Unmasking Contextual Stereotypes

Bartl, Marion; Nissim, Malvina; Gatt, Albert

Published in:
Proceedings of the Second Workshop on Gender Bias in Natural Language Processing

Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias

Marion Bartl
University of Groningen
University of Malta
marion.bartl.18@um.edu.mt

Malvina Nissim
University of Groningen
m.nissim@rug.nl

Albert Gatt
University of Malta
albert.gatt@um.edu.mt

Abstract

Contextualized word embeddings have been replacing standard embeddings as the representational knowledge source of choice in NLP systems. Since a variety of biases have previously been found in standard word embeddings, it is crucial to assess biases encoded in their replacements as well. Focusing on BERT (Devlin et al., 2018), we measure gender bias by studying associations between gender-denoting target words and names of professions in English and German, comparing the findings with real-world workforce statistics. We mitigate bias by fine-tuning BERT on the GAP corpus (Webster et al., 2018), after applying Counterfactual Data Substitution (CDS) (Maudslay et al., 2019). We show that our method of measuring bias is appropriate for languages such as English, but not for languages with a rich morphology and gender-marking, such as German. Our results highlight the importance of investigating bias and mitigation techniques cross-linguistically, especially in view of the current emphasis on large-scale, multilingual language models.

1 Introduction

The biases present in the large masses of language data that are used to train Natural Language Processing (NLP) models naturally leak into NLP systems. These systematic biases can have real-life consequences when such systems are e.g. used to rank the resumes of possible candidates for a vacancy in order to aid the hiring decision (Bolukbasi et al., 2016). If, for example, a model does not associate female terms with engineering professions, because these do not often co-occur in the same context in the training corpus, then the system is likely to rank male candidates for an engineering position higher than equally qualified female candidates.

As NLP applications reach more and more users directly (Sun et al., 2019), bias in NLP and as well as resulting societal implications, have become an area of research (Hovy and Spruit, 2016; Shah et al., 2019). The ACL conference includes a workshop on ethics in NLP since 2017 (Hovy et al., 2017) and one that specifically addresses gender bias since 2019 (Costa-jussà et al., 2019).

The present work contributes to promoting fairness in NLP by exploring methods to measure and mitigate gender bias in BERT (Devlin et al., 2018), a contextualized word embedding model. Its widespread and quick adoption by the research community as the backbone for a variety of tasks calls for an assessment of possible biases encoded in it.

Research Questions We combine researching how we can measure gender bias in BERT (RQ1) and how such potential gender bias can be mitigated (RQ2), with two further perspectives: a comparison with real-world statistics and a cross-lingual approach. We investigate whether gender bias in BERT is statistically related to actual women's workforce participation (RQ3), and whether a method that we successfully apply to assess gender bias in English is portable to a language with rich morphology and gender marking such as German, since such languages have proven challenging to existing methods (Gonen et al., 2019; Zmigrod et al., 2019; Zhou et al., 2019) (RQ4).

Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 1–16
Barcelona, Spain (Online), December 13, 2020.

Contributions This work makes the following contributions: (i) We present and release the Bias Evaluation Corpus with Professions (BEC-Pro), a template-based corpus in English and German, which we created to measure gender bias with respect to different profession groups. We make the dataset and code for all experiments publicly available at `https://github.com/marionbartl/gender-bias-BERT`. (ii) Through a more diverse sentence context in our corpus than in previous research, we confirm that the method of querying BERT's underlying MLM (Masked Language Model), proposed by Kurita et al. (2019), can be used for bias detection in contextualized word embeddings. (iii) We test our bias analysis on BERT against actual U.S. workforce statistics, which helps us to observe that the BERT language model does not only encode biases that reflect real-world data, but also those that are based on stereotypes. For bias mitigation, (iv) we show the success of a technique on BERT, which was previously applied on ELMo (Peters et al., 2018; Zhao et al., 2019). Finally, (v) we attempt the cross-lingual transfer of a bias measuring method proposed for English, and show how this method is impaired by the morphological marking of gender in German.

Bias Statement The present work focuses on gender bias specifically. Gender bias is the systematic unequal treatment on the basis of gender (Moss-Racusin et al., 2012; Sun et al., 2019). While we are treating gender as binary in this study, we are aware that this does not include people who identify as non-binary, which can create representational harm (Blodgett et al., 2020). In the context of our study of the BERT language model, gender bias occurs when one gender is more closely associated with a profession than another in language use, resulting in biased language models. Against the backdrop of gender participation statistics, we can assess whether a biased representation is related to the employment situation in the real world or based on stereotypes. In the latter case, this constitutes representational harm, because actual participation in the workforce is rendered invisible (Blodgett et al., 2020). Moreover, if word representations are used in downstream systems that affect hiring decisions, gender bias, irrespective of whether it is representative of real-world data, may lead to allocational harm, because male and female candidates are not equally associated with a profession from the start (Blodgett et al., 2020).

2 Background and Previous Work

Approaches to gender bias in contextualized word embeddings borrow techniques originally developed for standard embedding models. However, they need to rely on sentence contexts since contextualized word representations are conditioned on the sentence the word occurs in.

Previous research uses either templates (May et al., 2019; Kurita et al., 2019) or sentences randomly sampled from a corpus (Zhao et al., 2019; Basta et al., 2019). May et al. (2019) adapt the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) to pooled sentence representations, resulting in the SEAT (Sentence Encoder Association Test). However, the authors express concerns about the validity of this method. Zhao et al. (2019) analyze the gender subspace following Bolukbasi et al. (2016), and also classify the vectors of occupation words that occur in the same context with male and female pronouns, using coreference resolution as an extrinsic measure of gender bias. They mitigate bias via Counterfactual Data Augmentation (CDA) and neutralization.[1] Results show that CDA was more effective. Basta et al. (2019) measure gender bias by projection onto the gender direction (Bolukbasi et al., 2016) as well as clustering and classification, following Gonen and Goldberg (2019). Results from these adapted methods show that contextualized embeddings encode biases just like standard word embeddings (Zhao et al., 2019; Basta et al., 2019).

Instead of adapting bias measuring methods from standard word embeddings, Kurita et al. (2019) exploit the Masked Language Model (MLM), native to BERT (Devlin et al., 2018). Unlike ELMo (Peters et al., 2018) or GPT-2 (Radford et al., 2019), BERT learns contextualized word representations using a masked language modelling objective (Devlin et al., 2018), making the model bi-directional. Crucially, this makes it possible to obtain the probability of a single token in a sentence. Kurita et al. (2019) use the MLM to estimate the probability of a masked, gendered target word being associated with an attribute

[1] Neutralization means that at test time, gender-swapping is applied to an input sentence, and the ELMo representation for both sentences are averaged (Zhao et al., 2019).

word in a sentence. This method was shown to capture differences in association across the categories covered by Caliskan et al. (2017) in an interpretable way.

One of the problems of current research in NLP is that most work focuses on English (Hovy and Spruit, 2016; Sun et al., 2019). Methods developed to study gender bias in English do not translate well to languages that have grammatical gender, since grammatical gender can have a veiling effect on the semantics of a word. For example, Gonen et al. (2019) find that words with the same grammatical gender were regarded as more similar, as opposed to words that have similar meanings. This can occur, for instance, because gender agreement between articles and adjectives (Corbett, 2013) renders the contexts in which nouns with the same grammatical gender occur more similar. Gonen et al. (2019) also found that due to this grammatical gender bias, the debiasing method of Bolukbasi et al. (2016) was ineffective on Italian and German word embeddings. (2019) propose to use CDA as a debiasing method for gender-marking languages, because it is a pre-processing method and as such independent from the resulting vectors. The researchers measure gender bias extrinsically by using a neural language model, following Lu et al. (2018).

Present Work In the present research, we follow Kurita et al. (2019) in measuring gender bias. We apply their method of querying the MLM for a more diverse range of sentence templates from a professional context. Additionally, we base the choice of professions on workforce statistics, in order to compare bias to the real-world situation. For mitigating gender bias, we apply Maudslay et al.'s (2019) version of CDA to fine-tuning data for BERT, because it has shown promising results for both mitigating bias in English ELMo (Zhao et al., 2019) and in embeddings of morphologically rich languages (Zmigrod et al., 2019).

3 Data

In line with previous research (Kurita et al., 2019; Zhao et al., 2019; Basta et al., 2019), we measure gender bias in BERT using sentence templates. For this purpose we create the **Bias Evaluation Corpus with Professions (BEC-Pro)**, containing English and German sentences built from templates (Section 3.2). We also use two previously existing corpora, which are described in Section 3.1.

3.1 Existing Corpora

The Equity Evaluation Corpus (EEC) was developed by Kiritchenko and Mohammad (2018) as a benchmark corpus for testing gender and racial bias in NLP systems in connection with emotions. It contains 8,640 sentences constructed using 11 sentence templates with the variables <person>, which is instantiated by a male- or female-denoting NP; and <emotion word>, whose values can be one of the basic emotions. We use this corpus for preliminary bias assessment.[2] This corpus also inspired the structure of the BEC-Pro, and we borrow from it the person words used in our templates.

The GAP corpus (Webster et al., 2018) was developed as a benchmark for measuring gender bias in coreference resolution systems. It contains 8,908 ambiguous pronoun-name pairs in 4,454 contexts sampled from Wikipedia. An example sentence can be found in Figure 1. We use this corpus to fine-tune BERT (Section 4.6).

> The historical Octavia Minor's first husband was Gaius Claudius Marcellus Minor, and she bore him three children, Marcellus, Claudia Marcella Major and [Claudia Marcella Minor]; the [Octavia] in Rome is married to a nobleman named Glabius, with whom [she] has no children.

Figure 1: GAP example sentence

[2]The templates from the EEC corpus were used in preliminary experiments to test the validity of our method. For the sake of space, and to focus on the association with professions, we do not discuss here the results on the EEC data.

3.2 BEC-Pro

In order to measure bias in BERT, we created a template-based corpus in two languages, English and German. The sentence templates contain a gender-denoting noun phrase, or <person word>, as well as a <profession>.

We obtained 2019 data on gender and race participation for a detailed list of professions from the U.S. Bureau of Labor Statistics (2020)[3]. This overview shows, among others, the percentage of female employees for professions with more than 50,000 employed across the United States. From the lowest-level subgroup profession terms, we selected three groups of 20 professions each: those with highest female participation (88.3%-98.7%), those with lowest female participation (0.7%-3.3%), and those with a roughly 50-50 distribution of male and female employees (48.5%-53.3%). Profession terms were subsequently shortened to increase the likelihood that they would form part of the BERT vocabulary and make them easier to integrate in templates. For example, the phrase 'Bookkeeping, accounting, and auditing clerks', was shortened to 'bookkeeper'.

To maximize comparability, we translated the shortened English professions into both their masculine and feminine German counterparts, using the online dictionary *dict.cc*[4]. Translations were corrected by a native speaker of German. Feminine word forms were mostly created using the highly productive suffix *-in*. We note that feminine forms can have a low frequency, which can influence the probability assigned by the language model. The full list of German professions alongside their English original and shortened counterparts can be found in Tables 6-8 in the Appendix.

Following the template-based approach in the EEC (Kiritchenko and Mohammad, 2018), we created five sentence templates that include a person word, i.e. a noun phrase that describes a person and carries explicit gender information, and a profession term. These templates are shown in Table 1. The sentences were first constructed in English and then translated to German. Person words were taken from the EEC and translated into German.[5]

	English	German
1	<person>is a <profession>.	<person>ist <profession>.
2	<person> works as a <profession>.	<person>arbeitet als <profession>.
3	<person>applied for the position of <profession>.	<person>hat sich auf die Stelle als <profession>beworben.
4	<person>, the <profession>, had a good day at work.	<person>, die/der <profession>, hatte einen guten Arbeitstag.
5	<person>wants to become a <profession>.	<person>will <profession>werden.

Table 1: Sentence patterns for English and German

For example, in English, template 4 in Table 1 could generate the sentence '[My mother], the [firefighter], had a good day at work.' The same German template would then generate the sentence *[Meine Mutter], die [Feuerwehrfrau], hatte einen guten Arbeitstag.*

For each language, this led to a combined number of 5,400 sentences (5 sentence templates $\times$ 18 person words $\times$ 20 professions $\times$ 3 profession groups).

4 Method

4.1 Technical Specifications and Models

We use the Huggingface `transformers` library (Wolf et al., 2019) for `PyTorch` with a default random seed of 42 for all experiments (Adams, 2017). The model used for bias evaluation and fine-tuning is a pre-trained $\text{BERT}_{\text{BASE}}$ model (Devlin et al., 2018) with a language modelling head on top. For reasons of simplicity, this model will be referred to as *BERT language model* from here on. For English, the tokenizer and model are loaded with the standard pre-trained uncased $\text{BERT}_{\text{BASE}}$ model. Unlike in English, where capitalization for nouns is only relevant for proper names (which we do not use), in German

[3] `https://www.bls.gov/cps/cpsaat11.htm`

[4] `https://www.dict.cc/`

[5] The phrases 'this girl/this boy' were excluded, because they denote children and are therefore less likely to appear in sentences that refer to a professional context. Even though the word 'girl' is often used to refer to grown women, this does not apply to the word 'boy' to a similar extent.

capitalization is an integral part of the orthography (Stocker, 2012). For German we use the cased model provided by DBMDZ.[6]

4.2 Masking for Bias Evaluation

The method for measuring bias used in this work is based on the prediction of masked tokens and moreover relies on masking tokens to create potentially neutral settings to be used as prior. In all our experimental settings, *targets* are person words, and *attributes* are professions.

We apply masking to a sentence in three stages, illustrated in Table 2, and add the different masked versions to the BEC-Pro. Note that only target words (not determiners) are masked. If an attribute contains more than one token, all tokens of the respective phrase are masked individually.

original	My son is a medical records technician.
T masked	My `[MASK]` is a medical records technician.
A masked	My son is a `[MASK] [MASK] [MASK].`
T+A masked	My `[MASK]` is a `[MASK] [MASK] [MASK].`

Table 2: Masking example

4.3 Pre-processing

The inputs for both measuring and mitigating gender bias largely go through the same pre-processing steps. For GAP corpus instances, which can contain several sentences, we precede these steps by splitting instances into sentences. As a first step, the fixed input sequence length is determined as the smallest power of two greater than or equal to the maximum sequence length. In a second step, the inputs are tokenized by the pre-trained `BertTokenizer` and padded to the previously determined fixed sequence length. From the padded and encoded inputs, attention masks are created. Attention mask tensors have the same size as the input tensors. For each index of the input tensor, non-pad tokens are marked with a `1` and pad tokens with a `0` in the attention mask tensor.

4.4 Measuring Association Bias

Following Kurita et al. (2019), who take inspiration from the WEAT (Caliskan et al., 2017), we measure the influence of the attribute (A), which can be a profession or emotion, on the likelihood of the target (T), which denotes a male or female person: $P(T|A)$. It is assumed that in the BERT language model, the likelihood of a token is influenced by all other tokens in the sentence. Thus, we assume that the target likelihood is different depending on whether or not an attribute is present: $P(T) \neq P(T|A)$. Moreover, we assume that the likelihoods of male- and female-denoting targets are influenced differently by the same attribute word: $P(T_{female}|A) \neq P(T_{male}|A)$. Following Kurita et al. (2019), we will go on to call the probability of a target word in connection with an attribute word the *association* of the target with the attribute.

The sentence templates from the BEC-Pro (Section 3.2), are used to measure the association of target and attribute in a sentence. For measuring the association, we need to obtain the likelihood of the masked target from the BERT language model in two different settings: with the attribute masked (prior probability) and not masked (target probability). The prior and target probabilities are obtained by applying the softmax function to the logits that were predicted by the BERT language model for the position of the target in the sentence. This produces a probability distribution over the BERT vocabulary for that position in the sentence. We then obtain the (prior) probability of the respective target word by using its vocabulary index. The steps to calculate the association are shown in Figure 2.

For interpretation, a negative association between a target and an attribute means that the probability of the target is lower than the prior probability, i.e. the probability of the target *decreased* through the combination with the attribute. A positive association value means that the probability of the target *in-*

[6] `https://github.com/dbmdz/berts`

1. Take a sentence with a target and attribute word
   "He is a kindergarten teacher."

2. Mask the target word
   "[MASK] is a kindergarten teacher."

3. Obtain the probability of target word in the sentence
   $p_T = P(he = [MASK]|sent)$

4. Mask both target and attribute word. In compounds, mask each component separately.
   "[MASK] is a [MASK] [MASK]."

5. Obtain the prior probability, i.e. the probability of the target word when the attribute is masked
   $p_{prior} = P(he = [MASK]|masked\_sent)$

6. Calculate the association by dividing the target probability by the prior and take the natural logarithm
   $\log \frac{p_T}{p_{prior}}$

Figure 2: Procedure to calculate the log probability score, after Kurita et al. (2019).

id	hypothesis	expected observation
H1	There is a strong association of female (male) person-denoting noun phrases (NPs) with statistically female (male) professions, which is reduced through fine-tuning.	Positive association scores between female (male) NPs and statistically female (male) professions, which decrease after fine-tuning.
H2	There is a weak association of female (male) NPs with statistically male (female) professions, which is strengthened through fine-tuning.	Negative association scores between female (male) NPs and statistically male (female) professions, which increase after fine-tuning.
H3	There is no difference between the associations of female and male person-denoting NPs with statistically gender-balanced professions. Associations do not change much after fine-tuning.	Both association scores of female and male NPs have approx. the same value, which is likely located around zero. After fine-tuning, the association score does not deviate much from its original value.

Table 3: Hypotheses on associations between targets (person words) and attributes (professions) in the BEC-Pro.

creased through the combination with the attribute, with respect to the prior probability. Our hypotheses are summarized in Table 3.

4.5 Bias Mitigation

It has been shown that one of the more effective strategies for removing bias in traditional word embeddings involves modifying the training data instead of trying to change the resulting vector representation (Gonen and Goldberg, 2019). One such strategy is a derivative of CDA (Lu et al., 2018), Name-based Counterfactual Data Substitution (CDS) (Maudslay et al., 2019) in which the gender of words denoting persons in a training corpus is swapped in place in order to counterbalance bias. First names are exchanged as well.

To apply CDS in the context of English BERT, we use Maudslay et al.'s (2019) code for applying CDS to the GAP corpus (Webster et al., 2018). Subsequently, these gender-swapped data are used for fine-tuning the English BERT language model. Table 3 illustrates how we expect fine-tuning to influence associations in the English BERT language model. Since GAP instances are balanced between male and female genders, we expect this balance to be preserved after CDS, which would in turn influence male and female entities in the English BERT model to the same extent during fine-tuning.

4.6 Fine-tuning

For fine-tuning, each instance in the gender-swapped GAP corpus is tokenized into sentences. Subsequently, the sentences are pre-processed and attention masks are created. For training, the inputs need to undergo a masking procedure in order to be compatible with BERT's MLM. We follow the standard procedure for masking the inputs, as outlined by Devlin et al. (2018). The masking is carried out using

the `mask_tokens` function from code by Gururangan et al. (2020).[7] The unchanged input sentences then function as labels. For training, the instances are randomly sampled and a batch size of one is used. The model is trained for three epochs using an AdamW optimizer with a learning rate of 5×10^{-5} and a linear scheduler with warm-up. The fine-tuned model is subsequently used to carry out the exact same bias evaluation as outlined in Section 4.4.

5 Results

Table 4 displays the mean association scores between targets (person words) and attributes (professions) before and after fine-tuning the English BERT language model on the GAP corpus, to which CDS was applied (*pre-association* vs. *post-association*). The difference between these two association scores is used to perform the statistical analysis using the Wilcoxon signed-rank test (W) for all three profession groups individually. The effect size r is calculated following Rosenthal (1991) and Field et al. (2012). A positive difference score means that the association has increased after fine-tuning, a negative value indicates a decrease in association after fine-tuning.

5.1 Overall results

Similar to research by Rudinger et al. (2018), Table 4 contains pro- and anti-typical settings, which correspond to hypotheses H1 and H2, formulated in Table 3.

		pre	**post**	**diff.**	**Wilcoxon test**	
jobs	**person**	*mean*	*mean*	*mean*	*W*	*r*
B	f	-0.35	0.20	0.55	359188	-0.47
	m	0.05	0.07	0.01		
F	f	0.50	0.36	-0.14	96428	-0.32
	m	-0.68	-0.14	0.55		
M	f	-0.83	0.13	0.96	395974	-0.58
	m	0.16	0.21	0.05		

Table 4: Results for English association scores before (pre) and after fine-tuning (post). For jobs, B=balanced, F=female, M=male. In each row, N=900. All W tests are significant at p =2e-16.

In the pro-typical setting (H1), male (female) person words are paired with statistically male (female) profession terms. Conversely, in the anti-typical setting (H2), male (female) person words are paired with statistically female (male) profession terms. Table 4 shows that in fact, there are positive pre-association values in both pro-typical settings and negative pre-association values in both anti-typical settings, which confirms hypotheses H1 and H2. In other words, bias in BERT corresponds to real-world workforce statistics.

For the balanced professions, we expected that association values would not change much as a result of fine-tuning (H3). This hypothesis could only be confirmed for the male person words, while the female person words show a negative pre-association (-0.35) that changes to a positive post-association (0.20). This shows that male person words hold a neutral position with respect to gender-balanced professions. For female person terms, however, the negative pre-association shows that the gender-parity in the real world data is not reflected in the English BERT language model.

In general, male person words are relatively stable in BERT. Associations for these are less strong, i.e. less affected by the presence of the profession words, and also less affected by fine-tuning. These results correspond to Kurita et al.'s (2019) finding of strong male bias in BERT. Further support for this can be found in the results for the balanced profession group, which show similar behavior to those for the male group, though with lower absolute values. This suggests that workers in non-stereotypical professions are more likely to be talked about with male person terms.

[7] https://github.com/allenai/dont-stop-pretraining/blob/master/scripts/mlm_study.py

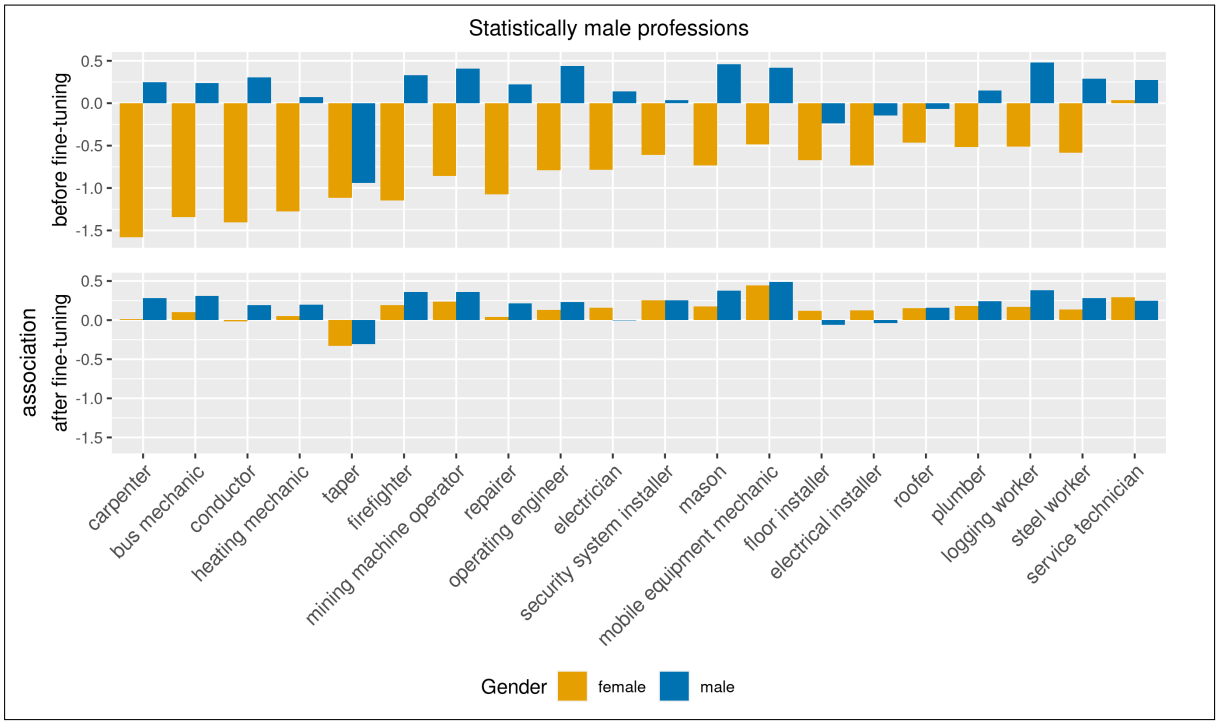

Figure 3: Pre- and post-associations of female and male person words with statistically male professions

In contrast, female person words have higher positive scores in pro-typical settings and lower negative scores in anti-typical settings, which are more susceptible to change after fine-tuning, resulting in positive scores for all professions after fine-tuning. On one hand, the more extreme association scores of female terms, as compared to male terms, illustrate them as more marked in language; on the other hand, it shows that the representations of female person words can be more easily adapted.

5.2 Profession results – English

This section zooms in on each individual profession group. The results for all profession groups are presented as two bar graphs, the upper graph showing the pre-associations and the lower showing the post-associations. The individual professions are ordered in descending order by the absolute difference in association before and after fine-tuning.

Male Professions Figure 3 shows the associations before and after fine-tuning for professions with predominantly male workers. It can be seen that there are nearly only negative associations before fine-tuning for female person words with these professions. After fine-tuning, the associations for female person words increase and almost all professions show a positive association with female person words. The male person words have small positive associations which do not change drastically after fine-tuning, in contrast to female profession terms. Generally, fine-tuning brings the association values of male and female person words closer, which indicates mitigation of gender bias. The exception to this trend is the word *taper*, whose behaviour can be attributed to the ambiguity of the term, whose more common sense is 'narrowing towards a point', rather than the profession.

Female Professions The results for the statistically female professions are summarized in Figure 4. Before fine-tuning, Figure 4 depicts very strong association values for more stereotypical professions, such as *housekeeper*, *nurse*, *receptionist* or *secretary*. For male person words, these associations are highly negative before fine-tuning and remain negative after. This could be due to the fact that the values were more extreme to begin with. Female person words show positive associations that are less extreme and have a narrower range after fine-tuning. In contrast, on the far right-hand side of Figure 4 (*paralegal*, *speech-language pathologist*, *billing clerk*, *dental hygienist*), the associations are very low for both female and make person terms, suggesting they are more gender-neutral in English BERT. Overall,

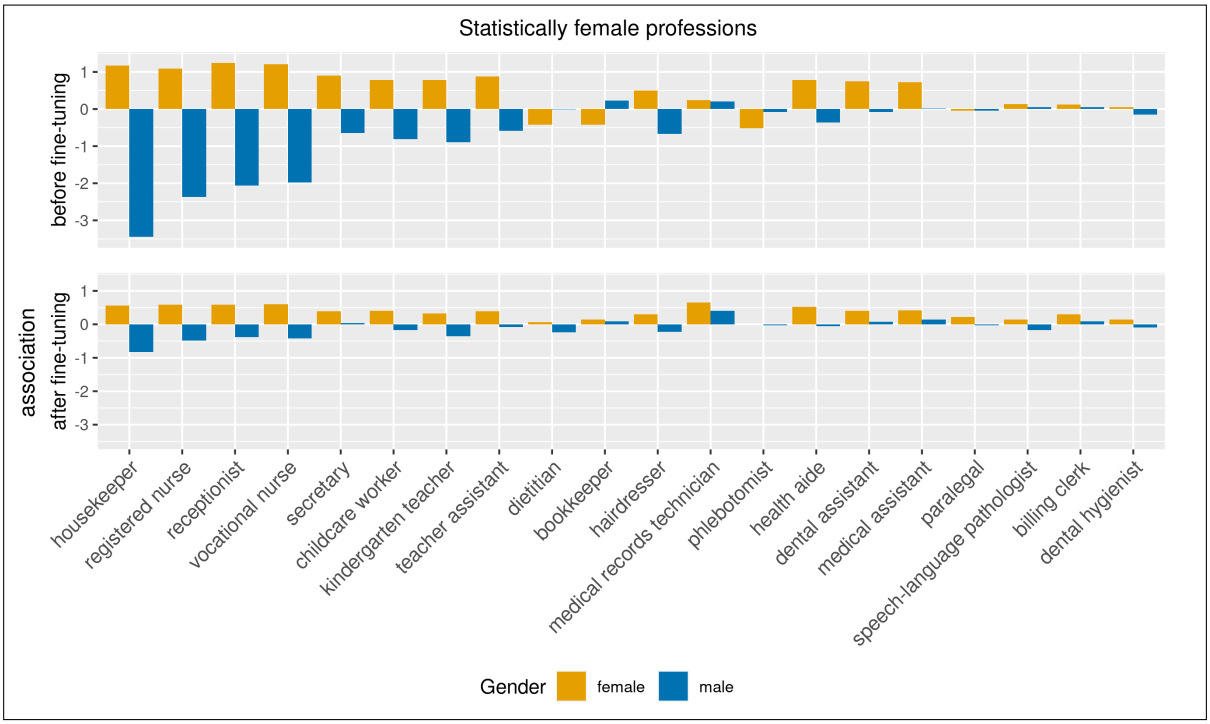

Figure 4: Pre-/post-associations of female and male person words with statistically female professions

Figure 4 shows that female bias was reduced, but the model still retained a preference for female person words in context with these professions, which corresponds to the real-world statistics.

Balanced Professions The results for the statistically balanced professions are displayed in Figure 5. They are especially interesting, because strong male or female biases do not correspond to real-world data and can be ascribed to language use in BERT's training data.

Figure 5 shows that the general trend for the associations of female person words with balanced professions before fine-tuning follows the results for statistically male professions: there are mostly negative pre-associations for female person words, which exposes bias in the English BERT language model. These associations mostly become positive after fine-tuning.

For male person words, the results show both negative and positive pre-associations. Professions with negative pre-associations for male person words are generally very specific (such as *electrical assembler* or *director of religious activities*), therefore, the negative associations may be due to low frequency of these terms. Professions with a positive pre-association for male person words are e.g. *crossing guard*, *medical scientist*, or *lifeguard*. These are more common, therefore, the positive association values reveal male-favoring bias in BERT for the professions in question. Figure 5 shows converging levels of association after fine-tuning, illustrating the method's effectiveness in mitigating gender bias.

5.3 Profession Results – German

Due to the ineffectiveness of the method for German, we only report on pre- and not on post-associations in Table 5. In order to statistically test the difference between associations for male and female person words, the Wilcoxon signed-rank test was again computed for each profession group separately.

Table 5 shows that the results across all three profession groups are highly similar: the mean associations for female person words have a value of around 2.1, and the values for male person words are around 1.4. This difference between the groups of person words is significant in all three profession groups with a medium effect size. Nevertheless, the fact that all three groups follow the same pattern indicates that the associations do not capture social gender bias. This can also be observed when looking at the pre-associations for the individual professions (We show them in Figure 6 in the Appendix).

The common pattern points to the main difference between the German and English profession terms:

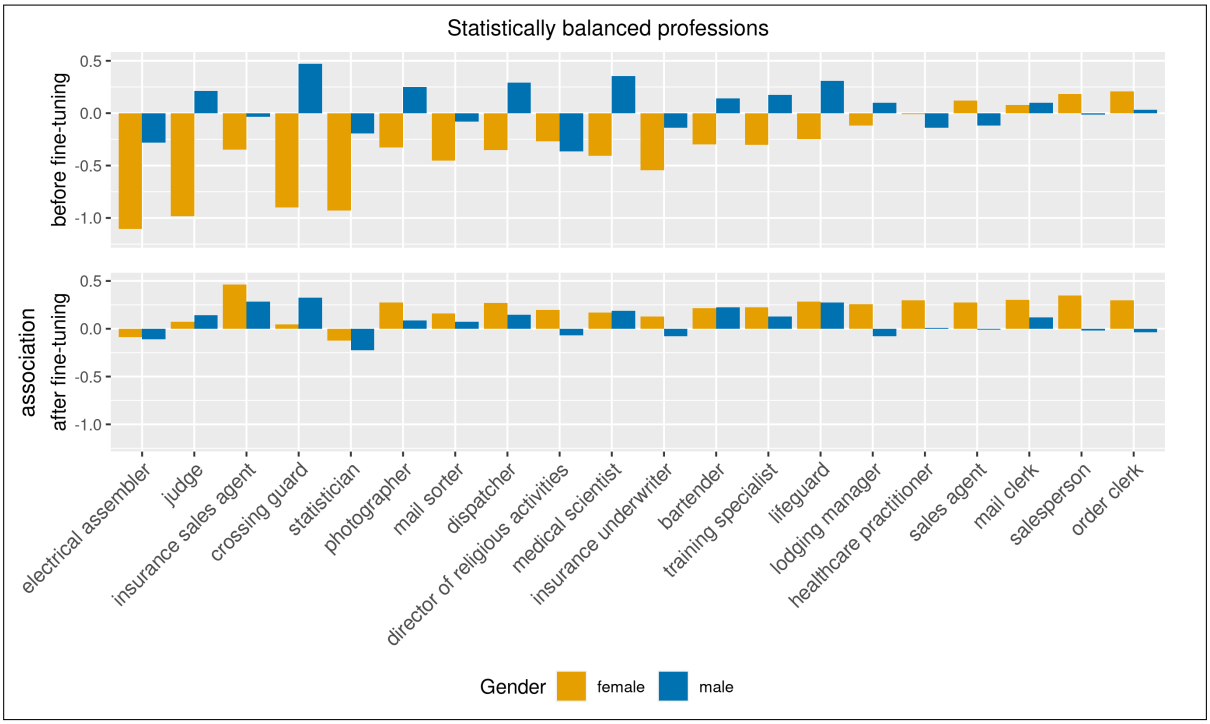

Figure 5: Pre-/post-associations of female and male person words with statistically balanced professions

		pre association		Wilcoxon		
jobs	**person**	*mean*	*sd*	*p*	*W*	*r*
B	f	2.14	2.4	2e-16	315,058	-0.34
	m	1.36	2.06			
F	f	2.05	2.45	2e-16	304,635	-0.31
	m	1.34	2.09			
M	f	2.14	2.46	2e-16	297,605	-0.29
	m	1.46	2.15			

Table 5: Results and statistical evaluation for German associations across professions and person words. For jobs, B=balanced, F=female, M=male. The number of instances for each row is 900.

German terms are divided into masculine and feminine forms (Section 3.2), because they agree with the grammatical gender of the corresponding person word. We believe that this grammatical difference generates similar association values across the three profession groups.

Specifically, the gender marker of the attribute (profession) influences the likelihood of the target (person word). The fact that the associations for female person words are consistently higher corresponds to the marking of the feminine noun form, e.g. with the suffix *-in*, which is attached to the unmarked masculine form. However, even though it is the unmarked word form, the masculine profession term also carries grammatical gender information, which we assume causes high positive associations across all profession groups.

6 Discussion and Conclusions

The goal of this work is to measure and mitigate gender bias in English and German BERT models (Devlin et al., 2018). For measuring gender bias, we use a method first proposed by Kurita et al. (2019): word probabilities taken from the BERT language model are used to calculate association bias between a gender-denoting target word and an attribute word, such as a profession. Our success in making gender bias in the English BERT model visible supports the establishment of the method as a unified metric. Moreover, we create the BEC-Pro (Bias Evaluation Corpus with Professions), a template-based corpus

set in a professional context, which includes professions from three different statistical groups as well as several male and female person words. With this corpus, which we make available to the community, we contribute to streamlining the visualization of gender bias in other contextualized word embedding models.

For mitigating gender bias, we first apply CDS (Maudslay et al., 2019) to the GAP corpus (Webster et al., 2018) and then fine-tune the English BERT language model on this corpus. We confirm Zhao et al.'s (2019) finding that CDA, or CDS in this case, is useful for mitigating gender bias in the English BERT model.

Using professions based on workforce statistics allows for a comparison of bias in the BERT language model with real-world data. We find that the English BERT language model reflects the real-world bias of the male- and female-typical profession groups through positive pro-typical associations and negative anti-typical associations before fine-tuning. After fine-tuning, we observe a reduction in association only for female person words and female-typical professions, but there is an increase in association in both anti-typical settings. This lends support to the effectiveness of our bias mitigation method.

However, we also observe that female person words have higher absolute pre-association values in both the pro- and anti-typical settings, and also show greater changes in post-association. One possible reason could be BERT's male bias, which has been previously investigated by Kurita et al. (2019). Male person terms seem to have a more stable position in BERT, which could cause their probabilities in the model to not vary much depending on the context and make them less susceptible to change through fine-tuning. Another explanation for female terms being more affected by fine-tuning could be that the GAP corpus contained somewhat more female pronouns and nouns, but especially first names, after CDS. Fine-tuning on a corpus with a slight surplus of female person words and first names could have made the likelihood of these terms more sensitive to change.

In the balanced profession group before fine-tuning, we observe that the BERT language model does not only encode biases that reflect real-world data, but also those that are based on stereotypes. Despite the fact that all of the balanced professions have an approximately even distribution of male and female employees in the U.S. (Bureau of Labor Statistics, 2020), there is a significantly lower, negative association for female person words before fine-tuning. This signifies that women's visibility in these professions is inhibited, i.e. that women are seen as less likely to carry out such a profession. In general, the associations in the balanced profession group behave similarly to those in the male-typical profession group. Thus, unless a profession is typically carried out by women, such as the professions *kindergarten teacher* or *nurse*, the default 'worker' is culturally seen as male.

Our results moreover show that a method that works well for English is not necessarily transferable to other languages. Since German is a gender-marking language, the agreement between the grammatical gender of the person word and the profession influences the associations. Still, the consistently higher associations of female person words compared to male person words illustrate the linguistic markedness of feminine word forms, as opposed to the default masculine forms, in German.

Furthermore, the fact that English and German both belong to the Germanic language family (Dryer and Haspelmath, 2013) highlights that (genetic) linguistic relatedness does not predict the cross-linguistic success of a method. Especially for a relatively new model such as BERT, developing language-specific methods to assess its limitations is crucial to prevent bias propagation to downstream applications in the language concerned. Our lack of success in transferring the method to German emphasizes the need for more typological variety in NLP research as well as language-specific solutions (Sun et al., 2019; Hovy and Spruit, 2016).

Naturally, there are a number of limitations of this work. We specifically focus on two, here. Firstly, we work with only one very specific English BERT model, namely the uncased $\text{BERT}_{\text{BASE}}$. There are many more contextualized word embedding models besides BERT, such as GPT-2 (Radford et al., 2019) or ELMo (Peters et al., 2018). Moreover, there have been various developments and enhancements of the initial BERT model, such as DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), or RoBERTa (Liu et al., 2019). Therefore, future work could focus on gender bias in a variety of models and investigate whether there are common patterns.

Secondly, the present work extensively relies on choices made by the researchers, due to the template-based method of measuring bias. On one hand, the method is dependent on curated lists of person words and profession terms, which already introduce human bias (Sun et al., 2019). We tried to partially counteract this bias by basing the choice of professions on recent labor statistics. On the other hand, the words in the templates themselves influence the target likelihood, because word representations in BERT are dependent on the entire sentence context (Devlin et al., 2018). Therefore, future research could include a broader variety of sentences, which could also be randomly sampled.

Acknowledgments

This work is based on the first author's master thesis, which was conducted under the ERASMUS Mundus Program Language and Communication Technologies (EMLCT). We would like to thank Rowan Hall Maudslay (Maudslay et al., 2019) and Ran Zmigrod (Zmigrod et al., 2019) for sharing their code. Moreover, we would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

References

Douglas Adams. 2017. *The Ultimate Hitchhiker's Guide to the Galaxy*, volume 6. Pan Macmillan.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Bureau of Labor Statistics. 2020. Labor force statistics from the current population survey, January. [Online; accessed 16-March-2020].

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

Greville G. Corbett. 2013. Number of genders. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors. 2019. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, August.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. Sage publications.

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. *arXiv preprint arXiv:1903.03862*.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? *arXiv preprint arXiv:1910.14161*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, April.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Robert Rosenthal. 1991. *Applied social research methods series, Vol. 6. Meta-analytic procedures for social research (Rev. ed.)*. Sage Publications, Inc. https://doi. org/10.4135/9781412984997.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Paul Stocker. 2012. Nouns. In *A Student Grammar of German*, page 11–29. Cambridge University Press.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. *arXiv preprint arXiv:1909.02224*.

Ran Zmigrod, Sebastian J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

Appendix

This Appendix contains the detailed list of all professions used in this paper in Tables 6, 7, and 8. The professions were chosen from a list of professions provided by the U.S. Bureau of Labor Statistics (2020) based on the percentage of women employed, shortened and subsequently translated into German masculine and feminine forms. Tables 6-8 are referred in Section 3.2, where we describe how our template-based BEC-Pro was created.

Additionally, we provide an illustration of the associations for the German individual professions of the three profession groups in Figure 6. Figure 6 is referred in Section 5.3 when discussing that the associations do not capture gender bias for the German model due to grammatical agreement.

original profession	**shortened profession**	**% women**	**German profession (masc./fem.)**
Nursing, psychiatric, and home health aides	health aide	88.3	Gesundheitsberater/Gesundheitsberaterin
Bookkeeping, accounting, and auditing clerks	bookkeeper	88.5	Buchhalter/Buchhalterin
Registered nurses	registered nurse	88.9	staatlich geprüfter Krankenpfleger/ staatlich geprüfte Krankenpflegerin
Maids and housekeeping cleaners	housekeeper	89.0	Haushälter/Haushälterin
Receptionists and information clerks	receptionist	89.3	Rezeptionist/Rezeptionist
Phlebotomists	phlebotomist	89.3	Phlebologe/Phlebologin
Billing and posting clerks	billing clerk	89.5	Fakturist/Fakturistin
Paralegals and legal assistants	paralegal	89.6	Rechtsanwaltsgehilfe/Rechtsanwaltsgehilfin
Teacher assistants	teacher assistant	89.7	Betreuungslehrer/Betreuungslehrerin
Licensed practical and licensed vocational nurses	vocational nurse	90.8	Berufskrankenpfleger/Berufskrankenpflegerin
Dietitians and nutritionists	dietitian	92.1	Ernährungsberater/Ernährungsberaterin
Hairdressers, hairstylists, and cosmetologists	hairdresser	92.3	Friseur/Friseurin
Medical assistants	medical assistant	92.7	Arzthelfer/Arzthelferin
Secretaries and administrative assistants	secretary	93.2	Sekretär/Sekretärin
Medical records and health information technicians	medical records technician	93.3	Medizintechniker/Medizintechnikerin
Childcare workers	childcare worker	93.4	Kinderbetreuer/Kinderbetreuerin
Dental assistants	dental assistant	94.9	Zahnarzthelfer/Zahnarzthelferin
Speech-language pathologists	speech-language pathologist	95.8	Logopäde/Logopädin
Dental hygienists	dental hygienist	96.0	Dentalhygieniker/Dentalhygienikerin
Preschool and kindergarten teachers	kindergarten teacher	98.7	Kindergärtner/Kindergärtnerin

Table 6: Shortening and translation of English female-typical profession terms into German masculine and feminine forms

original profession	shortened profession	% women	German profession (masc./fem.)
Drywall installers, ceiling tile installers, and tapers	taper	0.7	Trockenbaumonteur/Trockenbaumonteurin
Structural iron and steel workers	steel worker	0.9	Stahlarbeiter/Stahlarbeiterin
Miscellaneous vehicle and mobile equipment mechanics, installers, and repairers	mobile equipment mechanic	1.3	Mechaniker für mobile Geräte/ Mechanikerin für mobile Geräte
Bus and truck mechanics and diesel engine specialists	bus mechanic	1.5	Busmechaniker/Busmechanikerin
Heavy vehicle and mobile equipment service technicians and mechanics + Automotive service technicians and mechanics	service technician	1.5	Kfz-Servicetechniker/ Kfz-Servicetechnikerin
Heating, air conditioning, and refrigeration mechanics and installers	heating mechanic	1.5	Heizungsmechaniker/ Heizungsmechanikerin
Electrical power-line installers and repairers	electrical installer	1.6	Elektroinstallateur/Elektroinstallateurin
Operating engineers and other construction equipment operators	operating engineer	1.7	Betriebsingenieur/Betriebsingenieurin
Logging workers	logging worker	1.8	Holzfäller/Holzfällerin
Carpet, floor, and tile installers and finishers	floor installer	1.9	Bodenleger/Bodenlegerin
Roofers	roofer	1.9	Dachedecker/Dachdeckerin
Mining machine operators	mining machine operator	2.0	Bergbaumaschinentechniker/ Bergbaumaschinentechnikerin
Electricians	electrician	2.2	Elektriker/Elektrikerin
Automotive body and related repairers	repairer	2.2	Kfz-Mechaniker/Kfz-Mechanikerin
Railroad conductors and yardmasters	conductor	2.4	Schaffner/Schaffnerin
Pipelayers, plumbers, pipefitters, and steamfitters	plumber	2.7	Klempner/Klempnerin
Carpenters	carpenter	2.8	Zimmermann/Zimmerin
Security and fire alarm systems installers	security system installer	2.9	Installateur von Sicherheitssystemen/ Installateurin von Sicherheitssystemen
Cement masons, concrete finishers, and terrazzo workers	mason	3.0	Maurer/Maurerin
Firefighters	firefighter	3.3	Feuerwehrmann/Feuerwehrfrau

Table 7: Shortening and translation of English male-typical profession terms into German masculine and feminine forms

original profession	shortened profession	% women	German profession (masc./fem.)
Retail salespersons	salesperson	48.5	Verkäufer/Verkäuferin
Directors, religious activities and education	director of religious activities	48.6	Leiter religiöser Aktivitäten/ Leiterin religiöser Aktivitäten
Crossing guards	crossing guard	48.6	Verkehrslotse/Verkehrslotsin
Photographers	photographer	49.3	Fotograf/Fotografin
Lifeguards and other recreational, and all other protective service workers	lifeguard	49.4	Bademeister/Bademeisterin
Lodging managers	lodging manager	49.5	Herbergsverwalter/Herbergsverwalterin
Other healthcare practitioners and technical occupations	healthcare practitioner	49.5	Heilpraktiker/Heilpraktikerin
Advertising sales agents	sales agent	49.7	Vertriebsmitarbeiter/Vertriebsmitarbeiterin
Mail clerks and mail machine operators, except postal service	mail clerk	49.8	Postbeamter/Postbeamtin
Electrical, electronics, and electromechanical assemblers	electrical assembler	50.4	Elektro-Monteur/Elektro-Monteurin
Insurance sales agents	insurance sales agent	50.6	Versicherungskaufmann/Versicherungskauffrau
Insurance underwriters	insurance underwriter	51.1	Versicherungsvermittler/Versicherungsvermittlerin
Medical scientists	medical scientist	51.8	medizinischer Wissenschaftler/ medizinische Wissenschaftlerin
Statisticians	statistician	52.4	Statistiker/Statistikerin
Training and development specialists	training specialist	52.5	Ausbilder/Ausbilderin
Judges, magistrates, and other judicial workers	judge	52.5	Richter/Richterin
Bartenders	bartender	53.1	Barkeeper/Barkeeperin
Dispatchers	dispatcher	53.1	Fahrdienstleiter/Fahrdienstleiterin
Order clerks	order clerk	53.3	Auftragssachbearbeiter/Auftragssachbearbeiterin
Postal service mail sorters, processors, and processing machine operators	mail sorter	53.3	Postsortierer/Postsortiererin

Table 8: Shortening and translation of English balanced profession terms into German masculine and feminine forms

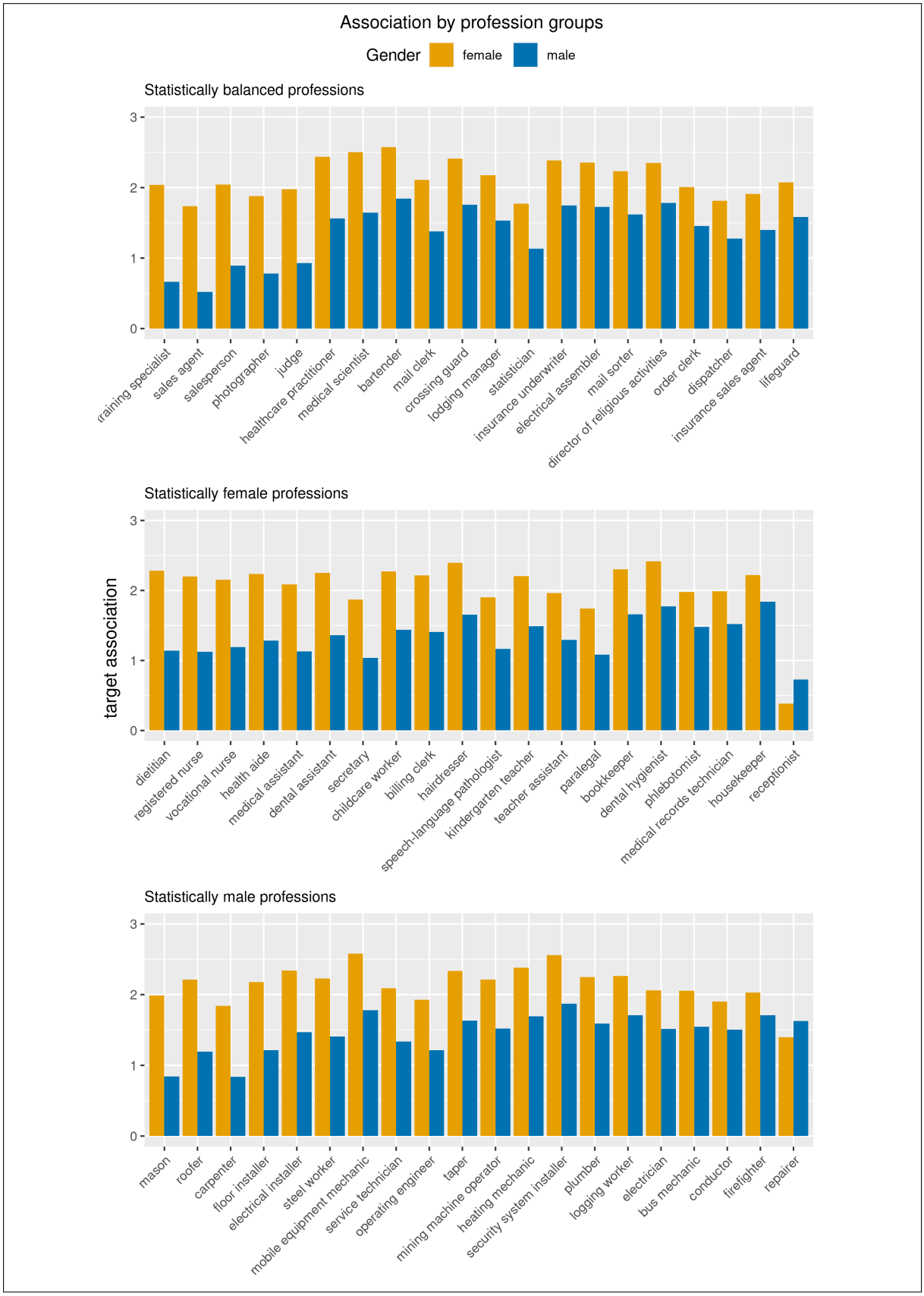

Figure 6: Mean associations for single professions of balanced, female and male profession groups for German BERT language model