from human feedback (RLHF) and instruction tuning to ensure outputs adhere to ethical considerations and fairness principles (Grattafiori et al., 2024).

Interestingly, Qwen 2.5, despite being a highly performing model specifically in Chinese NLP benchmarks (Yang et al., 2025), exhibits a strong bias in Chinese (26.40), whereas Qwen 2.5 produced a very low bias score in Thai. However, this does not suggest that the model has relatively less bias in Thai than in English, because the model's lack of comprehension in Thai leads to unintelligible or generic outputs which resulted in a lower calculated bias score. In contrast its proficiency in Chinese allows for a finer comprehension of the contexts, which in turn exposes more detectable biases. This finding reinforces the idea that bias is independent of overall performance and can persist, if not more represented, in larger language models with stronger performances, without alignment interventions. This supports the idea that bias is not merely a function of model size or performance, but rather a reflection of data composition, pretraining strategies, and alignment interventions (Bender et al., 2021).

Language-wise English has the lowest bias scores across all models (7.68 in 3); Russian, despite being a relatively high-resource language, has considerably high bias scores in XGLM and XLM-RoBERTa, particularly in gender and race-color. Chinese exhibits higher bias scores, despite being a highly resourced language. This may suggest that its character-based writing system and different sociocultural contexts contribute to increased biases. Indonesian scores better than expected, notably in XGLM (18.63) and 3 (9.13), possibly due to its simpler grammar and usage of Latin script, Thai exhibits the highest bias scores across models. This suggests that Thai, a low-resource language with a distinct script and structure, faces significant underrepresentation in training data, leading to increased biases.

On Religious Bias Unlike the other languages, Thai and Indonesian consistently exhibit significantly higher religious bias scores across all models. This may stem from the strong cultural emphasis on Buddhism in Thailand and Islam in Indonesia, both of which play a central role in their respective societies (Pew Research Center, 2017). These cultural influences could contribute to a more pronounced religious bias in Thai and Indonesian datasets, making it more challenging to mitigate in multilingual models.

On Gender Bias Butler (1990) argues, "*If gender itself is realized through grammatical conventions [...], then at the most fundamental epistemological level, the transformation of gender must involve a challenge to these grammatical structures.*" Russian exhibits stronger gender bias compared to other languages, likely due to its extensive gender-based inflection in nouns, adjectives, and verbs, *e.g.*, (he spoke) uses the masculine form of the verb, while (she spoke) uses the feminine form. This gender marking is not optional—every time a verb is used, it must agree with the subject's gender. The presence of explicit grammatical gender markers can reinforce gender bias in models by systematically encoding gendered distinctions. In contrast, Thai and Indonesian, which lack grammatical gender distinctions, do not show a clear pattern of gender bias. In Thai, for instance, (*khao*) can refer to both "he" and "she", while in Indonesian, *dia* serves as a gender-neutral pronoun, reducing the likelihood of systematic gender bias in models. This may contribute to lower gender bias compared to Russian, as gender distinctions in Thai are primarily does not affect verbs, adjectives and nouns. These findings suggest that linguistic structure and cultural norms shape the way bias is encoded in multilingual models.

English, despite having gendered pronouns, lacks the pervasive grammatical gender markers found in Russian, making its gender bias less structurally embedded. These findings suggest that linguistic structures play a crucial role in shaping how gender bias is encoded in multilingual models, with languages like Russian demonstrating a stronger inherent bias due to their grammatical framework, whereas in others, such as Chinese, Thai, and Indonesian, the effects are less obvious.

Key Takeaways

- Multilingual models without proper bias mitigation tend to exhibit higher biases in non-English, particularly low-resource languages, emphasizing the need for more diverse and representative training data.
- Model performance in a given language does not always correlate with its bias score; in some cases, higher performance is associated with stronger biases, while models with poor language support may produce low bias scores simply due to unintelligible or generic outputs.
- Recent LLMs show superior performance on bias in both English and non-English languages, indicating that the large scale might have provided the effectiveness of cross-lingual alignment.
- The disparity in bias scores across languages suggests that bias is not solely a result of training data quantity but may also be affected linguistic complexity, script differences, and cultural contexts. This has been observed in prior studies (Blodgett et al., 2020), where language-specific biases emerge due to imbalanced representation in training data.

4 Cross-lingual Transfer Debiasing

We have first examined how biases are presented across languages, and now we proceed to investigate whether other languages can benefit from English corpus-based debiasing methods. We use XLM-RoBERTa in this case as an example because there is a significant gap between English and other languages in terms of both bias severity and representational alignment. This discrepancy makes XLM-RoBERTa an illustrative case for investigating whether debiasing techniques developed for English can transfer effectively to other languages. To quantify the effect of debiasing techniques, we calculated the degree of bias reduction for each debiasing method as the relative percentage reduction:

$$\text{Reduction} = \frac{\mathbb{NBS} - \mathbb{NBS}'}{\mathbb{NBS}} \times 100\% \tag{4}$$

where $\mathbb{NBS}$ is baseline XLM-RoBERTa bias score, and $\mathbb{NBS}'$ is Bias score after applying mitigation.

4.1 Data

Following the similar experiments done by Reusens et al. (2023a), we select our debiasing experiment data from the Wikipedia dataset Meade et al. (2022), which contains cleaned

Language	**CDA**(%)	**DO**(%)	**SenDeb**(%)
English	-12.69	-9.55	-22.42
Chinese	-5.29	-2.67	-37.96
Russian	-6.30	-4.96	-23.86
Thai	-3.86	-4.61	-34.33
Indonesian	-5.06	-5.13	-22.93

Table 3: Debias Method Results across Different Languages with English Data

multilingual full-text Wikipedia articles. For the purpose of this research, we selected 10% of the data from Wikipedia's extensive database for experimentation. Through this approach, we obtained unsupervised data from 514,084 articles as a sufficient sample size for our bias assessment.

4.2 Methods

We adpoted three key strategies: CDA (Zmigrod et al., 2019; Webster et al., 2020), DO and SenDeb (Liang et al., 2020). Beyond the three, there are also other approaches such as Self-Debias, INLP and DR. However, Self-Debias is a post-hoc text generation debiasing procedure and cannot be used as a debiasing technique for downstream natural language understanding tasks; INLP, DR and SenDeb are all projection-based debiasing techniques, therefore we only chose to experiment with SenDeb. For a complete methodological-level description, please refer to appendix C.

- Counterfactual Data Augmentation (**CDA**) generates counterfactual samples to balance the dataset. In our experiment, we apply CDA to fine-tune on English data, measure bias shifts in English and four other languages. In the CDA process, for addressing gender bias, we selected common binary replacement pairs, such as *businessman* and *businesswoman*. For racial and religious bias, we used ternary replacement sets, such as *black*, *caucasian*, and *asian*, or *judaism*, *christianity*, and *islam*.
- Dropout Regularization (**DO**) randomly *drop* (temporarily remove) certain network nodes during training, forcing the model to learn more generalizable representations. In our experiment, we fine-tune XLM-RoBERTa using DO on English datasets and evaluate across multiple languages. In the DO debiasing process, we adjusted the model's hyperparameters and set the dropout probability of the hidden layers (hidden_dropout_prob) to 0.20 and the dropout probability of the attention heads (attention_probs dropout_prob) to 0.15.
- SentenceDebias (**SenDeb**) projects bias subspace to biased vectors, and extends debiased word vectors to full sentence representations. In our experiment, we first utilize the dataset processed in CDA, as described above, and then computed the bias subspace for the dataset. For each type of bias, we separately obtained and aligned the corresponding word vector representations. We computed the mean vector for each example, and subtracted it from each word vector to ensure data centering. We applied PCA to the aligned word vector representations and extracted the first principal component as the bias direction. During model inference, we applied projection correction of the bias direction to the last hidden layer to remove bias influence. Specifically, in the model's forward propagation, we perform debiasing operations on the final hidden state of the output layer.

4.3 Results & Analysis

Table 3 shows the percentage reduction in bias for different languages using CDA, DO, and SenDeb. Across all languages, the three debiasing methods demonstrated varying degrees of bias reduction, not just in English. This indicates that these debiasing techniques have cross-lingual applicability and effectiveness.

CDA showed particularly notable results in English. By training the XLM-RoBERTa model with English data, the model also showed bias reduction effects in other languages, but the overall effect remained primarily focused on English. DO showed relatively stable performance across languages, and its effect was most significant in English, while other languages were also positively impacted, demonstrating some debiasing capability. SenDeb showed the highest bias reduction effect across all languages. Surprisingly SenDeb did not show its best performance in English, unlike CDA and DO. Instead, it performed excellently across all languages, with the highest degrees of bias reduction in Thai and Chinese. As a direct debiasing technique, the SenDeb method might share a common bias subspace across languages. Therefore, this method was able to reduce bias regardless of languages. This could potentially be explained by the property of the word vector space, where vector relationships can approximate complex semantic and lexical relationships through linear operations (Drozd et al., 2016). In this way, word vectors can not only represent semantic information of words but also achieve analogical reasoning. Although the model was not specifically trained for such tasks, this relational structure of word vectors emerged naturally. In our experiments, we hypothesize that bias-level information, as relational properties, can be largely retained across languages, and overlaps in multilingual contexts.

This strong result calls back to previous work (Chang et al., 2022) that investigates how multilingual language models maintain a shared multilingual representation space while still encoding language-sensitive information in each language, by having language-sensitive and language-neutral axes naturally emerged within the representation space. For example, vector differences can represent semantic relationships such as gender and race to certain vector directions, and a bias vector subspace in English, when applied to other languages, can be seen as an approximation of the bias vector subspace of that language's own. This means that the bias subspace calculated from English can be easily applied to other languages, thereby providing useful tool for cross-lingual bias reduction.

5 Conclusion and Future Directions

This study focused on two main aspects: Evaluating bias in a multilingual setting and demonstrating the cross-lingual transferability of debiasing methods, offering new perspectives for ensuring fairness in multilingual models.

We proposed a new evaluation metric to enable fair comparisons between different models, constructed a multilingual bias evaluation dataset, consisting of languages with different resource status, and benchmarked major multilingual language models. We have found that for LLMs transfer learning of fairness is much better for non-English language compared to previous models, and suggested that there exists a nuance of social bias when dealing with different linguistic and cultural backgrounds. Our study demonstrates that high-resource language debias methods can be effectively transferred to low-resource languages for bias mitigation. **SenDeb** emerged as the most effective technique, suggesting that bias subspaces may share cross-linguistic properties, enabling cross-language debiasing. This finding opens new possibilities for developing universal debiasing methods across diverse languages.

Based on this work, more could be explored to further advance fairness in multilingual NLP, to improve inclusivity and ethical integrity of NLP worldwide:

- Expand and diversify Bias Evaluation Datasets, covering more languages and cultural contexts to enhance the comprehensiveness of bias evaluation.
- Conduct more detailed studies on different bias types and develop more language-aware debias methods. Also explore debiasing techniques more specifically tailored to causal LMs.
- Explain why bias subspace are shared mutually in various languages on a interpretability level and design better approach to align them. Develop a universal debiasing method that works equally well on a pan-linguistic scale.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in BERT. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 533–549, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.42. URL https://aclanthology.org/2021.emnlp-main.42/.

Senthil Kumar B, Pranav Tiwari, Aman Chandra Kumar, and Aravindan Chandrabose. Casteism in India, but not racism - a study of bias in word embeddings of Indian languages. In Kolawole Adebayo, Rohan Nanda, Kanishk Verma, and Brian Davis (eds.), *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pp. 1–7, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lateraisse-1.1/.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Judith Butler. *Gender Trouble*. Routledge, London, England, 1990.

Laura Cabello Piqueras and Anders Søgaard. Are pretrained multilingual models equally fair across languages? In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3597–3605, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.318/.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar (eds.), *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 90–106, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ltedi-1.11. URL https://aclanthology.org/2022.ltedi-1.11/.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations, 2022. URL https://arxiv.org/abs/2205.10964.

Wendy Chun. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. 11 2021. ISBN 9780262367264. doi: 10.7551/mitpress/14050.001.0001.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

Common Crawl. Common crawl language statistics, 2024. URL https://commoncrawl.github.io/cc-crawl-statistics/plots/languages. Accessed: 2024-05-16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In Yuji Matsumoto and Rashmi Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3519–3530, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1332/.

David Eberhard, Gary Simons, and Chuck Fennig. *Ethnologue: Languages of the World, 22nd Edition*. 02 2019.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2545–2568, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.201. URL https://aclanthology.org/2021.naacl-main.201.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*, 2020. doi: 10.18653/v1/2020.acl-main.560.

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. Casteist but not racist? quantifying disparities in large language model bias between india and the west, 2023.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In Imed Zitouni, Muhammad Abdul-Mageed, Houda Bouamor, Fethi Bougares, Mahmoud El-Haj, Nadi Tomeh, and Wajdi Zaghouani (eds.), *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pp. 192–199, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wanlp-1.17.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. Comparing biases and the impact of multilingual training across multiple languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10260–10280, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.634. URL https://aclanthology.org/2023.emnlp-main.634/.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*, 2020.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models, 2022. URL https://arxiv.org/abs/2112.10668.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models, 2022.

Rosalind Morris (ed.). *Can the Subaltern Speak?* Columbia University Press, 2009.

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

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8521–8531, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.583. URL https://aclanthology.org/2022.acl-long.583/.

Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 907–914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.114. URL https://aclanthology.org/2021.acl-short.114.

Pew Research Center. Religious composition by country, 2010–2050. https://www.pewresearch.org/religion/feature/religious-composition-by-country-2010-2050/, 2017. Accessed: 2025-03-27.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. *arXiv preprint arXiv:2310.10310*, 2023a.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques, 2023b. URL https://arxiv.org/abs/2310.10310.

Nihar Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13316–13330, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.842. URL https://aclanthology.org/2023.findings-acl.842.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp, 2021.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. Lost in translation: Loss and decay of linguistic richness in machine translation. *arXiv preprint arXiv:1906.12068*, 2019.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. A compact and language-sensitive multilingual translation method. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1213–1223, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1117. URL https://aclanthology.org/P19-1117.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi (eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5276–5284, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1531. URL https://aclanthology.org/D19-1531/.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*, 2019.

A $\mathbb{NBS}$: A Definition

This section provides a theoretical explanation of how social bias in language models can be evaluated. Nadeem et al. (2020) first proposed that bias in language models can be assessed using Masked Language Models (MLMs) Devlin et al. (2019), where bias is measured by predicting the probability of masked words. The specific measurement method is as follows:

Given a sentence s, s contains a specific social attribute (*e.g.*, ***Mr.*** *Li is a university professor.*), we can modify the words associated with that attribute (*e.g.*, ***Mrs.*** *Li is a university professor.*). Let $M = \{m_0, \ldots, m_n\}$ represent the modified words (***Mr.***, ***Mrs.***), $U = \{u_0, \ldots, u_l\}$ represent the unchanged words (***Li***, ***is***, ***a***...), then we have the modified sentence $s_i = U \cup m_i$.

Assuming the masked language model has parameters θ, we can measure the model's bias towards specific social attributes by masking the words in M and predicting their probabilities. By comparing the probabilities for different words $m_i \in M$, we can reasonably assess the probability of s_i in the language model:

$$P(s_i) = P(m_i | U; \theta) \tag{5}$$

However, Nangia et al. (2020) pointed out that the probability $P(m_i)$ would also affect model's prediction. This frequency bias does not necessarily indicate social bias in the language model itself. To address this issue, they proposed probability score $\mathbb{PS}$ evaluating the probability of unchanged words given the modified words by applying pseudo-log-likelihood estimation (Wang & Cho, 2019). For modified sentence s_i, words in U are masked one at a time until all u_j have been masked:

$$\mathbb{PS}(s_i) = \sum_{j=0}^{|U|} \log P(u_j \in U | U_{\setminus u_j}; m_i; \theta) \tag{6}$$

This score approximates the true conditional probability, measuring how strongly a model assigns higher likelihoods to stereotypical sentences.

While Nangia et al. (2020)'s method allows for intra-model comparisons, it does not facilitate bias comparisons across different models. Language models may have different architectures and training datasets, leading to varying internal weight distributions that affect their predictions under the same evaluation conditions.

To address this, this paper proposes normalizing each model's predictions by computing the average bias prediction score across different social attributes, using the formula below

(7). This method enables fair and consistent bias comparison across models and provides a comprehensive bias evaluation framework.

Because in the scope of this research we only look at modification where the modification is binary, we would simply the notation so as the original sentence is s and the modified is $\bar{s}$,

$$W_{avg} = \frac{1}{n} \sum_{l \in lang} \cdot \frac{1}{N} \sum_{k=1}^{N} \frac{\left|\mathbb{PS}(s_{l,k}) + \mathbb{PS}(\bar{s}_{l,k})\right|}{2} \tag{7}$$

$$\begin{aligned} \mathbb{NBS}(\theta) &= \frac{1}{W_{avg}} \cdot \frac{1}{N} \sum_{k=1}^{N} \left|\mathbb{PS}(s_{l,k}) - \mathbb{PS}(\bar{s}_{l,k})\right| \\ &= 2n \cdot \frac{\sum_{k=1}^{N} \left|\mathbb{PS}(s_{l,k}) - \mathbb{PS}(\bar{s}_{l,k})\right|}{\sum_{l \in lang} \sum_{k=1}^{N} \left|\mathbb{PS}(s_{l,k}) + \mathbb{PS}(\bar{s}_{l,k})\right|} \end{aligned} \tag{8}$$

where $n = |lang|$.

In the analysis, bias evaluations are conducted using this $\mathbb{NBS}(\theta)$ metric. The closer $\mathbb{NBS}$ is to 0, the lower the bias in the model. If $\mathbb{NBS}$ = 0, it indicates the model treats the two terms equally and exhibits no intrinsic social bias.

B Model Details

In this study we selected three pre-trained models: mBERT, XLM-RoBERTa, BLOOM, XGLM, Qwen and LLaMA. These models are widely used particularly in multilingual tasks.

mBERT (Devlin et al., 2019) is a multilingual version of BERT trained on Wikipedia data in 104 languages. Similar to English BERT, mBERT utilizes Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks for training. mBERT does not require language-specific adaptation, making it directly applicable to text analysis in multiple languages. In this experiment, we used the `google-bert/bert-base-multilingual-cased` version to evaluate bias across different languages.

XLM-RoBERTa Conneau et al. (2020) is a cross-lingual language model based on the RoBERTa architecture, trained on 100 languages. Unlike mBERT, XLM-RoBERTa uses a larger training dataset and longer training cycles, while removing the NSP task and relying solely on MLM. This allows XLM-RoBERTa to perform better in cross-lingual understanding tasks. Our experiments utilized the `FacebookAI/xlm-roberta-base` model.

XGLM (Lin et al., 2022) is a multilingual autoregressive language model designed to facilitate few-shot learning across multiple languages. Trained on a balanced corpus spanning 30 diverse languages and totaling 500 billion sub-tokens, XGLM aims to provide robust cross-linguistic generalization without requiring extensive task-specific finetuning. It follows a decoder-only Transformer architecture, making it well-suited for text generation and language modeling tasks. XGLM has demonstrated strong few-shot performance, highlighting its ability to adapt to new tasks with minimal supervision. In this experiment, we used the `facebook/xglm-564m` version to evaluate bias across different languages.

Gemma 3 (Team et al., 2024) models are available in various parameter sizes, including 1B, 4B, 12B, and 27B, and are designed for multimodal text and image processing. Gemma 3 follows an autoregressive Transformer architecture and supports a large 128K context window, making it suitable for tasks such as question answering, summarization, and reasoning. The model incorporates supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to enhance alignment with human preferences and safety considerations. With multilingual support spanning over 140 languages, Gemma 3 is optimized for global usability and can be further fine-tuned for domain-specific applications. In this experiment, we used the `google/gemma-3-1b-pt` version to evaluate bias across different languages.

Qwen2.5 (Yang et al., 2025) is the latest iteration in the Qwen series of large language models, offering improvements in knowledge retention, coding, and mathematical reasoning. Qwen2.5 models range from 0.5B to 72B parameters and provide enhanced instruction-following, long-text generation, and structured data comprehension. The models support over 29 languages, including Chinese, English, Russian and Indonesian. However, there is no report of Qwen 2.5 supporting Thai language. Qwen2.5 employs a Transformer architecture with RoPE, SwiGLU, and RMSNorm, along with Grouped-Query Attention (GQA) for efficiency. In this experiment, we used the `Qwen/Qwen2.5-0.5B` version to evaluate bias across different languages.

LLaMA 3 (Grattafiori et al., 2024) models are available in 1B and 3B parameter sizes and are optimized for multilingual dialogue, agentic retrieval, and summarization tasks. LLaMA 3 follows an autoregressive Transformer architecture and benefits from supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety. It supports a range of languages, with potential for further fine-tuning on additional languages. In this experiment, we used the `meta-llama/Llama-3.2-1B` version to evaluate bias across different languages.

C Debiasing Method Details

C.1 CDA

Counterfactual Data Augmentation (**CDA**) (Zhao et al., 2018; Zmigrod et al., 2019; Webster et al., 2020) is a bias mitigation technique that generates counterfactual samples to balance the dataset. It involves supplementing training data with modified sentences and evaluating the impact on bias reduction in low-resource language datasets. The process consists of the following steps:

1. Duplicating sentences that contain predefined biased attribute words.
2. Swapping biased attributes with their counterfactual counterparts (*e.g.*, replacing *he* with *she*).
3. Fine-tuning the model with the augmented dataset to reduce bias

Previous studies Webster et al. (2020) have shown that training English models (*e.g.*, ALBERT and BERT) on CDA-augmented datasets can significantly reduce bias. However, some research Reusens et al. (2023a) found that fine-tuning mBERT on English datasets actually increased bias in French and German. In our experiment, we apply CDA fine-tuning on English data, measure bias changes in English and four other languages, and determine whether English-based bias mitigation can be effectively transferred via multilingual learning.

C.2 DO

Dropout Regularization (**DO**) is a commonly used deep learning technique to prevent overfitting. The idea is to randomly *drop* (temporarily remove) certain network nodes during training, forcing the model to learn more generalizable representations. Since DO disrupts word association patterns in attention mechanisms, previous research (Webster et al., 2020) hypothesized that DO could also reduce gender and other types of bias. Studies found that DO fine-tuning effectively reduced bias in ALBERT and BERT, without modifying training data distribution or making explicit assumptions about bias patterns. Additionally, Reusens et al. (2023a) reported that DO fine-tuning in English reduced bias by 10% on French models.

In our experiment, we fine-tune XLM-RoBERTa using DO on English datasets and evaluate its impact on bias reduction across multiple languages.

C.3 SenDeb

SenDeb: Sentence Representation Vector Debiasing Algorithm Initialize the sentence (pretrained) encoder M_θ. Define bias types (e.g., binary gender: male g_m

and female g_f). Design the bias attribute lexicon $D = \{(w_1^{(i)}, \ldots, w_d^{(i)})\}_{i=1}^{m}$. $S = \bigcup_{i=1}^{m} \text{CONTEXTUALIZE}(w_1^{(i)}, \ldots, w_d^{(i)}) = \{(s_1^{(i)}, \ldots, s_d^{(i)})\}_{i=1}^{n}$ Integrate words into sentences $j \in [d]$ $R_j = \{M_\theta(s_j^{(i)})\}_{i=1}^{n}$ Obtain sentence vectors $V = \text{PCA}_k \left(\bigcup_{j=1}^{d} \bigcup_{w \in R_j} (w - \mu_i) \right)$ Compute the bias subspace each new sentence vector h $h_V = \sum_{j=1}^{k} \langle h, v_j \rangle v_j$ Compute projection onto the bias subspace $\hat{h} = h - h_V$ Subtract the projection SentenceDebias (**SenDeb**) (Liang et al., 2020) is a projection-based debiasing technique that extends debiased word vectors to full sentence representations. Previous research on bias mitigation tends to operate at the word level. However, supervised datasets are limited by vocabulary size (Bolukbasi et al., 2016), whereas the number of possible sentences is infinite, making it extremely difficult to precisely characterize bias-free sentences. Therefore, our approach converts these words into sentences to obtain feature representations from a pretrained sentence encoder. The following subsections describe the method used to address this problem. The specific implementation steps are as follows:

1. **Defining Bias Attributes**: For example, when characterizing gender bias, we use word pairs such as *(male, female)* to represent gender. Each tuple should consist of words that are semantically equivalent except for the bias attribute. Typically, for d-class bias attributes, the word pairs form a dataset $D = \{(w_1^{(i)}, \ldots, w_d^{(i)}))\}_{i=1}^{m}$ with m entries, where each entry $(w_1, \ldots, w_d)$ is a d-tuple.
2. **Computing the Subspace**: There exists a common bias subspace in all possible sentence representations. To accurately estimate this bias subspace, we should use as diverse sentence templates as possible to account for the word's position in surrounding contexts. In experiments, we retrieve attribute words from a corpus and place them into biased attribute sentences using a CDA-based approach, further obtaining their sentence representations. This results in a significantly expanded biased attribute sentence dataset S:

$$\begin{aligned} S &= \bigcup_{i=1}^{m} \text{CONTEXTUALIZE}(w_1^{(i)}, \ldots, w_d^{(i)}) \\ &= \{(s_1^{(i)}, \ldots, s_d^{(i)})\}_{i=1}^{n} \end{aligned} \tag{9}$$

   In an encoder M_θ parameterized by θ, the sentence representation vectors $R_j, j \in [d]$ satisfy $R_j = \{M_\theta(s_j^{(i)})\}_{i=1}^{n}$. Among all sentence representation vectors, we can estimate the bias subspace using Principal Component Analysis (PCA) (Abdi & Williams, 2010). Defining $\mu_j = \frac{1}{|R_j|} \sum_{w \in R_j} w$, and assuming that the first K dimensions of PCA define the bias subspace, the subspace $V = \{v_1, \ldots, v_k\}$ satisfies:

$$V = \text{PCA}_k \left(\bigcup_{j=1}^{d} \bigcup_{w \in R_j} (w - \mu_i) \right) \tag{10}$$

3. **Removing Subspace Projection**: By removing the projection onto this bias subspace, we can eliminate bias in general sentence representations. Given a sentence representation vector h, we first compute its projection h_V onto the bias subspace and then subtract this projection to obtain a vector $\hat{h}$ that is approximately bias-free and orthogonal to the bias subspace.

$$h_V = \sum_{j=1}^{k} \langle h, v_j \rangle v_j \tag{11}$$

$$\hat{h} = h - h_V \tag{12}$$

Beyond these three methods, there are also other approaches such as Self-Debias, Iterative Nullspace Projection (INLP), and DensRay (DR). However, since Self-Debias is a post-hoc

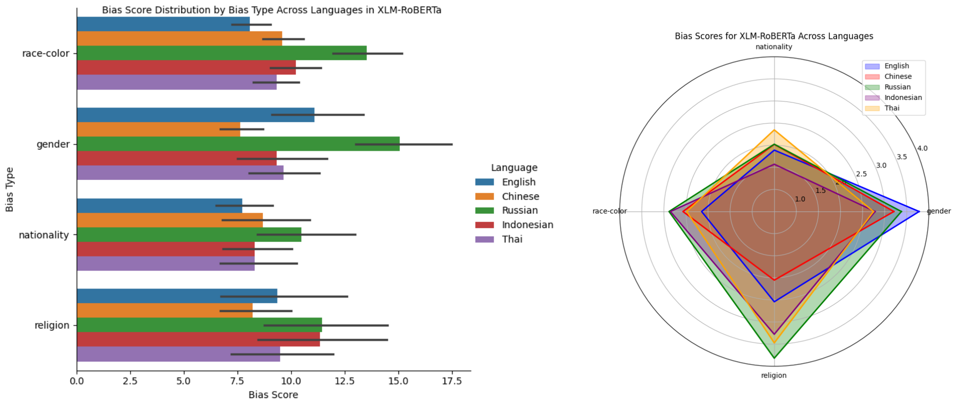

Figure 2: Bias Score Distribution by Bias Type Across Langauges in XLM-RoBERTa as bar charts and radar charts.

text generation debiasing procedure, it cannot be used as a debiasing technique for downstream natural language understanding tasks. Furthermore, `INLP`, `DR`, and `SenDeb` are all projection-based debiasing techniques, therefore we only chose to experiment with `SenDeb`. Through these three methods, we attempted to conduct experiments using multilingual models and report the bias indices before the debiasing experiments and the optimization improvements after the experiments were completed, in order to measure the effectiveness of the debiasing techniques.

D Bias Score Visualizations

In this section, we present additional visualizations of the bias scores computed across different models and languages. The figures provide a detailed breakdown of bias distributions for gender, nationality, race-color, and religion. Each plot illustrates how bias manifests in different linguistic contexts, allowing for a comparative analysis of bias trends across multilingual models.

Figures 2-3 depict the bias scores for various models: mBERT, XLM-RoBERTa, Qwen 2.5, XGLM and LLaMA 3 The x-axis represents the log-scaled bias score, while the y-axis categorizes bias types across different languages. To enhance readability, we have positioned the legends outside the main plots and adjusted the figure sizes accordingly.

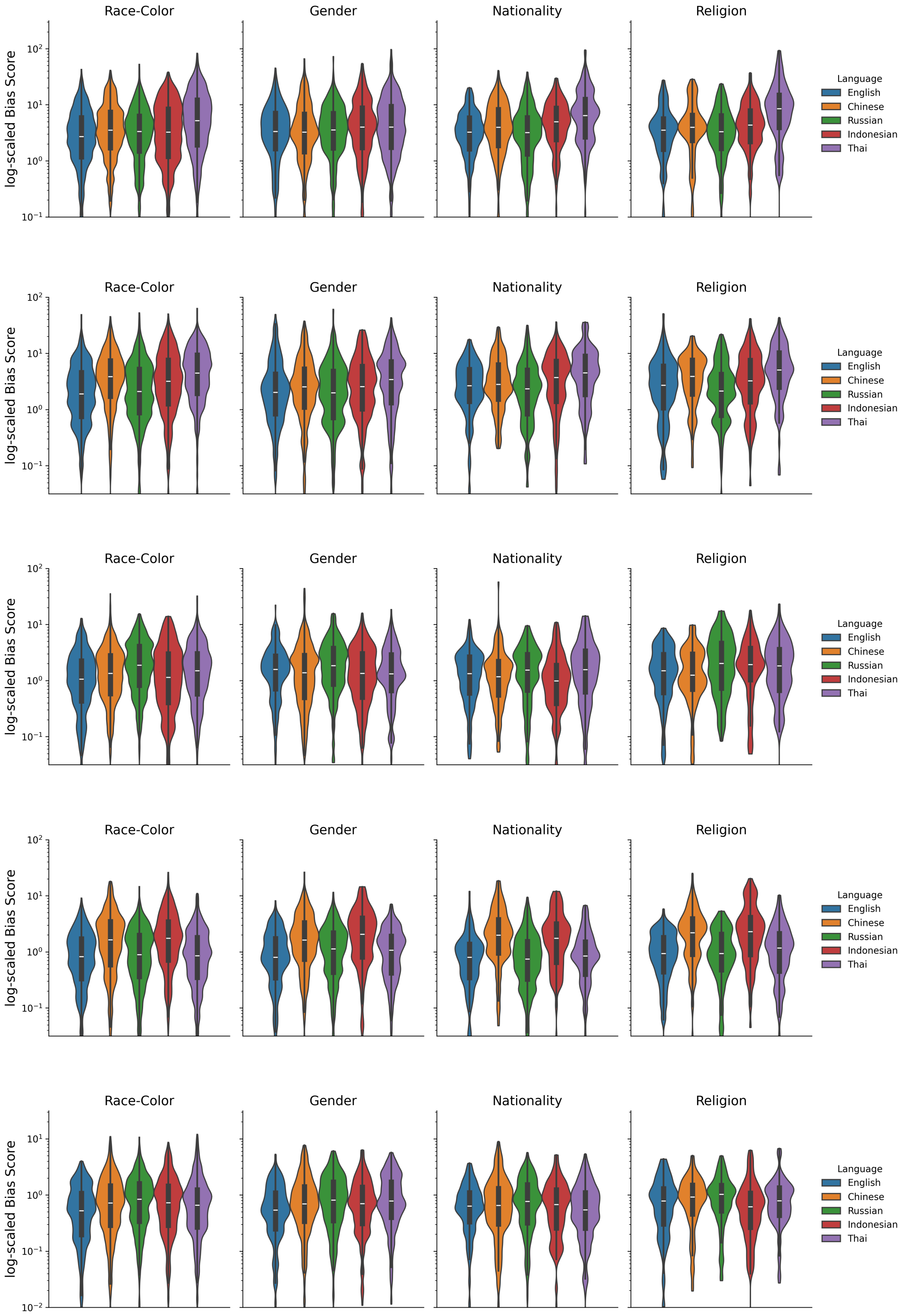

Figure 3: Bias Score Distribution Across Different Languages and Bias Categories in **mBERT**, **XLM-RoBERTa**, **XGLM**, **Qwen 2.5** , **LLaMA 3**, from top to bottom.