Bridging Fairness and Environmental Sustainability in Natural Language Processing

Marius Hessenthaler¹, Emma Strubell², Dirk Hovy³, Anne Lauscher⁴
¹Data and Web Science Group, University of Mannheim, Germany
²Language Technologies Institute, Carnegie Mellon University, U.S.
³MilaNLP, Bocconi University, Italy
⁴Data Science Group, University of Hamburg, Germany

marius.hessenthaler@web.de, strubell@cmu.edu,
dirk.hovy@unibocconi.it, anne.lauscher@uni-hamburg.de

Abstract

Fairness and environmental impact are important research directions for the sustainable development of artificial intelligence. However, while each topic is an active research area in natural language processing (NLP), there is a surprising lack of research on the interplay between the two fields. This lacuna is highly problematic, since there is increasing evidence that an exclusive focus on fairness can actually hinder environmental sustainability, and vice versa. In this work, we shed light on this crucial intersection in NLP by (1) investigating the efficiency of current fairness approaches through surveying example methods for reducing unfair stereotypical bias from the literature, and (2) evaluating a common technique to reduce energy consumption (and thus environmental impact) of English NLP models, knowledge distillation (KD), for its impact on fairness. In this case study, we evaluate the effect of important KD factors, including layer and dimensionality reduction, with respect to: (a) performance on the distillation task (natural language inference and semantic similarity prediction), and (b) multiple measures and dimensions of stereotypical bias (e.g., gender bias measured via the Word Embedding Association Test). Our results lead us to clarify current assumptions regarding the effect of KD on unfair bias: contrary to other findings, we show that KD can actually decrease model fairness.

1 Introduction

Fairness and environmental sustainability are critical to the future of human society, and, thus, also reflected by the United Nations’ 17 Sustainable Development Goals (e.g., Goal 5: Gender Equality, and Goal 13: Climate Action).¹ Accordingly, both topics are currently also active research areas in natural language processing (NLP).

On the one hand, several works have established that language representations are prone to encode and amplify stereotypical social biases (e.g., Bolukbasi et al., 2016), and, consequently, are a source of representational harm (Barocas et al., 2017; Hovy and Spruit, 2016; Shah et al., 2020). To address this issue and provide fairer language technologies, various approaches have developed methods for measuring bias (e.g., Caliskan et al., 2017; Nadeem et al., 2020; Nangia et al., 2020; Nozza et al., 2021, inter alia) as well as debiasing methods (e.g., Zhao et al., 2018; Dev and Phillips, 2019, inter alia).

On the other hand, recent advances in NLP have been fueled largely by increasingly computationally expensive pre-trained language models (PLMs). Whereas the original BERT base model has 110M parameters (Devlin et al., 2019), the Switch Transformer model, designed as a more efficient alternative to more recent PLMs, has over a trillion parameters (Fedus et al., 2022). While these models consistently obtain superior performance across a variety of NLP benchmarks (Wang et al., 2018, 2019), researchers have pointed out the increasing potential CO₂ emissions of these models. Strubell et al. (2019) estimated that pre-training a BERT base Transformer (Vaswani et al., 2017) using energy with average U.S. carbon intensity has emissions comparable to a passenger on a trans-American flight. More recent calculations confirm that the energy consumption of PLMs continues to grow along with their size (Dodge et al., 2022) and that consumption at inference time is non-negligible (Tambe et al., 2021). These findings have fueled the development of more environmentally sustainable NLP. For instance, tuning only a few new lightweight adapter layers instead of the whole architecture (e.g., Houlsby et al., 2019; Pfeiffer et al., 2020), and compressing models (e.g., Gupta and Agrawal, 2022) can reduce the energy consumption during training or inference.

However, while both fairness and sustainability are active research fields in our community,² it is

¹https://sdgs.un.org/goals

²See also the proceedings of dedicated workshops,
extremely surprising that there is so little work on the intersection of both aspects. We argue that this lack of focus is problematic, as some fairness approaches can jeopardize sustainability and sustainability approaches might hinder fairness.

For instance, Webster et al. (2020) propose a data-driven debiasing approach, which requires pre-training a fairer model from scratch. Thus, for each and every stereotype, a novel PLM must be trained, reducing environmental sustainability. Lauscher et al. (2021) pointed to potential issues of such an approach and proposed a modular, and therefore more sustainable method. In the other direction, recent work in computer vision has shown that compressed models are less robust, and can even amplify algorithmic bias (Hooker et al., 2020; Liebenwein et al., 2021). Ahia et al. (2021) investigated the relationship between pruning and low-resource machine translation, finding that pruning can actually aid generalization in this scenario by reducing undesirable memorization. However, aside from few works, there has been no systematic research on the interplay between the two fields in NLP.

Contributions. In this work, we acknowledge the potential for race conditions between fairness and environmental sustainability in NLP and call for more research on the interplay between the two fields. To shed light on the problem and to provide a starting point for fair and environmentally sustainable NLP, (1) we provide a literature overview and systematize a selection of exemplary fairness approaches according to their sustainability aspects. We show that the surveyed approaches require energy at various training stages and argue that fairness research should consider these aspects. (2) Based on work suggesting the potential of model compression to increase fairness (Xu and Hu, 2022), we take a closer look at knowledge distillation (KD; Hinton et al., 2015) as an example method targeting the environmental sustainability of language technology. In this approach, a (smaller) student model is guided by the knowledge of a (bigger) teacher model. We extensively analyze the effect of KD on intrinsic and extrinsic bias measures (e.g., Word Embedding Association Test (e.g., Caliskan et al., 2017), Bias-NLI (Dev et al., 2020)) across two tasks (Natural Language Inference and Semantic Similarity Prediction). We investigate important KD-factors, such as the number of hidden layers of the student and their dimensionality. Contrary to concurrent findings (Xu and Hu, 2022), we show that KD can actually decrease fairness. Thus, fairness in such sustainability approaches needs to be carefully monitored. We hope to inspire and inform future research into fair and environmentally sustainable language technology and make all code produced publicly available at: https://github.com/UhhDS/knowledge_distillation_fairness.

2 How Fairness Can Harm Sustainability

To illustrate the tight relationship between environmental sustainability and fairness in current NLP, we conduct an exemplary analysis of current mitigation approaches for unfair bias. Here, our goal is not to conduct an exhaustive survey, but to showcase when, why, and to what extent fairness approaches can be environmentally harmful.

2.1 Approach

We query the ACL Anthology for “debiasing” and “bias mitigation” and examine the first 20 results each. We focus on debiasing of unfair societal stereotypes in monolingual PLMs. Therefore, we exclude approaches on static embeddings, domain generalization, and solely multilingual PLMs. We also consider only papers that propose a novel adaptation or debiasing approach, and exclude papers that survey or benchmark mitigation methods (e.g., Meade et al., 2022). We remove any duplicates.

This approach left us with 8 relevant publications (out of the initial 40 ACL Anthology hits). To diversify the analysis pool, we added one more paper, based on our expert knowledge.

If a paper proposes multiple methods, we focus only on a single method. We apply a coarse-grained distinction between (a) projection-based, and (b) training-based methods. Projection-based methods follow an analytical approach in a manner similar to the classic hard debiasing (Bolukbasi et al., 2016). In contrast, training-based methods either rely on augmenting training sets (e.g., Zhao et al., 2018) or on a dedicated debiasing loss (e.g., Qian et al., 2019). For the training-based approaches, we additionally classify the stage where the authors demonstrate the debiasing.

---

As for instance common in the fact verification literature (e.g., Paul Panenghat et al., 2020)
Table 1: Overview of exemplary debiasing methods w.r.t. their efficiency. We provide information on the type of the approach (Projection vs. Training), and estimate their environmental impact in 3 classes (–) in different stages of the NLP-pipeline: 0. Pre-training, 1. Intermediate Training, 2. Fine-tuning, 3. Inference time, and Other.

### 2.2 Results and Discussion

We show the results of our analysis in Table 1.

**Underlying Debiasing Approach.** Our small survey yielded examples from a variety of approaches: the projection-based approaches are represented by (Karve et al., 2019), (Liang et al., 2020), and (Kaneko and Bollegala, 2021). These require generally only a small amount of energy ((Convolutional Neural Networks)) for the analytical computation, which, in some cases, is iteratively applied to improve debiasing performance (Ravfogel et al., 2020). In this case, each iteration will marginally decrease the efficiency. Kaneko and Bollegala (2021) explicitly couple their approach with the model fine-tuning. In contrast, the other 6 works belong to the category of training-based approaches. Here, Webster et al. (2020) and Lauscher et al. (2021) rely on CDA (Zhao et al., 2018) and Dinan et al. (2020a) use control codes to guide the biases. Barikeri et al. (2021) rely on a loss-based bias mitigation for equalizing the distance of opposing identity terms towards stereotypical attributes. Subramanian et al. (2021) use a two-player zero-sum game approach for enforcing fairness constraints and Guo et al. (2022) rely on a prompt-based approach.

**Training Stage.** For the projection-based approaches, the point in time of their application is not critical to their energy consumption. They can only be applied on a trained model (stages 1–3) and, in general, do not require much energy.

However, for the training-based approaches, the training stage is a vital factor: using them in pre-training (stage 0) corresponds to training a new model from scratch. The energy (and corresponding CO₂ emissions) to perform full PLM pretraining can vary widely. Recent estimates range from 37.3 kWh to train BERT small, to 103.5 MWh to train a 6B parameter Transformer language model (Lin et al., 2022). On the positive side, the model can then be used for a variety of applications without further debiasing, assuming that debiasing transfers (Jin et al., 2021). However, this assumption is under scrutiny (Steed et al., 2022).

Intermediate training requires less energy (3.1 kWh) as PLMs have already acquired representation capabilities. However, typically, all parameters are adjusted (e.g., 110M for BERT), and the question of transferability still applies.

Debiasing in the fine-tuning stage seems the most energy efficient (1% more parameters). They require less energy in the debiasing procedure (convolutional neural networks), but add a small overhead at inference time (ca. 1% more parameters). Whether or not they add overhead to the fine-tuning depends on whether developers tune the whole architecture.

Overall, we encourage NLP practitioners to consider the energy efficiency of their debiasing approach in addition to the effectiveness and usability. Energy and emission estimation tools can be used to better estimate the environmental impact of proposed approaches (e.g., Lacoste et al., 2019).

### 3 How Sustainability Can Harm Fairness

Xu and Hu (2022) hint at the potential of model compression to improve fairness. This finding holds promise for bridging the two fields. Unfortu-
nately, the authors partially use pre-distilled models (for which they cannot control the experimental setup), do not systematically investigate the important dimensions of compression (e.g., hidden size and initialization), and do not address the stochasticity of the training procedure. In contrast, Silva et al. (2021) and Ahn et al. (2022) demonstrate distilled models to be more biased, but either use off-the-shelf models, too, or focus on single bias dimensions and measures only. Gupta et al. (2022) start from the assumption that compression results in unfair models and show it for one setup. We provide the first thorough analysis of compression (using the example of knowledge distillation (KD; Hinton et al., 2015), employing multiple tasks, bias dimensions, and measures) and show that some of these previous assumptions do not hold.

3.1 Knowledge Distillation

The underlying idea of knowledge distillation (KD; Bucluă et al., 2006; Hinton et al., 2015) is to transfer knowledge from a (typically big, pre-trained, and highly regularized) teacher model to a (typically much smaller and untrained) student network. It has been shown that a student network which can learn from the teacher’s knowledge is likely to perform better than a small model trained without a teacher’s guidance. The knowledge transfer happens through effective supervision from the teacher, e.g., via comparing output probabilities (e.g., Hinton et al., 2015), comparing the intermediate features (e.g., Ji et al., 2021), and initializing the student’s layers from the teacher’s layers.

3.2 Experimental Setup

Throughout, we use the following setup.

Distillation Tasks, Data Sets, and Measures.

We test the effects of KD on two distillation tasks: 1) natural language inference (NLI) using the MNLI data set (Williams et al., 2018), and 2) semantic textual similarity (STS) prediction with the Semantic Textual Similarity-Benchmark (STS-B; Cer et al., 2017) data set. We chose these tasks since they are popular examples of downstream natural language understanding (NLU) tasks. There are also dedicated bias evaluation data sets and measures for the resulting models. For MNLI, we report the accuracy, and for STS the combined correlation score (average of the Pearson’s correlation coefficient and Spearman’s correlation coefficient).

Fairness Evaluation. Given that some of the existing measures have been shown to be brittle (e.g., Ethayarajh et al., 2019), we ensure the validity of our results by combining intrinsic with extrinsic measures for assessing stereotypical biases along four dimensions (gender, race, age, and illness).

Word Embedding Association Test (WEAT; Caliskan et al., 2017). WEAT is an intrinsic bias test that computes the differential association between two sets of target terms A (e.g., woman, girl, etc.), and B (e.g., man, boy, etc.), and two sets of stereotypical attribute terms X (e.g., art, poetry, etc.), and Y (e.g., science, math, etc.) based on the mean similarity of their embeddings:

\[
w(A, B, X, Y) = \sum_{a \in A} s(a, X, Y) - \sum_{b \in B} s(b, X, Y),
\]

with the association s of term t ∈ A or t ∈ B as

\[
s(t, X, Y) = \frac{1}{|X|} \sum_{x \in X} \cos(t, x) - \frac{1}{|Y|} \sum_{y \in Y} \cos(t, y).
\]

The final score is the effect size, computed as

\[
\frac{\mu(\{s(a, X, Y)\}_{a \in A}) - \mu(\{s(b, X, Y)\}_{b \in B})}{\sigma(\{s(t, X, Y)\}_{t \in A \cup B})},
\]

where \( \mu \) is the mean and \( \sigma \) is the standard deviation. To apply the measure, we follow Lauscher et al. (2021), and extract word embeddings from the PLM’s encoder, using the procedure proposed by Vulić et al. (2020). We use WEAT tests 3–106 which reflect racial (tests 3–5), gender (tests 6–8), illness (test 9), and age bias (test 10).

Sentence Embedding Association Test (SEAT; May et al., 2019). SEAT measures stereotypical bias in sentence encoders following the WEAT principle. However, instead of feeding words into the encoder, SEAT contextualizes the words of the test vocabularies via simple neutral sentence templates, e.g., “This is <word>.”, “<word> is here.”, etc. Accordingly, the final score is then based on comparing sentence representations instead of word representations. We use SEAT with the WEAT test vocabularies from tests 3–10, as before. Additionally, we use SEAT’s additional Heilman Double Bind (Heilman et al., 2004) Competent and Likable tests which reflect gender bias, and SEAT’s Angry Black Woman Stereotype (e.g., Madison, 2009) test, which reflects racial bias.

\[^6\]WEAT tests 1 and 2 consist of bias types which do not consider marginalized social groups (flowers vs. insects, and weapons vs. music instruments)
Bias-STS (Webster et al., 2020). The first extrinsic test is based on the Semantic Textual Similarity Benchmark (STS-B; Cer et al., 2017). The idea is to measure whether a model assigns a higher similarity to a stereotypical sentence pair $s_x = (s_{x1}, s_{x2})$ than to a counter-stereotypical pair $s_z = (s_{z1}, s_{z2})$. Webster et al. (2020) provide templates (e.g., “A [fill] is walking.”), which they fill with opposing gender identity terms (e.g., man, woman) and a profession term (e.g., nurse) from Rudinger et al. (2018) to obtain 16,980 gender bias test instances consisting of two sentence pairs (e.g., “A man is walking” vs. “A nurse is walking” and “A woman is walking” vs. “A nurse is walking”). We train the models on the STS-B training portion and collect the predictions on the created Bias-STS test set. We then follow Lauscher et al. (2021) and report the average absolute difference between the similarity scores of male and female sentence pairs.

Bias-NLI (Dev et al., 2020). Bias-NLI is another template-based test set, which allows for measuring the tendency of models to produce unfair stereotypical inferences in NLI. We train models on the MNLI training portions, and collect the predictions on the data set. It contains 1,936,512 instances, which we create using the authors’ original code as follows: we start from templates (“The <subject> <verb> a/an <object>”) and fill the the verb and object slots with activities (e.g., “bought a car”). To obtain a premise we fill the subject slot with an occupation (e.g., “physician”), and to obtain the hypothesis, we provide a gendered term as the subject (e.g., “woman”). The obtained premise-hypothesis pair (e.g., “physician bought a car”, “woman bought a car”) is neutral, as we can not make any assumption about the gender of the premise-subject. Accordingly, we can measure the bias in the model with the fraction neutral (FN) score — the fraction of examples for which the model predicts the neutral class — and as net neutral (NN) — the average probability that the model assigns to the neutral class across all instances. Thus, in contrast to the other measures, a higher FN or NN value indicates lower bias.

**Models and Distillation Procedure.** We start from BERT (Devlin et al., 2019) in base configuration (12 hidden layers, 12 attention heads per layer, hidden size of 768) available on the Huggingface hub (Wolf et al., 2020).\(^7\) We obtain teacher models from the PLM by optimizing BERT’s parameters on the training portions of the respective data sets. We train the models with Adam (Kingma and Ba, 2015) (cross-entropy loss for MNLI, mean-squared error loss for STS-B) for maximum 10 epochs and apply early stopping based on the validation set performance (accuracy for MNLI, combined correlation score for STS-B) with a patience of 2 epochs. We grid search for the optimal batch size $b_t \in \{16, 32\}$ and learning rate $\lambda_t \in \{2 \cdot 10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$. For ensuring validity of our results (Reimers and Gurevych, 2017) we conduct this procedure 3 times starting from different random initializations. As a result, for each of the two tasks, we obtain 3 optimized teacher models. For all distillation procedures, we use the TextBrewer (Yang et al., 2020) framework’s GeneralDistiller. We optimize the following hyperparameters: batch size $b_d \in \{64, 128\}$ and temperature $t_d \in \{4, 8\}$. We distill for maximum 60 epochs and apply early stopping based on the validation score with a patience of 4 epochs. If we initialize the students’ layers, we only apply the task-specific loss on the difference between the teacher’s and the student’s output. If no layers are initialized, we add a layer matching loss based on Maximum Mean Discrepancy (Huang and Wang, 2017). We use Adam with a learning rate of 1·10^{-4} (warm up over 10% of the total number of steps and linearly decreasing learning rate schedule).

**Dimensions of Analysis.** We focus on 3 dimensions: (1) we test the effect of reducing the number of layers of the student model and report results on students with 12–1 hidden layers for MNLI and 10–1 hidden layers for STS. All other parameters stay fixed: we set the hidden size to 768 and the number of attention heads per layer to 12 (as in the teacher). We either initialize all layers of the student randomly (for MNLI)\(^8\) or map teacher’s layers to student layers for the initialization (for MNLI and STS) according to the scheme provided in the Appendix. (2) The number of layers corresponds to a vertical reduction of the model size. Analogously, we study horizontal compression reflected by the hidden size of the layers. We analyze bias in students with a hidden size $h \in \{768, 576, 384, 192, 96\}$. Here, we fix the number of hidden layers to 4. We follow Turc et al. (2019) and set the number of self-attention heads

---

\(^7\)https://huggingface.com

---

\(^8\)Not mapping the layers, i.e., random initialization, yielded sub par performance for STS
Finally, we test the effect of the layer initialization. To this end, we constrain the student model to have 4 hidden layers, and a hidden size of 768. We then initialize each of the students layers $l_s \in [0, 4]$ (where 0 is the embedding layer) either individually or all together with the teacher’s layers $l_t \in [0, 12]$ for each experiment with the following mapping ($l_t \to l_s$): $0 \to 0, 3 \to 1, 6 \to 2, 9 \to 3,$ and $12 \to 4$. For all dimensions, we compare the students’ scores with the ones of the teacher model.

### 3.3 Results

We discuss the results of our KD analysis.

#### Varying the Number of Hidden Layers.

Figures 1a–1f show the MNLI distillation experiments, where we vary the number of student layers (without initializing them). We report the overall performance reflected by MNLI (accuracy) and the bias measured with Bias-NLI, WEAT (Tests 3–10), and SEAT (Tests 3–8, Heilman Double Bind Competent and Likable, and Angry Black Woman Stereotype). We provide the additional SEAT results (Tests 9 and 10) as well as the scores for the other tasks, STS and MNLI with initialization in the Appendix.

The accuracy indicates that we successfully ran the distillation (Figure 1a). Students with 12 hidden layers (no compression) reach roughly the same performance as their teachers. Generally, we observe that the performance variation among students is higher than among teachers, with the highest variation for students with 3 to 5 hidden layers.

Looking at the bias measures (see Figures 1b–1f), we note that the variation of the scores is even higher, especially among the teacher models. This observation suggests lower numerical stability of the bias measures tested. (The test set for Bias-NLI contains $\sim$2 Million instances, so this aspect cannot be attributed to lower test set sizes). Unsurprisingly, the bias results of the students are generally in roughly the same areas than the ones of their teachers. This shows that students inherit their teachers biases in the distillation process. Grouping the test results by measure (e.g., WEAT, etc.) and dimension (e.g., race) results in roughly the same patterns of biases measurable. E.g., in Figure 1f, the results of the aggregated tests 3, 4, and 5 follow the same pattern as the Angry Black Woman
Figure 2: Results for our KD analysis (varying hidden size) on MNLI (without initialization of student layers, 4 hidden layers). We depict (a) the accuracy on MNLI, (b) the fraction neutral and net neutral scores on Bias-NLI, (c) WEAT effect sizes averaged over tests 3, 4, 5 (race) and 6, 7, 8 (gender), (d) WEAT effect sizes for tests 9 (illness) and 10 (age), (e) SEAT effect sizes averaged over tests 6–8 (gender) and the Heilmann Double Bind tests, and (f) SEAT effect sizes for tests 3–5 (race) and the Angry Black Woman stereotype test. All results shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 96–768 hidden size student models.

Figure 3: Results MNLI-KD when varying initialization of the student layers. We depict (a) the accuracy on MNLI, (b) the FN and NN scores on Bias-NLI, and (c) WEAT effect sizes averaged over tests 3, 4, 5 (race) and 6, 7, 8 (gender). All results are averages with 90% confidence interval for the 3 teacher models (dashed lines) and students distilled from the teachers where either a single layer was initialized ([0], [1], [2], [3], or [4]) or all layers ([0, 1, 2, 3, 4]).

We hypothesize that this is due to the partially overlapping term sets. However, across measures and dimensions we find roughly the same bias behavior: students with 12 to 6 hidden layers often exhibit a higher bias than their teachers (for NLI, this corresponds to a lower FN)! The exception to this rule is WEAT test 9, illness. For most tests, the highest bias arises with 4 hidden layers. Students with lower number of layers are mostly less biased across all tests. However, from this point on, the accuracy on MNLI also drops more strongly. These findings are in stark contrast to the results of Xu and Hu (2022).

Varying the Hidden Size. We show the results of KD when varying the hidden size of the student models (number of hidden layers is fixed to 4) in Figures 2a–2f. As before, we provide additional
results in the Appendix. Generally, we note that the performance curve (Figure 2a) is again in-line with our expectations. As in the previous experiment we note high variations of the scores and students biases mostly seem to be located in roughly the same ball park as their teachers’ scores. However, we again note that the concrete behavior of the bias curves depends on bias measure and dimension. Interestingly, the curves when varying the hidden size look (with some exceptions) similar to the ones when varying the number of hidden layers. We thus hypothesize, that both vertical and horizontal compression have a similar effect on fairness.

Varying the Initialization. As a last aspect of our analysis, we look at the effect of initializing various layers of the student with the weights of the teacher. We depict some of the scores in Figures 3a–3c. Interestingly, changing the initialization has a large effect both on the MNLI accuracy, as well as on the bias measures. These findings highlight again that monitoring fairness after KD is crucial.

Overall, our findings show that the devil is in the detail. While generally, the amount of bias in the distilled models is inherited from the teacher’s biases and the biases measurable seem to roughly group by social bias dimension and measure, biases still need to be carefully tested. Most importantly, while Xu and Hu (2022) point at the potential of KD for increasing fairness, we cannot confirm this observation. In contrast, across most bias measures tested, the student models start from a higher amount of bias than the teacher.

A possible explanation for this behavior is that weak learners, i.e., models with limited capacity, generally show a stronger tendency to exploit biases in the data set during the learning process than models with higher capacity (Sanh et al., 2020).

4 Related Work

Fairness in NLP. There exists a plethora of works on increasing the fairness of NLP models, most prominently focused on the issue of unfair stereotypes in the models (e.g., Caliskan et al., 2017; Zhao et al., 2017; Dev et al., 2020; Nadeem et al., 2020, inter alia). We only provide an overview and refer the reader to more comprehensive surveys on the topic (e.g., Sun et al., 2019; Blodgett et al., 2020; Shah et al., 2020). Bolukbasi et al. (2016) were the first to point to the issue of stereotypes encoded in static word embeddings, which led to a series of works focused on measuring and mitigating these biases (e.g., Dev and Phillips, 2019; Lauscher et al., 2020a), as well as assessing the reliability of the tests (Gonen and Goldberg, 2019; Ethayarajh et al., 2019; Antoniak and Mimno, 2021; Delobelle et al., 2021; Blodgett et al., 2021). For instance, Caliskan et al. (2017) proposed the well-known WEAT. Recent works focus on measuring and mitigating bias in contextualized language representations (Kurita et al., 2019; Bordini and Bowman, 2019; Qian et al., 2019; Webster et al., 2020; Nangia et al., 2020; Sap et al., 2020) and in downstream scenarios, e.g., for dialog (e.g., Sheng et al., 2019; Nangia et al., 2020a; Barikeri et al., 2021), co-reference resolution (Zhao et al., 2018), and NLI (Rudinger et al., 2017; Dev et al., 2020). Similarly, researchers have explored multilingual scenarios (e.g., Lauscher and Glavaš, 2019; Lauscher et al., 2020c; Ahn and Oh, 2021), more fine-grained biases (Dinan et al., 2020b), and more biases, beyond the prominent sexism and racism dimensions (e.g., Zhao et al., 2018; Rudinger et al., 2018), like speciesist bias (Takeshita et al., 2022).

Sustainability in NLP. Strubell et al. (2019) have called for more awareness of NLP’s environmental impact. Reducing the energy consumption can be achieved through efficient pre-training (Di Liello et al., 2021), smaller models and employing less pre-training data considering the specific needs of the task at hand (e.g., Pérez-Mayos et al., 2021; Zhang et al., 2021). If a PLM is already in-place, one can rely on sample-efficient methods (e.g., Lauscher et al., 2020b), or refrain from fully fine-tuning the model (e.g., Houlsby et al., 2019; Pfieffer et al., 2021). Similarly, one can compress the models via distillation (e.g., Hinton et al., 2015; Sanh et al., 2019; He et al., 2021), pruning (e.g., Fan et al., 2019; Li et al., 2020; Wang et al., 2020), and quantization (e.g., Zhang et al., 2020), to increase energy-efficiency of later training stages or at inference time. A survey is provided by Gupta and Agrawal (2022). In the area of distillation, researchers have explored distillation in different setups, e.g., for a specific task (e.g., See et al., 2016), on a meta-level (e.g., He et al., 2021), or for a specific resource scenario (e.g., Wasserblat et al., 2020). Other efforts focused on accurate energy and emission measurement and provide tools for monitoring energy consumption (e.g., Lacoste et al., 2019; Cao et al., 2020). While most research in the area of NLP focuses on reducing operational costs, i.e., carbon emissions due to the energy re-
quired to develop and run models, downstream impacts of model deployment stand to have a much larger impact on the environment (Kaack et al., 2022). See Rolnick et al. (2022) for a detailed presentation of how machine learning can help to counter climate change more broadly, including a discussion of NLP applications.

**Bridging Fairness and Sustainability.** To the best of our knowledge, there are currently only few works that are located at the intersection of the two fields in NLP: Lauscher et al. (2021) proposed to use adapters for decreasing energy consumption during training-based debiasing and increasing the reusability of this knowledge, which has been proven effective by Holtermann et al. (2022). Recently, the unpublished work of Xu and Hu (2022) asks whether compression can improve fairness. In contrast, Silva et al. (2021) find that off-the-shelf distilled models, such as DistilBERT, exhibit higher biases, but do not provide a systematic evaluation of the effect of KD dimensions. Concurrent to our work, Ahn et al. (2022) demonstrate similar trends, but focus on gender bias (quantified through a single measure) and the number of hidden layers in the student, only. Starting from the assumption that compression can lead to biased models, Gupta et al. (2022) propose a fairness-increasing KD loss and demonstrate their baselines to be more biased. In a similar vein, Xu et al. (2021) discuss the robustness of BERT compression. In computer vision, researchers have shown that compression exacerbates algorithmic bias (e.g., Hooker et al., 2020). E.g., Liebenwein et al. (2021) demonstrate pruned models to be brittle to out-of-distribution points. Ahia et al. (2021) present the most relevant work in this space, exploring the low-resource double-bind: individuals with the least access to computational resources are also likely to have scarce data sources. They find that model pruning can lead to better performance on low-resource languages by reducing undesirable memorization of rare examples. This study represents a valuable step towards better understanding the intersection of fairness and sustainability. In this work, we argue that more research is needed to understand the complex relationships between the two fields.

### 5 Conclusion

Fairness and environmental sustainability are equally important goals in NLP. However, the vast majority of research in our community focuses exclusively on one of these aspects. We argue that bridging fairness and environmental sustainability is thus still an unresolved issue. To start bringing these fields together in a more holistic research on ethical issues in NLP, we conducted a two-step analysis: first, we provided an overview on the efficiency of exemplary fairness approaches. Second, we ran an empirical analysis of the fairness of KD, as a popular example of methods to enhance sustainability. We find that use of KD can actually decrease fairness, motivating our plea for research into joint approaches. We hope that our work inspires such research on the interplay between the two fields for fair and sustainable NLP.

**Acknowledgements**

This work is in part funded by the European Research Council under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). The work of Anne Lauscher is funded under the Excellence Strategy of the Federal Government and the Länder. At the time of writing, DH and AL were members of the Data and Marketing Insights unit of the Bocconi Institute for Data Science and Analysis. We thank the anonymous reviewers for their insightful comments.

**Limitations**

Our work deals with the general relationship between environmental sustainability and fairness. As a showcase, we explore the effect of KD on stereotypical bias measures. This does not imply that KD is the only or the most egregious method, and more research into other approaches is needed. In this context, we resorted to established bias measures, which treat gender as a binary variable. This is due to the limitations of the established data sets, some of which allow for measuring “classic” sexism and do not reflect the large spectrum of possible identities (Lauscher et al., 2022). We further acknowledge that, in this research, we only worked with examples to highlight the importance of considering both sustainability goals. We acknowledge that to truly understand the relationship between fairness and environmental sustainability, it requires more in-depth studies. We thus encourage future research to explore the interplay between the fields more, for other societal biases, including but not limited to queerphobia and non-binary exclusion (Dev et al., 2021), and for other sustainability and fairness approaches and aspects.
References


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


A Links to Data, Models, and Code Bases

We provide the links to datasets and code bases used in this work in Table 2. In all distillation experiments, we start from BERT in base configuration: https://huggingface.co/bert-base-uncased. Our code is provided in the GitHub repository linked in the main body of the manuscript.

A.1 Details on the Initialization

When varying the hidden layers of the student and fully initializing the layers, the layer initialization is dependent on the student's size. Following common practice, the mapping is spread across the layers of the teacher. We provide the mapping here ($l_t \rightarrow l_s$):

- 10 student layers: $0 \rightarrow 0, 2 \rightarrow 1, 3 \rightarrow 2, 4 \rightarrow 3, 5 \rightarrow 4, 6 \rightarrow 5, 7 \rightarrow 6, 8 \rightarrow 7, 9 \rightarrow 8, 10 \rightarrow 9, 12 \rightarrow 10$
- 8 student layers: $0 \rightarrow 0, 2 \rightarrow 1, 4 \rightarrow 2, 5 \rightarrow 3, 6 \rightarrow 4, 7 \rightarrow 5, 8 \rightarrow 6, 10 \rightarrow 7, 12 \rightarrow 8$
- 6 student layers: $0 \rightarrow 0, 2 \rightarrow 1, 4 \rightarrow 2, 6 \rightarrow 3, 8 \rightarrow 4, 10 \rightarrow 5, 12 \rightarrow 6$
- 5 student layers: $0 \rightarrow 0, 3 \rightarrow 1, 5 \rightarrow 2, 7 \rightarrow 3, 9 \rightarrow 4, 12 \rightarrow 5$
- 4 student layers: $0 \rightarrow 0, 3 \rightarrow 1, 6 \rightarrow 2, 9 \rightarrow 3, 12 \rightarrow 4$
- 3 student layers: $0 \rightarrow 0, 4 \rightarrow 1, 8 \rightarrow 2, 12 \rightarrow 3$
- 2 student layers: $0 \rightarrow 0, 6 \rightarrow 1, 12 \rightarrow 2$
- 1 student layer: $0 \rightarrow 0, 6 \rightarrow 1$

B Additional Results

We provide the additional results for our distillation experiments.

B.1 Varying the Number of Student Layers

The additional results for the MNLI distillation, i.e., the SEAT scores for tests 9 and 10 are shown in Figure 4. The MNLI results when initializing all layers are depicted in Figure 6.

B.2 Varying the Hidden Size

We provide the additional results for MNLI when varying the student’s hidden size (without initialization of student layers, 4 hidden layers) in Figure 5.

B.3 Varying the Initialization

We provide the additional results when varying the initialization for MNLI in Figure 7.

B.4 Semantic Textual Similarity

Finally, we also provide the results of our distillation on STS in Figure 8.
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Language Inference Data</td>
<td>MNLI</td>
<td><a href="https://huggingface.co/datasets/glue">https://huggingface.co/datasets/glue</a></td>
</tr>
<tr>
<td>Semantic Similarity Prediction Data</td>
<td>STS-B</td>
<td><a href="https://huggingface.co/datasets/glue">https://huggingface.co/datasets/glue</a></td>
</tr>
<tr>
<td>Intrinsic Bias Test Terms</td>
<td>WEAT</td>
<td><a href="https://www.science.org/doi/10.1126/science.aal4230">https://www.science.org/doi/10.1126/science.aal4230</a></td>
</tr>
<tr>
<td>Intrinsic Bias Test Sentences</td>
<td>SEAT</td>
<td><a href="https://github.com/W4ngatang/sent-bias">https://github.com/W4ngatang/sent-bias</a></td>
</tr>
<tr>
<td>General Code Base</td>
<td>Transformers</td>
<td><a href="https://transformer.huggingface.co">https://transformer.huggingface.co</a></td>
</tr>
<tr>
<td>Code Base for Distillation</td>
<td>TextBrewer</td>
<td><a href="https://github.com/airaria/TextBrewer">https://github.com/airaria/TextBrewer</a></td>
</tr>
</tbody>
</table>

Table 2: Links to the datasets and code bases used in our work.
Figure 6: Results for our KD analysis (number of hidden layers) on MNLI distillation with initialization of all layers. We depict (a) the accuracy on MNLI, (b) the fraction neutral and net neutral scores on Bias-NLI, (c) WEAT effect sizes averaged over tests 3, 4, 5 (race) and 6, 7, 8 (gender), (d) WEAT 9 and 10 (age and illness), (e) SEAT scores for tests 6, 7, 8 and the Heilmann Double Bind tests (gender), (f) SEAT scores for 3, 4, 5 and Angry Black Woman Stereotype, and (g) SEAT 9 and 10 (age and illness). All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 1–12 layer student models distilled from the teachers.
Figure 7: Results for our KD analysis (varying initialization of the student layers) on MNLI. We depict (a) WEAT Age & Illness, (b) SEAT Gender, (c) SEAT Race, and (d) SEAT Age & Illness. All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and student models distilled from the teachers where either a single layer was initialized ([1], [2], [3], or [4]) or all layers ([1, 2, 3, 4]).
Figure 8: Results for our KD analysis (number of hidden layers) on STS-B distillation with initialization of all layers. We depict (a) the correlation on STS (measured as average of the Pearson and Spearman correlation coefficients), (b) the average absolute difference on Bias-STS, (c) WEAT effect sizes averaged over tests 3, 4, 5 (race) and 6, 7, 8 (gender), (d) WEAT 9 and 10 (age and illness), (e) SEAT scores for tests 6, 7, 8 and the Heilmann Double Bind tests (gender), (f) SEAT scores for 3, 4, 5 and Angry Black Woman Stereotype, and (g) SEAT 9 and 10 (age and illness). All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 1–12 layer student models distilled from the teachers.