Iterative Multilingual Spectral Attribute Erasure

Anonymous ACL submission

Abstract

Multilingual representations embed words with similar meanings to share a common semantic space across languages, creating opportunities for transferring debiasing effects between languages. However, existing methods typically operate on individual languages, show limited transferability of debiasing effects across languages. We present MUSAL (MUltilingual Spectral Attribute removaL), which identifies and mitigates joint bias subspaces across multiple languages through iterative SVD-based truncation. Evaluating MUSAL across eight languages and five demographic dimensions, we demonstrate its effectiveness in both standard and zero-shot settings, where target language data is unavailable but linguistically similar languages can be used for debiasing. Our comprehensive experiments across diverse language models (BERT, LLaMA, Mistral) show MUSAL outperforms traditional monolingual and cross-lingual approaches while maintaining model utility. [1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable success across various domains, yet bias can emerge at multiple stages of training and deployment (Hovy and Prabhumoye, 2021; Chu et al., 2024), raising ethical concerns in downstream applications (Lauscher et al., 2021). Debiasing methods aim to mitigate this by reducing models' reliance on demographic patterns and promoting fairness across populations. Most approaches require pairing texts with authors' protected attributes to remove sensitive information from model representations (Reusens et al., 2023; Liang et al., 2020b). However, the difficulty of obtaining large-scale demographic labels has led most fairness studies to focus exclusively on English datasets (Orgad and Belinkov, 2023). To address this, multilingual debiasing leverages transfer learning to mitigate bias in a target language by incorporating information from multiple source languages. Existing approaches typically identify a small set of protected attribute directions—such as gender—in a single source language and apply debiasing to the target language by nullifying projections into these directions (Liang et al., 2020b). Methods include null space projection (Gonen et al., 2022), semantic gender shifting (Zhou et al., 2019), and aligning embeddings across representational spaces (Zhao et al., 2020). This line of work frames multilingual debiasing as a cross-lingual transfer problem: detecting bias in one language and applying the learned debiasing transformation to another. However, state-of-the-art methods remain limited in their ability to fully remove gender bias through transfer learning (Vashishtha et al., 2023), as they fail to account for cultural nuances and demographic variations across languages (Talat et al., 2022).

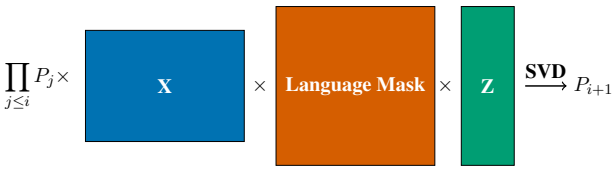

Figure 1: A visualization of MUSAL. A sequence of projections is created using SVD based on the input representations (r.v. $\mathbf{X}$), the guarded attributes (r.v. $\mathbf{Z}$) and a language mask that dictates which languages to use.

Gonen et al. (2022) showed that a joint gender subspace exists across languages, enabling cross-lingual gender prediction, but did not address how to neutralize this subspace for bias mitigation. Despite extensive work on cross-lingual debiasing, existing methods have yet to effectively identify and mitigate joint bias subspaces across multiple languages. Our work moves beyond single-language approaches by developing a method to identify and

[1] We will release our code and data upon acceptance

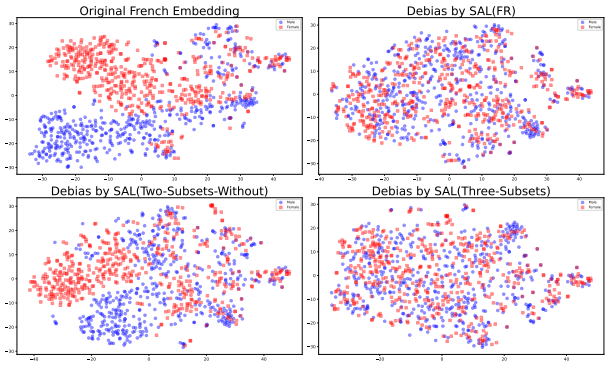

Figure 2: Visualization of gender bias in MBERT French embeddings using t-SNE. Top left: Original embeddings showing clear gender clustering. Top right and bottom: Results after applying MUSAL, demonstrating effective elimination of gender-based patterns through both monolingual (French-only) and cross-lingual (other languages) debiasing approaches.

neutralize these subspaces, particularly in linguistically similar languages. We introduce MUSAL (Multilingual Spectral Attribute removaL), a structural extension of Shao et al. (2023b). MUSAL iteratively debiases subsets of source languages using singular value decomposition (SVD) truncation, as illustrated in Figure 1. These subsets may overlap, and at each step, a shared subspace capturing the guarded attribute across languages is identified and neutralized. By progressively refining these subspaces, MUSAL ensures a more comprehensive removal of bias while preserving multilingual representations. This effectiveness is visualized in the t-SNE plots in Figure 2: while the original embeddings (top left) show clear gender clustering, applying MUSAL (remaining plots) successfully obscures these gender-based patterns. Notably, MUSAL achieves similar debiasing results whether using French data directly or only leveraging other languages (bottom right), demonstrating its ability to capture and neutralize shared bias patterns across languages. Our key contributions include:

- Introducing MUSAL, a method for identifying and mitigating shared bias subspaces across multiple languages. We validate its effectiveness on eight languages and five demographic dimensions using multilingual fairness benchmarks.
- Demonstrating that targeting joint bias subspaces enables superior zero-shot bias mitigation in linguistically similar but unseen languages, outperforming monolingual and cross-lingual debiasing methods.
- Establishing a comprehensive evaluation framework by comparing MUSAL with three state-of-the-art post-hoc debiasing methods across diverse languages, demographic attributes, and model families (LLaMA, Mistral, and BERT).
- Presenting the MSEFair (Multilingual Stack Overflow Fairness) dataset to further validate MUSAL and facilitate future research on fairness in non-English languages.

2 Multilingual Debiasing

2.1 Problem Formulation

For an integer n, we let $[n] = \{1, \ldots, n\}$. Let $\mathcal{L}$ be a set of languages indexed by integers. Let $\mathcal{L}_s \subseteq \mathcal{L}$ be a subset of source languages and $\mathcal{L}_t$ be a subset of target languages. Note we do not require that $\mathcal{L}_s \cap \mathcal{L}_t = \emptyset$.

We are assuming a joint multilingual representation space for the languages, where text from any language in $\mathcal{L}$ can be represented in a vector from that space in $\mathbb{R}^d$. We assume d-dimensional random vectors $\mathbf{X}_\ell$ for any $\ell \in \mathcal{L}$. These vectors vary over the representations.

Our goal is to use representations for languages from $\mathcal{L}_s$ to erase information about a random vector $\mathbf{Z}$ from representations of languages in $\mathcal{L}_t$. The algorithm is based on the SAL algorithm, and adds to it a structural component. The SAL algorithm erases protected attribute markings from neural representations by computing a cross-covariance matrix between the input representations and the protected attribute, and then projecting the input representations to the directions which least covary with the protected attribute.

2.2 The MUSAL Algorithm

Our algorithm, MUltingual Spectral Attribute removaL (or MUSAL) is based on the SAL algorithm presented by Shao et al. (2023b). Rather than relying on a single projection that erases information from the input representations based on the cross-covariance matrix between input representations and a guarded attribute, it creates a sequence of such projections, each corresponding to inputs from a predefined subset of languages.

More specifically, Figure 3 provides the MUSAL algorithm that we develop. The algorithm uses n_ℓ samples from the representations of each language in $\mathcal{L}_s$, $\mathbf{x}^{(\ell,i)}$ where $\ell \in \mathcal{L}_s$ and $i \in [n_\ell]$. In addition, there are corresponding samples $\mathbf{z}^{(\ell,i)}$. The algorithm also receives as input a sequence of possibly

Inputs: Samples $\mathbf{x}^{(\ell,i)}$, and $\mathbf{z}^{(\ell,i)}$, $\ell \in \mathcal{L}_s$ and $i \in [n_\ell]$, $\mathcal{L}_1, \ldots, \mathcal{L}_m$ subsets of $\mathcal{L}_s$.
Algorithm: (erase information based on the samples sequentially)

Initialize $\boldsymbol{P}^*$ to be the identity matrix.

Repeat the following for $j \in [m]$:

- Calculate $\boldsymbol{\Omega}$ as follows:

$$\boldsymbol{\Omega} \leftarrow \sum_{\ell \in \mathcal{L}_j} \sum_{i=1}^{n_\ell} \mathbf{x}^{(\ell,i)} (\mathbf{z}^{(\ell,i)})^\top,$$
$$\boldsymbol{\Omega} \leftarrow \boldsymbol{P}^* \boldsymbol{\Omega}.$$

- Calculate SVD on $\boldsymbol{\Omega}$ to calculate $(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V})$ with bottom k left singular vectors being $\boldsymbol{U}$.
- Update $\boldsymbol{P}^* \leftarrow \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{P}^*$.

Return: The erasure matrix $\boldsymbol{P}^*$.

Figure 3: The MUSAL algorithm.

overlapping subsets of $\mathcal{L}_s$, denoted $\mathcal{L}_1, \ldots, \mathcal{L}_m$. Each of this subset determines one possible way in the sequence to jointly remove bias. The sequence defines a spectrum of how to group languages which erasing information.

We explore the interplay between the different languages by grouping together the source languages in various ways. More specifically, we will focus in three specific settings:

- Monolingual or cross-lingual (we assume $|\mathcal{L}_t| = 1$): where $m = 1$ and $|\mathcal{L}_1| = 1$. This means we use one language (possibly different than the target language) to erase information.
- All subsets without the target languages: where $m = 2^{|\mathcal{L}_s \setminus \mathcal{L}_t|} - 1$, and the m subsets of $\mathcal{L}_s$ vary over all possible subsets of languages (except for the empty subset), excluding any target languages.
- All subsets with the target languages: where $m = 2^{|\mathcal{L}_s|} - 1$, and we use all subsets of the source languages except for the empty set.

Note that in the above the order the subsets, which is important to consider in the execution of MUSAL, is left underspecified. More of this is discussed in §2.3. In addition, note that we recover the SAL algorithm of Shao et al. (2023b) when $|\mathcal{L}| = 1$, $m = 1$ and $\mathcal{L}_s = \mathcal{L}_t = \mathcal{L}$.

2.3 Order Sensitivity in Sequential Projections

The final erasure matrix $\boldsymbol{P}^*$ is a product of projection matrices, and matrix multiplication is not commutative in general. Therefore, the order of applying these projections - whether we first remove bias using all languages jointly (global) and then refine language-specific components (local), or vice versa - could potentially affect the final debiased representations. Consider two possible orderings:

- Global-then-Local: First apply a projection using all languages L to identify and remove shared bias directions, followed by language-specific projections for each $\ell \in \mathcal{L}_s$.
- Local-then-Global: First apply individual projections for each language, then combine the insights to remove remaining shared bias components.

The global approach may capture broad bias patterns that are diluted when looking at languages individually, while the local-first approach may better preserve language-specific nuances. In our empirical analysis, we found that both orderings achieve similar debiasing performance across our evaluation tasks. This suggests that while the mathematical difference exists, the practical impact is limited - likely because the core bias directions are relatively stable regardless of the order of removal. Therefore, we report the Global-then-Local ordering in our main experiments.

2.4 A Fully Joint MUSAL Baseline

A straightforward approach to remove unwanted information from texts across multiple languages is to concatenate all the input representations from the different languages and their corresponding guarded attributes and feed them to an algorithm such as SAL. This corresponds to running MUSAL with $m = 1$, and $\mathcal{L}_1 = \mathcal{L}$. This one-shot reduction may be less effective than MUSAL in its full generality, which removes information iteratively on a per-language subset. We refer to this baseline as "FullyJoint."

Assume the guarded attribute has dimension one. The cross covariance matrix between the input representations for a specific language ℓ and the guarded attributes would have a rank o 1, and therefore, SVD yields a single direction $\mathbf{u}$ to remove.

Removing this direction from the space of the input representations (i.e., projecting onto the orthogonal complement of $\mathbf{u}$) eliminates all linear information linking the input representations from ℓ and the guarded attribute. If we concatenate all the texts and all the labels across languages, the overall label matrix is still of rank 1, with SVD leading to a single direction $\mathbf{u}'$.

However, if $\mathbf{u}'$ does not exactly coincide with each language-specific direction $\mathbf{u}$ for each language ℓ, then projecting the concatenated inputs will not eliminate the protected information present in each language. Some residual association between the input representations of ℓ and the protected attributes may remain, allowing a linear classifier to predict the protected attribute.

Thus, by removing the attribute-related direction iteratively for each language, we ensure that for every language ℓ the specific information linking the input representations and the protected attributes is fully removed.

3 Experiments

We explore the effectiveness of MUSAL in two scenarios. First, when the target language is included in the training set, we demonstrate that MUSAL, with additional languages, yields better results compared to monolingual debiasing. Second, when the target language is not part of the training set, we show that using multiple source languages via MUSAL outperforms the typical approach of conducting cross-lingual debiasing with a single source language. We experiment with three tasks: profession prediction (§3.1), hate speech recognition (§3.2) and helpfulness prediction (§3.3). For more information on the data sets split and statistics, see Appendix A. We use the profession prediction task to conduct preliminary experiments showing the limitations of current cross-lingual methods and the Fully Joint MUSAL Baseline.

Evaluation Metrics Our ultimate goal is to reduce bias while ensuring high downstream task performance. We measured disparities in classifier performance across different protected groups to quantify bias in language models. For instance, we compared the performance of male and female biographies in our profession prediction task. Specifically, we employed the True Positive Rate Gap (TPR-Gap), which calculates the difference in true positive rates between demographic groups, conditioned on the true class. A lower TPR-Gap indicates greater fairness, as it suggests the model performs similarly for both gender groups when predicting professions. We use accuracy to measure the downstream task performance.

Target	Baseline		EN - SAL		EN - INLP		EN - SentenceDebias	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased								
EN	80.5	15.4	↓0.1 80.4	↓1.9 13.5	80.5	↓0.2 15.2	↓0.2 80.3	↓0.1 15.3
DE	77.7	27.6	↑0.1 77.8	↓4.5 23.1	↑0.1 77.8	↓2.2 25.4	↓0.1 77.6	↓2.4 25.2
FR	72.7	22.8	↓0.1 72.6	↓0.8 22.0	72.7	↓0.5 22.3	↑0.1 72.8	↓0.5 22.3
Avg-Diff	Nan	Nan	↓0.03	↓2.4	↑0.03	↓2.9	↓0.07	↓1.0
Llama-3.1-8B								
EN	80.9	13.6	↓2.0 78.9	↓0.3 13.3	↓0.5 80.4	↑0.3 13.9	↑1.2 82.1	↑1.0 14.6
DE	79.8	26.8	↓0.3 79.5	↑0.2 27.0	79.8	26.8	↓0.2 79.6	↑0.2 27.0
FR	72.4	25.2	↑0.1 72.5	↑0.2 25.4	↑0.1 72.5	↑1.5 26.7	↑0.1 72.5	↑2.1 27.3
Avg-Diff	Nan	Nan	↓0.73	↑0.03	↓0.13	↑0.6	↑0.37	↑1.1
Mistral-7B-Instruct-v0.3								
EN	80.0	14.1	↓2.5 77.5	↓1.2 12.9	↑0.3 80.3	↓0.3 13.8	↑2.7 82.7	↓1.1 13.0
DE	77.3	23.3	↓0.2 77.1	↑0.3 23.6	77.3	23.3	↓0.2 77.1	↓0.1 23.2
FR	71.6	22.0	↑0.1 71.7	↓0.9 21.1	↓0.2 71.4	↑0.7 22.7	71.6	22.0
Avg-Diff	Nan	Nan	↓0.87	↓0.6	↑0.03	↑0.13	↑0.83	↓0.4

Table 1: Evaluation of post-hoc debiasing methods on multilingual BiasBios. The main task is profession prediction, while the TPR-Gap (True Positive Rate Gap) between males and females demonstrates the extrinsic bias in downstream tasks.

3.1 Fair Profession Prediction

Task and Data We use the Multilingual BiasBios dataset (Zhao et al., 2020), an extension of the original BiasBios dataset (De-Arteaga et al., 2019) that includes French, Spanish, and German biographies. The dataset was constructed by extracting biographies from Common Crawl using the template "NAME is an OCCUPATION-TITLE". Each biography is annotated with gender and profession labels. For our experiments, we used multilingual BERT (Devlin et al., 2019), Llama 3 (Grattafiori et al., 2024), Llama 3.1 (Meta, 2024a), Llama 3.2 (Meta, 2024b), Mistral 7B (Jiang et al., 2023) and Mistral Nemo (Mistral AI, 2024) to generate text representations.

3.1.1 Crosslingual Debiasing Results

To assess the cross-linguistic transferability of post-hoc debiasing methods, we evaluated three projection-based approaches: null-space projection (INLP) (Ravfogel et al., 2020), SVD-based Spectral Attribute Removal (SAL) (Shao et al., 2023b), and PCA-based SentenceDebias (Liang et al., 2020a). Table 1 presents the results using English as the source language, while the full results, which exhibit similar trends. While all methods maintain strong downstream task performance in MBERT, their effectiveness varies across model architectures. SAL demonstrate superior bias reduction in LLMs, with average results of 2.4% and 0.6% in MBERT and Mistral respectively. Based

Target	Baseline		Monolingual		Average - Crosslingual		FullyJoint		Two-Subsets-Without		Three-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Multilingual BERT												
EN	80.5	15.4	↓0.1 80.4	↓1.9 13.5	80.5	↑0.5 15.9	↓0.1 80.4	↓1.2 14.2	80.5	↑0.4 15.8	↓0.1 80.4	↓1.9 13.5
DE	77.7	27.6	↓0.3 77.4	↓0.3 27.3	↑0.1 77.8	↓2.2 25.4	↑0.1 77.8	↓4.6 23.0	↑0.1 77.8	↓4.0 23.6	↓0.2 77.5	↓2.1 25.5
FR	72.7	22.8	↓0.5 72.2	↓3.4 19.4	72.7	↓2.1 20.7	↓0.1 72.6	↓1.5 21.3	↑0.2 72.9	↓0.4 22.4	↓0.6 72.1	↓3.3 19.5
Avg-Diff	Nan	Nan	↓0.30	↓1.87	↑0.03	↓1.27	↓0.03	↓2.57	↑0.10	↓1.33	↓0.30	↓2.43
Llama-3.1-8B												
EN	80.9	13.6	↓2.0 78.9	↓0.3 13.3	↓0.6 80.3	↓0.4 13.2	↓1.9 79.0	↓0.6 13.0	↓0.6 80.3	↓0.3 13.3	↓2.0 78.9	↓0.8 12.8
DE	79.8	26.8	↓0.4 79.4	↓5.2 21.6	↓0.2 79.6	↑0.4 27.2	↓0.2 79.6	↑0.3 27.1	↓0.4 79.4	↑0.4 27.2	↓0.5 79.3	↓4.4 22.4
FR	72.4	25.2	72.4	↓5.9 19.3	↑0.1 72.5	↓0.2 25.0	↑0.1 72.5	↓1.3 23.9	↑0.1 72.5	↓0.9 24.3	72.4	↓4.5 20.7
Avg-Diff	Nan	Nan	↓0.8	↓3.80	↓0.23	↓0.07	↓0.67	↓0.53	↓0.3	↓0.27	↓0.83	↓3.23
Mistral-7B-Instruct-v0.3												
EN	80.0	14.1	↓2.5 77.5	↓1.2 12.9	↓0.3 79.7	↓0.1 14.0	↓2.3 77.7	↓1.1 13.0	↓0.3 79.7	↑0.2 14.3	↓2.3 77.7	↓1.3 12.8
DE	77.3	23.3	77.3	23.3	↓0.2 77.1	↑0.5 23.8	↓0.2 77.1	↑0.3 23.6	↓0.4 76.9	↓0.8 22.5	↓0.3 77.0	↓0.2 23.1
FR	71.6	22.0	↑0.1 71.7	↓3.7 18.3	↑0.1 71.7	↓0.7 21.3	↓0.1 71.5	↑0.3 22.3	↑0.1 71.7	↓1.7 20.3	↑0.2 71.8	↓3.8 18.2
Avg-Diff	Nan	Nan	↓0.8	↓1.63	↓0.13	↓0.10	↓0.87	↓0.07	↓0.20	↓0.77	↓0.80	↓1.77

Table 2: Evaluation of demographic bias mitigation on the multilingual BiasBios dataset using MBERT. Main shows hate speech detection accuracy, while TPR-GAP shows true positive rates between different demographic groups. Results compare Baseline, Monolingual (target language only), Four-Subsets-Without (excluding target language), and Five-Subsets (all languages) approaches across five languages.

on these findings, we selected SAL as our baseline and further report results on MUSAL as its structural variant in different settings.

3.1.2 MUSAL Results

With Target Language Consider Table 2. For two out of three LMs, incorporating information from additional languages using MUSAL ("Three-Subsets") further reduces bias compared to relying solely on the target language ("Monolingual"). While the FullyJoint approach slightly outperforms MUSAL for MBERT, MUSAL significantly outperforms FullyJoint for both LLMs. Regarding downstream task performance, all methods perform well, with only a relatively small drop in accuracy.

Without Target Language We observe that for all three LMs, MUSAL ("Two-Subsets-Without") outperforms the average cross-lingual debiasing method using a single source language in terms of debiasing. Both approaches minimally affect main-task performance while effectively reducing bias. See also results in Appendix 6.

3.2 Hate Speech Recognition

While BiasBios (De-Arteaga et al., 2019) focuses solely on gender bias, the Multilingual Twitter Hate Speech corpus (Huang et al., 2020) provides a more comprehensive evaluation framework across multiple demographic dimensions (gender, race, country, and age) and languages. Unlike BiasBios (De-Arteaga et al., 2019), which infer the authors' demographic information directly from the text itself, the Multilingual Twitter Hate Speech corpus proposed by Huang et al. (2020) derives demographic attributes from user profiles. This situation presents a more complex challenge compared to fairness datasets that allow demographic information to be easily guessed based on the text alone, which can often be accomplished even through simple keyword searches. When an author's information can be readily predicted from the text, any observed differences in text classification across demographic groups may be directly attributable to specific textual features, thereby undermining the reliability and independence of fairness evaluations.

Task and Data The Multilingual Twitter Hate Speech Recognition dataset (Huang et al., 2020) is a compilation of previously published dataset across five languages: English, Spanish, Italian, Polish, and Portuguese. For training, we used approximately 32,000, 1,900, 1,600, 6,800, and 800 samples from these respective languages. The test set sizes range from 20% to 25% of the corresponding training set sizes. Some results are excluded due to severe class imbalance in certain subsets. Complete data statistics are provided in Tables 7 and 8 in the Appendix. Huang et al. (2020) labeled the datasets by inferring the author attributes from user profiles across four demographic dimensions: gender (male/female), race (white/non-white), age (young/old), and country (US/non-US). The primary labels assigned to each tweet indicate whether it contains hate speech or not.

Evaluation Measures The fairness evaluation on sentiment recognition is same to profession prediction in §3.1. We quantify the biases in language

Target	Baseline		Monolingual		Average - Crosslingual		Four-Subsets-Without		Five-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Race										
EN	86.8	4.1	↓1.7 85.1	↓1.3 2.8	86.8	↑0.1 4.2	↓0.2 86.6	↓2.0 2.1	↓1.8 85.0	↓4.0 0.1
ES	63.7	10.2	↑0.4 64.1	↓0.1 10.1	↑0.5 64.2	↓0.2 10.0	↑1.2 64.9	↓9.6 0.6	↑1.2 64.9	↓6.5 3.7
IT	-	-	-	-	-	-	-	-	-	-
PL	91.3	6.2	91.3	↓1.2 5.0	91.3	↓0.3 5.9	↑0.2 91.5	↓5.6 0.6	↓0.3 91.0	↑0.6 6.8
PT	61.3	1.0	↓0.6 60.7	↑0.1 1.1	↓0.7 60.6	↑1.2 2.2	61.3	↑1.2 2.2	↑0.7 62.0	↑10.6 11.6
Avg-Diff	Nan	Nan	↓0.48	↓0.625	↓0.05	↓0.20	↑0.30	↓4.00	↓0.05	↑0.18
Gender										
EN	86.7	4.4	86.7	↑0.7 5.1	↑0.1 86.8	4.4	86.7	↑1.0 5.4	↑0.2 86.9	↓4.4 0.0
ES	63.7	3.6	↑0.4 64.1	↓0.2 3.4	↓0.1 63.6	3.6	↑0.4 64.1	↑1.9 5.5	↑0.9 64.6	↓1.0 2.6
IT	68.4	2.1	68.4	↓0.6 1.5	↓0.3 68.1	↓0.4 1.7	68.4	↑1.3 3.4	↑0.7 69.1	↑6.1 8.2
PL	88.2	11.6	↓0.2 88.0	↓8.8 2.8	88.2	↓0.1 11.5	↑0.2 88.4	↓10.0 1.6	↓0.6 87.6	↓11.6 0.0
PT	61.3	12.0	↑0.7 62.0	↑1.4 13.4	↓0.4 60.9	↑0.8 12.8	↑0.7 62.0	↑0.7 12.7	↓0.6 60.7	↓3.0 9.0
Avg-Diff	Nan	Nan	↑0.18	↓1.5	↓0.14	↑0.06	↑0.26	↓1.02	↑0.12	↓2.78
Age										
EN	86.7	9.2	↑0.3 87.0	↑0.5 9.7	↓0.2 86.5	↓0.2 9.0	↓0.6 86.1	↓4.3 4.9	↓0.3 86.4	↓7.2 2.0
ES	63.7	12.9	↓0.3 63.4	↓0.5 12.4	63.7	12.9	↑0.2 63.9	↓9.8 3.1	63.7	↓3.6 9.3
IT	68.2	3.6	↓0.3 67.9	↑0.3 3.9	68.2	↑0.1 3.7	↓0.5 67.7	↑4.9 8.5	↓1.5 66.7	↓3.3 0.3
PL	91.3	8.8	↓0.9 90.4	↓1.9 6.9	91.3	↓0.8 8.0	↓0.1 91.2	↓8.8 0.0	↓1.0 90.3	↓8.2 0.6
PT	61.3	17.6	↓0.6 60.7	↑1.5 19.1	↓0.9 60.4	↑0.5 18.1	↑0.7 62.0	↑5.8 23.4	↑0.7 62.0	↓5.7 11.9
Avg-Diff	Nan	Nan	↓0.36	↓0.02	↓0.22	↓0.08	↓0.06	↓2.44	↓0.42	↓5.6
Country										
EN	82.3	6.7	↓0.1 82.2	↓0.8 5.9	82.3	6.7	↓0.1 82.2	↓4.0 2.7	↓0.4 81.9	↓6.6 0.1
ES	65.1	5.1	65.1	↓0.2 4.9	↓0.1 65.0	↑0.4 5.5	↓0.4 64.7	↑6.0 11.1	↑0.3 65.4	↑3.3 8.4
IT	71.0	1.7	↑0.1 71.1	↓0.4 1.3	↓0.1 70.9	↑0.5 2.2	↓0.7 70.3	↑2.0 3.7	71.0	↑9.7 11.4
PL	-	-	-	-	-	-	-	-	-	-
PT	64.5	5.1	↓0.5 64.0	↓4.3 0.8	↑0.5 65.0	↑0.5 5.6	↑2.0 66.5	↓2.7 2.4	↓0.5 64.0	↑1.5 6.6
Avg-Diff	Nan	Nan	↓0.13	↓1.43	↑0.08	↑0.35	↑0.20	↑0.33	↓0.15	↑1.98

Table 3: Demographic bias mitigation results on multilingual HateSpeech dataset using MBERT, comparing monolingual and multilingual debiasing approaches. We exclude results for Italian in race bias evaluation and Poland in country bias evaluation due to severely imbalanced class distributions in these subsets that could lead to unreliable bias measurements. Detailed dataset statistics can be found in table 7 and table 8 in Appendix.

models by measuring the gap of true positive rate on hate speech recognition between different populations. A lower TPR gap indicates a fairer model.

3.2.1 Results

With Target Language Consider Table 3. MUSAL ("Five-Subsets") demonstrates stronger bias mitigation compared to monolingual debiasing for gender and age debiasing. For race bias, we achieve reductions of 4.0% for English and 6.5% for Spanish, compared to monolingual reductions of 1.3% and 0.1% respectively. Similarly significant improvements are seen for gender bias (11.6% reduction by MUSAL on Polish vs 8.8% monolingual) and age bias for all languages. Importantly, these improvements come with minimal impact on main task performance, with accuracy changes generally below 2%.

Without Target Language Even when target language data is unavailable, MUSAL ("Four-Subsets-Without") effectively reduces bias across different attributes. Using only non-target languages, we achieve bias reductions of 4.0% for race, 1.0% for gender, 2.4% for age on average. The effectiveness varies by demographic attribute, with nationality bias proving more challenging to mitigate - likely because it is less directly inferable from tweet content compared to other attributes. Notably, main task performance remains stable, demonstrating MUSAL's ability to preserve useful features while removing bias. See also results in Appendix B.2.

3.3 Multilingual Stack Exchange Fairness Benchmark

Previous debiasing research has primarily focused on Anglo-centric languages, with limited attention to non-Western contexts (Ramesh et al., 2023). A key unexplored question is whether debiasing effects transfer effectively across linguistically and culturally distant languages. While Vashishtha et al. (2023) extended DisCo (Webster et al., 2021) to Indian languages through human translation, this approach, like the translation of CrowS-Pairs, fails to capture culture-specific bias manifestations (Névéol et al., 2022). Moreover, existing studies have largely focused on well-explored tasks like sentiment analysis and profession prediction. To address these limitations, we explore bias in more abstract concepts using a Multilingual Stack Exchange dataset we developed. This dataset offers

Target	Baseline		Monolingual		Average - Crosslingual		Two-Subsets-Without		Three-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased										
EN	67.4	10.7	↓4.2 63.2	↓9.4 1.3	↓0.1 67.3	↑0.3 11.0	↓4.1 63.3	↓9.2 1.5	↓4.3 63.1	↓9.4 1.3
PT	78.3	16.0	↓19.6 58.7	↓14.6 1.4	78.3	↑0.2 16.2	↓19.8 58.5	↓15.5 0.5	↓19.6 58.7	↓15.5 0.5
RU	70.0	18.0	↓11.9 58.1	↓15.1 2.9	↓0.3 69.7	↑0.1 18.1	↓0.6 69.4	↑0.1 18.1	↓12.1 57.9	↓14.5 3.5
Avg-Diff	Nan	Nan	↓11.90	↓13.03	↓0.13	↑0.20	↓8.17	↓8.20	↓12.00	↓13.13
Llama-3.1-8B										
EN	68.4	11.2	↓4.3 64.1	↓6.0 5.2	68.4	11.2	↓4.4 64.0	↓6.3 4.9	↓4.4 64.0	↓6.1 5.1
PT	82.7	22.9	↓18.3 64.4	↓14.1 8.8	↓0.3 82.4	↓0.6 22.3	↓18.4 64.3	↓14.4 8.5	↓18.5 64.2	↓14.5 8.4
RU	72.1	16.4	↓8.4 63.7	↓10.3 6.1	↓0.1 72.0	↓0.1 16.3	↓0.1 72.0	↓0.4 16.0	↓8.6 63.5	↓10.5 5.9
Avg-Diff	Nan	Nan	↓10.33	↓10.13	↓0.13	↓0.23	↓7.63	↓7.03	↓10.5	↓10.37
Mistral-7B-Instruct-v0.3										
EN	66.8	9.5	↓3.7 63.1	↓5.8 3.7	66.8	↓0.2 9.3	↓3.6 63.2	↓5.9 3.6	↓3.7 63.1	↓6.3 3.2
PT	81.4	20.9	↓18.2 63.2	↓9.5 11.4	↓0.2 81.2	↑0.4 21.3	↓18.0 63.4	↓9.2 11.7	↓17.9 63.5	↓9.1 11.8
RU	70.5	14.6	↓8.2 62.3	↓9.2 5.4	↓0.1 70.4	↓0.2 14.4	↓0.2 70.3	↓0.6 14.0	↓8.4 62.1	↓9.4 5.2
Avg-Diff	Nan	Nan	↓10.03	↓8.17	↓0.10	0.00	↓7.27	↓5.23	↓10.00	↓8.27

Table 4: Reputation bias mitigation on SEFair dataset across English, Portuguese and Russian, comparing: Baseline, Monolingual (target language only), Average - Crosslingual (excluding target language), MUSAL on Two-Subsets-Without (excluding target language) and MUSAL Three-Subsets (using all languages). Main: helpfulness prediction accuracy; TPR-Gap: True positive rate gap between demographic groups.

two key advantages: it contains verified protected attributes. It represents authentic Russian language use rather than translations, providing a more reliable testbed for cross-lingual debiasing in non-Western contexts.

MSEFair includes content in English, Russian, and Portuguese, evaluated based on helpfulness and user reputation scores. While reputation scores are provided by Stack Exchange's community feedback system, they may introduce bias unrelated to the core task of assessing helpfulness. Our goal is to eliminate information about users' high or low reputation without compromising the performance of helpfulness prediction.

3.3.1 Results

With Target Language Consider Table 4. Our method MUSAL, which uses all available languages ("Three-Subsets"), demonstrates superior bias mitigation on average across all language models compared to monolingual debiasing using SAL. However, both methods cause significant damage to model utility for Portuguese and Russian. We want to bring to the community's attention that Stack Exchange helpfulness is a challenging topic to work on, as small changes in embeddings can lead to huge drops in classification performance.

Without Target Language While MUSAL's effectiveness ("Two-Subsets-Without") varies with linguistic similarity, it consistently outperforms cross-lingual debiasing approaches like SAL. However, for Russian, cross-lingual debiasing shows limited effectiveness, with minimal TPR-Gap reduction. This limitation likely stems from Russian's distinct linguistic features (Cyrillic script, complex case system, different word order), highlighting how structural language differences affect bias transfer. Notably, newer architectures like LLaMA and Mistral show improved cross-lingual debiasing performance compared to MBERT, suggesting better cross-lingual representation alignment in these models.

4 Related Work

This section briefly discusses the state-of-art debiasing methods, the limited work in multilingual debiasing and the challenges involved.

4.1 Debiasing Methods

Removing protected attributes from neural representations and debiasing includes early work about debiasing pre-trained language models (Mitchell et al., 2022; Schick et al., 2021), adding adversarial training objective (Elazar and Goldberg, 2018; Zhang et al., 2018; Xie et al., 2017; Ravfogel et al., 2022), adding counter factual data (Zmigrod et al., 2019; Webster et al., 2020), and post-hoc debiasing (Bolukbasi et al., 2016a; Gonen and Goldberg, 2019; Ravfogel et al., 2020; Belrose et al., 2023).

This paper focuses on post-hoc debiasing methods that remove protected attributes from embeddings by erasing components along bias directions. Post-hoc debiasing offers several advantages: it preserves model functionality without retraining

(Bolukbasi et al., 2016b), requires less computation than training-time methods (Mu et al., 2018; Karve et al., 2019), and provides insights into how biases manifest in model representations (Gonen and Goldberg, 2019).

However, previous methods primarily identify linear biased subspaces using dimensionality reduction techniques like PCA (Mu et al., 2018; Liang et al., 2020a), SVD (Bolukbasi et al., 2016b), and CCA (Liang et al., 2020b), then project embeddings to orthogonal directions. These approaches have been criticized for only masking surface-level biases while deeper patterns remain detectable (Gonen and Goldberg, 2019). Ravfogel et al. (2020) iteratively project representations to the null space of the biases classifier, removing bias, but also significantly damaging the model utility. SAL (Shao et al., 2023b,a) addresses these limitations by identifying directions with minimal covariance with protected attributes while preserving task-relevant information.

4.2 Debiasing for Multilingual Representations

Multilingual models embed words or contexts from different languages within a shared space, s.t. context with similar meanings are closer to each other. It enables transfer learning from one language to another. Debiasing multilingual representations is more challenging than debiasing monolingual representations. One reason is that each language has its own linguistic and cultural properties. (Ramesh et al., 2023). Just one example is the grammatical gender, which is one of nominal classification systems on nouns (Booij, 2010) that may not fully agree with biological sex. For example, German word for 'girl', Mädchen, is a neutral noun (Veeman et al., 2020). Another challenge is that the creation of multilingual representations can introduce new biases (Zhao et al., 2020). Gonen et al. (2022) demonstrated that gender components are neither fully shared across languages nor fully disjoint, which makes it hard to find the shared gender subspaces across languages.

Past work mainly focus on crosslingual debiasing, namely transferring debiasing effect from one language to another. Zhou et al. (2019) made a fundamental contribution by decomposing gender bias into grammatical and semantic components, enabling targeted debiasing through either word shifts along semantic gender dimensions or alignment with debiased English embeddings. Building on this, Zhao et al. (2020) explored different alignment targets and debias embeddings by equalising the distance between target words and protected attribute sets. Liang et al. (2020b) proposed a cross-lingual approach that manipulates distances between gender groups - maximizing intra-group distances while minimizing inter-group distances. In evaluating multilingual debiasing methods, Reusens et al. (2023) found SentenceDebias to be most effective for cross-lingual debiasing in mBERT when tested on CrowS-Pairs. However, Gonen et al. (2022) revealed that learning bias patterns from a single language is insufficient, as it fails to capture language-specific bias components in target languages.

The problem of multilingual debiasing requires a comprehensive approach. Vashishtha et al. (2023) observed limited transferability of debiasing effects across languages, especially from English to languages lacking Western context. The key objective is therefore to identify and neutralize shared bias components across languages while accounting for language-specific manifestations. Gonen et al. (2022) showed that such shared components exist and enable cross-lingual transfer of gender identification, suggesting that targeting these shared components could enable effective multilingual debiasing.

5 Conclusion

We have explored the challenges of debiasing multilingual representations by identifying and neutralizing joint linear bias subspaces across languages. Our proposed method, MUSAL, iteratively identifies and removes bias patterns using different language subsets, achieving more comprehensive bias mitigation than existing approaches. Through extensive experiments across eight languages and five demographic attributes, we demonstrate MUSAL's effectiveness in reducing bias while preserving model utility in state-of-the-art language models. Our key innovation lies in leveraging bias patterns from multiple languages simultaneously, enabling effective zero-shot debiasing where target language data is unavailable but linguistically similar languages can be used. MUSAL's framework is complementary to existing post-hoc debiasing methods, offering potential for adaptation to other multilingual debiasing approaches.

Limitations

First, our evaluation's reliance on binary demographic attributes (male/female, white/non-white, young/old) risks reinforcing harmful stereotypes and overlooking groups outside these binary classifications.

Second, using profile image inference APIs for demographic attributes and machine-translated datasets may misrepresent users' identities and obscure culture-specific bias patterns.

Third, our focus on European languages and post-processing debiasing limits wider applicability. Future work should explore non-Western languages (Vashishtha et al., 2023; Liang et al., 2020b) and other debiasing approaches across the NLP pipeline, including fine-tuning, alignment, and model editing (He et al., 2022).

Ethical Considerations

While MUSAL represents a breakthrough in multilingual debiasing, its deployment in real-world applications requires careful validation. Bias in language models is complex and context-dependent, and our approach, like any debiasing method, may have unintended consequences across different languages and cultural contexts. We emphasize that our results should not be a definitive solution or substitute for thorough evaluation in practical settings. Additionally, MUSAL should not serve as a superficial fix or "fig leaf" to mask deeper biases in AI systems. Responsible deployment necessitates continuous assessment, transparency, and engagement with affected communities.

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: perfect linear concept erasure in closed form. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Geert Booij. 2010. Construction morphology. *Language and linguistics compass*, 4(7):543–555.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *arXiv preprint arXiv:2404.01349*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

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.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits,

Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr

Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data.

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

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, 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.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020a. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

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

AI Meta. 2024a. Introducing llama 3.1: Our most capable models to date. *Meta AI Blog*, 12.

AI Meta. 2024b. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

Mistral AI. 2024. Mistral NeMo. Accessed: January 14, 2025.

Eric Mitchell, Peter Henderson, Christopher D Manning, Dan Jurafsky, and Chelsea Finn. 2022. Self-destructing models: Increasing the costs of harmful dual uses in foundation models. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.

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.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023a. Erasure of Unaligned Attributes from Neural Representations. *Transactions of the Association for Computational Linguistics*, 11:488–510.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023b. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia. 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.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada. Association for Computational Linguistics.

Hartger Veeman, Marc Allassonnière-Tang, Aleksandrs Berdicevskis, and Ali Basirat. 2020. Cross-lingual embeddings reveal universal and lineage-specific patterns in grammatical gender assignment. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 265–275, Online. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models.

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

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 585–596, Red Hook, NY, USA. Curran Associates Inc.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

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

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

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.

Language	Train Size	Test Size	# Professions	Gender Labels
English	295,044	98,379	28	Binary
Spanish	54,179	18,090	72	Binary
French	49,373	16,478	27	Binary

Table 5: Dataset statistics for multilingual BiasBios. Each sample contains a biography text paired with profession and gender labels. The main task is profession prediction, while gender information is used for bias evaluation through TPR-Gap.

A Multilingual BiasBios Details

Table 5 shows the data split for the multilingual BiasBios experiment across different languages and includes details about the protected attributes. Table 6 shows the complete results for eight different LLMs debiased by SAL, MUSAL ("Three-Subsets"), and MUSAL ("Two-Subsets-Without").

Target	Baseline		EN		De		FR		Two-Subsets-Without		Three-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased												
EN	80.5	15.4	↓0.1 80.4	↓1.9 13.5	80.5	↑0.5 15.9	80.5	↑0.4 15.8	80.5	↑0.4 15.8	↓0.1 80.4	↓1.9 13.5
DE	77.7	27.6	↑0.1 77.8	↓4.5 23.1	↓0.3 77.4	↓0.3 27.3	↑0.1 77.8	27.6	↑0.1 77.8	↓4.0 23.6	↓0.2 77.5	↓2.1 25.5
FR	72.7	22.8	↓0.1 72.6	↓0.8 22.0	72.7	↓0.7 22.1	↓0.5 72.2	↓3.4 19.4	↑0.2 72.9	↓0.4 22.4	↓0.6 72.1	↓3.3 19.5
Llama3-8B												
EN	81.1	12.6	↓1.8 79.3	↓0.2 12.4	↓0.8 80.3	↑0.2 12.8	↓0.6 80.5	↑0.2 12.8	↓0.9 80.2	↑0.3 12.9	↓1.9 79.2	↑0.2 12.8
DE	79.0	26.3	↓0.3 78.7	↓0.8 25.5	↓0.3 78.7	↑0.1 26.4	79.0	↓0.9 25.4	↓0.2 78.8	↓0.8 25.5	↓0.3 78.7	26.3
FR	72.7	25.7	↑0.2 72.9	↑1.0 26.7	↑0.1 72.8	↑0.5 26.2	↑0.2 72.9	↓4.0 21.7	72.7	↑0.4 26.1	↑0.2 72.9	↓4.2 21.5
Llama-3.1-8B												
EN	80.9	13.6	↓2.0 78.9	↓0.3 13.3	↓0.5 80.4	↓0.5 13.1	↓0.8 80.1	↓0.4 13.2	↓0.6 80.3	↓0.3 13.3	↓2.0 78.9	↓0.8 12.8
DE	79.8	26.8	↓0.3 79.5	↑0.2 27.0	↓0.4 79.4	↓5.2 21.6	↓0.2 79.6	↑0.5 27.3	↓0.4 79.4	↑0.4 27.2	↓0.5 79.3	↓4.4 22.4
FR	72.4	25.2	↑0.1 72.5	↑0.2 25.4	72.4	↓0.7 24.5	72.4	↓5.9 19.3	↑0.1 72.5	↓0.9 24.3	72.4	↓4.5 20.7
Llama-3.2-3B												
EN	80.2	13.3	↓1.4 78.8	↓1.7 11.6	↓0.5 79.7	↓0.7 12.6	↓0.2 80.0	↓0.7 12.6	↓0.6 79.6	↓1.0 12.3	↓1.3 78.9	↓1.3 12.0
DE	78.2	27.9	↓0.1 78.1	↑0.1 28.0	↓0.2 78.0	↓0.7 27.2	78.2	27.9	↓0.2 78.0	↑0.2 28.1	↓0.3 77.9	↓0.5 27.4
FR	71.1	16.4	↓0.3 70.8	↑0.6 17.0	71.1	↓1.7 14.7	↓0.1 71.0	↓1.0 15.4	↑0.1 71.2	↓0.8 15.6	↓0.2 70.9	↓1.0 15.4
Mistral-7B-Instruct-v0.3												
EN	80.0	14.1	↓2.5 77.5	↓1.2 12.9	↓0.4 79.6	↓0.1 14.0	↓0.3 79.7	↓0.2 13.9	↓0.3 79.7	↑0.2 14.3	↓2.3 77.7	↓1.3 12.8
DE	77.3	23.3	↓0.2 77.1	↑0.3 23.6	77.3	23.3	↓0.2 77.1	↑0.6 23.9	↓0.4 76.9	↓0.8 22.5	↓0.3 77.0	↓0.2 23.1
FR	71.6	22.0	↑0.1 71.7	↓0.9 21.1	71.6	↓0.6 21.4	↑0.1 71.7	↓3.7 18.3	↑0.1 71.7	↓1.7 20.3	↑0.2 71.8	↓3.8 18.2
Mistral-7B-v0.3												
EN	80.2	13.3	↓2.6 77.6	↓0.6 12.7	↓0.4 79.8	↑0.9 14.2	↓0.2 80.0	13.3	↓0.3 79.9	↑0.5 13.8	↓2.4 77.8	↓0.4 12.9
DE	78.4	27.3	↑0.1 78.5	↑0.2 27.5	↓0.7 77.7	↓1.1 26.2	↓0.1 78.3	↑0.2 27.5	↑0.1 78.5	27.3	↓0.5 77.9	↓1.3 26.0
FR	72.1	22.7	↑0.2 72.3	↓1.1 21.6	↑0.1 72.2	↓0.3 22.4	72.1	↓3.2 19.5	72.1	↑0.1 22.8	72.1	↓3.3 19.4
Mistral-Nemo-Base-2407												
EN	82.5	12.8	↓3.8 78.7	↑0.2 13.0	↓0.7 81.8	↑0.1 12.9	↓1.3 81.2	↑0.1 12.9	↓1.3 81.2	↑0.1 12.9	↓3.5 79.0	↓0.1 12.7
DE	79.4	31.0	↑0.2 79.6	31.0	↑0.2 79.6	↓1.0 30.0	↑0.2 79.6	31.0	↑0.1 79.5	↓0.1 30.9	79.4	↓1.6 29.4
FR	74.0	22.4	↑0.1 74.1	↓0.3 22.1	↑0.3 74.3	↓0.6 21.8	↑1.0 75.0	↓0.3 22.1	↑0.3 74.3	↓0.4 22.0	↑0.9 74.9	↑0.5 22.9
Mistral-Nemo-Instruct-2407												
EN	81.6	12.7	↓2.9 78.7	↓0.9 11.8	81.6	↓0.5 12.2	↓0.6 81.0	↑0.4 13.1	↓0.7 80.9	12.7	↓3.4 78.2	↓1.2 11.5
DE	78.5	26.5	↓0.1 78.4	↓0.1 26.4	↑0.3 78.8	↓0.8 25.7	↓0.1 78.4	↓0.5 26.0	↓0.2 78.3	↓0.2 26.3	↑0.2 78.7	↓1.3 25.2
FR	72.5	20.7	↑0.2 72.7	↓2.1 18.6	↑0.2 72.7	↓1.2 19.5	↑1.0 73.5	↑2.2 22.9	↑0.2 72.7	↓1.3 19.4	↑0.8 73.3	↑2.1 22.8

Table 6: Gender debiasing performance evaluation on BiasBios

B Multilingual Hate Speech Details

The detailed statistics for the multilingual Hate-Speech Dataset are presented in Table 7 (training and test set sizes) and Table 8 (class distribution across subsets). Comprehensive debiasing results comparing SAL and MUSAL across five languages and eight language models are provided in the appendix for each demographic attribute: Age bias (Table 9), Country bias (Table 10), Gender bias (Table 11) and Race bias (Table 12)

B.1 Multilingual Hate Speech Dataset Summary

Language	Gender	Race	Age	Country
English (en)	31691/8746	31408/8646	31691/8746	36159/7373
Spanish (es)	1900/410	1900/410	1900/410	1956/439
Italian (it)	1605/418	1598/418	1605/418	2388/644
Polish (pl)	6806/1446	5649/1235	6806/1446	2155/471
Portuguese (pt)	816/163	816/163	816/163	757/197

Table 7: Training/Test Sizes for Different Languages and Demographic Attributes

		Train		Dev		Test		Total	
Lang	**Bias**	**C0**	**C1**	**C0**	**C1**	**C0**	**C1**	**C0**	**C1**
EN	Gender	13,017	18,674	3,640	3,134	3,791	4,955	20,448	26,763
	Age	13,467	16,635	3,488	2,780	2,922	5,465	19,877	24,880
	Race	19,475	11,933	3,578	3,123	3,844	4,802	26,897	19,858
	Country	13,719	22,440	3,400	5,021	2,393	4,980	19,512	32,441
IT	Gender	1,091	514	256	103	290	128	1,637	745
	Age	791	809	178	180	229	189	1,198	1,178
	Race	1,568	30	354	3	413	5	2,335	38
	Country	610	1,778	107	345	79	565	796	2,688
PL	Gender	3,552	3,254	787	674	716	730	5,055	4,658
	Age	1,300	4,349	281	918	276	959	1,857	6,226
	Race	4,030	1,619	860	339	879	356	5,769	2,314
	Country	15	2,140	3	486	3	468	21	3,094
PT	Gender	682	134	76	74	60	103	818	311
	Age	737	79	92	58	68	95	897	232
	Race	634	182	84	66	86	77	804	325
	Country	341	416	88	110	66	131	495	657
ES	Gender	997	903	210	197	251	159	1,458	1,259
	Age	923	977	197	210	239	171	1,359	1,358
	Race	1,027	873	228	179	236	174	1,491	1,226
	Country	1,334	622	277	159	236	203	1,847	984

Table 8: Distribution of samples across different languages and protected attributes. C0 and C1 represent the two classes for each protected attribute.

B.2 Multilingual Hate Speech Results

Target	Baseline		EN		ES		IT		PL		PT		Four-Subsets-Without		Five-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased																
EN	86.7	9.2	↑0.3 87.0	↑0.5 9.7	↓0.5 86.2	↓0.4 8.8	86.7	9.2	↓0.5 86.2	↓0.2 9.0	86.7	↓0.2 9.0	↓0.6 86.1	↓4.3 4.9	↓0.3 86.4	↓7.2 2.0
ES	63.7	12.9	↑0.4 64.1	↑2.1 15.0	↓0.3 63.4	↓0.5 12.4	↑0.2 63.9	↓0.3 12.6	↑0.2 63.9	↑0.3 13.2	↓1.0 62.7	↑0.3 13.2	↑0.2 63.9	↓9.8 3.1	63.7	↓3.6 9.3
IT	68.2	3.6	↓0.3 67.9	↑0.3 3.9	68.2	3.6	↓0.3 67.9	↑0.3 3.9	68.2	↓0.3 3.3	↑0.2 68.4	↑0.4 4.0	↓0.5 67.7	↑4.9 8.5	↓1.5 66.7	↓3.3 0.3
PL	91.3	8.8	91.3	↓0.7 8.1	↓0.1 91.2	↓1.3 7.5	91.3	↓0.7 8.1	↓0.9 90.4	↓1.9 6.9	91.3	↓0.7 8.1	↓0.1 91.2	↓8.8 0.0	↓1.0 90.3	↓8.2 0.6
PT	61.3	17.6	↓1.2 60.1	↑1.2 18.8	↓1.2 60.1	↓0.6 17.0	↓1.2 60.1	↑1.2 18.8	61.3	17.6	↓0.6 60.7	↑1.5 19.1	↑0.7 62.0	↑5.8 23.4	↑0.7 62.0	↓5.7 11.9
Llama3-8B																
EN	79.6	7.7	↑0.8 80.4	↑0.6 8.3	↑0.1 79.7	↓0.2 7.5	↑0.1 79.7	7.7	↑0.2 79.8	7.7	↑0.3 79.9	↓0.2 7.5	↑0.2 79.8	↓7.5 0.2	↑1.1 80.7	↓3.6 4.1
ES	70.7	11.7	↑1.0 71.7	↑1.3 13.0	↓0.2 70.5	↓0.6 11.1	↑0.3 71.0	↑0.6 12.3	70.7	11.7	70.7	↑0.1 11.8	↓0.2 70.5	↑3.0 14.7	↓0.9 69.8	↓3.3 8.4
IT	69.9	8.0	↓0.3 69.6	↑0.3 8.3	↓0.3 69.6	↑0.3 8.3	↓0.5 69.4	↓0.7 7.3	↓0.3 69.6	↑0.3 8.3	↓0.3 69.6	↓0.3 7.7	↓0.8 69.1	↑1.0 9.0	69.9	↓0.6 7.4
PL	90.9	16.6	↑0.1 91.0	↓0.1 16.5	90.9	↓0.1 16.5	90.9	↓0.1 16.5	↓1.1 89.8	↓7.2 9.4	↑0.1 91.0	↓0.2 16.4	↑0.2 91.1	↓7.4 9.2	↓1.1 89.8	↓14.7 1.9
PT	57.1	2.1	↓0.7 56.4	↑0.9 3.0	↑1.2 58.3	↑1.9 4.0	↓0.7 56.4	↑0.9 3.0	57.1	2.1	57.1	2.1	57.1	↑3.2 5.3	↓1.9 55.2	↑11.1 13.2
Llama-3.1-8B																
EN	79.7	6.4	↑0.7 80.4	↑0.9 7.3	79.7	↓0.1 6.3	79.7	↑0.2 6.6	↑0.1 79.8	↑0.2 6.6	79.7	↓0.1 6.3	↑0.3 80.0	↓5.5 0.9	↑0.9 80.6	↓2.8 3.6
ES	72.9	8.6	↓0.5 72.4	↓1.5 7.1	↓0.2 72.7	↓1.0 7.6	72.9	↓1.1 7.5	↓0.5 72.4	↓0.5 8.1	↓0.2 72.7	↓2.8 5.8	↑0.5 73.4	↓0.8 7.8	↑0.8 73.7	↑3.2 11.8
IT	66.5	9.5	66.5	↓0.8 8.7	↑0.7 67.2	↓1.4 8.1	↑0.7 67.2	↓0.5 9.0	↑1.0 67.5	↓0.2 9.3	↑0.5 67.0	↓0.1 9.4	↑0.5 67.0	↓1.2 8.3	↑0.5 67.0	↓0.2 9.3
PL	90.3	14.6	90.3	14.6	90.3	14.6	↑0.1 90.4	↑0.6 15.2	↓0.3 90.0	↓6.3 8.3	↑0.1 90.4	↑0.6 15.2	90.3	↓11.2 3.4	↓0.7 89.6	↓14.6 0.0
PT	59.5	2.8	↑0.6 60.1	↑0.5 3.3	↑0.6 60.1	↓1.4 1.4	59.5	2.8	59.5	2.8	↑0.6 60.1	↑0.5 3.3	↓0.6 58.9	↓0.8 2.0	↓0.6 58.9	↑6.1 8.9
Llama-3.2-3B																
EN	79.7	6.4	↑0.7 80.4	↑0.6 7.0	79.7	↑0.1 6.5	↑0.1 79.8	↓0.1 6.3	↑0.1 79.8	↓0.1 6.3	↑0.2 79.9	↑0.2 6.6	↑0.2 79.9	↑4.6 11.0	↑0.7 80.4	↓6.2 0.2
ES	68.3	12.1	↓0.3 68.0	↓1.2 10.9	68.3	↑1.0 13.1	68.3	12.1	↑0.2 68.5	↑0.9 13.0	↓0.7 67.6	↑0.6 12.7	↓0.5 67.8	↑0.7 12.8	68.3	↓1.3 10.8
IT	67.9	5.8	67.9	↓0.3 5.5	↑0.3 68.2	↑1.2 7.0	↓0.4 67.5	↓0.1 5.7	↓0.2 67.7	↑0.2 6.0	↓0.2 67.7	↑0.2 6.0	↓1.4 66.5	↓2.9 2.9	↓0.2 67.7	↑1.6 7.4
PL	90.7	17.2	90.7	17.2	90.7	↓0.7 16.5	90.7	↓0.7 16.5	↓0.7 90.0	↓1.5 15.7	↑0.2 90.9	↓0.1 17.1	90.7	↑20.4 37.6	↓1.1 89.6	↓1.4 15.8
PT	57.1	9.7	57.1	↑2.0 11.7	57.1	9.7	57.1	9.7	↓1.3 55.8	↑1.9 11.6	↓0.7 56.4	↑0.9 10.6	↑2.4 59.5	↑0.6 10.3	↓0.7 56.4	↑1.5 11.2
Mistral-7B-Instruct-v0.3																
EN	79.6	6.7	↑0.9 80.5	↑1.7 8.4	79.6	↑0.5 7.2	↑0.3 79.9	6.7	↓0.1 79.5	↑0.7 7.4	79.6	↑0.3 7.0	↑0.7 80.3	↑2.4 9.1	↑0.9 80.5	↓6.6 0.1
ES	64.9	6.7	↓1.0 63.9	↑0.3 7.0	↑0.5 65.4	↑2.1 8.8	↓0.5 64.4	↓0.3 6.4	↓0.8 64.1	↑1.0 7.7	↓0.8 64.1	↓1.0 5.7	↓0.8 64.1	↑0.2 6.9	↓3.2 61.7	↑2.6 9.3
IT	66.0	3.5	66.0	3.5	↓0.2 65.8	↓0.2 3.3	↑0.3 66.3	↓0.3 3.2	↓0.2 65.8	↓0.5 3.0	↓0.2 65.8	↓0.2 3.3	↓0.2 65.8	↓1.3 2.2	↓0.4 65.6	↓0.9 2.6
PL	91.0	19.5	↑0.2 91.2	19.5	↑0.3 91.3	↑1.2 20.7	↓0.1 90.9	↓0.6 18.9	↓0.8 90.2	↓3.8 15.7	↑0.2 91.2	↑1.3 20.8	↑0.1 91.1	↓11.3 8.2	↓0.6 90.4	↓19.4 0.1
PT	59.5	17.2	59.5	17.2	↓0.6 58.9	↓1.0 16.2	↓0.6 58.9	↓1.0 16.2	59.5	17.2	↓0.6 58.9	↑1.8 19.0	↓1.2 58.3	↓3.3 13.9	59.5	↓13.9 3.3
Mistral-7B-v0.3																
EN	79.8	7.1	↑0.6 80.4	↑0.4 7.5	↓0.1 79.7	↓0.1 7.0	79.8	↓0.4 6.7	79.8	7.1	↓0.2 79.6	↓0.3 6.8	↓0.3 79.5	↑3.0 10.1	↑0.5 80.3	↓5.5 1.6
ES	67.1	12.8	67.1	↓1.4 11.4	↓0.3 66.8	↓0.8 12.0	↓0.3 66.8	12.8	↓0.5 66.6	↑1.7 14.5	67.1	↓0.6 12.2	↑0.5 67.6	↑4.8 17.6	↓0.3 66.8	↑2.6 15.4
IT	65.8	5.6	65.8	↓0.1 5.5	↓0.2 65.6	↑0.5 6.1	↑0.5 66.3	↓1.0 4.6	65.8	↓0.1 5.5	65.8	↓0.8 4.8	65.8	↓2.5 3.1	↑0.7 66.5	↑1.5 7.1
PL	90.4	17.8	↑0.2 90.6	↓0.2 17.6	90.4	17.8	90.4	↓0.8 17.0	↓0.9 89.5	↓5.8 12.0	↓0.1 90.3	↑0.1 17.9	90.4	↓10.9 6.9	↓0.7 89.7	↑2.8 20.6
PT	57.7	12.4	57.7	12.4	57.7	12.4	57.7	12.4	57.7	12.4	57.7	12.4	↑0.6 58.3	↑3.6 16.0	↑0.6 58.3	↓4.4 8.0
Mistral-Nemo-Base-2407																
EN	78.9	6.2	↑0.4 79.3	↑1.3 7.5	↓0.4 78.5	↑0.4 6.6	↓0.4 78.5	↑0.4 6.6	↓0.2 78.7	↑0.2 6.4	↓0.3 78.6	6.2	↓0.5 78.4	↑0.8 7.0	↑0.5 79.4	↓3.0 3.2
ES	71.7	16.2	↓0.7 71.0	↑1.4 17.6	↑1.2 72.9	↑1.9 18.1	↑0.5 72.2	↑2.1 18.3	71.7	↑1.1 17.3	↑0.3 72.0	↓0.3 15.9	↓0.2 71.5	↓4.6 11.6	↑1.0 72.7	↑3.0 19.2
IT	64.8	5.8	↑0.5 65.3	↑1.1 6.9	↑0.3 65.1	↓1.4 4.4	↓0.4 64.4	5.8	↓0.2 64.6	↑0.5 6.3	↑0.5 65.3	↑0.1 5.9	↑0.3 65.1	↓2.6 3.2	↑1.0 65.8	↓4.9 0.9
PL	91.3	20.9	↓0.3 91.0	↓0.7 20.2	↓0.1 91.2	↓0.6 20.3	↓0.1 91.2	↓0.6 20.3	↓1.0 90.3	↓3.4 17.5	↓0.1 91.2	20.9	↓0.4 90.9	↓9.6 11.3	↓1.1 90.2	↓6.5 14.4
PT	63.2	8.8	63.2	8.8	↓0.6 62.6	↓1.6 7.2	63.2	↓3.0 5.8	↑0.6 63.8	↓1.4 7.4	63.2	8.8	63.2	↓2.2 6.6	63.2	↓0.9 7.9
Mistral-Nemo-Instruct-2407																
EN	78.9	6.0	↑0.6 79.5	↑1.0 7.0	↑0.2 79.1	↑0.2 6.2	↑0.1 79.0	↓0.2 5.8	↑0.2 79.1	6.0	↑0.2 79.1	↑0.5 6.5	78.9	↓5.9 0.1	↑0.7 79.6	↑6.3 12.3
ES	71.0	10.2	↓0.5 70.5	↑1.2 11.4	↓0.8 70.2	↑1.2 11.4	↓0.5 70.5	↑0.9 11.1	↓1.0 70.0	↓0.6 9.6	↓0.3 70.7	↑1.6 11.8	↓0.3 70.7	↑1.5 11.7	↓1.7 69.3	↓3.9 6.3
IT	65.3	8.0	↓0.2 65.1	↓0.6 7.4	↓0.2 65.1	↓0.1 7.9	65.3	↓0.3 7.7	65.3	8.0	65.3	8.0	↑1.0 66.3	↑0.5 8.5	↑0.3 65.6	↑1.3 9.3
PL	90.9	18.9	↓0.1 90.8	↑0.1 19.0	90.9	18.9	90.9	18.9	↓0.7 90.2	↓1.9 17.0	90.9	18.9	↓0.2 90.7	18.9	↓0.7 90.2	↓3.4 15.5
PT	65.0	2.4	↓0.6 64.4	↓0.9 1.5	↓0.6 64.4	↓0.9 1.5	↓1.2 63.8	↑0.9 3.3	↑0.6 65.6	↑4.3 6.7	↓0.6 64.4	↓0.9 1.5	65.0	↓1.0 1.4	65.0	↑5.2 7.6

Table 9: Age debiasing performance evaluation on HateSpeech.

Target	Baseline		EN		ES		IT		PL		PT		Four-Subsets-Without		Five-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased																
EN	82.3	6.7	↓0.1 82.2	↓0.8 5.9	82.3	6.7	82.3	↓0.1 6.6	82.3	↑0.1 6.8	↑0.1 82.4	↓0.1 6.6	↓0.1 82.2	↓4.0 2.7	↓0.4 81.9	↓6.6 0.1
ES	65.1	5.1	↓0.4 64.7	↑0.8 5.9	65.1	↓0.2 4.9	↓0.2 64.9	↑0.2 5.3	↓0.2 64.9	↓0.2 4.9	↑0.3 65.4	↑0.8 5.9	↓0.4 64.7	↑6.0 11.1	↑0.3 65.4	↑3.3 8.4
IT	71.0	1.7	↓0.2 70.8	↑0.4 2.1	71.0	↑0.8 2.5	↑0.1 71.1	↓0.4 1.3	↓0.2 70.8	↑0.4 2.1	↓0.2 70.8	↑0.4 2.1	↓0.7 70.3	↑2.0 3.7	71.0	↑9.7 11.4
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	64.5	5.1	64.5	↑1.5 6.6	↑1.0 65.5	5.1	64.5	5.1	↑1.0 65.5	↑0.5 5.6	↓0.5 64.0	↓4.3 0.8	↑2.0 66.5	↓2.7 2.4	↓0.5 64.0	↑1.5 6.6
Llama3-8B																
EN	77.1	6.0	77.1	↓1.0 5.0	77.1	↑0.4 6.4	77.1	6.0	77.1	6.0	↓0.1 77.0	↑0.3 6.3	↓0.1 77.0	↓4.8 1.2	↓0.2 76.9	↓1.3 4.7
ES	66.7	10.0	↑1.6 68.3	↓0.2 9.8	↑0.3 67.0	↑0.1 10.1	66.7	10.0	66.7	10.0	↑0.7 67.4	↓1.3 8.7	↑1.4 68.1	↓9.1 0.9	↑1.0 67.7	↑2.1 12.1
IT	70.5	12.4	↑0.5 71.0	↑0.4 12.8	↑0.5 71.0	↑0.8 13.2	↑0.3 70.8	↑0.2 12.6	↑0.2 70.7	↑0.3 12.7	↑0.6 71.1	↑0.7 13.1	↑0.5 71.0	↓4.2 8.2	↑1.2 71.7	↓9.4 3.0
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	67.5	2.9	↑0.5 68.0	↑0.5 3.4	↑1.0 68.5	↑1.5 4.4	↑1.0 68.5	↓0.2 2.7	↓0.5 67.0	↑0.1 3.0	↑0.5 68.0	↑2.5 5.4	↑1.0 68.5	↑1.0 3.9	67.5	↑2.6 5.5
Llama-3.1-8B																
EN	77.1	5.7	↓0.2 76.9	↓0.6 5.1	77.1	5.7	↓0.1 77.0	↑0.1 5.8	77.1	↑0.4 6.1	77.1	↓0.2 5.5	77.1	↓2.6 3.1	↓0.3 76.8	↑1.1 6.8
ES	70.4	22.2	↓0.5 69.9	↑2.8 25.0	↓0.2 70.2	↓0.6 21.6	↓0.5 69.9	22.2	70.4	22.2	↓0.5 69.9	↑0.4 22.6	↓0.7 69.7	↓17.7 4.5	70.4	↓5.9 16.3
IT	68.6	12.7	↑0.3 68.9	↑0.5 13.2	↑0.2 68.8	↓0.1 12.6	68.6	↓0.4 12.3	68.6	↓0.4 12.3	↑0.3 68.9	↓0.1 12.6	68.6	↓8.8 3.9	↓0.3 68.3	↓6.9 5.8
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	66.5	5.4	↑0.5 67.0	↓1.4 4.0	↑1.0 67.5	↓1.6 3.8	66.5	5.4	↑0.5 67.0	↓1.4 4.0	↑0.5 67.0	↓0.2 5.2	↑1.0 67.5	↑1.0 6.4	↑1.5 68.0	↓2.3 3.1
Llama-3.2-3B																
EN	76.5	4.8	↓0.4 76.1	↓0.6 4.2	76.5	↑0.3 5.1	76.5	↑0.2 5.0	↑0.1 76.6	4.8	76.5	↑0.5 5.3	76.5	↓4.6 0.2	↓0.2 76.3	↓0.3 4.5
ES	67.7	8.9	↓0.5 67.2	↑0.4 9.3	↓0.5 67.2	↓0.6 8.3	67.7	↓1.0 7.9	↑0.2 67.9	↑0.7 9.6	67.7	8.9	↑0.4 68.1	↓2.4 6.5	↓1.2 66.5	↓7.4 1.5
IT	66.1	15.6	66.1	↓0.9 14.7	↑0.2 66.3	↓1.0 14.6	↓0.1 66.0	↓0.4 15.2	↓0.1 66.0	↓0.9 14.7	↓0.1 66.0	↓1.3 14.3	↓0.6 65.5	↑5.9 21.5	↓0.4 65.7	↑5.9 21.5
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	67.0	5.5	↓1.0 66.0	↓0.9 4.6	↓2.0 65.0	↑1.3 6.8	↓1.0 66.0	↑2.2 7.7	↓0.5 66.5	↑1.3 6.8	↑0.5 67.5	↑0.7 6.2	67.0	↓2.7 2.8	67.0	↓3.9 1.6
Mistral-7B-Instruct-v0.3																
EN	75.2	6.1	↓0.2 75.0	↓1.2 4.9	75.2	↑0.1 6.2	75.2	↓0.1 6.0	↑0.1 75.3	↑0.3 6.4	75.2	↑0.2 6.3	75.2	↓5.8 0.3	↓0.2 75.0	↓5.1 1.0
ES	60.6	13.5	↑0.9 61.5	↑0.5 14.0	↑0.7 61.3	↓1.1 12.4	↑0.2 60.8	↓0.4 13.1	↑0.2 60.8	↑0.5 14.0	↑0.4 61.0	↓1.0 12.5	↑0.9 61.5	↓8.6 4.9	↑1.4 62.0	↓0.8 12.7
IT	68.9	7.7	68.9	↑0.4 8.1	68.9	7.7	68.9	7.7	↓0.3 68.6	↓0.3 7.4	↑0.7 69.6	↑0.1 7.8	↑0.4 69.3	↑4.5 12.2	↑0.2 69.1	↑10.8 18.5
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	60.9	0.7	60.9	0.7	60.9	↑1.7 2.4	↑0.5 61.4	↑2.3 3.0	↓0.5 60.4	↑0.7 1.4	↑0.5 61.4	↑1.4 2.1	↓1.0 59.9	↑3.8 4.5	60.9	↑0.1 0.8
Mistral-7B-v0.3																
EN	75.5	4.4	75.5	↓0.7 3.7	75.5	↓0.1 4.3	↑0.1 75.6	↑0.4 4.8	↑0.1 75.6	↑0.2 4.6	↓0.1 75.4	↓0.1 4.3	↑0.1 75.6	↓0.1 4.3	↓0.1 75.4	↓0.7 3.7
ES	64.5	2.7	↑0.4 64.9	↑0.7 3.4	64.5	↓0.6 2.1	64.5	2.7	64.5	2.7	↓0.3 64.2	↑0.6 3.3	↓0.5 64.0	↑0.6 3.3	↓0.3 64.2	↓0.5 2.2
IT	67.2	11.0	↓0.1 67.1	↓0.4 10.6	↑0.2 67.4	↓0.4 10.6	↓0.1 67.1	↓0.4 10.6	↓0.3 66.9	↓1.2 9.8	↑0.3 67.5	↓0.4 10.6	↑0.3 67.5	↑5.3 16.3	↑0.3 67.5	↑2.7 13.7
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	66.0	5.4	66.0	5.4	↓2.0 64.0	↓2.5 2.9	66.0	5.4	66.0	5.4	66.0	↓0.3 5.1	↓2.0 64.0	↓4.9 0.5	↓2.0 64.0	↑4.4 9.8
Mistral-Nemo-Base-2407																
EN	75.8	4.8	↓0.1 75.7	↓0.6 4.2	↑0.2 76.0	↓0.1 4.7	↓0.1 75.7	↓0.1 4.7	↑0.1 75.9	4.8	↑0.1 75.9	4.8	↑0.1 75.9	↓4.0 0.8	↑0.1 75.9	↓4.7 0.1
ES	66.1	10.4	↑0.4 66.5	↑1.6 12.0	↑0.4 66.5	↑0.8 11.2	↑0.6 66.7	↓1.3 9.1	↓0.3 65.8	↑0.2 10.6	↓0.3 65.8	↑0.5 10.9	↑0.4 66.5	↓7.1 3.3	↑1.6 67.7	↓0.1 10.3
IT	69.6	12.7	↑0.4 70.0	↑0.7 13.4	↑0.7 70.3	↑0.6 13.3	↑0.1 69.7	↑0.4 13.1	69.6	↓0.4 12.3	↑0.1 69.7	↓0.1 12.6	↑0.6 70.2	↓5.6 7.1	↓0.3 69.3	↓7.4 5.3
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0
PT	72.6	3.8	↑0.5 73.1	↑1.0 4.8	↓0.5 72.1	↓1.5 2.3	↑0.5 73.1	↓0.2 3.6	72.6	↓1.9 1.9	↓1.5 71.1	3.8	↓0.5 72.1	↑23.3 27.1	↓1.5 71.1	↑9.9 13.7
Mistral-Nemo-Instruct-2407																
EN	76.7	4.6	↓0.7 76.0	↓0.5 4.1	↑0.2 76.9	4.6	↑0.1 76.8	↓0.1 4.5	↑0.1 76.8	↑0.1 4.7	↑0.1 76.8	↓0.2 4.4	↓0.6 76.1	↑0.4 5.0	↓1.3 75.4	↓4.3 0.3
ES	70.4	10.2	↓0.2 70.2	↓2.1 8.1	↑0.2 70.6	↓3.2 7.0	↓0.7 69.7	10.2	↓0.2 70.2	↑0.3 10.5	↓0.9 69.5	↑1.1 11.3	↓0.9 69.5	↑1.6 11.8	↓0.7 69.7	↑0.2 10.4
IT	67.7	12.6	67.7	↓0.5 12.1	67.7	12.6	↑0.3 68.0	↓0.6 12.0	↑0.2 67.9	↑0.4 13.0	↑0.2 67.9	↓0.1 12.5	↑0.3 68.0	↓9.6 3.0	↑1.4 69.1	↓11.6 1.0
PL	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	0.0	99.6	↑0.4 0.4
PT	70.1	9.8	70.1	9.8	↓0.6 69.5	↑3.9 13.7	70.1	↑2.8 12.6	70.1	↑2.8 12.6	70.1	↑3.4 13.2	↑0.5 70.6	↓0.1 9.7	↓1.1 69.0	↑11.5 21.3

Table 10: Country debiasing performance evaluation on HateSpeech.

Target	Baseline		EN		ES		IT		PL		PT		Four-Subsets-Without		Five-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased																
EN	86.7	4.4	86.7	↑0.7 5.1	86.7	4.4	86.7	4.4	↑0.1 86.8	↓0.1 4.3	86.7	4.4	86.7	↑1.0 5.4	↑0.2 86.9	↓4.4 0.0
ES	63.7	3.6	↑0.2 63.9	↑1.0 4.6	↑0.4 64.1	↓0.2 3.4	63.7	3.6	63.7	3.6	↓0.5 63.2	↓0.9 2.7	↑0.4 64.1	↑1.9 5.5	↑0.9 64.6	↓1.0 2.6
IT	68.4	2.1	↓0.5 67.9	↓1.4 0.7	↓0.2 68.2	↓0.1 2.0	68.4	↓0.6 1.5	↓0.5 67.9	↓0.1 2.0	68.4	2.1	68.4	↑1.3 3.4	↑0.7 69.1	↑6.1 8.2
PL	88.2	11.6	↓0.1 88.1	11.6	↑0.2 88.4	11.6	88.2	↓0.4 11.2	↓0.2 88.0	↓8.8 2.8	88.2	11.6	↑0.2 88.4	↓10.0 1.6	↓0.6 87.6	↓11.6 0.0
PT	61.3	12.0	↓0.6 60.7	↑0.7 12.7	↓1.2 60.1	↑2.4 14.4	↓0.6 60.7	↑1.2 13.2	↑0.7 62.0	↓1.2 10.8	↑0.7 62.0	↑1.4 13.4	↑0.7 62.0	↑0.7 12.7	↓0.6 60.7	↓3.0 9.0
Llama3-8B																
EN	79.4	4.0	↑0.2 79.6	↑0.4 4.4	↓0.1 79.3	↑0.1 4.1	79.4	4.0	↓0.1 79.3	↓0.1 3.9	↓0.1 79.3	↑0.2 4.2	79.4	↑1.4 5.4	↑0.5 79.9	↓0.1 3.9
ES	70.7	5.5	↑0.3 71.0	↑1.0 6.5	↑0.3 71.0	↓1.2 4.3	↑0.5 71.2	↑0.9 6.4	70.7	↓1.1 4.4	↑0.3 71.0	↑1.0 6.5	↑0.5 71.2	↑3.4 8.9	↓0.2 70.5	↓0.6 4.9
IT	69.4	3.7	↑0.2 69.6	3.7	↑0.2 69.6	↓0.8 2.9	↓0.7 68.7	↓1.6 2.1	↑0.5 69.9	↓0.8 2.9	↑0.5 69.9	↓0.8 2.9	↑0.2 69.6	3.7	69.4	↓1.6 2.1
PL	88.4	15.3	↓0.2 88.2	15.3	↓0.2 88.2	↑0.1 15.4	↓0.1 88.3	15.3	↓1.2 87.2	↓2.5 12.8	↓0.3 88.1	↓0.4 14.9	↓0.2 88.2	↑5.3 20.6	↓0.9 87.5	↓14.1 1.2
PT	57.1	2.5	57.1	2.5	↑1.2 58.3	↑1.5 4.0	↓0.7 56.4	↑1.9 4.4	57.1	2.5	↓0.7 56.4	↑0.2 2.7	↓1.3 55.8	↓1.0 1.5	↓3.7 53.4	↓0.5 2.0
Llama-3.1-8B																
EN	79.0	3.1	↑0.3 79.3	↑0.6 3.7	79.0	↑0.3 3.4	79.0	↑0.1 3.2	79.0	↑0.1 3.2	79.0	↓0.1 3.0	79.0	↑0.4 3.5	↑0.1 79.1	↓1.0 2.1
ES	72.9	1.4	72.9	↑2.4 3.8	↑0.3 73.2	↑0.2 1.6	↓0.7 72.2	↑2.1 3.5	72.9	↑0.8 2.2	↓1.2 71.7	↑0.9 2.3	72.9	↑0.6 2.0	72.9	↑2.0 3.4
IT	65.6	1.1	↑0.2 65.8	↑0.1 1.2	↑0.2 65.8	↓0.4 0.7	↑0.2 65.8	↑1.0 2.1	↑0.7 66.3	1.1	↓0.3 65.3	1.1	↓0.3 65.3	↑1.5 2.6	↑0.4 66.0	1.1
PL	87.0	11.6	↓0.1 86.9	11.6	87.0	11.6	87.0	↑0.4 12.0	↓0.3 86.7	↓5.4 6.2	↑0.1 87.1	↑0.4 12.0	↑0.2 87.2	↑4.1 15.7	↓0.3 86.7	↑15.2 26.8
PT	59.5	2.9	↑0.6 60.1	↓0.8 2.1	↑0.6 60.1	↓0.8 2.1	59.5	2.9	59.5	2.9	↓0.6 58.9	↓0.8 2.1	↓0.6 58.9	↓1.8 1.1	↓3.1 56.4	↓1.4 1.5
Llama-3.2-3B																
EN	79.5	3.3	↑0.1 79.6	↑0.5 3.8	79.5	↓0.1 3.2	79.5	3.3	79.5	3.3	↑0.1 79.6	↑0.1 3.4	↑0.1 79.6	↓0.6 2.7	↑0.2 79.7	↓0.6 2.7
ES	68.3	2.4	↓0.3 68.0	↓0.7 1.7	↓0.5 67.8	↓1.1 1.3	↓0.3 68.0	↑0.5 2.9	↓0.3 68.0	↑1.3 3.7	↓0.5 67.8	↑1.2 3.6	↓0.3 68.0	↓1.3 1.1	↑0.5 68.8	↓0.9 1.5
IT	68.2	4.3	↓1.0 67.2	↓0.7 3.6	↓0.5 67.7	↓0.8 3.5	↓1.5 66.7	4.3	↓0.3 67.9	↓0.8 3.5	68.2	4.3	68.2	↑1.4 5.7	↓2.2 66.0	↓1.5 2.8
PL	87.8	8.2	↑0.3 88.1	↑5.8 14.0	↑0.4 88.2	↑6.2 14.4	↑0.3 88.1	↑0.5 8.7	↓0.8 87.0	↑1.2 9.4	↑0.1 87.9	↑5.8 14.0	87.8	↑9.8 18.0	↓0.9 86.9	↑12.7 20.9
PT	56.4	11.1	↑0.7 57.1	↓0.5 10.6	↑0.7 57.1	↓0.5 10.6	↑1.3 57.7	↑0.8 11.9	↓1.2 55.2	↓0.2 10.9	↑1.9 58.3	↓2.9 8.2	56.4	↑0.8 11.9	↑2.5 58.9	↓0.7 10.4
Mistral-7B-Instruct-v0.3																
EN	79.9	2.5	↑0.1 80.0	↑1.1 3.6	79.9	2.5	↑0.1 80.0	↑0.3 2.8	79.9	2.5	↑0.1 80.0	↑0.2 2.7	79.9	↑2.4 4.9	↓0.2 79.7	↓1.8 0.7
ES	64.6	5.1	↓0.5 64.1	↑0.7 5.8	64.6	5.1	↓0.7 63.9	↓0.2 4.9	↓0.2 64.4	↑0.5 5.6	64.6	↑0.2 5.3	↓0.5 64.1	↓3.3 1.8	↓1.2 63.4	↓4.8 0.3
IT	66.3	2.5	↓0.3 66.0	↑0.8 3.3	↑0.2 66.5	↑0.2 2.7	↓1.0 65.3	↑1.7 4.2	↑0.2 66.5	↓0.7 1.8	66.3	↑3.2 5.7	↑0.4 66.7	↑3.1 5.6	66.3	↑2.2 4.7
PL	87.5	15.8	↑0.1 87.6	15.8	87.5	15.8	87.5	15.8	↑0.1 87.6	↓0.1 15.7	87.5	↓0.4 15.4	↓0.2 87.3	↑4.6 20.4	↓0.1 87.4	↑3.4 19.2
PT	59.5	10.2	59.5	10.2	59.5	10.2	59.5	10.2	59.5	10.2	59.5	10.2	59.5	↑4.9 15.1	↓0.6 58.9	↑3.4 13.6
Mistral-7B-v0.3																
EN	79.6	3.6	↑0.4 80.0	↑1.0 4.6	↓0.1 79.5	↓0.2 3.4	79.6	3.6	↓0.1 79.5	↓0.2 3.4	↑0.2 79.8	↑0.1 3.7	↓0.1 79.5	↓2.2 1.4	↑0.1 79.7	3.6
ES	67.1	6.8	↓0.3 66.8	↑2.3 9.1	↓0.3 66.8	↑2.4 9.2	↓0.3 66.8	↑0.7 7.5	↓0.8 66.3	↑0.6 7.4	↓1.0 66.1	↑0.6 7.4	↑0.5 67.6	↓2.9 3.9	↑0.5 67.6	↓3.1 3.7
IT	65.6	5.7	65.6	5.7	65.6	5.7	↑0.7 66.3	5.7	65.6	5.7	↓0.8 64.8	↓0.7 5.0	↑1.1 66.7	↑0.3 6.0	↑0.9 66.5	↓1.2 4.5
PL	87.6	12.0	87.6	↓0.4 11.6	87.6	↓0.4 11.6	↑0.1 87.7	12.0	↓0.7 86.9	↑0.2 12.2	87.6	↓0.1 11.9	↓0.1 87.5	↓10.8 1.2	↓0.6 87.0	↑5.1 17.1
PT	57.7	8.3	57.7	8.3	57.7	8.3	57.7	8.3	57.7	8.3	↑0.6 58.3	↓0.3 8.0	↑1.2 58.9	↓6.5 1.8	↑1.2 58.9	↓4.0 4.3
Mistral-Nemo-Base-2407																
EN	79.0	3.9	↓0.6 78.4	↑0.2 4.1	79.0	↓0.2 3.7	↑0.1 79.1	↓0.5 3.4	79.0	↓0.5 3.4	79.0	↓0.3 3.6	↓0.3 78.7	↓2.2 1.7	↓0.8 78.2	↓2.4 1.5
ES	72.2	3.1	↓1.2 71.0	↑1.6 4.7	↑0.5 72.7	↑1.0 4.1	72.2	↑2.0 5.1	↓0.2 72.0	↑1.5 4.6	↓0.7 71.5	↑1.7 4.8	↑0.2 72.4	↓2.3 0.8	↑0.5 72.7	↓1.8 1.3
IT	65.8	1.6	65.8	1.6	65.8	↑2.3 3.9	↓0.2 65.6	↑3.4 5.0	↓0.2 65.6	↑0.7 2.3	65.8	1.6	↑0.2 66.0	↑3.7 5.3	↑0.9 66.7	↑1.4 3.0
PL	88.5	16.1	↑0.1 88.6	↑0.8 16.9	88.5	↑0.4 16.5	↑0.1 88.6	↑0.4 16.5	↓0.3 88.2	↓1.3 14.8	88.5	↑0.4 16.5	↑0.2 88.7	↑4.1 20.2	↓0.1 88.4	↑7.6 23.7
PT	63.2	14.0	↓0.6 62.6	↓1.3 12.7	↓0.6 62.6	↓1.3 12.7	↓0.6 62.6	↓1.3 12.7	↑0.6 63.8	↓1.0 13.0	63.2	↓2.1 11.9	63.2	↓8.4 5.6	63.2	↓11.4 2.6
Mistral-Nemo-Instruct-2407																
EN	79.5	2.9	↓0.2 79.3	↑0.7 3.6	↓0.1 79.4	↓0.1 2.8	↓0.2 79.3	↑0.3 3.2	↑0.2 79.7	↑0.2 3.1	↓0.2 79.3	↑0.3 3.2	↓0.1 79.4	↑0.2 3.1	↓0.2 79.3	↓2.7 0.2
ES	71.0	5.4	71.0	↑0.9 6.3	71.0	↑1.3 6.7	↓0.3 70.7	5.4	71.0	↑1.3 6.7	↓0.3 70.7	↓0.2 5.2	↓0.3 70.7	↑3.1 8.5	↓0.5 70.5	↓4.2 1.2
IT	65.8	4.9	↑0.2 66.0	↑0.4 5.3	65.8	4.9	↑0.2 66.0	↑1.0 5.9	65.8	4.9	↑0.5 66.3	↑0.7 5.6	↑0.2 66.0	↓2.1 2.8	65.8	↑0.1 5.0
PL	87.6	10.7	↑0.2 87.8	↑0.1 10.8	↑0.1 87.7	10.7	↑0.1 87.7	10.7	↓0.4 87.2	↑3.7 14.4	↑0.2 87.8	10.7	↑0.4 88.0	↑1.7 12.4	87.6	↓9.2 1.5
PT	64.4	16.8	64.4	↓1.2 15.6	64.4	16.8	64.4	16.8	↑0.6 65.0	↓1.0 15.8	↓0.6 63.8	↑0.1 16.9	64.4	↓0.7 16.1	↓1.8 62.6	↑5.4 22.2

Table 11: Gender debiasing performance evaluation on HateSpeech.

Target	Baseline		EN		ES		IT		PL		PT		Four-Subsets-Without		Five-Subsets	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Mbert-uncased																
EN	86.8	4.1	↓1.7 85.1	↓1.3 2.8	86.8	4.1	86.8	4.1	86.8	4.1	↓0.1 86.7	↑0.3 4.4	↓0.2 86.6	↓2.0 2.1	↓1.8 85.0	↓4.0 0.1
ES	63.7	10.2	↑0.7 64.4	↓0.8 9.4	↑0.4 64.1	↓0.1 10.1	↑0.9 64.6	↑0.1 10.3	↑0.2 63.9	10.2	63.7	10.2	↑1.2 64.9	↓9.6 0.6	↑1.2 64.9	↓6.5 3.7
IT	68.4	32.6	↓0.2 68.2	↓0.1 32.5	↓0.2 68.2	↓0.6 32.0	↓0.9 67.5	↓1.6 31.0	68.4	32.6	↓0.2 68.2	↓0.6 32.0	↓0.9 67.5	↑25.9 58.5	↓1.4 67.0	↓31.0 1.6
PL	91.3	6.2	91.3	6.2	↑0.1 91.4	6.2	91.3	↓0.6 5.6	91.3	↓1.2 5.0	91.3	↓0.6 5.6	↑0.2 91.5	↓5.6 0.6	↓0.3 91.0	↑0.6 6.8
PT	61.3	1.0	↑0.7 62.0	↑2.9 3.9	↓0.6 60.7	↑0.1 1.1	↓1.2 60.1	↑0.2 1.2	↓1.8 59.5	↑1.5 2.5	↓0.6 60.7	↑0.1 1.1	61.3	↑1.2 2.2	↑0.7 62.0	↑10.6 11.6
Llama3-8B																
EN	79.3	4.4	↓1.7 77.6	↓1.6 2.8	↑0.1 79.4	↑0.1 4.5	79.3	↑0.2 4.6	79.3	↑0.2 4.6	79.3	4.4	79.3	↓3.8 0.6	↓1.2 78.1	↓4.2 0.2
ES	70.7	7.0	↑0.8 71.5	↑0.9 7.9	70.7	7.0	↑0.8 71.5	↑1.8 8.8	↑0.3 71.0	↓0.1 6.9	↑0.3 71.0	↑0.9 7.9	↑0.5 71.2	↑1.5 8.5	↑1.3 72.0	↓0.7 6.3
IT	69.9	43.3	↓0.3 69.6	↑0.1 43.4	↓0.3 69.6	↑0.1 43.4	↓1.5 68.4	↓1.7 41.6	↓0.3 69.6	↓0.5 42.8	↓0.3 69.6	↑0.7 44.0	↓0.3 69.6	↓17.1 26.2	↓0.8 69.1	↑1.1 44.4
PL	91.0	16.3	91.0	↓0.6 15.7	91.0	↓0.6 15.7	↓0.1 90.9	↓0.6 15.7	91.0	↓0.6 15.7	91.0	↓0.6 15.7	↓0.1 90.9	↓6.4 9.9	↑0.2 91.2	↓3.5 12.8
PT	57.1	18.0	57.1	18.0	57.1	18.0	↓0.7 56.4	↑1.3 19.3	57.1	18.0	57.1	18.0	↓1.9 55.2	↓2.4 15.6	↓0.7 56.4	↓4.2 13.8
Llama-3.1-8B																
EN	79.0	3.1	↓1.8 77.2	↓2.2 0.9	79.0	↑0.1 3.2	↑0.1 79.1	↓0.1 3.0	79.0	↓0.1 3.0	79.0	3.1	↑0.1 79.1	↓3.0 0.1	↓1.6 77.4	↓1.0 2.1
ES	72.9	1.9	↓0.2 72.7	↓0.2 1.7	↓0.5 72.4	1.9	↓0.2 72.7	↓0.4 1.5	↓0.2 72.7	↑0.1 2.0	↓1.2 71.7	↓0.1 1.8	↓0.5 72.4	↑2.0 3.9	↑0.3 73.2	↑2.9 4.8
IT	66.7	33.0	66.7	33.0	↑0.5 67.2	↑0.5 33.5	↑0.8 67.5	↑1.6 34.6	↑0.5 67.2	↑0.5 33.5	↑0.5 67.2	↑0.5 33.5	↑1.2 67.9	↑2.8 35.8	↑0.5 67.2	↑4.8 37.8
PL	90.4	14.5	↓0.1 90.3	↓0.6 13.9	90.4	↓0.6 13.9	↓0.1 90.3	↓0.6 13.9	↓0.1 90.3	↓0.6 13.9	↑0.3 90.7	↑1.2 15.7	↓0.1 90.3	↑1.8 16.3	↓0.1 90.3	↓8.0 6.5
PT	59.5	12.1	↑0.6 60.1	↓1.2 10.9	↑1.2 60.7	↓2.3 9.8	59.5	12.1	59.5	12.1	↑0.6 60.1	↓1.2 10.9	↑0.6 60.1	↓3.1 9.0	59.5	↓1.3 10.8
Llama-3.2-3B																
EN	79.5	3.6	↓2.0 77.5	↓2.9 0.7	↓0.1 79.4	↑0.3 3.9	79.5	↑0.2 3.8	79.5	↑0.1 3.7	↑0.1 79.6	3.6	↓0.1 79.4	↓3.2 0.4	↓2.0 77.5	↓1.8 1.8
ES	68.3	3.6	68.3	3.6	↓0.3 68.0	↑0.2 3.8	↓1.0 67.3	↑0.8 4.4	↑0.2 68.5	↑1.1 4.7	↓0.3 68.0	↑0.6 4.2	↑0.2 68.5	↑3.8 7.4	↓0.5 67.8	↑1.4 5.0
IT	68.9	38.3	↓1.2 67.7	↓1.6 36.7	↓0.5 68.4	↓1.1 37.2	↓1.7 67.2	↓1.1 37.2	↓0.5 68.4	↓0.5 37.8	↓0.7 68.2	↓0.6 37.7	↓1.9 67.0	↓6.1 32.2	↓1.4 67.5	↓5.8 32.5
PL	90.9	16.3	90.9	16.3	↓0.2 90.7	↓0.6 15.7	↓0.3 90.6	↓0.6 15.7	90.9	↑0.6 16.9	90.9	16.3	↓0.5 90.4	↑21.8 38.1	↓0.1 90.8	↓6.0 10.3
PT	56.4	5.4	↑0.7 57.1	↑1.3 6.7	↑0.7 57.1	↓4.1 1.3	56.4	5.4	56.4	↑1.0 6.4	↑0.7 57.1	↑0.9 6.3	↓0.6 55.8	↓1.5 3.9	↑0.7 57.1	↑1.7 7.1
Mistral-7B-Instruct-v0.3																
EN	79.9	2.8	↓3.1 76.8	↓0.8 2.0	79.9	↑0.5 3.3	↓0.1 79.8	↑0.1 2.9	↓0.1 79.8	↑0.1 2.9	↓0.1 79.8	↑0.1 2.9	↓0.4 79.5	↑1.4 4.2	↓3.2 76.7	↑1.7 4.5
ES	64.6	8.0	↓0.9 63.7	↓0.9 7.1	64.6	8.0	↓0.2 64.4	↑1.7 9.7	↑0.5 65.1	↑1.4 9.4	↓0.5 64.1	↓0.2 7.8	64.6	↓2.9 5.1	↓1.2 63.4	↓0.3 7.7
IT	66.0	39.4	↑0.3 66.3	↑0.1 39.5	↑0.3 66.3	↑0.1 39.5	↓1.4 64.6	↓2.1 37.3	↓0.2 65.8	↓0.4 39.0	↓0.2 65.8	↓0.4 39.0	↑0.3 66.3	↓8.0 31.4	↓0.2 65.8	↑10.2 49.6
PL	91.0	18.7	↑0.3 91.3	18.7	↑0.2 91.2	↑1.2 19.9	↓0.1 90.9	18.7	91.0	↓0.6 18.1	↑0.3 91.3	↑0.6 19.3	91.0	↓7.8 10.9	↓0.1 90.9	↑8.5 27.2
PT	59.5	8.8	59.5	8.8	59.5	8.8	↓0.6 58.9	↑1.0 9.8	↓0.6 58.9	↑1.0 9.8	59.5	8.8	↓0.6 58.9	↓2.4 6.4	↓2.4 57.1	↓3.3 5.5
Mistral-7B-v0.3																
EN	79.5	4.9	↓1.6 77.9	↓1.8 3.1	79.5	↓0.2 4.7	79.5	4.9	↑0.1 79.6	↓0.1 4.8	79.5	4.9	↓0.2 79.3	↓3.7 1.2	↓1.7 77.8	↓1.4 3.5
ES	67.1	3.5	↓0.3 66.8	↓0.2 3.3	67.1	3.5	↓0.3 66.8	↑0.3 3.8	↓0.8 66.3	↑0.5 4.0	↓0.5 66.6	↓1.2 2.3	↓1.2 65.9	↑1.0 4.5	↑0.7 67.8	↑3.6 7.1
IT	66.3	41.1	↓0.3 66.0	↓0.5 40.6	66.3	↓0.4 40.7	66.3	↓1.2 39.9	↓0.3 66.0	↓0.1 41.0	66.3	↓0.4 40.7	66.3	↓8.1 33.0	↓0.3 66.0	↓27.5 13.6
PL	90.4	16.9	↑0.1 90.5	16.9	↓0.2 90.2	↓0.6 16.3	↓0.2 90.2	↓0.6 16.3	↑0.1 90.5	↑0.6 17.5	↓0.1 90.3	↓0.6 16.3	90.4	↑6.7 23.6	↑0.4 90.8	↑17.1 34.0
PT	57.7	2.9	57.7	2.9	↑0.6 58.3	↑2.2 5.1	57.7	2.9	↑0.6 58.3	↑2.2 5.1	57.7	2.9	↑0.6 58.3	↓0.2 2.7	↑0.6 58.3	↑0.2 3.1
Mistral-Nemo-Base-2407																
EN	79.3	5.3	↓3.3 76.0	↓1.5 3.8	↓0.9 78.4	↑0.5 5.8	↓0.5 78.8	5.3	↓1.2 78.1	↑0.6 5.9	↓1.1 78.2	↑0.5 5.8	↓1.2 78.1	↓3.6 1.7	↓3.4 75.9	↓3.6 1.7
ES	72.2	8.9	↓0.7 71.5	↓1.3 7.6	72.2	↓2.1 6.8	↓0.2 72.0	↓0.9 8.0	↑0.2 72.4	↓0.1 8.8	↓1.0 71.2	↓0.3 8.6	↓1.2 71.0	↓0.8 8.1	↓1.0 71.2	↓8.4 0.5
IT	65.1	40.6	↑0.9 66.0	↑1.0 41.6	↑0.5 65.6	↑0.5 41.1	↓0.5 64.6	↑0.1 40.7	65.1	40.6	↑0.5 65.6	↑0.5 41.1	↑0.5 65.6	↑3.4 44.0	↑0.5 65.6	↓26.0 14.6
PL	91.3	19.9	91.3	19.9	↓0.1 91.2	↓0.6 19.3	↓0.1 91.2	↓0.6 19.3	↓0.4 90.9	↓1.8 18.1	91.3	19.9	↓0.2 91.1	↑6.7 26.6	↓0.1 91.2	↓19.8 0.1
PT	63.2	7.6	↓0.6 62.6	↓1.1 6.5	↓0.6 62.6	↓1.1 6.5	↓0.6 62.6	↓1.1 6.5	63.2	7.6	↑0.6 63.8	↓1.3 6.3	↓0.6 62.6	↑0.1 7.7	63.2	↓3.8 3.8
Mistral-Nemo-Instruct-2407																
EN	79.5	5.3	↓2.7 76.8	↓3.4 1.9	↓0.1 79.4	↑0.4 5.7	↓0.2 79.3	↓0.1 5.2	↓0.2 79.3	↓0.2 5.1	79.5	5.3	↓0.1 79.4	↓1.9 3.4	↓2.7 76.8	↓1.4 3.9
ES	71.0	4.4	↓0.5 70.5	↑0.9 5.3	71.0	4.4	↓0.5 70.5	4.4	↓0.5 70.5	↓0.2 4.2	↓0.5 70.5	↓1.1 3.3	↓1.0 70.0	↑6.6 11.0	↓1.0 70.0	↑5.1 9.5
IT	65.3	44.8	↓0.5 64.8	↓0.3 44.5	65.3	↓1.2 43.6	↓0.7 64.6	↓2.0 42.8	65.3	44.8	65.3	44.8	65.3	↓25.9 18.9	65.3	↓26.4 18.4
PL	90.9	18.1	90.9	18.1	90.9	18.1	90.9	18.1	90.9	18.1	90.9	18.1	↓0.1 90.8	↓7.6 10.5	↑0.1 91.0	↑4.0 22.1
PT	64.4	3.9	↑0.6 65.0	↓0.7 3.2	64.4	↓2.4 1.5	64.4	↓2.4 1.5	↑1.9 66.3	↑3.4 7.3	64.4	3.9	↑0.6 65.0	↑9.2 13.1	↓0.6 63.8	↑12.6 16.5

Table 12: Race debiasing performance evaluation on HateSpeech.