FairDistillation: Mitigating Stereotyping in Language Models

Pieter Delobelle[1] (✉) and Bettina Berendt[1,2]

[1] Department of Computer Science, KU Leuven, Belgium
Leuven.AI institute, Belgium
`{firstname.lastname}@kuleuven.be`
[2] Faculty of Electrical Engineering and Computer Science, TU Berlin, Germany
Weizenbaum Institute, Germany

Abstract. Large pre-trained language models are successfully being used in a variety of tasks, across many languages. With this ever-increasing usage, the risk of harmful side effects also rises, for example by reproducing and reinforcing stereotypes. However, detecting and mitigating these harms is difficult to do in general and becomes computationally expensive when tackling multiple languages or when considering different biases. To address this, we present FAIRDISTILLATION: a cross-lingual method based on knowledge distillation to construct smaller language models while controlling for specific biases. We found that our distillation method does not negatively affect the downstream performance on most tasks and successfully mitigates stereotyping and representational harms. We demonstrate that FairDistillation can create fairer language models at a considerably lower cost than alternative approaches.

Keywords: Knowledge distillation, Fairness, BERT, Language models

1 Introduction

Pre-trained transformer-based Language Models (LMs), like BERT [14], are not only pushing the state-of-the-art across many languages, they are also being deployed in various services, ranging from machine translation to internet search [14, 20, 30]. However, these deployed language models have been shown to exhibit problematic behaviour. For instance, BERT and other models (i) replicate gender stereotypes [1, 12, 28], (ii) exhibit dubious racial correlations [28] and (iii) reproduce racial stereotypes [24]. These behaviours are all present in pre-trained models that are used in a wide range of applications, which are referred to as *downstream tasks*.

Without precautions, downstream tasks could use such problematic behaviour to make biased predictions. LMs are generally finetuned for such tasks, where allocation harms (i.e. allocating or withholding a resource) might occur [3]. These can originate from the fine-tuning dataset or the pre-trained model or a combination of both. We focus on the pre-training, where representation harms (i.e. encoding stereotypes) can occur in the pre-trained LMs [3, 32].

Multiple methods have been proposed to reduce representational harms in language models [1, 32], These methods are based on *pre-processing* of the data, for example *Counterfactual Data Augmentation* (CDA) [21] or *Counterfactual Data Substitution* (CDS) [16]. In both cases, gendered words in input sequences are replaced by a predefined counterfactual, e.g. "*He is a doctor*" → "*She is a doctor*". CDA can significantly increase the training dataset, with longer training times as a consequence, so CDS-based methods replace input sequences instead. Nevertheless, both techniques require retraining the model with an augmented dataset, instead of leveraging the efforts done to train the original model.

We propose a framework for mitigating representational harms based on knowledge distillation [17], which we demonstrate on gender stereotypes. Our approach uses existing language models as a teacher, which provides a richer training signal and does not require retraining from scratch. To prevent the transfer of learnt correlations to new LMs, our framework replaces CDA's augmentation strategy with probabilistic rules between tokens. Since our approach can be performed at a fraction of the original training cost and also creates smaller models, it becomes more feasible to create domain-specific bias-controlled LMs.

In this paper, we start in section 2 with an overview of language models and fairness interventions (subsection 2.1). In section 3, we present our method to create debiased language models, which we call FairDistillation. section 4 describes the evaluation set-up and section 5 presents the results. section 6 gives an overview of future work and ethical considerations and we conclude in section 7.

2   Background

BERT [14] is a language model that is trained in two phases: (i) self-supervised *pre-training* with a Masked Language Modeling (MLM) objective and afterwards (ii) supervised *finetuning* for downstream tasks. The intuition behind the first learning task is that learning to reconstruct missing words in a sentence helps with capturing interesting semantics—and because this relies on co-occurrences it also captures stereotypes. Formally, a token x_m in the input sequence $x_1, \ldots, x_N$ is replaced by a masked token (`<mask>`) and the MLM objective is to predict the original token x_m based on the context $\mathbf{x} = x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_N$, following

$$\max_{\theta} \sum_{i=1}^{N} \mathbf{1}_{x_i = x_m} \log\left(P\left(x_i \mid \mathbf{x}; \theta\right)\right)$$

with $\mathbf{1}_{x_i = x_m}$ as an indicator function whether the token is correctly predicted. This training setup results in a good estimator of the contextualized probability of a word $P\left(x_i \mid \mathbf{x}; \theta\right)$. Aside from the MLM objective, the original BERT model also incorporated a Next Sentence Prediction (NSP) objective. Liu et al. [20] later concluded that the NSP objective did not improve training and removed it

when constructing RoBERTa. Because of this, we do not further consider this objective during distillation or evaluation.

After pre-training, the newly obtained model can be reused and finetuned for different classification and regression tasks, like sentiment analysis. Finetuning requires different datasets, that can also introduce biases that are referred to as *extrinsic* biases [11]. Mitigating extrinsic biases in downstream tasks is out of scope for this work. Nevertheless, since LMs are used both for downstream tasks and for generating contextualized embeddings, mitigating intrinsic biases is still crucial.

2.1 Mitigating intrinsic biases

Bolukbasi et al. [5] presented two intrinsic debiasing methods based on removing the observed *gender axis* in static word embeddings. Mitigating problematic correlations is more challenging for LMs because of the contextualization that models like BERT incorporate. This means that word representations from LMs cannot be considered in isolation, so mitigation strategies for word embeddings cannot be applied.

Models like BERT can only generate meaningful representations for a given sequence, so for this reason, mitigation strategies have mostly been based on Counterfactual Data Augmentation (CDA) [1, 16, 21, 32]. This strategy augments the pre-training dataset with sequences where certain words, like pronouns or names, are swapped.

Unfortunately, this requires re-training the model from scratch, which can be extremely costly and with many negative side effects [2].

One of few mitigation strategies that does not alter the training data was also presented by Webster et al. [32], namely using dropout as a regularisation method against problematic correlations. Regularisation as a means to mitigate problematic correlations thus seems a feasible option, but albeit effective, the method still requires retraining the model from scratch. It should also be noted that these efforts are mostly focused on English. Results of performing CDA on a German model were less successful, likely due to gender marking [1].

2.2 Knowledge distillation

Knowledge distillation is a method to transfer learnt knowledge from one model—originally proposed as an ensemble of models—to another, usually smaller model [7, 17]. Buciluă et al. [7] introduced this technique as model compression with an ensemble of models that are used to label a dataset. This was later adapted for neural networks [17]. The teacher outputs a label probability distribution z_i where some labels have higher probabilities, for example names or pronouns are more likely than verbs in the sentence "`<mask>` is a doctor.". To incorporate this information, a variation of the softmax function (Equation 1) can be used with a temperature T to tune the importance of these labels.
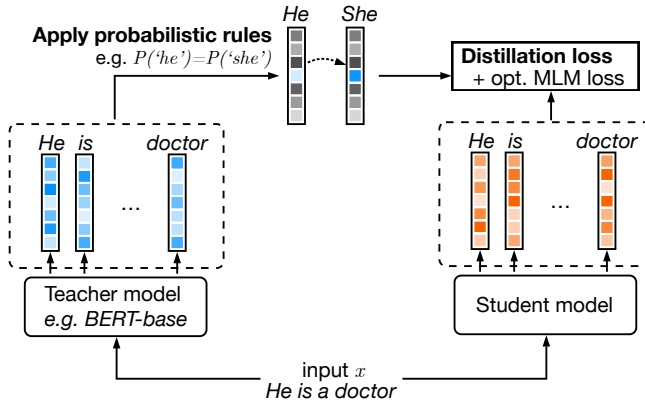
Fig. 1. Overview of the training procedure with FairDistillation for a single input sequence in English.

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}. \tag{1}$$

Sanh et al. [27] focus on the distillation of the MLM task from pre-trained LMs. Their models, DistilBERT and DistilRoBERTa, are trained on a linear combination of a distillation loss $\mathcal{L}_{ce}$ with the softening function from Equation 1, the original MLM loss $\mathcal{L}_{mlm}$, and additionally a cosine loss $\mathcal{L}_{cos}$ for the last hidden states, following

$$\mathcal{L} = \alpha_{ce}\mathcal{L}_{ce} + \alpha_{mlm}\mathcal{L}_{mlm} + \alpha_{cos}\mathcal{L}_{cos}.$$

TinyBERT [18] takes the same approach but also proposes a set of loss functions that perform distillation on (i) the embeddings layer, (ii) each of the transformer layers, and (iii) the prediction layer for specific tasks. These different loss functions make TinyBERT perform slightly better than DistilBERT, but these functions require additional transformations to be learnt. In addition, if the student and teacher have a different number of layers, a mapping function is also required to transfer the knowledge between both.

3 FairDistillation

In this section, we introduce FairDistillation, a method to mitigate problematic correlations in pre-trained language models. We first present the distillation architecture (subsection 3.1) and afterwards, we will discuss the probabilistic rules that our method relies on (subsection 3.2).

3.1   Architecture

Our method trains a newly initialized model (the student) from an already trained model (the teacher). Often, the teacher model has already been evaluated for biases, for example stereotypical gender norms for professions, which can lead to representational harms [3]. In this example, models like BERT-base [14] predict that the input sentence "`<mask>` *is a doctor.*" should be filled with '*He*' instead of '*She*'. The LM encoded that the token '*He*' is more frequent in the training dataset, both in isolation and in combination with words like '*doctor*'.

To prevent representational harms from being encoded in the final model, we apply a set of user-specified rules to the predictions of the original model. By doing so, we can train a new model with these predictions. Predictions of a teacher model provide a richer training signal and thus require less training time compared to CDA and CDS [17, 18, 27]. Moreover, we can simultaneously reduce the student's model size to improve both training and inference times, which boils down to knowledge distillation as is done for DistilBERT [27].

Figure 1 illustrates our method, which consists of 5 steps. First, an input sequence x is passed to both the teacher and the student model, both with an MLM prediction head. Second, the MLM predictions of the teacher model are passed to the rule engine. Third, the predictions for certain tokens, like '*He*', are modified based on the provided rules. Figure 1 demonstrates how a rule where we assume equal probabilities $P('He' \mid x) = P('She' \mid x)$ alters the MLM prediction, which we discuss more in-depth later in this section. Fourth, both MLM outputs, after applying possible rules to the teacher outputs, are used to calculate the distillation loss $\mathcal{L}_{ce}$ between the teacher and student outputs. Finally, the MLM outputs of the student model can also be used to calculate an additional loss term $\mathcal{L}_{mlm}$ to train the student model in the same manner as the original model.

Student architecture. The student models use the same base architecture as the teacher models, but with 6 attention layers instead of the typical 12 layers, following Sanh et al. [27]. The weights are initialized at random, which we prefer over smarter initialization strategies [27] to prevent an accidental transfer of problematic correlations. We also reuse the teacher's tokenizer for the student, since these are already specifically constructed for the targeted language and no complex token translation is needed.

Applying probabilistic rules. The MLM head outputs a vector for each position in the input sequence, so for BERT-base this means at most 512 vectors. Each value in this vector represents the probability that a token fits in this position. Consequently, there will be 30,522 values for BERT-base-uncased. We assume that some probabilities should be equal, like $P('He' \mid$ "`<mask>` is a doctor") $= P('She' \mid$ "`<mask>` is a doctor"), so our method can enforce these kind of equality rules.

During distillation, our method applies these equality rules to all the MLM outputs of the teacher. For efficiency reasons, the tokens of interest are translated into a small lookup table at the start of the distillation loop so that applying

each rule only requires a few lookup operations. The corresponding values of the tokens are set to the mean of both values. Consequently, the outcome is also normalized and each prediction still sums up to 1.

Currently, our method only supports equalization between two or more tokens. We did experiment with implementing these and more complex rules in ProbLog, a probabilistic logic programming language [10], but this proved to be unfeasible because of inference times that frequently exceeded 0.5 seconds per training example. Nevertheless, future work could focus on adding more complex rules that also depend for example on context or on part-of-speech tags to distinguish between adjectives ('His car' → 'Her car') or pronouns ('. . . is his' → '. . . is hers').

Knowledge distillation. We follow the DistilBERT [27] distillation method, as discussed in section 2. FairDistillation applies a set of rules to affect the distillation loss, but the student not only learns from the distillation task, but also from the MLM task. It is possible to concurrently train on this MLM objective for little additional cost. Although this can be another source of problematic correlations, we opted to use this loss without correcting any associations. We reason that the contextual probability for a single input sequence can also be a useful signal.

3.2 Obtaining probabilistic rules

Until now, we used a running example of a probabilistic rule where the contextualized probability, as generated by the teacher LM, has to be equal for two tokens, namely '*He*' and '*She*'. CDA achieves something similar by augmenting the dataset based on word mappings [15, 21]. These mappings are very similar to our probabilistic rules; in fact, AugLy, a popular data augmentation framework [26], has the same mapping[3] that we use for our running example in the context of gender bias.

Depending on which biases one wants to mitigate, different sets of rules are required. We focus in this work on gender bias, so we rely on the same kind of rules as CDA. Simple rules to balance predictions highlight the robustness of our method and do not require lists of professions, which come with their own issues and biases [4]. However, creating more fine-grained, domain-specific rules might improve our results. Such rules could aim at balancing, for example, profession titles or proper names.

4 Experimental setup

We evaluate our method in two Indo-European languages: (i) English and (ii) Dutch, of which the results are discussed further in section 5. Both languages have their own set of models, pre-training corpora and evaluation datasets, which we briefly cover in this section. The evaluation of gender biases is also

---

[3] `https://raw.githubusercontent.com/facebookresearch/AugLy/main/augly/assets/text/gendered_words_mapping.json`

highly language-dependent and to illustrate generalization of our method beyond English, we also used a monolingual model for Dutch [12] with an architecture similar to RoBERTa [20]. We opted for this language since it has some interesting, challenging characteristics, namely it is one of only two languages with cross-serial dependencies that make it non-context free, with the other one being Swiss-German [6]. It also has gendered suffixes for some, yet not all, nouns. This affects such evaluations since these rely on implicit associations between nouns (e.g. for professions). However, grammatical gender can also be an opportunity to evaluate how e.g. gendered professions align with the workforce [1] or with equal opportunity policies. We compare our method based on three popular metrics that we discuss in this section, Delobelle et al. [11] provides a more comprehensive overview of intrinsic fairness measures.

SEAT. The Word Embedding Association Test (WEAT) [8] measures associations between target words ('*He*', '*She*', . . .) and attribute words ('doctor', 'nurse', . . .). Between the embeddings of each target and attribute word, a similarity measure like cosine similarity can be used to quantify the association between word pairs. To add context, SEAT uses some 'semantically bleached' template sentences [24].

LPBS. Kurita et al. [19] observe that using SEAT for the learned BERT embeddings fails to find many statistically significant biases, which is addressed in the presented *log probability bias score* (LPBS). This score computes a probability p_{tgt} for a target token t (e.g. '*He*' or '*She*') from the distribution of the masked position X_m following

$$p_{tgt} = P\left(X_m = t \mid \mathbf{x}; \theta\right),$$

for a template sentence, e.g. "`<mask>` is a doctor". Since the prior likelihood $P(X_m = t)$ can skew the results, the authors correct for this by calculating a template prior p_{prior} by additionally masking the token(s) with a profession or another attribute x_p, following

$$p_{prior} = P\left(X_m = t \mid \mathbf{x}\backslash\{x_p\}; \theta\right).$$

Both probabilities are combined in a measure of association $\log \frac{p_{tgt}}{p_{prior}}$ and the bias score is the difference between the association measures for two targets, like '*He*' and 'She'. Kurita et al. [19] applied their method to the original English BERT model [14] and found statistically significant differences for all categories of the WEAT templates.

DisCo. Webster et al. [32] also utilize templates to evaluate possible biases which their approach also mitigates (see subsection 2.1). As an intrinsic measure, the authors present *discovery of correlations* (DisCo). Compared to previously discussed metrics, this metric measures the difference in predictions for the attribute token x_p when varying gendered tokens (i.e. '`<P>` *is a* `<mask>`' for different pronouns or names instead of '`<mask>` *is a* `<P>`' with different professions).

In the remainder of this section, we discuss our evaluations of English and Dutch in their respective subsections, where we define the used datasets, models and language-specific evaluation aspects.

4.1   English setup

The first model we use as a teacher is the original uncased BERT model (`BERT-base-uncased`) as released by Devlin et al. [14], which is also the most-studied LM with regard to gender stereotypes. This model was trained on the Toronto Bookcorpus and Wikipedia, but the Toronto Bookcorpus is no longer publicly available anymore and thus hinders reproduction. For this reason, Jiao et al. [18] use only Wikipedia. We used a portion of the English section of the OSCAR corpus [25] to keep the training dataset size similar. More specifically, we used the first two shards of the unshuffled version. We recognise that there is a mismatch between the domains of the Bookcorpus and OSCAR, but we believe this is acceptable to increase reproducibility.

As introduced in subsection 3.2, we use a set of gendered pronouns and define which ones should have the same probability. Since we use the uncased variant, we only need to define one set of rules, since '*She*' and '*she*' result in the same token.

The tokenization method used by BERT, WordPiece, splits words and adds a merge symbol (e.g. 'word' + '##piece'), so no special care is required. For RoBERTa [20], which uses Byte Pair Encoding (BPE), a word boundary symbol is used. Consequently, a word can have different tokens and representations depending if a space, punctuation mark, mask token or sequence start token are in front of the target token. Since the Dutch model uses BPE, we will revisit this issue in subsection 4.2.

Evaluation. To evaluate our method's performance trade-off, we finetune the obtained model on the Internet Movie Database (IMDB) sentiment analysis task [22], which was also done by BERT [14] and DistilBERT [27]. The dataset contains 25k training examples, from which we used 5k as a separate validation set, and another 25k test sequences. This is a high-level task where no gendered correlations should be used for predictions. Predicting entailment is a high-level task covered multiple times in GLUE [31], on which we also evaluated our method with the pre-trained model. For a description of this benchmark and all datasets, we refer to Wang et al. [31].

Bias evaluation. To evaluate possible problematic correlations with regard to gender stereotypes, we compute DisCo and LPBS, which we introduce at the beginning of this section. We use the Employee Salary dataset[4] [19]. Following Kurita et al. [19], we filter on the top 1000 highest-earning instances as a proxy for prestigious jobs and test this for the same two templates ('`<mask>` is a `<P>`'

---

[4] `https://github.com/keitakurita/contextual_embedding_bias_measure/blob/master/notebooks/data/employeesalaries2017.csv`

and '`<mask>` can do `<P>`'). However, we additionally filter digits from the job titles and remove duplicate titles, to not skew the results towards more popular professions.

4.2   Dutch setup

We use a Dutch RoBERTa-based model called RobBERT [12] as a teacher, more specifically `robbert-v2-dutch-base`. This model was pre-trained on the Dutch section of the shuffled version of OSCAR[5]. Similar to the distilled version of RobBERT [13], we select a 1GB portion of the OSCAR corpus (using `head`, 2.5%) to illustrate the ability to perform successful knowledge distillation with only a small fraction of the data required in comparison to the pre-trained model.

To create our model, we used a defined a set of rules based on the gendered pronouns 'Hij' and 'Zij' ('*He*' and '*She*'). The tokens corresponding to these pronouns were grouped based on capitalization and included spaces, since the BPE tokenizer includes a word boundary character at the beginning of some tokens. Our method then used these rules to equalise the distributions predicted by the teacher during distillation, which we performed for 3 epochs. This took approximately 40 hours per epoch on a Nvidia 1080 Ti and a traditionally distilled model required the same time, indicating our method has very limited effect on training time.

Evaluation. We compare the model created with FairDistillation to RobBERT and RobBERTje [12, 13] on the same set of benchmark tasks: (i) sentiment analysis on book reviews (DBRD) [29], (ii) NER, (iii) POS tagging, and (iv) natural language inference with SICK-NL [33].These tasks are fairly high-level sequence-labelling tasks that can exhibit allocational harms, such as the predictive difference for sentiment analysis that was illustrated by Delobelle et al. [12].

Bias evaluation. We also evaluate numerically using the LPBS and DisCo metrics, but the RobBERT LM has also been evaluated by the authors on gender stereotyping using a different technique. This evaluation technique is based on a set of templates and a translated set of professions[6] from Bolukbasi et al. [5]. These professions have a perceived gender (e.g. 'actress' is a female profession and 'surveyor' is neutral), which can be correlated with the predictions by the model. The authors rank the tokens based on the predicted probability instead of using this probability directly. Interestingly, a correlation was not considered problematic, but male pronouns were predicted higher on average, even for by definition female professions (e.g. 'nun'). To compare these results, we recreate the same plot and report the Mean Ranking Difference (MRD). We focus on the gendered pronouns '*zij*' ('*she*') and '*hij*' ('*he*') for our evaluation.

---

[5] `https://oscar-corpus.com`
[6] `https://people.cs.kuleuven.be/~pieter.delobelle/data.html`

Table 1. English results on IMDB (sentiment analysis), GLUE [31], and two bias measures. Following Devlin et al. [14], we report F_1 scores for QQP and MRPC, Spearman correlations for STS-B, and accuracy for all other tasks. Results reported by Devlin et al. [14] on the GLUE dashboard are indicated with an obelisk (†), while the results from [27] are also on the GLUE dev set, indicated with an asterisk (*). For LPBS, positive values represent more stereotypical associations, and for DisCo, lower values are more favorable.

Model	IMDB	GLUE								Bias	
		MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	LPBS	DisCo
BERT [14]	93.5	84.0†	71.2†	90.5†	93.5†	52.1†	85.8†	88.9†	66.4†	1.16	-0.48
DistilBERT [27]	92.82	82.2*	88.5*	89.2*	91.3*	51.3*	86.9*	87.5*	59.9*	-0.27	-0.55
FairDistillation	85.5 ± 0.4	80.1	82.1	86.6	90.6	38.5	84.0	85.1	59.6	**-0.16**	**0.25**

5　Results

In this section, we present the results of the experiments (section 4). We discuss English (subsection 5.1) and Dutch (subsection 5.2) results separately. We also performed experiments on French using the CamemBERT model [23], but we chose to ommit those results due to our limited understanding of the language, which we address further in section 6.

To eliminate any possible effect from hyperparameter assignments on the results, we ran each finetuning training 10 times with random hyperparameter assignments. We varied the (i) learning rate, (ii) weight decay, and (iii) the number of gradient accumulation steps to effectively scale the batch size while still fully utilizing the GPU. The full set of hyperparameters is listed in Table 3 in the supplementary materials. For the Dutch benchmarks and for the English IMDB, we select the best-performing model based on the *validation* set and present the results on the held-out *test* set. The results from the GLUE benchmark are from the dev set, which were also the results reported by [27].

Unless indicated otherwise, all training runs are done on a single Nvidia 1080 Ti with 11 GB VRAM. All models are also comparably sized, with 66M trainable parameters each. This is 50% of the model size of the teachers.

5.1　English results

We observe that problematic correlations are reduced on all three metrics, as is shown in Table 1. One interesting observation—which also holds for Dutch—is that distillation in itself is already successful in mitigating these correlations. This might be related to regularization as a method to control correlations [32], but we leave this for a future study.

On the IMDB task, our model suffers a 10% accuracy drop, which is significant. However, as noted in section 4, we used a smaller training set for finetuning than BERT and DistilBERT, because we created a separate validation set from the original training set.

Table 2. Dutch results on several benchmarks, namely Dutch Book Reviews (DBRD, sentiment analysis), named entity recognition (NER), part-of-speech (POS), tagging, and language inference (SICK-NL). We report bias as measured with LPBS and DisCo and additionally the mean ranking difference (MRD), which measures the preference of a language model to fill in male tokens (negative score) or female tokens (positive score). Benchmarks are reported with accuracy with 95% CI, except for the NER task, where we report the F_1 score. Results indicated with † were reported by Delobelle et al. [12]. For MRD, smaller ranking differences are more favorable, for LPBS, positive values represent more stereotypical associations, and for DisCo, lower values are more favorable.

| Model | Params | Benchmark scores | | | | Bias | | |
		DBRD	NER	POS	SICK-NL	MRD	LPBS	DisCo
RobBERT [12]	116 M	$94.4 \pm 1.0^\dagger$	$89.1^\dagger$	$96.4 \pm 0.4^\dagger$	84.2 ± 1.0	-7.47	1.13	-0.29
RobBERTje [13]	74 M	92.5 ± 1.1	82.7	$\mathbf{95.6} \pm 0.4$	$\mathbf{83.4} \pm 1.0$	-6.66	**-0.45**	-0.41
FairDistillation	74 M	92.1 ± 1.1	82.7	95.4 ± 0.4	82.4 ± 1.1	**-3.98**	1.14	**-0.08**

For GLUE, the results are in line with distilBERT. We do observe some diminished scores, notably CoLA, but the overall trade-off is limited.

Unlike the other models, we performed our FairDistillation method on 4 Nvidia V100's for 3 epochs, which took 70h per epoch. Finetuning was done on an Nvidia 1080 Ti for 4 epochs for IMDB, which took approximately 1h per run and was replicated 10 times. For GLUE, we report the dev results and did not do any hyperparameter search. We used the same hyperparameters as distilBERT [27], who also report the development set results.

5.2 Dutch results

Both the distilled RobBERT model and the model obtained with FairDistillation perform only slightly worse (within 97.5% of the original model) for both downstream tasks (see Table 2). Both models have only half the parameters compared to the original RobBERT model and are thus faster to train and deploy, making this a decent trade-off between model size and predictive performance. With no significant differences in performance between the distilled model and our FairDistillation model, this highlights the potential of our method.

With regards to the bias evaluation, we observe a reduction between the original model and ours (Table 2): correlations are significantly reduced as measured by DisCo and the mean ranking of female associated tokens improved by 3.5 tokens. The only exception is LPBS [19], which incorporates a correction based on the prior probability of a token. Our method effectively corrects this prior, while still allowing the context to affect individual results with the MLM objective. Since Dutch has gendered nouns for some professions, a correlation is not necessarily undesirable, but the prior is (e.g. assuming all physicians are men). Further graphical analysis of the predicted rankings for the third person singular pronouns confirms this, as shown in Figure 2. These charts reveal that
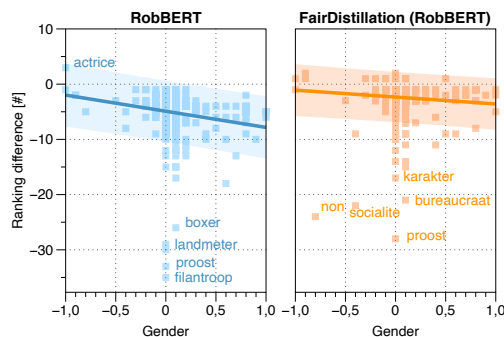
Fig. 2. Differences in predictions for the Dutch template '`<mask>` *is een* `<P>`' for our and the original RobBERT model. The 'gender' axis ranges from words associated with female professions (left) to words associated with male professions (right). A positive ranking difference indicates '*She*' is predicted before '*He*'.

most professions are now less associated with the masculine pronoun. When considering which pronoun is ranked higher (i.e. above or below $y = 0$), this result is even more pronounced. RobBERT only predicted a feminine pronoun for a single profession [12], while with our method this increased to 15 professions.

6   Limitations and ethical considerations

Despite the promising results, there are several potential improvements possible to our methods, as well as some ethical considerations. First, we rely on facts that express probabilities for a single token at a time. For gender stereotyping, this is sufficient as the vocabulary usually contains the tokens of interest. However, this is not the case for many other problematic correlations, especially those affecting minority groups. Tokens that are interesting here, like names, are not in the tokenizer's vocabulary because this is created based on occurrence counts. Addressing this limitation would require extending our method to support facts that span multiple tokens.

Second, the effects of our method on 'low-level' grammatical tasks require further study in English, as we focused on GLUE and sentiment analysis. The used Dutch benchmarks do cover more tasks and seem to indicate favourable reductions.

Third, our work focuses on binary gender stereotypes. This leaves out a wide range of people who do not identify as such and although our method supports equalization over more than two tokens, this might be challenging if the intended words span multiple tokens. This also poses a challenge for generalizing our method beyond gender bias, since this frequently involves words that are not a single token in BERT's vocabulary.

Fourth, since none of the authors is a native speaker of a language like French, we only performed a limited, exploratory evaluation of our methods

with CamemBERT [23]. Our method appears successful, with improved scores of 0.04 (DisCo) and -0.85 (LPBS) compared to -1.15 (DisCo) and 1.99 (LPBS) for the original CamemBERT model [23]. The performance of the LM was also still high, with XNLI [9] scores 75.6 compared to 82.5 for CamemBERT. However, constructing correct probabilistic rules and evaluating them is tricky for non-native speakers. For example, the female variant of a profession can refer to a woman practising said profession, but also to the spouse of a man with this profession. When discussing these results with native French speakers from Wallonia, Belgium and from northern France, we realised that we are not well-suited to address this. We thus leave a more comprehensive evaluation across languages as future work.

Finally, by presenting a method to remove correlations with gender stereotypes in pre-trained language models, we risk it being used as a 'rubber stamp' to absolve model creators from their responsibilities. Therefore, we urge creators to critically analyse LMs within the social context that these models will be deployed in, both with respect to both the pre-trained and the finetuned model.

7   Conclusion

We introduced a method called FairDistillation that allows to use probabilistic rules during knowledge distillation. We showed that this can effectively mitigate gender stereotypes in language models. Our method demonstrates that knowledge distillation of language models with probabilistic rules is a possible alternative to re-training in order to reduce representational harms. Even though comes at a slight cost for some downstream tasks, but we find that the overall cost is limited and can mostly be attributed to the distillation process.

Acknowledgements

References

[1] Bartl, M., Nissim, M., Gatt, A.: Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. arXiv:2010.14534 [cs] (Oct 2020)

[2] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021)

[3] Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476, Association for Computational Linguistics, Online (Jul 2020), `https://doi.org/10.18653/v1/2020.acl-main.485`, URL `https://www.aclweb.org/anthology/2020.acl-main.485`

[4] Blodgett, S.L., Lopez, G., Olteanu, A., Sim, R., Wallach, H.: Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. 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), pp. 1004–1015 (2021)

[5] Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs, stat] (Jul 2016)

[6] Bresnan, J., Kaplan, R.M., Peters, S., Zaenen, A.: Cross-serial dependencies in Dutch. In: The formal complexity of natural language, pp. 286–319, Springer (1982)

[7] Buciluă, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535–541 (2006)

[8] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017), ISSN 0036-8075, `https://doi.org/10.1126/science.aal4230`, URL `https://science.sciencemag.org/content/356/6334/183`

[9] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: XNLI: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2018)

[10] De Raedt, L., Kimmig, A., Toivonen, H.: ProbLog: A probabilistic prolog and its application in link discovery. In: IJCAI, vol. 7, pp. 2462–2467, Hyderabad (2007)

[11] Delobelle, P., Tokpo, E.K., Calders, T., Berendt, B.: Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. arXiv preprint arXiv:2112.07447 (2021)

[12] Delobelle, P., Winters, T., Berendt, B.: RobBERT: A Dutch RoBERTa-based Language Model. In: Findings of ACL: EMNLP 2020 (Sep 2020)

[13] Delobelle, P., Winters, T., Berendt, B.: RobBERTje: A distilled Dutch BERT model. Computational Linguistics in the Netherlands Journal **11**, 125–140 (Feb 2022), URL `https://www.clinjournal.org/clinj/article/view/131`

[14] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), `https://doi.org/10.186 53/v1/N19-1423`

[15] Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP. arXiv preprint arXiv:2105.03075 (2021)

[16] Hall Maudslay, R., Gonen, H., Cotterell, R., Teufel, S.: 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), pp. 5267–5275, Association for Computational Linguistics, Hong Kong, China (Nov 2019), `https://doi. org/10.18653/v1/D19-1530`, URL `https://www.aclweb.org/anthology /D19-1530`

[17] Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [cs, stat] (Mar 2015)

[18] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for natural language understanding. In: Findings of ACL: EMNLP 2020 (2020)

[19] Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 166–172, Association for Computational Linguistics, Florence, Italy (Aug 2019), `http s://doi.org/10.18653/v1/W19-3823`, URL `https://www.aclweb.org/a nthology/W19-3823`

[20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs] (Jul 2019)

[21] Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A.: Gender Bias in Neural Natural Language Processing, pp. 189–202. Springer International Publishing, Cham (2020), ISBN 978-3-030-62077-6, URL `https://doi.or g/10.1007/978-3-030-62077-6_14`

[22] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150, Association for Computational Linguistics, Portland, Oregon, USA (June 2011), URL `http://www.aclweb.org/antholo gy/P11-1015`

[23] Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: A tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219, Association for Computational Linguistics, Online (Jul 2020), `https://doi.org/10.18653/v1/2020.acl -main.645`

[24] May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. 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), pp. 622–628, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), `https://doi.org/10.18653/v1/N19-1063`

[25] Ortiz Suárez, P.J., Romary, L., Sagot, B.: A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1703–1714, Association for Computational Linguistics, Online (Jul 2020), URL `https://www.aclweb.org/anthology/2020.acl-main.156`

[26] Papakipos, Z., Bitton, J.: Augly: Data augmentations for robustness. arXiv preprint arXiv:2201.06494 (2022)

[27] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In: NeurIPS EMC^2 Workshop (2019)

[28] Tan, Y.C., Celis, L.E.: Assessing social and intersectional biases in contextualized word representations. In: Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 13230–13241, Curran Associates, Inc. (2019)

[29] van der Burgh, B., Verberne, S.: The merits of Universal Language Model Fine-tuning for Small Datasets – a case with Dutch book reviews. arXiv:1910.00896 [cs] (Oct 2019), URL `http://arxiv.org/abs/1910.008 96`

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, Curran Associates, Inc. (2017)

[31] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding (2019), in the Proceedings of ICLR.

[32] Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Petrov, S.: Measuring and Reducing Gendered Correlations in Pre-trained Models. arXiv:2010.06032 [cs] (Oct 2020)

[33] Wijnholds, G., Moortgat, M.: SICKNL: A dataset for Dutch natural language inference. arXiv preprint arXiv:2101.05716 (2021), URL `https://arxiv.or g/abs/2101.05716`