Iterative Multilingual Spectral Attribute Erasure

Shun Shao[1] **Yftah Ziser**[3,4] **Zheng Zhao**[2] **Yifu Qiu**[2]
Shay B. Cohen[2] **Anna Korhonen**[1]
[1]University of Cambridge [2]University of Edinburgh
[3]NVIDIA Research [4]University of Groningen
ss3047@cam.ac.uk yziser@nvidia.com zheng.zhao@ed.ac.uk
yifu.qiu@ed.ac.uk scohen@inf.ed.ac.uk alk23@cam.ac.uk

Abstract

Multilingual representations embed words with similar meanings to share a common semantic space across languages, creating opportunities to transfer debiasing effects between languages. However, existing methods for debiasing are unable to exploit this opportunity because they operate on individual languages. We present Iterative Multilingual Spectral Attribute Erasure (IMSAE), which identifies and mitigates joint bias subspaces across multiple languages through iterative SVD-based truncation. Evaluating IMSAE 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 that IMSAE outperforms traditional monolingual and cross-lingual approaches while maintaining model utility. [1]

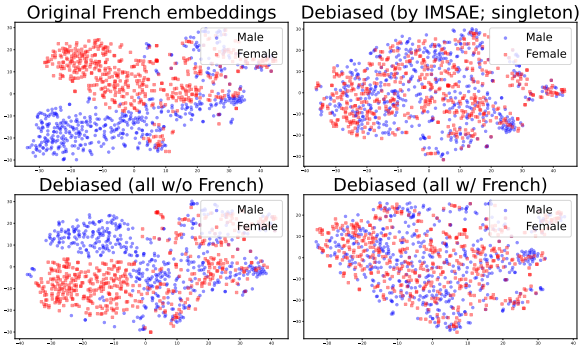

Figure 1: Visualization of gender bias in mBERT French embeddings using t-SNE (blue, male; red, female). Top left: Original embeddings showing clear gender clustering. Top right and bottom: Results after applying IMSAE, demonstrating effective elimination of gender-based patterns through both monolingual (French-only) and cross-lingual (other languages) debiasing approaches. For additional LLM embedding visualizations across different datasets and models, see Figure 3.

1 Introduction

Large language models (LLMs) may make biased decisions or generate unwanted outputs at various 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. 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, because large-scale demographic labels are difficult to obtain, most fairness studies have focused 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., 2019b), 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 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).

While prior work has shown the existence of

[1]Code is available at https://github.com/jasonshaoshun/IMSAE.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29219–29244
November 4-9, 2025

joint gender subspaces across languages (Gonen et al., 2022), how to effectively leverage these subspaces for cross-lingual bias mitigation remains an open question. To address this gap, we propose Iterative Multilingual Spectral Attribute Erasure (IMSAE). IMSAE is a structural extension of the monolingual debiasing method by Shao et al. (2023b). IMSAE is specifically designed to address the issues of directly applying monolingual techniques in multilingual contexts (see §4 for a detailed discussion). IMSAE iteratively debiases subsets of source languages using singular value decomposition (SVD) truncation. These subsets may overlap, and at each step, a shared subspace capturing the guarded attribute across languages is identified and neutralized. Our approach can debias representations without direct access to target language data. IMSAE addresses this challenge in a principled manner by leveraging shared linguistic structures. This effectiveness is visualized in the t-SNE plots in Figure 1: while the original embeddings (top left) show clear gender clustering, applying IMSAE (remaining plots) successfully obscures these gender-based patterns. IMSAE reduces bias in French embeddings both with (bottom left) and without (bottom right) access to French data, demonstrating its ability to identify and mitigate bias patterns across languages.

In addition, to properly evaluate our approach, we introduce the Multilingual Stack Exchange Fairness (MSEFair) dataset, which offers two advantages: verified protected attributes and authentic Russian language usage rather than translations. This dataset can serve as a robust evaluation framework for cross-lingual debiasing techniques in linguistically diverse, non-Western linguistic contexts (Ramesh et al., 2023; Vashishtha et al., 2023).

We validate IMSAE's effectiveness across eight languages and five demographic dimensions using comprehensive multilingual fairness benchmarks, including our newly developed MSEFair dataset. Our evaluation compares IMSAE against three state-of-the-art post-hoc debiasing methods across diverse language families and multiple model architectures (Llama, Mistral, and BERT), demonstrating consistent performance improvements particularly in zero-shot scenarios.

2 Problem Formulation and Notation

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. We do not require $\mathcal{L}_s \cap \mathcal{L}_t = \emptyset$.

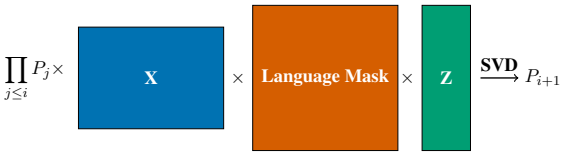

Figure 2: A visualization of IMSAE. A sequence of projections is created using SVD based on the input representations (r.v. $\mathbf{X}$), the guarded attributes (r.v. $\mathbf{Z}$) and a language mask that dictates which languages to use.

We assume 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. Recent work has demonstrated effective compression of words and definitions into shared multilingual spaces (Chen et al., 2024).

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 inspired by the SAL algorithm of Shao et al. (2023b), and adds 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 agree with the protected attribute.

3 The IMSAE Algorithm

Our algorithm, IMSAE, is based on the SAL algorithm presented by Shao et al. (2023b). Rather than relying on a single projection that is derived from the cross-covariance matrix between input representations and a guarded attribute and that removes information from monolingual input representations, the method creates a sequence of such projections, each corresponding to inputs from a predefined subset of languages.

Figure 4 shows the IMSAE algorithm. 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 overlapping subsets of $\mathcal{L}_s$, denoted $\mathcal{L}_1, \ldots, \mathcal{L}_m$. Each of these subsets determines one possible way in the sequence to jointly remove bias. The sequence defines a spectrum of

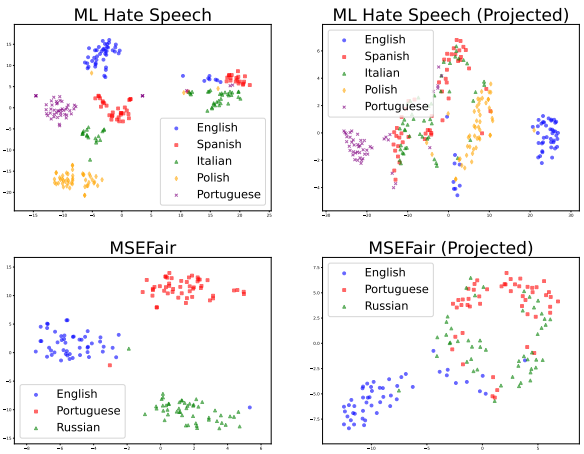

Figure 3: Visualization of Llama embeddings across the Multilingual Hate Speech and Multilingual SEFair datasets. "Projected" denotes the projection of the embeddings against the corresponding $\mathbf{Z}$ variable (using SVD). The projection shows that there is a natural mixture both originally and when projecting the representations against Z. Once we project X onto the direction of the protected attribute, the level of mixture increases, but it remains separable, validating the assumption about mixture of different languages.

how to group languages to erase information.

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

- Monolingual or cross-lingual (we assume $|\mathcal{L}_t| = 1$): where $m = 1$ and $|\mathcal{L}_1| = 1$. This means that we use one language (possibly different from the target language) to erase information.
- All subsets without 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 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 are in, which is important to consider in the execution of IMSAE, is left underspecified. More of this is discussed in §3.1. We recover (and therefore generalize through a structural extension) the SAL algorithm of Shao et al. (2023b) when $|\mathcal{L}| = 1$, $m = 1$ and $\mathcal{L}_s = \mathcal{L}_t = \mathcal{L}$.

Our algorithm has a matrix formulation, as an iterative projection algorithm with an SVD on a masked version of the cross-covariance matrix. The subsets of $\mathcal{L}$ essentially select examples from the entire set of examples in the pool, as depicted in Figure 2. See also similar discussion by Osborne et al. (2016).

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 4: The IMSAE algorithm. IMSAE targets attribute-specific bias through cross-covariance between representations $\mathbf{X}$ and protected attributes $\mathbf{Z}$, using SVD to remove only attribute-correlated directions.

In practice, we set $k = 2$ for all experiments since binary protected attributes (e.g., male/female) yield a rank-2 cross-covariance matrix, eliminating the need for hyperparameter tuning. All evaluation uses identical LogisticRegression settings (max_iter=1000) to ensure consistent comparison across methods.

3.1 Order Sensitivity in Sequential Projections

The final erasure matrix $\boldsymbol{P}^*$ is a product of projection matrices, and such matrix multiplication is not commutative in general. Therefore, the order in which these projections are applied matters. We consider two orderings: (a) Global-then-Local: First apply a projection using all languages $\mathcal{L}_s$ to identify and remove shared bias directions, followed by language-specific projections for each $\ell \in \mathcal{L}_s$; (b) Local-then-Global: the reverse.

The global-first 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.

4 Justification and the Fully Joint Baseline with Its Drawback

The following can be skimmed through in a first read. We proceed with providing an intuition for a justification of the IMSAE algorithm, especially when a target language is missing from training. Central to our analysis is a two-fold assumption:

Assumption 1. *There exists C **meta-linguistic** random vectors, $p(\overline{\mathbf{X}}_i)$ and a latent variable $\mathbf{H}$ such that for any language $\ell \in \mathcal{L}$ there exists constants $\zeta_{\ell,h,1}, \ldots, \zeta_{\ell,h,C}$ for any h value of $\mathbf{H}$ where:*

$$\mathbb{E}[\mathbf{X}_\ell \mid h] = \sum_i \zeta_{\ell,h,i} \mathbb{E}[\overline{\mathbf{X}}_i \mid h],$$

and $\mathbf{X}_\ell$ and $\mathbf{Z}$ are conditionally independent given $\mathbf{H}$.

For the condition on the latent variable $\mathbf{H}$, see Shao et al. (2023a). The reference to meta-linguistic variables implies that any representation of a specific language (at least in expectation) can be represented as a combination of representations of some core prototypical meta-linguistic vectors. Figure 3 demonstrates how various languages may cluster in the embedding space of two datasets (Hate Speech and MSEFair, discussed in §6.2 and §5), with several centroids. The mixture of centroids is even more noticeable when considering the projection of $\mathbf{X}$ against $\mathbf{Z}$ (Projected). We could have a stronger mixture condition on the probability distribution, but it is sufficient to require the expectation condition for our needs. We expect $C < \min\{|\mathcal{L}|, d\}$.

We also assume a set of coefficients over $\mathbf{X}$, the r.v. vector from all languages, such that $p(\mathbf{X} \mid h) = \sum_{h,\ell} \mu_{h,\ell} p(\mathbf{X}_\ell \mid h)$. This is a reasonable assumption, as all it implies is that for a specific subset $\mathcal{L}_j$, the resulting vector $\mathbf{X}$ on which we operate has a mixture distribution over the languages based on the relative frequency of each language in the data (Zhou et al., 2019a; Zhao et al., 2023). Such coefficient could be zero, for example, if a language ℓ does not appear in $\mathcal{L}_j$.[2]

The implication of Assumption 1 together with the mixture of languages requirement is that for an $\mathcal{L}_j$, the covariance $\mathbf{E}[\mathbf{X}\mathbf{Z}^\top]$ between a vector $\mathbf{X}$ from that distribution and $\mathbf{Z}$ can be rewritten as:

$$\sum_\ell \sum_h \sum_i \mu_{h,\ell} \zeta_{h,\ell,i} \mathbf{E}[\overline{\mathbf{X}}_i \mid h] \cdot \mathbf{E}[\mathbf{Z}^\top \mid h].$$

Combining the indices $(h, \ell) = k$ together, this means that for any $\mathcal{L}_j$, there exist coefficients $\eta_{k,i}$ such that:

$$\mathbf{E}[\mathbf{X}\mathbf{Z}^\top] = \sum_{k,i} \eta_{k,i} \mathbf{E}[\overline{\mathbf{X}}_i \mid h] \mathbb{E}[\mathbf{Z}^\top \mid h].$$

Each specific subset of languages, as presented by the algorithm in Figure 4, provides a different set of coefficients $\eta_{k,i}$. This is also true for the target language, even when missing from the data. Therefore, running the algorithm provides an erasure matrix that iteratively erases directions that correspond to different coefficients $\eta_{k,i}$. This means that the final matrix is more robust to the specific coefficient of the left-out target language, which is also based on some combination of $\eta_{k,i}$, hopefully targeted by one of $\mathcal{L}_j$.[3]

We note that while it may seem like we are identifying the intersection of the projections ranges, that is not the case, and intersecting them would require, for example, using Ben-Israel's algorithm or using repeated projection (Ben-Israel, 2015). We did not experiment with these approaches, though it might be worth considering that in future work.

The FullyJoint Baseline A straightforward baseline to erase information from representations across multiple languages is to concatenate all the input representations from the different languages and their corresponding guarded attributes and feed them to an erasure algorithm. For example, running IMSAE with $m = 1$, and $\mathcal{L}_1 = \mathcal{L}$. This one-shot reduction may be less effective than IMSAE in its full generality, which erases information iteratively on a per-language subset. We refer to this baseline as *FullyJoint*.

[2] Empirically, each $\mathcal{L}_j$ defines a different distribution over $\mathbf{X}$ because of the different frequencies of the languages.

[3] This leaves an idea for further exploration, where not only subsets of the languages are taken, but also subsets of different sizes of the data for these languages, yielding different coefficients.

We turn to explain a specific case of the analysis above with FullyJoint, illustrating why $m > 1$ is needed. If $\mathbf{Z}$ is essentially a scalar (represented possibly as a one-hot vector or otherwise), the rank of any of the covariance matrices between $\mathbf{X}$ and $\mathbf{Z}$ is 1. With the FullyJoint baseline, $\mathcal{L}_1$ corresponds to a specific value of $\eta_{k,i}$ and a single erasure direction $\mathbf{u}$. These quantities do not have to coincide with the values of using, for example, a singleton subset containing one language. Some residual association between the input representations of each language and the protected attributes may remain, allowing a linear classifier to predict the protected attribute. Thus, by erasing the attribute-related direction iteratively through different subsets (unlike FullyJoint), for example, including singletons, we ensure a more robust erasure.

5 Multilingual Stack Exchange Fairness Dataset

This section introduces MSEFair, a challenging dataset curated to support experimentation with non-English languages, which are often more difficult to transfer to.

5.1 Motivation

Previous debiasing research has focused on Anglocentric text, with limited attention to non-Western contexts (Ramesh et al., 2023). A key open question is whether it is possible to effectively debias across distant languages. While Vashishtha et al. (2023) extended DisCo (Webster et al., 2021) to Indian languages through human translation, this approach fails to capture culture-specific bias (Névéol et al., 2022). Moreover, existing studies have largely focused on specific tasks like sentiment analysis and profession prediction. To address these issues, we introduce the Multilingual Stack Exchange dataset. This dataset offers two key advantages: it contains verified protected attributes; and, it represents authentic Russian language use rather than translations, providing a more reliable testbed in non-Western contexts.

5.2 Data Collection for MSEFair

In addition to its English-language sites, the Stack Exchange platform also hosts localized versions of its most popular site, Stack Overflow in Russian and Portuguese. We curated posts (questions and answers) from the users of those websites. We use the user reputation as the protected attribute for predicting post helpfulness. We classify users in the top 1% of reputation as high-reputation users and those in the bottom 98% as low-reputation users.[4] As the Stack Exchange platform provides user reputation, we consider it a reliable protected attribute. For the helpfulness prediction, we classify posts with four or more upvotes as helpful and posts with zero upvotes as not helpful. We aim to classify posts as helpful or not (x's) regardless of their authors reputation (z's). Another challenge posed by this dataset is the correlation between the protected attribute and the downstream task label, as upvotes are a major factor in determining user reputation on Stack Exchange platforms. This correlation makes it difficult to remove information about the protected attribute without discarding information essential for the downstream task. The dataset statistics are provided in Table 1.

Name	Train	Val	Test	Total
English	4,058	2,029	14,205	20,292
Russian	2,370	1,186	8,300	11,856
Portuguese	1,670	835	5,847	8,352

Table 1: Statistics for the MSEFair datasets.

6 Experiments

We explore the effectiveness of IMSAE in two scenarios. First, when the target language is included in the training set, we demonstrate that IMSAE, 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 IMSAE outperforms the typical approach of conducting cross-lingual debiasing with a single source language. We experiment with three tasks: profession prediction (§6.1), hate speech recognition (§6.2) and helpfulness prediction (§6.3). For more information on the data sets split and statistics, see Appendix A.

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 example, we compared the performance of male and

[4]The reputation distribution in the Stack Exchange network is highly skewed, where users with a reputation of 1 are in the top 80% of users in Stack Overflow, for example - https://tinyurl.com/2uhn7u52.

female biographies in our profession prediction task. Specifically, we used 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.

Models We benchmark against eight open-weight models. Our chosen models represent varied architectures, parameter sizes, multilingual capabilities, performance levels, and bias tendencies: 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). Due to page constraints, we focus on three representative models in the main paper: mBERT, Llama-3.1, and Mistral-7B-Instruct-v0.3, with comprehensive results for all eight models provided in Appendix A. In our methodology, we used the final hidden state representation from the model's last layer as input for probing experiments on both the primary and bias detection tasks, enabling evaluation of bias before and after debiasing.

6.1 Fair Profession Prediction

Task and Data We use the Multilingual BiasBios dataset (Zhao et al., 2020), an extension of the BiasBios dataset (De-Arteaga et al., 2019) with 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.

6.1.1 Crosslingual Debiasing Results

We evaluated three erasure approaches: null-space projection (INLP; Ravfogel et al. 2020), SVD-based erasure (SAL; Shao et al. 2023b), and SentenceDebias (Liang et al., 2020a). Table 2 presents the results using English as the source language, while the full results show similar trends. While all methods maintain strong downstream task performance in mBERT, their effectiveness varies across model architectures. SAL demonstrates superior bias reduction in LLMs, with average results of 2.4% and 0.6% in mBERT and Mistral respectively. Based on these findings, we selected SAL as our reference and further report results on IMSAE as its structural variant in different settings.

Target	Baseline		SAL (EN)		INLP (EN)		SentenceDebias (EN)	
	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	↓0.2 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
Llama-3.1-8B								
EN	81.1	13.7	↓2.1 79.0	↓0.7 13.0	↓0.7 80.4	↑0.2 13.9	↓0.5 80.6	↓0.4 13.3
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	72.5	25.0	↑0.1 72.6	↑0.1 25.1	72.5	↑1.1 26.1	↓0.1 72.4	↑1.0 26.0
Mistral-7B-Instruct-v0.3								
EN	80.5	14.0	↓2.7 77.8	↓1.3 12.7	↓0.2 80.3	↓0.2 13.8	↓0.3 80.2	↓0.2 13.8
DE	77.3	23.3	↓0.2 77.1	↑0.3 23.6	↓0.1 77.2	23.3	↓0.2 77.1	↓0.1 23.2
FR	71.6	23.1	↑0.2 71.8	↓1.4 21.7	↓0.2 71.4	↓0.4 22.7	↓0.2 71.4	↓2.1 21.0

Table 2: Evaluation of post-hoc debiasing methods on the multilingual BiasBios dataset. 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.

6.1.2 IMSAE Results

With Target Language Consider Table 3. For two out of three LMs, incorporating information from additional languages using IMSAE ("Three-Subsets") further reduces bias on average compared to relying solely on the target language ("Monolingual"). While the FullyJoint approach slightly outperforms IMSAE for mBERT, IMSAE 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 two out of three LMs, IMSAE ("Subsets w/o") outperforms the average cross-lingual debiasing method in terms of debiasing. Both approaches minimally affect main-task performance while effectively reducing bias. See also results in Appendix A.

6.2 Hate Speech Recognition

The Multilingual Twitter Hate Speech corpus (Huang et al., 2020) provides data across multiple demographic dimensions (gender, race, country, and age) and languages (English, Spanish, Italian, Polish, and Portuguese). The demographic attributes are derived directly from user profiles, making the demographic attributes reliable.

Task and Data 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 9 and 10 in the appendix. Huang et al. (2020) labeled

Target	Baseline		SAL (Monolingual)		SAL (Avg-Crosslingual)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (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	↑0.4 15.8	↓0.1 80.4	↓1.2 14.2	80.5	↑0.2 15.6	↓0.1 80.4	↓1.8 13.6
DE	77.7	27.6	↓0.3 77.4	↓0.3 27.3	↑0.1 77.8	↓2.2 25.4	↑0.2 77.9	↓4.6 23.0	↑0.1 77.8	↓1.9 25.7	↓0.1 77.6	↓2.2 25.4
FR	72.7	22.8	↓0.5 72.2	↓3.4 19.4	72.7	↓0.7 22.1	↓0.2 72.5	↓1.5 21.3	↑0.1 72.8	↓0.6 22.2	↓0.5 72.2	↓3.2 19.6
Llama-3.1-8B												
EN	81.1	13.7	↓2.1 79.0	↓0.7 13.0	↓0.8 80.3	↓0.1 13.6	↓2.0 79.1	↓1.1 12.6	↓1.1 80.0	↓0.5 13.2	↓2.1 79.0	↓0.5 13.2
DE	79.8	26.8	↓0.3 79.5	↓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	↓0.2 79.6	↓4.6 22.2
FR	72.5	25.0	↓0.1 72.4	↓5.7 19.3	↑0.1 72.6	↓0.4 24.6	↓0.1 72.4	↓0.4 24.6	72.5	↓0.5 24.5	↓0.1 72.4	↓3.7 21.3
Mistral-7B-Instruct-v0.3												
EN	80.5	14.0	↓2.7 77.8	↓1.3 12.7	↓0.5 80.0	14.0	↓2.8 77.7	↓0.8 13.2	↓0.7 79.8	↓0.1 13.9	↓2.8 77.7	↓1.3 12.7
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	↓0.3 77.0	↓0.8 22.5	↓0.3 77.0	↑0.4 23.7
FR	71.6	23.1	↑0.2 71.8	↓4.9 18.2	↑0.1 71.7	↓1.2 21.9	71.6	↓1.2 21.9	↓0.1 71.5	↓1.8 21.3	↑0.1 71.7	↓5.1 18.0

Table 3: Evaluation of demographic bias mitigation on the multilingual BiasBios dataset. Main shows profession prediction accuracy, while TPR-GAP shows true positive rates between different demographic groups. Results compare Baseline, Monolingual (target language only), Average - Crosslingual, IMSAE on FullyJoint (4), IMSAE on Two-Subsets-Without (excluding target language) and IMSAE Three-Subsets (using all languages).

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.

6.2.1 Results

With Target Language Consider Table 4. IMSAE ("Five-Subsets") demonstrates stronger bias mitigation compared to monolingual debiasing for gender and age debiasing. For race bias, we achieve reductions of 4.1% for English and 6.2% for Spanish, compared to monolingual reductions of 1.3% and 1.2% respectively. Similarly significant improvements are seen for gender bias (11.6% reduction by IMSAE 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 or equal to 2%.

Without Target Language Even when target language data is unavailable, IMSAE ("Subsets w/o") effectively reduces bias across different attributes. Using only non-target languages, we achieve maximum bias reductions of 9.2% for race, 5.2% for gender, 9.0% for age, 5.4% for country. Gender debiasing proves particularly challenging in cross-lingual scenarios due to fundamental differences in gender representation across languages, especially when grammatical gender in some languages creates structural barriers to transfer. The main task performance remains stable, demonstrating IMSAE's ability to preserve useful features while removing bias. See also results in Appendix B.2.

6.3 Multilingual Stack Exchange Fairness Benchmark

We use the MSEFair dataset, for which a detailed description is provided in §5.

6.3.1 Results

With Target Language Consider Table 5. Our method IMSAE, 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 The performance of IMSAE's ("Subsets w/o") varies with linguistic similarity, but outperforms SAL and FullyJoint (§4). However, for Russian, cross-lingual debiasing shows limited effectiveness, with small TPR-Gap reduction. This limitation may stem from Russian's linguistic features affecting bias transfer (Cyrillic script, complex case system, different word order). Newer architectures (Llama and Mistral) show improved cross-lingual debiasing performance compared to mBERT, suggesting better cross-lingual representation alignment in these models.

7 Related Work

We focus in our related work on debiasing in the context of multilingual representations. Debiasing multilingual representations is harder due to

Target	Baseline		Monolingual		SAL (Avg-Crosslingual)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (Five-Subsets)	
	Main	TPR-Gap	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	↓0.3 86.5	↑0.8 4.9	↓0.1 86.7	↑1.9 6.0	↓2.0 84.8	↓4.1 0.0
ES	63.7	10.2	↑0.4 64.1	↓0.1 10.1	↑0.4 64.1	↓0.2 10.0	↑0.4 64.1	↓0.1 10.1	63.7	↓9.2 1.0	↓0.3 63.4	↓1.7 8.5
IT	-	-	-	-	-	-	-	-	-	-	-	-
PL	91.3	6.2	91.3	↓1.2 5.0	91.3	↓0.3 5.9	91.3	6.2	91.3	↓4.7 1.5	↓0.2 91.1	↓6.2 0.0
PT	61.3	1.0	↓0.6 60.7	↑0.1 1.1	↓0.7 60.6	↑1.0 2.0	61.3	↑1.4 2.4	↑0.7 62.0	↑5.0 6.0	↑1.3 62.6	↑11.3 12.3
Gender												
EN	86.7	4.4	86.7	↑0.7 5.1	86.7	4.4	86.7	4.4	86.7	↓4.3 0.1	↓0.1 86.6	↓4.3 0.1
ES	63.7	3.6	↑0.4 64.1	↓0.2 3.4	↓0.1 63.6	3.6	↑0.9 64.6	↑0.1 3.7	63.7	↑2.3 5.9	↓0.3 63.4	↑0.7 4.3
IT	68.4	2.1	68.4	↓0.6 1.5	↓0.3 68.1	↓0.3 1.8	↓0.2 68.2	↓1.7 0.4	↑0.3 68.7	↑1.0 3.1	↓0.5 67.9	↓0.8 1.3
PL	88.2	11.6	↓0.2 88.0	↓8.8 2.8	88.2	↓0.1 11.5	88.2	11.6	↓0.1 88.1	↑0.4 12.0	↓0.2 88.0	↓11.6 0.0
PT	61.3	12.0	↑0.7 62.0	↑1.4 13.4	61.0	12.6	↓1.2 60.1	↑1.8 13.8	↑1.3 62.6	↓5.2 6.8	↑3.1 64.4	↑2.4 14.4
Age												
EN	86.7	9.1	↑0.3 87.0	↑0.4 9.5	↓0.2 86.5	↓0.1 9.0	86.7	↑0.2 9.3	86.7	↓9.0 0.1	↑0.2 86.9	↓9.0 0.1
ES	63.7	12.9	↓0.3 63.4	↓0.5 12.4	63.7	↑0.5 13.4	↓0.3 63.4	↓0.4 12.5	↑0.2 63.9	↓8.0 4.9	↑0.4 64.1	↓6.3 6.6
IT	68.2	3.6	↓0.3 67.9	↑0.3 3.9	68.2	↑0.1 3.7	↓0.3 67.9	↑0.3 3.9	↓1.2 67.0	↓2.8 0.8	↓0.5 67.7	↓0.7 2.9
PL	91.3	8.8	↓0.9 90.4	↓1.9 6.9	91.3	↓0.7 8.1	↓0.1 91.2	↓1.3 7.5	↓0.4 90.9	↓5.0 3.8	↓0.9 90.4	↓8.8 0.0
PT	61.3	17.6	↓0.6 60.7	↑1.5 19.1	↓0.7 60.6	↑0.4 18.0	61.3	↓1.9 15.7	↑1.3 62.6	↑3.4 21.0	↑1.3 62.6	↓3.8 13.8
Country												
EN	82.3	6.7	↓0.1 82.2	↓0.8 5.9	82.3	6.7	82.3	6.7	↓0.1 82.2	↓5.4 1.3	82.3	↓6.6 0.1
ES	65.1	5.1	65.1	↓0.2 4.9	↓0.1 65.0	↑0.3 5.4	↓0.4 64.7	↑1.8 6.9	↓0.6 64.5	↓0.4 4.7	↓0.4 64.7	↓3.0 2.1
IT	71.0	1.7	↑0.1 71.1	↓0.4 1.3	↓0.1 70.9	↑0.4 2.1	↓0.2 70.8	↑0.4 2.1	71.0	↑1.6 3.3	↓0.5 70.5	↓1.1 0.6
PL	-	-	-	-	-	-	-	-	-	-	-	-
PT	64.5	5.1	↓0.5 64.0	↓4.3 0.8	↑0.4 64.9	↑0.4 5.5	↑2.0 66.5	↑4.4 9.5	↑2.0 66.5	↓4.9 0.2	↓1.0 63.5	↓4.2 0.9

Table 4: Demographic bias mitigation results on the Multilingual Hate Speech dataset using mBERT, comparing monolingual and multilingual debiasing approaches. Main: Hate speech prediction accuracy; TPR-Gap: True positive rate gap between demographic groups. 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 9 and Table 10 (appendix).

Target	Baseline		Monolingual		SAL (Avg-Crosslingual)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (Three-Subsets)	
	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap	Main	TPR-Gap
Multilingual BERT												
EN	67.5	10.7	↓4.3 63.2	↓9.4 1.3	↓0.2 67.3	↑0.3 11.0	↓0.3 67.2	10.7	↓4.2 63.3	↓9.3 1.4	↓4.4 63.1	↓9.5 1.2
PT	78.3	16.2	↓19.6 58.7	↓14.8 1.4	78.3	16.2	↓0.1 78.2	↓0.8 15.4	↓19.7 58.6	↓14.5 1.7	↓19.8 58.5	↓15.6 0.6
RU	70.0	18.0	↓11.9 58.1	↓15.1 2.9	↓0.3 69.7	↑0.1 18.1	↓0.2 69.8	↓0.3 17.7	↓0.6 69.4	↓0.1 17.9	↓12.3 57.7	↓14.7 3.3
Llama-3.1-8B												
EN	68.4	11.2	↓4.3 64.1	↓6.0 5.2	68.4	11.2	↓0.1 68.3	↑0.2 11.4	↓4.3 64.1	↓6.1 5.1	↓4.5 63.9	↓5.9 5.3
PT	82.6	22.4	↓18.2 64.4	↓13.6 8.8	↓0.2 82.4	↓0.1 22.3	↓0.7 81.9	↓0.6 21.8	↓18.1 64.5	↓13.1 9.3	↓18.3 64.3	↓13.6 8.8
RU	72.1	16.4	↓8.4 63.7	↓10.4 6.0	↓0.1 72.0	↓0.1 16.3	72.1	↑0.2 16.6	↓0.1 72.0	↓0.3 16.1	↓8.7 63.4	↓10.6 5.8
Mistral-7B-Instruct-v0.3												
EN	66.8	9.5	↓3.7 63.1	↓5.8 3.7	66.8	↓0.1 9.3	↓0.1 66.7	↓0.5 9.0	↓3.6 63.2	↓6.3 3.2	↓3.8 63.0	↓6.4 3.1
PT	81.4	20.9	↓18.2 63.2	↓9.5 11.4	↓0.2 81.2	↑0.4 21.3	↓0.1 81.3	↓0.5 20.4	↓18.1 63.3	↓9.7 11.2	↓17.9 63.5	↓8.2 12.7
RU	70.5	14.7	↓8.2 62.3	↓9.3 5.4	↓0.1 70.4	↓0.3 14.4	70.5	↓0.3 14.4	↓0.1 70.4	↓0.9 13.8	↓8.4 62.1	↓9.4 5.3

Table 5: Reputation bias mitigation on the MSEFair dataset across English, Portuguese and Russian, comparing monolingual and multilingual debiasing approaches. Main: helpfulness prediction accuracy; TPR-Gap: True positive rate gap between demographic groups.

language-specific traits (Ramesh et al., 2023), mismatches like grammatical vs. biological gender (Booij, 2010; Veeman et al., 2020), biases introduced during language alignment (Zhao et al., 2020), and the partial overlap of gender components across languages (Gonen et al., 2022). Past work on multilingual debiasing has largely focused on cross-lingual transfer—applying debiasing from one language to another. Zhou et al. (2019b) laid the groundwork by separating grammatical and semantic gender bias, enabling targeted debiasing via semantic shifts or alignment with English embeddings. Zhao et al. (2020) refined this by equalizing distances between target words and protected sets. Liang et al. (2020b) proposed maximizing intra-group and minimizing inter-group distances across genders. Reusens et al. (2023) found SentenceDebias most effective on mBERT for cross-lingual debiasing. However, Gonen et al. (2022) showed that relying on a single source language misses language-specific bias, limiting effectiveness. Multilingual debiasing requires a comprehen-

sive approach. Vashishtha et al. (2023) found limited debiasing transfer, particularly from English to languages without a Western context. Recent activation editing methods including Sparse Activation Editing (Zhao et al., 2025) and SEA (Qiu et al., 2024) address alignment and bias at inference time, while IMSAE specifically targets joint demographic bias subspaces across multiple languages through iterative erasure.

8 Conclusion

We have studied debiasing multilingual representations by identifying joint linear bias subspaces across languages. Our proposed method, IMSAE, iteratively identifies and removes bias patterns using different language subsets. The method leverages bias patterns from multiple languages, enabling effective zero-shot debiasing where target language data is unavailable but linguistically similar languages can be used. Through experiments across eight languages and five demographic attributes, we demonstrate IMSAE's effectiveness in reducing bias while preserving model utility in state-of-the-art language models. In addition, we have introduced the MSEFair dataset to support future multilingual fairness research further.

Limitations

While IMSAE provides a promising approach to multilingual debiasing, it relies on the assumption of a shared embedding space (Wendler et al., 2024; Fierro et al., 2025), which may be an issue with highly divergent languages.

In addition, our evaluations mostly focus on European languages and a narrow range of demographic attributes, leaving open questions about IMSAE's ability to generalize to typologically diverse languages. This may compromise its performance for such languages and, ultimately, language communities with limited resources. Beyond demographic bias, fairness considerations span diverse AI applications including embodied systems (Li et al., 2024).

Furthermore, our analysis focuses on the final representations of the LLMs. While prior work shows that information is distributed across layers (Zhao et al., 2024), we leave a layer-wise analysis to future work.

Ethical Considerations

Our reliance on datasets that use binary demographic categories, such as male/female or white/non-white, risks reinforcing reductive stereotypes and marginalizing non-binary and intersectional identities. For a discussion, see Dev et al. (2021) and Cao and Daumé III (2020). As we consider deploying IMSAE in real-world applications, it is essential to recognize that bias in language models is complex and context-dependent. Like any debiasing method, IMSAE may have unintended consequences, particularly across different languages and cultural settings. It should not be treated as a superficial fix to mask deeper systemic issues in AI systems. Instead, responsible deployment demands ongoing validation, transparency, and meaningful engagement with the communities affected by these technologies.

Acknowledgments

We thank the reviewers for their helpful comments and feedback. SS and AK acknowledge the support of the UKRI Frontier grant EP/Y031350/1. We gratefully acknowledge an Apple AI/ML scholarship awarded to YQ. ZZ is supported by the UKRI Centre for Doctoral Training in Natural Language Processing (EP/S022481/1). We are grateful for UKRI's support (SC) through the provision of compute resources at the University of Birmingham (Baskerville). We thank the Stack Overflow team for making their data easily available to download.

References

Adi Ben-Israel. 2015. Projectors on intersection of subspaces. *Contemporary Mathematics*, 636:41–50.

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

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

Pinzhen Chen, Zheng Zhao, and Shun Shao. 2024. Cher at KSAA-CAD 2024: Compressing words and definitions into the same space for Arabic reverse dictionary. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 686–691, Bangkok, Thailand. Association for Computational Linguistics.

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.

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

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.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts?

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 et al. 2024. The Llama 3 herd of models.

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.

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.

Yuanchao Li, Lachlan Urquhart, Nihan Karatas, Shun Shao, Hiroshi Ishiguro, and Xun Shen. 2024. Beyond voice assistants: Exploring advantages and risks of an in-car social robot in real driving scenarios.

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.

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.

Dominique Osborne, Shashi Narayan, and Shay B. Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay Cohen. 2024. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987.

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.

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.

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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers.

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.

Runcong Zhao, Chengyu Cao, Qinglin Zhu, Xiucheng Lv, Shun Shao, Lin Gui, Ruifeng Xu, and Yulan He. 2025. Sparse activation editing for reliable instruction following in narratives.

Zheng Zhao, Yftah Ziser, and Shay B Cohen. 2024. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Zhao, Yftah Ziser, Bonnie Webber, and Shay Cohen. 2023. A joint matrix factorization analysis of multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12764–12783, Singapore. Association for Computational Linguistics.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019a. Density matching for bilingual word embedding. 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 1588–1598, Minneapolis, Minnesota. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019b. 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.

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 6: 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.

Appendices

We include below further results that can complement and complete the results in the main part of the paper. The appendix can be skimmed on a first read, and is required only for a more in-depth analysis of IMSAE for those who are interested.

A Multilingual BiasBios Details

This appendix provides comprehensive information about the Multilingual BiasBios dataset used in our profession prediction experiments. Table 6 includes detailed statistics on data distribution across languages, demographic attributes, and experimental configurations. Table 7 shows the complete results for eight different LLMs debiased by SAL, IMSAE ("Three-Subsets"), and IMSAE ("Two-Subsets-Without"). Table 8 compares different post-hoc debiasing methods (SAL, INLP, and SentenceDebias) in crosslingual settings, showing their relative effectiveness when applied across language boundaries.

Target	Baseline		SAL (EN)		SAL (DE)		SAL (FR)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (Three-Subsets)	
	Main	TPR-Gap	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	↑0.4 15.8	80.5	↑0.3 15.7	↓0.1 80.4	↓1.2 14.2	80.5	↑0.2 15.6	↓0.1 80.4	↓1.8 13.6
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	↑0.2 77.9	↓4.6 23.0	↑0.1 77.8	↓1.9 25.7	↓0.1 77.6	↓2.2 25.4
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 72.5	↓1.5 21.3	↑0.1 72.8	↓0.6 22.2	↓0.5 72.2	↓3.2 19.6
Llama3-8B														
EN	81.2	13.3	↓2.2 79.0	↓0.8 12.5	↓1.1 80.1	↓0.6 12.7	↓0.5 80.7	↓0.5 12.8	↓2.1 79.1	13.3	↓1.1 80.1	↓0.2 13.1	↓2.2 79.0	↓0.4 12.9
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	↓0.3 78.7	↓0.8 25.5	↓0.3 78.7	↓0.8 25.5	↓0.4 78.6	↑0.2 26.5
FR	72.8	26.1	↑0.1 72.9	↑0.2 26.3	72.8	↑1.2 27.3	72.8	↓4.5 21.6	72.8	↑1.1 27.2	72.8	26.1	↓0.1 72.7	↓4.6 21.5
Llama-3.1-8B														
EN	81.1	13.7	↓2.1 79.0	↓0.7 13.0	↓0.9 80.2	↓0.1 13.6	↓0.8 80.3	↓0.2 13.5	↓2.0 79.1	↓1.1 12.6	↓1.1 80.0	↓0.5 13.2	↓2.1 79.0	↓0.5 13.2
DE	79.8	26.8	↓0.3 79.5	↑0.2 27.0	↓0.3 79.5	↓5.2 21.6	↓0.2 79.6	↑0.5 27.3	↓0.2 79.6	↑0.3 27.1	↓0.4 79.4	↑0.4 27.2	↓0.2 79.6	↓4.6 22.2
FR	72.5	25.0	↑0.1 72.6	↑0.1 25.1	72.5	↓0.9 24.1	↓0.1 72.4	↓5.7 19.3	↓0.1 72.4	↓0.4 24.6	72.5	↓0.5 24.5	↓0.1 72.4	↓3.7 21.3
Llama-3.2-3B														
EN	79.9	13.0	↓1.1 78.8	↓1.4 11.6	↓0.2 79.7	↓0.5 12.5	↓0.2 79.7	↓0.6 12.4	↓1.1 78.8	↓1.0 12.0	↓0.5 79.4	↓0.6 12.4	↓1.1 78.8	↓1.5 11.5
DE	78.2	27.9	↓0.1 78.1	27.9	↓0.3 77.9	↓0.6 27.3	78.2	27.9	↓0.1 78.1	27.9	↓0.2 78.0	↑0.2 28.1	↓0.4 77.8	↓0.5 27.4
FR	71.2	16.3	↓0.2 71.0	↑1.1 17.4	71.2	↓1.4 14.9	↓0.2 71.0	↓0.9 15.4	↓0.2 71.0	↑0.3 16.6	71.2	↓1.0 15.3	↓0.2 71.0	↓1.5 14.8
Mistral-7B-Instruct-v0.3														
EN	80.5	14.0	↓2.7 77.8	↓1.3 12.7	↓0.4 80.1	14.0	↓0.6 79.9	↓0.1 13.9	↓2.8 77.7	↓0.8 13.2	↓0.7 79.8	↓0.1 13.9	↓2.8 77.7	↓1.3 12.7
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	↓0.2 77.1	↑0.3 23.6	↓0.3 77.0	↓0.8 22.5	↓0.3 77.0	↑0.4 23.7
FR	71.6	23.1	↑0.2 71.8	↓1.4 21.7	71.6	↓1.0 22.1	↑0.2 71.8	↓4.9 18.2	71.6	↓1.2 21.9	↓0.1 71.5	↓1.8 21.3	↑0.1 71.7	↓5.1 18.0
Mistral-7B-v0.3														
EN	80.9	13.9	↓2.9 78.0	↓1.2 12.7	↓0.9 80.0	↑0.1 14.0	↓1.2 79.7	13.9	↓2.9 78.0	↓0.7 13.2	↓0.7 80.2	↓0.1 13.8	↓3.0 77.9	↓0.8 13.1
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	↑0.2 78.6	↑0.2 27.5	↓0.1 78.3	27.3	↓0.2 78.2	↓1.2 26.1
FR	72.2	23.7	↑0.1 72.3	↓2.8 20.9	↑0.1 72.3	↓0.7 23.0	↓0.1 72.1	↓4.4 19.3	72.2	↓0.5 23.2	72.2	↓1.1 22.6	↓0.1 72.1	↓4.4 19.3
Mistral-Nemo-Base-2407														
EN	82.2	13.1	↓3.0 79.2	↓0.2 12.9	↓0.7 81.5	↓0.2 12.9	↓0.2 82.0	↓0.4 12.7	↓3.2 79.0	↓0.1 13.0	↓0.8 81.4	↑0.1 13.2	↓3.8 78.4	↓0.6 12.5
DE	79.4	31.0	↑0.2 79.6	31.0	↑0.1 79.5	↓1.0 30.0	↑0.2 79.6	↑0.1 31.1	↑0.2 79.6	31.0	↑0.1 79.5	↓0.1 30.9	↑0.2 79.6	↓1.0 30.0
FR	73.9	22.0	↑0.2 74.1	↑0.9 22.9	↑0.4 74.3	↑1.3 23.3	↑1.0 74.9	↓0.3 21.7	↑0.3 74.2	↑0.7 22.7	↑0.2 74.1	↓1.8 20.2	↑1.0 74.9	↑1.0 23.0
Mistral-Nemo-Instruct-2407														
EN	81.1	12.5	↓2.2 78.9	↓0.6 11.9	↑0.3 81.4	↓0.1 12.4	↑0.3 81.4	↓0.4 12.1	↓2.5 78.6	↓1.2 11.3	81.1	12.5	↓2.6 78.5	↓0.9 11.6
DE	78.4	26.6	78.4	↓0.2 26.4	↑0.4 78.8	↓1.1 25.5	↓0.1 78.3	↓0.6 26.0	↑0.1 78.5	↓0.2 26.4	↑0.1 78.5	↓0.3 26.3	↑0.6 79.0	↓1.3 25.3
FR	72.8	21.8	↓0.2 72.6	↓1.9 19.9	↓0.1 72.7	↓1.9 19.9	↑0.6 73.4	↑1.1 22.9	↑0.1 72.9	↓2.4 19.4	↓0.1 72.7	↓2.8 19.0	↑0.4 73.2	↑0.7 22.5

Table 7: Gender debiasing performance evaluation on BiasBios across eight language models. We report both main task accuracy and TPR-Gap reduction for each model architecture and debiasing approach. IMSAE (Three-Subsets) demonstrates superior cross-lingual performance compared to monolingual methods, particularly for newer LLM architectures.

Target	Baseline		SAL (EN)		SAL (DE)		SAL (FR)		INLP (EN)		INLP (DE)		INLP (FR)		SentenceDebias (EN)		SentenceDebias (DE)		SentenceDebias (FR)	
	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext
mBERT-uncased																				
EN	80.5	15.4	↓0.1 80.4	↓1.9 13.5	80.5	↑0.4 15.8	80.5	↑0.3 15.7	80.5	↓0.2 15.2	80.5	↓0.1 15.3	80.5	↓0.5 14.9	↓0.2 80.3	↓0.1 15.3	80.5	↓0.1 15.3	80.5	↓0.3 15.1
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	↓2.2 25.4	↑0.1 77.8	↓1.5 26.1	77.7	27.6	↓0.1 77.6	↓2.4 25.2	↓1.1 76.6	↓0.5 27.1	↑0.2 77.9	↓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	72.7	↓0.5 22.3	↑0.1 72.8	↓0.7 22.1	↓0.2 72.5	↓1.1 21.7	↑0.1 72.8	↓0.5 22.3	↓0.2 72.5	↓0.1 22.7	↓0.2 72.5	↓1.3 21.5
Llama3-8B																				
EN	81.2	13.3	↓2.2 79.0	↓0.8 12.5	↓1.1 80.1	↓0.6 12.7	↓0.5 80.7	↓0.5 12.8	↓0.3 80.9	↓0.6 12.7	↓0.3 80.9	↓0.5 12.8	↓0.6 80.6	↓0.5 12.8	↓0.4 80.8	↓0.7 12.6	↓0.7 80.5	↓0.5 12.8	↓0.5 80.7	↓0.6 12.7
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	79.0	26.3	↓0.2 78.8	↑0.1 26.4	↑0.1 79.1	26.3	↓0.1 78.9	↑0.1 26.4	↓0.9 78.1	↑0.7 27.0	↑0.1 79.1	↑1.3 27.6
FR	72.8	26.1	↑0.1 72.9	↑0.2 26.3	72.8	↑1.2 27.3	72.8	↓4.5 21.6	↓0.1 72.7	↑1.0 27.1	↓0.1 72.7	↑1.6 27.7	↓0.1 72.7	↓0.8 25.3	↑0.1 72.9	↓0.7 25.4	72.8	↑0.3 26.4	72.8	↑0.8 26.9
Llama-3.1-8B																				
EN	81.1	13.7	↓2.1 79.0	↓0.7 13.0	↓0.9 80.2	↓0.1 13.6	↓0.8 80.3	↓0.2 13.5	↓0.7 80.4	↑0.2 13.9	↓0.5 80.6	↓0.2 13.5	↓0.7 80.4	13.7	↓0.5 80.6	↓0.4 13.3	↓0.3 80.8	↑0.1 13.8	↓0.9 80.2	13.7
DE	79.8	26.8	↓0.3 79.5	↑0.2 27.0	↓0.3 79.5	↓5.2 21.6	↓0.2 79.6	↑0.5 27.3	79.8	26.8	↑0.1 79.9	26.8	79.8	26.8	↓0.2 79.6	↑0.2 27.0	↓0.8 79.0	↓2.7 24.1	79.8	↓0.1 26.7
FR	72.5	25.0	↑0.1 72.6	↑0.1 25.1	72.5	↓0.9 24.1	↓0.1 72.4	↓5.7 19.3	72.5	↑1.1 26.1	72.5	↑2.0 27.0	↑0.1 72.6	↑0.7 25.7	↓0.1 72.4	↑1.0 26.0	↓0.1 72.4	↑0.2 25.2	↑0.2 72.7	↑0.2 25.2
Llama-3.2-3B																				
EN	79.9	13.0	↓1.1 78.8	↓1.4 11.6	↓0.2 79.7	↓0.5 12.5	↓0.2 79.7	↓0.6 12.4	79.9	↑0.2 13.2	↑0.2 80.1	13.0	79.9	↓0.1 12.9	↓0.1 79.8	13.0	↑0.1 80.0	↑0.2 13.2	↑0.6 80.5	↑0.4 13.4
DE	78.2	27.9	↓0.1 78.1	27.9	↓0.3 77.9	↓0.6 27.3	78.2	27.9	78.2	27.9	78.2	27.9	78.2	27.9	78.2	27.9	↓1.0 77.2	↓1.6 26.3	78.2	27.9
FR	71.2	16.3	↓0.2 71.0	↑1.1 17.4	71.2	↓1.4 14.9	↓0.2 71.0	↓0.9 15.4	↓0.2 71.0	↑0.7 17.0	↓0.1 71.1	↑1.1 17.4	↓0.2 71.0	↑1.3 17.6	71.2	↑2.1 18.4	↓0.1 71.1	↓0.6 15.7	↓0.5 70.7	↓1.2 15.1
Mistral-7B-Instruct-v0.3																				
EN	80.5	14.0	↓2.7 77.8	↓1.3 12.7	↓0.4 80.1	14.0	↓0.6 79.9	↓0.1 13.9	↓0.2 80.3	↓0.2 13.8	↓0.2 80.3	↑0.3 14.3	↑0.1 80.6	↓0.1 13.9	↓0.3 80.2	↓0.2 13.8	↓0.5 80.0	14.0	↓0.4 80.1	↓0.2 13.8
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	↓0.1 77.2	23.3	77.3	23.3	↓0.1 77.2	23.3	↓0.2 77.1	↓0.1 23.2	↓1.1 76.2	↓0.4 22.9	77.3	23.3
FR	71.6	23.1	↑0.2 71.8	↓1.4 21.7	71.6	↓1.0 22.1	↑0.2 71.8	↓4.9 18.2	↓0.2 71.4	↓0.4 22.7	↑0.1 71.7	↓0.3 22.8	71.6	↓1.2 21.9	↓0.2 71.4	↓2.1 21.0	71.6	↓1.6 21.5	↓0.3 71.3	↓2.7 20.4
Mistral-7B-v0.3																				
EN	80.9	13.9	↓2.9 78.0	↓1.2 12.7	↓0.9 80.0	↑0.1 14.0	↓1.2 79.7	13.9	NaN	NaN	↓0.3 80.6	13.9	↓0.4 80.5	↓0.1 13.8	↓0.3 80.6	↓0.1 13.8	↑0.5 81.4	↑0.7 14.6	↓0.3 80.6	↑0.3 14.2
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	78.4	27.3	↓0.1 78.3	↑0.4 27.7	78.4	27.3	↑0.1 78.5	↓0.1 27.2	↓0.3 78.1	↑1.0 28.3	78.4	27.3
FR	72.2	23.7	↑0.1 72.3	↓2.8 20.9	↑0.1 72.3	↓0.7 23.0	↓0.1 72.1	↓4.4 19.3	↓0.1 72.1	↑1.2 24.9	72.2	↓0.5 23.2	72.2	↓0.1 23.6	72.2	↓0.5 23.2	↓0.1 72.1	↓0.2 23.5	↓0.3 71.9	↓4.6 19.1
Mistral-Nemo-Base-2407																				
EN	82.2	13.1	↓3.0 79.2	↓0.2 12.9	↓0.7 81.5	↓0.2 12.9	↓0.2 82.0	↓0.4 12.7	↓0.2 82.0	↓0.7 12.4	↓0.5 81.7	↓0.7 12.4	↑0.1 82.3	↓0.2 12.9	↓0.4 81.8	↓0.3 12.8	↓0.5 81.7	↑0.5 13.6	↓0.4 81.8	↓0.4 12.7
DE	79.4	31.0	↑0.2 79.6	31.0	↑0.1 79.5	↓1.0 30.0	↑0.2 79.6	↑0.1 31.1	↑0.1 79.5	↑0.1 31.1	79.4	↑0.1 31.1	79.4	31.0	↑0.1 79.5	31.0	↓0.1 79.3	↑0.6 31.6	79.4	↑0.1 31.1
FR	73.9	22.0	↑0.2 74.1	↑0.9 22.9	↑0.4 74.3	↑1.3 23.3	↑1.0 74.9	↓0.3 21.7	↑0.2 74.1	↑1.0 23.0	↓0.3 73.6	↓0.6 21.4	↓0.1 73.8	↑0.9 22.9	↓0.2 73.7	↓1.1 20.9	↓0.1 73.8	↓0.5 21.5	↑0.1 74.0	↓0.1 21.9
Mistral-Nemo-Instruct-2407																				
EN	81.1	12.5	↓2.2 78.9	↓0.6 11.9	↑0.3 81.4	↓0.1 12.4	↑0.3 81.4	↓0.4 12.1	↑0.2 81.3	↑0.5 13.0	↑0.4 81.5	↑0.5 13.0	↑0.5 81.6	↑0.2 12.7	↑0.3 81.4	↑0.1 12.6	↓0.4 80.7	↑0.3 12.8	↑0.1 81.2	↑0.2 12.7
DE	78.4	26.6	78.4	↓0.2 26.4	↑0.4 78.8	↓1.1 25.5	↓0.1 78.3	↓0.6 26.0	78.4	26.6	↓0.3 78.1	↑2.7 29.3	↓0.1 78.3	↑4.1 30.7	↓0.4 78.0	↑3.1 29.7	↑0.3 78.7	↑5.8 32.4	↑0.1 78.5	↑0.2 26.8
FR	72.8	21.8	↓0.2 72.6	↓1.9 19.9	↓0.1 72.7	↓1.9 19.9	↑0.6 73.4	↑1.1 22.9	↓0.1 72.7	↓1.3 20.5	↓0.1 72.7	↓1.0 20.8	↓0.1 72.7	↓0.6 21.2	↓0.5 72.3	↓3.0 18.8	↓0.2 72.6	↓2.1 19.7	↓0.1 72.7	↓1.5 20.3

Table 8: Comparative analysis of crosslingual gender debiasing approaches on BiasBios, showing performance across different source-target language pairs. This analysis reveals how IMSAE leverages multilingual representations more effectively than single-source transfer methods.

B Multilingual Hate Speech Details

This appendix presents detailed information about our experiments on the Multilingual Twitter Hate Speech corpus. The detailed statistics for the Multilingual Hate Speech Dataset are presented in Table 9 (training and test set sizes) and Table 10 (class distribution across subsets). Comprehensive debiasing results comparing SAL and IMSAE across five languages and eight language models are provided for each demographic attribute: Age bias (Table 11), Country bias (Table 12), Gender bias (Table 13) and Race bias (Table 14).

B.1 Dataset Summary

This section provides the detailed statistics of the Multilingual Hate Speech dataset, including the number of samples and distribution across different demographic attributes.

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 9: Training and test set sizes for different languages and demographic attributes in the Multilingual Hate Speech dataset. The diverse distribution across languages enables robust evaluation of cross-lingual transfer capabilities.

Lang	Bias	Train		Dev		Test		Total	
		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 10: Distribution of samples across different languages and protected attributes in the Multilingual Hate Speech dataset. C0 and C1 represent the two classes for each protected attribute. Note the inherent class imbalance in certain language-attribute combinations.

B.2 IMSAE Debiasing Results

This section presents the complete results of applying IMSAE for bias mitigation across different demographic attributes in the Multilingual Hate Speech dataset. Our evaluation compares both standard and zero-shot settings by: Age bias (Table 11), Country bias (Table 12), Gender bias (Table 13) and Race bias (Table 14).

Target	Baseline		SAL (EN)		SAL (ES)		SAL (IT)		SAL (PL)		SAL (PT)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (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	Main	TPR-Gap
mBERT-uncased																		
EN	86.7	9.1	↑0.3 87.0	↑0.4 9.5	↓0.5 86.2	↓0.3 8.8	86.7	↑0.1 9.2	↓0.5 86.2	↓0.1 9.0	86.7	↓0.1 9.0	86.7	↑0.2 9.3	86.7	↓9.0 0.1	↑0.2 86.9	↓9.0 0.1
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	↓0.3 63.4	↓0.4 12.5	↑0.2 63.9	↓8.0 4.9	↑0.4 64.1	↓6.3 6.6
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	↓0.3 67.9	↑0.3 3.9	↓1.2 67.0	↓2.8 0.8	↓0.5 67.7	↓0.7 2.9
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	↓1.3 7.5	↓0.4 90.9	↓5.0 3.8	↓0.9 90.4	↓8.8 0.0
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	61.3	↓1.9 15.7	↑1.3 62.6	↑3.4 21.0	↑1.3 62.6	↓3.8 13.8
Llama3-8B																		
EN	79.6	7.8	↑0.8 80.4	↑0.6 8.4	79.6	↓0.4 7.4	↑0.1 79.7	↓0.1 7.7	↑0.1 79.7	↓0.2 7.6	↑0.3 79.9	↓0.3 7.5	↑0.1 79.7	7.8	↑0.1 79.7	↓7.3 0.5	↑1.0 80.6	↓0.4 7.4
ES	70.7	11.7	↑0.3 71.0	↑0.6 12.3	↓0.2 70.5	↓0.6 11.1	↑0.3 71.0	↑0.6 12.3	70.7	11.7	↑0.5 71.2	↑0.5 12.2	70.7	11.7	70.7	↓1.1 10.6	↓0.2 70.5	↑1.2 12.9
IT	69.9	8.0	↓0.3 69.6	↑0.3 8.3	69.9	8.0	↓0.5 69.4	↓0.7 7.3	↑0.2 70.1	↓0.9 7.1	↓0.3 69.6	↓0.3 7.7	↓0.5 69.4	↑1.2 9.2	↓0.3 69.6	↓0.1 7.9	↑0.2 70.1	↓3.8 4.2
PL	91.0	17.2	91.0	↓0.7 16.5	↓0.1 90.9	↓0.6 16.6	91.0	↓0.1 17.1	↓1.2 89.8	↓7.8 9.4	91.0	↓0.8 16.4	↓0.1 90.9	↓0.7 16.5	91.0	↑3.7 20.9	↓1.3 89.7	↓9.1 8.1
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	↓1.3 55.8	↑1.4 3.5	↓1.3 55.8	↑3.2 5.3	↓0.7 56.4	↑8.3 10.4
Llama-3.1-8B																		
EN	79.7	6.5	↑0.7 80.4	↑0.7 7.2	79.7	↓0.1 6.4	79.7	↑0.2 6.7	↑0.1 79.8	↑0.1 6.6	79.7	↓0.2 6.3	↑0.1 79.8	↓0.3 6.2	↑0.1 79.8	↓2.1 4.4	↑1.0 80.7	↓4.1 2.4
ES	72.9	8.6	↓0.2 72.7	↑0.1 8.7	↓0.2 72.7	↓1.0 7.6	72.9	↓1.1 7.5	↓0.5 72.4	↓0.5 8.1	72.9	↓2.2 6.4	72.9	↑0.7 9.3	↓0.9 72.0	↑2.6 11.2	↓0.5 72.4	↓2.7 5.9
IT	66.5	9.5	↑0.2 66.7	↓1.2 8.3	↑0.5 67.0	↓1.0 8.5	↑0.7 67.2	↓0.5 9.0	↑0.5 67.0	↓1.0 8.5	↑0.5 67.0	↓0.1 9.4	↑0.2 66.7	↓1.2 8.3	↑0.7 67.2	↓0.9 8.6	66.5	↑2.2 11.7
PL	90.4	15.2	↓0.1 90.3	↓0.6 14.6	90.4	15.2	90.4	15.2	↓0.4 90.0	↓6.9 8.3	↓0.1 90.3	↓0.6 14.6	↓0.1 90.3	↓0.6 14.6	90.4	↑2.1 17.3	↓0.7 89.7	↓11.4 3.8
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	59.5	↑1.5 4.3	↑0.6 60.1	↑1.8 4.6	↓1.8 57.7	↑1.7 4.5
Llama-3.2-3B																		
EN	79.7	6.2	↑0.6 80.3	↑0.8 7.0	↑0.1 79.8	↑0.3 6.5	79.7	6.2	↑0.1 79.8	↑0.1 6.3	↑0.1 79.8	↑0.3 6.5	79.7	↑0.4 6.6	79.7	↓6.0 0.2	↑0.7 80.4	↓6.0 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	↓0.5 67.8	↑0.9 13.0	68.3	12.1	68.3	↓0.5 11.6	↑0.2 68.5	↓2.3 9.8
IT	67.7	5.0	67.7	↑0.3 5.3	↑0.5 68.2	↑2.0 7.0	↓0.2 67.5	↑0.7 5.7	↑0.2 67.9	↑1.8 6.8	↑0.2 67.9	↑1.5 6.5	↑0.2 67.9	↑0.5 5.5	67.7	↑3.5 8.5	↓0.2 67.5	↑0.8 5.8
PL	90.9	17.1	↓0.2 90.7	↑0.1 17.2	↓0.2 90.7	↓0.6 16.5	↓0.3 90.6	↓0.6 16.5	↓0.9 90.0	↓1.4 15.7	90.9	17.1	↓0.2 90.7	↑0.1 17.2	↓0.3 90.6	↓10.7 6.4	↓0.9 90.0	↑7.4 24.5
PT	56.4	8.9	↑0.7 57.1	↑2.8 11.7	↑0.7 57.1	↑0.8 9.7	↑0.7 57.1	↑0.8 9.7	↓1.2 55.2	↑1.5 10.4	56.4	↑1.7 10.6	↓1.2 55.2	↑1.5 10.4	↑1.3 57.7	↑7.7 16.6	↑0.7 57.1	↑11.1 20.0
Mistral-7B-Instruct-v0.3																		
EN	79.4	7.1	↑1.1 80.5	↑1.3 8.4	↑0.2 79.6	↑0.1 7.2	↑0.3 79.7	↑0.1 7.2	↑0.1 79.5	↑0.3 7.4	↑0.2 79.6	↓0.1 7.0	↑0.2 79.6	↑0.2 7.3	↑0.1 79.5	↑0.6 7.7	↑0.7 80.1	↓5.6 1.5
ES	64.6	7.2	↓0.7 63.9	↓0.2 7.0	↑0.8 65.4	↑1.6 8.8	↓0.2 64.4	↓0.8 6.4	↓0.5 64.1	↑0.5 7.7	↓0.5 64.1	↓1.5 5.7	↓1.2 63.4	↓0.1 7.1	64.6	7.2	↓1.2 63.4	↑13.0 20.2
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	↑0.5 66.5	↑0.1 3.6	↓0.2 65.8	↓0.2 3.3	↓0.2 65.8	↓0.5 3.0	66.0	↑1.4 4.9	↓0.2 65.8	↓0.1 3.4
PL	91.0	19.5	91.0	19.5	↑0.2 91.2	↑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	↑0.7 20.2	↑0.2 91.2	↑6.8 26.3	↓0.7 90.3	↓0.3 19.2
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	59.5	17.2	↓0.6 58.9	↓0.4 16.8	↓1.2 58.3	↑0.2 17.4
Mistral-7B-v0.3																		
EN	79.8	7.1	↑0.6 80.4	↑0.4 7.5	↓0.1 79.7	7.1	79.8	↓0.3 6.8	↓0.1 79.7	↑0.1 7.2	↓0.2 79.6	↓0.3 6.8	↓0.1 79.7	↓0.4 6.7	↓0.3 79.5	↓0.7 6.4	↑0.6 80.4	↑2.2 9.3
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	↓0.3 66.8	↑1.8 14.6	67.1	↑0.2 13.0	67.1	↓0.6 12.2	67.1	↓7.0 5.8	↓1.0 66.1	↓11.5 1.3
IT	65.8	5.6	65.8	↓0.9 4.7	↓0.2 65.6	↑0.5 6.1	↑0.2 66.0	↓0.6 5.0	65.8	↓0.1 5.5	65.8	↓0.8 4.8	↑0.2 66.0	↓1.3 4.3	↑0.7 66.5	↓2.1 3.5	65.8	↓3.7 1.9
PL	90.4	17.8	↑0.3 90.7	↑0.4 18.2	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	17.8	↑0.2 90.6	↓17.2 0.6	↓0.7 89.7	↑4.0 21.8
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	57.7	12.4	57.7	↑0.7 13.1	57.7	12.4
Mistral-Nemo-Base-2407																		
EN	78.9	6.4	↑0.4 79.3	↑1.1 7.5	↓0.3 78.6	↑0.1 6.5	↓0.5 78.4	↑0.2 6.6	78.9	↓0.1 6.3	↓0.3 78.6	↓0.3 6.1	↓0.5 78.4	↓0.4 6.0	↓0.4 78.5	↑3.7 10.1	↑0.3 79.2	↑1.8 8.2
ES	72.2	16.3	↓1.2 71.0	↑1.3 17.6	↑0.7 72.9	↑1.8 18.1	72.2	↑2.0 18.3	↑0.2 72.4	↑0.6 16.9	↓0.5 71.7	↓1.0 15.3	↓0.7 71.5	↑1.5 17.8	↓0.2 72.0	↓3.9 12.4	↑0.2 72.4	↓0.4 15.9
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	↑1.0 6.8	↑0.5 65.3	↑0.7 6.5	↑0.5 65.3	↑4.9 10.7
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	↓0.2 91.1	↓0.7 20.2	↓0.4 90.9	↓8.1 12.8	↓1.1 90.2	↓12.6 8.3
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	↓3.0 5.8	↑1.2 64.4	↑2.5 11.3	63.2	↓2.0 6.8
Mistral-Nemo-Instruct-2407																		
EN	79.3	5.9	↑0.3 79.6	↑1.1 7.0	↓0.2 79.1	↑0.1 6.0	↓0.2 79.1	↑0.2 6.1	↓0.2 79.1	↑0.1 6.0	↓0.3 79.0	↑0.5 6.4	↓0.3 79.0	↑0.6 6.5	↓0.2 79.1	↓2.0 3.9	↑0.4 79.7	↓5.2 0.7
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	↓0.5 70.5	↑0.2 10.4	↓0.3 70.7	↓3.4 6.8	↓2.2 68.8	↓2.6 7.6
IT	65.3	8.0	↓0.2 65.1	↓0.6 7.4	↓0.2 65.1	↓0.1 7.9	65.3	↓0.4 7.6	65.3	8.0	65.3	↑0.5 8.5	↓0.2 65.1	↓0.6 7.4	↑0.5 65.8	↑0.9 8.9	↓0.5 64.8	↓1.2 6.8
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	↓0.3 90.6	↑0.1 19.0	90.9	↓2.5 16.4	↓0.9 90.0	↑1.1 20.0
PT	64.4	1.5	64.4	1.5	64.4	1.5	↓0.6 63.8	↑1.8 3.3	↑1.2 65.6	↑2.1 3.6	64.4	1.5	↑0.6 65.0	↑1.0 2.5	64.4	↑3.1 4.6	64.4	↑1.7 3.2

Table 11: Comprehensive evaluation of age bias mitigation on the Multilingual Hate Speech dataset across eight language models. We report hate speech prediction accuracy (Main) and True Positive Rate gap (TPR-Gap) between demographic groups for each model and debiasing approach. IMSAE (Five-Subsets) consistently achieves stronger bias reduction compared to monolingual and standard cross-lingual approaches.

Target	Baseline		SAL (EN)		SAL (ES)		SAL (IT)		SAL (PL)		SAL (PT)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (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	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	82.3	6.7	82.3	6.7	↓0.1 82.2	↓5.4 1.3	82.3	↓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	↑1.8 6.9	↓0.6 64.5	↓0.4 4.7	↓0.4 64.7	↓3.0 2.1
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	↓0.2 70.8	↑0.4 2.1	71.0	↑1.6 3.3	↓0.5 70.5	↓1.1 0.6
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	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	↑4.4 9.5	↑2.0 66.5	↓4.9 0.2	↓1.0 63.5	↓4.2 0.9
Llama3-8B																		
EN	77.1	6.0	77.1	↓1.0 5.0	77.1	↑0.4 6.4	77.1	6.0	↓0.1 77.0	↑0.1 6.1	↓0.1 77.0	↑0.3 6.3	77.1	↑0.1 6.1	↓0.2 76.9	↓3.5 2.5	↓0.3 76.8	↓4.7 1.3
ES	66.7	10.0	↑1.2 67.9	↑1.0 11.0	↑0.3 67.0	↑0.1 10.1	66.7	10.0	66.7	10.0	↑0.7 67.4	↓1.3 8.7	↑0.7 67.4	↑0.6 10.6	↑1.2 67.9	↓6.0 4.0	↑1.4 68.1	↑2.9 12.9
IT	70.5	12.4	↑0.5 71.0	↑0.4 12.8	↑0.6 71.1	↑0.7 13.1	↑0.3 70.8	↑0.2 12.6	↑0.2 70.7	↑0.3 12.7	↑0.8 71.3	↑0.6 13.0	↑0.3 70.8	↑0.6 13.0	↑1.2 71.7	↓1.1 11.3	↑0.6 71.1	↓6.5 5.9
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	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	↑0.5 68.0	↑0.5 3.4	67.5	↓0.4 2.5	↓0.5 67.0	↓1.6 1.3
Llama-3.1-8B																		
EN	77.1	5.7	↓0.2 76.9	↓0.6 5.1	77.1	↓0.1 5.6	↓0.1 77.0	↑0.2 5.9	77.1	↑0.2 5.9	↓0.1 77.0	5.7	77.1	↑0.1 5.8	↑0.1 77.2	↓5.6 0.1	↓0.4 76.7	↓4.0 1.7
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	↓0.2 70.2	↑0.3 22.5	↑0.2 70.6	↑0.6 22.8	↓0.2 70.2	↓9.3 12.9	70.4	↓6.8 15.4
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	↑0.2 68.8	↓0.1 12.6	↑0.3 68.9	↓0.1 12.6	↑0.3 68.9	↓1.1 11.6	↓0.3 68.3	↓2.1 10.6
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	99.6	0.0
PT	66.5	5.4	66.5	↓3.7 1.7	↑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	↑0.5 67.0	↓1.4 4.0	↑1.5 68.0	↓0.3 5.1	↑0.5 67.0	↓0.1 5.3
Llama-3.2-3B																		
EN	76.5	4.8	↓0.4 76.1	↓0.7 4.1	76.5	↑0.4 5.2	↑0.1 76.6	↑0.3 5.1	↑0.1 76.6	↓0.1 4.7	76.5	↑0.4 5.2	76.5	↑0.3 5.1	↓0.2 76.3	↓4.1 0.7	↓0.4 76.1	↓3.2 1.6
ES	67.7	8.9	67.7	↓0.6 8.3	↓0.5 67.2	↓0.6 8.3	↓0.5 67.2	↑1.0 9.9	↑0.2 67.9	↑0.7 9.6	67.7	8.9	↓0.3 67.4	↑0.4 9.3	↑0.4 68.1	↓1.9 7.0	↑0.6 68.3	↓1.5 7.4
IT	66.1	15.6	66.1	↓0.9 14.7	66.1	↓0.9 14.7	↓0.1 66.0	↓0.4 15.2	↓0.1 66.0	↓0.9 14.7	66.1	↓0.9 14.7	66.1	↓0.9 14.7	↓0.4 65.7	↑0.2 15.8	↑0.4 66.5	↓4.1 11.5
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	99.6	0.0
PT	67.0	5.5	↓1.0 66.0	↓0.9 4.6	↓1.5 65.5	↑3.3 8.8	↓1.0 66.0	↑2.2 7.7	67.0	5.5	↑0.5 67.5	↑0.7 6.2	67.0	5.5	↑0.5 67.5	↓2.7 2.8	↓1.0 66.0	↑6.0 11.5
Mistral-7B-Instruct-v0.3																		
EN	75.2	6.1	↓0.2 75.0	↓1.2 4.9	75.2	6.1	75.2	↓0.1 6.0	↑0.1 75.3	↑0.3 6.4	75.2	↑0.2 6.3	↑0.1 75.3	6.1	↑0.1 75.3	↓2.0 4.1	↓0.4 74.8	↓2.7 3.4
ES	60.6	13.5	↑0.9 61.5	↑0.5 14.0	↑0.4 61.0	↓1.1 12.4	↑0.2 60.8	↓0.4 13.1	↑0.2 60.8	↑0.5 14.0	↑0.4 61.0	13.5	↑1.1 61.7	↑0.3 13.8	↑0.7 61.3	↓9.4 4.1	↑1.4 62.0	↑0.2 13.7
IT	68.9	7.7	↓0.1 68.8	↑0.5 8.2	68.9	7.7	68.9	7.7	↓0.1 68.8	↑0.1 7.8	↑0.5 69.4	↑0.1 7.8	↑0.2 69.1	↑0.8 8.5	↑0.4 69.3	↑6.1 13.8	↑0.4 69.3	↑5.8 13.5
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	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	↑0.5 61.4	↑1.4 2.1	↓1.0 59.9	↑3.0 3.7	↑0.5 61.4	↑3.7 4.4
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	4.4	↑0.1 75.6	↑0.2 4.6	↑0.2 75.7	↓0.8 3.6	↓0.1 75.4	↓0.4 4.0
ES	64.5	2.7	↑0.4 64.9	↑0.7 3.4	64.5	↓0.6 2.1	↓0.3 64.2	↑0.6 3.3	64.5	2.7	↓0.3 64.2	↑0.6 3.3	↑0.6 65.1	↑0.2 2.9	↑0.2 64.7	↑4.3 7.0	↓0.3 64.2	↑1.6 4.3
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	↑0.2 67.4	↓0.8 10.2	67.2	11.0	↑0.3 67.5	↓6.9 4.1	↑0.3 67.5	↓2.7 8.3
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	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	66.0	5.4	↓2.0 64.0	↓1.7 3.7	↓1.5 64.5	↓1.8 3.6
Mistral-Nemo-Base-2407																		
EN	75.8	4.8	75.8	↓0.6 4.2	↑0.1 75.9	4.8	↓0.1 75.7	↓0.2 4.6	↑0.1 75.9	↑0.1 4.9	↑0.1 75.9	↑0.1 4.9	↓0.1 75.7	↓0.7 4.1	↑0.2 76.0	↓3.2 1.6	↓0.1 75.7	↓4.3 0.5
ES	66.1	10.4	↑0.4 66.5	↑1.6 12.0	↑0.4 66.5	↑0.8 11.2	↑0.9 67.0	↓1.5 8.9	66.1	↓0.6 9.8	↓0.3 65.8	↑0.5 10.9	↑0.2 66.3	↑0.5 10.9	↑0.2 66.3	↑2.2 12.6	↑0.9 67.0	↓0.5 9.9
IT	69.6	12.7	↑0.4 70.0	↑0.7 13.4	↑0.3 69.9	↓0.1 12.6	↑0.3 69.9	↑0.3 13.0	↓0.2 69.4	↓0.4 12.3	69.6	12.7	↑0.1 69.7	↑0.4 13.1	↑0.1 69.7	↓6.9 5.8	↑0.7 70.3	↓11.7 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	0.0	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	72.6	3.8	↓1.5 71.1	↑11.1 14.9	↓1.5 71.1	↑10.6 14.4
Mistral-Nemo-Instruct-2407																		
EN	76.7	4.6	↓0.7 76.0	↓0.4 4.2	↑0.3 77.0	↓0.2 4.4	↑0.1 76.8	↑0.1 4.7	↑0.1 76.8	4.6	↓0.1 76.6	↓0.1 4.5	↓0.5 76.2	↑0.3 4.9	↓0.6 76.1	↓2.2 2.4	↓1.0 75.7	↓3.3 1.3
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	↓0.2 70.2	↓2.1 8.1	↓0.5 69.9	↓7.8 2.4	↓0.5 69.9	↑7.6 17.8
IT	67.7	12.6	↑0.2 67.9	↓0.5 12.1	67.7	12.6	↑0.2 67.9	↓0.5 12.1	↑0.2 67.9	↑0.4 13.0	↑0.2 67.9	↓0.1 12.5	↑0.3 68.0	↑0.3 12.9	↑0.5 68.2	↑0.2 12.8	↑0.5 68.2	↓9.0 3.6
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	99.6	0.0
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	↑2.3 12.1	70.1	↑4.8 14.6	↓0.6 69.5	↑4.1 13.9

Table 12: Comprehensive evaluation of Country bias mitigation on the Multilingual Hate Speech dataset across eight language models. We report hate speech prediction accuracy (Main) and True Positive Rate gap (TPR-Gap) between demographic groups for each model and debiasing approach. IMSAE (Five-Subsets) consistently achieves stronger bias reduction compared to monolingual and standard cross-lingual approaches.

Target	Baseline		SAL (EN)		SAL (ES)		SAL (IT)		SAL (PL)		SAL (PT)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (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	Main	TPR-Gap
mBERT-uncased																		
EN	86.7	4.4	86.7	↑0.7 5.1	86.7	4.4	86.7	↓0.1 4.3	↑0.1 86.8	↓0.1 4.3	86.7	4.4	86.7	4.4	86.7	↓4.3 0.1	↓0.1 86.6	↓4.3 0.1
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	↑0.9 64.6	↑0.1 3.7	63.7	↑2.3 5.9	↓0.3 63.4	↑0.7 4.3
IT	68.4	2.1	↓0.5 67.9	↓1.4 0.7	↓0.5 67.9	↓0.1 2.0	68.4	↓0.6 1.5	↓0.5 67.9	↓0.1 2.0	68.4	2.1	↓0.2 68.2	↓1.7 0.4	↑0.3 68.7	↑1.0 3.1	↓0.5 67.9	↓0.8 1.3
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	88.2	11.6	↓0.1 88.1	↑0.4 12.0	↓0.2 88.0	↓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	↓1.2 60.1	↑1.8 13.8	↑1.3 62.6	↓5.2 6.8	↑3.1 64.4	↑2.4 14.4
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	↑0.1 4.1	79.4	4.0	↓0.1 79.3	↑0.2 4.2	↓0.1 79.3	↓0.1 3.9	↑0.3 79.7	↓3.6 0.4	↑0.2 79.6	↓2.5 1.5
ES	70.7	5.5	↑0.3 71.0	↑1.0 6.5	70.7	↓0.6 4.9	↑0.5 71.2	↑0.9 6.4	70.7	↓1.1 4.4	↑0.3 71.0	↑1.0 6.5	↑0.3 71.0	↑1.0 6.5	↑0.8 71.5	↑2.9 8.4	70.7	↓0.3 5.2
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	↑0.2 69.6	3.7	↑0.2 69.6	3.7	↑0.5 69.9	↑0.1 3.8	↑0.7 70.1	↑3.0 6.7	↓0.5 68.9	↓2.3 1.4
PL	88.4	15.3	↓0.1 88.3	15.3	↓0.1 88.3	15.3	88.4	15.3	↓1.2 87.2	↓2.5 12.8	↓0.3 88.1	↓0.4 14.9	↓0.1 88.3	↓0.4 14.9	↓0.1 88.3	↑5.7 21.0	↓1.4 87.0	↓10.7 4.6
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	↓1.8 0.7	↓0.7 56.4	↓1.5 1.0	↓3.7 53.4	↑0.3 2.8
Llama-3.1-8B																		
EN	79.0	3.1	↑0.3 79.3	↑0.6 3.7	79.0	↑0.2 3.3	↓0.1 78.9	↑0.1 3.2	79.0	↑0.1 3.2	79.0	↓0.1 3.0	79.0	↑0.1 3.2	79.0	↑6.0 9.1	↑0.3 79.3	↓2.9 0.2
ES	72.9	1.4	↓0.2 72.7	↑3.1 4.5	↑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	↑0.7 2.1	↓0.7 72.2	↑2.4 3.8	72.9	↑0.8 2.2	↑1.0 73.9	↑1.7 3.1
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	↑0.2 65.8	↓0.4 0.7	↓0.3 65.3	1.1	↑0.2 65.8	↑0.1 1.2	↑0.7 66.3	↑3.4 4.5	↑0.7 66.3	↑1.1 2.2
PL	87.0	11.6	↓0.1 86.9	↑0.4 12.0	↓0.1 86.9	↓0.4 11.2	↑0.1 87.1	↑0.4 12.0	↓0.3 86.7	↓5.4 6.2	87.0	↑0.4 12.0	↑0.1 87.1	↑0.8 12.4	↓0.1 86.9	↑7.9 19.5	↓0.2 86.8	↓6.6 5.0
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 60.1	↓0.8 2.1	59.5	↓0.9 2.0	↓1.2 58.3	↑2.7 5.6
Llama-3.2-3B																		
EN	79.5	3.3	↑0.1 79.6	↑0.5 3.8	79.5	↓0.1 3.2	↑0.1 79.6	3.3	79.5	↑0.1 3.4	↑0.1 79.6	3.3	79.5	↓0.1 3.2	79.5	↓1.7 1.6	↑0.1 79.6	↓2.9 0.4
ES	68.3	2.4	↓0.3 68.0	↓0.7 1.7	↓0.5 67.8	↓1.1 1.3	↓0.5 67.8	2.4	↓0.3 68.0	↑1.3 3.7	↓0.5 67.8	↑1.2 3.6	68.3	2.4	↑0.2 68.5	↓0.4 2.0	↓0.5 67.8	↑0.3 2.7
IT	68.2	4.3	↓1.0 67.2	↓0.7 3.6	↓0.7 67.5	↓1.6 2.7	↓1.5 66.7	4.3	↓0.3 67.9	↓1.9 2.4	↓0.3 67.9	↓0.8 3.5	↓1.2 67.0	↓0.9 3.4	↓1.5 66.7	↓1.2 3.1	↓1.9 66.3	↑0.9 5.2
PL	87.8	8.2	↑0.3 88.1	↑5.8 14.0	↑0.2 88.0	↑5.8 14.0	↑0.4 88.2	↑0.4 8.6	↓0.8 87.0	↑1.2 9.4	↑0.4 88.2	↑5.8 14.0	↑0.2 88.0	↑5.8 14.0	↑0.3 88.1	↑15.1 23.3	↓0.7 87.1	↓5.3 2.9
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	↓0.6 55.8	↑0.7 11.8	↑1.9 58.3	↓2.9 8.2	↑0.7 57.1	↓0.5 10.6	56.4	↑1.0 12.1	↑1.3 57.7	↑2.3 13.4
Mistral-7B-Instruct-v0.3																		
EN	79.9	2.5	↑0.1 80.0	↑1.1 3.6	↓0.1 79.8	2.5	↑0.1 80.0	↑0.2 2.7	↑0.1 80.0	↑0.1 2.6	↑0.1 80.0	↑0.2 2.7	↑0.2 80.1	↑0.2 2.7	79.9	↓0.3 2.2	79.9	↓1.2 1.3
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	↓0.5 64.1	↑0.8 5.9	↓0.5 64.1	↑1.4 6.5	↓0.9 63.7	↓2.6 2.5	↓1.7 62.9	↓2.4 2.7
IT	66.3	2.5	↑0.4 66.7	↑0.1 2.6	↑0.2 66.5	↑0.2 2.7	↓1.0 65.3	↑1.7 4.2	↑0.2 66.5	↑0.2 2.7	66.3	↑3.2 5.7	66.3	↑2.5 5.0	↓0.3 66.0	↑4.5 7.0	↓0.5 65.8	↑2.0 4.5
PL	87.5	15.8	↑0.1 87.6	15.8	87.5	15.8	↑0.1 87.6	15.8	↑0.1 87.6	↓0.1 15.7	87.5	↓0.4 15.4	↑0.1 87.6	↑0.4 16.2	↓0.2 87.3	↓1.1 14.7	↓0.2 87.3	↓14.6 1.2
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	↓0.6 58.9	↑0.1 10.3	↓0.6 58.9	↓0.7 9.5	↓0.6 58.9	↑4.6 14.8
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	↑0.1 3.7	↓0.1 79.5	↓0.1 3.5	↑0.2 79.8	3.6	79.6	↑0.2 3.8	↑0.3 79.9	↓3.1 0.5	↑0.3 79.9	↑1.3 4.9
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	↓0.8 66.3	↓0.1 6.7	↓0.3 66.8	↑1.6 8.4	67.1	↓0.9 5.9	↓0.8 66.3	↓0.9 5.9
IT	65.6	5.7	↑0.2 65.8	5.7	↓0.5 65.1	↓0.8 4.9	↑0.7 66.3	5.7	65.6	5.7	65.6	5.7	65.6	↑0.8 6.5	↑0.2 65.8	↓3.1 2.6	↑1.4 67.0	↑1.3 7.0
PL	87.6	12.0	87.6	↓0.4 11.6	87.6	12.0	↑0.1 87.7	12.0	↓0.7 86.9	↑0.2 12.2	87.6	↓0.1 11.9	87.6	12.0	87.6	↓3.3 8.7	↓0.8 86.8	↓9.3 2.7
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	57.7	8.3	57.7	↑1.6 9.9	↑1.2 58.9	↓0.8 7.5
Mistral-Nemo-Base-2407																		
EN	79.0	3.9	↓0.7 78.3	↑0.2 4.1	↓0.7 78.3	↓0.5 3.4	↑0.2 79.2	↓0.4 3.5	↑0.3 79.3	↓0.2 3.7	↓0.1 78.9	↓0.4 3.5	↓0.6 78.4	↓0.1 3.8	↓0.1 78.9	↑0.7 4.6	↓0.1 78.9	↓3.6 0.3
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	↓0.7 71.5	↑1.3 4.4	↓1.0 71.2	↑8.0 11.1	↓1.2 71.0	↑1.0 4.1
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 65.6	↓0.8 0.8	↓0.2 65.6	↑3.3 4.9	↑0.9 66.7	↑2.7 4.3
PL	88.5	16.1	88.5	↑0.4 16.5	88.5	↑0.4 16.5	↑0.1 88.6	↑0.4 16.5	↓0.3 88.2	↓1.3 14.8	88.5	16.1	↑0.2 88.7	↑1.2 17.3	↑0.1 88.6	↓10.3 5.8	↓0.3 88.2	↓15.5 0.6
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	↓0.6 62.6	↓1.3 12.7	↓1.2 62.0	↓6.7 7.3	↑0.6 63.8	↑3.2 17.2
Mistral-Nemo-Instruct-2407																		
EN	79.5	2.9	↓0.2 79.3	↑0.6 3.5	↓0.1 79.4	↓0.2 2.7	↓0.1 79.4	2.9	79.5	2.9	↓0.1 79.4	↓0.1 2.8	↓0.1 79.4	↓0.1 2.8	↓0.2 79.3	↓2.3 0.6	↓0.2 79.3	↓2.4 0.5
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	↓0.8 70.2	↑1.6 7.0	↓0.3 70.7	↓2.0 3.4	↓0.3 70.7	↓2.5 2.9
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	65.8	4.9	65.8	↓2.5 2.4	↓0.5 65.3	↓3.5 1.4
PL	87.6	10.7	↑0.2 87.8	↑0.1 10.8	87.6	10.7	↑0.1 87.7	10.7	↓0.5 87.1	↑3.7 14.4	↑0.2 87.8	↑0.1 10.8	↑0.2 87.8	↑0.1 10.8	↑0.3 87.9	↓1.6 9.1	↓0.5 87.1	↓8.4 2.3
PT	64.4	16.8	↑0.6 65.0	↓2.5 14.3	64.4	16.8	64.4	16.8	↑1.2 65.6	↓1.2 15.6	↓0.6 63.8	↑0.1 16.9	64.4	↓1.2 15.6	↓0.6 63.8	↓1.8 15.0	↓1.8 62.6	↓2.3 14.5

Table 13: Comprehensive evaluation of gender bias mitigation on the Multilingual Hate Speech dataset across eight language models. We report hate speech prediction accuracy (Main) and True Positive Rate gap (TPR-Gap) between demographic groups for each model and debiasing approach. IMSAE (Five-Subsets) consistently achieves stronger bias reduction compared to monolingual and standard cross-lingual approaches.

Target	Baseline		SAL (EN)		SAL (ES)		SAL (IT)		SAL (PL)		SAL (PT)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (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	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	↓0.3 86.5	↑0.8 4.9	↓0.1 86.7	↑1.9 6.0	↓2.0 84.8	↓4.1 0.0
ES	63.7	10.2	↑0.7 64.4	↓0.8 9.4	↑0.4 64.1	↓0.1 10.1	↑0.7 64.4	10.2	↑0.2 63.9	10.2	63.7	10.2	↑0.4 64.1	↓0.1 10.1	63.7	↓9.2 1.0	↓0.3 63.4	↓1.7 8.5
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	↓0.2 68.2	↓0.6 32.0	↓0.5 67.9	↑23.2 55.8	↓0.5 67.9	↓25.2 7.4
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	91.3	6.2	91.3	↓4.7 1.5	↓0.2 91.1	↓6.2 0.0
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	↑1.4 2.4	↑0.7 62.0	↑5.0 6.0	↑1.3 62.6	↑11.3 12.3
Llama3-8B																		
EN	79.3	4.4	↓1.7 77.6	↓1.7 2.7	79.3	↑0.1 4.5	↓0.1 79.2	↑0.2 4.6	79.3	↑0.2 4.6	↑0.1 79.4	↑0.1 4.5	79.3	↑0.1 4.5	↑0.1 79.4	↓4.3 0.1	↓1.6 77.7	↓2.4 2.0
ES	70.7	7.0	↑0.5 71.2	↑0.9 7.9	70.7	7.0	↑0.5 71.2	↑0.8 7.8	↑0.3 71.0	↓0.1 6.9	↑0.8 71.5	↑1.8 8.8	↑0.5 71.2	↑1.9 8.9	↓0.2 70.5	↑0.7 7.7	↑0.5 71.2	↑0.5 7.5
IT	69.9	43.3	↓0.3 69.6	↑0.1 43.4	↓0.3 69.6	↓0.5 42.8	↓1.5 68.4	↓1.7 41.6	69.9	43.3	↓0.3 69.6	↑0.7 44.0	↓0.5 69.4	↑0.2 43.5	↓0.5 69.4	↓6.4 36.9	↓1.7 68.2	↓32.7 10.6
PL	91.0	16.3	91.0	↓0.6 15.7	↓0.1 90.9	↓0.6 15.7	↑0.1 91.1	16.3	91.0	↓0.6 15.7	91.0	↓0.6 15.7	91.0	↓0.6 15.7	91.0	↑10.5 26.8	↑0.1 91.1	↓4.2 12.1
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	↓0.7 56.4	↓1.7 16.3	↓1.3 55.8	↓4.4 13.6	↓1.3 55.8	↓2.0 16.0
Llama-3.1-8B																		
EN	79.0	3.1	↓1.8 77.2	↓2.2 0.9	79.0	↑0.1 3.2	79.0	↓0.2 2.9	79.0	3.1	79.0	3.1	79.0	↑0.1 3.2	↑0.1 79.1	↑0.1 3.2	↓2.0 77.0	↓0.6 2.5
ES	72.9	1.9	↓0.2 72.7	↓0.2 1.7	↓0.7 72.2	1.9	↓0.2 72.7	↓0.4 1.5	↓0.2 72.7	↑0.1 2.0	↓1.4 71.5	↓0.5 1.4	↑0.5 73.4	↑0.7 2.6	↓0.2 72.7	↑2.3 4.2	72.9	↑2.9 4.8
IT	66.7	33.0	↑0.3 67.0	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	66.7	33.0	↑1.0 67.7	↓12.5 20.5	↑1.0 67.7	↑1.6 34.6
PL	90.4	14.5	↓0.1 90.3	↓0.6 13.9	90.4	14.5	90.4	14.5	↓0.1 90.3	↓0.6 13.9	90.4	↑0.6 15.1	↓0.1 90.3	↓0.6 13.9	90.4	↓9.9 4.6	90.4	↓9.5 5.0
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	↓3.4 8.7	↑1.2 60.7	↑2.3 14.4	↓0.6 58.9	↓4.3 7.8
Llama-3.2-3B																		
EN	79.5	3.6	↓2.0 77.5	↓2.9 0.7	79.5	↑0.2 3.8	79.5	3.6	79.5	3.6	↑0.1 79.6	↑0.1 3.7	↓0.1 79.4	↑0.3 3.9	79.5	↓3.5 0.1	↓2.0 77.5	↓1.7 1.9
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	↓0.5 67.8	↑0.8 4.4	68.3	↑0.5 4.1	↑0.5 68.8	↑1.4 5.0
IT	68.9	38.3	↓1.2 67.7	↓1.6 36.7	↓1.0 67.9	↓1.1 37.2	↓1.7 67.2	↓1.1 37.2	↓0.7 68.2	↓0.6 37.7	↓0.7 68.2	↓0.6 37.7	↓1.2 67.7	↓1.6 36.7	↓1.0 67.9	↓0.4 37.9	↓1.2 67.7	↓17.7 20.6
PL	90.9	16.3	90.9	16.3	↓0.1 90.8	16.3	↓0.3 90.6	↓0.6 15.7	90.9	↑0.6 16.9	90.9	16.3	90.9	16.3	↓0.2 90.7	↓10.3 6.0	↓0.1 90.8	↑19.4 35.7
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	↑1.6 7.0	56.4	↓0.4 5.0	↓0.6 55.8	↑1.8 7.2
Mistral-7B-Instruct-v0.3																		
EN	79.9	2.8	↓3.1 76.8	↓0.9 1.9	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	↓0.1 79.8	↑0.2 3.0	↓0.3 79.6	↑0.5 3.3	↓3.1 76.8	↓1.1 1.7
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	↓0.7 63.9	↓0.2 7.8	↓0.5 64.1	↑0.1 8.1	↑0.3 64.9	↓1.6 6.4	↓1.7 62.9	↓1.2 6.8
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	66.0	39.4	66.0	39.4	↓0.2 65.8	↓0.4 39.0	↓0.9 65.1	↓14.0 25.4	↓0.7 65.3	↑2.8 42.2
PL	91.0	18.7	↑0.3 91.3	↑0.6 19.3	91.0	18.7	91.0	18.7	91.0	↓0.6 18.1	↓0.1 90.9	18.7	↑0.1 91.1	↑0.6 19.3	↑0.2 91.2	↑1.9 20.6	↓0.1 90.9	↓8.4 10.3
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	59.5	8.8	↓0.6 58.9	↓5.2 3.6	↓1.2 58.3	↑4.3 13.1
Mistral-7B-v0.3																		
EN	79.5	4.9	↓1.6 77.9	↓1.8 3.1	79.5	↓0.1 4.8	79.5	↑0.1 5.0	79.5	↓0.3 4.6	79.5	↑0.1 5.0	79.5	↓0.1 4.8	↑0.1 79.6	↓3.7 1.2	↓1.6 77.9	↓2.0 2.9
ES	67.1	3.5	67.1	↓0.1 3.4	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	↓0.5 66.6	↓0.8 2.7	↓1.2 65.9	↑0.1 3.6	↓0.3 66.8	↑5.6 9.1
IT	66.3	41.1	↓0.3 66.0	↓0.5 40.6	↑0.2 66.5	↑0.1 41.2	66.3	↓1.2 39.9	↓0.3 66.0	↓0.1 41.0	↓0.3 66.0	↓0.9 40.2	↑0.2 66.5	↑0.1 41.2	↓0.5 65.8	↓22.8 18.3	↑0.7 67.0	↓16.1 25.0
PL	90.4	16.9	90.4	↓0.6 16.3	↓0.2 90.2	↓0.6 16.3	↓0.2 90.2	↓0.6 16.3	↑0.1 90.5	↑0.6 17.5	90.4	16.9	↓0.1 90.3	↓0.6 16.3	90.4	↓10.8 6.1	↑0.1 90.5	↓15.1 1.8
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	57.7	2.9	↑0.6 58.3	↑1.8 4.7	↑0.6 58.3	↓0.4 2.5
Mistral-Nemo-Base-2407																		
EN	79.3	5.3	↓3.3 76.0	↓1.4 3.9	↓1.0 78.3	↑0.4 5.7	↓0.8 78.5	↓0.1 5.2	↓1.4 77.9	↑0.5 5.8	↓1.1 78.2	5.3	↓0.8 78.5	↓0.4 4.9	↓0.5 78.8	↓5.2 0.1	↓3.5 75.8	↓4.6 0.7
ES	72.2	8.9	↓0.5 71.7	↓1.4 7.5	72.2	↓2.1 6.8	↓0.2 72.0	↓0.9 8.0	↓0.2 72.0	↓0.9 8.0	↓1.0 71.2	↓0.3 8.6	72.2	↓1.1 7.8	72.2	↓7.3 1.6	↓0.2 72.0	↓7.8 1.1
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	↑0.9 66.0	↑1.0 41.6	↑0.5 65.6	↓22.7 17.9	↑1.6 66.7	↓22.9 17.7
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	↓1.2 18.7	↓0.3 91.0	↑4.4 24.3	↓0.2 91.1	↓16.9 3.0
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	↑0.6 63.8	↓1.3 6.3	↑0.6 63.8	↓1.3 6.3	63.2	↓2.2 5.4	↑1.2 64.4	↓6.3 1.3	↑0.6 63.8	↑2.0 9.6
Mistral-Nemo-Instruct-2407																		
EN	79.5	5.3	↓2.8 76.7	↓3.3 2.0	↓0.3 79.2	↑0.3 5.6	↓0.2 79.3	↓0.2 5.1	↓0.2 79.3	↓0.1 5.2	↑0.2 79.7	5.3	↓0.3 79.2	↑0.1 5.4	↓0.2 79.3	↑1.5 6.8	↓2.6 76.9	↓4.8 0.5
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	↓0.3 70.7	↓0.5 3.9	↓1.2 69.8	↑5.2 9.6	↓0.5 70.5	↓0.8 3.6
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	↓0.5 64.8	↑0.3 45.1	↑0.3 65.6	↓0.7 44.1	↑0.3 65.6	↓43.7 1.1
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	90.9	↑0.6 18.7	↓0.2 90.7	↑8.5 26.6	↑0.1 91.0	↑3.0 21.1
PT	64.4	3.9	↑0.6 65.0	↓0.7 3.2	↓0.6 63.8	↓1.4 2.5	64.4	↓2.4 1.5	↑1.2 65.6	↑4.5 8.4	64.4	3.9	↑1.2 65.6	↓0.3 3.6	↓0.6 63.8	↑3.5 7.4	↑0.6 65.0	↑4.2 8.1

Table 14: Comprehensive evaluation of race bias mitigation on the Multilingual Hate Speech dataset across eight language models. We report hate speech prediction accuracy (Main) and True Positive Rate gap (TPR-Gap) between demographic groups for each model and debiasing approach. IMSAE (Five-Subsets) consistently achieves stronger bias reduction compared to monolingual and standard cross-lingual approaches.

B.3 Crosslingual Baseline Performance

This section evaluates the performance of three state-of-the-art crosslingual debiasing methods: SAL (Table 15), INLP (Table 16), and SentenceDebias (Table 17) on age debiasing across different source-target language pairs. Due to page limitations, we only include detailed results for age debiasing here; however, the complete results for country, gender, and race debiasing across all methods have been uploaded to our GitHub repository for reference.

Target	Baseline		SAL (EN)		SAL (ES)		SAL (IT)		SAL (PL)		SAL (PT)	
	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext
mBERT-uncased												
EN	86.7	9.1	↑0.3 87.0	↑0.4 9.5	↓0.5 86.2	↓0.3 8.8	86.7	↑0.1 9.2	↓0.5 86.2	↓0.1 9.0	86.7	↓0.1 9.0
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
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
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
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
Llama3-8B												
EN	79.6	7.8	↑0.8 80.4	↑0.6 8.4	79.6	↓0.4 7.4	↑0.1 79.7	↓0.1 7.7	↑0.1 79.7	↓0.2 7.6	↑0.3 79.9	↓0.3 7.5
ES	70.7	11.7	↑0.3 71.0	↑0.6 12.3	↓0.2 70.5	↓0.6 11.1	↑0.3 71.0	↑0.6 12.3	70.7	11.7	↑0.5 71.2	↑0.5 12.2
IT	69.9	8.0	↓0.3 69.6	↑0.3 8.3	69.9	8.0	↓0.5 69.4	↓0.7 7.3	↑0.2 70.1	↓0.9 7.1	↓0.3 69.6	↓0.3 7.7
PL	91.0	17.2	91.0	↓0.7 16.5	↓0.1 90.9	↓0.6 16.6	91.0	↓0.1 17.1	↓1.2 89.8	↓7.8 9.4	91.0	↓0.8 16.4
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
Llama-3.1-8B												
EN	79.7	6.5	↑0.7 80.4	↑0.7 7.2	79.7	↓0.1 6.4	79.7	↑0.2 6.7	↑0.1 79.8	↑0.1 6.6	79.7	↓0.2 6.3
ES	72.9	8.6	↓0.2 72.7	↑0.1 8.7	↓0.2 72.7	↓1.0 7.6	72.9	↓1.1 7.5	↓0.5 72.4	↓0.5 8.1	72.9	↓2.2 6.4
IT	66.5	9.5	↑0.2 66.7	↓1.2 8.3	↑0.5 67.0	↓1.0 8.5	↑0.7 67.2	↓0.5 9.0	↑0.5 67.0	↓1.0 8.5	↑0.5 67.0	↓0.1 9.4
PL	90.4	15.2	↓0.1 90.3	↓0.6 14.6	90.4	15.2	90.4	15.2	↓0.4 90.0	↓6.9 8.3	↓0.1 90.3	↓0.6 14.6
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
Llama-3.2-3B												
EN	79.7	6.2	↑0.6 80.3	↑0.8 7.0	↑0.1 79.8	↑0.3 6.5	79.7	6.2	↑0.1 79.8	↑0.1 6.3	↑0.1 79.8	↑0.3 6.5
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	↓0.5 67.8	↑0.9 13.0
IT	67.7	5.0	67.7	↑0.3 5.3	↑0.5 68.2	↑2.0 7.0	↓0.2 67.5	↑0.7 5.7	↑0.2 67.9	↑1.8 6.8	↑0.2 67.9	↑1.5 6.5
PL	90.9	17.1	↓0.2 90.7	↑0.1 17.2	↓0.2 90.7	↓0.6 16.5	↓0.3 90.6	↓0.6 16.5	↓0.9 90.0	↓1.4 15.7	90.9	17.1
PT	56.4	8.9	↑0.7 57.1	↑2.8 11.7	↑0.7 57.1	↑0.8 9.7	↑0.7 57.1	↑0.8 9.7	↓1.2 55.2	↑1.5 10.4	56.4	↑1.7 10.6
Mistral-7B-Instruct-v0.3												
EN	79.4	7.1	↑1.1 80.5	↑1.3 8.4	↑0.2 79.6	↑0.1 7.2	↑0.3 79.7	↑0.1 7.2	↑0.1 79.5	↑0.3 7.4	↑0.2 79.6	↓0.1 7.0
ES	64.6	7.2	↓0.7 63.9	↓0.2 7.0	↑0.8 65.4	↑1.6 8.8	↓0.2 64.4	↓0.8 6.4	↓0.5 64.1	↑0.5 7.7	↓0.5 64.1	↓1.5 5.7
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	↑0.5 66.5	↑0.1 3.6	↓0.2 65.8	↓0.2 3.3
PL	91.0	19.5	91.0	19.5	↑0.2 91.2	↑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
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
Mistral-7B-v0.3												
EN	79.8	7.1	↑0.6 80.4	↑0.4 7.5	↓0.1 79.7	7.1	79.8	↓0.3 6.8	↓0.1 79.7	↑0.1 7.2	↓0.2 79.6	↓0.3 6.8
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	↓0.3 66.8	↑1.8 14.6	67.1	↑0.2 13.0
IT	65.8	5.6	65.8	↓0.9 4.7	↓0.2 65.6	↑0.5 6.1	↑0.2 66.0	↓0.6 5.0	65.8	↓0.1 5.5	65.8	↓0.8 4.8
PL	90.4	17.8	↑0.3 90.7	↑0.4 18.2	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
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
Mistral-Nemo-Base-2407												
EN	78.9	6.4	↑0.4 79.3	↑1.1 7.5	↓0.3 78.6	↑0.1 6.5	↓0.5 78.4	↑0.2 6.6	78.9	↓0.1 6.3	↓0.3 78.6	↓0.3 6.1
ES	72.2	16.3	↓1.2 71.0	↑1.3 17.6	↑0.7 72.9	↑1.8 18.1	72.2	↑2.0 18.3	↑0.2 72.4	↑0.6 16.9	↓0.5 71.7	↓1.0 15.3
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
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
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
Mistral-Nemo-Instruct-2407												
EN	79.3	5.9	↑0.3 79.6	↑1.1 7.0	↓0.2 79.1	↑0.1 6.0	↓0.2 79.1	↑0.2 6.1	↓0.2 79.1	↑0.1 6.0	↓0.3 79.0	↑0.5 6.4
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
IT	65.3	8.0	↓0.2 65.1	↓0.6 7.4	↓0.2 65.1	↓0.1 7.9	65.3	↓0.4 7.6	65.3	8.0	65.3	↑0.5 8.5
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
PT	64.4	1.5	64.4	1.5	64.4	1.5	↓0.6 63.8	↑1.8 3.3	↑1.2 65.6	↑2.1 3.6	64.4	1.5

Table 15: Performance comparison of SAL age debiasing approaches on the Multilingual Hate Speech dataset, showing effectiveness across different source-target language pairs. These baseline methods show limited cross-lingual transfer compared to IMSAE's multilingual approach.

Target	Baseline		EN - INLP		ES - INLP		IT - INLP		PL - INLP		PT - INLP	
	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext
mBERT-uncased												
EN	86.7	9.1	↑0.3 87.0	↑0.7 9.8	↑0.1 86.8	↓0.1 9.0	86.7	↑0.1 9.2	86.7	9.1	↑0.1 86.8	↓0.2 8.9
ES	63.7	12.9	↑0.4 64.1	↓0.5 12.4	63.7	↑1.0 13.9	↑0.4 64.1	↑0.1 13.0	↑0.2 63.9	↓0.3 12.6	63.7	↑0.6 13.5
IT	68.2	3.6	68.2	3.6	↑0.2 68.4	↓0.2 3.4	↑0.2 68.4	↓0.2 3.4	68.2	3.6	↑0.2 68.4	↑0.9 4.5
PL	91.3	8.8	91.3	↓0.7 8.1	91.3	8.8	91.3	8.8	↓0.6 90.7	↓1.3 7.5	91.3	8.8
PT	61.3	17.6	↑0.7 62.0	↓1.5 16.1	↓0.6 60.7	↓0.4 17.2	↓1.2 60.1	↑1.2 18.8	61.3	17.6	61.3	↑2.0 19.6
Llama3-8B												
EN	79.6	7.8	↓0.4 79.2	↓0.6 7.2	79.6	↓0.2 7.6	79.6	↓0.2 7.6	↑0.2 79.8	↑0.1 7.9	↑0.1 79.7	↓0.2 7.6
ES	70.7	11.7	70.7	11.7	70.7	11.7	70.7	11.7	70.7	11.7	70.7	11.7
IT	69.9	8.0	69.9	8.0	69.9	8.0	↓0.5 69.4	↑1.2 9.2	↓0.3 69.6	↑1.0 9.0	69.9	8.0
PL	91.0	17.2	↓0.1 90.9	↓0.6 16.6	↓0.1 90.9	↓0.6 16.6	↓0.1 90.9	↓0.6 16.6	↑0.1 91.1	↑0.6 17.8	91.0	17.2
PT	57.1	2.1	↓0.7 56.4	↑0.9 3.0	57.1	2.1	57.1	2.1	57.1	2.1	↑3.0 60.1	↑7.1 9.2
Llama-3.1-8B												
EN	79.7	6.5	↓0.6 79.1	↓0.1 6.4	↑0.1 79.8	6.5	↓0.1 79.6	6.5	79.7	↓0.2 6.3	79.7	6.5
ES	72.9	8.6	↓0.5 72.4	↓0.3 8.3	72.9	8.6	72.9	↓2.3 6.3	↓0.9 72.0	↑0.3 8.9	72.9	↓1.2 7.4
IT	66.5	9.5	66.5	9.5	66.5	9.5	↑0.5 67.0	↓0.1 9.4	66.5	9.5	66.5	9.5
PL	90.4	15.2	90.4	15.2	↓0.1 90.3	↓0.6 14.6	90.4	15.2	90.4	15.2	↓0.1 90.3	↓0.6 14.6
PT	59.5	2.8	59.5	2.8	59.5	2.8	↑0.6 60.1	↑0.5 3.3	59.5	2.8	59.5	↑1.7 4.5
Llama-3.2-3B												
EN	79.7	6.2	↑0.1 79.8	↑0.1 6.3	79.7	↑0.2 6.4	79.7	↑0.1 6.3	79.7	↑0.2 6.4	79.7	↑0.2 6.4
ES	68.3	12.1	68.3	12.1	68.3	12.1	68.3	12.1	↓0.3 68.0	↓0.8 11.3	↓0.3 68.0	↓0.8 11.3
IT	67.7	5.0	↑0.5 68.2	↑0.2 5.2	67.7	5.0	67.7	↑0.7 5.7	↑0.5 68.2	↑1.3 6.3	↑0.7 68.4	↑1.5 6.5
PL	90.9	17.1	↓0.2 90.7	↑0.1 17.2	90.9	17.1	↓0.1 90.8	17.1	↓0.2 90.7	↑1.4 18.5	↓0.2 90.7	↓0.6 16.5
PT	56.4	8.9	↑0.7 57.1	↑0.8 9.7	↑1.3 57.7	↑2.0 10.9	56.4	↑1.9 10.8	↑0.7 57.1	↑0.8 9.7	↑1.3 57.7	↑3.9 12.8
Mistral-7B-Instruct-v0.3												
EN	79.4	7.1	↑0.7 80.1	↓0.1 7.0	79.4	↑0.1 7.2	79.4	↑0.1 7.2	↑0.2 79.6	↓0.6 6.5	↑0.2 79.6	↓0.4 6.7
ES	64.6	7.2	64.6	7.2	64.6	↓1.2 6.0	64.6	7.2	↑0.3 64.9	↓0.5 6.7	↑0.3 64.9	↓0.5 6.7
IT	66.0	3.5	66.0	3.5	66.0	3.5	↓0.2 65.8	↓0.5 3.0	↓0.7 65.3	↓0.4 3.1	↓0.4 65.6	↓0.7 2.8
PL	91.0	19.5	91.0	19.5	91.0	19.5	↓0.1 90.9	19.5	↓0.1 90.9	19.5	↓0.1 90.9	19.5
PT	59.5	17.2	↓0.6 58.9	↓1.0 16.2	59.5	17.2	59.5	17.2	59.5	17.2	↓2.4 57.1	↓4.2 13.0
Mistral-7B-v0.3												
EN	79.8	7.1	↑0.3 80.1	↓0.2 6.9	↓0.1 79.7	↓0.2 6.9	79.8	7.1	79.8	↓0.1 7.0	79.8	↓0.1 7.0
ES	67.1	12.8	67.1	12.8	67.1	12.8	↓0.3 66.8	↓0.8 12.0	↑0.2 67.3	↓0.7 12.1	↓0.5 66.6	↓0.8 12.0
IT	65.8	5.6	65.8	↓0.1 5.5	↓0.2 65.6	↑0.4 6.0	↑0.5 66.3	↓0.2 5.4	65.8	↓0.1 5.5	↑0.2 66.0	↓1.3 4.3
PL	90.4	17.8	90.4	17.8	90.4	17.8	↓0.2 90.2	↓0.5 17.3	↑0.2 90.6	↓0.1 17.7	90.4	17.8
PT	57.7	12.4	57.7	12.4	57.7	12.4	57.7	12.4	57.7	12.4	↓1.3 56.4	↑0.7 13.1
Mistral-Nemo-Base-2407												
EN	78.9	6.4	↑0.2 79.1	↑0.1 6.5	78.9	6.4	↑0.3 79.2	↓0.1 6.3	↓0.1 78.8	↑0.3 6.7	78.9	↓0.3 6.1
ES	72.2	16.3	↓0.5 71.7	↑1.0 17.3	72.2	↑2.1 18.4	↑0.2 72.4	↑0.6 16.9	72.2	↑1.1 17.4	↓0.2 72.0	↑0.5 16.8
IT	64.8	5.8	64.8	5.8	↓0.2 64.6	↑0.5 6.3	↑0.5 65.3	↑0.1 5.9	64.8	5.8	↓0.2 64.6	↑0.5 6.3
PL	91.3	20.9	↓0.1 91.2	↓0.6 20.3	91.3	20.9	↓0.1 91.2	↓0.6 20.3	↓0.2 91.1	↓1.3 19.6	↓0.1 91.2	↓0.6 20.3
PT	63.2	8.8	63.2	8.8	63.2	8.8	63.2	8.8	63.2	8.8	63.2	↓3.0 5.8
Mistral-Nemo-Instruct-2407												
EN	79.3	5.9	↓0.3 79.0	↓0.3 5.6	↓0.1 79.2	↑0.3 6.2	↓0.2 79.1	↑0.3 6.2	79.3	↑0.5 6.4	↓0.2 79.1	↓0.1 5.8
ES	71.0	10.2	↓0.3 70.7	↑0.3 10.5	↓0.3 70.7	↑0.3 10.5	↓0.3 70.7	↑0.6 10.8	↓0.3 70.7	↑0.3 10.5	↓0.3 70.7	↑0.6 10.8
IT	65.3	8.0	65.3	8.0	65.3	8.0	65.3	8.0	65.3	8.0	↓0.2 65.1	↓0.1 7.9
PL	90.9	18.9	90.9	18.9	90.9	18.9	90.9	18.9	90.9	18.9	90.9	18.9
PT	64.4	1.5	↓1.2 63.2	↑3.6 5.1	↑0.6 65.0	↑0.9 2.4	64.4	1.5	↓0.6 63.8	↑4.0 5.5	↑0.6 65.0	↑1.0 2.5

Table 16: Performance comparison of INLP age debiasing approaches on the Multilingual Hate Speech dataset, showing effectiveness across different source-target language pairs. These baseline methods show limited cross-lingual transfer compared to IMSAE's multilingual approach.

Target	Baseline		EN - SentenceDebias		ES - SentenceDebias		IT - SentenceDebias		PL - SentenceDebias		PT - SentenceDebias	
	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext
mBERT-uncased												
EN	86.7	9.1	↑0.3 87.0	↑0.9 10.0	86.7	9.1	86.7	9.1	86.7	9.1	↑0.1 86.8	9.1
ES	63.7	12.9	↑0.2 63.9	↑1.4 14.3	↑1.4 65.1	↑1.8 14.7	↑0.2 63.9	↑0.3 13.2	63.7	↑0.6 13.5	63.7	↑0.6 13.5
IT	68.2	3.6	↓0.3 67.9	↑0.3 3.9	↓0.3 67.9	↓0.7 2.9	↑0.2 68.4	↑0.9 4.5	68.2	↓1.0 2.6	68.2	3.6
PL	91.3	8.8	↓0.1 91.2	↓1.3 7.5	↓0.1 91.2	↓0.7 8.1	91.3	8.8	↓0.2 91.1	↓0.7 8.1	91.3	8.8
PT	61.3	17.6	↓0.6 60.7	↓0.4 17.2	↓0.6 60.7	↓0.4 17.2	↑0.7 62.0	↑1.8 19.4	↓0.6 60.7	↓0.4 17.2	↑1.3 62.6	↑2.2 19.8
Llama3-8B												
EN	79.6	7.8	↑0.5 80.1	↑0.6 8.4	79.6	↓0.2 7.6	↑0.1 79.7	↓0.3 7.5	79.6	7.8	79.6	↑0.1 7.9
ES	70.7	11.7	70.7	11.7	70.7	11.7	70.7	11.7	70.7	11.7	70.7	11.7
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	8.0	69.9	8.0	↓0.3 69.6	↑0.3 8.3
PL	91.0	17.2	91.0	↓0.7 16.5	↓0.1 90.9	↓0.6 16.6	↓0.1 90.9	17.2	↓0.2 90.8	↑0.6 17.8	91.0	↓0.7 16.5
PT	57.1	2.1	57.1	2.1	↑0.6 57.7	↑0.8 2.9	57.1	2.1	57.1	2.1	↑0.6 57.7	↑5.9 8.0
Llama-3.1-8B												
EN	79.7	6.5	↑0.4 80.1	↑1.3 7.8	↓0.1 79.6	↓0.3 6.2	79.7	6.5	79.7	↑0.1 6.6	↑0.1 79.8	↑0.1 6.6
ES	72.9	8.6	↓0.2 72.7	↓1.9 6.7	↓0.2 72.7	↓0.9 7.7	↑0.3 73.2	↓0.5 8.1	↓0.5 72.4	↓0.3 8.3	72.9	8.6
IT	66.5	9.5	↑0.5 67.0	↓1.0 8.5	↑0.2 66.7	↓0.5 9.0	↑0.5 67.0	↓0.6 8.9	66.5	9.5	↑0.2 66.7	↓1.3 8.2
PL	90.4	15.2	90.4	15.2	90.4	15.2	90.4	15.2	90.4	↑0.7 15.9	↓0.1 90.3	↓0.6 14.6
PT	59.5	2.8	↑0.6 60.1	↑0.5 3.3	59.5	2.8	59.5	2.8	↑0.6 60.1	↑0.5 3.3	↓2.4 57.1	↑0.3 3.1
Llama-3.2-3B												
EN	79.7	6.2	↑0.1 79.8	↑1.2 7.4	79.7	↑0.2 6.4	↓0.1 79.6	↑0.2 6.4	79.7	↑0.2 6.4	↑0.1 79.8	↑0.2 6.4
ES	68.3	12.1	68.3	12.1	↓0.7 67.6	↓0.3 11.8	↓0.3 68.0	↓0.2 11.9	↓0.3 68.0	↑0.2 12.3	↓0.5 67.8	↓0.1 12.0
IT	67.7	5.0	67.7	↑1.2 6.2	↑0.2 67.9	↑0.8 5.8	↓0.2 67.5	↑2.3 7.3	↑0.5 68.2	↑0.2 5.2	↑0.2 67.9	↑0.5 5.5
PL	90.9	17.1	90.9	17.1	↓0.2 90.7	↑0.1 17.2	↓0.1 90.8	17.1	↓0.5 90.4	↑1.2 18.3	90.9	17.1
PT	56.4	8.9	↑1.3 57.7	↑2.0 10.9	↑1.3 57.7	↑4.0 12.9	↓0.6 55.8	↑2.7 11.6	↑0.7 57.1	↑2.8 11.7	↑0.7 57.1	↑1.2 10.1
Mistral-7B-Instruct-v0.3												
EN	79.4	7.1	↑0.6 80.0	↑0.5 7.6	79.4	↑0.2 7.3	↑0.2 79.6	↓0.4 6.7	↑0.2 79.6	↓0.4 6.7	↑0.6 80.0	↓0.1 7.0
ES	64.6	7.2	↓0.5 64.1	↑1.0 8.2	64.6	↑2.4 9.6	↑0.3 64.9	↓0.5 6.7	↑0.3 64.9	↓0.5 6.7	↑1.0 65.6	↑0.6 7.8
IT	66.0	3.5	66.0	3.5	↓0.4 65.6	↓0.7 2.8	↓0.7 65.3	↓0.1 3.4	↑0.5 66.5	↑0.1 3.6	66.0	↓0.6 2.9
PL	91.0	19.5	91.0	19.5	↑0.2 91.2	↑0.6 20.1	↑0.1 91.1	19.5	↑0.3 91.3	19.5	91.0	19.5
PT	59.5	17.2	59.5	17.2	59.5	17.2	↓0.6 58.9	↓1.0 16.2	59.5	17.2	↓3.1 56.4	↓6.3 10.9
Mistral-7B-v0.3												
EN	79.8	7.1	↑0.5 80.3	↓0.4 6.7	79.8	↓0.1 7.0	79.8	↓0.1 7.0	79.8	↑0.2 7.3	↓0.1 79.7	↓0.2 6.9
ES	67.1	12.8	↑0.2 67.3	↓1.7 11.1	↓0.5 66.6	↓2.7 10.1	↓0.3 66.8	↓0.8 12.0	↓0.5 66.6	↓1.7 11.1	67.1	12.8
IT	65.8	5.6	↑0.2 66.0	↑0.4 6.0	65.8	↓0.1 5.5	↑0.5 66.3	5.6	↑0.5 66.3	↓1.0 4.6	65.8	↓0.1 5.5
PL	90.4	17.8	90.4	17.8	90.4	17.8	90.4	17.8	90.4	↑0.1 17.9	↑0.1 90.5	↓0.1 17.7
PT	57.7	12.4	57.7	12.4	57.7	12.4	57.7	12.4	↑0.6 58.3	↑0.8 13.2	57.7	↓2.1 10.3
Mistral-Nemo-Base-2407												
EN	78.9	6.4	↑0.7 79.6	↑1.0 7.4	↓0.5 78.4	↑0.1 6.5	↓0.1 78.8	↓0.1 6.3	78.9	6.4	↓0.2 78.7	↑0.1 6.5
ES	72.2	16.3	72.2	↑1.1 17.4	↓0.2 72.0	↑1.4 17.7	72.2	↑1.1 17.4	↓0.5 71.7	↑3.2 19.5	↑0.5 72.7	↑1.2 17.5
IT	64.8	5.8	64.8	5.8	64.8	5.8	↑1.2 66.0	↑2.5 8.3	64.8	5.8	↑0.3 65.1	↓0.4 5.4
PL	91.3	20.9	↓0.3 91.0	↓0.7 20.2	91.3	20.9	91.3	20.9	↓0.4 90.9	↓1.9 19.0	91.3	20.9
PT	63.2	8.8	↓0.6 62.6	↓1.6 7.2	↓0.6 62.6	↓1.6 7.2	63.2	8.8	63.2	↓3.0 5.8	↑1.2 64.4	↓4.1 4.7
Mistral-Nemo-Instruct-2407												
EN	79.3	5.9	↑0.1 79.4	↑1.0 6.9	79.3	↑0.3 6.2	↓0.3 79.0	↑0.1 6.0	↑0.1 79.4	↑0.1 6.0	↓0.2 79.1	↑0.4 6.3
ES	71.0	10.2	↑0.2 71.2	↑0.4 10.6	↓1.2 69.8	↑0.4 10.6	↓0.5 70.5	↑0.9 11.1	↓0.5 70.5	↑0.9 11.1	71.0	↑0.4 10.6
IT	65.3	8.0	↑0.5 65.8	↑0.1 8.1	↓0.2 65.1	↓0.1 7.9	↑0.5 65.8	↑0.3 8.3	65.3	8.0	↑0.3 65.6	↑0.5 8.5
PL	90.9	18.9	90.9	18.9	90.9	18.9	90.9	18.9	↓0.5 90.4	↓0.6 18.3	90.9	18.9
PT	64.4	1.5	↓1.2 63.2	↑3.6 5.1	↑0.6 65.0	↑1.0 2.5	↓1.2 63.2	↑1.9 3.4	↑0.6 65.0	↑1.0 2.5	↓1.2 63.2	↑6.4 7.9

Table 17: Performance comparison of SentenceDebias age debiasing approaches on the Multilingual Hate Speech dataset, showing effectiveness across different source-target language pairs. These baseline methods show limited cross-lingual transfer compared to IMSAE's multilingual approach.

C MSEFair

This appendix provides additional results and analysis for our Multilingual Stack Exchange Fairness (MSEFair) dataset. Table 18 presents the full reputation bias mitigation results on the MSEFair dataset across English, Portuguese and Russian, comparing Baseline, Crosslingual, IMSAE on Two-Subsets-Without (excluding target language) and IMSAE Three-Subsets (using all languages). Table 19 shows a comprehensive comparison of crosslingual reputation debiasing approaches on MSEFair across different source-target language pairs.

Target	Baseline		SAL (EN)		SAL (PT)		IMSAE (RU)		IMSAE (FullyJoint)		IMSAE (Subsets w/o)		IMSAE (Three-Subsets)	
	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext
mBERT-uncased														
EN	67.5	10.7	↓4.3 63.2	↓9.4 1.3	↓0.1 67.4	↑0.3 11.0	↓0.3 67.2	↑0.2 10.9	↓0.3 67.2	10.7	↓4.2 63.3	↓9.3 1.4	↓4.4 63.1	↓9.5 1.2
PT	78.3	16.2	↑0.1 78.4	↓0.1 16.1	↓19.6 58.7	↓14.8 1.4	↓0.1 78.2	16.2	↓0.1 78.2	↓0.8 15.4	↓19.7 58.6	↓14.5 1.7	↓19.8 58.5	↓15.6 0.6
RU	70.0	18.0	↓0.4 69.6	↓0.1 17.9	↓0.2 69.8	↑0.2 18.2	↓11.9 58.1	↓15.1 2.9	↓0.2 69.8	↓0.3 17.7	↓0.6 69.4	↓0.1 17.9	↓12.3 57.7	↓14.7 3.3
Llama3-8B														
EN	68.1	11.7	↓4.4 63.7	↓6.6 5.1	↑0.1 68.2	11.7	68.1	↓0.1 11.6	↓0.1 68.0	↓0.4 11.3	↓4.3 63.8	↓6.9 4.8	↓4.5 63.6	↓6.9 4.8
PT	82.6	21.1	↓0.3 82.3	↓0.1 21.0	↓18.0 64.6	↓11.4 9.7	82.6	↑0.4 21.5	↓0.4 82.2	↓0.3 20.8	↓17.9 64.7	↓11.5 9.6	↓17.8 64.8	↓11.2 9.9
RU	71.6	17.6	↓0.1 71.5	17.6	↓0.1 71.5	↓0.2 17.4	↓8.0 63.6	↓9.8 7.8	71.6	↑0.2 17.8	↓0.2 71.4	↓0.2 17.4	↓8.1 63.5	↓10.0 7.6
Llama-3.1-8B														
EN	68.4	11.2	↓4.3 64.1	↓6.0 5.2	68.4	↓0.2 11.0	68.4	↑0.1 11.3	↓0.1 68.3	↑0.2 11.4	↓4.3 64.1	↓6.1 5.1	↓4.5 63.9	↓5.9 5.3
PT	82.6	22.4	↓0.2 82.4	↓0.4 22.0	↓18.2 64.4	↓13.6 8.8	↓0.3 82.3	↑0.1 22.5	↓0.7 81.9	↓0.6 21.8	↓18.1 64.5	↓13.1 9.3	↓18.3 64.3	↓13.6 8.8
RU	72.1	16.4	↓0.1 72.0	↓0.2 16.2	↓0.2 71.9	16.4	↓8.4 63.7	↓10.4 6.0	72.1	↑0.2 16.6	↓0.1 72.0	↓0.3 16.1	↓8.7 63.4	↓10.6 5.8
Llama-3.2-3B														
EN	66.4	8.9	↓4.1 62.3	↓4.6 4.3	↑0.1 66.5	↓0.2 8.7	↑0.1 66.5	↑0.2 9.1	66.4	↑0.2 9.1	↓4.1 62.3	↓4.8 4.1	↓4.0 62.4	↓4.5 4.4
PT	82.1	17.7	82.1	↑0.3 18.0	↓19.4 62.7	↓11.7 6.0	↓0.1 82.0	17.7	↑0.1 82.2	↑0.1 17.8	↓19.7 62.4	↓12.5 5.2	↓19.6 62.5	↓11.5 6.2
RU	71.7	17.1	71.7	17.1	↓0.2 71.5	↑0.1 17.2	↓10.7 61.0	↓11.9 5.2	↓0.3 71.4	↑0.1 17.2	↓0.2 71.5	↑0.1 17.2	↓10.5 61.2	↓12.6 4.5
Mistral-7B-Instruct-v0.3														
EN	66.8	9.5	↓3.7 63.1	↓5.8 3.7	66.8	↓0.2 9.3	↓0.1 66.7	↓0.3 9.2	↓0.1 66.7	↓0.5 9.0	↓3.6 63.2	↓6.3 3.2	↓3.8 63.0	↓6.4 3.1
PT	81.4	20.9	↓0.3 81.1	↑0.3 21.2	↓18.2 63.2	↓9.5 11.4	↓0.2 81.2	↑0.5 21.4	↓0.1 81.3	↓0.5 20.4	↓18.1 63.3	↓9.7 11.2	↓17.9 63.5	↓8.2 12.7
RU	70.5	14.7	70.5	↓0.8 13.9	↓0.2 70.3	↑0.1 14.8	↓8.2 62.3	↓9.3 5.4	70.5	↓0.3 14.4	↓0.1 70.4	↓0.9 13.8	↓8.4 62.1	↓9.4 5.3
Mistral-7B-v0.3														
EN	67.5	10.6	↓4.2 63.3	↓4.7 5.9	67.5	↓0.1 10.5	67.5	↓0.2 10.4	↓0.1 67.4	10.6	↓4.1 63.4	↓5.0 5.6	↓4.1 63.4	↓4.9 5.7
PT	82.4	20.1	↓0.1 82.3	20.1	↓18.9 63.5	↓10.5 9.6	↓0.2 82.2	↑0.1 20.2	82.4	↓0.3 19.8	↓18.8 63.6	↓10.8 9.3	↓19.0 63.4	↓11.0 9.1
RU	71.2	16.2	↓0.2 71.0	↓0.6 15.6	↓0.1 71.1	↓0.2 16.0	↓9.7 61.5	↓9.6 6.6	↓0.3 70.9	16.2	↓0.2 71.0	↓0.7 15.5	↓9.6 61.6	↓10.1 6.1
Mistral-Nemo-Base-2407														
EN	68.4	11.0	↓4.2 64.2	↓5.6 5.4	↓0.2 68.2	↓0.2 10.8	↓0.1 68.3	↑0.3 11.3	↓0.2 68.2	↓0.2 10.8	↓4.3 64.1	↓5.6 5.4	↓4.1 64.3	↓5.7 5.3
PT	83.1	21.4	↓0.2 82.9	↓0.4 21.0	↓18.2 64.9	↓14.8 6.6	↓0.2 82.9	↑0.2 21.6	↓0.5 82.6	↓0.3 21.1	↓18.3 64.8	↓15.0 6.4	↓18.4 64.7	↓14.9 6.5
RU	72.3	16.6	↓0.2 72.1	↓0.1 16.5	↓0.3 72.0	↑0.2 16.8	↓10.8 61.5	↓11.3 5.3	↓0.2 72.1	↑0.1 16.7	↓0.2 72.1	↑0.4 17.0	↓11.0 61.3	↓11.2 5.4
Mistral-Nemo-Instruct-2407														
EN	67.9	10.4	↓3.6 64.3	↓5.4 5.0	↓0.2 67.7	↓0.2 10.2	67.9	↓0.2 10.2	↓0.3 67.6	↓0.7 9.7	↓3.7 64.2	↓5.4 5.0	↓3.7 64.2	↓5.6 4.8
PT	82.4	20.8	↑0.1 82.5	↑0.6 21.4	↓18.5 63.9	↓16.6 4.2	↓0.1 82.3	↑0.4 21.2	↓0.3 82.1	20.8	↓18.6 63.8	↓16.5 4.3	↓18.8 63.6	↓16.0 4.8
RU	72.0	15.6	↓0.2 71.8	↓0.1 15.5	↓0.1 71.9	↓0.5 15.1	↓9.1 62.9	↓11.1 4.5	↓0.1 71.9	↓0.9 14.7	↓0.2 71.8	↓0.2 15.4	↓9.1 62.9	↓11.1 4.5

Table 18: Full reputation bias mitigation results on the MSEFair dataset across English, Portuguese, and Russian, comparing Baseline, Crosslingual, IMSAE on FullyJoint, IMSAE without target language, and IMSAE with all languages. We report helpfulness prediction accuracy (Main) and True Positive Rate gap (TPR-Gap) between low and high reputation users. Results demonstrate IMSAE's effectiveness in reducing bias while maintaining acceptable task performance.

Target	Baseline		SAL (EN)		SAL (PT)		SAL (RU)		INLP (EN)		INLP (PT)		INLP (RU)		SentenceDebias (EN)		SentenceDebias (PT)		SentenceDebias (RU)	
	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext	Main	Ext
mBERT-uncased																				
EN	67.5	10.7	↓4.3 63.2	↓9.4 1.3	↓0.1 67.4	↑0.3 11.0	↓0.3 67.2	↑0.2 10.9	↓0.4 67.1	↓0.8 9.9	↓0.3 67.2	↓0.2 10.5	↓0.2 67.3	↓0.2 10.5	↓2.9 64.6	↑1.0 11.7	↓0.2 67.3	↓0.1 10.6	↓0.1 67.4	↓0.1 10.6
PT	78.3	16.2	↑0.1 78.4	↓0.1 16.1	↓19.6 58.7	↓14.8 1.4	↓0.1 78.2	16.2	78.3	↓0.1 16.1	↓0.7 77.6	↓0.6 15.6	↓0.1 78.2	16.2	↓0.1 78.2	↓0.2 16.0	↓1.5 76.8	↓1.4 14.8	78.3	16.2
RU	70.0	18.0	↓0.4 69.6	↓0.1 17.9	↓0.2 69.8	↑0.2 18.2	↓11.9 58.1	↓15.1 2.9	↓0.1 69.9	↓0.7 17.3	↓0.1 69.9	↓0.4 17.6	↓1.0 69.0	↓1.9 16.1	70.0	↓0.7 17.3	70.0	↑0.1 18.1	↑0.1 70.1	↓0.1 17.9
Llama3-8B																				
EN	68.1	11.7	↓4.4 63.7	↓6.6 5.1	↑0.1 68.2	11.7	68.1	↓0.1 11.6	↓0.1 68.0	↓0.4 11.3	↑0.1 68.2	↑0.1 11.8	68.1	11.7	↓1.4 66.7	↓1.0 10.7	68.1	11.7	68.1	11.7
PT	82.6	21.1	↓0.3 82.3	↓0.1 21.0	↓18.0 64.6	↓11.4 9.7	82.6	↑0.4 21.5	82.6	21.1	↓0.3 82.3	↓1.0 20.1	82.6	21.1	↓0.1 82.5	↓0.2 20.9	↑0.3 82.9	↑0.3 21.4	82.6	21.1
RU	71.6	17.6	↓0.1 71.5	17.6	↓0.1 71.5	↓0.2 17.4	↓8.0 63.6	↓9.8 7.8	↑0.1 71.7	↑0.1 17.7	↑0.1 71.7	↑0.1 17.7	↓0.4 71.2	↓0.5 17.1	71.6	↑0.4 18.0	71.6	↑0.2 17.8	↓0.2 71.4	↓0.4 17.2
Llama-3.1-8B																				
EN	68.4	11.2	↓4.3 64.1	↓6.0 5.2	68.4	↓0.2 11.0	68.4	↑0.1 11.3	68.4	↑0.7 11.9	↓0.1 68.3	↓0.1 11.1	68.4	11.2	↓1.2 67.2	↑0.1 11.3	68.4	11.2	68.4	↑0.1 11.3
PT	82.6	22.4	↓0.2 82.4	↓0.4 22.0	↓18.2 64.4	↓13.6 8.8	↓0.3 82.3	↑0.1 22.5	↑0.1 82.7	↑0.4 22.8	↓1.5 81.1	↓1.5 20.9	82.6	↑0.5 22.9	↓0.1 82.5	↓0.3 22.1	↑0.1 82.7	↑0.5 22.9	↑0.1 82.7	↑0.2 22.6
RU	72.1	16.4	↓0.1 72.0	↓0.2 16.2	↓0.2 71.9	16.4	↓8.4 63.7	↓10.4 6.0	72.1	↑0.2 16.6	↓0.1 72.0	16.4	↓0.3 71.8	↓1.1 15.3	72.1	↓0.3 16.1	72.1	↓0.2 16.2	↓0.1 72.0	↓0.1 16.3
Llama-3.2-3B																				
EN	66.4	8.9	↓4.1 62.3	↓4.6 4.3	↑0.1 66.5	↓0.2 8.7	↑0.1 66.5	↑0.2 9.1	↓0.1 66.3	↑0.7 9.6	66.4	8.9	66.4	↑0.1 9.0	↓0.6 65.8	↓1.0 7.9	66.4	8.9	66.4	↑0.1 9.0
PT	82.1	17.7	82.1	↑0.3 18.0	↓19.4 62.7	↓11.7 6.0	↓0.1 82.0	17.7	82.1	↑0.2 17.9	↓0.5 81.6	↑0.3 18.0	↓0.1 82.0	↑0.2 17.9	82.1	↑0.1 17.8	↓0.1 82.0	↑0.7 18.4	82.1	↑0.1 17.8
RU	71.7	17.1	71.7	17.1	↓0.2 71.5	↑0.1 17.2	↓10.7 61.0	↓11.9 5.2	↓0.1 71.6	↓0.1 17.0	71.7	↓0.1 17.0	↓1.1 70.6	↓1.8 15.3	↓0.1 71.6	↓0.3 16.8	↓0.1 71.6	↓0.2 16.9	↓0.4 71.3	↓0.1 17.0
Mistral-7B-Instruct-v0.3																				
EN	66.8	9.5	↓3.7 63.1	↓5.8 3.7	66.8	↓0.2 9.3	↓0.1 66.7	↓0.3 9.2	↑0.2 67.0	↓0.4 9.1	↓0.1 66.7	↓0.4 9.1	66.8	↓0.1 9.4	↓0.9 65.9	↓1.1 8.4	66.8	↓0.1 9.4	66.8	9.5
PT	81.4	20.9	↓0.3 81.1	↑0.3 21.2	↓18.2 63.2	↓9.5 11.4	↓0.2 81.2	↑0.5 21.4	↑0.1 81.5	↑0.1 21.0	↓0.7 80.7	↑0.2 21.1	81.4	↑0.2 21.1	81.4	↓0.1 20.8	81.4	20.9	81.4	20.9
RU	70.5	14.7	70.5	↓0.8 13.9	↓0.2 70.3	↑0.1 14.8	↓8.2 62.3	↓9.3 5.4	↑0.1 70.6	↑0.1 14.8	↑0.1 70.6	↓0.2 14.5	↓0.5 70.0	↓0.6 14.1	↑0.1 70.6	14.7	↑0.1 70.6	↓0.1 14.6	↑0.1 70.6	↓0.2 14.5
Mistral-7B-v0.3																				
EN	67.5	10.6	↓4.2 63.3	↓4.7 5.9	67.5	↓0.1 10.5	67.5	↓0.2 10.4	↑0.1 67.6	↓0.2 10.4	67.5	↑0.2 10.8	67.5	↑0.1 10.7	↓1.4 66.1	↓0.6 10.0	67.5	10.6	67.5	↑0.1 10.7
PT	82.4	20.1	↓0.1 82.3	20.1	↓18.9 63.5	↓10.5 9.6	↓0.2 82.2	↑0.1 20.2	↑0.1 82.5	20.1	↓0.8 81.6	↑0.7 20.8	↑0.1 82.5	20.1	↑0.1 82.5	↑0.1 20.2	82.4	↓0.3 19.8	82.4	↑0.1 20.2
RU	71.2	16.2	↓0.2 71.0	↓0.6 15.6	↓0.1 71.1	↓0.2 16.0	↓9.7 61.5	↓9.6 6.6	↑0.1 71.3	↓0.3 15.9	↑0.1 71.3	↑0.1 16.3	↓0.2 71.0	↓0.7 15.5	↑0.1 71.3	↑0.2 16.4	71.2	16.2	↑0.1 71.3	↓0.2 16.0
Mistral-Nemo-Base-2407																				
EN	68.4	11.0	↓4.2 64.2	↓5.6 5.4	↓0.2 68.2	↓0.2 10.8	↓0.1 68.3	↑0.3 11.3	↓0.2 68.2	↑0.2 11.2	68.4	↑0.4 11.4	68.4	↑0.2 11.2	↓0.2 68.2	↑0.7 11.7	68.4	↑0.3 11.3	↓0.1 68.3	↑0.4 11.4
PT	83.1	21.4	↓0.2 82.9	↓0.4 21.0	↓18.2 64.9	↓14.8 6.6	↓0.2 82.9	↑0.2 21.6	↓0.1 83.0	↑0.1 21.5	↓1.0 82.1	↓1.2 20.2	83.1	21.4	↓0.1 83.0	21.4	83.1	↑0.4 21.8	↓0.1 83.0	↑0.2 21.6
RU	72.3	16.6	↓0.2 72.1	↓0.1 16.5	↓0.3 72.0	↑0.2 16.8	↓10.8 61.5	↓11.3 5.3	72.3	16.6	↓0.1 72.2	↓0.1 16.5	↓0.3 72.0	↑0.3 16.9	72.3	↑0.1 16.7	72.3	16.6	↓0.3 72.0	↑0.2 16.8
Mistral-Nemo-Instruct-2407																				
EN	67.9	10.4	↓3.6 64.3	↓5.4 5.0	↓0.2 67.7	↓0.2 10.2	67.9	↓0.2 10.2	↓0.1 67.8	↓0.2 10.2	↓0.1 67.8	↓0.1 10.3	67.9	↑0.1 10.5	↓0.8 67.1	↓1.4 9.0	↓0.2 67.7	↓0.1 10.3	↓0.1 67.8	10.4
PT	82.4	20.8	↑0.1 82.5	↑0.6 21.4	↓18.5 63.9	↓16.6 4.2	↓0.1 82.3	↑0.4 21.2	82.4	↑0.3 21.1	↓1.0 81.4	↓1.6 19.2	82.4	↑0.4 21.2	82.4	↑0.2 21.0	↑0.1 82.5	↑1.0 21.8	↑0.2 82.6	↑0.3 21.1
RU	72.0	15.6	↓0.2 71.8	↓0.1 15.5	↓0.1 71.9	↓0.5 15.1	↓9.1 62.9	↓11.1 4.5	↓0.1 71.9	↑0.2 15.8	↓0.1 71.9	↓0.3 15.3	↓0.5 71.5	↓0.2 15.4	72.0	↓0.2 15.4	72.0	↓0.2 15.4	↓0.2 71.8	↓0.1 15.5

Table 19: Comprehensive comparison of crosslingual reputation debiasing approaches on MSEFair, showing performance across different source-target language combinations for various debiasing methods. This analysis reveals the limitations of traditional cross-lingual approaches when dealing with typologically different languages like Russian.