The Association of Gender Bias with BERT

Measuring, Mitigating and Cross-lingual Portability

Marion Bartl

MSc. Dissertation

Department of Artificial Intelligence
Institute of Linguistics and Language Technology
Faculty of Information and Communication Technology

University of Malta

September, 2020

Supervisors:
Prof. Albert Gatt, Institute of Linguistics and Language Technology, University of Malta
Prof. Malvina Nissim, Department of Communication and Information Sciences,
University of Groningen

Submitted in partial fulfilment of the requirements for the degree of
Master of Science in Human Language Science and Technology (HLST)

FACULTY OF
INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as “the unacknowledged use, as one’s own work, of work of another person, whether or not such work has been published” (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master’s dissertation submitted is my own work, except where acknowledged and referenced.

I understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Student Name:	Marion Bartl
Course Code:	CSA5310 HLST Dissertation
Title of work:	The Association of Gender Bias with BERT: Measuring, Mitigating and Cross-lingual Portability

Signature of Student: Marion Bartl

Date: January 11, 2021

Supervisors

Prof. Malvina Nissim

Department of Communication and Information Sciences

University of Groningen

and

Prof. Albert Gatt

Institute of Linguistics and Language Technology

University of Malta

Acknowledgements

This was not how I imagined my last semester as a university student to go. I did not imagine that I would be writing my thesis in the midst of a pandemic, with the university closed, just like most other things that belonged to my daily life in Groningen. But as much as life had changed, my task remained the same: I was still required to conduct programming experiments, evaluate them, and interpret my findings in written form. All of this, apart from weekly meetings with my supervisors, required solitary work, rendered even more solitary without cafés, study rooms or libraries open, that usually provided a space for working alone together. Therefore, I have got all the more reason to thank all of my friends and my family for their support during this time. In particular, I would like to thank my friend Gaetana Ruggiero, who was in the same fix, and who I could confide in with all of my little anxieties, fears, and doubts. Beginning with our year in Malta, we have contested (university) life together and I am very lucky to have had her by my side. Moreover, I would like to thank my friend and roommate at the time Blanca Calvo Figueras, who kept me company during the lockdown, listened to my thesis trouble, and made taking a break even more enjoyable.

Last but not least, I would like to thank my supervisors, Malvina Nissim in Groningen and Albert Gatt in Malta, for their flexibility, their guidance, their calm and their overall support. Like most people, they had to deal with the effects of COVID-19 on top of their usual workload, which included home office, online teaching and, in Malvina's case, home schooling her kids as well. I am grateful, that they made time for me between all of that. Our online meetings always gave me new ideas and a fresh take on my thesis project.

With the project, I am continuing a line of research I was introduced to in 2016, through a seminar at UCLA entitled "Language and Gender". Since then, my interest in feminist linguistics has grown into research on fairness in Natural Language Processing (NLP), made possible through my admission into the LCT Erasmus Mundus Master's program. I am glad that I went down this road and I hope that I will get the opportunity to follow my research interest again in the future.

In closing, I would like to take a look back at the beginning of my university career. Here, my last thanks go out to my parents, who, without question, supported me in pursuing a non-conventional Bachelor's degree in English and Empiric Linguistics. It was definitely the right choice.

Abstract

The development of BERT (Devlin et al., 2018) and other contextualized word embeddings (Radford et al., 2019; Peters et al., 2018) brought about a significant performance increase for many NLP applications. For this reason, contextualized embeddings are replacing standard embeddings as the semantic knowledge base in NLP systems. Since a variety of biases were previously found in standard word embeddings (Caliskan et al., 2017), it is crucial to take a step back and assess biases encoded in their replacements as well. This work focuses on gender bias in BERT, aiming to measure bias, compare this bias with real-world statistics and subsequently mitigate it. Gender bias is measured through associations between gender-denoting target words and professional terms (Kurita et al., 2019). For mitigating gender bias, we first apply Counterfactual Data Substitution (CDS) (Maudslay et al., 2019) to the GAP corpus (Webster et al., 2018) and then fine-tune BERT on these data.

Since these methods for measuring and mitigating bias were originally developed for English, we also adopt a cross-lingual perspective and test whether the approach is portable to German. Unfortunately, we find that grammatical gender in German strongly influences the associations between target and attribute words, which makes it impossible to measure gender bias using the same methodology applied for English. Therefore, further experiments to mitigate gender bias in the German BERT model are discarded.

On one hand, we find that gender bias in the English BERT model is reflective of both real-world data and gender stereotypes. We mitigate this gender bias through fine-tuning on data to which CDS was applied. We hope that our positive results for English can contribute to the development of standardized methods to deal with gender bias in contextualized word embeddings. On the other hand, the fact that these methods do not work for German supports previous research calling for more language-specific work in NLP (Gonen et al., 2019; Sun et al., 2019). In light of BERT's rising popularity, finding appropriate methods to measure and mitigate bias continues to be an essential task.

Contents

1 Introduction

Gender bias in everyday language is a well-known phenomenon that has been studied and debated since the 1960s (Leavy, 2018). Solutions for eliminating biased wording have been found as well. In English, the use of 'he' as the generic form to refer to all genders can e.g. be replaced by the generic singular 'they', which leaves the gender of the referent open (Martyna, 1978). In a sentence: 'If an employee wants to get promoted, *they* need to work hard'. In German, it is not as easy to replace a generic masculine form, because German marks grammatical gender. Pronouns need to agree with the gender of their referent, which can either have masculine, feminine or neutral gender (Stocker, 2012). However, despite a more difficult grammatical situation in German, there are still initiatives to create a more inclusive language. Instead of using only the masculine gender for the entire sentence, some publications have adopted e.g. a 'gender star' (*) or 'gender gap' (_) to include not only the masculine and feminine grammatical forms, but also everyone who does not feel represented by the gender binary (Motschenbacher, 2014; AG Feministisch Sprachhandeln, 2015). The original translation for the sentence 'If [an employee] wants to get promoted, [they] need to work hard' traditionally includes the generic masculine for the subject noun, as well as the corresponding indefinite article and pronoun: *Wenn [ein Angestellter] befördert werden möchte, muss [er] hart arbeiten.* With the gender gap, the same sentence can be formulated more inclusively: *Wenn [ein_e Angestellte_r] befördert werden möchte, muss [er_sie] hart arbeiten.*

We thus see language changes that promote inclusion of all genders and counteract male-as-norm gender bias (Motschenbacher, 2014). However, this language change is ongoing and not accepted or adopted everywhere. Moreover, language does not only exist on an individual level. It is not enough to petition the local newspaper to adopt a gender-fair language policy, even though it is a step in the right direction. Nowadays, large masses of language data from various sources are used to train Natural Language Processing (NLP) systems, which then encode biases present in language (Shah et al., 2019). In these large amounts of data, biases are not only expressed through choice of word, but more importantly through word frequency or word co-occurrence frequency

(Sun et al., 2019). For example, if most instances of the word ‘nurse’ in a corpus have female referents and thus co-occur with female pronouns and female first names, then, naturally, a model will learn to strongly associate nurses with the female gender.

Now, one could argue that a preference to associate certain words with one gender over the other is not necessarily problematic, because it merely reflects the reality depicted in the training corpus. In fact, this *descriptive* approach (Shah et al., 2019) is used by studies in the social sciences to visualize, for example, meaning shifts of words over a span of several decades (Garg et al., 2018). Nevertheless, systematic biases can have real-life consequences when such a system is e.g. used to rank the resumés of possible candidates for a vacancy in order to aid the hiring decision (Bolukbasi et al., 2016). If, for example, the model used does not associate female terms with engineering professions, because these do not often co-occur in the same context within the training corpus, then the ranking system is likely to rank male candidates for an engineering position higher than equally qualified female candidates. At the point where the male candidate is hired, the dynamics at play are mutually reinforcing: there are still more male than female engineers, which means that the profession remains more closely associated with the male gender. Thus, a ranking system will keep ranking resumés of male engineers as more relevant. In cases like that of a resumé ranking system, we thus need to adopt a *normative* approach (Shah et al., 2019) and try to reduce the bias, because all candidates, no matter their gender, ethnicity, or religion, should get equal opportunities in an application process.

Unfortunately, gender biases in NLP systems often go unnoticed or are overlooked, because automated systems, unlike humans, are meant to be able to make objective decisions (Kiritchenko and Mohammad, 2018). What’s more, it is nearly impossible for the individuals impacted by these decisions to track the decision making process without access to the relevant HR tools or, beyond that, without being a computational linguist. Therefore, developers have the responsibility to be aware of, measure, and, if possible, reduce or eliminate biases in NLP systems. As Kiritchenko and Mohammad (2018) put it: “[...] we cannot absolve ourselves of the ethical implications of the systems we build.”

The present work aims to contribute to promoting fairness in NLP by studying gender bias. Here, we would like to add the general disclaimer that we are treating gender as binary in the study, but are aware that this constitutes a practical simplification of a much more nuanced situation in the real world. Specifically, this study aims at exploring methods to measure and mitigate gender bias in BERT (Devlin et al., 2018), a contextualized word embedding model. Word embeddings are vector representations of words that carry semantic information (Jurafsky and Martin, 2020, Chapter 6). Contextualized word embeddings have the additional asset that the representation of a word is conditioned on its context, i.e. the sentence the respective word occurs in. Therefore, contextualized word embedding models are replacing standard word embeddings as representational knowledge source included in the majority of NLP systems (Kurita et al., 2019). Their widespread and quick adaptation by the research community consequently calls for an assessment of the biases encoded in them.

Our two central research questions are:

RQ1. How can we measure gender bias in BERT?

RQ2. How can potential gender bias in BERT be mitigated?

Additionally to these research objectives, we also offer two further perspectives: a comparison with real-world statistics and a cross-lingual approach. Thus, we arrive at two secondary research questions:

RQ3. Does gender bias in BERT correspond to statistics of women's workforce participation?

RQ4. Is a method that was developed to assess gender bias in an English BERT model portable to a German BERT model?

For answering RQ3, we obtain professions with high and low percentages of female workers, as well as professions with roughly a 50/50 ratio in the United States.[1] The latter, balanced professions are displayed in Figure 1. Some of the professions in Figure 1, such as *judge* or *statistician*, are commonly seen as predominantly male, however, the statistics show that they are in fact balanced for gender. Therefore, including labor statistics in the

---

[1] A full overview of U.S. workforce participation statistics for the three profession groups (male, female, balanced) can be found in Figure 6 in Appendix A.2

analysis can help to assess which of the biases contained in BERT correspond to real-world data and which rely on common preconceptions or stereotypes (Caliskan et al., 2017).

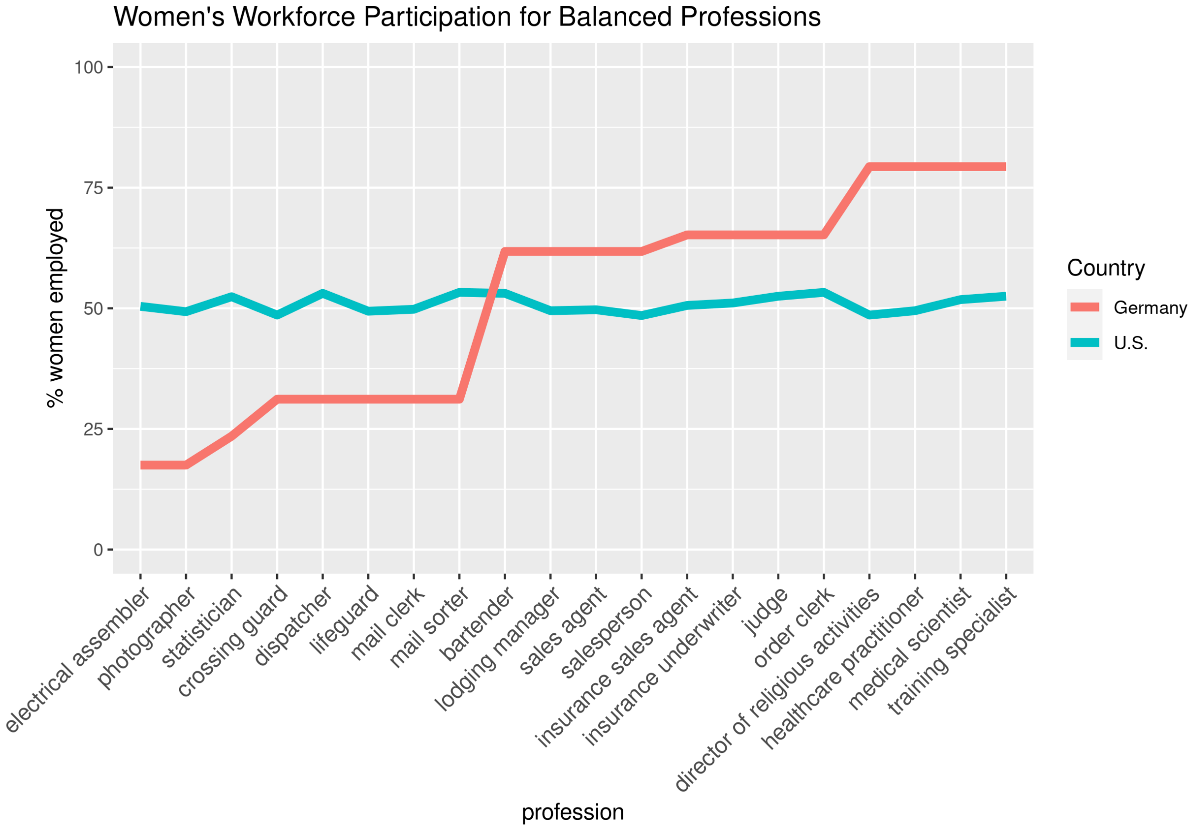

Figure 1: The displayed professions are part of a gender-balanced profession group, because all have around 50% of female employees, according to 2019 U.S. labor statistics (Bureau of Labor Statistics (BLS), 2020). To our knowledge, a similarly detailed statistic does not exist for Germany. We therefore obtained the displayed German percentages from statistics for the overarching occupational categories to which the professions belong (Statistisches Bundesamt, 2020).

With reference to RQ4, we test whether the method we use to mitigate bias in the English BERT model can also be applied to the German BERT model, thus providing an additional cross-linguistic viewpoint. This viewpoint is often missing from NLP research, because most work focuses on English (Sun et al., 2019). This focus is mainly caused by "commercial incentives" (Hovy and Spruit, 2016), because English, as a worldwide *lingua franca*, opens up the largest market for NLP applications. Therefore, existing tools for English are often transferred to other languages, with varying success (Gonen et al., 2019;

Zmigrod et al., 2019).

For example, Lu et al. (2018), Zhao et al. (2018), and Maudslay et al. (2019) use Counterfactual Data Augmentation (CDA) to reduce gender bias in English word embedding models. CDA interferes directly on the training data by 'swapping' the gender of human-denoting nouns that carry gender information. As an example, 'he' is replaced by 'she' and vice versa, 'mother' is replaced by 'father' and vice versa, and so on. In order to apply the same method to gender-marking languages such as Spanish or Hebrew, Zmigrod et al. (2019) develop an additional supporting model. This model is used to infer which words are dependents of the original target word (such as articles or adjectives), and consequently also need to be subjected to a gender change. Naturally, this re-inflection does not have a 100 percent success rate, resulting in a loss of grammaticality in the training corpus.

The example of CDA illustrates that a successful method for English cannot be assumed to seamlessly transfer to other languages, especially languages with a rich morphology (Gonen et al., 2019; Zmigrod et al., 2019). This point is especially relevant for newly developed technologies such as BERT (Devlin et al., 2018), which receive extensive attention amongst the NLP community (Kurita et al., 2019). In fact, our experiments show that our chosen method of measuring gender bias in BERT, which was developed for English by Kurita et al. (2019), does not transfer to German. Lacking a working method to measure bias, we were unable to carry out further experiments of mitigating gender bias in German BERT.

The present work is structured as follows: Section 2 first gives an overview on efforts to promote fairness in NLP, and then introduces the reader to standard and contextualized word embedding models. This includes previous approaches to (gender) bias in these models. Section 3 covers data and material that are used to measure gender bias on one hand, and mitigate gender bias on the other hand. In Section 3, we also present the Bias Evaluation Corpus with Professions (BEC-Pro), a template-based corpus in English and German, which we create to measure gender bias with respect to different profession groups. The corpus is freely available[2] and can be used for future research on BERT or

---

[2] `https://github.com/marionbartl/gender-bias-BERT/`

other representational knowledge sources.

The methodology for our experiments is presented in Section 4: in order to measure gender bias, we use the BERT language model and calculate the associations between words that carry gender information (‘he’, ‘mother’, etc.) and different professions. Our method was originally proposed by Kurita et al. (2019). As mentioned earlier, the method of measuring associations is not effective for German, which is why experiments to mitigate gender bias in BERT are only carried out for English. In these experiments, we fine-tune the English BERT model on data to which CDA was applied. This approach is based on work by Zhao et al. (2019), who apply CDA for the ELMo model (Peters et al., 2018), which precedes BERT as a contextualized word embedding model. We present the results for all experiments as well as their discussion in Section 5. Finally, Section 6 reviews findings of the present work as well as contributions to the field. We also deliberate on limitations and offer directions for future research.

2 Background

In this chapter we summarize previous research on bias, standard word embeddings and contextualized word embeddings. In Section 2.1, we begin by giving a general overview of bias and fairness in NLP as a research direction. Section 2.2 details the concept of traditional, or standard, word embeddings, the ideas behind them, and their functionality. We then zoom in on bias in standard word embeddings. Thereafter, Section 2.3 explains the development of contextualized word embeddings, their structure, as well as improvements over standard word embeddings. The chapter concludes with recent approaches to measuring and mitigating gender bias in contextualized word embeddings in Section 2.3.2. This open research problem then constitutes the focus of the present work.

2.1 Bias and Fairness in Natural Language Processing

Fairness in Natural Language Processing (NLP) has become a topic of greater interest through a position paper by Hovy and Spruit (2016). Previously, ethical considerations were thought to be superfluous, because NLP does not directly involve human subjects (Hovy and Spruit, 2016). However, as NLP applications reach more and more users directly (Sun et al., 2019), biases in these applications and in the field in general, as well as their societal implications, are becoming an area of research (Shah et al., 2019). For example, since 2016, the ACL (Association for Computational Linguistics) conference includes a workshop on ethics in NLP (Hovy et al., 2017). Since 2019, there is also a workshop at the ACL conference that specifically addresses gender bias (Costa-jussà et al., 2019).

The present work focuses on bias, more specifically gender bias. Gender bias is the systematic unequal treatment on the basis of gender (Moss-Racusin et al., 2012; Sun et al., 2019). In NLP applications, the effects of gender bias can e.g. be seen when an image captioning algorithm always predicts that a person sitting behind a computer is a man and a person standing in a kitchen is a woman (Hendricks et al., 2018). Sun et al. (2019) observe that “[t]he study of gender bias in NLP is still relatively nascent and consequently lacks unified metrics and benchmarks for evaluation.” (Sun et al., 2019)

To counteract this lack of unified metrics, Shah et al. (2019) create a conceptual framework to systematically address bias in NLP. For a more detailed description of the framework we refer the reader to the original paper. Still, we would like to touch on a few important points that are made.

Firstly, since language is the subject of NLP, models are bound to learn systematic biases that are present in language. However, even if bias is near-inevitable, it is important to be aware of the kind of biases models have in order to assess possible negative outcomes (Shah et al., 2019). These could for example constitute a performance loss due to the misclassification of underrepresented demographic groups, or the propagation of harmful stereotypes, as in the image captioning example above (Sun et al., 2019). On the other hand, biases in NLP models can also be used to showcase how language encodes stereotypes and societal attitudes, as well as their shift across time (Garg et al., 2018).

Secondly, bias needs to be seen with reference to the application of the model. For example, if the same embedding model is used as part of a part-of-speech (POS) tagger and also as part of a resumé ranking system, it is possible that this model shows bias in the latter case but not the former (Shah et al., 2019). Therefore, it is necessary to view bias in light of the application in question.

Thirdly, Shah et al. (2019) stress the need to differentiate between the effect that bias has on model performance, and the origins of bias. Most fully integrated NLP systems have a pipeline structure and bias can have its source in any of the building blocks of this pipeline. These include the training data, training labels, the algorithm itself, or additional semantic resources such as word embeddings (Shah et al., 2019; Sun et al., 2019).

The present research addresses gender bias in BERT, a contextualized word embedding model (Devlin et al., 2018). We will further expand on (contextualized) word embeddings in the following Sections 2.2 and 2.3. Shah et al. (2019) refer to bias in word embeddings as *semantic bias*. Semantic bias is a special case, because as a first element in the pipeline it can lead to other biases ‘downstream’ (Shah et al., 2019). Moreover, when using pre-trained embedding models, semantic bias can again originate in either the training data or the embedding algorithm itself. For the present work, we assume that bias measured in the BERT embedding model originates in the training data, which consist of large amounts of text. The trained BERT model thus encapsulates systematic biases contained in these language data (Kurita et al., 2019).

2.2 Traditional Word Embeddings

2.2.1 Overview

Word embeddings are vector representations of words that are derived from a large collection of text and carry semantic information (Jurafsky and Martin, 2020, Chapter 6). They were developed following the distributional hypothesis: words with the same meanings will occur in the same contexts (Joos, 1950; Harris, 1954). Thus, word embeddings are also referred to as Distributional Semantic Models (DSMs) (Baroni et al., 2014).

In their most basic form, word vectors are obtained by counting how many times a target word co-occurs with other words of the vocabulary V within a fixed-size context window of n words to the left and right of the target word (Jurafsky and Martin, 2020, Chapter 6). These word vectors have as many dimensions as words in the vocabulary, i.e. their length equals the vocabulary size $|V|$. Thus, the whole co-occurrence matrix has the size $|V| \times |V|$. These raw count DSMs can be further optimized by applying reweighting algorithms, such as *tf-idf* or *Pointwise Mutual Information* (PMI). Moreover, the dimensions can be reduced by applying compression algorithms, such as *Singular Value Decomposition* (SVD) (Baroni et al., 2014). The resulting dense vectors make computations easier and they generalize better in NLP tasks (Jurafsky and Martin, 2020, Chapter 6).

However, research found that dense vector representations of words can also be learned in an unsupervised fashion from large collections of text by using a log-linear model (Mikolov et al., 2013). The classification follows either of two objectives: predict the context words given a word (skip-gram algorithm), or predict the word given the context words (continuous bag of words (CBOW) algorithm) (Mikolov et al., 2013). The weights learned by the classifier then constitute the word embeddings. These learned word embeddings have been shown to consistently outperform earlier count-based methods while being more efficient at the same time (Baroni et al., 2014; Pennington et al., 2014). Thus, they replaced count-based methods as standard way of representing words in NLP applications (Baroni et al., 2014; Mikolov et al., 2013). In the following, mentions of word embeddings will therefore refer to the learned word vectors.

The cosine similarity of word embedding vectors can be used as a proxy to measure word similarity: the assumption is that the closer two vectors are, the more semantically similar are the words (Jurafsky and Martin, 2020, Chapter 6). Following this idea, word embeddings can also be used to perform a limited amount of human reasoning. For example, one can produce analogies of the form A *is to* B *as* C *is to* D by subtracting the vector of B from the vector of A and adding the vector of C. The nearest neighbors of the resulting vector should then represent semantically fitting candidates for word D (Mikolov et al., 2013). An example analogy of this form is the analogy between countries and their capitals, such as '*Netherlands* is to *Amsterdam* as *Germany* is to D'. Here, the vector D should be close to the vector for the word *Berlin*, because Berlin is the capital of Germany.

The initial prediction-based word embedding model by Mikolov et al. (2013) was released under the name *word2vec*. Pennington et al. (2014) develop the equally successful *GloVe* embedding model, which optimizes the word embeddings based on co-occurrence statistics of words. Bojanowski et al. (2016) further extend this line of research by developing *fasttext* word embeddings, which combine vectors of the character n-grams that make up a word and thus carry sub-word information. Using character n-grams also allows for creating vectors for out-of-vocabulary (OOV) words, which need to be represented by, for example, a vector with zeroes in the *word2vec* and *GloVe* models (Bojanowski et al., 2016).

However, word embeddings do not only carry thesaurus-like functionality such as finding similar words or analogies between words, but they constitute an important building block in modern NLP applications. Word embeddings serve as a semantic knowledge base and are used to encode inputs in most NLP systems (Jurafsky and Martin, 2020, Chapter 6). Training these word embeddings can also be done simultaneously while training a neural network on the target task. This creates word embeddings specific to the training data and task. However, especially when there is little available training data, pre-trained embeddings, i.e. embeddings that were already trained on a large amount of text, are a valuable resource (Qi et al., 2018).

Even though standard word embeddings have significantly advanced a large variety of NLP tasks, there are two main limitations: firstly, they do not capture polysemy or homonymy (Peters et al., 2018). That means that two words with the same spelling but different meanings will be represented by the same vector. As an example, Jurafsky and Martin (2020, Chapter 6) cite the polysemous word 'mouse', which can have two senses, among others: It can either mean a small animal or a device to move the cursor on a computer screen (Princeton University, 2010). A traditional word embedding model would assign the same vector to instances of both word senses, making it e.g. difficult for a search engine to retrieve the right entries for a query that includes the word 'mouse' (Jurafsky and Martin, 2020). Contextualized word representations such as BERT (Devlin et al., 2018) or ELMo (Peters et al., 2018), which are addressed in Section 2.3, provide a solution to this problem.

Secondly, learned DSMs are not human-readable. Though the dimensionality of the semantic space (i.e. the length of the vectors) is pre-defined, the individual dimensions do not have a meaning themselves. Thus, word embeddings only gain human-readable meaning through similarity to other vectors. This makes it harder to spot misrepresentations without explicitly looking for them and introducing bias through the researcher.

2.2.2 Bias in Word Embeddings

Word embeddings are trained on large amounts of text. Therefore, they incorporate human biases and stereotypes present in these texts (Bolukbasi et al., 2016; Caliskan et al., 2017; Sun et al., 2019). While some of the biases designate general human attitudes, such as 'music instruments are more pleasant than weapons', other biases are stereotyped towards social groups and can be harmful (Caliskan et al., 2017). Such biases are e.g. racial and gender biases. We note that most of the work discussed in the following section focuses on gender bias, which is partly because the present research focuses on gender bias and partly because gender bias receives more attention from the research community than racial or ethnic bias. For example, the search for 'gender bias' in the ACL Anthology[3]

---

[3] `https://www.aclweb.org/anthology/`

returns about 3,000 results while a search for 'race bias' returns close to 900 results. For recent approaches to racial, ethnic and religious biases see e.g. Kiritchenko and Mohammad (2018) and Manzini et al. (2019).

As a part of research on fairness in NLP, there are various efforts to quantify bias in word embeddings on one hand, and 'de-bias' word embeddings on the other hand (Hovy and Spruit, 2016). In the following, we first focus on intrinsic and extrinsic measuring efforts and then move on to different strategies to mitigate and eliminate bias. We present prominent approaches alongside recent research on their limitations.

Measuring Bias Mikolov et al. (2013) first present the analogy task as a way to evaluate the effectiveness of word embeddings. Bolukbasi et al. (2016) show that analogies can also capture biases, such as in 'man is to computer programmer as woman is to homemaker'. However, Nissim et al. (2020) argue for reconsideration of analogies as a means to measure bias. While analogies are easy to understand, especially for people outside the field of NLP or Artificial Intelligence (AI), they rely on subjective choices of input words, and might not return unbiased analogies due to algorithm constraints or vocabulary cutoffs, among others (Nissim et al., 2020).

A more reliable method, the Word Embedding Association Test (WEAT), is proposed by Caliskan et al. (2017). It is derived from psychology's Implicit Association Test (IAT) (Greenwald et al., 1998) and measures the association between two sets of target words (e.g. African American and European American names) and two sets of attribute words (e.g. pleasant and unpleasant words). The calculation of associations relies on the cosine similarity of two word vectors and is tested for statistical significance with a permutation test (Caliskan et al., 2017). The researchers find that this method could not only replicate human biases without social significance ('flowers are more pleasant than insects'), but also show that GloVe word embeddings (Pennington et al., 2014) encompass gender and racial biases, such as 'female terms are more closely associated with the arts while male terms are more closely associated with science' (Caliskan et al., 2017). Moreover, Caliskan et al. (2017) draw a comparison between their results and real-world data. Using 2015 U.S. labor statistics, they find that the association of an occupation word with words of the female gender is strongly correlated with the percentage of women who work in the the respective occupation.

Further methods of making bias in word embeddings visible are clustering the embeddings of implicitly biased words, and training a classifier to predict the implicit bias of a word. Gonen and Goldberg (2019) use these methods to assess the efficiency of commonly used de-biasing methods that focus on gender bias. Clustering with k-means shows that words carrying stereotypical gender associations, such as 'delicate' or 'jock', are grouped according to gender. Furthermore, they train an SVM classifier with the word embeddings of 500 male- and 500 female-stereotyped words, which reaches >85% accuracy before and after de-biasing (Gonen and Goldberg, 2019).

What is more, bias cannot only be measured intrinsically, i.e. using the word embeddings themselves. It is also important to study the performance of NLP tasks that build on word embeddings and see how bias is propagated 'downstream'. This constitutes the extrinsic evaluation. Previous research finds bias in a number of systems. For example, in Sentiment Analysis (SA) systems, the intensity score of the sentiment is found to be dependent on the race or gender mentioned (Kiritchenko and Mohammad, 2018). Stanovsky et al. (2019) observe that a machine translation system shows a better performance on stereotypical scenarios. The same is found for gendered pronoun resolution (Rudinger et al., 2018; Zhao et al., 2018). Beyond that, bias in word embeddings can also measure social phenomena, such as demographic changes and resulting shifts in word meaning (Garg et al., 2018), or assess possible implications in high-stakes settings such as the classification of profession in online biographies (De-Arteaga et al., 2019).

Mitigating Bias Approaches to mitigating gender bias in word embeddings can be divided into pre-processing and post-processing (Gonen and Goldberg, 2019), which refers to whether the de-biasing occurs before or during training of the word embeddings, or after. The most widely adapted post-processing methods were developed by Bolukbasi et al. (2016) and Zhao et al. (2018). Bolukbasi et al. (2016) determine the 'gender direction' of a vector space by using the vectors of gender-pairs (such as *she/he*, *woman/man*, etc.) and then project all non-gender-denoting nouns (gender-neutral nouns) to be orthogonal to the gender direction. Zhao et al.'s (2018) approach is to train a classifier that gathers all of the gender information in the vectors' last dimensions, which can subsequently be removed.

While these two methods are widely adopted and still receive ample attention as the most popular bias removal techniques, Gonen and Goldberg (2019) show that bias in word embeddings has more dimensions than captured by the gender direction. Clustering the de-biased word embeddings, they show that gender information can still be recovered from the embeddings of words that carry underlying gender associations (Gonen and Goldberg, 2019).

A different, pre-processing, method for mitigating gender bias is Counterfactual Data Augmentation (CDA), which was established by Lu et al. (2018). CDA intervenes on the training data: based on a list of gendered word pairs, person-denoting nouns are swapped for their opposites and the newly generated sentence is added to the corpus. This is also referred to as gender-swapping (Zhao et al., 2019). For example, the sentence 'the [man] programmed at [his] computer' becomes 'the [woman] programmed at [her] computer'. In order to prevent the generation of nonsensical sentences, Lu et al. (2018) additionally stop the swapping of pronouns if they refer to a named entity in the sentence. Maudslay et al. (2019), however, believe that omitting sentences with named entities omits potentially stereotyped settings from 'treatment'. Therefore, they propose to swap first names as well and create a list of 2,500 male-female name pairs that are matched based on their gender-specificity (*Kim* vs. *Rose*) and frequency (*Sybil* vs. *James*) (Maudslay et al., 2019). Moreover, instead of adding the swapped sentences to the corpus, they swap gendered words in place. The full method is called Name-based Counterfactual Data Substitution (Maudslay et al., 2019), which we refer to as CDS in this work for reasons of simplicity.

2.2.3 Multilingual Approaches and Gender

One of the problems of current research in NLP, as identified by Hovy and Spruit (2016) and Sun et al. (2019), is that most work focuses on English. In research on gender bias in word embeddings, methods that were developed for English do not translate well to languages that have grammatical gender. In the following, we first give an overview of gender as a grammatical category and then differentiate between grammatical gender and gender as a social construct. Subsequently, we move on to discuss word embeddings for gender-marking languages as well as strategies for bias mitigation in these word embeddings.

Grammatical and social gender Grammatical gender has the function of distinguishing different noun classes and is characterized by syntactic agreement: the dependents of a noun agree with this noun in gender (Corbett, 2013a). This includes the agreement of anaphoric pronouns with their antecedents. Grammatical gender-systems always have a "semantic 'core'" (Corbett, 2013b): the division into grammatical genders is linked to one or several semantic features of the respective nouns. Often, this feature is biological sex, but there are also systems that are "based on some notion of animacy" (Corbett, 2013b). When a system is sex-based, the grammatical categories of gender are called *feminine* and *masculine*, as opposed to the categories *female* and *male* of biological sex.

The assignment of nouns into one or another grammatical gender group is either partially or fully based on their semantic properties (Corbett, 2013c). For example, in sex-based systems, sex-differentiable entities are mostly assigned to either masculine or feminine gender. The remaining nouns, the *semantic residue*, are either assigned a gender based on further semantic features (this includes grouping all nouns that are not sex-differentiable into one category), or based on phonological or morphological properties (Corbett, 2013c). For a more general, comprehensive take on the grammatical category of gender in languages of the world see Corbett (1991).

Now that we have discussed grammatical gender, it is important to take a closer look at gender as a social category, and the interaction between the two. Previously, it was mentioned that grammatical gender systems can be sex-based (Corbett, 2013b). This introduces the differentiation between *sex*, which Corbett (2013b) relates to biology, and *gender*, which can either have a grammatical or sociological sense. In the following, we concentrate on the latter. Contrary to earlier notions of sex as 'natural' and static, social gender is viewed to be constructed in the context of social interaction. The sociologists

West and Zimmerman (1987) do not only differentiate between sex and gender, but make distinctions between *sex*, *sex category* and *gender*, which they define as follows:

> *Sex* is a determination made through the application of socially agreed upon biological criteria for classifying persons as females or males. The criteria for classification can be genitalia at birth or chromosomal typing before birth, and they do not necessarily agree with one another. Placement in a *sex category* is achieved through application of the sex criteria, but in everyday life, categorization is established and sustained by the socially required identificatory displays that proclaim one's membership in one or the other category. In this sense, one's sex category presumes one's sex and stands as proxy for it in many situations, but sex and sex category can vary independently; that is, it is possible to claim membership in a sex category even when the sex criteria are lacking. *Gender*, in contrast, is the activity of managing situated conduct in light of normative conceptions of attitudes and activities appropriate for one's sex category. Gender activities emerge from and bolster claims to membership in a sex category.

West and Zimmerman (1987) point out that all three categories *sex*, *sex category* and *gender* are human creations. Thus, they leave behind the notion of biological determinism, i.e. sex determining a person's gender and consequently their personal traits, interests and behavior. Nevertheless, they acknowledge that, similar to grammatical gender, social gender is rooted in and still stands in relation to the binary distinction into male and female sex. However, the constructivist conception of social gender also includes gender identities that are non-binary or fluctuating. It is therefore necessary to draw clear boundaries between grammatical gender, which classifies nouns into predefined categories, and social gender, which was constructed by humans to classify people and is constantly performed in a societal context (West and Zimmerman, 1987).

These two kinds of gender can overlap, but do not necessarily have to, as illustrated by the German examples in Table 1. German has three genders: masculine, feminine and neutral (Stocker, 2012). *Das Mädchen* 'the girl' is grammatically neutral (due to the diminutive marker *-chen*) but the gender of the referent is female. Here, a morphological feature takes precedence over a semantic feature for grammatical gender assignment (Corbett, 2013c). *Die Lampe* 'the lamp' is grammatically feminine, but since the word refers to an object and not an animate entity, it does not have a social gender. In the case of *die Mutter* 'the mother', grammatical gender (feminine) and social gender (female) correspond with each other.

Table 1: These German examples illustrate that grammatical gender and social gender can, but need not correspond.

German noun	**translation**	**grammatical gender**	**social gender**
die Mutter	the mother	feminine	female
das Mädchen	the girl	neutral	female
die Lampe	the lamp	feminine	-

Word Embeddings in Gender-marking Languages In the context of word embeddings, grammatical gender can have a veiling effect on the semantics of a word, which include social gender. Gonen et al. (2019) perform the word similarity task with word embeddings in Italian, which is a gender-marking language. They find that words with the same grammatical gender are regarded as more similar, as opposed to words that have similar meanings. However, this does not come as a surprise, since word embeddings are computed based on word co-occurrence. In a gender-marking language, e.g. articles and adjectives agree with the grammatical gender of the noun they belong to (Corbett, 2013a), therefore the contexts of nouns of the same grammatical gender will be very similar. Gonen et al. (2019) also find that due to this grammatical gender bias, the debiasing method of Bolukbasi et al. (2016) is ineffective on Italian and German word embeddings.

Along these lines, Zmigrod et al. (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 trained on the de-biased embeddings, thereby following Lu et al. (2018). More specifically, Zmigrod et al. (2019) measure the log likelihood of phrases that contain a declined version of the word 'engineer' and neutral vs. stereotyped adjectives (e.g. in Spanish: *La ingeniera hermoso* vs. *El ingeniero hermoso* 'The beautiful (female/male) engineer'). The researchers find that CDA reduces stereotyping in the language models of four languages within their probe experiment. What is important to mention here is that Zmigrod et al. (2019) also develop a novel method for CDA in gender-marking languages. While in English, a gender-denoting noun can easily be swapped for another, gender-marking languages have to preserve the morpho-syntactic agreement with dependent words, as described earlier. Therefore, Zmigrod et al. (2019) develop a Markov random field model to infer the change of dependents at the gender change of a noun.

2.3 BERT: Transforming the World of NLP

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) is a model that learns contextualized word representations from large collections of text (Devlin et al., 2018). In this model, the vector representation of a word is dependent on the sentence the word belongs to. This is a major improvement over standard word embeddings, because polysemy and homonymy can be modelled. Adding a linear layer on top of a pre-trained BERT model, BERT can be fine-tuned to perform specific tasks. As a result, BERT pushed several benchmarks in NLP and became widely adopted by the community as a replacement for standard word embeddings (Devlin et al., 2018; Basta et al., 2019).

In the following, we first provide an overview over BERT's architecture, training objectives and improvements over previous work. We then move on to recent literature discussing gender bias in BERT.

2.3.1 Overview

The development of BERT relies on several previous ideas. First off, BERT consists of a stack of encoders, which are taken from the encoder-decoder architecture of the Transformer model (Vaswani et al., 2017). The Transformer, more specifically its method of self-attention, proves to be better at modelling word dependencies with a greater distance between them than recurrent neural network (RNN) architectures (Alammar, 2018). RNNs have been previously used to obtain contextualized word representations, such as in the ELMo model (Peters et al., 2018), a predecessor of BERT.

Secondly, BERT is trained as a language model and uses the 'knowledge' about language it gains this way for other tasks during the additional fine-tuning phase (Alammar, 2018). This process is called *transfer learning* and was introduced as an effective method by Howard and Ruder (2018).

Thirdly, instead of a standard language modelling (LM) objective, which is e.g. used for the contextual word embedding models ELMo (Peters et al., 2018) and GPT-2 (Radford et al., 2019), BERT is trained with a masked language modelling (MLM) objective (Devlin et al., 2018). For each training instance, an algorithm randomly chooses 15% of the tokens to be predicted, of which 80% are replaced by the mask token `[MASK]`, 10% are replaced by a random token, and another 10% are left unchanged. This masking procedure allows for the self-attention layers of the transformer encoder to have access to the sentence context left and right of the word that is predicted, allowing for the transformer to be bidirectional (Devlin et al., 2018). By contrast, traditional LMs are uni-directional, because they predict the next word left-to-right based on the previously seen words (Jurafsky and Martin, 2020, Chapter 7). If a traditional LM were to include the entire sentence context in the prediction of the next word, then the model could already 'look ahead' to see the words it will predict at a future time step (Devlin et al., 2018). For the MLM this is not a concern, because the tokens of a sentence are processed independently (Jurafsky and Martin, 2020, Chapter 9) and the majority of tokens that are both predicted and serve as context for the prediction of other tokens are masked or replaced by a random token in the original input.

Lastly, BERT also employs binary Next Sentence Prediction (NSP), i.e. predicting whether or not the next sentence comes after the previous sentence (Devlin et al., 2018). For this, the next sentence is replaced by a random sentence from the training corpus 50% of the time. Devlin et al. (2018) state that NSP helps the model to capture relations between sentences, which is e.g. important for Question Answering (QA). However, Liu et al. (2019) show that, in fact, removing the task of NSP does not hurt performance. Therefore, Lan et al. (2019) replace NSP with predicting the order of two consecutive sentences in their model ALBERT, which improves over the original BERT model.

On the basis of the discussed improvements over standard word embeddings, BERT reached top performance in a variety of NLP tasks at its release (Devlin et al., 2018).

Since then, the basic model is being optimized in various ways, by e.g. training longer on more data (Liu et al., 2019), or by downsizing and thus making computation less expensive (Sanh et al., 2019). In addition, there are ongoing efforts to train BERT models for different languages than English. For German, which is the secondary focus of this research, two basic BERT models have been released: one by deepset.ai[4] and the other one by the MDZ Digital Library team (dbmdz) at the Bavarian State Library[5].

2.3.2 Bias in Contextualized Word Embeddings

Given the fact that contextualized word embeddings such as BERT replace standard word embeddings in many NLP applications, the study of bias has naturally been extended to these as well (Zhao et al., 2019). Again, the following section is mainly concerned with gender bias due to the focus of this work.

For mitigating or measuring gender bias in contextualized embeddings, it is not possible to simply use approaches for standard word embeddings, since contextualized embeddings do not have singular word representations (Zhao et al., 2019). Instead, the representation of a word is conditioned on the sentence it occurs in. Therefore, previous work utilizes sentence contexts in order to obtain vectors for words within the sentence. These sentences can either be template-based (May et al., 2019; Kurita et al., 2019) or randomly sampled from a corpus (Zhao et al., 2019; Basta et al., 2019).

Using the contextualized word embeddings from these sentences, all previously discussed techniques for measuring as well as mitigating gender bias are being applied for contextual embeddings: May et al. (2019) adapt the WEAT (Caliskan et al., 2017) for pooled sentence representations, resulting in the SEAT (Sentence Encoder Association Test). However, the authors express concerns about the validity of this method since it leads to mixed results (May et al., 2019). 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. Moreover, Zhao et al. (2019) train a coreference resolution system to measure gender bias extrinsically and sub-

---

[4] `https://deepset.ai/german-bert`

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

sequently also test two bias mitigation methods: CDA and neutralization.[6] Results show that CDA is more effective than neutralization in reducing gender bias in the coreference resolution system. 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). Regarding the question of whether contextualized embeddings are less biased, Basta et al. (2019) find e.g. less direct gender bias (as measured by closeness to the gender direction) and less tight clusters of biased words. Still, predicting the implicit gender of words is more accurate and words carrying implicit gender bias can be more easily grouped with words that explicitly express gender (Basta et al., 2019). Moreover, Zhao et al. (2019) find that while a coreference resolution system trained with ELMo embeddings performs better than one trained with GloVe embeddings, it can also be seen that bias slightly increased.

Instead of adapting bias measuring methods from standard word embeddings, Kurita et al. (2019) go another way and make use of the MLM, which is used to train BERT (Devlin et al., 2018). They obtain the likelihood of a masked target word denoting gender in a sentence context with an attribute word, such as a profession (i.e. '`[MASK]` (target) is a *nurse* (attribute).'). This is called the *association* between target and attribute. Since the representation of one word depends on all others in the same sentence, differences in association are interpretable (Kurita et al., 2019).

Kurita et al. (2019) take the target and attribute words from Caliskan et al. (2017) and embed them in simple sentence patterns. Then, they measure the *log probability bias score*, i.e. the logarithm of the difference between the association of two corresponding targets in a sentence context with an attribute. The researchers also compute the WEAT for BERT by obtaining the embedding for the attribute in the context of a simple sentence. While the WEAT for BERT does not show statistically significant biases, the novel method of querying the MLM shows that the differences in association are significant across all categories covered by Caliskan et al. (2017). This shows that the method of using the

---

[6] 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).

MLM is more sensitive to the biases encoded in BERT than the WEAT (Kurita et al., 2019). The researchers apply their method again with the pronouns *he/she* as targets, and as attributes positive and negative traits, high-paying professions, and words describing skills. This experiment shows that BERT has a strong tendency (∼80%) to associate all of the tested attributes with the male pronoun, which points to a strong male bias in BERT (Kurita et al., 2019).

3 Data and Material

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 created the **Bias Evaluation Corpus with Professions (BEC-Pro)**. This corpus is designed to measure gender bias for different groups of professions and contains English and German sentences built from templates. The process of creating the corpus will be detailed in Section 3.1. In addition, we also use two previously available corpora, the EEC and the GAP corpus, for bias evaluation in a different domain and fine-tuning BERT, respectively. These corpora are introduced and discussed in Section 3.2

3.1 Bias Evaluation Corpus with Professions

In order to measure bias in BERT, we created a bi-lingual corpus, the BEC-Pro. The corpus comprises sentence templates in English and German, which contain a gender-denoting noun phrase, or <person word>, as well as a <profession>. The professions were chosen based on U.S. employment participation statistics (Bureau of Labor Statistics (BLS), 2020), making it possible to study linguistic biases against the backdrop of the real-world situation in the workforce. This section details the process of choosing the profession terms and constructing the corpus with sentence templates, alongside descriptions of the translation process into German.

3.1.1 Job Selection

English We obtained 2019 data on gender and race participation for a detailed list of professions from the U.S. Bureau of Labor Statistics (BLS) (2020)[7]. This overview shows, among others, the percentage of female employees for professions with more than 50,000 employed across the United States. Since the professions are overall divided by occupational sectors into groups and sub-groups, we only took professions at the lowest-level sub-group for all top-level groups in order to obtain the names of single professions.

From these profession terms, we then selected three groups of 20 professions each: those with highest female participation, those with lowest female participation, and those

---

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

with a roughly 50-50 distribution of male and female employees. Since the statistic only includes two genders, male and female, we will refer to the professions with low female participation as 'male', to those with high female participation as 'female', and to the third group as 'balanced'. In the statistically male professions, the percentages of women employed range from 0.7% (drywall installers, ceiling tile installers, and tapers) to 3.3% (firefighters). The statistically female professions have female workforce participation ranging from 88.3% (nursing, psychiatric, and home health aides) to 98.7% (preschool and kindergarten teachers). In the balanced professions, the percentages of women range from 48.5% (retail salespersons) to 53.3% (postal service mail sorters, processors, and processing machine operators).

In an additional pre-processing step, we shortened the obtained profession terms. This was done, because even the most low-level profession descriptions often include two or more professions. The choice, which of the terms is used for the shortened form was made subjectively by the researcher. For example, the phrase 'Bookkeeping, accounting, and auditing clerks', was shortened to 'bookkeeper'. The shortening makes the terms more likely to be a part of the BERT vocabulary and easier to integrate into sentence templates. The final lists are displayed in Table 2, arranged in descending order by percentage of female employees. The full list of the original and shortened English profession terms, alongside the percentage of women employed, can be found in Table 11 in Appendix A.1.

German In order to preserve comparability, we decided to translate the shortened English professions into German. Since German nouns denote grammatical gender, the professions were translated into both the masculine and feminine word forms. Translation was done with the help of the online tool *DeepL Translator*[8] and corrected by the researcher, who is a native speaker of German. The corrections were aided by the English-German online dictionary *dict.cc*.[9]

In German, as in most gender-marking languages, the masculine form is *unmarked*. Feminine nouns can generally be formed by attaching the suffix *-in* to the masculine form. Thus, feminine nouns formed this way are *marked*, as opposed to the *unmarked* (default)

---

[8] `https://www.deepl.com/translator`

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

Table 2: Simplified English profession terms, ordered in descending order by the percentage of women employed, according to the 2019 U.S. Labor Statistic (Bureau of Labor Statistics (BLS), 2020).

female	**male**	**balanced**
health aide	taper	salesperson
bookkeeper	steel worker	director of religious activities
registered nurse	mobile equipment mechanic	crossing guard
housekeeper	bus mechanic	photographer
receptionist	service technician	lifeguard
phlebotomist	heating mechanic	lodging manager
billing clerk	electrical installer	healthcare practitioner
paralegal	operating engineer	sales agent
teacher assistant	logging worker	mail clerk
vocational nurse	floor installer	electrical assembler
dietitian	roofer	insurance sales agent
hairdresser	mining machine operator	insurance underwriter
medical assistant	electrician	medical scientist
secretary	repairer	statistician
medical records technician	conductor	training specialist
childcare worker	plumber	judge
dental assistant	carpenter	bartender
speech-language pathologist	security system installer	dispatcher
dental hygienist	mason	order clerk
kindergarten teacher	firefighter	mail sorter

masculine form. The strategy of adding the suffix *-in* was used to obtain the feminine forms of most of the 60 occupations in the occupations list. However, there were some exceptions:

- *Krankenpfleger/in* (nurse):

  Since the job of nurse was traditionally carried out by women, the traditional word for nurse is *Krankenschwester*, which is a compound of the words *Kranke* 'sick people' and *Schwester* 'sister' in the sense of 'nun'. The word thus carries a semantic gender marker. When trying to obtain the masculine form, the conversion of *Schwester*'sister' into *Bruder* 'brother' is ungrammatical (**Krankenbruder*). The more commonly used term nowadays is *Krankenpfleger*, in which the second element *Pfleger* means ‘carer’. *Krankenpfleger* is a masculine noun and can take the feminine suffix *-in* to create its feminine equivalent *Krankenpflegerin.*

- *Versicherungskaufmann/-frau* (insurance sales agent):
  Similar to *Krankenschwester* ‘nurse’, the translation for ‘insurance sales agent’, *Versicherungskaufmann* includes the semantic gender marker *-mann* ‘-man’. Therefore, the feminine word form cannot be derived by adding the suffix *-in.* However, in this case, the suffix *-mann* ‘-man’ can be replaced by *-frau* ‘-woman’ to create the feminine word form.

- *Zimmermann/Zimmerin* (carpenter):
  The German word for ‘carpenter’, *Zimmermann*, presents a case similar to *Versicherungskaufmann*, however, in this case the suffix *-mann* ‘-man’ cannot be substituted for *-frau* ‘woman’. The online dictionary we used advocates the use of *Zimmerin* instead.

- *Barkeeper/in* (bartender):
  Barkeeper is a loanword from English. However, since German marks grammatical gender and the ending of the word *Barkeeper* mirrors the German masculine person suffix *-er*, its gender neutrality in English is lost in German. The feminine form takes the suffix *-in.* This further illustrates the productivity of the feminine suffix *-in.*

Moreover, since suffixation with *-in* is a very productive strategy, it was easily applied to professions in the statistically male category which are naturally masculine nouns (*Heizungsmechaniker-in* ‘heating mechanic-[fem.]’, *Servicetechniker-in* ‘service technician-[fem.]’, etc.). Nevertheless, the feminine profession words created in such a way are likely to have a low frequency, which can have an influence on the probability assigned by the language model. The full list of German professions alongside their English counterparts can be found in Table 12 in Appendix A.1.

3.1.2 Creation of the BEC-Pro

The creation of the Bias Evaluation Corpus with Professions (BEC-Pro) was inspired by sentence templates from the Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad, 2018). The EEC is a template-based corpus developed to test racial and gender bias in NLP systems with sentences containing emotion words. It will be discussed in more detail in the following Section 3.2.1. For the BEC-Pro, 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. The sentences were first constructed in English and then translated to German. The full list can be found in Table 3. In the German sentence template no. 4, either the masculine article *der* or the feminine article *die* are chosen based on the gender of the person word. For example, in English, template no. 4 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.*

Table 3: Sentence patterns for English and German.

no.	English	German
1	<person word>is a <profession>.	<person word>ist <profession>.
2	<person word> works as a <profession>.	<person word>arbeitet als <profession>.
3	<person word>applied for the position of <profession>.	<person word>hat sich auf die Stelle als <profession>beworben.
4	<person word>, the <profession>, had a good day at work.	<person word>, die/der <profession>, hatte einen guten Arbeitstag.
5	<person word>wants to become a <profession>.	<person word>will <profession>werden.

The person words were directly taken from the EEC (Kiritchenko and Mohammad, 2018). However, 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. Moreover, in the context of American English, 'boy' carries negative racial connotations. Therefore, we assume that the probabilities of the two words appearing in sentences that mention professions are biased from the start and we excluded them.

Again, the list of person words was translated into German and the full list is displayed in Table 4.

Table 4: Person words in English and German.

female		**male**	
English	**German**	**English**	**German**
she	sie	he	er
this woman	diese Frau	this man	dieser Mann
my sister	meine Schwester	my brother	mein Bruder
my daughter	meine Tocher	my son	mein Sohn
my wife	meine Frau	my husband	mein Mann
my girlfriend	meine Freundin	my boyfriend	mein Freund
my mother	meine Mutter	my father	mein Vater
my aunt	meine Tante	my uncle	mein Onkel
my mom	meine Mama	my dad	mein Papa

The person terms were then used together with the professions described in Section 3.1.1 to form sentences out of the described templates. For all three profession groups (statistically female, male and balanced), each sentence template (Table 3.1.2) was combined with all person words (Table 4) and all profession words from the respective group. For each language, this led to a combined number of 1,800 sentences per profession group (18 person words × 5 sentence templates × 20 professions) and a combined corpus size of 5,400 sentences for all three profession groups. The corpus files as well as the code used to create the corpus are available at `https://github.com/marionbartl/gender-bias-BERT/`.

3.2 Previously Available Corpora

Besides creating a new bilingual corpus to evaluate gender bias in a BERT language model or in any other embedding or knowledge source, we also use external corpora in order to evaluate bias and fine-tune the BERT language model. The EEC (Kiritchenko and Mohammad, 2018) is used to carry out an exploratory analysis to test the method of measuring bias in the English BERT model by means of association. GAP (Webster et al., 2018) is used as a fine-tuning corpus for the English BERT model after CDS was applied to it.

3.2.1 Equity Evaluation Corpus

The EEC was developed by Kiritchenko and Mohammad (2018) as a benchmark corpus for testing bias in NLP systems. Initially, it was used to assess gender and racial bias in sentiment analysis (SA) systems by checking whether sentiment intensity scores allocated by the systems are dependent on the tested variables. However the researchers propose that the EEC could also be used to examine bias in other kinds of NLP applications besides SA. In this work, we use the EEC to test whether our method of measuring gender bias in BERT can be applied to the emotional domain, since our main focus lies on profession terms.

The EEC is template-based with short and grammatically simple sentences. There are a total of eleven templates, seven of which include two placeholder variables: <person>, which can take values like ‘my sister’ or ‘he’, and <emotion word>, such as ‘discouraged’ or ‘relieved’. The other four templates only include the variable <person> to serve as emotionally neutral sentences. The variable <person> can either be instantiated by female and male noun phrases to examine gender bias, or by African American and European American first names that serve as a proxy for examining racial bias. The values for <emotion word> correspond to the four basic emotions (anger, fear, joy, sadness). The emotion words can either describe an emotional state or an emotional situation, and they differ in intensity. Overall, the corpus contains 8,640 sentences, which can easily be divided into groups by choosing (levels of) variables, due to its tabular nature.

As a first use case, the EEC was employed to assess bias in 219 SA systems submitted to the *SemEval-2018 Shared Task 1: Affect in Tweets* (Mohammad et al., 2018). The researchers find that the sentiment intensity scores of sentences differ significantly across the variables gender and race in over 75% of the systems (Kiritchenko and Mohammad, 2018).

3.2.2 GAP Corpus

Webster et al. (2018) developed GAP in order to provide a benchmark corpus for coreference resolution systems, in particular as a means of evaluating gender bias in these systems. The corpus contains 8,908 ambiguous pronoun-name pairs in 4,454 contexts sampled from Wikipedia. The following Examples (1) and (2) show two instances from the GAP corpus. Following Webster et al. (2018), we annotated the ambiguous pronoun in boldface and the two possible coreferents in italics. The correct coreferent is additionally underlined. In Example (1), the first of the possible coreferent names is referenced by the pronoun, while in Example (2) it is the second.

(1) She is best-known for her role as Elizabeth Bellamy in the 1970s TV drama series Upstairs, Downstairs. She was educated at St. Maur International School, in Yokohama, Japan, the oldest international school in Asia. *Pagett* played the title role in a 1977 BBC adaptation of *Anna Karenina*, and **she** gave a memorable performance in David Nobbs's TV series ' A Bit of a Do '.

(2) 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.

The researchers decided to focus on ambiguous pronoun resolution, because previous corpora under-represented ambiguous pronoun resolution in favor of resolution by string-matching (Webster et al., 2018). Moreover, well-known coreference datasets also under-represent feminine pronouns, leading to a performance bias that favors male entities. For example, in the OntoNotes dataset (Weischedel et al., 2011), a popular general purpose coreference corpus, feminine pronouns make up under 25% of all pronouns (Zhao et al., 2018). Following the Winograd schema (Levesque et al., 2012), Zhao et al. (2018) and Rudinger et al. (2018) developed gender-balanced corpora for assessing gender bias in coreference systems, however, these sentences are artificially constructed and do not represent real-world input to coreference systems well (Webster et al., 2018). The development of GAP therefore makes three contributions: the corpus contains naturally occurring text sampled from Wikipedia, it focuses on ambiguous pronouns, and it contains a 1:1 ratio of male and female pronoun-name pairs.

4 Methods

This chapter describes and discusses the methods used to measure and mitigate gender bias contained in pre-trained English and German BERT models. We first outline the steps that were taken to pre-process the inputs for evaluating and mitigating bias in Section 4.1. Section 4.2 explains the method of measuring bias in the BERT language model using the likelihood of a target word in sentence context with an attribute (Kurita et al., 2019). Subsequently, in Section 4.3 we outline our approach to reducing the measured bias by fine-tuning English BERT on data to which Counterfactual Data Substitution (CDS) (Maudslay et al., 2019) was applied.

4.1 Preprocessing

4.1.1 Technical Specifications

All operations up to and including prediction with the BERT language model are performed in python 3.8. Statistical evaluations as well as graphical visualization of the results are carried out in R 3.6. We use the Huggingface `transformers` library for python (Wolf et al., 2019) to work with BERT, using the provided `PyTorch` (Paszke et al., 2019) environment. For reasons of reproducibility, we use a random seed for all experiments, which defaults to 42 (Adams, 2017). The code and data for all experiments can be found at `https://github.com/marionbartl/gender-bias-BERT`.

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 is referred to as *BERT language model* from here on. The model is loaded by calling the `BertForMaskedLM` class with one of the provided pre-trained BERT models. Following the `transformers` workflow, the same pre-trained model is also used to instantiate the corresponding `BertTokenizer`.

For English, the tokenizer and model are loaded with the standard pre-trained uncased $\text{BERT}_{\text{BASE}}$ model (`bert-base-uncased`). For German, we use a cased $\text{BERT}_{\text{BASE}}$ model trained by DBMDZ[10] (`bert-base-german-dbmdz-cased`). We choose to use the cased model for German, because German nouns are capitalized (Stocker, 2012), which makes

---

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

capitalization an integral part of German orthography. For example, the English sentence "The brown fox jumps over the lazy dog" translates to "Der braune Fuchs springt über den faulen Hund". It can be seen that both the translations for 'fox' *Fuchs* and 'dog' *Hund* are written with an uppercase letter. Lowercasing these would constitute a orthographic mistake.

In written English, on the other hand, capitalization mainly marks proper nouns as well as the beginning of a sentence. Since the sentences used for prediction in this research are independent from each other and do not contain proper nouns, we disregard capitalization for the task at hand and use the most basic English BERT model.

4.1.2 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. Masking means that the respective tokens were replaced by the string '`[MASK]`', which is the mask token commonly employed in BERT models (Devlin et al., 2018). This section focuses on the masking of tokens in the BEC-Pro.

As described in Section 3.1.2, the sentences in the BEC-Pro contain a person word as well as an occupation. The person word functions as the target, the occupation term constitutes the attribute of the sentence (Kurita et al., 2019). The process of calculating the association of the target with the attribute will be described in detail in Section 4.2.1.

We apply masking to a sentence in three stages and add the different masked versions to the BEC-Pro. This process adds three columns to the dataset, which contain the original sentence with (1) only targets masked (2) only attributes masked and (3) both targets and attributes masked.

In the sentence 'My son is a medical records technician.', the word 'son' functions as the target and the phrase 'medical records technician' as the attribute. When the target word is contained in a phrase, like 'son' in the phrase 'my son', only the actual target word is masked: 'my `[MASK]`'. If an attribute contains more than one token, all tokens of the respective phrase are masked individually, as in the example of 'medical records technician' in Table 5. Table 5 illustrates the full masking process.

Table 5: Masking example.

original	My son is a medical records technician.
target masked	My `[MASK]` is a medical records technician.
attribute masked	My son is a `[MASK] [MASK] [MASK]`.
target & attribute masked	My `[MASK]` is a `[MASK] [MASK] [MASK]`.

4.1.3 Processing of Inputs

The BEC-Pro is used to measure bias in BERT while a gender-swapped corpus is used to fine-tune the BERT language model and thus mitigate bias. Because the same model is used in connection with both kinds of corpora, the input sentences largely go through the same pre-processing steps. However, since instances in the GAP corpus consist of several sentences, each instance is first split into sentences using the `nltk` library (Bird et al., 2009).

As a first step, the input sequence length is determined. Since the inputs are first processed by BERT as two-dimensional tensors, the input sequences need to have the same length, which is achieved in the subsequent step through either cutting sequences that are too long, or adding `[PAD]` tokens to sequences that are too short. The latter process is called padding. However, first this fixed sequence length needs to be determined. This can e.g. be done by simply taking the length of the longest input, or maximum sequence length (max_len), as fixed sequence length (fix_len). Here, we want the fixed sequence length to be a power of two, i.e. a number of the form 2^n. More specifically, we want the smallest power of two greater or equal to the maximum sequence length:

$$fix\_len = 2^n : 2^n \geq max\_len \tag{1}$$

For this, n is determined as the ceiling value of the binary logarithm of the maximum sequence length.

$$n = \lceil \log_2 max\_len \rceil \tag{2}$$

The length of a sequence is first approximated by splitting the string containing a sentence by whitespaces. Then, the longest sequence of all inputs is used as max_len. Determining the fixed sequence length as a power of two is done in order to make computations simpler.

Naturally, setting the sequence length in such a way results in two different input lengths for the evaluation corpus and the fine-tuning corpus. This is, however, not a problem for BERT, which is built with a transformer architecture and therefore able to process variable length inputs (Vaswani et al., 2017).

In a second step, the inputs are tokenized and the tokens are converted into their corresponding indices in the BERT vocabulary. As mentioned in Section 4.1.1, we use the pre-trained `BertTokenizer` from the `transformers` library (Wolf et al., 2019). This tokenizer recognizes masked tokens in the input sentences and adds the special token `[CLS]` to the beginning and the special token `[SEP]` to the end of each sequence. Moreover, the `BertTokenizer` is based on the SentencePiece tokenizer[11], which does not only find word- but also subword-boundaries and thus includes data-driven morphological tokenization. The encoded inputs are then padded to the previously determined maximum sequence length.

From the padded and encoded inputs, attention masks are created, which help the model distinguish meaningful tokens from padding tokens[12]. 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.

4.2 Measuring Association Bias

There is no standard way of measuring bias in word embeddings. The most widely adopted methods use the gender direction (Bolukbasi et al., 2016), the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), or clustering (Gonen and Goldberg, 2019). Another way to measure bias in standard word embeddings is through language models (Lu et al., 2018; Zmigrod et al., 2019) trained on these embeddings. For example, bias can be made explicit by measuring and comparing the probabilities of sentences that only differ in a word related to the dimension of bias to be investigated (Zmigrod et al., 2019).

Since BERT is built by training a masked language model (Devlin et al., 2018), it stands to reason that this underlying language model can be used to measure bias in

---

[11] `https://github.com/google/sentencepiece`

[12] `https://huggingface.co/transformers/glossary.html#attention-mask`

BERT. The following sections illustrate how the language model was used to measure associations between target and attribute words based on the approaches of Caliskan et al. (2017) and Kurita et al. (2019).

4.2.1 Target Association

Following Kurita et al. (2019), who take inspiration from the Word Embedding Association Test (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:

$$P(T|A) \tag{3}$$

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 whether or not an attribute is present:

$$P(T) \neq P(T|A) \tag{4}$$

Moreover, we assume that the likelihoods of male and female-denoting targets are influenced differently by the same attribute word:

$$P(T_{female}|A) \neq P(T_{male}|A) \tag{5}$$

Following Kurita et al. (2019), we 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, which are described in Section 3.1.2, as well as templates from the EEC, 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.

In the following, the steps needed to calculate the association are described and illustrated with an example sentence. We would like to note here that Kurita et al. (2019) call the association the *increased log probability score.*

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 the target word in the sentence
   $p_T = P(he = [MASK]|sentence)$

4. Mask both the target and the attribute word. If the attribute word is a compound, 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
   $p_{prior} = P(he = [MASK]|masked\_sentence)$

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

At the interpretation stage, 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 decreases through the combination with the attribute. By contrast, a positive association value means that the probability of the target increases through the combination with the attribute, with respect to the prior probability.

With reference to the associations between person words and profession groups in our BEC-Pro, we formulate the following hypotheses:

H1: There is a strong association of female (male) person-denoting noun phrases (NPs) with statistically female (male) professions.

In this case, we expect to observe positive association scores between female (male) NPs and statistically female (male) professions.

H2: There is a weak association of female (male) NPs with statistically male (female) professions.
In this case, we expect to observe negative association scores between female (male) NPs and statistically male (female) professions.

H3: There is no difference between the associations of female and male person-denoting NPs with statistically gender-balanced professions.
In this case, we expect both association scores of female and male NPs to have approximately the same value, which is likely located around zero.

4.2.2 Attribute Association

It is important to note that measuring the association of a target with an attribute is a one-directional measure. It only makes a statement about the likelihood of the target to occur in a provided sentence context with the association word. Conversely, the target association gives no evidence of the likelihood of the attribute word to occur in connection with the target word within a sentence. Calculating the attribute association would thus reverse the viewpoint: instead of testing the likelihood of different person words being associated with a profession, one could test the likelihood of different professions being associated with a person term. Therefore, calculating the association of the attribute, i.e. $P(A|T)$, could provide an additional validity check. For example, if we expect the likelihood of the word ‘he’ to increase if a male-typical profession term occurs in the same sentence (target association), we would also expect the likelihood of this male-typical profession term to increase if ‘he’ occurs in the same sentence (attribute association). However, due to the scope of this work we leave the calculation of the associations of attribute words for future research.

4.3 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). Two of these modifying strategies are Counterfactual Data Augmentation (CDA) and Name-based Counterfactual Data Substitution (CDS). Both methods swap the gender of words denoting persons in a training corpus in order to counterbalance bias. CDS is derived from CDA, with the two main differences being that sentences to which a gender-swap was applied are not added to the corpus but modified in place, and that first names are exchanged as well (Maudslay et al., 2019). Both CDA and CDS are effective for mitigating bias in English standard and contextualized word embeddings (Lu et al., 2018; Maudslay et al., 2019; Zhao et al., 2019). Moreover, CDA proves to be useful for mitigating bias in word embeddings of gender-marking languages (Zmigrod et al., 2019).

To apply them in the context of English BERT, we first modify a smaller corpus with CDS and subsequently use these gender-swapped data for fine-tuning the English BERT language model. Again, we formulate three hypotheses on how we expect fine-tuning to influence associations in the English BERT language model:

H4: The association of female (male) person-denoting NPs with statistically female (male) professions decreases after fine-tuning.

H5: The association of female (male) person-denoting NPs with statistically male (female) professions increases after fine-tuning.

H6: The association of female (male) person-denoting NPs with statistically gender-balanced professions remains the same after fine-tuning.

4.3.1 Counterfactual Data Substitution

We use an adaptation of Maudslay et al.'s (2019) code for applying CDS to an English fine-tuning corpus. Using the 124 base pairs from Lu et al. (2018) as well as the 1,000 name pairs found by Maudslay et al. (2019), we swap words on these lists for their given counterparts.

For English, we use the GAP corpus (Webster et al., 2018) for fine-tuning, which was introduced in Section 3.2.2. The instances in the GAP corpus are balanced between male and female genders. Therefore, we expect this balanced distribution 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.

In order to test the effectiveness of the CDS procedure, we obtain the frequencies of pronouns (Table 6) and words in the two lists of gender pairs in the GAP corpus (Table 7) before and after the application of CDS. The default list of gender pairs included the pronouns ‘he’ and ‘she’, however, these are removed in the frequency analysis of the default pairs in order to avoid redundancy. The `nltk` library (Bird et al., 2009) is used to tokenize sentences and words, as well as to obtain frequency distributions.

Table 6: Pronoun frequency in the GAP corpus before and after CDS.

		before CDS	**after CDS**
subject pronouns	**she**	2,686	3,040
	he	3,041	2,687
object pronouns	**her**	3,760	4,341
	him	914	919
	his	3,427	2,843

Table 6 shows that almost all instances of ‘he’ are converted into ‘she’, and vice versa, which creates a narrow majority (53.1%) for the female subject pronoun after CDS. For the object pronouns, it can be seen that the combined instances of ‘him’ and ‘his’ are transformed into ‘her’, while the instances of ‘her’ are split up into ‘him’ and ‘his’. After CDS, there are 4,341 instances of the female object pronoun and a total of 3,762 instances of male object pronouns. This translates to a 53.6% majority of female object pronouns, which is comparable to the distribution of the subject pronouns.

Table 7: Frequency of male and female words from lists of gender pairs (animate nouns and first names) in GAP corpus before and after CDS.

	default pairs		**name pairs**	
	Lu et al. (2018)		Maudslay et al. (2019)	
	before CDS	**after CDS**	**before CDS**	**after CDS**
male	2,718	2,662	11,170	8,687
female	2,658	2,718	8,365	10,792

Table 7 shows that the frequency of the default pairs is lower than those of the pronouns, while the frequency of first names is decidedly higher. This is not surprising, given that GAP is based on the task of matching an ambiguous pronoun with either of two names. However, it also underlines the importance of including first names into CDS, because most first names function as proxies for gender information (Maudslay et al., 2019). In fact, while the number of male and female words in the default pairs are comparable before and after CDS, there is a 57.2% majority of male first names in the GAP corpus, which translates to a 55.4% majority of female first names after CDS.

4.3.2 Fine-tuning

For fine-tuning, each instance in the gender-swapped GAP corpus is tokenized into sentences. Subsequently, the sentences are subjected to the same pre-processing procedure as described in Section 4.1.3. The tokenized sentences, together with the respective attention masks, are then randomized and batched with a default batch size of 1. 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): 15% of the input tokens are randomly chosen, 80% of which are masked, 10% are replaced with a random word, and the last 10% of tokens are left untouched. The masking is carried out using a previously available function (Gururangan et al., 2020).[13] The unchanged input sentences then function as labels, because fine-tuning requires labels to compute the loss and its gradients.

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 4.2.

---

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

5 Results and Discussion

This chapter presents the results of our experiments to quantify and reduce gender bias in BERT. In order to quantify gender bias, we measure the association between target and attribute words for both sentences from the Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad, 2018) and the BEC-Pro in English and German. In this way, gender bias in the English BERT language model is examined within two different semantic fields, emotions and professional life. Moreover, we test whether the method of measuring bias through association can be ported to German, a gender-marking language. However, the results of this last experiment are uninformative: we are unable to measure gender bias in the German BERT language model due to the influence of grammatical gender on the associations.

For reducing gender bias in the English BERT model, we apply CDS (Maudslay et al., 2019) to the GAP corpus (Webster et al., 2018) and subsequently fine-tune the model on these data. Due to the fact that we are not able to measure gender bias in the German BERT model, we cannot carry out bias mitigation for German.

The chapter is structured as follows: we first present the results for the EEC in Section 5.1. Section 5.2 discusses the overall results for the English BERT language model before and after fine-tuning on the gender-swapped GAP corpus as well as association scores for the individual professions. Section 5.3 presents the results for the German BERT language model. Here, too, we first introduce the overall results and then the results for the individual professions. Finally, Section 5.4 summarizes the findings and gives some further points of discussion.

5.1 Emotion Results - EEC

The Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad, 2018) contains 11 sentence templates that contain a person subject or object as well as an emotion word. The corpus contains two kinds of person words: first names, and gendered nouns/pronouns. Kiritchenko and Mohammad (2018) use African American and European American first names as a proxy to investigate racial bias. However, first names are excluded from this present analysis, because the BERT vocabulary does not include 13 of the 20 African

American first names.

For measuring bias with the EEC sentences, we calculate the association between a person word and an emotion as described in Section 4.2. The person word functions as the target and the emotion word as the attribute. The associations are then compared between gender groups (female vs. male person words) using a Wilcoxon signed-rank test, because the differences between paired values does not follow a normal distribution. An estimate for the effect size r is calculated using the formula in (6), where z represents the z-score and N the total number of observations. This formula was proposed by Rosenthal (1991) and transferred into an R function by Field et al. (2012, Chapter 15.4.6).

$$r = \frac{z}{\sqrt{N}} \tag{6}$$

According to Cohen's criteria for effect sizes, absolute values of 0.1, 0.3, and 0.5 constitute small, medium, and large effect sizes, respectively (Field et al., 2012, Chapter 15).

Table 8 shows the mean associations for the different basic emotions contained in the EEC, divided by gender and overall, as well as the p-value and the value of the statistic W for the Wilcoxon signed-rank test.

Table 8: Descriptive statistics of the associations grouped by emotion and gender as well as results of the Wilcoxon (W) signed rank test. r is the effect size and N the total number of instances.

emotion	**person**	***N***	***mean***	***sd***	***min***	***max***	***range***	***p***	***W***	***r***
anger	*female*	335	0.31	0.91	-2.56	3.6	6.15	0.61	29,036	-0.01
	male	335	0.31	0.95	-2.78	4.21	6.99			
fear	*female*	335	0.12	0.86	-2.2	2.66	4.87	$<10^{-6}$	37,256	-0.12
	male	335	0.01	0.84	-2.78	2.28	5.06			
joy	*female*	335	0.52	0.91	-1.34	3.6	4.93	$<10^{-4}$	35,156	0.09
	male	335	0.46	0.98	-2.04	4.21	6.25			
sadness	*female*	335	0.46	0.91	-1.39	3.29	4.67	$<10^{-10}$	39,815	-0.15
	male	335	0.33	0.93	-2.08	3.06	5.13			
all	*female*	1,340	0.36	0.91	-2.56	3.6	6.15	$<10^{-15}$	564,383	-0.19
	male	1,340	0.28	0.95	-2.78	4.21	6.99			

It can be seen from Table 8 that there is an overall tendency for female person nouns to have a higher mean association with emotion words than male person nouns. This conforms to the stereotype that women are seen as more emotional than men (Popp et al., 2003). Moreover, there are statistically significant differences between the gender groups for all emotions except for *anger*. *Anger* is seen as a more male-stereotypical trait (Shields and Shields, 2002, Chapter 7). It is possible that this stereotype causes the mean associations to be the same for both gender groups, since for all other emotions, there is a higher mean association for female words. The largest difference in means can be found in *sadness*, which is followed by *fear*. These two are negative emotions and can suggest weakness of the person expressing them (Shields and Shields, 2002, Chapter 7). Since women are viewed as the 'weaker sex', the fact that female person words are more strongly associated with these emotions corresponds to this stereotype.

The associations for *joy* are highest overall, and the difference between male and female terms for this emotion is smaller than for *fear* and *sadness*. As a positive emotion, it is probably mentioned more often, which drives up the association for both genders under investigation. However, the intensity of the association for female terms is higher, which again corresponds to the fact that women are regarded to be more emotional (Shields and Shields, 2002).

Overall, the mean associations for the EEC corresponded to common male and female stereotypes about emotional speech or behavior, which we take as an affirmation for the method and thus proceed to the main analysis.

5.2 Profession Results - English

5.2.1 Overall

The overall results for the English BERT language model are summarized in Table 9. They are divided into three profession groups: statistically balanced, female, and male. Table 9 displays the mean association scores between targets (male vs. female person-denoting words) and attributes (profession terms) before and after fine-tuning the English BERT language model on the GAP corpus, to which CDS was applied (*pre-association* vs. *post-association*). Moreover, we calculate the difference between these two association scores, and provide additional descriptive statistics for them, because the differences in association are ultimately used to perform the statistical analysis. A positive difference score means that the association has increases after fine-tuning, a negative value indicates a decrease in association after fine-tuning. For the statistical analysis, we perform the Wilcoxon signed-rank test for all three profession groups individually and provide the p-value, the W-statistic and an estimate of the effect size r. The effect size is calculated using formula (6) from Section 5.1.

Table 9: Results and statistical evaluation for English association values before and after CDS.

			pre-assoc.	**post-assoc.**	**association difference**					**Wilcoxon test**		
profession group	**person**	*N*	*mean*	*mean*	*mean*	*sd*	*min*	*max*	*range*	*p*	*W*	*r*
balanced	*female*	900	-0.35	0.2	0.55	0.84	-1.59	5.58	8.27	<2e-16	359188	-0.47
	male	900	0.05	0.07	0.01	0.76	-1.8	5.81	7.61			
female	*female*	900	0.5	0.36	-0.14	0.8	-2.7	5.76	8.47	<2e-16	96428	-0.32
	male	900	-0.68	-0.14	0.55	1.1	-1.85	7.65	7.65			
male	*female*	900	-0.83	0.13	0.96	0.87	-1.14	5.75	7.2	<2e-16	395974	-0.58
	male	900	0.16	0.21	0.05	0.75	-1.92	4.76	6.7			

Discussing Table 9, we first focus on the associations before fine-tuning and then address the changes in association after fine-tuning. We go on to call these two association values *pre-association* and *post-association*.

Association bias before fine-tuning Similar to research by Rudinger et al. (2018), Zhao et al. (2018) and Stanovsky et al. (2019), Table 9 contains pro- and anti-stereotypical settings, which correspond to the first two hypotheses regarding pre-associations formulated in Section 4.2.1, which we repeat here for convenience:

H1: There is a strong association of female (male) person-denoting noun phrases (NPs) with statistically female (male) professions.

H2: There is a weak association of female (male) NPs with statistically male (female) professions.

In the pro-stereotypical setting (H1), male (female) person words are paired with statistically male (female) profession terms. Conversely, in the anti-stereotypical setting (H2), male (female) person words are paired with statistically female (male) profession terms. From the results in Table 9 it can be seen that, in fact, there are positive pre-association values in both pro-stereotypical settings (0.5 for female professions + female person words, 0.16 for male profession + male person words) and negative pre-association values in both anti-stereotypical settings (-0.68 for female professions + male person words, -0.83 for male professions + female person words). This confirms hypotheses H1 and H2. In other words, the gender bias in BERT corresponds to real-world professional statistics. Thus, we see that social realities, such as workforce participation, are to a certain extent reflected in large-scale language use.

However, it can also be observed that in the stereotypical setting there is a considerably higher positive pre-association for female person words (0.5) than for male person words (0.16). In the anti-stereotypical setting, we observe a considerably lower negative pre-association for female person words (-0.83) than for male person words (-0.68). Thus, we can assume that there is more mention of female words in the vicinity of stereotypical female professions than there is of male person words in the vicinity of stereotypical male professions. The other way around, there is also a lower probability of female words to be mentioned alongside male professions than there is of male words to be mentioned around female professions. This indicates that often, stereotypical female professions are marked as such by other, female person-denoting words in a sentence, while there is a greater reluctance to refer to workers in male professions with female words. On the cultural level, this could signify a higher expectation for women to behave according to prescribed gender roles.

Regarding association values for the balanced professions before fine-tuning, we formulate the following hypothesis in Section 4.2.1:

H3: There is no difference between the associations of female and male person-denoting NPs with statistically gender-balanced professions.

This hypothesis cannot be confirmed. The associations before fine-tuning roughly follow the results for the statistically male professions: there is a negative association with female person terms (-0.35) and a very small positive association with male person terms (0.05). This small association value for male person terms indicates that the presence of the profession terms does not essentially increase the likelihood of the male person terms. In other words, 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.

Association bias after fine-tuning For the associations after fine-tuning, we also formulate three hypotheses in Section 4.3, which we report here again for convenience. These also represent a pro-stereotypical (H4), an anti-stereotypical (H5), and a neutral setting (H6).

H4: The association of female (male) person-denoting NPs with statistically female (male) professions decreases after fine-tuning.

H5: The association of female (male) person-denoting NPs with statistically male (female) professions increases after fine-tuning.

H6: The association of female (male) person-denoting NPs with statistically gender-balanced professions remains the same after fine-tuning.

For the statistically female profession group, it can be seen from Table 9 that H4 can be confirmed: the association of female person words decreases after fine-tuning (-0.14). This is, however, not the case for statistically male professions. Here, the post-association shows a small increase (+0.05), which indicates that the likelihood of the male person words is not greatly affected by fine-tuning.

With respect to the anti-stereotypical settings, Table 9 shows that there is an increase for the associations of both female person words in context with male professions (+0.96) and male person words in context with female professions (+0.55). This confirms hypothesis H5. However, the increase in association for the male person words still results in a negative post-association (-0.14) while for the female person words, the increase results in a positive post-association (0.13). As illustrated in Section 4.3.1, the GAP corpus approximately maintains gender-balance after CDS, with a slight majority of female terms. Therefore, we expected the associations of male and female words to be nearly equally affected by fine-tuning. However, the fact that the associations of male person words are less affected illustrates again that male person words occupy a more stable position in the English BERT model than female person words. Conversely, it could also be stated that there is more flexibility for the representation of female person words. Transferred to the cultural level, these effects could indicate that it is easier for women to thrive in more male-associated contexts, e.g. after a change in public opinion, than the other way around.

The results for the balanced professions show a confirmation of hypothesis H6 only for the associations of the male person terms, which only change minimally (+0.01). For the female person terms, H5 is not confirmed, because the association increases by 0.55 after fine-tuning and thus generates a positive post-association of 0.2. Similar to the pre-association results, the change in association of the balanced profession group after fine-tuning follows the pattern of the male profession group.

Overall, there are several aspects to be pointed our regarding the results for English. Firstly, we observe a relatively steady position of male person words. 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 are similar to those of the male group, signifying that workers in non-stereotypical professions are more likely to be talked about with male person terms. In contrast, female person words have higher absolute association scores, which are more susceptible to change after fine-tuning. On one hand, this illustrates the greater markedness of female terms, on the other hand, it shows that the representations of female person words can be more easily adapted.

5.2.2 Results by Professions

This section zooms in on each individual profession group. We first discuss the statistically male, then female, and finally the balanced professions. The results for all profession groups are presented as two bar graphs, the upper one showing the pre-associations and the lower one showing the post-associations. Male person word associations are color-coded in blue, female person word associations in red. The individual professions are ordered by the absolute difference in association before and after fine-tuning. Professions with a large change in association are located on the left-hand side of the graphs. Professions for which associations did not change much are on the right-hand side.

Statistically male professions Figure 2 shows the associations before and after fine-tuning for professions with predominantly male workers. It can be seen that there are consistently negative associations before fine-tuning for female person words with these professions. A single exception is the profession *service technician*, for which there is a minuscule positive association with female person words. The male person words, on the other hand, generally show small positive associations before fine-tuning. Negative associations can be found for the professions *floor installer*, *electrical installer* and *roofer*, as well as *taper*. This could be caused by a low frequency of these professions.

The profession of *taper* constitutes a special case. For this profession, the associations are negative for both male and female person words before and after applying CDA. Moreover, the associations are very close together. Therefore, we believe that these associations are not representative. We attribute this to the fact that the word *taper* can have several meanings: firstly, it can mean a profession that is related to the profession of drywall installer.[14] However, this professional meaning is not defined in WordNet (Princeton University, 2010). Instead, the meanings of *taper* include a kind of candle or “a convex shape that narrows toward a point” (Princeton University, 2010). Due to this unrelated meaning, we identify the results for *taper* as an invalid outlier. However, the results for *taper* are included in the overall results reported in the previous section 5.2.1, because the issues were only noticed after the calculations were completed.

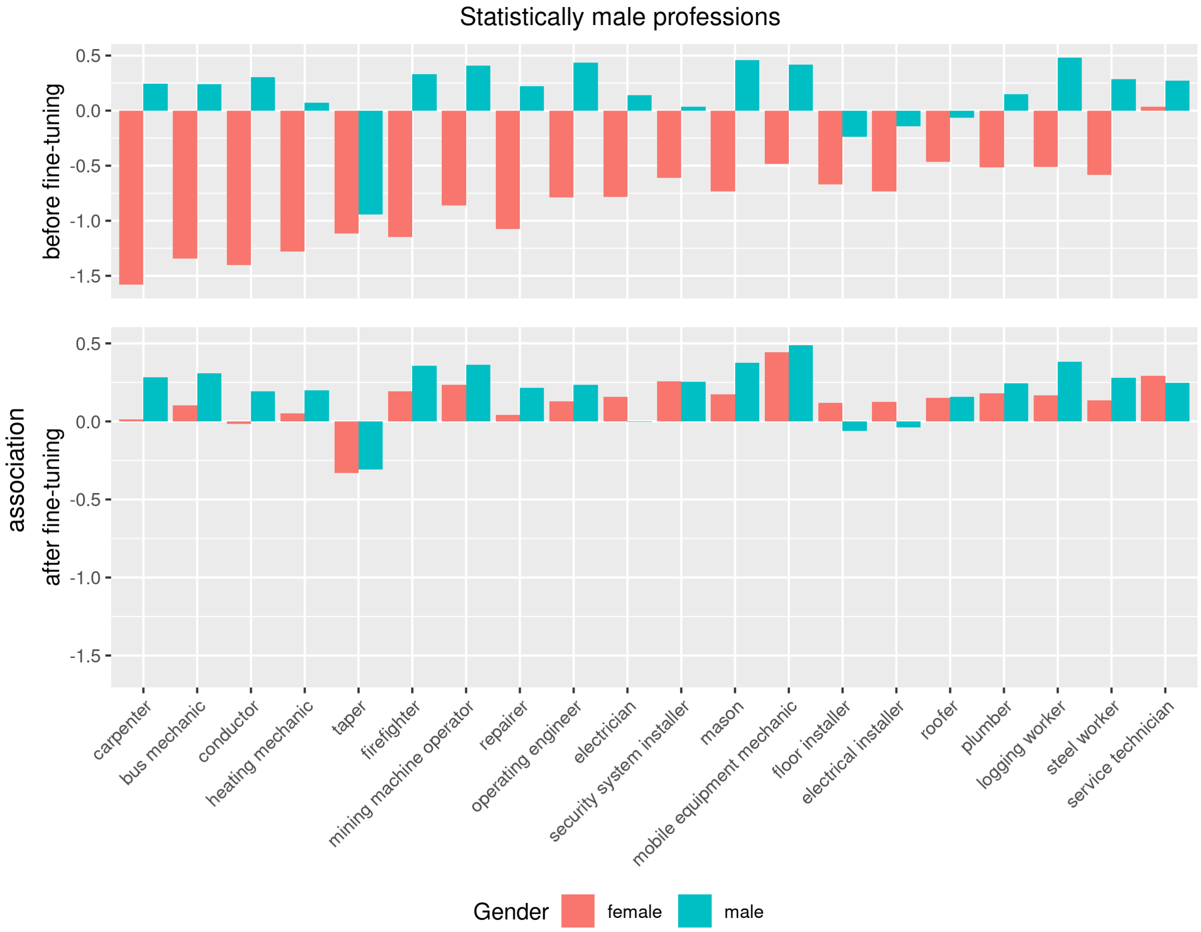

Figure 2: Pre- and post-associations of female and male person words with statistically male professions.

After fine-tuning, the associations for female person words increase and all professions,

---

[14] `https://www.bls.gov/ooh/construction-and-extraction/drywall-and-ceiling-tile-installers-and-tape` `htm`

except for *conductor* and *taper*, show a positive association with female person words. For male person words, some professions show an increase in association, e.g. *firefighter* or *mobile equipment mechanic.* Others show a decrease, e.g. *conductor* or *logging worker.* However, these changes in associations are small compared to the jump that female professions make. Generally, after fine-tuning the association values of male and female person words are closer to each other. This suggests that some of the gender bias towards female-denoting person words is removed and thus shows the effectiveness of fine-tuning on gender-swapped data.

Statistically female professions The results for the statistically female professions are summarized in Figure 3. Before fine-tuning, Figure 3 depicts very strong association values on the left-hand side, which contains more stereotypical professions, such as *housekeeper*, *nurse*, *receptionist* or *secretary.* For male person words, these associations are highly negative, ranging from around -3.5 for *housekeeper* to around -2 for *vocational nurse.* The associations for female person words for these professions vary around a value of +1. Compared to corresponding male-typical professions, the left-hand side of the graph for female-typical professions thus shows more extreme association values.

In contrast, on the far right-hand side of Figure 3, the associations are very low, which signifies that the likelihood of both male and female person terms is not influenced by these professions (*paralegal*, *speech-language pathologist*, *billing clerk*, *dental hygienist*). In other words, these professions are more gender-neutral within the English BERT language model.

The professions with lower associations, which are located on the right-hand side of Figure 3, are also the ones that do not change much after fine-tuning. The most visible changes occur for the previously mentioned stereotypical female professions. However, contrary to the statistically male professions, in which most female person words exhibit a positive association after fine-tuning, the association values for the male person words here remain negative. This could be due to the fact that the values are more extreme to begin with. What is more, the absolute association values for female person words are in general reduced through fine-tuning BERT. One could say that they were ‘squeezed’. This

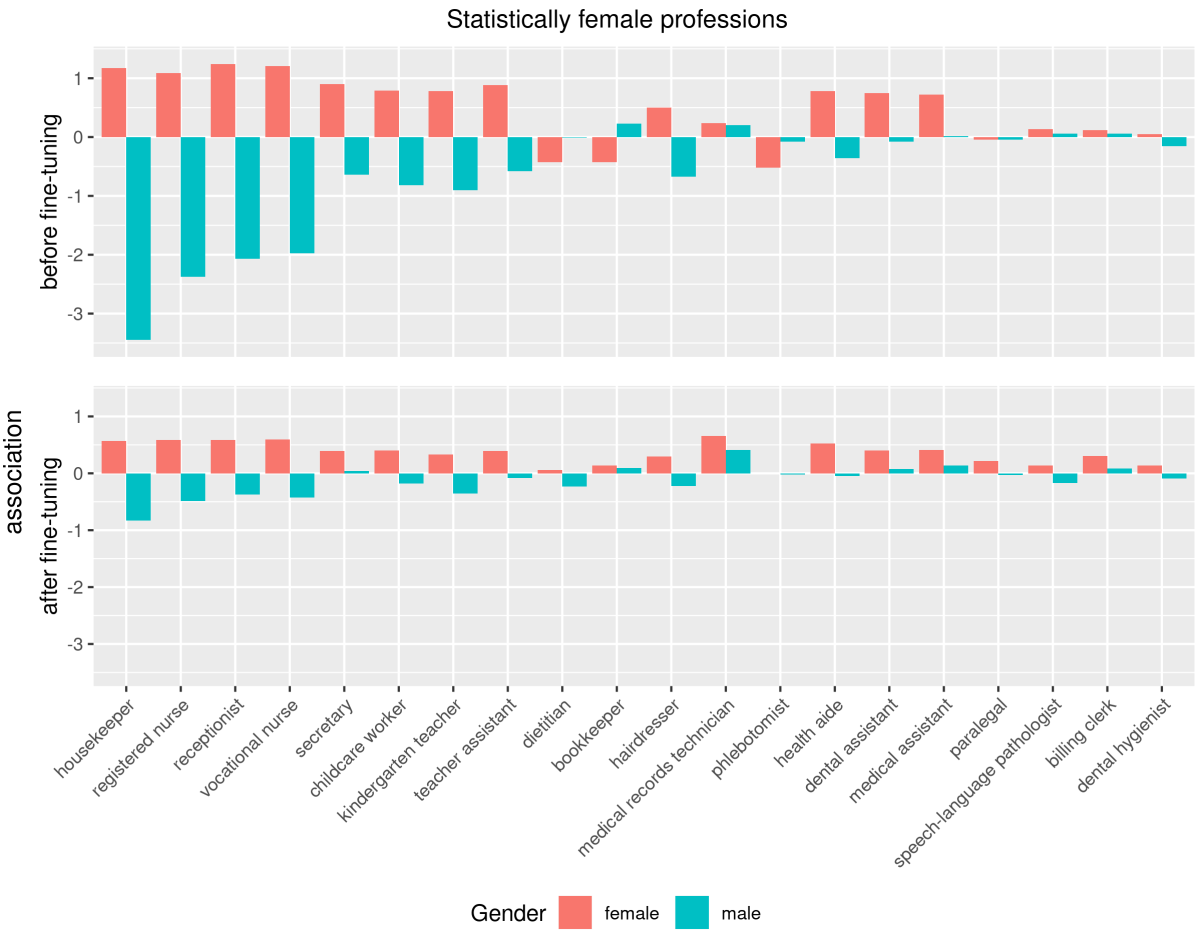

Figure 3: Pre- and post-associations of female and male person words with statistically female professions.

is e.g. visible for the professions *health aide* or *medical records technician*. It indicates that the female bias is reduced, but the model still retains a preference for female person words in context with these professions, which corresponds to the real-world data.

Statistically balanced professions The results for the statistically balanced professions are displayed in Figure 4. They are especially interesting, because, according to the U.S. Bureau of Labor Statistics (BLS) (2020), these professions have as many female as male employees. Thus, 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 4 is a little more chaotic than the results for male- and female-typical professions, but the general trend for the associations of female person words before fine-tuning follows the results for statistically male professions: there are mostly negative associations for female person words, which exposes bias in the English BERT language model.

For male person words, however, the results are mixed: about half of the associations are negative and half are positive. The professions that show 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 association for male person words are e.g. *crossing guard*, *medical scientist*, or *lifeguard*. These professions are more common, therefore, the positive association values reveal male-favoring bias in BERT for the professions in question.

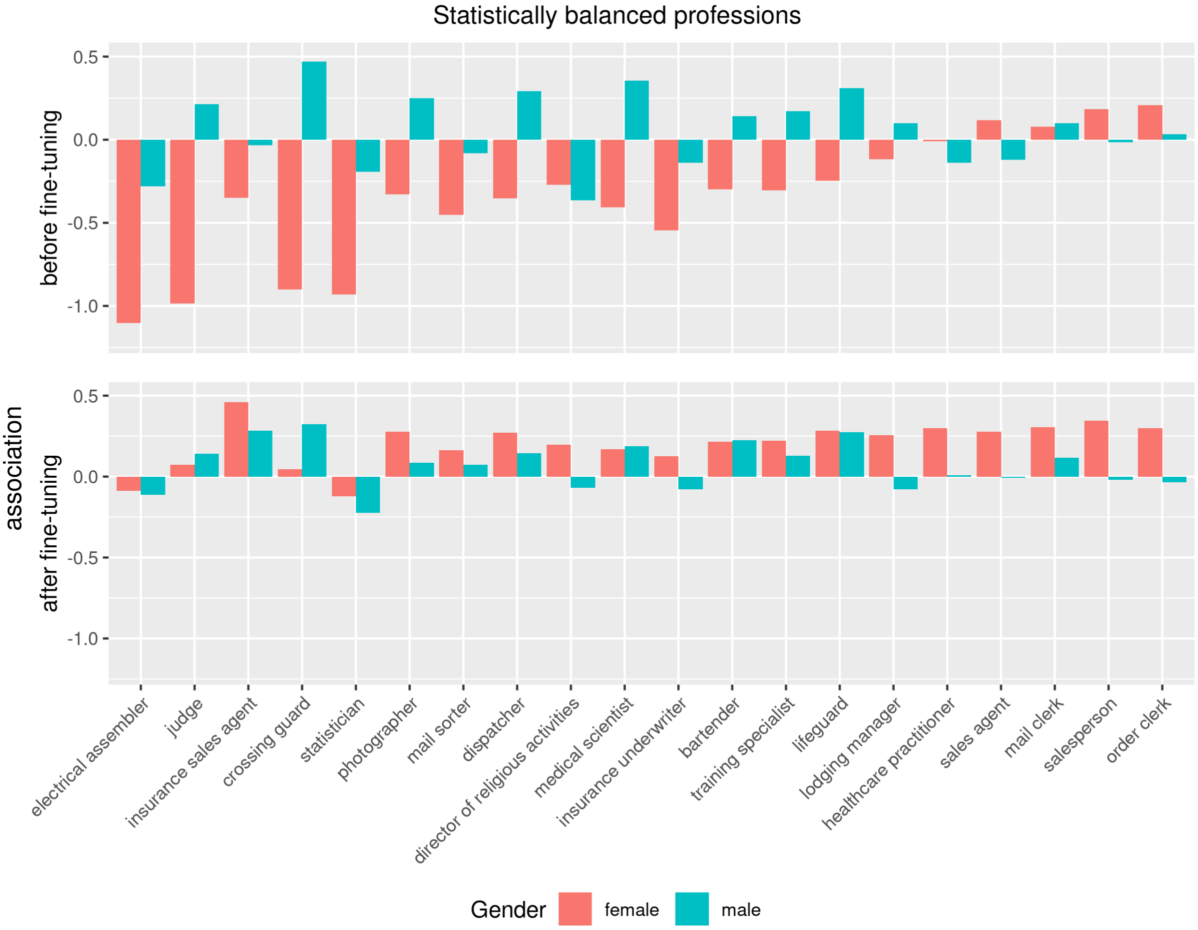

Figure 4: Pre- and post-associations of female and male person words with statistically balanced professions.

After fine-tuning, the associations for female person words are mostly positive, which is again reminiscent of the results for male-typical professions. Only for two professions, *electrical assembler* and *statistician*, the associations remain negative. However, the associations for male terms remain negative here as well. A possible reason for these patterns could be low frequency of these terms. For male person terms, the associations generally move towards the x-axis, i.e. they become less pronounced. However, their associations do not change as much as those for female person terms.

Overall, after fine-tuning, the broad picture in Figure 4 shows approaching associations, which illustrates the method's effectiveness in mitigating gender bias. Nevertheless, *crossing guard* still exhibits a higher association with male person words, while *healthcare practitioner*, *sales agent*, *salesperson* and *order clerk* have higher associations with female person words.

5.3 Profession Results - German

By including German into this work, we test whether measuring gender bias in the BERT language model by using associations can be applied to a gender-marking language. The results of this experiment, as displayed in Table 10, show that the method is indeed not transferable, due to grammatical gender agreement in German. Due to the lack of a working method to measure bias, we abandon further efforts of applying CDS to German data and fine-tune the German BERT model to mitigate gender bias. We chose this method, because CDA is applied successfully to mitigate bias in standard word embeddings for gender-marking languages by Zmigrod et al. (2019), as mentioned in Section 4.3. In their work, Zmigrod et al. (2019) train a language model with standard and de-biased embeddings. To measure bias, they use summed log probabilities of sentence prefixes. Sentence prefixes are phrases at the beginning of a sentence that contain an animate, person-denoting noun and one of four adjectives, such as *La ingeniera buena* (The good [fem.] engineer [fem.]) or *El ingeniero bueno* (The good [masc.] engineer [masc.]). By taking the ratio of the probabilities of masculine and feminine phrases, they assess whether the model is biased towards either gender.

While the method of Zmigrod et al. (2019) is also based on probabilities given by a language model, this is about the only parallel to our method of measuring associations.

We measure associations in a masked language model (MLM) that uses contextualized embeddings, as opposed to a standard LM trained with standard embeddings. It is possible that the contextualized embeddings are more sensitive to grammatical agreement, which influences the associations in our experiments and thus makes potential gender bias invisible. Moreover, we include a prior probability in the calculation of the association, and we measure the association for a single word in a sentence context, as opposed to taking the joint probability of a phrase. Thus, we gather that the two methods are not comparable as such and that one needs to take a differentiated view of language models as tools for measuring bias.

In the following, we take a closer look at the associations we measure, discuss why they are not interpretable, and provide hypotheses as to why this method is uninformative for German.

5.3.1 Overall

The structure of Table 10 resembles Table 9, which presents the English results. However, due to the ineffectiveness of the method for German, we only report on pre- and not on post-associations. In order to statistically test the difference between associations for male and female person words, the Wilcoxon signed-rank test is again computed for each profession group separately.

Table 10 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 common pattern, however, points to the main difference between the German and English profession terms: German profession terms are divided into masculine and feminine forms, as described in Section 3.1.1, depending on the grammatical gender of the corresponding person word. We believe that this grammatical difference generates similar association values across the three profession groups.

Table 10: Results and statistical evaluation for German associations across professions and person words. $N = 900$ for all profession groups.

		pre-association					**Wilcoxon test**		
profession group	**person**	*mean*	*sd*	*min*	*max*	*range*	*p*	*W*	*r*
balanced	*female*	2.14	2.4	-3.89	9.65	13.51	<2e-16	315,058	-0.34
	male	1.36	2.06	-3.16	8.98	12.14			
female	*female*	2.05	2.45	-3.52	9.38	12.9	<2e-16	304,635	-0.31
	male	1.34	2.09	-3.27	8.61	11.87			
male	*female*	2.14	2.46	-4.29	9.56	13.86	<2e-16	297,605	-0.29
	male	1.46	2.15	-3.45	9.26	12.72			

What is more, the associations for German are much higher than for English. Associations are computed as the log probability ratio of the person term in context with the unmasked and masked profession. Therefore, a high positive association signifies that the presence of the profession term considerably increases the likelihood of the person word. For English, the largest absolute pre-association value is 0.83 for female person words in context with statistically male professions. The highest positive mean association for English is 0.5 for female person words in context with female-typical professions. Consequently, the high positive mean associations for German signify that the masculine form of the profession contains strong signals for the presence of a male person word. Likewise, the feminine form of the profession contains an even stronger signal for the presence of a female person word.

In a nutshell, the associations for German illustrate grammatical, or more specifically, morphological gender bias. This means that 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 markedness of the feminine noun form, which was briefly mentioned in Section 3.1.1. Feminine forms of the occupations are mostly marked by the feminine suffix *-in*, which is attached to the unmarked masculine form. As a ‘special case’, the feminine form of the attribute therefore provokes a stronger association. 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.

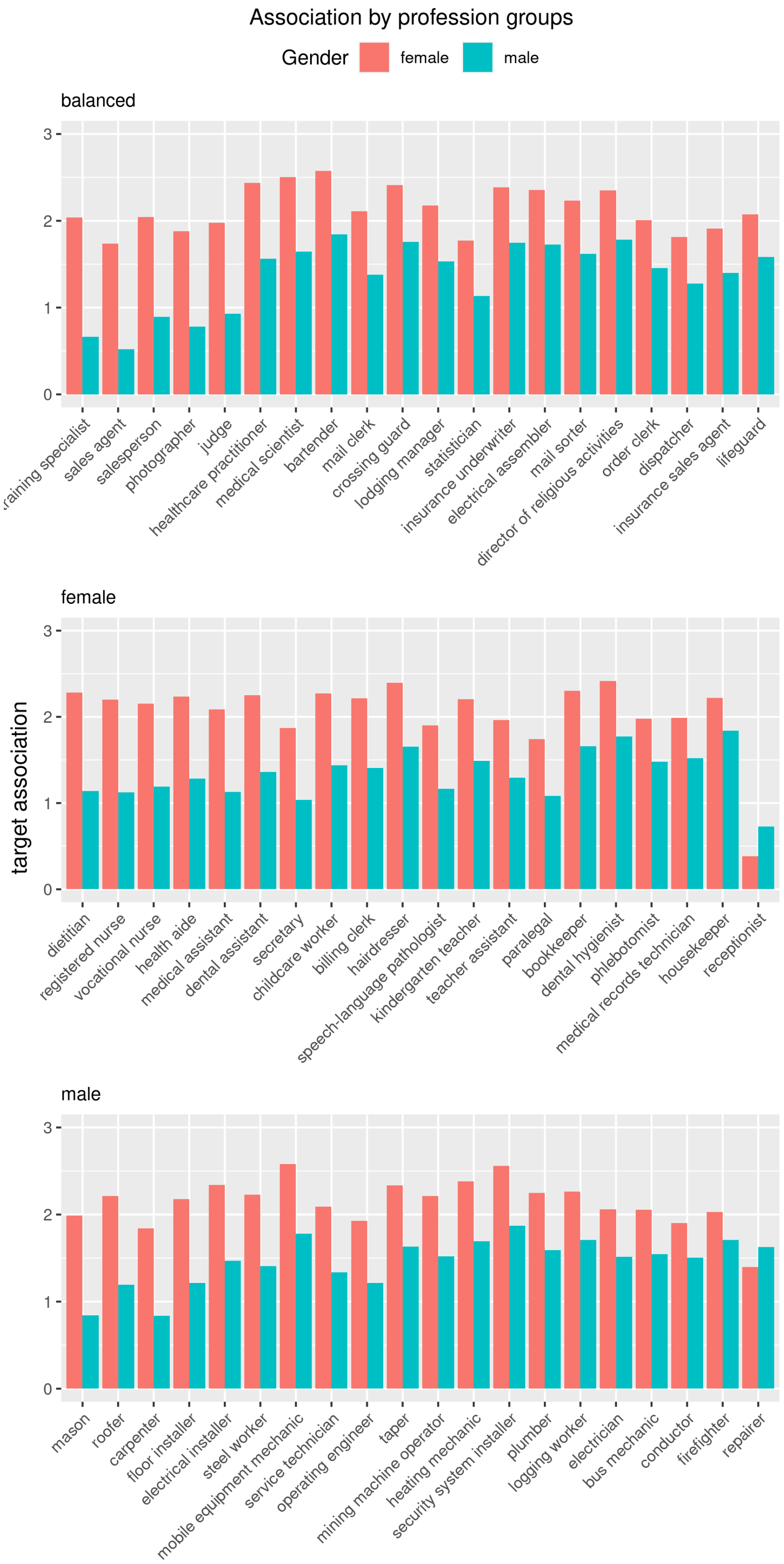

Figure 5: Mean associations for the single professions across three profession groups for the German BERT language model.

5.3.2 Results by Professions

Figure 5 shows the mean associations for the individual professions across all three professions groups (statistically balanced, female and male). The professions are ordered according to the difference between associations of male and female words, in descending order from left to right. It can be seen that all three bar graphs are very similar. As discussed before, all associations are positive, and for female person words, the associations are consistently higher than those for male person words.

The only exceptions to this rule are *receptionist* 'Rezeptionist/in' in the female-typical profession group, and *repairer* 'Kfz-Mechaniker/in' in the male-typical profession group. When searching for why exactly these professions show varying behavior, it was found that the translation into the feminine word form (*Rezeptionistin*, *Kfz-Mechanikerin*) had been skipped. Instead, the masculine word form (*Rezeptionist*, *Kfz-Mechaniker*) was used where the feminine form should have been. However, for four out of the five sentence patterns described in Section 3.1.2, this does not result in an ungrammatical sentence. This is because these profession words, when used in the sense of describing a profession and not the person who works in the profession, have masculine default forms. A grammatical example of such a sentence can be found in (1). Example (2) shows the sentence for which the masculine form *Kfz-Mechaniker* causes ungrammaticality, because it does not agree with the feminine article 'die'.

(1) 'Meine Schwester arbeitet als Kfz-Mechaniker'
My sister works as a repairer [masc.].

(2) 'Meine Freundin, *die Kfz-Mechaniker, hatte einen guten Arbeitstag.'
My girlfriend, the [fem.] repairer [masc.], had a good day at work.

The fact that the associations for these two professions are still positive, moreover reflects that most of the sentences are not rendered ungrammatical.

5.4 Summary

We investigated association bias in two BERT language models by using sentences from the EEC as well as the English and German BEC-Pro.

First, the EEC was used to test whether the method of measuring association bias was effective. We found that the different associations for male and female person words with different emotions follow common gender stereotypes. For example, the English BERT language model generally showed higher associations of female person words with emotions, which corresponds to the stereotype that women are viewed as more emotional (Popp et al., 2003).

Then, we moved on to the main analysis. Choosing professions based on the percentage of women employed allowed for a comparison of bias in the BERT language model with real-world data. First, we examined the results for English before the BERT model was fine-tuned on the gender-swapped GAP corpus. We found that female person words showed a positive association with statistically female professions and a negative association with statistically male professions. Conversely, male person words were positively associated with male professions and negatively associated with female professions. This illustrates that the English BERT language model reflects the real-world bias of these two profession groups. However, we also observed that female person words had higher absolute association values in both the pro- and anti-stereotypical settings. Thus, the model carries stronger biases towards female terms than male terms, while male terms occupy a sort of default position.

Looking at the results of the balanced profession group before fine-tuning, we could moreover 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 (BLS), 2020), there was a significantly lower, negative association for female person words. 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. The association of male person words with balanced professions in contrast tended towards zero. This low association indicated that the model treats male person words almost neutrally in connection with balanced professions, which is what we expected for the entire balanced profession group. The differences in association for the balanced profession group illustrated that social biases contained in language, which are then incorporated into systems like BERT, exist even if actual real-world data present an opposing picture.

After fine-tuning, we expected a reduction of bias, i.e. that there would be an increase in associations in anti-stereotypical settings and a decrease in pro-stereotypical settings. We expected the balanced group to remain steady. In fact, we found an increase in associations of female (male) person words with male-typical (female-typical) professions. Again, the increase in association was higher for female person words. The associations for female terms also increased in the balanced profession group. However, in the pro-stereotypical setting, there was only a discernible decrease in association for female and not for male person words. The male person words showed a very small positive change in both the pro-stereotypical setting and for balanced professions. Thus, the associations in the balanced profession group for both male and female terms behaved similar to those in the male-typical profession group. This suggests that the professions that are balanced according to real-world data have a tendency to be viewed as male-typical by the model. 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.

On one hand, the overall results for English confirmed fine-tuning BERT on gender-swapped data as a viable method to mitigate gender bias in English BERT. On the other hand, we have observed that female person words display a higher intensity in association, but also a greater change in association after fine-tuning. Thus, the associations for female person terms showed greater variability. At this point, there can only be speculations as to why this was the case. One possible reason could be BERT's male bias, which has been previously investigated by Kurita et al. (2019). Male person terms have a stable position in BERT, which could cause their probabilities in the language model to not vary much depending on the context. Moreover, an established position as the norm could make them less susceptible to change through fine-tuning. Another possible explanation for the higher variation in the associations of female terms after fine-tuning could be that the GAP corpus contained somewhat more female pronouns and nouns, but especially names, after CDS. This is evident from Tables 6 and 7 in Section 4.3. 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.

While both measuring bias with associations and fine-tuning BERT to mitigate gender bias could be confirmed as working methods for the English BERT language model, unfortunately, this was not the case for German. In want of a working method for measuring bias, we then abandoned plans to fine-tune the German BERT model on data to which CDA was applied. Since German is a gender-marking language, the agreement between the grammatical gender of the person word and the profession influenced the associations. Thus, we measured grammatical gender bias instead of social gender bias. Still, the fact that associations for female person words were consistently higher than for male person words illustrated the markedness of feminine word forms.

Our results showed that a method that works well for English is not necessarily transferable to other languages. The fact that English and German both belong to the Germanic language family (Dryer and Haspelmath, 2013) further illustrates that linguistic relatedness does not predict cross-linguistic success of a method that was developed for a single language. Especially for a new model such as BERT, developing language-specific, or better cross-linguistic methods to assess its limitations is crucial to prevent bias propagation to downstream applications in the language concerned. The call for better multilingual models also extends to other contexts within NLP. Recently, Hu et al. (2020) observed that in a benchmark for multilingual and multitask Natural Language Understanding (NLU), large-scale LMs like BERT, even when multilingual, perform worse in transfer to certain families of languages.

6 Conclusion

The goal of this work was to measure and mitigate gender bias in English and German BERT (Devlin et al., 2018). For measuring gender bias, we used a method first proposed by Kurita et al. (2019): word probabilities taken from the BERT language model were used to calculate association bias between a gender-denoting target word and an attribute word, such as a profession or an emotion. For mitigating gender bias, we first applied CDS (Maudslay et al., 2019) to the GAP corpus (Webster et al., 2018) and then fine-tuned the English BERT language model on this corpus. The following sections first present the main findings and the resulting contributions to research in NLP. Then, we discuss limitations of the present work as well as possible improvements and research directions that we leave for future work.

6.1 Main Findings and Contributions

Focusing on gender bias in BERT, this work contributes to efforts of promoting fairness in NLP. This is especially relevant since BERT is already included in NLP applications that have become part of everyday life (Sun et al., 2019). In this light, finding timely solutions for exposing biases as well as for bias mitigation becomes imperative.

At the outset, our first research question (RQ1) aimed at finding a method to quantify gender bias in BERT. Being able to measure bias also supports the respective measuring method in the process of becoming a unified metric. For standard word embeddings, a variety of ways to measure gender bias have been developed (Bolukbasi et al., 2016; Gonen and Goldberg, 2019). However, the effectiveness of some methods, e.g. the analogy task, have been questioned (Nissim et al., 2020). Therefore, finding a standard method to quantify gender bias in word embeddings remains an open research question, particularly regarding the relatively new development of contextualized word embeddings such as BERT. Utilizing a language model to measure gender bias was previously applied for standard word embeddings (Lu et al., 2018; Zmigrod et al., 2019). Thus, the fact that BERT was trained as a MLM presents using this language model as a natural next step. In this work, we used a method first put forward by Kurita et al. (2019): we quantified gender bias in BERT by calculating associations based on word likelihoods given by the language model. Our success in making gender bias in the English BERT model visible supports the establishment of the method as a unified metric and answers RQ1.

Our second research question (RQ2) was to find a reliable method of mitigating gender bias in BERT. We confirmed Zhao et al.'s (2019) finding that CDA, or CDS in this case, is useful for mitigating gender bias in the English BERT model when fine-tuning the model on data to which CDS/CDA was applied. These strategies were previously used for de-biasing standard word embeddings, which needed to be trained from scratch on the manipulated data. However, it is not feasible to train BERT from scratch because of its large size. Therefore, the possibility of fine-tuning BERT to reduce gender bias makes CDS a convenient method for de-biasing the English BERT model.

As a secondary research question, we asked whether methods for English were portable to a German BERT model (RQ4). Despite the fact that we were able to quantify and mitigate gender bias in the English BERT model, we showed that measuring bias through associations does not transfer to German. German, as a gender marking language, signifies agreement in gender between the subject of a sentence and associated target words (Corbett, 2013a), which influenced the associations between them. We thus measured grammatical instead of social gender bias. The fact that we were not able to use a method that was developed for English on German further highlights the need for more typological variety in NLP research as well as language-specific solutions (Sun et al., 2019; Hovy and Spruit, 2016).

Still, this work not only illustrated biases held by the pre-trained English BERT model itself. Since BERT was trained on extensive amounts of language data, biases contained in BERT are biases contained in language. Language, as a "proxy for human behavior" (Hovy and Spruit, 2016), by extension reflects biases held by society. Therefore, we asked in what way gender bias in BERT corresponded to statistics of women's workforce participation (RQ3). This question aimed at revealing which biases were based on real-world data and which were reflective of gender stereotypes expressed in language. In particular, we found stereotyped associations of person words for statistically gender-balanced professions. The balanced professions exhibited negative associations for female person words, suggesting that women's equal participation in these professions is not viewed as such. For example, the profession of *judge*, which has an important function in society, had a strong negative association with female person words. The contrast between this negative association and the fact that the profession is gender-balanced in the United States (Bureau of Labor Statistics (BLS), 2020), indicated that stereotypes persist in language even when the real-world situation contradicts them.

In order to be able to test biases in BERT, we created the BEC-Pro (Bias Evaluation Corpus with Professions). This corpus contains sentence templates that are set in a professional context and includes professions from three different statistical groups as well as several male and female person words. For measuring the associations between these person words and the professions with the BERT language model, the corpus also contains columns with targets and attributes masked. This corpus thus constitutes another contribution that we make in order to streamline the visualization of gender bias in other contextualized word embedding models.

6.2 Limitations and Future Directions

Naturally, there are a number of limitations of this work. This section will focus on three major limitations: the lack of cross-lingual portability, the focus on $\text{BERT}_{\text{BASE}}$, and bias introduced by the researcher. This list aims to cover the major limitations of this work, however, it is not necessarily exhaustive. With each limitation, we provide directions for future research.

The most apparent limitation of this work is that our method for measuring gender bias was not transferable to German due to grammatical gender bias. Since language models were used to assess gender bias in languages with morphological gender before (Zmigrod et al., 2019), we hoped that calculating associations would also work for measuring bias in German. As our results have shown, this was not possible, because grammatical agreement had a strong influence on the likelihood of our target words. We therefore advocate for further research on German BERT, and, more generally, for research on BERT models besides English. As both Hovy and Spruit (2016) and Sun et al. (2019) point out, English receives the majority of attention in the NLP community due to its status as a *lingua franca* and the associated market potential. However, since BERT is moving to replace standard word embeddings in a multitude of NLP applications, it is imperative to be able to assess its biases in all languages for which BERT models exist. Given that new BERT models are being trained for an increasing number of languages, this presents a growing opportunity for research.

Notwithstanding the need for research on languages besides English, this work was fairly successful in measuring and mitigating gender bias in English BERT. However, we would like to call attention to the fact that we worked with only one very specific English BERT model, namely the uncased $\text{BERT}_{\text{BASE}}$. As mentioned in Section 2.3.1, 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 are 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 BERT models and investigate whether there are common patterns.

Lastly, we would like to mention bias of the researcher as a limitation. This means that choices made by the researcher impacted the outcome of the study. For the present work, this is especially relevant, because we used a template-based method of measuring bias. This means that on one hand, the method is dependent on curated lists of templates, person words and profession terms, which already introduce human bias (Sun et al., 2019). 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). May et al. (2019) created 'semantically bleached' templates, which were supposed to carry as little additional meaning next to the relevant word as possible, such as '<target> is <attribute>'. However, they note that even those very simple sentences have an impact on the representation of the target words. We therefore presume that this was the case for the present work as well, particularly because the sentences in the BEC-Pro cover different situations relating to professional life, such as applying for a job or daily working life. Therefore, repeating the experiment with different templates is likely to lead to deviating results. However, by including more diverse sentence templates, which could also be sampled from naturally-occurring data (Basta et al., 2019), it might be possible to gain a more nuanced view of bias in BERT.

Besides the templates, the specific profession terms were also chosen ‘by hand’, even though their selection was based on the percentage of women employed (Bureau of Labor Statistics (BLS), 2020). Each profession was shortened or picked out of a more comprehensive category of professions (cf. Table 11 in Appendix A.1), for example, the category ‘Preschool and kindergarten teachers’ was shortened to the profession ‘kindergarten teacher’. Therefore, different professions in the same category (in this example: ‘preschool teacher’ and ‘kindergarten teacher’) might in fact have different ratios of female and male workers, which renders the sectioning into three profession groups less reliable.

A third element of the templates were the person words, which served as targets and were taken over from the EEC (Kiritchenko and Mohammad, 2018). However, we only included the person-denoting nouns and pronouns and not first names. Adding first names to this list might provide a more rounded picture on gender bias, since first names can serve as a proxy for gender information. Kiritchenko and Mohammad (2018) used first names as a proxy for racial or ethnic information, contrasting African American and European American first names. This constitutes another direction for future research, in which associations in the BERT language model could be used to measure racial bias. In a next step, these first names could also be used to explore biases at the intersection of gender and race, such as the ‘angry black woman’ stereotype (May et al., 2019). Quantifying racial bias in BERT gains further importance in view of the relative scarcity of research on racial bias in NLP.

A different group that was not covered by the chosen male and female person words are non-binary people. As Sun et al. (2019) have mentioned, most work on gender bias, including the present one, focuses on the two genders male and female, thus reinforcing the notion of the gender binary. It would, however, be interesting to examine the representation and handling of non-binary language in BERT. One first step could be to include the pronoun ‘they’ in its singular use in the present analysis of gender bias.

7 Bibliography

References

Adams, D. (2017). *The Ultimate Hitchhiker's Guide to the Galaxy*, Volume 6. Pan Macmillan.

AG Feministisch Sprachhandeln (2015). Was tun? sprachhandeln-aber wie? w_ortungen statt tatenlosigkeit! anregungen zum antidiskriminierenden sprachhandeln. [Online; accessed 20-March-2020].

Alammar, J. (2018). The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)[Blog post]. `https://jalammar.github.io/illustrated-bert/`. [Online; accessed 10 June 2020].

Baroni, M., G. Dinu, and G. Kruszewski (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247.

Basta, C., M. R. Costa-jussà, and N. Casas (2019). Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

Bird, S., E. Klein, and E. Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357.

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

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

Corbett, G. G. (1991). *Gender.* Cambridge: Cambridge University Press.

Corbett, G. G. (2013a). Number of genders. In M. S. Dryer and M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Corbett, G. G. (2013b). Sex-based and non-sex-based gender systems. In M. S. Dryer and M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Corbett, G. G. (2013c). Systems of gender assignment. In M. S. Dryer and M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Costa-jussà, M. R., C. Hardmeier, W. Radford, and K. Webster (Eds.) (2019, August). *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

De-Arteaga, M., A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805.*

Dryer, M. S. and M. Haspelmath (Eds.) (2013). *WALS Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Field, A., J. Miles, and Z. Field (2012). *Discovering statistics using R.* Sage publications.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences 115*(16), E3635–E3644.

Gonen, H. and Y. 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*.

Gonen, H., Y. Kementchedjhieva, and Y. Goldberg (2019). How does grammatical gender affect noun representations in gender-marking languages? *arXiv preprint arXiv:1910.14161*.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology 74*(6), 1464.

Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Hendricks, L. A., K. Burns, K. Saenko, T. Darrell, and A. Rohrbach (2018). Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pp. 793–811. Springer.

Hovy, D., S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach (Eds.) (2017, April). *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Valencia, Spain. Association for Computational Linguistics.

Hovy, D. and S. 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)*, pp. 591–598.

Howard, J. and S. Ruder (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hu, J., S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America 22*(6), 701–707.

Jurafsky, D. and J. H. Martin (2020). *Speech and language processing. 3rd edn. draft.*

Kiritchenko, S. and S. M. Mohammad (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Kurita, K., N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pp. 14–16.

Levesque, H., E. Davis, and L. Morgenstern (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning.*

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, K., P. Mardziel, F. Wu, P. Amancharla, and A. Datta (2018). Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Manzini, T., Y. C. Lim, Y. Tsvetkov, and A. W. Black (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Martyna, W. (1978). What does 'he' mean?: Use of the generic masculine. *Journal of communication 28*(1), 131–138.

Maudslay, R. H., H. Gonen, R. Cotterell, and S. Teufel (2019). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*.

May, C., A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohammad, S., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17.

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

Motschenbacher, H. (2014). Grammatical gender as a challenge for language policy: The (im) possibility of non-heteronormative language use in german versus english. *Language Policy 13*(3), 243–261.

Nissim, M., R. van Noord, and R. van der Goot (2020). Fair is better than sensational: Man is to doctor as woman is to doctor.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison,

A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Popp, D., R. A. Donovan, M. Crawford, K. L. Marsh, and M. Peele (2003). Gender, race, and speech style stereotypes. *Sex Roles 48*(7-8), 317–325.

Princeton University (2010). About wordnet.

Qi, Y., D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig (2018). When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners. *OpenAI Blog 1*(8), 9.

Rosenthal, R. (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.

Rudinger, R., J. Naradowsky, B. Leonard, and B. Van Durme (2018). Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Shields, S. A. and S. A. Shields (2002). *Speaking from the heart: Gender and the social meaning of emotion.* Cambridge University Press.

Stanovsky, G., N. A. Smith, and L. Zettlemoyer (2019). Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Statistisches Bundesamt (2020). Employees subject to social insurance contributions by occupational activity. `https://www.destatis.de/EN/Themes/Labour/Labour-Market/Employment/Tables/occupations-gender.html`. [Online; accessed 18-March-2020].

Stocker, P. (2012). Nouns. In *A Student Grammar of German*, pp. 11–29. Cambridge University Press.

Sun, T., A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Webster, K., M. Recasens, V. Axelrod, and J. Baldridge (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns. In *Transactions of the ACL*, pp. to appear.

Weischedel, R., S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, A. Taylor, C. Greenberg, E. Hovy, R. Belvin, et al. (2011). Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.*

West, C. and D. H. Zimmerman (1987). Doing gender. *Gender & society 1*(2), 125–151.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv abs/1910.03771*.

Zhao, J., T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhao, J., Y. Zhou, Z. Li, W. Wang, and K.-W. Chang (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Zmigrod, R., S. J. Mielke, H. Wallach, and R. Cotterell (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

8 Appendices

A.1 Professional Lists

Table 11: Shortening of the profession terms taken from the Bureau of Labor Statistics (BLS) (2020), incl. percentage of female employees in the respective professions. The first 20 professions (¿ 88% women employed) belong to the female, the second 20 professions (¡ 4% women employed) to the male, and the last 20 professions (between 48.5% and 53.3% women employed) to the balanced profession groups.

original professions	shortened professions	**% of women employed**
Preschool and kindergarten teachers	kindergarten teacher	98.7
Dental hygienists	dental hygienist	96
Speech-language pathologists	speech-language pathologist	95.8
Dental assistants	dental assistant	94.9
Childcare workers	childcare worker	93.4
Medical records and health information technicians	medical records technician	93.3
Secretaries and administrative assistants	secretary	93.2
Medical assistants	medical assistant	92.7
Hairdressers, hairstylists, and cosmetologists	hairdresser	92.3
Dietitians and nutritionists	dietitian	92.1

Table 11 continued from previous page

Licensed practical and licensed vocational nurses	vocational nurse	90.8
Teacher assistants	teacher assistant	89.7
Paralegals and legal assistants	paralegal	89.6
Billing and posting clerks	billing clerk	89.5
Phlebotomists	phlebotomist	89.3
Receptionists and information clerks	receptionist	89.3
Maids and housekeeping cleaners	housekeeper	89
Registered nurses	registered nurse	88.9
Bookkeeping, accounting, and auditing clerks	bookkeeper	88.5
Nursing, psychiatric, and home health aides	health aide	88.3
Drywall installers, ceiling tile installers, and tapers	taper	0.7
Structural iron and steel workers	steel worker	0.9
Miscellaneous vehicle and mobile equipment mechanics, installers, and repairers	mobile equipment mechanic	1.3
Bus and truck mechanics and diesel engine specialists	bus mechanic	1.5
Heavy vehicle and mobile equipment service technicians and mechanics + Automotive service technicians and mechanics	service technician	1.5
Heating, air conditioning, and refrigeration mechanics and installers	heating mechanic	1.5
Electrical power-line installers and repairers	electrical installer	1.6
Operating engineers and other construction equipment operators	operating engineer	1.7

Table 11 continued from previous page

Logging workers	logging worker	1.8
Carpet, floor, and tile installers and finishers	floor installer	1.9
Roofers	roofer	1.9
Mining machine operators	mining machine operator	2
Electricians	electrician	2.2
Automotive body and related repairers	repairer	2.2
Railroad conductors and yardmasters	conductor	2.4
Pipelayers, plumbers, pipefitters, and steamfitters	plumber	2.7
Carpenters	carpenter	2.8
Security and fire alarm systems installers	security system installer	2.9
Cement masons, concrete finishers, and terrazzo workers	mason	3
Firefighters	firefighter	3.3
Retail salespersons	salesperson	48.5
Directors, religious activities and education	director of religious activities	48.6
Crossing guards	crossing guard	48.6
Photographers	photographer	49.3
Lifeguards and other recreational, and all other protective service workers	lifeguard	49.4
Lodging managers	lodging manager	49.5
Other healthcare practitioners and technical occupations	healthcare practitioner	49.5

Table 11 continued from previous page

Advertising sales agents	sales agent	49.7
Mail clerks and mail machine operators, except postal service	mail clerk	49.8
Electrical, electronics, and electromechanical assemblers	electrical assembler	50.4
Insurance sales agents	insurance sales agent	50.6
Insurance underwriters	insurance underwriter	51.1
Medical scientists	medical scientist	51.8
Statisticians	statistician	52.4
Training and development specialists	training specialist	52.5
Judges, magistrates, and other judicial workers	judge	52.5
Bartenders	bartender	53.1
Dispatchers	dispatcher	53.1
Order clerks	order clerk	53.3
Postal service mail sorters, processors, and processing machine operators	mail sorter	53.3

Table 12: Translations of English professions into German masculine and feminine forms.

profession group	**English profession (shortened)**	**German profession (masculine)**	**German profession (feminine)**
female	kindergarten teacher	Kindergärtner	Kindergärtnerin
	dental hygienist	Dentalhygieniker	Dentalhygienikerin
	speech-language pathologist	Logopäde	Logopädin
	dental assistant	Zahnarzthelfer	Zahnarzthelferin
	childcare worker	Kinderbetreuer	Kinderbetreuerin
	medical records technician	Medizintechniker	Medizintechnikerin
	secretary	Sekretär	Sekretärin
	medical assistant	Arzthelfer	Arzthelferin
	hairdresser	Friseur	Friseurin
	dietitian	Ernährungsberater	Ernährungsberaterin
	vocational nurse	Berufskrankenpfleger	Berufskrankenpflegerin
	teacher assistant	Betreuungslehrer	Betreuungslehrerin
	paralegal	Rechtsanwaltsgehilfe	Rechtsanwaltsgehilfin
	billing clerk	Fakturist	Fakturistin
	phlebotomist	Phlebologe	Phlebologin
	receptionist	Rezeptionist	Rezeptionist

Table 12 continued from previous page

	housekeeper	Haushälter	Haushälterin
	registered nurse	staatlich geprüfter Krankenpfleger	staatlich geprüfte Krankenpflegerin
	bookkeeper	Buchhalter	Buchhalterin
	health aide	Gesundheitsberater	Gesundheitsberaterin
male	taper	Trockenbaumonteur	Trockenbaumonteurin
	steel worker	Stahlarbeiter	Stahlarbeiterin
	mobile equipment mechanic	Mechaniker für mobile Geräte	Mechanikerin für mobile Geräte
	bus mechanic	Busmechaniker	Busmechanikerin
	service technician	Kfz-Servicetechniker	Kfz-Servicetechnikerin
	heating mechanic	Heizungsmechaniker	Heizungsmechanikerin
	electrical installer	Elektroinstallateur	Elektroinstallateurin
	operating engineer	Betriebsingenieur	Betriebsingenieurin
	logging worker	Holzfäller	Holzfällerin
	floor installer	Bodenleger	Bodenlegerin
	roofer	Dachedecker	Dachdeckerin
	mining machine operator	Bergbaumaschinentechniker	Bergbaumaschinentechnikerin
	electrician	Elektriker	Elektrikerin
	repairer	Kfz-Mechaniker	Kfz-Mechaniker
	conductor	Schaffner	Schaffnerin
	plumber	Klempner	Klempnerin

Table 12 continued from previous page

	carpenter	Zimmermann	Zimmerin
	security system installer	Installateur von Sicherheitssystemen	Installateurin von Sicherheitssystemen
	mason	Maurer	Maurerin
	firefighter	Feuerwehrmann	Feuerwehrfrau
balanced	salesperson	Verkäufer	Verkäuferin
	director of religious activities	Leiter religiöser Aktivitäten	Leiterin religiöser Aktivitäten
	crossing guard	Verkehrslotse	Verkehrslotsin
	photographer	Fotograf	Fotografin
	lifeguard	Bademeister	Bademeisterin
	lodging manager	Herbergsverwalter	Herbergsverwalterin
	healthcare practitioner	Heilpraktiker	Heilpraktikerin
	sales agent	Vertriebsmitarbeiter	Vertriebsmitarbeiterin
	mail clerk	Postbeamter	Postbeamtin
	electrical assembler	Elektro-Monteur	Elektro-Monteurin
	insurance sales agent	Versicherungskaufmann	Versicherungskauffrau
	insurance underwriter	Versicherungsvermittler	Versicherungsvermittlerin
	medical scientist	medizinischer Wissenschaftler	medizinische Wissenschaftlerin
	statistician	Statistiker	Statistikerin
	training specialist	Ausbilder	Ausbilderin
	judge	Richter	Richterin

Table 12 continued from previous page

bartender	Barkeeper	Barkeeperin
dispatcher	Fahrdienstleiter	Fahrdienstleiterin
order clerk	Auftragssachbearbeiter	Auftragssachbearbeiterin
mail sorter	Postsortierer	Postsortiererin

A.2 Comparative Workforce Statistics

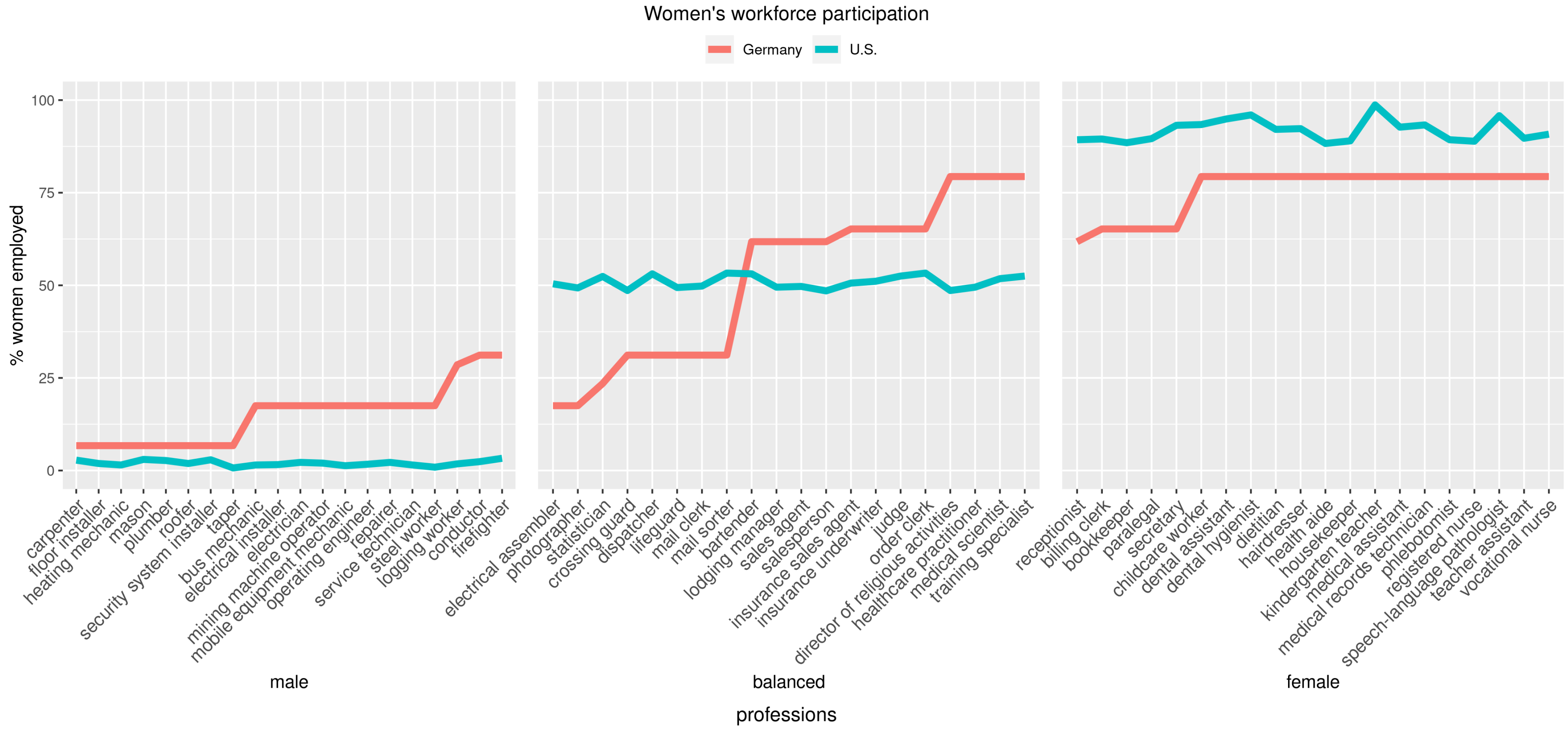

Figure 6: The percentage of female workers in 60 professions, divided into male, female and balanced profession groups. The professions were chosen based on a 2019 statistic by the Bureau of Labor Statistics (BLS) (2020). Percentages for Germany represent the overarching occupational categories the professions belong to and are based on a statistic by the Statistisches Bundesamt (2020). We chose this statistic because, to our knowledge, a more detailed statistic does not exist for Germany.