Don't Forget About Pronouns: Removing Gender Bias in Language Models without Losing Factual Gender Information

Anonymous ACL submission

Abstract

The representations in large language models contain various types of gender information. We focus on two types of such signals in English texts: factual gender information, which is a grammatical or semantic property, and gender bias, which is the correlation between a word and specific gender. We can disentangle the model's embeddings and identify components encoding both information with probing. We aim to diminish the representation of stereotypical bias while preserving factual gender signal. Our filtering method shows that it is possible to decrease the bias of gender-neutral profession names without deteriorating language modeling capabilities. The findings can be applied to language generation and understanding to mitigate reliance on stereotypes while preserving gender agreement in coreferences.

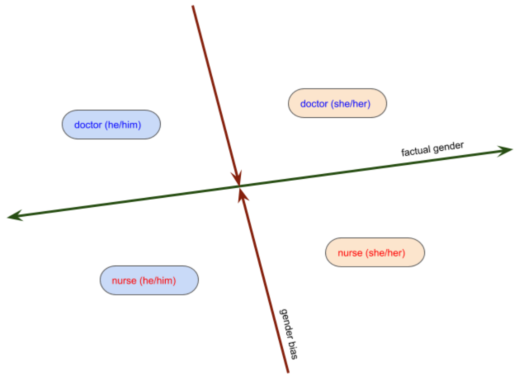


Figure 1: A schema presenting the difference between gender bias and grammatical gender in pronouns. We want to transform the representations to remove the former and preserve the latter.

1 Introduction

Neural networks are successfully applied in natural language processing. While they achieve state-of-the-art results on various tasks, their decision process is not yet fully explained (Lipton, 2018). It is often the case that neural networks base their prediction on spurious correlations learned from large uncurated datasets. An example of such spurious tendency is gender bias, even the most accurate models tend to associate some words with a specific gender unjustly (Zhao et al., 2018a; Stanovsky et al., 2019). The representations of profession names tend to be closely connected with the stereotypical gender of their holders. When the model encounters the word "nurse", it will tend to use female pronouns ("she", "her") when referring to this person in the generated text. This tendency is reversed for words such as "doctor", "professor", or "programmer", which are male-biased.

That means that the neural model is not reliable enough to be applied in high-stakes language processing tasks such as connecting job offers to applicants' CVs (De-Arteaga et al., 2019). If the underlying model was biased, the high-paying jobs, which are stereotypically associated with men, could be inaccessible for female candidates. The challenge is to ensure that the model's predictions are fair.

The recent works on the topics aimed to diminish the role of gender bias by feeding examples of unbiased text and training the network (de Vassimon Manela et al., 2021) or transforming the representations of the neural networks post-hoc (without any additional training) (Bolukbasi et al., 2016). However, those works relied on the notion that to debias representation all gender signal needs to be eliminated. It is not always the case, pronouns and a few other words (e.g.: "king" - "queen"; "boy" - "girl") have factual information about gender. A few works separately considered gendered words and exempted them from de-biasing (Zhao et al., 2018b; Kaneko and Bollegala, 2019). In contrast to these approaches, we focus on contextual word embeddings. In contextual representations, we want to preserve the factual gender information for gender-neutral words when it is indicated by context, e.g., personal pronoun. This sort of information needs to be maintained in the representations. In language modeling, the network needs to be consistent about the gender of a person if it was revealed earlier in the text. The model's ability to encode factual gender information is crucial for that purpose.

We propose a method for disentangling the factual gender information and gender bias encoded in the representations. We think that semantic gender information (from pronouns) is encoded in the network distinctly from the stereotypical bias of gender-neutral words fig. 1. To examine that we apply orthogonal probe, which proved useful, e.g., in separating semantic and syntactic information encoded in the neural model (Limisiewicz and Mareček, 2021). Then we filter out the bias subspace from the embedding space and keep the subspace encoding factual gender information. We show that this method performs well in both desired properties: decreasing the network's reliance on bias while retaining knowledge about factual gender.

1.1 Terminology

We consider two types of gender information encoded in text:

- **Factual gender** is the grammatical (pronouns "he", "she", "her", etc.) or semantic ("boy", "girl", etc.) feature of specific word. It can be also indicated by a correference link. We will call words with factual gender as *gendered* in contrast to *gender-neutral* words.
- **Gender bias** is the connection between word and specific gender with which it is usually associated, regardless of factual premise. We will refer to words with gender bias as *biased* in contrast to *non-biased*.

Please note that those definitions do not preclude the existence of biased gender-neutral words. In that case, we consider bias stereotypical and aim to mitigate it in our method. On the other hand, we want to preserve bias in gendered words.

2 Methods

We aim to remove the influence of gender-biased words while keeping the information about factual gender in the sentence from pronouns. We focus on interactions of gender bias and factual gender information in coreference cues of the following form:

[NOUN] examined the farmer for injuries because [PRONOUN] was caring.

In English, we can expect to obtain the factual gender of the noun from the pronoun. We expect that revealing one of the words in this coreference link model should impact the prediction of the other. Therefore we can name two casual effects:

$$C_I \quad \text{Noun} \rightarrow \text{Pronoun}$$

$$C_{II} \quad \text{Pronoun} \rightarrow \text{Noun}$$

For gender-neutral nouns, the effect on predicting masked pronouns would be primarily correlated with their gender bias. While the second causality is more useful, as it reveals factual gender information and can improve the masked token prediction of a gendered word. We define two conditional probability distributions associated with those casual effects.

$$\begin{gathered} P_I(y_{Pronoun}|X, b) \\ P_{II}(y_{Noun}|X, g) \end{gathered} \tag{1}$$

Where y is a token predicted in the position of pronoun and noun, respectively; X is the context for masked language modeling. b and g are bias and factual gender factors. We model the bias factor by including a gender-neutral biased word in the noun position. Below we present examples for introducing female and male bias: [1]

Example 1:

b_f **The nurse** examined the farmer for injuries because [PRONOUN] was caring.

b_m **The doctor** examined the farmer for injuries because [PRONOUN] was caring

Similarly, factual gender factor is modeled by introducing a pronoun with a specific gender in the sentence:

Example 2:

g_f [NOUN] examined the farmer for injuries because **she** was caring.

g_m [NOUN] examined the farmer for injuries because **he** was caring.

Our aim is to diminish the role of bias in the prediction of pronouns of a specific gender. On the other hand, the gender indicated in pronouns

[1] We use [NOUN] and [PRONOUN] tokens for a better explanation, in practice, they both are masked by the same mask token, e.g. [MASK] in BERT.

can be useful in the prediction of a gendered noun, Mathematically speaking, we want to drop the conditionality on bias factor in P_I from eq. (1), while keeping the conditionality on gender factor in P_{II}.

$$\begin{aligned} P_I(y_{Pronoun}|X, b) &\equiv P_I(y_{Pronoun}|X) \\ P_{II}(y_{Noun}|X, g) &\not\equiv P_{II}(y_{Noun}|X) \end{aligned} \tag{2}$$

To decrease the effect of gender signal from the words other than pronoun and noun, we introduce a baseline example, where both pronoun and noun tokens are masked:

Example 3:

∅ [NOUN] examined the farmer for injuries because [PRONOUN] was caring.

2.1 Evaluation of Bias

Manifestation of gender bias may vary significantly from model to model and can be attributed mainly to the choice of pre-training corpora and also training regime. We define *gender preference* in a sentence by the ratio between the probability of predicting male and female pronouns:

$$GP(X) = \frac{P_I([PRONOUN_m]|X)}{P_I([PRONOUN_f]|X)} \tag{3}$$

To estimate the gender bias of a profession name, we compare the gender preference in a sentence where profession word is masked (example 1 from the previous paragraph) and not masked (example 3). We define *relative gender preference*:

$$RGP(NOUN) = \log(GP(X_{NOUN})) - \log(GP(X_{\varnothing})) \tag{4}$$

X_{NOUN} denotes contexts in which noun is not masked (example 1), and $X_{\varnothing}$ corresponds to example 3. We take the logarithm, so the results around zero would mean that revealing noun does not affect *gender preference*.[2]

2.2 Disentangling Gender Signals with Orthogonal Probe

To coerce not biased prediction eq. (2), we focus on the internal representation of the model. We aim to identify the particular subspaces in the representation of the language models that encode the casual effects C_I and C_{II}. For that purpose, we utilize *orthogonal structural probes* proposed by (Limisiewicz and Mareček, 2021).

In structural probing, the pairs of vectors are transformed, so that distance between projected embedding approximates a linguistic feature, e.g., distance in a dependency tree (Hewitt and Manning, 2019). In our case, we want to approximate the gender information introduced by a gendered pronoun f (factual) and gender-neutral noun b (bias). The f takes the values -1 for female pronouns and 1 for male ones. b is RGP for a noun.

Our orthogonal probe consists of three trainable components:

- O: *orthogonal transformation*, mapping representation to new coordinate system.
- SV: *scaling vector*, element-wise scaling dimensions in a new coordinate systems. Dimensions that store probed information are identified by finding large scaling coefficients.
- i: *intercept* shifting the representation.

The probing objective is following:

$$\begin{aligned} ||SV_I \odot (O \cdot (h_{b,P} - h_{\varnothing,P})) - i_I||_d \approx b \\ ||SV_{II} \odot (O \cdot (h_{g,N} - h_{\varnothing,N})) - i_{II}||_d \approx g \end{aligned} \tag{5}$$

Where, $h_{b,P}$ is the vector representation of masked pronoun in example 1; $h_{g,N}$ is the vector representation of masked noun in example 2; vectors $h_{\varnothing,P}$ and $h_{\varnothing,N}$ are the representations of masked pronoun and noun respectively in example 3.

To account for negative values of target factors in eq. (5), we generalize distance metric to negative values in the following way:

$$||\overrightarrow{v}||_d = ||\max(\overrightarrow{0}, \overrightarrow{v})||_2 - ||\min(\overrightarrow{0}, \overrightarrow{v})||_2 \tag{6}$$

We jointly probe for both objectives (orthogonal transformation is shared). (Limisiewicz and Mareček, 2021) observed that the resulting scaling vector after optimization tends to be sparse, and thus they allow to find the subspace of the embedding space that encodes particular information.

2.3 Filtering Algorithm

The backbone of our debiasing strategy is diminishing the role of bias factor to the predictions we

[2]The *relative gender preference* was inspired by *total effect* measure proposed by Vig et al. (2020).

need to filter it out from the representations. Particularly, we assume that, when $||h_{b,P} - h_{\varnothing,P}|| \rightarrow 0$ then $P_I(y_{Pronoun}|X, b) \rightarrow P_I(y_{Pronoun}|X)$

We can diminish the information by masking the dimensions with a corresponding scaling vector coefficient larger than small ϵ.[3] The bias filter is defined as:

$$F_{-b} = \overrightarrow{\mathbb{1}}[abs(SV_I) < \epsilon], \tag{7}$$

where $abs(\cdot)$ is element-wise absolute value and $\overrightarrow{\mathbb{1}}$ is element-wise indicator. We apply this vector to the representations of hidden layers:

$$\hat{h} = O^T \cdot (F_{-b} \odot (O \cdot h) + abs(SV_I) \odot i_I) \tag{8}$$

To preserve factual gender information, we propose an alternative version of the filter. The dimension is kept when its importance (measured by the absolute value of scaling vector coefficient) is higher in probing for factual gender than in probing for bias. We define factual gender preserving filter as:

$$F_{-b,+g} = F_{-b} + \overrightarrow{\mathbb{1}}[\epsilon \leq abs(SV_I) < abs(SV_{II})] \tag{9}$$

The filtering is performed as in eq. (8) We analyze the number of overlapping dimensions in two scaling vectors in Section 3.2.

3 Experiments and Results

We examine the representation of two BERT models (base-cased: 12 layers, 768 embedding size; and large-cased: 24 layers, 1024 embedding size, Devlin et al. (2019)), and ELECTRA (base-generator: 12 layers, 256 embedding size Clark et al. (2020)). All the models are Transformer encoders trained on the masked language modeling objective.

3.1 Evaluation of Gender Bias in Language Models

Before constructing a de-biasing algorithm, we evaluate the bias in the prediction of tree language models.

We evaluate the gender bias in language models on 104 professional words from the WinoBias dataset Zhao et al. (2018a). The authors analyzed the data from the US job market and annotated 20 professions with the highest share of woman as stereotypically female, and 20 professions with the highest share of men as stereotypically male.

We run the inference on the prompts in five formats presented in table 2 and estimate with equation eq. (4). To obtain the bias of the word in the model, we take mean $RGP(NOUN)$ computed on all prompts.

3.1.1 Results

We compare our results with the list of stereotypical words from the annotation of Zhao et al. (2018a). Similarly, we pick up to 20 nouns with the highest and positive RGP as male-biased and up to 20 nouns with the lowest and negative RGP as female-biased. These lists differ for models.

In table 2, we present the most biased words according to three models. Noticeably, there are minor differences between empirical and annotated bias. Especially word "salesperson" considered male-biased based on job market data was one of the most skewed toward female gender in 2 out of 3 models. The full results of the evaluation can be found in the appendix.

3.2 Probing for Gender Bias and Factual Gender Information

We optimize the joint probe, where orthogonal transformation is shared, while scaling vectors and intercepts are task specific. The probing objective is to approximate: $\boldsymbol{C_I}$) gender bias ($b = RGP$); and $\boldsymbol{C_{II}}$) factual gender information (f).

We use WinoMT dataset[4] Stanovsky et al. (2019) which is derivate of WinoBias dataset Zhao et al. (2018a). The in this dataset examples are harder to solve than in our evaluation prompts table 1. Each sentence contains two potential antecedents. We probe on top of each of the model's layers. We introduce another dataset for probing because we want to separate probe optimization and evaluation data. Moreover, we want to identify the encoding of gender bias and factual gender information in more diverse contexts.

We split the dataset into train, development and test set with non-overlapping nouns, mostly profession names. They contain 62, 21, and 21 unique nouns, corresponding to 2474, 856, and 546 sentences. The splits are designed to balance male and female-biased words in each of them.

The primary purpose of probing is to construct bias filters based on the values of scaling vectors

[3] We take epsilon equal to 10^{-12}. Our results weren't particularly vulnerable to this parameter, we show the analysis in the appendix.

[4] The dataset was originally introduced to evaluate gender bias in machine translation

Prompt	PRONOUN	PRONOUN 2
[PRONOUN] is [NOUN].	She He	
[PRONOUN] was [NOUN].	She He	
[PRONOUN]works as [NOUN].	She He	
[PRONOUN] job is [NOUN].	Her His	
[NOUN]said that [PRONOUN] loves [PRONOUN 2] job.	he she	her his
[NOUN] said that [PRONOUN] hates [PRONOUN 2] job.	she he	her his

Table 1: List of evaluation prompts used in the evaluation of *relative gender preference*.

Most Female Biased				Most Male Biased			
NOUN	N Models	Avg. RGP	Annotated	NOUN	N Models	Avg. RGP	Annotated
housekeeper	3/3	-2.009	female	carpenter	3/3	0.870	male
nurse	3/3	-1.840	female	farmer	3/3	0.753	male
receptionist	3/3	-1.602	female	guard	3/3	0.738	male
hairdresser	3/3	-0.471	female	sheriff	3/3	0.651	male
librarian	2/3	-0.279	female	firefighter	3/3	0.779	**neutral**
victim	2/3	-0.102	**neutral**	driver	3/3	0.622	male
child	2/3	-0.060	**neutral**	mechanic	2/3	0.719	male
salesperson	2/3	-0.056	**male**	engineer	2/3	0.645	**neutral**

Table 2: Evaluated empirical bias in analyzed Masked Language Models. Column number shows the count of models for which the word was considered biased. Annotated is the bias assigned in Zhao et al. (2018a) based on the job market data.

corresponding to F_{-b} and $F_{-b,+g}$ to perform our de-biasing transformation eq. (7) on the last layers of the model.

3.2.1 Results

The probes on the top layer give good approximation of factual gender – pearson correlation between predicted ans gold values in the range from 0.928 to 0.946 . Pearson correlation for bias was high for BERT base (0.876), BERT large (0.94.6), and lower for ELECTRA (0.451%).[5]

We have identified the dimensions encoding conditionality $\boldsymbol{C_I}$ and $\boldsymbol{C_{II}}$. In Figure 2, we present the number of dimensions selected for each objective and their overlap. We see that bias is encoded sparsely in 18 to 80 dimensions, those coordinates will be filtered out eq. (7), optionally keeping some of the overlapping dimensions, based on the eq. (9).

3.3 Filtering Gender Bias

We filter the bias dimension in the representations of the models' top layers and again evaluate the RGP for all professions. We monitor the following metrics to measure the overall improvement of the de-biasing algorithm on the set of 104 gender-neutral nouns S_{GN}:

$$MSE_{GN} = \frac{1}{|S_{GN}|} \sum_{w \in S_{GN}} RGP(w)^2 \qquad (10)$$

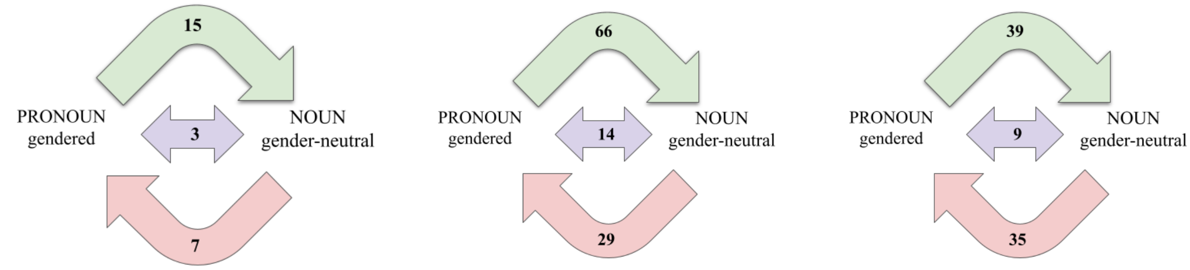
Mean squared error show how far from zero is RGP . The advantage of this metric is that the bias of some word cannot be compensated by the opposite bias of others. The main objective of debiasing is to minimize mean squared error.

$$MEAN_{GN} = \frac{1}{|S_{GN}|} \sum_{w \in S_{GN}} RGP(w) \qquad (11)$$

Mean shows whether the model is skewed toward predicting specific gender. In cases when the mean is close to zero, but MSE is high we can tell that there is no general preference of the model toward one gender, but the individual words are biased.

$$VAR_{GN} = MSE_{GN} - MEAN_{GN}^2 \qquad (12)$$

[5]For ELECTRA, we observed higher correlation of the bias probe on penultimate layer 0.668%.



(a) BERT base (out of 768 dimensions) (b) BERT large (out of 1024 dimensions) (c) ELECTRA (out of 256 dimensions)

Figure 2: Number of selected dimensions by the probe only for each of the tasks $\boldsymbol{C_I}$ (red arrow), $\boldsymbol{C_{II}}$ (green arrow), and shared for both tasks (purple arrow).

Setting	FL	MSE gendered	MSE gender-neutral	$MEAN$ gender-neutral	VAR gender-neutral
BERT B	-	**6.177**	0.504	0.352	0.124
-bias	1	2.914	0.136	**-0.056**	0.133
	2	2.213	**0.102**	-0.121	**0.088**
+f. gender	1	3.780	0.184	-0.067	0.180
	2	2.965	0.145	-0.144	0.124
ELECTRA	-	**1.360**	0.367	0.163	0.340
-bias	1	0.100	0.124	0.265	0.054
	2	0.048	**0.073**	0.200	**0.033**
+f. gender	1	0.901	0.186	**0.008**	0.185
	2	0.488	0.101	-0.090	0.093
BERT L	-	1.363	0.099	0.235	0.044
-bias	1	0.701	0.051	0.166	0.024
	2	0.267	0.015	0.069	0.011
	4	0.061	0.033	0.162	**0.007**
+f. gender	1	**1.156**	0.057	0.145	0.036
	2	0.755	0.020	**0.011**	0.020
	4	0.292	**0.010**	0.037	0.009
AIM:		$\uparrow$	$\downarrow$	≈ 0	$\downarrow$

Table 3: Aggregation of *relative gender preference* in prompts for gendered and gender-neutral nouns. FL denotes the number of the model's top layers for which filtering was performed.

Variance is a similar measure to MSE. It is useful to show the spread of RGP when the mean is non-zero.

Additionally, we introduce a set of 26 gendered nouns (S_G) for which we expect to observe non-zero RPG. We monitor MSE to diagnose whether semantic gender information is preserved in debiasing:

$$MSE_G = \frac{1}{|S_G|} \sum_{w \in S_G} RGP(w) \qquad (13)$$

3.3.1 Results

In Table 3 we observe that in all cases, gender bias measured by MSE_{GN} decreases after filtering of bias subspace. The filtering on more than one layer usually further brings this metric down. It is important to note that the original model differs in the extent to which their predictions are biased. The mean square error is the lowest for BERT large (0.099), noticeably it is lower than in other analyzed models after de-biasing (except for ELECTRA after 2-layer filtering 0.073).

The predictions of all the models are skewed toward predicting male pronoun when the noun is revealed. The values of $MEAN_{GN}$ in the range from 0.235 to 0.352 can be translated to the increase in the probability of male pronouns by 29% - 42% in comparison to the probability of female pronouns. Most of the pronouns used in the evaluation were professional names. Therefore, we think that this result is the manifestation of the stereotype that career-related words tend to be associated with men.

After filtering BERT base becomes slightly skewed toward female pronouns ($MEAN_{GN} < 0$). For two remaining models to decrease $MEAN_{GN}$, it is advisable to do not filter out factual gender signal.

Another advantage of keeping factual gender representation is the preservation of the bias in semantically gendered nouns, i.e., MSE_G.

3.4 How Bias Filtering Affect Masked Language Modeling?

We examine whether filtering affects the model's performance on the original task. For that pur-

Setting	FL	Accuracy		
		BERT L	BERT B	ELECTRA
Original	-	0.516	0.526	0.499
-bias	1	0.515	0.479	0.429
	2	0.504	0.474	0.434
	4	0.479	-	-
+f. gender	1	0.515	0.479	0.434
	2	0.510	0.480	0.433
	4	0.489	-	-

Table 4: Top 1 accuracy for all tokens in EWT UD.

Setting	FL	Accuracy		
		Overall	Male	Female
BERT L	-	0.799	0.**816**	0.781
-bias	1	0.690	0.757	0.624
	2	0.774	0.804	0.744
	4	0.747	0.770	0.724
+f. gender	1	0.754	0.782	0.726
	2	0.785	0.801	0.769
	4	**0.801**	0.807	**0.794**
-f. gender	1	0.725	0.775	0.675
	2	0.763	0.788	0.738
	4	0.545	0.633	0.458
BERT B	-	**0.732**	**0.752**	**0.712**
-bias	1	0.632	0.733	0.531
	2	0.597	0.706	0.487
+f. gender	1	0.659	0.734	0.584
	2	0.620	0.690	0.549
-f. gender	1	0.634	0.662	0.606
	2	0.604	0.641	0.567
ELECTRA	-	0.652	0.680	0.624
-bias	1	0.506	0.731	0.280
	2	0.485	0.721	0.249
+f. gender	1	**0.700**	**0.757**	**0.642**
	2	0.691	0.721	0.661
-f. gender	1	0.395	0.660	0.129
	2	0.473	0.708	0.239

Table 5: Top 1 accuracy for masked pronouns in GAP dataset.

pose, we evaluate top 1 prediction accuracy for the masked tokens in test set from English Web Treebank UD (Silveira et al., 2014) with 2077 sentences. We evaluate the capability of the model to infer the personal pronoun based on the context. We use the GAP Coreference Dataset (Webster et al., 2018) with 8908 paragraphs. In each test case, we mask a pronoun referring to a person usually mentioned by their name. In the sentences gender can be easily inferred from the name, in some cases the texts also contain un-masked gender pronouns.

3.4.1 Results: All Tokens

The results in Table 4 show that filtering out bias dimensions affect performance on masked language modeling task only slightly.

3.4.2 Results:Personal Pronouns in GAP

In GAP dataset we observe a more meaningful drop in results after debiasing. The deterioration can be alleviated by omitting factual gender dimensions in the filter. For BERT large and ELECTRA this setting can even bring improvement over the original model. Our explanation of this phenomenon is that filtering can decrease the confounding information from stereotypically biased words that affect the prediction of correct gender.

In this experiment, we also examine the filter which removes all factual-gender dimensions. The transformation significantly decreases the accuracy. However, we still obtain relatively good results, i.e., on par with resutlts in Table 4. Thus, we conjecture that the gender signal is still left in the model despite filtering.

4 Related Work

In recent years, much focus was put on evaluating and countering bias in language representations or word embeddings. Bolukbasi et al. (2016) observed the distribution of Word2Vec embeddings (Mikolov et al., 2013) encode gender bias. They tried to diminish its role by projecting the embeddings along so-called "gender dimension", that separate gendered words such as *he* and *she*. They measure the bias as cosine similarity between an embedding and the gender dimension.

$$GenderDirection \approx \overrightarrow{he} - \overrightarrow{she} \tag{14}$$

(Zhao et al., 2018b) propose a method to diminish differentiation of word representations in the gender dimension during training of the GloVe embeddings (Pennington et al., 2014). Nevertheless, the following analysis of Gonen and Goldberg (2019) argued that these approaches remove bias only partially and showed that bias is encoded in the multi-dimensional subspace of the embedding space. The issue can be resolved by projecting in multiple dimensions to further nullify the role of gender in the representations (Ravfogel et al., 2020). Dropping all the gender-related information, e.g., the distinction between feminine and masculine pronouns can be detrimental to gender-sensitive applications. Kaneko and Bollegala (2019) proposed a de-biasing algorithm that gendered information in gendered words.

In this work, we both remove bias from multiple dimensions and protect gendered words. Unlike, previously mentioned approaches we work with contextual embeddings of language models. In recent research on contextualized models, (Vig et al., 2020) investigated bias in the representation of the contextual model (GPT-2 Radford et al. (2019)). They used casual mediation analysis to identify components of the model responsible for encoding bias. Nadeem et al. (2021) proposed a method of evaluation bias (including gender) with counterfactual test examples, to some extent similar to our prompts.

Recently, Stanczak and Augenstein (2021) summarized the research on evaluation and mitigation of gender bias in the survey of 304 papers.

5 Discussion and Limitations

It is important to note that in our filtering method, we focus on filtering out stereotypical bias while keeping factual gender information in the representations. Therefore, the gender is easily recoverable from the pre-processed embeddings.

This aspect makes our method not applicable to downstream tasks that use gender-biased data. For instance, in the task of predicting a profession based on a person's biography (De-Arteaga et al., 2019), there are different proportions of men and women among holders of specific professions. A classifier trained on de-biased but not de-gendered embeddings would learn to rely on gender property in its predictions.

We think that de-biasing of the proposed type can find application in language generation. In a generation, gender agreement between antecedents needs to be kept. On the other hand, gender should not be assigned based on the presence of stereotypically biased words in the context. This issue is especially grave in machine translation when translating from English to languages that widely denote gender grammatically (Stanovsky et al., 2019).

Admittedly, in our results, we see that the proposed method based on *orthogonal probes* does not fully remove gender bias from the representations section 3.3. Even though our method typically identifies multiple dimensions encoding bias and factual gender information, there is no guarantee that all such dimensions will be filtered. Noticeably, the de-biased BERT base still underperform off-the-shelf BERT large in terms of MSE_{GN}. The reason behind this particular method was its ability to disentangle the representation of two language signals, in our case: gender bias and factual gender information.

Lastly, the probe can only recreate linear transformation, while in a non-linear system such as Transformer, the signal can be encoded non-linearly. Therefore, even when we remove the whole bias subspace, the information can be recovered in the next layer of the model (Ravfogel et al., 2020).

6 Conclusions

We propose a new insight on gender information in contextual language representations. In de-biasing, we focus on the trade-off between removing stereotypical bias while preserving the semantic and grammatical information about the gender of a word from its context. Our evaluation of gender bias showed that three analyzed masked language models (BERT large, BERT based, and ELECTRA) are biased and skewed toward predicting male gender for profession names. To mitigate this issue, we disentangle stereotypical bias from factual gender information. Our filtering method is able to remove the former and preserve the latter. As a result, we decrease the bias in predictions of language models without significant deterioration of their performance in masked language modeling task.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency*.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Tomasz Limisiewicz and David Mareček. 2021. Introducing orthogonal constraint in structural probes. 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 428–442, Online. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue*, 16(3):31–57.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Epsilon	MSE	MSE	$MEAN$	VAR
	gendered	gender-neutral		
10^{-2}	0.762	0.083	0.233	0.029
10^{-4}	0.756	0.081	0.230	0.028
10^{-6}	0.764	0.074	0.213	0.029
10^{-8}	0.738	0.078	0.225	0.027
10^{-10}	0.721	0.082	0.234	0.027
10^{-12}	0.701	0.051	0.166	0.024
10^{-14}	0.709	0.043	0.138	0.023
10^{-16}	0.770	0.023	0.013	0.022

Table 6: Tuning of filtering threshold ϵ. Results for filterin bias in the last layer of BERT large.

A Technical Details

We use batches of size 10. Optimization is conducted with Adam (Kingma and Ba, 2015) with initial learning rate 0.02 and meta parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We use learning rate decay and early-stopping mechanism with decay factor 10. The training is stopped after three consecutive epochs not resulting in the improvement of validation loss learning rate updates not resulting in a new minimum, the training is stopped. We clip each gradient's norm at $c = 1.0$. The orthogonal penalty was set to $\lambda_O = 0.1$.

We implemented the network in TensorFlow 2 (Abadi et al., 2015). The code will be available at GitHub.

A.1 Computing Infrastructure

We optimized probes on a GPU core *GeForce GTX 1080 Ti*. Training a probe on top on one layer of BERT large takes about 5 minutes.

A.2 Number of Paramters in the Probe

The number of the parameters in the probe depends on the model's embedding size emb_{size}. The *orthogonal transformation* matrix consist of emb_{size}^2; both *intercept* and *scalling vector* have emb_{size} parameters. All together, the size of the probe equals to $emb_{size}^2 + 4 \cdot emb_{size}$.

B Details about Datasets

WinoMT is distributed under MIT licences; EWT UD under Creative Commons 4.0 license; GAP under Apache 2.0 license.

C Results for Different Filtering Thresholds

In table 6 we show how choice of filtering threshold ϵ affect the results of our method for BERT large.

We decided to pick the threshold equal to 10^{-12}, as lowering it brough only minor improvement in MSE_{GN}.

D Evaluation of Bias in Language Models

We present the list of 26 gendered words and their empirical bias in table 7. Following tables tables 8 and 9 show the evaluation results for 104 gender-neutral words.

NOUN	Relative Gender Preference				NOUN	Relative Gender Preference			
	BERT base	BERT large	ELECTRA	Avg.		BERT base	BERT large	ELECTRA	Avg.
Female Gendered					Male Gendered				
councilwoman	-4.262	-2.050	-0.832	-2.381	wizard	0.972	0.314	0.237	0.508
policewoman	-4.428	-1.710	-0.928	-2.355	manservant	0.974	0.493	0.115	0.527
princess	-3.486	-1.598	-1.734	-2.273	steward	0.737	0.495	0.675	0.636
actress	-3.315	-1.094	-2.319	-2.242	spokesman	0.846	0.591	0.515	0.651
chairwoman	-4.020	-1.818	-0.629	-2.156	waiter	1.003	0.473	0.639	0.705
waitress	-2.806	-1.167	-2.475	-2.150	priest	0.988	0.442	0.928	0.786
busimesswoman	-3.202	-1.696	-1.096	-1.998	actor	1.366	0.392	0.632	0.797
queen	-2.752	-0.910	-2.246	-1.969	prince	1.401	0.776	0.418	0.865
spokeswoman	-2.543	-2.126	-1.017	-1.895	policeman	1.068	0.514	1.202	0.928
stewardess	-3.484	-2.215	0.089	-1.870	king	1.399	0.658	0.772	0.943
maid	-3.092	-0.822	-1.452	-1.788	chairman	1.140	0.677	1.069	0.962
witch	-2.068	-0.706	-1.476	-1.416	councilman	1.609	1.040	0.419	1.023
nun	-2.472	-0.974	-0.613	-1.353	businessman	1.829	0.549	0.985	1.121

Table 7: List of gendered nouns with evaluated bias in three analyzed models (RGP).

NOUN	Relative Gender Preference				Bias Class			
	BERT base	BERT large	ELECTRA	Avg.	BERT base	BERT large	ELECTRA	Annotated
housekeeper	-2.813	-0.573	-2.642	-2.009	female	female	female	female
nurse	-2.850	-0.568	-2.103	-1.840	female	female	female	female
receptionist	-1.728	-0.776	-2.302	-1.602	female	female	female	female
hairdresser	-0.400	-0.228	-0.785	-0.471	female	female	female	female
librarian	0.019	-0.088	-0.768	-0.279	neutral	female	female	female
assistant	-0.477	0.020	-0.117	-0.192	female	neutral	neutral	female
secretary	-0.564	0.024	-0.027	-0.189	female	neutral	neutral	female
victim	-0.075	0.091	-0.323	-0.102	female	neutral	female	neutral
teacher	0.129	0.175	-0.595	-0.097	neutral	neutral	female	female
therapist	0.002	0.016	-0.233	-0.072	neutral	neutral	female	neutral
child	-0.100	0.073	-0.154	-0.060	female	neutral	female	neutral
salesperson	-0.680	-0.206	0.719	-0.056	female	female	male	male
practitioner	0.150	0.361	-0.621	-0.037	neutral	neutral	female	neutral
client	-0.157	0.250	-0.165	-0.024	female	neutral	female	neutral
dietitian	0.175	0.003	-0.143	0.012	neutral	neutral	female	neutral
cook	-0.150	0.141	0.048	0.013	female	neutral	neutral	male
educator	0.278	0.144	-0.375	0.015	neutral	neutral	female	neutral
cashier	0.009	0.041	0.017	0.023	neutral	neutral	neutral	female
customer	-0.401	0.328	0.142	0.023	female	neutral	neutral	neutral
attendant	-0.157	0.226	0.010	0.027	female	neutral	neutral	female
designer	0.200	0.173	-0.232	0.047	neutral	neutral	female	female
cleaner	0.151	0.099	-0.089	0.053	neutral	neutral	neutral	female
teenager	0.343	0.088	-0.210	0.074	neutral	neutral	female	neutral
passenger	0.015	0.151	0.100	0.089	neutral	neutral	neutral	neutral
guest	0.162	0.258	-0.150	0.090	neutral	neutral	female	neutral
someone	0.026	0.275	0.082	0.128	neutral	neutral	neutral	neutral
student	0.307	0.281	-0.195	0.131	neutral	neutral	female	neutral
clerk	0.107	0.216	0.105	0.143	neutral	neutral	neutral	female
visitor	0.471	0.273	-0.280	0.155	neutral	neutral	female	neutral
counselor	0.304	0.165	0.009	0.159	neutral	neutral	neutral	female
editor	0.244	0.161	0.081	0.162	neutral	neutral	neutral	female
resident	0.528	0.300	-0.304	0.174	neutral	neutral	female	neutral
patient	0.009	0.305	0.217	0.177	neutral	neutral	neutral	neutral
homeowner	0.422	0.158	-0.002	0.192	neutral	neutral	neutral	neutral
advisee	0.175	0.252	0.168	0.199	neutral	neutral	neutral	neutral
psychologist	0.259	0.232	0.124	0.205	neutral	neutral	neutral	neutral
nutritionist	0.474	0.134	0.020	0.210	neutral	neutral	neutral	neutral
dispatcher	0.250	0.118	0.284	0.217	neutral	neutral	neutral	neutral
tailor	0.572	0.382	-0.250	0.235	neutral	male	female	female
employee	0.124	0.228	0.371	0.241	neutral	neutral	neutral	neutral
owner	0.044	0.213	0.493	0.250	neutral	neutral	neutral	neutral
advisor	0.339	0.271	0.148	0.253	neutral	neutral	neutral	neutral
witness	0.287	0.319	0.187	0.264	neutral	neutral	neutral	neutral
writer	0.497	0.237	0.060	0.265	neutral	neutral	neutral	female
undergraduate	0.575	0.148	0.075	0.266	neutral	neutral	neutral	neutral
veterinarian	0.616	0.007	0.209	0.278	neutral	neutral	neutral	neutral
pedestrian	0.446	0.226	0.170	0.281	neutral	neutral	neutral	neutral
investigator	0.518	0.228	0.120	0.289	neutral	neutral	neutral	neutral
hygienist	0.665	0.274	-0.040	0.300	neutral	neutral	neutral	neutral
buyer	0.529	0.190	0.183	0.300	neutral	neutral	neutral	neutral
supervisor	0.257	0.228	0.426	0.304	neutral	neutral	neutral	male
worker	0.151	0.267	0.511	0.310	neutral	neutral	neutral	neutral
bystander	0.786	0.117	0.072	0.325	male	neutral	neutral	neutral

Table 8: List of gender-neutral nouns with their evaluated bias RGP. Female and male bias classes are assigned for 20 lowest negative and 20 highest positive RGP values. Annotated bias from Zhao et al. (2018a). Part 1 of 2.

NOUN	Relative Gender Preference				Bias Class			
	BERT base	BERT large	ELECTRA	Avg.	BERT base	BERT large	ELECTRA	Annotated
chemist	0.579	0.311	0.107	0.332	neutral	neutral	neutral	neutral
administrator	0.428	0.236	0.350	0.338	neutral	neutral	neutral	neutral
examiner	0.445	0.281	0.296	0.341	neutral	neutral	neutral	neutral
broker	0.376	0.358	0.295	0.343	neutral	neutral	neutral	neutral
instructor	0.413	0.196	0.436	0.348	neutral	neutral	neutral	neutral
developer	0.536	0.338	0.172	0.349	neutral	neutral	neutral	male
technician	0.312	0.362	0.400	0.358	neutral	neutral	neutral	neutral
baker	0.622	0.287	0.178	0.362	neutral	neutral	neutral	female
planner	0.611	0.341	0.147	0.366	neutral	neutral	neutral	neutral
bartender	0.628	0.282	0.293	0.401	neutral	neutral	neutral	neutral
paramedic	0.787	0.094	0.333	0.405	male	neutral	neutral	neutral
protester	0.722	0.498	0.019	0.413	neutral	male	neutral	neutral
specialist	0.501	0.363	0.392	0.419	neutral	male	neutral	neutral
electrician	0.935	0.283	0.076	0.431	male	neutral	neutral	neutral
physician	0.438	0.359	0.502	0.433	neutral	neutral	neutral	male
pathologist	0.817	0.307	0.181	0.435	male	neutral	neutral	neutral
analyst	0.645	0.315	0.361	0.440	neutral	neutral	neutral	male
appraiser	0.729	0.305	0.302	0.445	neutral	neutral	neutral	neutral
onlooker	0.978	0.093	0.274	0.448	male	neutral	neutral	neutral
janitor	0.702	0.493	0.174	0.456	neutral	male	neutral	male
mover	0.717	0.407	0.253	0.459	neutral	male	neutral	male
chef	0.682	0.348	0.352	0.460	neutral	neutral	neutral	neutral
lawyer	0.696	0.271	0.421	0.462	neutral	neutral	neutral	male
paralegal	0.829	0.247	0.313	0.463	male	neutral	neutral	neutral
doctor	0.723	0.355	0.322	0.467	neutral	neutral	neutral	neutral
auditor	0.654	0.329	0.504	0.496	neutral	neutral	neutral	female
officer	0.465	0.463	0.584	0.504	neutral	male	male	neutral
surgeon	0.368	0.417	0.733	0.506	neutral	male	male	neutral
programmer	0.543	0.304	0.684	0.510	neutral	neutral	male	neutral
scientist	0.568	0.427	0.548	0.514	neutral	male	neutral	neutral
painter	0.721	0.298	0.555	0.525	neutral	neutral	male	neutral
pharmacist	0.862	0.244	0.495	0.534	male	neutral	neutral	neutral
laborer	0.996	0.557	0.058	0.537	male	male	neutral	male
machinist	0.821	0.449	0.361	0.544	male	male	neutral	neutral
architect	0.790	0.243	0.609	0.547	male	neutral	male	neutral
taxpayer	0.785	0.525	0.339	0.550	male	male	neutral	neutral
chief	0.595	0.472	0.628	0.565	neutral	male	male	male
inspector	0.631	0.344	0.726	0.567	neutral	neutral	male	neutral
plumber	1.186	0.468	0.205	0.620	male	male	neutral	neutral
construction worker	0.770	0.326	0.769	0.622	male	neutral	male	male
driver	0.847	0.415	0.603	0.622	male	male	male	male
manager	0.456	0.346	1.084	0.628	neutral	neutral	male	male
engineer	0.562	0.385	0.987	0.645	neutral	male	male	neutral
sheriff	0.850	0.396	0.708	0.651	male	male	male	male
CEO	0.701	0.353	0.989	0.681	neutral	neutral	male	male
mechanic	0.752	0.307	1.098	0.719	male	neutral	male	male
guard	0.907	0.586	0.720	0.738	male	male	male	male
accountant	0.610	0.291	1.350	0.750	neutral	neutral	male	female
farmer	1.044	0.477	0.736	0.753	male	male	male	male
firefighter	1.294	0.438	0.604	0.779	male	male	male	neutral
carpenter	0.934	0.415	1.263	0.870	male	male	male	male

Table 9: List of gender-neutral nouns with their evaluated bias RGP. Female and male bias classes are assigned for 20 lowest negative and 20 highest positive RGP values. Annotated bias from Zhao et al. (2018a). Part 2 of 2.