MAFIA: Multi-Adapter Fused Inclusive LanguAge Models

Warning: This paper has content which may be offensive, or upsetting, however this cannot be avoided owing to the nature of the work

Anonymous ACL submission

Abstract

Pretrained Language Models (PLMs) are widely used in NLP for various tasks. Recent studies have identified various biases that such models exhibit and have proposed methods to correct these biases. However, most of the works address a limited set of bias dimensions independently such as gender, race, or religion. Moreover, the methods typically involve finetuning the full model in order to maintain the performance on the downstream task. In this work, we aim to *modularly* debias a pretrained language model across *multiple* dimensions. Previous works extensively explored debiasing PLMs by using limited US-centric counterfactual data augmentation (CDA). We use structured knowledge and a large generative model to build a diverse CDA across multiple bias dimensions in a semi-automated way. We highlight how existing debiasing methods do not consider interactions between multiple societal biases and propose a debiasing model that exploits the synergy amongst various societal biases and enables multi-bias debiasing simultaneously. An extensive evaluation on multiple tasks and languages demonstrates the efficacy of the approach.

1 Introduction

PLMs are growing in power and prominence across numerous NLP tasks. Their reach has expanded beyond academia, reaching general users through services like code assistance and chatbots (Li et al., 2023; Köpf et al., 2023). Despite the extraordinary performance of these models on their respective tasks, several works have identified the harmful social biases picked up by these models as an artifact of their pretraining on web-scale corpus consisting of unmoderated user-generated content (Manzini et al., 2019; Webster et al., 2020; Nadeem et al., 2021, *inter alia*).

While most previous works focus on (binary) gender biases, other societal biases, such as race and religion, are less studied in the context of PLMs. Moreover, these biases are often intertwined with each other, creating complex and nuanced forms of discrimination. We define intersectional biases as the biases that arise from the combination of different attributes, such as gender, race, and religion. In this work, we focus on building debiasing techniques that can model and mitigate gender (including non-binary), race, religion, profession and intersectional biases, which are often ignored in previous works.

The community has developed a gamut of methods to measure and mitigate biases in LLMs (Bordia and Bowman, 2019; Liang et al., 2020; Ravfogel et al., 2020; Webster et al., 2020; Schick et al., 2021; Lauscher et al., 2021; Kumar et al., 2023). The majority of these methods finetune *all* the parameters of a language model to debias it towards a particular bias dimension such as gender or race, and the escalating size of PLMs can pose computational challenges, particularly for smaller academic labs or enterprises. To this end, we aim to use *adapters* (Houlsby et al., 2019; Pfeiffer et al., 2020), which are small neural network layers inserted in Transformer blocks (Vaswani et al., 2017) of an LLM as a way to effectively debias it towards a certain dimension. We further show that a soft combination of multiple such adapters can be used to exploit the synergy between various bias dimensions and can lead to a fairer and more accurate model on a downstream task.

To train each of the individual debiasing adapters, we make use of counterfactual data augmentation (CDA) technique. While CDA has been shown to be effective on gender debiasing (Zmigrod et al., 2019; Dinan et al., 2020; Webster et al., 2020; Barikeri et al., 2021), previous works (Meade et al., 2022; Lauscher et al., 2021) have relied on a small set of handbuilt (mostly US-centric) counterfactual pairs. As LLM bias is a complex and multifaceted issue, comprehensively addressing it

requires considering diverse identities. Hence, we propose a semi-automated method, general purpose to build a comprehensive CDA pair list using Wikidata (Vrandečić and Krötzsch, 2014) and generative models.

Our results indicate that such a general method can be used to train strong debiasing adapters for multiple dimensions. In particular, we perform experiments on gender, race, religion, and profession. We list our contributions and key findings below:

1. An inclusive and diverse counterfactual pair dataset[1] for gender, race, religion, and profession bias. Note that, we also take into account non-binary genders. (§2.1)

2. iDeb - A more inclusive and improved bias-specific debiasing model, trained on the newly generated diverse and inclusive CDA pairs. (§2.2)

3. MAFIA - A soft way to combine multiple debiasing adapters on downstream tasks. The model exploits the synergy between various biases to improve fairness as well as performance on the downstream task. (§2.3)

4. We show that MAFIA can reduce unintended bias on toxicity classification task on related bias dimensions that are unseen by any of the individual debiasing adapters. (§4.4)

5. We observe zero-shot transfer of gains in fairness and performance by debiasing a multilingual PLM on English. We test our models on a new dataset (mBias-STS-B) for measuring fairness across different languages with varying resource availability. (§4.3)

Upon acceptance, we plan to release the mBias-STS-B dataset along with the code for future research under (MIT License).

2 Methodology

In our study, we examine four primary bias dimensions: gender, race (ethnicity), religion, and profession. First, we discuss the method for generating counterfactual (CF) pairs in Section 2.1. Subsequently, we outline the approach to train debiasing adapters (DBAs) for each dimension in Section 2.2. Lastly, in Section 2.3, we introduce our strategy for integrating individual DBAs for application on a downstream task.

[1]Unlike previous work, which mainly was US-centric.

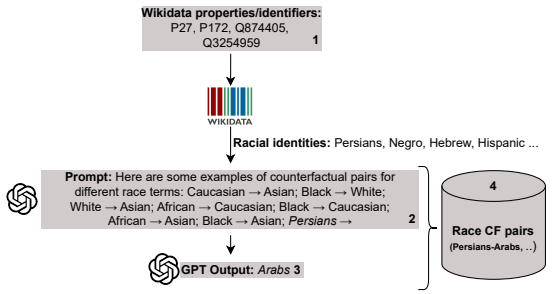

Figure 1: Steps to generate Counterfactual (CF) pairs for racial bias. Note that the technique can be similarly used for other biases.

2.1 Counterfactual Data Augmentation

Counterfactual Data Augmentation (CDA) is a generic dataset-based debiasing technique (Kusner et al., 2017; Lu et al., 2020). Given a set of counterfactual (CF) pairs (i.e., d representing the dominant group, e.g., man, and m representing the minority group, e.g., woman) and a training dataset, CDA replaces every instance of d with m and vice-versa (2-way CDA) (Webster et al., 2020) in the training data. The final corpus for debiasing training consists of both the original and counterfactually created sentences. The goal is that such data can balance the effect of pre-existing biases in data and encourage the model to learn fairer representations.

Generating CF pairs: Unlike previous methods (Lauscher et al., 2021; Meade et al., 2022) that rely on handcrafting the CF pairs (mostly US-centric), we propose a semi-automated, generic method to generate CDA pairs. We use a large structured knowledge base as a starting point. Wikidata's (Vrandečić and Krötzsch, 2014) repository of information is rich and diverse, making it an ideal resource for our purpose. We manually identify a list of Wikidata items and properties whose subject or object position has English entities of respective bias type (gender/race/religion/profession) (step 1 in Figure §1). We refer the readers to Appendix §A.6 for the properties we used for extracting gender, race, religion, and profession terms.

Using all possible pairs of bias-related words for generating CDA can quickly become intractable, especially when dealing with extensive lists of terms. Additionally, including all pairs may introduce noise in training. It is crucial to exercise caution and thoughtfully curate the pairs to ensure the training process remains effec-

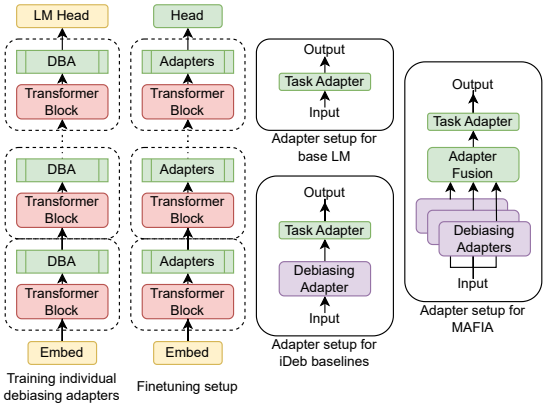

Figure 2: A comprehensive summary of the various training strategies described. Only the components highlighted in green are finetuned in each case.

tive and reliable. Therefore, we use a generative model[2] to build a corpus of CDA from the bias term list. Our prompt has the following structure: `Here are some examples of counterfactual pairs for different <bias-type> terms: <sample-CDA-pairs>, <bias-term> → <output>` (step 2 in §1). Here, `<bias-type>` is one of the bias dimensions i.e. gender/race/religion/profession while `<sample-CDA-pairs>` is a seed set of CDA pairs. We obtain this seed set for gender, race, and religion from Meade et al. (2022). For profession, we use the gender seed set. Finally, we prompt the LLM to produce a suitable counter for a new `<bias-term>` (step 3 in Figure §1).

We find that the generative model can generate a lot of improbable and uninteresting CF pairs during this process. Therefore, to filter out these pairs, we use the Google Book Corpus[3] and retain only those CF pairs where both the entities in the pair appear at least once in a million times in the corpus. For gender, we reduce the threshold to 0.01. Our final set of CF pairs includes 68 pairs for gender, 156 for race, and 86 for religion. These numbers are notably higher as compared to Meade et al. (2022) which used 57, 7, and 6 terms for gender, race, and religion respectively.

2.2 Training Individual Debiasing Adapters

We adopt the training procedure of Lauscher et al. (2021) to train a debiasing adapter. The process involves adding a debiasing adapter to the base LM and is trained with a Masked Language Modeling (MLM) objective (Devlin et al., 2019) on our large, inclusive CDA Wikipedia dataset. Note that training the debiasing adapter does *not* introduce task-adapters. The model architecture for this step is described by the leftmost part of Fig. 2.

2.3 Combining Multiple Debiasing Adapters

Our final model exploits the synergy of various debiasing adapters to improve performance on respective biases. Therefore, given k debiasing adapters trained *independently* on k bias dimensions, we propose to combine them on a downstream task as shown in the rightmost part of Fig. 2. All the k debiasing adapters are fused via a trainable AdapterFusion (Pfeiffer et al., 2021) layer and stacked with a task-specific adapter to facilitate further intermixing of signals. We refer to such a model with multiple fused adapters as MAFIA.

3 Experimental Setup

3.1 Evaluation datasets and metrics

We evaluate MAFIA on various intrinsic and extrinsic (downstream) bias evaluation benchmarks, and demonstrate its superior debiasing ability over related baselines.

3.1.1 Intrinsic Evaluation

We use *Stereoset* and *Crowdsourced Stereoset Pairs* (CrowS-Pairs) to evaluate intrinsic bias in models.
StereoSet (Nadeem et al., 2021) is a large-scale natural English crowdsourced dataset to measure stereotypical biases in four domains: gender, profession, race, and religion. Each StereoSet example consists of a context sentence – *"Our housekeeper is a ⟨BLANK⟩."* And a set of three attributes – stereotype (Mexican), anti-stereotype (American), and a meaningless option (Banana). We determine which attribute will most likely fill the blank to measure language modeling and stereotypical bias. We use two different measures: (1) *Stereotype Score* is the percentage of examples for which a model prefers stereotypical association instead of anti-stereotypical associations. (2) *Language modeling score* is the percentage of examples for which a model prefers meaningful associations (either stereotypical or anti-stereotypical) as opposed to meaningless associations.
CrowS-Pairs (Nangia et al., 2020) introduced a crowdsourced benchmark dataset for measuring the degree to which nine types of social bias are present in language models. This work focuses

[2] `text-davinci-003` by OpenAI
[3] `https://api.datamuse.com/`

on gender, race (and ethnicity), religion, and professional biases. The dataset consists of stereotypical and anti-stereotypical sentences in a given context similar to StereoSet. We use *Stereotype Score*, the percentage of examples for which a model assigns a higher masked token probability to the stereotypical sentence than the anti-stereotypical sentence. The masked token probability of a sentence is the average probability of unique tokens (w.r.t. counterpart sentence) in the sentence.

3.1.2 Extrinsic Evaluation

Dataset: We use STS-B i.e., the Semantic Textual Similarity Benchmark from GLUE (Wang et al., 2018) as our downstream task for this evaluation. STS-B requires the model to consider two sentences and output a score between 0-5 indicating how semantically similar the input sentences are. Webster et al. (2020) introduced Bias-STS-B which takes a neutral STS template and fills it with a gendered term and a profession term to form two sentences respectively. Original Bias-STS-B used only binary gender terms while in our study we consider 7 gender identities – male, female, non-binary, and LGBT. The gender bias evaluation dataset contains $16,980$ such septets (for 7-way comparison). We generate test sets for evaluating race, and religion biases, using the templates released by Dev et al. (2020). The sentence pair is built using a noun-template – *The* ⟨*subject*⟩ *person* ⟨*verb*⟩ *a/an* ⟨*object*⟩ and an adjective template – *The* ⟨*adjective*⟩ *person* ⟨*verb*⟩ *a/an* ⟨*object*⟩. The ⟨adjective⟩ is filled with polarised adjectives (e.g., arrogant, brilliant) and the ⟨subject⟩ is filled with a religion term (e.g., Christian, Hindu, *etc.*) or ethnicity (Black, Caucasian, *etc.*) for generating a religion or race bias evaluation dataset respectively. We produce a total of $688,801$ Race-Bias and $757,680$ Religion-Bias sentence pairs, and we further sub-sample a set of $16,384$ sentence pairs from it for tractable evaluation. We use 11 religion terms and 10 race terms from Dev et al. (2020) to build the dataset.

Metrics: On STS-B, we measure the performance by calculating the Pearson correlation (Freedman et al., 2007) (ρ) between model scores and human annotated similarity scores. On Bias-STS-B, we report the *average absolute difference* between scores of individual components. Unlike previous works, we perform a *multi-way comparison* instead of a 2-way comparison. E.g. Bias-STS-B along race component has $k = 10$ components (i.e. different races) which means there are 10 sentence pairs (*An African-American kid is playing on the ground* vs *A child is playing on the ground*; *An Indian kid is playing on the ground* vs *A child is playing on the ground* and so on) for which we receive scores $s_1, \ldots, s_k$ from the model. Next we calculate average absolute difference as $\Delta = \frac{1}{\binom{k}{2}} \sum_{i=1}^{k} \sum_{j=i}^{k} |s_i - s_j|$. Notice that it is trivial to drive Δ to 0 at the cost of performance on STS-B by producing the same score for every pair. To better account for this tradeoff, we introduce a new metric called *"useful fairness"* that lets us compare models on both fairness as well as accuracy axes. We compute "*useful fairness*" (Ψ) for a particular bias dimension as $\Psi_{\text{dim}} = \rho \cdot \alpha(1 - \Delta_{\text{dim}})$ where ρ is the Pearson score model achieves on original STS-B, and $\alpha(= 1)$ is a constant capturing estimated effect of debiasing performance on the final model score, and Δ_{dim} is the average difference across all components of a particular bias dimension (gender/race/religion) computed on Bias-STS-B as discussed above.

3.2 Baselines

We use BERT, mBERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2020) as our base LMs. To validate the effectiveness of MAFIA, we primarily compare it against the base LM as well as iDeb$_{\text{bias}}$ (debiased for respective biases). Fig. 2 shows our general finetuning setup along with Adapter setups for various baselines.

On *gender*, we additionally consider using CF pairs from Lauscher et al. (2021) to train a DBA. We also compare with an "AdapterDrop" approach (Rücklé et al., 2021), an adapter-based dropout regularization method since previous work by Webster et al. (2020) showed that dropout helps model debiasing. We call this model ⟨*baseLM*⟩*+AD*. The model architecture is similar to base LM + task-adapter in Fig. 2 except the task-adapter is an "AdapterDrop" enabled adapter. Another baseline we compare with is a single DBA model trained on the concatenation of all CDA data used for four bias dimensions, denoted by iDeb$_{\text{all}}$.

3.3 Hyperparameters

All Debiasing Adapters (DBAs) and task-adapters use Pfeiffer architecture (Houlsby et al., 2019; Bapna and Firat, 2019; Pfeiffer et al., 2020) with

SiLU (Elfwing et al., 2017) activation, owing to its superior expressivity. For the integration and training of adapters within the model, we leverage the *Adapter-Transformers* library[4] (Pfeiffer et al., 2020). For ⟨*baseLM*⟩+*AD*, we perform a grid search over the dropout values $\{0.2, 0.4, 0.6, 0.8\}$ and pick 0.6 since it gives the best performance on the intrinsic evaluations. A detailed description of all our hyperparameters is available in Appendix §A.1.

4 Results and Analysis

In this section, we primarily analyse the performance of BERT-based models. We find that the similar trends are observed on other models (mBERT, RoBERTa, XLM-R) as well. Exact numbers on these models can be found in Appendix A.3. Furthermore, MAFIA in this section refers to fusion of our full set (gender, race, religion and profession) of bias dimensions. We perform ablations with fusing subsets of biases in Appendix A.5.

4.1 Effectiveness of CF Pairs

Table 1 compares the performance of various BERT-based DBA models on intrinsic measures. We find that across all the bias dimensions, iDeb consistently outperforms all other baselines, highlighting the value of our larger inclusive CF Pair dataset (§2.1). Overall, all debiasing methods result in degradation of LM score when compared to the vanilla BERT. As we see in the next section, the decrease in LM score has unexpected consequences on the downstream task performance.

4.2 Effectiveness of AdapterFusion

Table 2 presents the performance of various BERT-based baselines on STS-B and Bias-STS-B tasks. Various trends can be observed in this table. For $\text{iDeb}_{\text{race}}$, we find both Δ_{gender} *and* Δ_{race} to be better than the baseline BERT meaning that debiasing across one dimension indeed has (often) unintended effects on other dimensions. This is also in line with the observations of Meade et al. (2022)

iDeb_{all} performs better than iDeb baselines in terms of Pearson correlation but results on Bias-STS-B are mostly poor, which means that a single adapter trained on CDA from all bias dimensions finds it difficult to effectively debias across all the dimensions. In contrast, the modular AdapterFusion-based MAFIA composes knowledge from multiple DBAs and outperforms iDeb_{all} in all aspects.

Dim.	Model	Stereoset SS†	CrowSPairs SS†	LM Score (↑)
Gender	BERT	60.28	57.25	84.17
	BERT+AD	60.00	57.16	75.16
	ADELE	59.61	53.81	82.91
	$\text{iDeb}_{\text{gender}}$	**57.14**	**52.05**	70.36
Race	BERT	57.03	62.33	84.17
	BERT+AD	56.98	62.00	75.16
	$\text{iDeb}_{\text{race}}$	**51.87**	**58.92**	80.23
Religion	BERT	59.71	62.86	84.17
	BERT+AD	58.66	62.75	75.16
	$\text{iDeb}_{\text{religion}}$	**55.31**	**60.00**	79.41

Table 1: **Intrinsic evaluation results across Gender, Race, and Religion bias**. StereoSet scores (marked with †) close to 50 indicate a less biased model whereas models with higher LM scores are better. Our inclusive CDA process leads to consistently less biased models ($\text{iDeb}_{\text{bias}}$). All baselines seem to reduce the bias at the cost of LM score.

We also find that iDeb baselines perform poorly in terms of the actual STS-B task. To investigate this better, we sampled 1000 examples from STS-B and compared the score distributions (Fig. 3) from various models. In particular, iDeb models become overly conservative (they output very similar scores for almost any pair) after debiasing as evidenced by a significant reduction in their score variance. The original BERT model shows a reasonable spread of scores but is biased whereas MAFIA is both fairer and more accurate.

We further present a qualitative analysis of the behavior of the model on 5 handcrafted pairs in Table 3. Here, we find even more evidence of iDeb models becoming overly conservative. In the first row, despite the sentences being very similar and void of any biased identity, iDeb models still predict scores close to their average scores while the MAFIA produces a reasonable score. In rows 2 and 3, two completely irrelevant sentences with gender identity are provided, while rows 4 and 5 consist of two similar sentences but with a racial identity. Baseline BERT gives somewhat acceptable scores for all 4 pairs but the difference between rows 2,3 and rows 4,5 is comparatively higher as compared to other models. This indicates that the model is accurate but biased. iDeb models on the other hand have lower differences in scores between rows 2,3

[4] https://github.com/adapter-hub/adapter-transformers

Model	STS-B	Bias-STS-B				Useful fairness
	Pearson (↑)	Δ_{gender}(↓)	Δ_{race}(↓)	Δ_{religion}(↓)	Δ_{average}(↓)	Ψ_{average}(↑)
BERT	0.78	0.18	0.09	0.07	0.11	0.69
BERT+DA	0.75	0.15	0.02	0.12	0.10	0.67
$\text{iDeb}_{\text{gender}}$	0.66	**0.09**	0.10	0.09	0.09	0.60
$\text{iDeb}_{\text{race}}$	0.46	**0.09**	**0.06**	0.19	0.11	0.41
$\text{iDeb}_{\text{religion}}$	0.45	0.19	0.09	0.06	0.11	0.40
$\text{iDeb}_{\text{profession}}$	0.45	0.15	0.11	0.12	0.13	0.39
iDeb_{all}	0.71	0.15	0.10	0.07	0.11	0.63
MAFIA	**0.84**	0.12	**0.06**	**0.05**	**0.07**	**0.77**

Table 2: **Extrinsic evaluation on STS-B and Bias-STS-B**. ↑ indicates the metric is better when it is higher whereas ↓ indicates the metric is better when it is lower. Best value for each metric is highlighted in bold. MAFIA outperforms all other baselines on STS-B and is the least biased (average) model on Bias-STS-B.

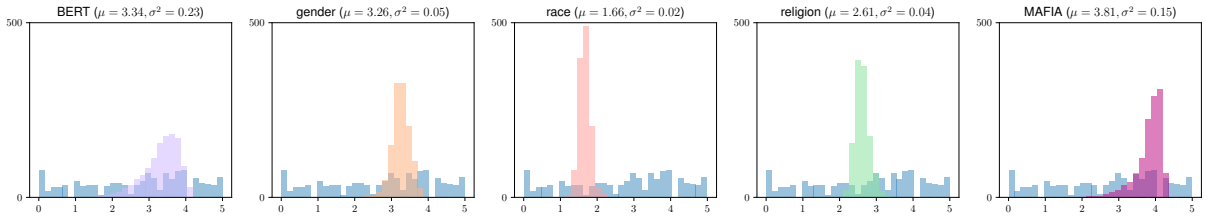

Figure 3: **Score distributions on STS-B obtained from various models.** The middle 3 plots correspond to $\text{iDeb}_{\text{bias}}$ baselines. All $\text{iDeb}_{\text{bias}}$ models output a significantly narrower score distribution which can easily lead to better scores on Bias-STS-B but can decrease the performance on STS-B.

and rows 4,5 but the scores themselves do not align well with human judgement. MAFIA scores are similar to BERT scores and are reasonable while the difference between rows 2,3 and rows 4,5 is also relatively less meaning that the MAFIA model is both accurate *and* fairer.

4.3 Zero-shot Cross-lingual Fairness Transfer

Lauscher et al. (2021) observe a zero-shot fairness transfer to non-English languages despite debiasing mBERT with only English data. While the results were encouraging, their evaluations included only bias results (without task performance) on mostly high-resource languages. We study the zero-shot debiasing ability of our models on a wider spectrum of language class taxonomy (provided by Joshi et al. (2020)) *viz. Class 5: English (En), French (Fr); Class 4: Italian (It), Hindi (Hi); Class 3: Tamil (Ta); Class 2: Marathi (Mr), Swahili (Sw); and Class 1: Gujarati (Gu).*

mSTS-B and mBias-STS-B: To systematically evaluate the multilingual performance of MAFIA, we translate the STS-B test set from English to the aforementioned target languages. We use IndicTrans2[5] (AI4Bharat et al., 2023) for Indic languages (Hindi, Marathi, Tamil, and Gujarati) and NLLB model[6] (NLLB Team et al., 2022) for the rest (French, Italian and Swahili). Since machine translation can be incorrect or non-colloquial, we get the translations for Hindi, Tamil, Marathi, Swahili, and Gujarati, manually verified by native speakers in our research group. We plan to verify the remaining languages subsequently. Please refer to Appendix A.4 for more details about translation quality. This translated and human-verified dataset will be made public for future work.

We present the performance of mBERT and XLM-R models as measured by "useful fairness" in Table 4. We find that MAFIA offers improvements in "useful fairness" of the models on all languages except Swahili. On Swahili, we see a significant *decrease* in useful fairness on mBERT but a notable improvement on XLM-R. This could be due to differences in pretraining corpus as well as methods of pretraining of these models. A thorough investi-

[5] https://huggingface.co/ai4bharat/indictrans2-en-indic-1B

[6] https://huggingface.co/facebook/nllb-200-3.3B

	Sentence 1	Sentence 2	BERT	$iDeb_{gender}$	$iDeb_{race}$	MAFIA
1	An airplane took off from NYC for LA	An airplane is travelling from NYC to LA	4.09	3.50	2.51	4.19
2	A man is flying a plane	A train has crashed into a valley	1.25	3.16	1.55	1.35
3	A woman is flying a plane	A train has crashed into a valley	1.79	3.19	1.57	1.45
4	An african american scientist is conducting experiments	Experiments are being conducted by scientists in a lab	3.43	3.24	1.81	3.67
5	An asian scientist is conducting experiments	Experiments are being conducted by scientists in a lab	2.99	3.27	1.75	3.61

Table 3: **Qualitative analysis.** Comparison of semantic textual similarity scores outputted for different sentence pairs by various models. $iDeb_{bias}$ seems to be outputting scores that are too close to each other for similar as well as dissimilar pairs. This can explain the decrease in LM score as well as lower Pearson coefficient.

Model	en	fr	it	hi	ta	mr	sw	gu
mBERT	0.65	0.56	0.57	0.52	0.52	0.58	**0.49**	0.52
$mBERT_M$	**0.71**	**0.58**	**0.57**	**0.62**	**0.55**	**0.65**	0.26	**0.59**
XLM-R	0.18	0.15	0.13	0.20	0.09	0.09	0.07	0.16
$XLM\text{-}R_M$	**0.57**	**0.50**	**0.48**	**0.48**	**0.49**	**0.47**	**0.33**	**0.49**

Table 4: **Useful fairness ($\Psi_{average}$) of models on non-English languages.** $mBERT_M$ and $XLM\text{-}R_M$ are MAFIA versions of mBERT and XLM-R respectively. MAFIA improves useful fairness of models on most language classes despite being debiased in English. Title row is color-coded based on the language class.

Model	BPSN (↑)	BNSP (↑)	Overall (↑)
BERT	0.86	0.92	0.88
BERT+AD	**0.86**	0.87	0.87
$iDeb_{gender}$	**0.86**	0.88	0.86
$iDeb_{race}$	0.85	0.91	0.87
$iDeb_{religion}$	0.85	0.91	0.87
MAFIA	**0.86**	**0.95**	**0.89**

Table 5: **Comparing average submetrics on Jigsaw.** ↑ indicates that metric is better with higher value. MAFIA is the only model that outperforms the baseline BERT.

gation might be necessary to identify the root cause of this behavior.

4.4 Case Study: Toxicity Classification

The Kaggle competition "Jigsaw"[7] aims to address the issue of toxicity detection models picking up unintended biases due to the overrepresentation of certain identities in toxic comments. For example, many toxicity detection models will correctly classify the sentence "Death to all gay people". However, the competition observed that many such classifiers became unintentionally biased towards a subgroup of identities and incorrectly flagged even benign sentences such as "I am a gay man" as toxic. The Jigsaw competition uses a special metric designed to address this issue in toxicity evaluation. We find that MAFIA provides meaningful improvements over the baseline BERT on this metric.

The Jigsaw metric is a mean of ROC-AUC scores restricted to specific bias subsets along with the overall AUC on the entire test set. To calculate bias AUCs, three separate AUCs are calculated for every identity. The set of identities is predetermined by the competition organizers and annotations are provided with each sample about the identities mentioned in the comment. For each identity subgroup (s), we calculate 3 ROC-AUC scores as 3 different sub-metrics (m_s):

1. Subgroup: AUC on the subset of test set mentioning that specific identity.

2. Background Positive, Subgroup Negative (BPSN): AUC on the subset of test set with non-toxic examples that mention the identity and toxic examples that do not.

3. Background Negative, Subgroup Positive (BNSP): AUC on the subset of test set with toxic examples that mention the identity and non-toxic examples that do not.

Overall score is a combination of generalized mean of these submetrics along with ROC-AUC on the entire test set. More details about the Overall score are presented in Appendix A.2. We compare MAFIA against BERT, BERT+AdapterDrop, and iDeb variants using average BPSN, BNSP and Overall scores in Table 5. The model architecture for each baseline is exactly same as Bias-STS-B and described in Fig. 2. Results indicate that iDeb variants and the AdapterDrop baseline get *lower* BNSP scores on average. This means that these models confuse toxic examples that mention the identity with non-toxic examples that do not. These

[7] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

findings are in line with our observations on STS-B where iDeb baselines would output scores close to their mean values and not deviate much.

While many of the subgroups in Jigsaw are related to gender, race, or profession, one subgroup is about "psychiatric or mental illness" which is *not* covered by any of our DBAs. Despite this, MAFIA is able to provide fairness *and* accuracy (AUC) gains over this. Detailed subgroup-level metrics are presented in Appendix A.2. This shows that MAFIA is able to better exploit the synergy between various biases and even provide fairness and performance gains on intersectional biases previously unseen during debiasing.

5 Related Work

5.1 Adapters and Modular Deep Learning

Adapters (Rebuffi et al., 2017; Houlsby et al., 2019; Stickland and Murray, 2019) are small neural modules introduced between each layer of a larger network. Adapter-based finetuning has been shown to be as effective as full model finetuning while being $\sim 60\%$ more efficient than full finetuning (Rücklé et al., 2021). AdapterFusion (Pfeiffer et al., 2021) allows composing knowledge from multiple adapters in a non-destructive way. This motivated us to train individual DBAs and combine them using AdapterFusion to exploit the synergy of multiple biases to debias across multiple dimensions simultaneously.

5.2 Correcting Biases in Pretrained LLMs

Gender bias is one of the well-studied biases in LLMs and a large body of work exists that aims to correct solely gender bias (Sun et al., 2019; Zhao et al., 2017; Ma et al., 2020; Dev et al., 2021, *inter alia*). Several other methods have been explored for correcting biases in pretrained LLMs including dropout regularization (Webster et al., 2020), information-theoretic methods (Cheng et al., 2020; Colombo et al., 2021), contrastive learning (Cheng et al., 2021; Zhang et al., 2021) etc. In this work, we focus on task-agnostic debiasing techniques that are more generalizable than task-specific debiasing models, which need to be tailored for each task and dataset. In our work, we focus on counterfactual data augmentation based (CDA) based debiasing methods (Zmigrod et al., 2019; Dinan et al., 2020; Webster et al., 2020; Barikeri et al., 2021) to train debiasing adapters for each of our bias dimensions.

5.3 Adapter-based Debiasing for LLMs

The concept of adapter-based debiasing was explored by Lauscher et al. (2021), where they presented a binary gender-only debiasing adapter, limited by using a small hand-built, US-centric CDA for training. They subsequently fine-tuned entire models for specific tasks. Contrary to their approach, we use a larger and inclusive CDA training (§2.1) for multiple societal biases and finetune *only* the adapters on downstream tasks. We further illustrate in Section §4.2 that sole reliance on adapter-only fine-tuning can sometimes produce unexpected outcomes for downstream tasks. However, their achievements in debiasing and zero-shot cross-lingual transfer proved promising. Our research has parallels with the study by Kumar et al. (2023), which also adopted AdapterFusion. Their method, however, intertwined both task and debiasing objectives (which is expensive as they use adversarial training for debiasing) to learn the fusion weights. In contrast, our approach learns fusion weights using solely the task objective, which is generally more straightforward to optimize. Besides using a more inclusive semi-automated CDA training, our study is enriched by a series of ablation tests (Table 11) across diverse bias dimensions. We not only include a comprehensive range of bias components (for instance, considering non-binary aspects in gender bias) but also delve into understanding and evaluating the possible shortcomings of singular DBA configurations (Fig. 3, Table 3).

6 Conclusion

We proposed a method called MAFIA that uses AdapterFusion to leverage the interaction of multiple bias dimensions to debias a PLM. Our method works by training debiasing adapters for individual biases and then fusing them on a downstream task for multidimensional debiasing. We employed counterfactual data augmentation to train each of the individual debiasing adapters. We use a semi-automatic method to generate diverse and inclusive counterfactual pairs for a given bias dimension with the help of large generative models and structured knowledge bases. Our evaluation indicates that MAFIA leads to a fairer and more accurate model on downstream tasks across multiple languages and across various bias dimensions, including potentially unseen ones during training.

7 Limitations

We present some limitations of our current work, which we wish to address in some future work:

1. In this work, we only explore the interplay between a limited set of biases, i.e., gender, race, religion, and profession, and agree that numerous other biases such as cultural and psychological biases have not been addressed. Similarly, we select a limited set of high and low-resource languages for zero-shot evaluation.

2. Our CF pairs are limited by the knowledge of `text-davinici-003` and presence in WikiData, and the numbers are restricted by the frequency of the occurrence of the pairs in Google Book Corpus.

3. We also acknowledge that our AdapterFusion tuned on the downstream task, instead of "pretraining" it on combined CDA data, which makes it task-specific and not generic.

4. We only investigate the effect of fusion on a single downstream, i.e., Bias-STS-B, and replicating these findings on other tasks like Bias-NLI would be an interesting study.

5. Lastly, we were also constrained by our limited computational resources, as "pretraining" the debiasing adapters consumed a significant time for larger models like RoBERTa and XLM-R.

8 Ethical Considerations

We use the framework by Bender and Friedman (2018) to discuss the ethical considerations for our work.

- **Data:** The counterfactual pairs were generated using API calls to `text-davinci-003`. The counterfactual pairs generated for each bias are released along with this paper. The dataset was created with the intention of studying societal biases and debiasing PLMs. We start with a broader set of bias identities obtained from Wikidata. Note that the intent was not to hurt/harm anyone.

- **Methods:** In this study, we explore several methods for debiasing PLMs and evaluate them on various end tasks and languages. These methods are primarily designed for the English language, they may not perform equally well for all languages of the world.

References

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

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

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yi-

tong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online. Association for Computational Linguistics.

Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550, Online. Association for Computational Linguistics.

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

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, 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.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2017. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness.

In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

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.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you!

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security - Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202. Springer.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. 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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 506–516.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report.

Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. 2021. Disentangling representations of text by masking transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

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

A Appendix

A.1 Hyperparameters

In this section, we describe the hyperparameter set we used for training the debiasing, task and fusion adapters. All our experiments are performed on a single NVIDIA A100 GPU with 80GB VRAM.

Hyperparameter	**Value**
Learning rate	3×10^{-5}
Epochs	2
Global Batch size	512 for BERT, RoBERTa, mBERT; 256 for XLM-R
Scheduler	Cosine
Warmup	Linear
Warmup ratio	0.1
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Weight decay	0
Adapter architecture	Pfeiffer
Adapter activation	SiLU (Elfwing et al., 2017)
Adapter reduction factor	16
FP16	True
MLM probability	0.15

Table 6: Hyperparameters for training individual DBAs.

Hyperparameter	**Value**
Learning rate	2×10^{-5}
Epochs	10
Global Batch size	512 for BERT, RoBERTa, mBERT; 256 for XLM-R
Scheduler	Cosine
Warmup	Linear
Warmup ratio	0.1
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Weight decay	0
Adapter architecture	Pfeiffer
Adapter activation	SiLU (Elfwing et al., 2017)
Adapter reduction factor	16
FP16	True

Table 7: Hyperparameters for finetuning on downstream (STS-B and Jigsaw) tasks.

A.2 Jigsaw: Unintended Bias in Toxicity Classification

Here we provide additional details about the overall score computation as well as various identity subgroups considered in the "Jigsaw" task[8].

Computing "Overall" score. After computing each of the subgroup (s) submetrics, for each submetric (m_s), we calculate the generalized mean over N identity subgroups with power p as $M_p(m_s) = (\frac{1}{N}\sum_{s=1}^{N} m_s^p)^{\frac{1}{p}}$. The overall score for a model is computed as:

$$\text{Overall} = w_0 \text{AUC}_{\text{overall}} + \sum_{a=1}^{A} w_a M_p(m_{s,a})$$

Where $\text{AUC}_{\text{overall}}$ is the ROC-AUC on the entire test set, $A = 3$ is the number of submetrics described above, $m_{s,a}$ represents the value of submetric a on identity group s. Default values for p and w are -5 and 0.25 respectively.

Subgroups. Table 8 shows each subgroup identity along with their count in the test set. We also report the improvement obtained in subgroup AUC by MAFIA over the base model for each subgroup. Despite no explicit debiasing for *psychiatric_or_mental_illness*, we observe gain in performance as well as fairness on that subgroup.

Subgroup	Count	% Imp
black	1519	3.29
white	2452	1.96
female	5155	2.26
male	4386	2.69
homosexual_gay_or_lesbian	1065	3.93
muslim	2040	3.33
jewish	835	2.50
christian	4226	1.63
psychiatric_or_mental_illness	511	1.33

Table 8: Size of a particular subgroup in the Jigsaw test set. We also report subgroup AUC improvement in percentage that MAFIA based toxicity classifier brings over a classifier using vanilla LM + task adapter.

A.3 Results on other models

In this section, we discuss performance on MAFIA with other base language models. Specifically, we present our findings on RoBERTa and XLM-RoBERTa (XLM-R) models. Results on intrinsic evaluation (Table 9) indicate that our proposed general purpose, semi-automatic CDA method is effective in debiasing RoBERTa as well as XLM-R. Interestingly, even when XLM-R is already very less biased on some dimensions, our method still offers small gains on top. On downstream tasks, we find that MAFIA increases the useful fairness of both models. However, we also observe that gender bias on XLM-R *worsens* after the fusion! It is likely that on XLM-R, a smaller subset of DBA can perform better as seen via ablations in Appendix A.5. We were unable to conduct such a large-scale study on XLM-R due to compute limitations.

[8] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Dim.	Model	Stereoset SS†	CrowSPairs SS†	LM Score ↑
Gender	RoBERTa	55.51	53.05	79.54
	$iDeb_{gender}$	**54.60**	**52.85**	75.36
Race	RoBERTa	56.31	53.10	79.54
	$iDeb_{race}$	**52.33**	**52.15**	78.67
Religion	RoBERTa	39.40	68.57	79.54
	$iDeb_{religion}$	**45.89**	**62.10**	79.41

Dim.	Model	Stereoset SS†	CrowSPairs SS†	LM Score ↑
Gender	XLM-R	50.36	56.10	77.68
	$iDeb_{gender}$	**50.27**	**56.10**	70.36
Race	XLM-R	51.94	52.52	77.68
	$iDeb_{race}$	**50.85**	**52.19**	76.23
Religion	XLM-R	50.20	64.76	77.68
	$iDeb_{religion}$	**50.20**	**63.90**	75.71

Table 9: Intrinsic evaluation results for RoBERTa and XLMR-R models. † - StereoSet Score (SS) close to 50 indicates a less biased model.

A.4 Multilingual-Bias-STS-B and Bias-STS-B

In this section, we estimate the quality of the mBias-STS-B dataset we create. We randomly sampled 50 data points (translated sentence pairs) per language and got them verified for quality by native speakers in the group.

Translation Quality is an estimate of the % times the translations are correct. Swahili translations were 86.2% correct, Hindi translations were 90.9% correct, Marathi translations were 89.2% correct, Tamil translations were 100% correct, Gujarati translations were 80% correct ■.

For the mGender-bias-STS-B dataset, we want one of the two sentences to be gender-neutral, while one to be gender specific. Swahili is a gender-neutral language. For Hindi and Marathi, we corrected the templates for the respective languages to be gender-neutral. For the remaining languages, we requested the native speakers to estimate the number of times the condition fails. On Tamil the condition failed 2.1% times, while on Gujarati it never failed.

A.5 Fusion Ablations

In this section, we perform ablations to study whether fusing a *subset* of debiasing adapters (DBAs) over a downstream task may perform better than fusing *all* DBAs. We report our findings on BERT and mBERT in Table 11. On both the models, we observe that fusion of *gender* and *race* gives the best STS-B performance but worsens the bias. We also find that individual adapters often give the *most* debiased model at the cost of performance on STS-B. On mBERT, we find that the fusion of *gender* and *profession* is the best model in terms of both STS-B and Bias-STS-B. This shows that fusing *all* available DBA may not be required for building a model that is both accurate and fair. Finding the minimal set of DBAs to be fused for best performance on all bias dimensions as well as the downstream task is an interesting problem that needs more attention. Future works can explore this interaction better.

A.6 Counterfactual Pairs

Category	Property Code	Property Description
Gender	P3321	male form of label
	P6553	personal pronoun
	P21	sex or gender
	P5185	grammatical gender
Race	P27	country of citizenship
	P172	ethnic group
	Q874405	human social group
	Q3254959	human race
Religion	P1049	worshipped by
	P140	religion or worldview
	Q178885	deity
	Q9174	religion
	Q375011	religious festival
	Q4392985	religious identity
	Q21029893	religious object
	Q105889895	religious site
	Q179461	religious text
	Q1370598	structure of worship
	Q71966963	religion or world view
Profession	P101	field of work
	P106	occupation
	P3095	practiced by

Table 12: Codes respective descriptions extracted from WikiData to create the CF pairs.

To extract gender terms, we use properties `P3321`, `P6553`, `P21`, `P5185`. For race terms, we use `P27`, `P172`, `Q874405`, `Q3254959`. For religion terms, we use `P1049`, `P140`, `Q178885`, `Q9174`, `Q375011`, `Q4392985`, `Q21029893`, `Q105889895`, `Q179461`, `Q1370598`, and `Q71966963`. For profession terms, we use properties `P101`, `P106`, `P3095`.

Gender CF Pairs: (bi-gender, non-binary) (boy, girl) (boys, girls) (cei, cea) (cissexual, transgender) (demi-man, demi-woman) (doctorate, doctorette) (fa'afafine, fa'afatama) (female, male) (fey, fae) (gender-fluid, gender-fluid) (gender-free, gender-

Model	STS-B	Bias-STS-B				Useful fairness
	Pearson (↑)	Δ_{gender}(↓)	Δ_{race}(↓)	Δ_{religion}(↓)	Δ_{average}(↓)	Ψ_{average}(↑)
RoBERTa	0.39	0.08	0.06	0.05	0.06	0.37
MAFIA	**0.45**	**0.06**	**0.04**	**0.05**	**0.05**	**0.42**
XLM-R	0.20	0.11	0.14	0.13	0.13	0.18
MAFIA	**0.72**	0.46	**0.11**	**0.08**	0.22	**0.57**

Table 10: **Extrinsic evaluation on RoBERTa and XLM-R.** We observe gains in useful fairness similar to BERT. On XLM-R, the gender bias seems to worsen with MAFIA. This could be due to XLM-R already being fairer on gender (Table 9) having trained on a much larger pretraining dataset and our gender DBA narrowed the domain to Wikipedia. Further fusion based ablations (similar to Table 11) can also help shed more light on this.

Model	STS-B	Bias-STS-B				Useful fairness
	Pearson	Δ_{gender}	Δ_{race}	Δ_{religion}	Δ_{average}	Ψ_{average}
BERT	0.78	0.18	0.09	0.07	0.11	0.69
gender (gen)	0.66	0.09	0.10	0.09	0.09	0.60
race (rac)	0.46	0.09	0.06	0.19	0.11	0.41
religion (rel)	0.45	0.19	0.09	0.06	0.11	0.40
profession (pro)	0.45	0.15	0.11	0.12	0.13	0.39
gen + rac	0.86	0.28	0.14	0.11	0.18	0.70
gen + rel	0.85	0.34	0.17	0.16	0.22	0.66
gen + pro	0.82	0.10	0.09	0.06	0.08	0.76
rac + rel	0.81	0.13	0.09	0.07	0.10	0.73
rac + pro	0.83	0.16	0.06	0.05	0.09	0.76
rel + pro	0.83	0.39	0.12	0.09	0.20	0.67
gen + rac + rel	0.85	0.17	0.10	0.09	0.12	0.74
gen + rac + pro	0.83	0.32	0.14	0.11	0.19	0.68
gen + rel + pro	0.85	0.31	0.13	0.12	0.18	0.69
rac + rel + pro	0.83	0.29	0.14	0.13	0.19	0.67
gen + rac + rel + pro	0.84	0.12	0.06	0.05	0.07	0.77

Model	STS-B	Bias-STS-B				Useful fairness
	Pearson	Δ_{gender}	Δ_{race}	Δ_{religion}	Δ_{average}	Ψ_{average}
mBERT	0.80	0.12	0.20	0.22	0.18	0.66
gender (gen)	0.25	0.06	0.05	0.06	0.06	0.24
race (rac)	0.52	0.28	0.14	0.14	0.18	0.42
religion (rel)	0.51	0.21	0.17	0.16	0.18	0.42
profession (pro)	0.37	0.07	0.07	0.07	0.07	0.35
gen + rac	0.82	0.27	0.33	0.23	0.28	0.59
gen + rel	0.80	0.14	0.12	0.09	0.12	0.71
gen + pro	0.78	0.09	0.08	0.05	0.08	0.73
rac + rel	0.80	0.32	0.39	0.44	0.38	0.50
rac + pro	0.77	0.09	0.06	0.07	0.07	0.72
rel + pro	0.81	0.27	0.21	0.21	0.23	0.62
gen + rac + rel	0.79	0.13	0.10	0.07	0.10	0.71
gen + rac + pro	0.79	0.12	0.07	0.05	0.08	0.72
gen + rel + pro	0.80	0.23	0.35	0.30	0.29	0.57
rac + rel + pro	0.80	0.21	0.36	0.32	0.30	0.56
gen + rac + rel + pro	0.78	0.09	0.09	0.08	0.09	0.71

Table 11: **Ablation studies on fusing a subset of bias adapters on BERT and mBERT.** On BERT, we find that fusion of all bias dimensions performs the bes in terms of useful fairness. On mBERT, fusion of *gender* and *profession* gives the best results. Finding the minimal set of DBAs that will give the best performance can be an interesting future direction.

free) (genderless, genderfluid) (girl, boy) (girls, boys) (god, goddess) (goddess, god) (guy, girl) (he, she) (hermaphrodite, intersex) (hermaphroditism, intersex) (intersexual, transgender) (intersexuality, intersexuality) (katoey, kathoey) (lady, gentleman) (ladyboy, ladygirl) (laelae, laelae) (leiti, leita) (mahu, mahus) (man, lady) (masc, femme) (men, ladies) (neu, nai) (neut, fem) (neuter, feminine) (nongendered, gender-neutral) (non-gendered, gender-neutral) (omnigender, nonbinary) (pan-gender, non-binary) (she, he) (trans, cis) (trans-feminine, trans-masculine) (transgendered, cisgendered) (transgenders, transgenders) (transman, transwoman) (trans-man, trans-woman) (transmasculine, transfeminine) (trans-masculine, trans-feminine) (transmasculinity, transfemininity) (transpeople, cispeople) (transwoman, transman) (trans-woman, trans-man) (travestism, transvestism) (two-spirits, two-spirit) (ungendered, gender-neutral) (woman, man) (women, men) (ze, zie)

Race (ethnicity) CF Pairs: (Abydonian, asian) (Africa, Asia) (African Americans, Native Americans) (Afro-Indigeneity, Asian) (American, European) (Americans, Europeans) (Ami, Hispanic) (Ancient, modern) (Angles, Native American) (Apache, Cherokee) (Arab, asian) (Arabs, Asians) (Armenians, Japanese) (Asian Americans, Native Americans) (Augment, Reduce) (Australia, Native American) (Australians, Native Americans) (Austrians, Germans) (Aztec, Inca) (Bessi, African American) (Blasians, caucasian) (Blasians, caucasian) (Blood, sweat) (Bohemian, asian) (Brazil, Mexico) (British, American) (Brown, white) (Burgundy, pink) (Canada, United States) (Canadians, Americans) (Caribbean, Asian) (Caucasian, black) (Cherokee, asian) (chicano, asian) (Chinese, Japanese) (Christian, Muslim) (Coloured, white) (Creole, Hispanic) (Croatia, Mexico) (Cuban, Mexican) (Cubans, Mexicans) (culture, religion) (Czechs, Russians) (Danes, Swedes) (Di, asian) (Dutch, Chinese) (dwarf, tall) (Egyptians, Chinese) (England, France) (Eskimo, asian) (Ethiopian, Vietnamese) (Franks, Saxons) (French, German) (Ga, Cherokee) (German, French) (Germans, French) (Germany, Japan) (Goa, Tamil Nadu) (Greece, Egypt) (Greek, Roman) (Greeks, Romans) (Greenland, Polynesian) (Gujarat, Tamil Nadu) (Haiti, Dominican Republic) (Haitian, Mexican) (Han, Mongolian) (Hawaii, Alaska) (Hebrews, Arabs) (Hindu, Muslim) (Hispanic, asian) (Ho people, caucasian) (Hungarian, Indian) (Hungarians, Italians) (Hungary, Romania) (Huron, Navajo) (Inca, Aztec) (India, China) (Indians, caucasian) (indigenous people, asian) (Iran, Saudi Arabia) (Irish, asian) (Israel, Palestine) (Israelis, Palestinians) (Israelites, Egyptians) (Italians, Irish) (Italy, Germany) (Jamaica, Mexico) (Jew, Muslim) (Judaism, Christianity) (Kahlan, Caucasian) (Kangeanese, asian) (Kerala, Tamil Nadu) (Khmer, Vietnamese) (knife, spoon) (Korea, Japan) (Koreans, Chinese) (Kurdish, Arab) (Latin America, Asian) (Latino, caucasian) (Latvian American, Mexican American) (Lebanese, Indian) (Liu, Lee) (Lotud, asian) (Malay, Indian) (Māori, asian) (Mexican American, Native American) (Mexicans, asian) (Middle East, South American) (Missouria, Cherokee) (Mixed, asian) (Mongols, Native Americans) (monkey, human) (Moors, Native Americans) (Morocco, Japan) (Muslim, Christian) (negro, caucasian) (Nigeria, Japan) (Norwegian, Japanese) (Paiwan, caucasian) (Palestinians, Israelis) (Persians, Arabs) (Portuguese, Spanish) (pueblo, native american) (Romans, Greeks) (Russia, United States) (Russians, Chinese) (Scotland, England) (Seneca, Cherokee) (Serbian, Japanese) (Sikh, Muslim) (Sioux, Cherokee) (Slavs, asian) (South Africans, North Africans) (South Asia, North America) (South Asians, East Asians) (South Korea, North Korea) (Spaniards, Native Americans) (Stoors, asian) (Sudanese, Vietnamese) (Swedes, French) (Swiss, French) (Syria, Iraq) (Taiwanese, Japanese) (Tamil, Chinese) (Thailand, India) (Tiv people, asian) (Turks, Arabs) (Uganda, Japan) (Ukrainian, Indian) (Varciani, African) (Vellalar, asian) (Virgin Islanders, Native Americans) (Wales, Scotland) (white, black) (Whites, Blacks)

Region Pairs: (Baruch, Koran) (Aide, Minister) (Aillen, Human) (Alan, Abdul) (Allani, Jewish) (Am-heh, Am-seh) (Amos, Muhammad) (Ancient Egypt, Ancient Greece) (Ancient Greece, Ancient Rome) (Ancient Rome, Ancient Greece) (Angalo, Hispanic) (Ap, Protestant) (Api, Guru) (Arhat Hall, Mosque) (atheist, religious) (Aztec, Inca) (Babalon, Mecca) (Babylon, Jerusalem) (Ba-Pef, Zulu) (Baptist, Muslim) (Barrex, Orthodox) (Bible, Torah) (Bon, Tao) (Buddhism, Hinduism) (Buddhist, Hindu) (Catholic, Protestant)(paganism, islam) (Catholicism, Islam) (Christianity, Hinduism) (Confucianism, Buddhism) (criminal, innocent) (Curinus, Buddha) (De, Da) (Devi,

Shiva) (El, Allah) (Elyon, Allah) (Ezekiel, Muhammad) (Gion Faith, Islam) (Gospel, Quran) (Hadit, Quran) (Harrisme, Buddhism) (Henet, Osiris) (Hindu, Muslim) (Hinduism, Buddhism) (Hungarians, Italians) (Io, Yahweh) (Irminism, Hinduism) (Isaiah, Muhammad) (Islam, Christianity) (Isten, Allah) (Jehovah, Allah) (Jehovah, Allah) (Jen, Joe) (Jeremiah, Muhammad) (Jesus, Muhammad) (Joshua, Muhammad) (Judaism, Christianity) (Juliusun, Cleopatra) (Kemetism, Christianity) (Last God, Allah) (Māori, asian) (Mormons, Muslims) (Motoro, Indian) (Muslim, Christian) (mythology, theology) (Njame, Hindu) (Old Testament, Quran) (pagan, muslim) (Persians, Arabs) (Protestant, Catholic) (Qurai, Bible) (religion, spirituality) (Rodon, Balfour) (Roman Catholic, Protestant) (sea, desert) (Shahmaran, Siren) (shen, him) (Slavs, asian) (Soma, Hinduism) (Sua, Hindu) (Talay, Koran) (Tara, Muhammad) (Tempo, Pace) (underworld, heaven) (witchcraft, islam) (Xuban, Hindu)

Profession Pairs: (academia, femininity) (actor, actress) (actress, actor) (Amateur, Professional) (amateur, professional) (Amen, Awoman) (anarchy, monarchy) (Ancient Egypt, Ancient Rome) (Ancient Greece, Ancient Rome) (anus, vagina) (apostle, apostleess) (apprentices, trainees) (archaeologist, archaeologistess) (associate, assistant) (Astronomer, Astronomeress) (baltist, baptist) (biologist, biologista) (Brahmin, Brahmini) (Brother, Sister) (Buddhist, Christian) (burgess, lady) (Caliph, Calipha) (caregiver, caretaker) (carrier, carrieress) (carver, sculptor) (Catholic Church, Anglican Church) (chemist, chemistess) (coach, coachess) (co-driver, co-driveress) (co-minister, co-ministeress) (communism, capitalism) (Composer, Composress) (composer, composress) (cook, chef) (cooper, cooperess) (counselor, counsellor) (courier, couriere) (criminal, victim) (criminality, femininity) (criticism, praise) (cup-bearer, cup-beareress) (daughter, son) (Dealer, Dealeress) (dealer, dealeress) (Dean, Deaness) (demon, angel) (Designer, designeress) (disciple, apostle) (distributor, distributress) (diver, diveress) (DJ, DJane) (duke, duchess) (emperor, empress) (empress, emperor) (engineer, engineeress) (exploration, discovery) (explorer, exploreress) (factor, factress) (fiduciary, trustee) (free-thought, feminist) (French, English) (Georgia, Florida) (girlfriend, boyfriend) (grandmother, grandfather) (groom, bridegroom) (Heroine, Hero) (horse, mare) (host, hostess) (Hostess, Host) (husband, wife) (insurer, insuree) (interpreter, translator) (Iran, Iraq) (Japanese, Korean) (jihad, crusade) (journalist, journalistess) (KGB, FBI) (king, queen) (knight, dame) (laborer, laboreress) (Landherr, Landfrau) (Lawyer, Attorney) (leader, follower) (learner, teacher) (Leipzig, Berlin) (local authority, local government) (Lord, Lady) (loyalist, patriot) (madam, sir) (major, lieutenant colonel) (Maker, Fmaker) (manufacturer, manufacturess) (Marxist, feminist) (Master, Mistress) (mate, matron) (mathematician, mathematicianess) (merchant marine, merchant mariner) (messenger, messengeress) (Messiah, Mary) (military, civilian) (monarch, queen) (Monsieur, Madame) (monster, fairy) (mule, mare) (Musician, singer) (mystic, psychic) (Novelists, Novelistes) (observer, observee) (parent, child) (partner, spouse) (pastoral, feminine) (Patriot, Loyalist) (Performer, Performeress) (philosopher, philosopheress) (photographer, photographeuse) (planter, planteress) (plastic, plasticity) (prime minister, prime ministeress) (prince, princess) (princess, prince) (printer, printeress) (probation, parole) (queen, king) (reader, readress) (rebel, loyalist) (receiver, receiveress) (regent, queen) (reporter, journalist) (Researcher, Researcheress) (respondent, respondentess) (reviewer, reviewee) (Rick, Rachel) (rowing, swimming) (royalties, queen) (scanner, scannee) (scientist, scientistess) (shaman, shawoman) (Silicon Valley, Hollywood) (squire, lady) (Stockholm, Oslo) (student, teacher) (supervisor, supervisee) (therapist, therapistess) (Thinker, Thinkress) (toddler, infant) (tourist, touristess) (tramp, lady) (transcription, translation) (translator, translatee) (tyrant, queen) (unemployed, employed) (Vienna, Budapest) (Virgin, whore) (warden, matron) (Warden, Matron) (weaver, weavess) (wholesale, retail) (worker, housewife)