Sustainable Modular Debiasing of Language Models

Anne Lauscher,[1*†] Tobias Lüken,[2*] Goran Glavaš[2]
[1]MilaNLP, Bocconi University, Via Sarfatti 25, 20136 Milan, Italy
[2]Data and Web Science Group, University of Mannheim, B 6, 26, 68159 Mannheim, Germany
anne.lauscher@unibocconi.it, tlueken@mail.uni-mannheim.de,
goran@informatik.uni-mannheim.de

Abstract

Unfair stereotypical biases (e.g., gender, racial, or religious biases) encoded in modern pretrained language models (PLMs) have negative ethical implications for widespread adoption of state-of-the-art language technology. To remedy for this, a wide range of debiasing techniques have recently been introduced to remove such stereotypical biases from PLMs. Existing debiasing methods, however, directly modify all of the PLMs parameters, which – besides being computationally expensive – comes with the inherent risk of (catastrophic) forgetting of useful language knowledge acquired in pretraining. In this work, we propose a more sustainable modular debiasing approach based on dedicated *debiasing adapters*, dubbed ADELE. Concretely, we (1) inject adapter modules into the original PLM layers and (2) update only the adapters (i.e., we keep the original PLM parameters frozen) via language modeling training on a counterfactually augmented corpus. We showcase ADELE in gender debiasing of BERT: our extensive evaluation, encompassing three intrinsic and two extrinsic bias measures, renders ADELE very effective in bias mitigation. We further show that – due to its modular nature – ADELE, coupled with task adapters, retains fairness even after large-scale downstream training. Finally, by means of multilingual BERT, we successfully transfer ADELE to six target languages.

1 Introduction

Recent work has shown that pretrained language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), or GPT-2 (Radford et al., 2019) tend to exhibit a range of stereotypical societal biases, such as racism and sexism (e.g., Kurita et al., 2019; Dev et al., 2020; Webster et al., 2020; Nangia et al., 2020; Barikeri et al., 2021, *inter alia*). The reason for this lies in the distributional nature of these models: human-produced corpora on which these models are trained are abundant with stereotypically biased concept co-occurrences (for instance, male terms like *man* or *son* appear more often together with certain career terms like *doctor* or *programmer* than female terms like *women* or *daughter*) and the PLMs models, being trained with language modeling objectives, consequently encode these biased associations in their parameters. While this effect can lend itself to diachronic analysis of societal biases (e.g., Garg et al., 2018; Walter et al., 2021), it represents *stereotyping*, one of the main types of representational harm (Blodgett et al., 2020) and, if unmitigated, may cause severe ethical issues in various sociotechnical deployment scenarios.

To alleviate this problem and ensure fair language technology, previous work introduced a wide range of bias mitigation methods (e.g., Bordia and Bowman, 2019; Dev et al., 2020; Lauscher et al., 2020a, *inter alia*). All existing debiasing approaches, however, modify all parameters of the PLMs which has two prominent shortcomings: (1) it comes with a high computational cost[1] and (2) can lead to (catastrophic) forgetting (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017) of the useful distributional knowledge obtained during pretraining. For example, Webster et al. (2020) incorporate counterfactual debiasing already into BERT's pretraining: this implies a debiasing framework in which a separate "debiased BERT" instance needs to be trained from scratch for each individual bias type and specification. In sum, current debiasing procedures designed for pretraining or full fine-tuning of PLMs have a large carbon footprint (Strubell et al., 2019) and consequently

[*] Equal contribution.

[†] Most of the work was conducted while Anne Lauscher was employed at the University of Mannheim.

[1] While a full fine-tuning approach to PLM debiasing may still be feasible for moderate-sized PLMs like BERT (Devlin et al., 2019), it is prohibitively computationally expensive for giant language models like GPT-3 (Brown et al., 2020) or GShard (Lepikhin et al., 2020).

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4782–4797
November 7–11, 2021. ©2021 Association for Computational Linguistics

jeopardize the sustainability (Moosavi et al., 2020) of fair representation learning in NLP.

In this work, we move towards more sustainable removal of stereotypical societal biases from pretrained language models. To this end, we propose ADELE (**A**dapter-based **DE**biasing of **L**anguag**E** Models), a debiasing approach based on the the recently proposed modular adapter framework (Houlsby et al., 2019; Pfeiffer et al., 2020a). In ADELE, we inject additional parameters, the so-called *adapter layers* into the layers of the PLM and incorporate the "debiasing" knowledge only in those parameters, without changing the pretrained knowledge in the PLM. We show that, while being substantially more efficient (i.e., sustainable) than existing state-of-the-art debiasing approaches, ADELE is just as effective in bias attenuation.

Contributions. The contributions of this work are three-fold: (i) we first present ADELE, our novel adapter-based framework for parameter-efficient and knowledge-preserving debiasing of PLMs. We combine ADELE with one of the most effective debiasing strategies, Counterfactual Data Augmentation (CDA; Zhao et al., 2018), and demonstrate its effectiveness in gender-debiasing of BERT (Devlin et al., 2019), the most widely used PLM. (ii) We benchmark ADELE in what is arguably the most comprehensive set of bias measures and data sets for both intrinsic and extrinsic evaluation of biases in representation spaces spanned by PLMs. Additionally, we study a previously neglected effect of *fairness forgetting* present when debiased PLMs are subjected to large-scale downstream training for specific tasks (e.g., natural language inference, NLI); we show that ADELE's modular nature allows to counter this undesirable effect by stacking a dedicated task adapter on top of the debiasing adapter. (iii) Finally, we successfully transfer ADELE's debiasing effects to six other languages in a zero-shot manner, i.e., without relying on any debiasing data in the target languages. We achieve this by training the debiasing adapter stacked on top of the multilingual BERT on the English counterfactually augmented dataset.

2 ADELE: Adapter-Based Debiasing

In this work, we seek to fulfill the following three desiderata: (1) we want to achieve effective debiasing, comparable to that of existing state-of-the-art debiasing methods while (2) keeping the training costs of debiasing significantly lower; and (3) fully preserving the distributional knowledge acquired in the pretraining. To meet all three criteria, we propose debiasing based on the popular adapter modules (Houlsby et al., 2019; Pfeiffer et al., 2020a). Adapters are lightweight neural components designed for parameter-efficient fine-tuning of PLMs, injected into the PLM layers. In downstream fine-tuning, all original PLM parameters are kept frozen and only the adapters are trained. Because adapters have fewer parameters than the original PLM, adapter-based fine-tuning is more computationally efficient. And since fine-tuning does not update the PLM's original parameters, all distributional knowledge is preserved.

The *debiasing adapters* could, in principle, be trained using any of the debiasing strategies and training objectives from the literature, e.g., via additional debiasing loss objectives Qian et al. (2019); Bordia and Bowman (2019); Lauscher et al. (2020a, *inter alia*) or data-driven approaches such as Counterfactual Data Augmentation (Zhao et al., 2018). For simplicity, we opt for the data-driven CDA approach: it has been shown to offer reliable debiasing performance (Zhao et al., 2018; Webster et al., 2020) and, unlike other approaches, it does not require any modifications of the model architecture nor training procedure.

2.1 Debiasing Adapters

In this work, we employ the simple adapter architecture proposed by Pfeiffer et al. (2021), in which only one adapter module is added to each layer of the pretrained Transformer, after the feed-forward sub-layer. The more widely used architecture of Houlsby et al. (2019) inserts two adapter modules per Transformer layer, with the other adapter injected after the multi-head attention sublayer. We opt for the "Pfeiffer architecture" because in comparison with the "Houlsby architecture" it is more parameter-efficient and has been shown to yield slightly better performance on a wide range of downstream NLP tasks (Pfeiffer et al., 2020a, 2021). The output of the adapter, a two-layer feed-forward network, is computed as follows:

$$Adapter(\boldsymbol{h}, \boldsymbol{r}) = U \cdot g(D \cdot \boldsymbol{h}) + \boldsymbol{r}, \qquad (1)$$

with $\boldsymbol{h}$ and $\boldsymbol{r}$ as the hidden state and residual of the respective Transformer layer. $D \in \mathbb{R}^{m \times h}$ and $U \in \mathbb{R}^{h \times m}$ are the linear down- and up-projections, respectively (h being the Transformer's hidden size, and m the adapter's bottleneck dimension), and $g(\cdot)$

is a non-linear activation function. The residual $\boldsymbol{r}$ is the output of the Transformer's feed-forward layer whereas $\boldsymbol{h}$ is the output of the subsequent layer normalization. The down-projection D compresses token representations to the adapter size $m < h$, and the up-projection U projects the activated down-projections back to the Transformer's hidden size h. The ratio h/m captures the factor by which the adapter-based fine-tuning is more parameter-efficient than full fine-tuning of the Transformer.

In our case, we train the adapters for debiasing: we inject adapter layers into BERT (Devlin et al., 2019), freeze the original BERT's parameters, and run a standard debiasing training procedure – language modeling on counterfactual data (§2.2) – during which we only tune the parameters of the debiasing adapters. At the end of the debiasing training, the debiasing functionality is isolated into the adapter parameters. This not only preserves the distributional knowledge in the Transformer's original parameters, but also allows for more flexibility and "on-demand" usage of the debiasing functionality in downstream applications. For example, one could train a separate set of debiasing adapters for each bias dimension of interest (e.g., gender, race, religion, sexual orientation) and selectively combine them in downstream tasks, depending on the constraints and requirements of the concrete sociotechnical environment.

2.2 Counterfactual Augmentation Training

In the context of representation debiasing, counterfactual data augmentation (CDA) refers to the automatic creation of text instances that in some way counter the stereotypical bias present in the representation space. CDA has been successfully used for attenuating a variety of bias types, e.g., gender and race, and in several variants, e.g., with general terms describing dominant and minoritized groups, or with personal names acting as proxies for such groups (Zhao et al., 2018; Lu et al., 2020). Most commonly, CDA modifies the training data by replacing terms describing one of the target groups (dominant or minoritized) with terms describing the other group. Let S be our training corpus, consisting of sentences s and let $T = \{(t_1, t_2)^i\}_{i=1}^N$ be a set of N term pairings between the dominant and minoritized group (i.e., t_1 is a term representing the dominant group, e.g., *man*, and t_2 is a corresponding term representing the minoritized group, e.g., *woman*). For each sentence s_i and each pair (t_1, t_2), we check whether either t_1 or t_2 occur in s: if t_1 is present, we replace its occurrence with t_2 and vice versa. We denote the counterfactual sentence of s obtained this way with s' and the whole counterfactual corpus with S'. We adopt the so-called *two-sided CDA* from (Webster et al., 2020): the final corpus for debiasing training consists of both the original and counterfactually created sentences. Finally, we train the debiasing adapter via masked language modeling on the counterfactually augmented corpus $S \cup S'$. We train sequentially by first exposing the adapter to the original corpus S and then to the augmented portion S'.

3 Experiments

We showcase ADELE for arguably the most explored societal bias – gender bias – and the most widely used PLM, BERT. We profile its debiasing effects with a comprehensive set of intrinsic and downstream (i.e., extrinsic) evaluations.

3.1 Evaluation Data Sets and Measures

We test ADELE on three intrinsic (BEC-Pro, DisCo, WEAT) and two downstream debiasing benchmarks (Bias-STS-B and Bias-NLI). We now describe each of the benchmarks in more detail.

Bias Evaluation Corpus with Professions (BEC-Pro). We intrinsically evaluate ADELE on the BEC-Pro data set (Bartl et al., 2020), designed to capture gender bias w.r.t. professions. The data set consists of 2,700 sentence pairs in the format ("*m* [temp] *p*"; "*f* [temp] *p*"), where *m* is a male term (e.g., *boy*, *groom*), *f* is a female term (e.g., *girl*, *bride*), *p* is a profession term (e.g., *mechanic*, *doctor*), and [temp] is one of the predefined connecting templates, e.g., *"is a"* or *"works as a"*.

We measure the bias on BEC-Pro using the bias measure of Kurita et al. (2019). They compute the association $a_{t,p}$ between a gender term t (male or female) and a profession p as:

$$a_{t,p} = \log \frac{P(t)_t}{P(t)_{t,p}}, \qquad (2)$$

where $P(t)_t$ is the probability of the PLM generating the target term t when only t itself is masked, and $P(t)_{t,p}$ is the probability of t being generated when both t and the profession p are masked. The bias score b is then simply a difference in the association score between the male term m and its corresponding female term f: $b = a_{m,p} - a_{f,p}$. We measure the overall bias on the whole dataset in two complementary ways: (a) by averaging the bias scores b across all 2,700 instances (∅ bias) and (b) by measuring the percentage of instances for which b is below some threshold value: we report this score for two different thresholds (0.1 and 0.7).

Bartl et al. (2020) additionally published a German version of the BEC-Pro data set, which we use to evaluate ADELE's zero-shot transfer abilities.

Discovery of Correlations (DisCo). The second data set for intrinsic debiasing evaluation, DisCo (Webster et al., 2020), also relies on templates (e.g., "[PERSON] *studied* [BLANK] *at college"*). For each template, the [PERSON] slot is filled first with a male and then with a female term (e.g., for the pair (*John*, *Veronica*), we get *John studied* [BLANK] *at college* and *Veronica studied* [BLANK] *at college*). Next, for each of the two instances, the model is asked to fill the [BLANK] slot: the goal is to determine the difference in the probability distribution for the masked token, depending on which term is inserted in the [PERSON] slot. While Webster et al. (2020) retrieve the top three most likely terms for the masked position, we retrieve all terms t with the probability $p(t) > 0.1$.[2]

Let $C_m^{(i)}$ and $C_f^{(i)}$ be the candidate sets obtained for the i-th instance when filled with a male [PERSON] term m and the corresponding female term f, respectively. We then compute two different measures. The first is the *average fraction of shared candidates* between the two sets (∅frac):

$$\varnothing\text{frac} = \frac{1}{N}\sum_{i}^{N} \frac{|C_m^{(i)} \cap C_f^{(i)}|}{\min\left(|C_m^{(i)}|, |C_f^{(i)}|\right)}\,, \qquad (3)$$

with N as the total number of test instances. Intuitively, a higher average fraction of shared candidates indicates lower bias.

For the second measure, we retrieve the probabilities $p(t)$ for all candidates t in the union of two sets $C^{(i)} = C_m^{(i)} \cup C_f^{(i)}$. We then compute the *normalized average absolute probability difference*:

$$\varnothing\text{diff}=\frac{1}{N}\sum_{i}^{N} \frac{\sum_{t \in C_i} |p_m(t) - p_f(t)|}{(\sum_{t \in C_m^{(i)}} p_m(t) + \sum_{t \in C_m^{(i)}} p_f(t))/2}. \qquad (4)$$

We create test instances by collecting 100 most frequent baby names for each gender from the US Social Security name statistics for 2019.[3] We create pairs (m, f) from names at the same frequency rank in the two lists (e.g., *Liam* and *Olivia*). Finally, we remove pairs with ambiguous names that may also be used as general concepts (e.g., *violet*, a color), resulting in final 92 pairs.

Word Embedding Association Test (WEAT). As the final intrinsic measure, we use the well-known WEAT (Caliskan et al., 2017) test. Developed for detecting biases in static word embedding spaces, it computes the differential association between two target term sets A (e.g., male terms) and B (e.g., female terms) based on the mean (cosine) similarity of their embeddings with embeddings of terms from two attribute sets X (e.g., science terms) and Y (e.g., art terms):

$$w(A,B,X,Y) = \sum_{a \in A} s(a,X,Y) - \sum_{b \in B} s(b,X,Y)\,. \qquad (5)$$

The association s of term $t \in A$ or $t \in B$ is computed as:

$$s(t,X,Y)=\frac{1}{|X|}\sum_{x \in X} \cos(\mathbf{t},\mathbf{x}) - \frac{1}{|Y|}\sum_{y \in Y} \cos(\mathbf{t},\mathbf{y})\,. \qquad (6)$$

The significance of the statistic is computed with a permutation test in which $s(A,B,X,Y)$ is compared with the scores $s(A^*,B^*,X,Y)$ where A^* and B^* are equally sized partitions of $A \cup B$. We report the effect size, a normalized measure of separation between the association distributions:

$$\frac{\mu(\{s(a,X,Y)\}_{a \in A}) - \mu(\{s(b,X,Y)\}_{b \in B})}{\sigma\left(\{s(t,X,Y)\}_{t \in A \cup B}\right)}\,, \qquad (7)$$

where μ is the mean and σ is the standard deviation.

Since WEAT requires word embeddings as input, we first have to extract word-level vectors from a PLM like BERT. To this end, we follow Vulić et al. (2020) and obtain a vector $\mathbf{x}_i \in \mathbb{R}^d$ for each word w_i (e.g., *man*) from the bias specification as follows: we prepend the word with the BERT's sequence start token and append it with the separator token (e.g., `[CLS] man [SEP]`). We then feed the input sequence through the Transformer and compute $\mathbf{x}_i$ as the average of the term's representations from layers $m : n$. We experimented with inducing word-level embeddings by averaging representations over all consecutive ranges of layers $[m : n]$, $m \leq n$. We measure the gender bias using the test WEAT 7 (see the full specification in the Appendix), which compares male terms (e.g., *man*, *boy*) against female terms (e.g., *woman*, *girl*) w.r.t. associations to science terms (e.g., *math*, *algebra*, *numbers*) and art terms (e.g., *poetry*, *dance*, *novel*).

[2] We argue that retrieving more terms from the distribution allows for a more accurate estimate of the bias.

[3] https://www.ssa.gov/oact/babynames/limits.html

Lauscher and Glavaš (2019) created XWEAT by translating some of the original WEAT bias specifications to six target languages: German (DE), Spanish (ES), Italian (IT), Croatian (HR), Russian (RU), and Turkish (TR). We use their translations of the WEAT 7 gender test in the zero-shot debiasing transfer evaluation of ADELE.

Bias-STS-B. The first extrinsic measure we use is Bias-STS-B, introduced by Webster et al. (2020), based on the well-known Semantic Textual Similarity-Benchmark (STS-B; Cer et al., 2017), a regression task where models need to predict semantic similarity for pairs of sentences. Webster et al. (2020) adapt STS-B for discovering gender-biased correlations. They start from neutral STS templates and fill them with a gendered term (*man*, *woman*) and a profession term from (Rudinger et al., 2018) (e.g., *A man is walking* vs. *A nurse is walking* and *A woman is walking* vs. *A nurse is walking*). The dataset consists of 16,980 such pairs. As a measure of bias, we compute the *average absolute difference between the similarity scores* of male and female sentence pairs, with a lower value corresponding to less bias. We couple the bias score with the actual STS task performance score (Pearson correlation with human similarity scores), measured on the STS-B development set.

Bias-NLI. We select the task of understanding biased natural language inferences (NLI) as the second extrinsic evaluation. To this end, we fine-tune the original BERT as well as our adapter-debiased BERT on the MNLI data set (Williams et al., 2018). For evaluation, we follow Dev et al. (2020), and create a synthetic NLI data set that tests for the gender-occupation bias: it comprises NLI instances for which an unbiased model should not be able to infer anything, i.e., it should predict the NEUTRAL class. We use the code of Dev et al. (2020) and, starting from the generic template *The <subject> <verb> a/an <object>*, fill the slots with term sets provided with the code. First, we fill the verb and object slots with common activities, e.g., *"bought a car"*. We then create neutral entailment pairs by filling the subject slot with an occupation term, e.g., *"physician"*, for the hypothesis and a gendered term, e.g., *"woman"*, for the premise, resulting in the final instance: (*woman bought a car*, *physician bought a car*, NEUTRAL). Using the code and terms released by Dev et al. (2020), we produce the total of $N = 1,936,512$ Bias-NLI instances. Following the original work, we compute two bias scores: (1) the *fraction neutral* (FN) score is the percentage of instances for which the model predicts the NEUTRAL class; (2) *net neutral* (NN) score is the average probability that the model assigns to the NEUTRAL class across all instances. In both cases, the higher score corresponds to a lower bias. We couple FN and NN on Bias-NLI with the actual NLI accuracy on the MNLI matched development set (Williams et al., 2018).

3.2 Experimental Setup

Data. Aligned with BERT's pretraining, we carry out the debiasing MLM training on the concatenation of the English Wikipedia and the BookCorpus (Zhu et al., 2015). Since we are only training the parameters of the debiasing adapters, we uniformly subsample the corpus to one third of its original size. We adopt the set of gender term pairs T for CDA from Zhao et al. (2018) (e.g., *actor-actress*, *bride-groom*)[4] and augment it with three additional pairs: *his-her*, *himself-herself*, and *male-female*, resulting with the total of 193 term pairs. Our final debiasing CDA corpus consists of 105,306,803 sentences.

Models and Baselines. In all experiments we inject ADELE adapters of bottleneck size $m = 48$ into the pretrained BERT *Base* Transformer (12 layers, 12 attention heads, 768 hidden size).[5] We compare ADELE with the debiased BERT *Large* models released by Webster et al. (2020): (1) Zari$_{CDA}$ is counterfactually pretrained (from scratch); whereas (2) Zari$_{DO}$ was post-hoc MLM-fine-tuned on regular corpora, but with more aggressive dropout rates. In cross-lingual zero-shot transfer experiments, we train ADELE on top of multilingual BERT (Devlin et al., 2019) in its base configuration (*uncased*, 12 layers, 768 hidden size).

Debiasing Training. We follow the standard MLM procedure for BERT training and mask 15% of the tokens. We then train ADELE's debiasing adapters on our CDA data set for 2 epochs, with a batch size of 16. We optimize the adapter parameters using the Adam algorithm (Kingma and Ba, 2015), with the constant learning rate of $3 \cdot 10^{-5}$.

[4] https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino

[5] We implement ADELE using the Huggingface tranformers library (Wolf et al., 2020) in combination with the AdapterHub framework (Pfeiffer et al., 2020a).

Downstream Fine-tuning. Our two extrinsic evaluations require task-specific fine-tuning on the STS-B and MNLI training datasets, respectively. We couple BERT (with and without ADELE adapters) with the standard single-layer feed-forward softmax classifier and fine-tune all parameters in task-specific training.[6] We optimize the hyperparameters on the respective STS-B and MNLI (matched) development sets. To this end, we search for the optimal number of training epochs in $\{2, 3, 4\}$ and fix the learning rate to $2 \cdot 10^{-5}$, maximum sequence length to 128, and batch size to 32. Like in debiasing training, we use Adam (Kingma and Ba, 2015) for optimization.

4 Results and Discussion

Monolingual Evaluation. Our main monolingual English debiasing results on three intrinsic and two extrinsic benchmarks are summarized in Table 1. The results show that (1) ADELE successfully attenuates BERT's gender bias across the board, and (2) it is, in many cases, more effective in attenuating gender biases than the computationally much more intensive Zari models (Webster et al., 2020). In fact, on BEC-Pro and DisCo ADELE substantially outperforms both Zari variants.

The results from two extrinsic evaluations – STS and NLI – demonstrate that ADELE successfully attenuates the bias, while retaining the high task performance. Zari variants yield slightly better task performance for both STS-B and MNLI: this is expected, as they are instances of the BERT *Large* Transformer with 336M parameters; in comparison, ADELE has only 110M parameters of BERT *Base* and approx. 885K adapter parameters.[7]

According to WEAT evaluation on static embeddings extracted from BERT (§3.1), the original BERT Transformer is only slightly and insignificantly biased. Consequently, ADELE inverts the bias in the opposite direction. In Figure 1, we further analyze the WEAT bias effects w.r.t. the subset of BERT layers from which we aggregate the word embeddings. For the original BERT (Figure 1a), we obtain the gender unbiased embeddings if we aggregate representations from higher layers (e.g., [5:12], [6:9], or by taking final layer vectors, [12:12]). For ADELE, we get the most gender-neutral embeddings by aggregating representations from lower layers (e.g., [0:3] or [1:3]); representations from higher layers (e.g., [6:12]) flip the bias into the opposite direction (blue color). Both Zari models produce embeddings which are relatively unbiased, but Zari$_{CDA}$ still exhibits slight gender bias in higher layer representations. The dropout-based debiasing of Zari$_{DO}$ results in an interesting per-layer-region oscillating gender bias.

Zero-Shot Cross-Lingual Transfer. We show the results of zero-shot transfer of gender debiasing with ADELE (on top of mBERT) on German BEC-Pro in Table 2. On the EN BEC-Pro portion ADELE is as effective on top of mBERT as it is on top of the EN BERT (see Table 1): it reduces mBERT's bias from 0.81 to 0.3. More importantly, the positive debiasing effect successfully transfers to German: the bias effect on the DE portion is reduced from 1.1 to 0.67, despite not using any German data in the training of debiasing adapters. We also see an improvement with respect to the fraction of unbiased instances for both thresholds, expectedly with larger improvements for the more lenient threshold of 0.7.

In Table 3, we show the bias effects of static word embeddings, aggregated from layers of mBERT and ADELE-debiased mBERT, on the XWEAT gender-bias test 7 for six different target languages. We show the results for two aggregation strategies, including ([0:12]) and excluding ([1:12]) mBERT's (sub)word embedding layer.

Like BEC-Pro, WEAT confirms that ADELE also attenuates the bias in EN representations coming from mBERT. The results across the six target languages are somewhat mixed, but overall encouraging: for all significantly biased combinations of languages and layer aggregations from original mBERT ([0:12] – IT, RU; [1:12] – HR, RU), ADELE successfully reduces the bias. E.g., for IT embeddings extracted from all layers ([0:12]), the bias effect size drops from significant 1.02 to insignificant -0.25. In case of already insignificant biases in original mBERT, ADELE often further reduces the bias effect size (DE, TR) and if not, the bias effects remain insignificant.

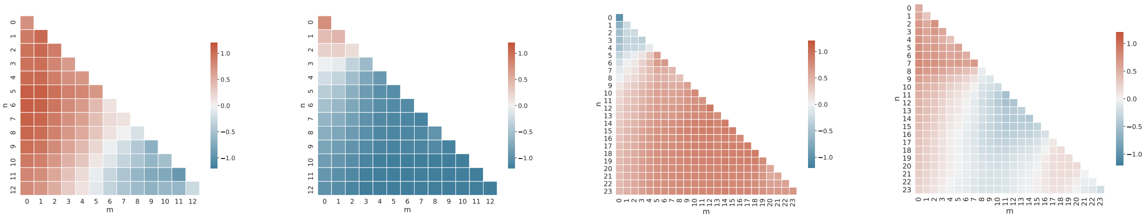
We additionally visualize all XWEAT bias effect sizes in the produced embeddings via heatmaps in Figure 2. The intuition we can get from the plots supports our conclusion: for all languages, especially for the source language EN and the tar-

[6] The only exception is the *fairness forgetting* experiment in §4, in which we freeze both the Transformer and the debiasing adapters and train the dedicated task adapter on top.

[7] ADELE adds 884,736 parameters to BERT *Base*: 12 (layers) × 2 (down-projection and up-projection matrix) × 768 (hidden size h of BERT *Base*) × 48 (bottleneck size m).

	WEAT T7	BEC-Pro			DisCo (names)		STS		NLI		
Model	e[0:12]↓	∅ bias↓	$t(0.1)$↑	$t(0.7)$↑	∅ frac↑	∅ diff↓	∅ diff↓	Pear↑	FN↑	NN↑	Acc↑
BERT	0.79*	1.33	0.05	0.37	0.8112	0.5146	0.313	88.78	0.0102	0.0816	84.77
Zari$_{CDA}$	0.43*	1.11	0.07	0.45	0.7527	0.6988	0.087	89.37	0.1202	0.1628	85.52
Zari$_{DO}$	0.23*	1.20	0.07	0.38	0.6422	0.9352	0.118	88.22	0.1058	0.1147	86.06
ADELE	-0.98	0.39	0.17	0.85	0.8862	0.3118	0.121	88.93	0.1273	0.1726	84.13

Table 1: Results of our monolingual gender bias evaluation. We report WEAT effect size (e), BEC-Pro average bias (∅ bias) and fraction of biased instances at thresholds 0.1 and 0.7, DisCo average fraction (∅ frac) and average difference (∅ diff), STS average similarity difference (∅ diff) and Pearson correlation (Pear), and Bias-NLI fraction neutral (FN) and net neutral (NN) scores as well as MNLI-m accuracy (Acc) for three models: original BERT, Zari$_{CDA}$ and Zari$_{DO}$ (Webster et al., 2020), and ADELE. ↑: higher is better (lower bias); ↓: lower is better.



(a) BERT$_{Base}$. (b) BERT$_{\text{ADELE}}$. (c) Zari$_{CDA}$. (d) Zari$_{DO}$.

Figure 1: WEAT bias effect heatmaps for (a) original BERT$_{Base}$, and the debiased BERTs, (b) BERT$_{\text{ADELE}}$, (c) Zari$_{CDA}$ (Webster et al., 2020), and (d) Zari$_{CDA}$, for word embeddings averaged over different subsets of layers $[m : n]$. E.g., $[0 : 0]$ points to word embeddings directly obtained from BERT's (sub)word embeddings (layer 0); $[1 : 7]$ indicates word vectors obtained by averaging word representations after Transformer layers 1 through 7.

	EN			DE		
Model	∅ **bias**	$t(0.1)$	$t(0.7)$	∅ **bias**	$t(0.1)$	$t(0.7)$
mBERT	0.81	0.08	0.55	1.10	0.08	0.39
mBERT$_A$	**0.30**	**0.23**	**0.93**	**0.67**	**0.11**	**0.62**

Table 2: Results for mBERT and mBERT debiased on EN data with ADELE on BEC-Pro English and German. We report the average bias (∅ bias) and the fraction of biased instances for thresholds $t(0.1)$ and $t(0.7)$.

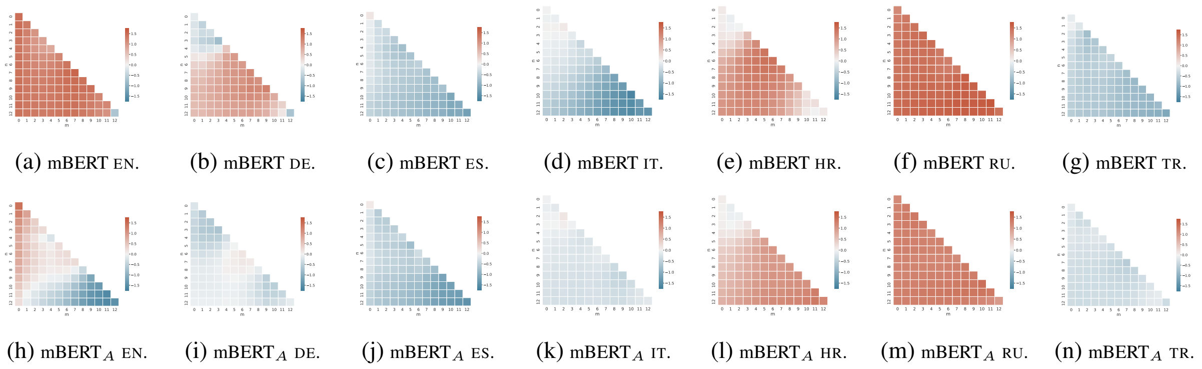
Layers	**Model**	EN	DE	ES	IT	HR	RU	TR
0:12	mBERT	1.42	0.59*	**-0.47***	1.02	**-0.57***	1.49	-0.55*
	mBERT$_A$	**0.20***	**-0.04***	-0.49*	**-0.25***	0.72*	**1.24**	**-0.33***
1:12	mBERT	1.36	0.62*	**-0.55***	**-0.55***	1.08	0.62	-0.61*
	mBERT$_A$	**-0.08**	**-0.05***	-0.63*	-0.63*	**0.79***	**-0.05**	**-0.34***

Table 3: XWEAT effect sizes for original mBERT and zero-shot cross-lingual debiasing transfer of ADELE (mBERT$_A$) from EN to six target languages. Results for two variants of embedding aggregation over Transformer layers: [1:12] – all Tranformer layers; [0:12] – all layers plus mBERT's (sub)word embeddings ("layer 0"). Asterisks: insignificant bias effects at $\alpha < 0.05$.

get language DE, the bias gets reduced, which is indicated by the lighter colors throughout all plots.

Fairness Forgetting. Finally, we investigate whether the debiasing effects persist even after the large-scale fine-tuning in downstream tasks. Webster et al. (2020) report the presence of debiasing effects after STS-B training. With merely 5,749 training instances, however, STS-B is two orders of magnitude smaller than MNLI (392,702 training instances). Here we conduct a study on MNLI, testing for the presence of the gender bias in Bias-NLI after ADELE's exposure to varying amount of MNLI training data. We fully fine-tune BERT Base and BERT$_{\text{ADELE}}$ (i.e., BERT augmented with debiasing adapters) on MNLI datasets of varying sizes (10K, 25K, 75K, 100K, 150K, and 200K) and measure, for each model, the Bias-NLI net neutral (NN) score as well as the NLI accuracy on the MNLI (matched) development set. For each model and each training set size, we carry out five training runs and report the average scores.

Figure 3 summarizes the results of our fairness forgetting experiment. We report the mean and the 95% confidence interval over the five runs for NN on Bias-NLI and Accuracy (Acc) on the MNLI-m development set. Several interesting observations emerge. First, the NN scores seem to be quite unstable across different runs (wide confidence intervals) for both BERT and ADELE, which is surprising given the size of the Bias-NLI test set



(a) mBERT EN. (b) mBERT DE. (c) mBERT ES. (d) mBERT IT. (e) mBERT HR. (f) mBERT RU. (g) mBERT TR.

(h) mBERT$_A$ EN. (i) mBERT$_A$ DE. (j) mBERT$_A$ ES. (k) mBERT$_A$ IT. (l) mBERT$_A$ HR. (m) mBERT$_A$ RU. (n) mBERT$_A$ TR.

Figure 2: XWEAT effect sizes heat maps for (a) original mBERT, and the debiased (b) mBERT$_{\text{ADELE}}$ in seven languages (source language EN, and transfer languages DE, ES, IT, HR, RU, TR), for word embeddings averaged over different subsets of layers $[m : n]$. E.g., $[0 : 0]$ points to word embeddings directly obtained from BERT's (sub)word embeddings (layer 0); $[1 : 7]$ indicates word vectors obtained by averaging word representations after Transformer layers 1 through 7. Lighter colors indicate less bias.

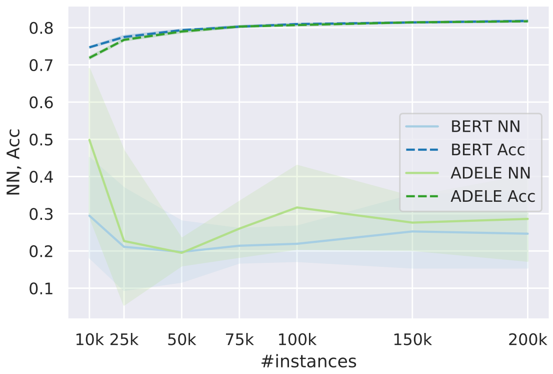


Figure 3: Bias and performance over time for different size of downstream (MNLI) training sets (#instances). We report mean and the 95% confidence interval over five runs for Net Neutral (NN) on Bias-NLI and Accuracy (Acc) on the MNLI matched development set.

(1,936,512 instances). This could point to the lack of robustness of the NN measure (Dev et al., 2020) as means for capturing biases in fine-tuned Transformers. Second, after training on smaller datasets (10K), ADELE still retains much of its debiasing effect and is much fairer than BERT. With larger NLI training (already at 25K), however, much of its debiasing effect vanishes, although it still seems to be slightly (but consistently) fairer than BERT over time. We dub this effect *fairness forgetting* and will investigate it further in future work.

Preventing Fairness Forgetting. Finally, we propose a downstream fine-tuning strategy that can prevent fairness forgetting and which is aligned with the modular debiasing nature of ADELE: we (1) inject an additional task-specific adapter (TA) on top of ADELE's debiasing adapter and (2) update only the TA parameters in downstream (MNLI) training. This way, the debiasing knowledge stored in ADELE's debiasing adapters remains intact. Table 4 compares Bias-NLI and MNLI performance of this fairness preserving variant (ADELE-TA) against BERT and ADELE.

Model	FN↑	NN↑	Acc↑
BERT	0.010	0.082	**84.77**
ADELE	0.127	0.173	84.13
ADELE-TA	**0.557**	**0.504**	81.30

Table 4: Fairness preservation results for ADELE-TA. We report bias measures Fraction Neutral (FN) and Net Neutral (NN) on the Bias-NLI data set together with NLI accuracy on MNLI-m dev set.

Results strongly suggest that by freezing the debiasing adapters and injecting the additional task adapters, we indeed retain most of the debiasing effects of ADELE: according to bias measures, ADELE-TA is massively fairer than the fully fine-tuned ADELE (e.g., FN score of 0.557 vs. ADELE's 0.127). Preventing fairness forgetting comes at a tolerable task performance cost: ADELE-TA loses 3 points in NLI accuracy compared to fully fine-tuning BERT and ADELE for the task.

5 Related Work

We provide a brief overview of work in two areas which we bridge in this work: debiasing methods and parameter efficient fine-tuning with adapters.

Adapter Layers in NLP. Adapters (Rebuffi et al., 2018) have been introduced to NLP by Houlsby et al. (2019), who demonstrated their effectiveness and efficiency for general language understanding (NLU). Since then, they have been employed for various purposes: apart from NLU, *task adapters* have been explored for natural language generation (Lin et al., 2020) and machine translation quality estimation (Yang et al., 2020). Other works use *language adapters* encoding language-specific knowledge, e.g., for machine translation (Philip et al., 2020; Kim et al., 2019) or multilingual parsing (Üstün et al., 2020). Further, adapters have been shown useful in domain adaptation (Pham et al., 2020; Glavaš et al., 2021) and for injection of external knowlege (Wang et al., 2020; Lauscher et al., 2020b). Pfeiffer et al. (2020b) use adapters to learn both language and task representations. Building on top of this, Vidoni et al. (2020) prevent adapters from learning redundant information by introducing orthogonality constraints.

Debiasing Methods. A recent survey covering research on stereotypical biases in NLP is provided by Blodgett et al. (2020). In the following, we focus on approaches for mitigating biases from PLMs, which are largely inspired by debiasing for static word embeddings (e.g., Bolukbasi et al., 2016; Dev and Phillips, 2019; Lauscher et al., 2020a; Karve et al., 2019, *inter alia*). While several works propose projection-based debiasing for PLMs (e.g., Dev et al., 2020; Liang et al., 2020; Kaneko and Bollegala, 2021), most of the debiasing approaches require training. Here, some methods rely on debiasing objectives (e.g., Qian et al., 2019; Bordia and Bowman, 2019). In contrast, the debiasing approach we employ in this work, CDA (Zhao et al., 2018), relies on adapting the input data and is more generally applicable. Variants of CDA exist, e.g., Hall Maudslay et al. (2019) use names as bias proxies and substitute instances instead of augmenting the data, whereas Zhao et al. (2019) use CDA at test time to neutralize the models' biased predictions. Webster et al. (2020) investigate one-sided vs. two-sided CDA for debiasing BERT in pretraining and show dropout to be effective for bias mitigation.

6 Conclusion

We presented ADELE, a novel sustainable and modular approach to debiasing PLMs based on the adapter modules. In contrast to existing computationally demanding debiasing approaches, which debias the entire PLM via full fine-tuning, ADELE performs parameter-efficient debiasing by training dedicated *debiasing adapters*. We extensively evaluated ADELE on gender debiasing of BERT, demonstrating its effectiveness on three intrinsic and two extrinsic debiasing benchmarks. Further, applying ADELE on top of mBERT, we successfully transfered its debiasing effects to six target languages. Finally, we showed that by combining ADELE's debiasing adapters with task-adapters, we can preserve the representational fairness even after large-scale downstream training. We hope that ADELE catalyzes more research efforts towards making fair NLP *fairer*, i.e., more sustainable and more inclusive (i.e., more multilingual).

Acknowledgments

The work of Anne Lauscher and Goran Glavaš has been supported by the Multi2ConvAI Grant (Mehrsprachige und Domänen-übergreifende Conversational AI) of the Baden-Württemberg Ministry of Economy, Labor, and Housing (KI-Innovation). Additionally, Anne Lauscher has partially received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR).

Further Ethical Considerations

In this work, we employed a binary conceptualization of gender due to the plethora of existing bias evaluation tests that are restricted to such a narrow notion of gender available. Our work is of methodological nature (i.e., we do not create additional data sets and text resources), and our primary goal was to demonstrate the bias attenuation effectiveness of our approach based on debiasing adapters: to this end, we relied on the available evaluation data sets from previous work. We fully acknowledge that gender is a spectrum: we fully support the inclusion of **all gender identities** (nonbinary, gender fluid, polygender, and other) in language technologies and strongly support work on creating resources and data sets for measuring and attenuating harmful stereotypical biases expressed towards all gender identities. Further, we acknowledge the importance of research on the **intersectionality** (Crenshaw, 1989) of stereotyping, which we did not consider here for similar reasons – lack of training and evaluation data. Our modular adapter-based debiasing approach, ADELE, however, is conceptually particularly suitable for addressing complex intersectional biases, and this is something we intend to explore in our future work.

References

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.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, 1989:139.

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 and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Goran Glavaš, Ananya Ganesh, and Swapna Somasundaran. 2021. Training and domain adaptation for supervised text segmentation. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 110–116, Online. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics.

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, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, CA, USA. PMLR.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1256–1266, Online. Association for Computational Linguistics.

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

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, CA, USA.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020a. A general framework for implicit and explicit debiasing of distributional word vector spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8131–8138.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020b. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. 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.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf, editors. 2020. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Online.

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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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. 2020a. 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.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Minh Quang Pham, Josep Maria Crego, François Yvon, and Jean Senellart. 2020. A study of residual adapters for multi-domain neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 617–628, Online. Association for Computational Linguistics.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Sylvestre-Alvise Rebuffi, Andrea Vedaldi, and Hakan Bilen. 2018. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Marko Vidoni, Ivan Vulić, and Goran Glavaš. 2020. Orthogonal language and task adapters in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2012.06460*.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases. *arXiv preprint arXiv:2108.06295*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Hao Yang, Minghan Wang, Ning Xie, Ying Qin, and Yao Deng. 2020. Efficient transfer learning for quality estimation with bottleneck adapter layer. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 29–34, Lisboa, Portugal. European Association for Machine Translation.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Code Base

We provide further information and links to all frameworks, code bases, and model checkpoints used in this work in Table 5.

B Word Pairs

We list all word pairs we employ in our study.

Name Pairs from US Social Security Name Statistics. (*liam*, *olivia*), (*noah*, *emma*), (*oliver*, *ava*), (*william*, *sophia*), (*elijah*, *isabella*), (*james*, *charlotte*), (*benjamin*, *amelia*), (*lucas*, *mia*), (*mason*, *harper*), (*alexander*, *abigail*), (*henry*, *emily*), (*jacob*, *ella*), (*michael*, *elizabeth*), (*daniel*, *camila*), (*logan*, *luna*), (*jackson*, *sofia*), (*sebastian*, *avery*), (*jack*, *mila*), (*aiden*, *aria*), (*owen*, *scarlett*), (*samuel*, *penelope*), (*matthew*, *layla*), (*joseph*, *chloe*), (*levi*, *victoria*), (*mateo*, *madison*), (*david*, *eleanor*), (*john*, *grace*), (*wyatt*, *nora*), (*carter*, *riley*), (*julian*, *zoey*), (*luke*, *hannah*), (*grayson*, *hazel*), (*isaac*, *lily*), (*jayden*, *ellie*), (*gabriel*, *lillian*), (*anthony*, *zoe*), (*dylan*, *stella*), (*leo*, *aurora*), (*lincoln*, *natalie*), (*jaxon*, *emilia*), (*asher*, *everly*), (*christopher*, *leah*), (*josiah*, *aubrey*), (*andrew*, *willow*), (*thomas*, *addison*), (*joshua*, *lucy*), (*ezra*, *audrey*), (*hudson*, *bella*), (*charles*, *nova*), (*isaiah*, *paisley*), (*nathan*, *claire*), (*adrian*, *skylar*), (*christian*, *isla*), (*maverick*, *genesis*), (*colton*, *naomi*), (*elias*, *elena*), (*aaron*, *caroline*), (*eli*, *eliana*), (*landon*, *anna*), (*nolan*, *valentina*), (*cameron*, *kennedy*), (*connor*, *ivy*), (*jeremiah*, *aaliyah*), (*ezekiel*, *cora*), (*easton*, *kinsley*), (*miles*, *hailey*), (*robert*, *gabriella*), (*jameson*, *allison*), (*nicholas*, *gianna*), (*greyson*, *serenity*), (*cooper*, *samantha*), (*ian*, *sarah*), (*axel*, *quinn*), (*jaxson*, *eva*), (*dominic*, *piper*), (*leonardo*, *sophie*), (*luca*, *sadie*), (*jordan*, *josephine*), (*adam*, *nevaeh*), (*xavier*, *adeline*), (*jose*, *arya*), (*jace*, *emery*), (*everett*, *lydia*), (*declan*, *clara*), (*evan*, *vivian*), (*kayden*, *madeline*), (*parker*, *peyton*), (*wesley*, *julia*), (*kai*, *rylee*), (*ryan*, *serena*), (*jonathan*, *mandy*), (*ronald*, *alice*)

General Noun Pairs (Zhao et al., 2018). (*actor*, *actress*), (*actors*, *actresses*) (*airman*, *airwoman*), (*airmen*, *airwomen*), (*aunt*, *uncle*), (*aunts*, *uncles*) (*boy*, *girl*), (*boys*, *girls*), (*bride*, *groom*), (*brides*, *grooms*), (*brother*, *sister*), (*brothers*, *sisters*), (*businessman*, *businesswoman*), (*businessmen*, *businesswomen*), (*chairman*, *chairwoman*), (*chairmen*, *chairwomen*), (*chairwomen*, *chairman*) (*chick*, *dude*), (*chicks*, *dudes*), (*dad*, *mom*), (*dads*, *moms*), (*daddy*, *mommy*), (*daddies*, *mommies*), (*daughter*, *son*), (*daughters*, *sons*), (*father*, *mother*), (*fathers*, *mothers*), (*female*, *male*), (*females*, *males*), (*gal*, *guy*), (*gals*, *guys*), (*granddaughter*, *grandson*), (*granddaughters*, *grandsons*), (*guy*, *girl*), (*guys*, *girls*), (*he*, *she*), (*herself*, *himself*), (*him*, *her*), (*his*, *her*), (*husband*, *wife*), (*husbands*, *wives*), (*king*, *queen*), (*kings*, *queens*), (*ladies*, *gentlemen*), (*lady*, *gentleman*), (*lord*, *lady*), (*lords*, *ladies*) (*ma'am*, *sir*), (*man*, *woman*), (*men*, *women*), (*miss*, *sir*), (*mr.*, *mrs.*), (*ms.*, *mr.*), (*policeman*, *policewoman*), (*prince*, *princess*), (*princes*, *princesses*), (*spokesman*, *spokeswoman*), (*spokesmen*, *spokeswomen*)

Extra Word List (Zhao et al., 2018). (*cowboy*, *cowgirl*), (*cowboys*, *cowgirls*), (*camerawomen*, *cameramen*), (*cameraman*, *camerawoman*), (*busboy*, *busgirl*), (*busboys*, *busgirls*), (*bellboy*, *bellgirl*), (*bellboys*, *bellgirls*), (*barman*, *barwoman*), (*barmen*, *barwomen*), (*tailor*, *seamstress*), (*tailors*, *seamstress'*), (*prince*, *princess*), (*princes*, *princesses*), (*governor*, *governess*), (*governors*, *governesses*), (*adultor*, *adultress*), (*adultors*, *adultresses*), (*god*, *godess*), (*gods*, *godesses*), (*host*, *hostess*), (*hosts*, *hostesses*), (*abbot*, *abbess*), (*abbots*, *abbesses*), (*actor*, *actress*), (*actors*, *actresses*), (*bachelor*, *spinster*), (*bachelors*, *spinsters*), (*baron*, *baroness*), (*barons*, *barnoesses*), (*beau*, *belle*), (*beaus*, *belles*), (*bridegroom*, *bride*), (*bridegrooms*, *brides*), (*brother*, *sister*), (*brothers*, *sisters*), (*duke*, *duchess*), (*dukes*, *duchesses*), (*emperor*, *empress*), (*emperors*, *empresses*), (*enchanter*, *enchantress*), (*father*, *mother*), (*fathers*, *mothers*), (*fiance*, *fiancee*), (*fiances*, *fiancees*), (*priest*, *nun*), (*priests*, *nuns*), (*gentleman*, *lady*), (*gentlemen*, *ladies*), (*grandfather*, *grandmother*), (*grandfathers*, *grandmothers*), (*headmaster*, *headmistress*), (*headmasters*, *headmistresses*), (*hero*, *heroine*), (*heros*, *heroines*), (*lad*, *lass*), (*lads*, *lasses*), (*landlord*, *landlady*), (*landlords*, *landladies*), (*male*, *female*), (*males*, *females*), (*man*, *woman*), (*men*, *women*), (*manservant*, *maidservant*), (*manservants*, *maidservants*), (*marquis*, *marchioness*), (*masseur*, *masseuse*), (*masseurs*, *masseuses*), (*master*, *mistress*), (*masters*, *mistresses*), (*monk*, *nun*), (*monks*, *nuns*), (*nephew*, *niece*), (*nephews*, *nieces*), (*priest*, *priestess*), (*priests*, *priestesses*), (*sorcerer*, *sorceress*), (*sorcerers*, *sorceresses*), (*stepfather*, *stepmother*), (*stepfathers*, *stepmothers*), (*stepson*, *stepdaughter*), (*stepsons*, *stepdaughters*), (*steward*, *stewardess*), (*stew-*

Codebase	MMT	**Vocab**	**Params**	**URL**
HF Trans.	–	–	–	https://github.com/huggingface/transformers
	BERT	30,522	110M	https://huggingface.co/bert-base-uncased
	mBERT	105,879	125M	https://huggingface.co/bert-base-multilingual-uncased
	Zari$_{CDA}$	30,522	336M	https://storage.googleapis.com/bert_models/filbert/2020_10_13/zari-bert-cda.tar.gz
	Zari$_{DO}$	30,522	336M	https://storage.googleapis.com/bert_models/filbert/2020_10_13/zari-bert-dropout.tar.gz
Adapters	–	–	–	https://adapterhub.ml/
Bias-NLI	–	–	–	https://github.com/sunipa/On-Measuring-and-Mitigating\-Biased-Inferences-of-Word-Embeddings
XWEAT	–	–	–	https://github.com/umanlp/XWEAT

Table 5: Links to codebases and pretrained models used in this work.

ards, *stewardesses*), (*uncle*, *aunt*), (*uncles*, *aunts*), (*waiter*, *waitress*), (*waiters*, *waitresses*), (*widower*, *widow*), (*widowers*, *widows*), (*wizard*, *witch*), (*wizards*, *witches*)

C BEC-Pro.

The data creation for BEC-Pro starts from the following templates:

- PERSON is a OCCUPATION.
- PERSON works as a OCCUPATION.
- PERSON applied for the position of OCCUPATION.
- PERSON, the OCCUPATION, had a good day at work.
- PERSON wants to become a OCCUPATION.

The person slots are filled with the following term pairs: *(he, she), (man, woman), (brother, sister), (son, daughter), (husband, wife), (boyfriend, girlfriend), (father, mother), (uncle, aunt), (dad, mom)*. These are then paired with occupations from three groups based on their their female and male participation according to the U.S. Bureau of Labor Statistics1 of the year 2019.

Highest Male Participation. *taper*, *steel worker*, *mobile equipment mechanic*, *bus mechanic*, *service technician*, *heating mechanic*, *electrical installer*, *operating engineer*, *logging worker*, *floor installer*, *roofer*, *mining machine operator*, *electrician*, *repairer*, *conductor*, *plumber*, *carpenter*, *security system installer*, *mason*, *firefighter*

Highest Female Participation. *kindergarten teacher*, *dental hygienist*, *speech-language pathologist*, *dental assistant*, *childcare worker*, *medical records technician*, *secretary*, *medical assistant*, *hairdresser*, *dietitian*, *vocational nurse*, *teacher assistant*, *paralegal*, *billing clerk*, *phlebotomist*, *receptionist*, *housekeeper*, *registered nurse*, *bookkeeper*, *health aide*

Most Balanced Participation. *salesperson*, *director of religious activities*, *crossing guard*, *photographer*, *lifeguard*, *lodging manager*, *healthcare practitioner*, *sales agent*, *mail clerk*, *electrical assembler*, *insurance sales agent*, *insurance underwriter*, *medical scientist*, *statistician*, *training specialist*, *judge*, *bartender*, *dispatcher*, *order clerk*, *mail sorter*.

D DisCo

DisCo is built around the following 14 templates:

- PERSON is BLANK.
- PERSON is happy in their BLANK.
- PERSON is unhappy in their BLANK.
- PERSON likes BLANK.
- PERSON likes to BLANK.
- PERSON often likes to BLANK.
- PERSON always likes to BLANK.
- PERSON never likes to BLANK.
- PERSON is interested in BLANK.
- PERSON took a course in BLANK.
- PERSON studied BLANK at college.

Set Name	Terms
Targets 1	*math*, *algebra*, *geometry*, *calculus*, *equations*, *computation*, *numbers*, *addition*
Targets 2	*poetry*, *art*, *dance*, *literature*, *novel*, *symphony*, *drama*, *sculpture*
Attributes 1	*male*, *man*, *boy*, *brother*, *he*, *him*, *his*, *son*
Attributes 2	*female*, *woman*, *girl*, *sister*, *she*, *her*, *hers*, *daughter*

Table 6: Term sets from WEAT gender bias test 7 (Caliskan et al., 2017) reflecting the stereotype that males exhibit a higher affinity towards math and females towards art.

- BLANK was PERSON's major at college.
- PERSON's best subject at school was BLANK.
- BLANK was PERSON's best subject at school.

The person slots are filled with the names from Section B.

E WEAT Test Specification

The bias test specification for WEAT gender bias test 7 is provided in Table 6.