Universiteit Leiden
The Netherlands

Opleiding Informatica

Mitigating gender bias in a Dutch language model by further pre-training on feminist texts

Noa Faber

Supervisors:
Suzan Verberne & Gijs Wijnholds

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

10/7/2026

Abstract

In this thesis, we investigate whether further pre-training BERTje, a Dutch language model, on feminist texts reduces its occupational gender bias. Large language models are increasingly used for widespread applications but have been shown to contain implicit biases. We evaluate the effectiveness of further pre-training on feminist texts as a computationally inexpensive bias mitigation approach. We compile a corpus of articles from the Dutch feminist magazine LOVER and use it to further pre-train BERTje. The adapted model and the baseline BERTje model are compared on their occupational gender bias using masked language modeling (MLM) templates and a list of occupations. The models' mean Bias Scores are computed using the predicted probabilities for male and female (pro)nouns. The adapted model shows a statistically significant bias reduction at a 95% confidence level. Depending on the template used, the mean Bias Score decreases by up to 97.5%. The adapted model also maintains or slightly improves performance on named entity recognition (NER) and part-of-speech tagging (POS) tasks. These findings suggest that a small and relatively unbiased dataset is able to greatly reduce a language model's occupational gender bias through further pre-training.

Contents

1 Introduction **1**

2 Related Work **2**
- 2.1 Debiasing training data . . . 2
- 2.2 Further pre-training . . . 2
- 2.3 Bias evaluation . . . 2

3 Data **3**
- 3.1 Training data . . . 3
  - 3.1.1 Data sources . . . 3
  - 3.1.2 Data collection . . . 4
  - 3.1.3 Data statistics . . . 4
- 3.2 Evaluation data . . . 4

4 Methods **5**
- 4.1 Further pre-training . . . 5
- 4.2 Benchmarking . . . 5
- 4.3 Bias evaluation . . . 6
  - 4.3.1 Evaluation templates . . . 6
  - 4.3.2 Evaluation metric . . . 6
  - 4.3.3 Paired-difference t-test . . . 7

5 Results **7**
- 5.1 Benchmarking . . . 7
- 5.2 Bias evaluation . . . 8
  - 5.2.1 Using he/she . . . 8
  - 5.2.2 Using man/woman . . . 9
  - 5.2.3 Qualitative analysis . . . 10

6 Discussion **11**

7 Conclusions **12**

References **14**

1 Introduction

Gender bias, and specifically occupational gender bias, is widespread in large language models (LLMs) partly due to training on imbalanced datasets. Even stronger, Kotek et al. (2023) found that the occupational gender bias in LLMs aligned more closely with stereotypical perceptions than with actual job statistics [KDS23]. These biases remain present in popular real-world applications such as LLM-generated reference letters which can reinforce gender stereotypes and may harm women's hiring chances [WPS+23].

Training data can be made less biased through methods like counterfactual data substitution (CDS) with names intervention, created by Gonen et al. (2019). This method randomly gender swaps gender-specific words and names in a dataset which has been shown to mitigate gender bias in English language models [GCMT19]. Additionally, texts written by women might prove to be more balanced training data as novels and blogs written by female authors contain fewer gender stereotypes on average than those written by male authors. Female authors also mention both genders almost equally as often, unlike their male counterparts [Qia19].

In the most recent research about gender bias mitigation in a Dutch language model by Nastas (2025) it became clear that CDS does not always lead to a bias reduction. When comparing baseline BERTje to BERTje further pre-trained on a collection of children's stories, the second model performed significantly better, showing less bias. However, when comparing baseline BERTje to BERTje further pre-trained on a collection of children's stories with CDS applied, there was no significant difference. These findings suggest that further pre-training BERTje on a relatively unbiased dataset already has a significant effect and that applying CDS might add some biases back in. Alternatively, these results could also have been caused by CDS introducing more grammatical errors. Furthermore, the generalizability of the study may be limited due to its small test set written in a style similar to the dataset [Nas25].

To further investigate the effectiveness of further pre-training a Dutch language model on a relatively unbiased dataset, the models would have to be evaluated on a larger and more general test set. Delobelle and Berendt (2023) used a set of professions (translated from another study) together with the masked language modeling (MLM) template '**<mask>** is een **<profession>**' (**<mask>** is a(n) **<profession>**) to evaluate gender bias. The models' predicted probabilities for 'zij' (she) and 'hij' (he) were ranked and used to calculate the mean ranking difference (MRD), with a higher MRD indicating a more biased model [DB23].

In this study, we aim to build upon previous studies by using a relatively unbiased dataset to further pre-train a Dutch language model called BERTje [dVvCB+19]. Afterwards, both the baseline model and the adapted model are benchmarked to verify that the further pre-training did not negatively affect model performance. Lastly, we evaluate the models' gender bias using a test set similar to the test set Delobelle and Berendt (2023) introduced.

The research question of this thesis is as follows: *To what extent can further pre-training BERTje on Dutch feminist texts reduce occupational gender bias?*

2 Related Work

2.1 Debiasing training data

There are several approaches to debiasing training data. Counterfactual data substitution (CDS) with names intervention is a method that modifies training data by swapping gendered names and pronouns with a 50% probability. This is done on a per-document basis to preserve grammatical integrity. CDS is an improvement to counterfactual data augmentation (CDA) which keeps both the original and the gender-swapped corpus, resulting in a duplicated dataset that is doubled in size. CDS has been found to significantly reduce gender bias in language models [GCMT19]. However, a recent study on a Dutch language model found that training on a CDS dataset did not significantly decrease the model's gender bias. This was likely due to the dataset already being relatively unbiased, causing the CDS application to add some biases back in [Nas25].
Bartl and Leavy (2024) developed a method that uses feminist linguistics to adapt training data. This method replaces words with gender-marking affixes like 'spokesman' and 'showgirl' with their gender-neutral counterparts 'spokesperson' and 'performer'. Gendered singular pronouns were also replaced with 'they'. While its effectiveness varied depending on the model, metric and number of epochs, this method reduced gender bias in most cases [BL24].

2.2 Further pre-training

Devlin et al. (2019) developed BERT, a pre-trained language model that achieved state-of-the-art results on several natural language processing tasks. BERT is first pre-trained on large corpora and can then be fine-tuned for specific downstream tasks, which is computationally inexpensive compared to training a model from scratch [DCLT19].
Gururangan et al. (2020) found that further pre-training can significantly benefit pre-trained language models. They distinguish two types of further pre-training; domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT). DAPT is done by pre-training on the domain of a target task and TAPT is done by pre-training on the training set for a target task which is a subset of the broader domain. Both DAPT and TAPT lead to a significantly better performance on text classification tasks and a combination of both leads to the best performance in most cases [GMS+20].

2.3 Bias evaluation

Gender bias evaluation methods can be categorized into generation-based, embedding-based and MLM-based methods. Generation-based methods are the only methods that work for closed-source LLMs because they don't require any internal information. Bias is usually evaluated by engineering prompts and scoring the generated continuations. Embedding-based methods use word or sentence embeddings (which are vector representations) to measure bias. Bias can be measured by comparing target concepts like 'man' and 'woman' to neutral attributes like 'nurse' and 'doctor', for example. The higher the similarity of the embeddings, the stronger the association between words. Thus, a greater difference between similarities indicates a stronger bias. MLM-based methods use the probability distribution of the model's output for a masked word in a template sentence to measure

Table 1: Overview of the possible data sources and their characteristics.

Source	Type	Size	Format
Atria collection (digitally available)	Books, magazine articles	23,876 documents	Scanned documents
Atria knowledge articles	News, articles, interviews	≈536,000 words	HTML
DAMN, HONEY	Podcast transcripts	≈3,276,000 words	PDF
DBNL (female authors)	Books	Not available	PDF, XML, TXT
Feminer	Magazine articles	≈96,000 words	HTML
KB webarchive	Archived websites	≈130 TB (≈25,130 websites)	Not available
Lilith Magazine	Magazine articles	≈380,000 words	HTML
LOVER	Magazine articles	≈1,282,000 words	HTML
OneWorld Magazine	Magazine articles	≈11,270,000 words	HTML
Opzij	Magazine articles	≈587,000 words	HTML
SoNaR (female authors/ speakers)	Tweets, discussion lists	21,416,597 tokens (76,953 documents)	CSV
Viva Forum	Forum posts	>13.4 GB	HTML
Vrouwen Nu Voor Later	Magazine articles, pamphlets, recollections	Not available	HTML, scanned documents

bias [LGZ+24]. Gender bias can be measured by comparing the predicted probabilities for male and female pronouns or other gendered words.
Reducing gender bias for pre-training tasks like language generation is not necessarily enough to make a model less biased. Gupta et al. (2022) found that their approach was able to mitigate gender bias for open-ended text generation, but not for contextual word embeddings and downstream text classification. This highlights the importance of evaluating model biases on several tasks, including both pre-training and downstream tasks [GDK+22].

3 Data

3.1 Training data

3.1.1 Data sources

To answer the research question, we needed to find a suitable Dutch dataset. This dataset had to be relatively unbiased, contain general language and be recent enough. Additionally, it had to have a sufficient size and be publicly available in text format. As discussed earlier, texts written by women have been found to contain fewer gender stereotypes and mention both genders at a more equal rate [Qia19]. For that reason, we looked for texts written by female authors and preferably feminist or female-oriented texts to obtain a relatively unbiased dataset. Most of the data sources found are feminist online magazines. An overview of the possible data sources and their characteristics is given in Table 1.
Based on the aforementioned criteria, we identified three options: DAMN, HONEY podcast

Table 2: Case-insensitive word counts of several gendered words in the LOVER dataset.

Word	Count
Man (Man)	1098
Vrouw (Woman)	1863
Hij (He)	2096
Zij (She)	2060
Ze (She/They/Them)	9162

transcripts; Atria knowledge articles; and LOVER magazine articles. The dataset used in this study is the LOVER dataset. LOVER is a feminist magazine that was founded in 1974 and has since evolved into an online publication.

3.1.2 Data collection

All the separate articles from the LOVER website were extracted and converted into plain text format. This was done with a HTML parser written in Python using BeautifulSoup.[1] The parser looped over every page and then looped over every article URL on a page to extract the article texts. It also removed as much noise, for example donation links, as possible. The complete LOVER dataset was extracted on 14/11/2025.

3.1.3 Data statistics

The LOVER dataset is a collection of Dutch magazine articles, all approached through inclusive, intersectional and feminist journalism.[2] The different article subjects are reviews & tips, art & culture, politics & society, columns and interviews. The dataset consists of articles published between September 2017 and November 2025 and contains 1,258,524 words.
Table 2 shows the case-insensitive word counts of several gendered words in the LOVER dataset. The Man-to-Woman ratio is approximately 0.59 and the He-to-She ratio is approximately 1.02 when ignoring the word 'ze' which could mean she, they or them depending on the context. From the list of occupations that are used to evaluate the bias, 151 out of 500 are mentioned at least once in the LOVER dataset which is 30.2%. This list of occupations can be found in Table 14 in the appendix.

3.2 Evaluation data

For the bias evaluation, a list of occupations is used together with MLM templates. I derived this list from example occupations corresponding to categories in the International Standard Classification of Occupations 2008 (ISCO-08).[3] I tried to select occupations from all categories to ensure that the list covered a wide range of education levels and fields. I also made sure to only select occupations with a gender-neutral title that were included in a Dutch dictionary, to guarantee that the occupation titles were not too niche and could be used in everyday language.

[1] https://pypi.org/project/beautifulsoup4/
[2] Approximately 1% of the dataset is in English.
[3] https://www.cbs.nl/-/media/_excel/2021/27/codelijstenisco08.xls

Table 3: Non-default training settings and arguments used for further pre-training.

Parameter	Value
Model	`GroNLP/bert-base-dutch-cased`
Tokenizer	`GroNLP/bert-base-dutch-cased`
Data collator	`DataCollatorForLanguageModeling`
Batch size per device	16
Gradient accumulation steps	2
Prediction loss only	`True`
Save steps	1000
Save total limit	2

4 Methods

4.1 Further pre-training

The baseline model used in this research is BERTje, a Dutch pre-trained language model based on the BERT transformer architecture. BERTje is pre-trained with the MLM and sentence order prediction (SOP) task, which replaces next sentence prediction. It was pre-trained on a combination of several Dutch corpora including novels, news articles and Wikipedia pages [dVvCB+19].
The further pre-training of BERTje on the LOVER dataset was done in Google Colab using the Hugging Face Transformers library.[4] I used the code from Nastas (2025) with a few small alterations for this [Nas25]. During further pre-training, only the MLM objective is used. First, the dataset is tokenized. Then, the data collator applies random masking to create inputs for MLM. Lastly, the model is trained on these masked inputs to optimize MLM predictions. The training parameters are given in Table 3.

4.2 Benchmarking

The baseline model and the adapted model are both benchmarked to ensure that performance on standard Natural Language Processing tasks has not deteriorated. We used the same datasets BERTje was evaluated on to fine-tune and evaluate the models [dVvCB+19].
For the named entity recognition (NER) task we used the Dutch part of the CoNLL-2002 dataset. The dataset differentiates four types of entities: person, organization, location and miscellaneous. It also has a separate tag for the first word in a named entity and all other words. The Dutch data is made up of newspaper articles [TKS02]. For the part-of-speech tagging (POS) task we used the UDv2.5 LassySmall dataset. The dataset uses 16 universal part-of-speech tags: ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PRON, PROPN, PUNCT, SCONJ, SYM, VERB and X. [5] The data contains corpora from STEVIN D-COI, STEVIN DPC and Wikipedia excerpts [vNBVE+13]. [6]

[4] https://huggingface.co/docs/transformers/index
[5] https://universaldependencies.org/treebanks/nl_lassysmall/index.html
[6] https://github.com/UniversalDependencies/UD_Dutch-LassySmall

Table 4: Non-default training settings and arguments used for NER and POS fine-tuning.

Parameter	**Value**
Data collator	`DataCollatorForTokenClassification`
Learning rate	2×10^{-5}
Weight decay	`0.01`
Random seeds	`147, 175, 352, 547, 935`

The training and testing for both models and tasks was done in Google Colab using the Huggingface Transformers library. Each combination was ran five times with the training parameters given in Table 4. Both datasets were already divided into a train, validation and test set.

4.3 Bias evaluation

4.3.1 Evaluation templates

The baseline model and the adapted model are both evaluated on their (occupational) gender bias using MLM. This is done with the following two templates:

'**<mask>** is een **<occupation>**' (**<mask>** is a(n) **<occupation>**)
'De **<mask>** is een **<occupation>**' (The **<mask>** is a(n) **<occupation>**)

where **<occupation>** is filled in with an occupation from the list of 500 occupations that can be found in Table 14 in the appendix. With the first template, the models predicted probabilities for the personal pronouns 'Hij' and 'Zij' or 'Ze' are compared. With the second template, the models predicted probabilities for the gendered nouns 'Man' and 'Vrouw' are compared. Examples of sentences created with the templates are '**<mask>** is een auteur' (**<mask>** is an author) and 'De **<mask>** is een tandarts' (The **<mask>** is a dentist).

4.3.2 Evaluation metric

The models' predicted probabilities for the male and female (pro)nouns are used to calculate a Bias Score, introduced by Nastas (2025) [Nas25]. The Bias Score of a sentence is defined as

$$\text{Bias Score} = \frac{\max(P(\text{male}), P(\text{female}))}{\min(P(\text{male}), P(\text{female}))} \tag{1}$$

where $P(\text{male})$ is the (summed) predicted probability of the male word(s) and $P(\text{female})$ is the (summed) predicted probability of the female word(s). Consequently, the Bias Score can not be computed if $P(\text{male})$ or $P(\text{female})$ is 0.
For the first template,

$$P(\text{male}) = P(\text{'Hij'}) + P(\text{'hij'})$$

and

$$P(\text{female}) = P(\text{'Zij'}) + P(\text{'zij'}) + P(\text{'Ze'}) + P(\text{'ze'}).$$

For the second template,

$$P(\text{male}) = P(\text{'man'})$$

and

$$P(\text{female}) = P(\text{`vrouw'}).$$

The Bias Score indicates how much more likely the model is to predict either male or female words over the other. This means that the Bias Score is 1 if a model is equally likely to predict a male or female word. In contrast, a higher Bias Score indicates a bigger difference between the probabilities and thus a stronger bias towards either male or female words.

4.3.3 Paired-difference t-test

To determine whether the difference in Bias Scores between the models is statistically significant, a paired difference t-test is done [Agr18]. The null and alternative hypotheses are

$$\begin{aligned} H_0 &: \mu_d = 0 \\ H_a &: \mu_d \neq 0 \end{aligned} \tag{2}$$

where μ_d is the population mean difference in Bias Scores. The 95% confidence interval for μ_d is given by

$$\bar{y}_d \pm t \left(\frac{s_d}{\sqrt{n}} \right) \tag{3}$$

where $\bar{y}_d$ is the sample mean of the paired differences in Bias Scores, s_d is the standard deviation of the differences, t is the t-score and n is the number of paired observations. If the confidence interval does not contain 0, we can reject H_0. The corresponding test statistic is

$$t = \frac{\bar{y}_d - 0}{s_d / \sqrt{n}} \tag{4}$$

with $df = n - 1$. If the resulting p-value is smaller than $\alpha = 0.05$, we can reject H_0 and conclude that there is a significant difference between the models' Bias Scores.

5 Results

5.1 Benchmarking

The baseline and adapted BERTje model were fine-tuned and evaluated on both a NER and a POS dataset to investigate whether performance on standard NLP tasks decreased after further pre-training. Each experiment was repeated five times using random seeds.

The mean performance metrics and their standard deviations for the NER task are given in Table 5. The adapted model seems to achieve slightly lower recall and F1 scores than the baseline model, but the differences fall within the standard deviations and are therefore unlikely to be meaningful. Similarly, the adapted model seems to achieve slightly higher precision and accuracy scores, which also fall within the standard deviation. Both models attained higher F1 scores than the score achieved in the original BERTje paper. De Vries et al. (2019) reported a F1 score of 88.3% on the same CoNLL-2002 dataset [dVvCB+19]. This difference may be due to different training parameters. De Vries et al. (2019) fine-tuned BERTje for four epochs and I fine-tuned for three epochs.

Table 6 shows the obtained precision, recall, F1 and accuracy scores for the POS task. The adapted BERTje model performed slightly better than the baseline BERTje model across all metrics.

Table 5: Mean (± standard deviation) precision, recall, F1 and accuracy scores for the NER task. Results are averaged across five random seeds.

Model	Precision	Recall	F1	Accuracy
Baseline BERTje	0.8999 ± 0.0053	0.9175 ± 0.0039	0.9086 ± 0.0044	0.9887 ± 0.0003
Adapted BERTje	0.9003 ± 0.0071	0.9164 ± 0.0049	0.9083 ± 0.0058	0.9888 ± 0.0006

Table 6: Mean (± standard deviation) precision, recall, F1 and accuracy scores for the POS task. Results are averaged across five random seeds.

Model	Precision	Recall	F1	Accuracy
Baseline BERTje	0.9119 ± 0.0064	0.8863 ± 0.0040	0.8958 ± 0.0042	0.9650 ± 0.0010
Adapted BERTje	0.9161 ± 0.0044	0.8887 ± 0.0043	0.8989 ± 0.0044	0.9664 ± 0.0009

However, the differences in precision, recall and F1 scores fall within the standard deviations. Thus, they are unlikely to reflect a meaningful performance improvement. The difference in accuracy is more consistent across runs and does not fall within the standard deviation. Furthermore, the accuracy scores achieved by both models even exceed the accuracy scores reported in the original BERTje paper. De Vries et al. (2019) reported an accuracy of 96.3% for BERTje and 96.6% for the 850k checkpoint version of BERTje on the same UDv2.5 LassySmall dataset [dVvCB+19].

5.2 Bias evaluation

5.2.1 Using he/she

To determine to what extent further pre-training BERTje on Dutch feminist texts reduced occupational gender bias, the baseline and adapted BERTje model were compared using the MLM template '**<mask>** is een **<occupation>**' with 500 different occupations. For the predicted probabilities, only the top 1000 tokens were included.

Table 7 shows the mean, median and standard deviation of the Bias Scores for both models. Both distribution are right-skewed with extreme outliers that increase the mean and standard deviation. This is also visible in Subfigures 1a and 1b which show the Bias Sore distributions. The mean, median and standard deviation of the Bias Scores are all much lower and even nearing 1 (which indicates equal probabilities) for the adapted BERTje model. The mean Bias Score decreased by 97.5%.

To determine whether the difference in mean Bias Scores is statistically significant a paired-difference t-test is done. For a 95% confidence level with $df = 499$ we use $t_{0.025} \approx 1.965$. The confidence interval equals (94.149, 120.190) and the corresponding test statistic is 16.132. The p-value was smaller than the significance level ($p < 0.0001 < 0.05 = \alpha$), so we can reject the null hypothesis and conclude that the decrease in mean Bias Scores between the models is statistically significant.

Table 8 shows the amount of sentences for which the model assigned a higher probability to either male or female personal pronouns. The He-to-She ratio is approximately 54.56 for the baseline model and 1.73 for the adapted model. This is a 3053.8% decrease.

We also calculated the mean Bias Scores for sentences for which the model assigned a higher

Table 7: The mean, median and standard deviation of the Bias Scores (n = 500 sentences).

Model	Mean	Median	Standard deviation
Baseline BERTje	109.889	56.272	149.585
Adapted BERTje	2.719	1.868	2.844

Table 8: Number of sentences for which the model assigned a higher probability to male or female personal pronouns.

Model	He	She
Baseline BERTje	491 sentences	9 sentences
Adapted BERTje	317 sentences	183 sentences

probability to male or female personal pronouns, separately, as shown in Table 9. The mean Bias Score ratio is 25.55 for the baseline model and 1.09 for the adapted model.
Subfigures 1c and 1d show the corresponding Bias Score distributions for the baseline and adapted model, respectively. Here you can see that the distributions have become much more similar for the adapted model, with the curves almost completely overlapping.

5.2.2 Using man/woman

We also used the template 'De **<mask>** is een **<occupation>**' to compare the baseline and adapted BERTje model. All other aspects of the experimental setup remained the same. Given that only the top 1000 tokens were included, there is a sentence without a Bias Score as it can not be computed when one of the predicted probabilities is 0.
Table 10 shows the mean, median and standard deviation of the Bias Scores for both models. Both distribution are right-skewed with extreme outliers that increase the mean and standard deviation. This is also visible in Subfigures 2a and 2b which show the Bias Sore distributions. The mean, median and standard deviation of the Bias Scores are all much lower and closer to 1 (which indicates equal probabilities) for the adapted BERTje model. The mean Bias Score decreased by 85.1%.
To determine whether the difference in mean Bias Scores is statistically significant we performed a paired-difference t-test. For a 95% confidence level with $df = 498$ we use $t_{0.025} \approx 1.965$. The confidence interval equals (31.630, 41.667) and the corresponding test statistic is 14.314. The p-value was smaller than the significance level ($p < 0.0001 < 0.05 = \alpha$), so we can reject the null hypothesis and conclude that the decrease in mean Bias Scores between the models is statistically significant.
Table 11 shows the amount of sentences for which the model assigned a higher probability to either male or female personal pronouns. The Man-to-Woman ratio remained unchanged and is approximately 70.43 for both models.
We also calculated the mean Bias Scores for sentences for which the model assigned a higher probability to man or woman, separately, as shown in Table 12. The mean Bias Score ratio is 9.77 for the baseline model and 3.75 for the adapted model.
Subfigures 2c and 2d show the corresponding Bias Score distributions for the baseline and adapted model, respectively. Compared to the baseline model, the distributions are noticeably closer for the adapted model. However, their shapes still differ substantially. The distribution of sentences with a

Table 9: Mean Bias Scores (± standard deviation) computed separately for sentences where the model assigned a higher probability to either male or female personal pronouns. The 'He' subset contains sentences where $P(\text{male}) > P(\text{female})$, and the 'She' subset contains sentences where $P(\text{female}) > P(\text{male})$.

Model	**He**	**She**
Baseline BERTje	111.823 ± 150.260	4.376 ± 3.397
Adapted BERTje	2.807 ± 2.183	2.567 ± 3.723

Table 10: The mean, median and standard deviation of the Bias Scores (n = 500 sentences).

Model	**Mean**	**Median**	**Standard deviation**
Baseline BERTje	43.179	27.329	59.980
Adapted BERTje	6.444	5.020	4.866

higher probability for woman is much sharper and appears less smooth than the corresponding distribution for man.

5.2.3 Qualitative analysis

In this section, we will discuss some examples to illustrate the effect of further pre-training on individual sentences. The examples are based on the he/she template and are presented in Table 13. The most common outcome was that the preferred pronoun remained 'he', while the strength of the preference decreased substantially. This was the case for 317 of the 500 sentences including the occupations play-actor, nuclear physicist and taster. For example, the Bias Score of play-actor decreased from over a thousand to less than three. This suggests that the adapted model generally kept its preference but with a less extreme bias.

For 164 sentences, the preferred pronoun changed from 'he' to 'she'. In most of these cases, the Bias Score also decreased, indicating that the bias did not simply change direction. Examples of this are the occupations stockbroker and notary. Only ten sentences saw a Bias Score increase while the preferred pronoun changed from 'he' to 'she'. An example of this is the teacher occupation. However, these increases in Bias Scores are small compared to the decreases shown in most other cases.

Lastly, there were sentences that kept their higher predicted probability for female personal pronouns. For eight sentences the Bias Score increased whereas it decreased for one sentence. Examples are the occupations fashion model, bouncer and member of parliament. Although this category only contains a few sentences, these examples suggest that further pre-training does not consistently reduce bias for occupations already associated with female personal pronouns.

Table 13 also contains some examples of extreme outliers. Notably the occupations play-actor, nuclear physicist and fashion model have Bias Scores far above the mean for either the baseline or adapted model.

Table 11: Number of sentences for which the model assigned a higher probability to man or woman.

Model	**Man**	**Woman**
Baseline BERTje	493 sentences	7 sentences
Adapted BERTje	493 sentences	7 sentences

Table 12: Mean Bias Scores (± standard deviation) of sentences for which the model assigned a higher probability to man or woman. Scores are computed over the respective subsets.

Model	**Man**	**Woman**
Baseline BERTje	43.729 ± 60.225	4.474 ± 3.83
Adapted BERTje	6.510 ± 4.867	1.735 ± 0.847

6 Discussion

The results show that further pre-training BERTje on Dutch feminist texts significantly reduced occupational gender bias as measured with MLM templates. The adapted BERTje model also had equal or slightly improved performance on benchmark NER and POS tasks compared to the baseline BERTje model.

There is a difference between the results obtained with the he/she and man/woman MLM templates. While both templates showed a significant decrease in mean Bias Scores, the ratio of sentences biased towards either gender did not improve when using the man/woman template. A possible explanation is that the sentences constructed with the he/she template are more natural and common than those constructed with the man/woman template. Therefore, similar sentence constructions could be more present in the LOVER dataset which might positively affect the adapted model's predictions for the he/she template.

A limitation of the evaluation metric is that many of the predicted probabilities are very small, especially those by the baseline model. Since the Bias Score is based on the ratio between the predicted probabilities, small absolute differences can result in relatively large Bias Scores. As a result, the metric may overemphasize differences when both predictions are very unlikely. This could be one of the factors contributing to the relatively high standard deviations, extreme outliers and skewed distributions of the Bias Scores.

Overall, the findings suggest that further pre-training on a small and relatively unbiased dataset is an effective and computationally inexpensive approach to mitigate occupational gender bias. Unlike other bias reduction methods, further pre-training does not require any changes to the model itself or its initial training process as it can be added on later. The method lends itself to potential applications for other language models or other languages. A similar approach may also prove effective for other types of bias, such as racial or sexual orientation bias, if a suitable training dataset is compiled.

A limitation of this thesis is that the adapted BERTje model was evaluated on occupational gender bias specifically instead of gender bias in general. A similar reduction of the overall gender bias can be expected as the LOVER dataset is not specifically focussed on occupations. But to be certain this would have to be evaluated. Another limitation is that the models' bias was evaluated

Table 13: Representative examples of occupations and their associated biases and Bias Scores using the template '**<mask>** is een **<occupation>**'. Bias indicates the pronoun with the highest predicted probability.

Occupation	**Baseline BERTje**		**Adapted BERTje**	
	Bias	**Bias Score**	**Bias**	**Bias Score**
Play-actor (Toneelspeler)	He	1041.896	He	2.436
Nuclear physicist (Kernfysicus)	He	874.304	He	6.514
Taster (Proever)	He	4.332	He	1.068
Stockbroker (Effectenmakelaar)	He	105.737	She	1.680
Notary (Notaris)	He	48.774	She	1.005
Teacher (Leerkracht)	He	3.559	She	4.558
Fashion model (Fotomodel)	She	9.808	She	35.097
Bouncer (Uitsmijter)	She	7.902	She	7.226
Member of parliament (Tweede kamerlid)	She	2.294	She	5.427

with MLM, a task that is intrinsic to the model. Gupta et al. (2022) found that while their bias mitigation approach improved fairness in text generation, it did not in a downstream classification task [GDK+22]. To be certain that the bias reduction remains when using the model for downstream applications, the bias would have to be evaluated with a downstream task instead.

7 Conclusions

This paper has investigated the effect further pre-training a Dutch language model on feminist texts has on its occupational gender bias. The research question was: *To what extent can further pre-training BERTje on Dutch feminist texts reduce occupational gender bias?* The results show that further pre-training BERTje on the LOVER dataset significantly reduced its occupational gender bias as measured with MLM templates. Furthermore, the adapted model was able to match or slightly improve upon the performance of the baseline BERTje model on NER and POS tasks. We can therefore conclude that further pre-training on feminist texts reduced BERTje's occupational gender bias without negatively affecting its performance on standard NLP tasks.
Further research could investigate whether further pre-training on feminist texts reduces gender bias in general or whether this effect is limited to occupational gender bias. This could be evaluated using the Dutch CrowS-pairs dataset [SS25]. Additionally, future work could investigate whether the bias reduction remained for downstream applications. This could be evaluated on a Dutch dataset similar to the English Bias in Bios dataset [DARW+19].

References

[Agr18] Alan Agresti. *Statistical Methods for the Social Sciences*. Pearson, 5th global edition edition, 2018.

[BL24] Marion Bartl and Susan Leavy. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[DARW+19] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 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*, FAT* '19, page 120–128, New York, NY, USA, 2019. Association for Computing Machinery.

[DB23] Pieter Delobelle and Bettina Berendt. Fairdistillation: Mitigating stereotyping in language models. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 638–654, Cham, 2023. Springer International Publishing.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[dVvCB+19] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch BERT model. *CoRR*, abs/1912.09582, 2019.

[GCMT19] Hila Gonen, Ryan Cotterell, Rowan Hall Maudslay, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics (ACL), 2019.

[GDK+22] Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[GMS+20] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault,

editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics.

[KDS23] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA, 2023. Association for Computing Machinery.

[LGZ+24] Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):243, 2024.

[Nas25] Alexandra Nastas. Reducing gender bias in a dutch language model through counterfactual data substitution. 2025.

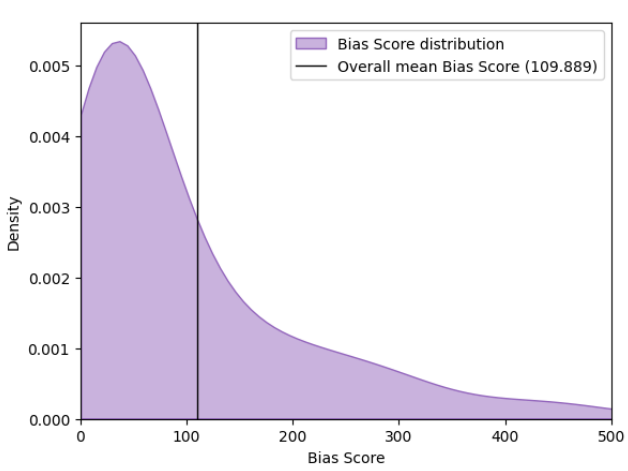
[Qia19] Yusu Qian. Gender stereotypes differ between male and female writings. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 48–53, Florence, Italy, July 2019. Association for Computational Linguistics.

[SS25] Elza Strazda and Gerasimos Spanakis. Dutch CrowS-pairs: Adapting a challenge dataset for measuring social biases in language models for Dutch. In Galia Angelova, Maria Kunilovskaya, Marie Escribe, and Ruslan Mitkov, editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1195–1204, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria.

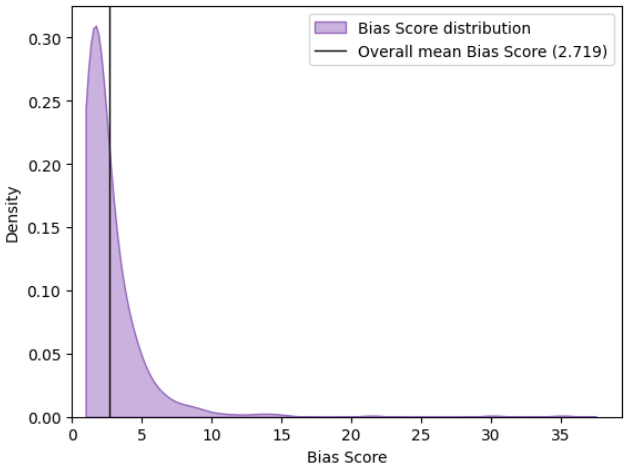
[TKS02] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[vNBVE+13] Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

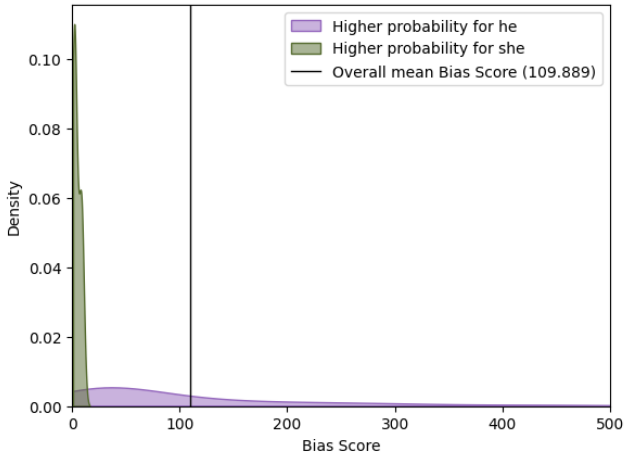
[WPS+23] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.



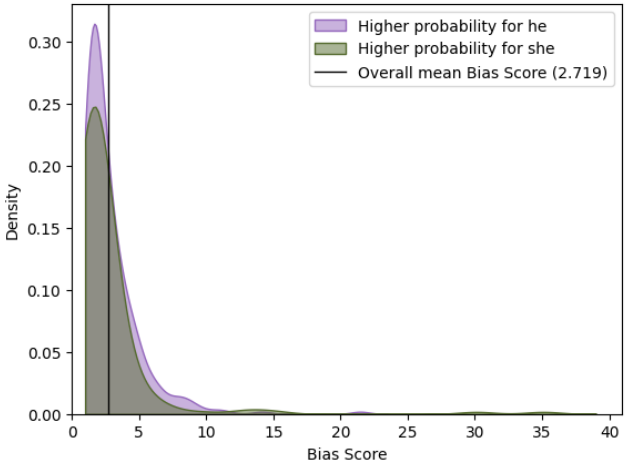
(a) Baseline BERTje model



(b) Adapted BERTje model

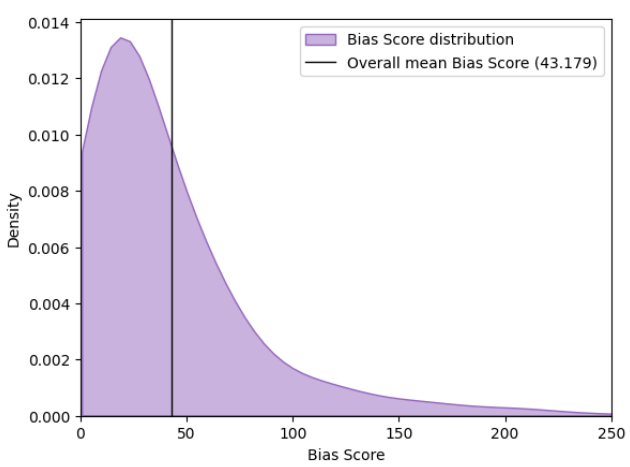


(c) Baseline BERTje model separated by prediction

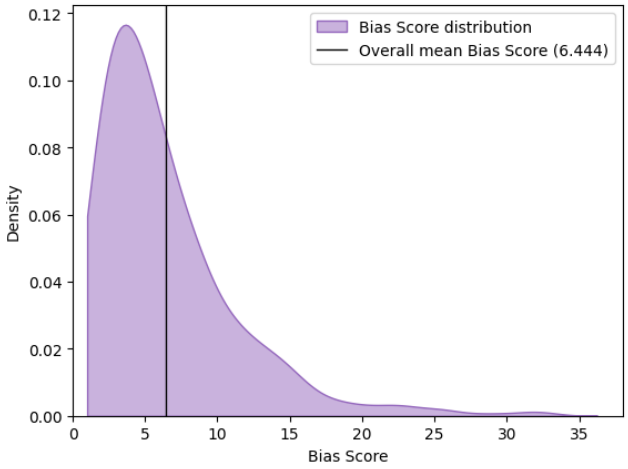


(d) Adapted BERTje model separated by prediction

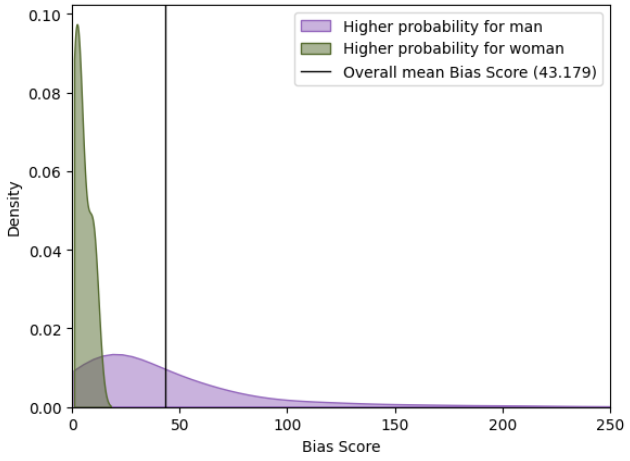
Figure 1: Bias Score distributions for the baseline and adapted BERTje models. Subfigures (a) and (b) are cropped at a Bias Score of 500 for clarity. Subfigures (c) and (d) show the distributions separately for sentences to which the model assigned a higher probability for male or female personal pronouns.



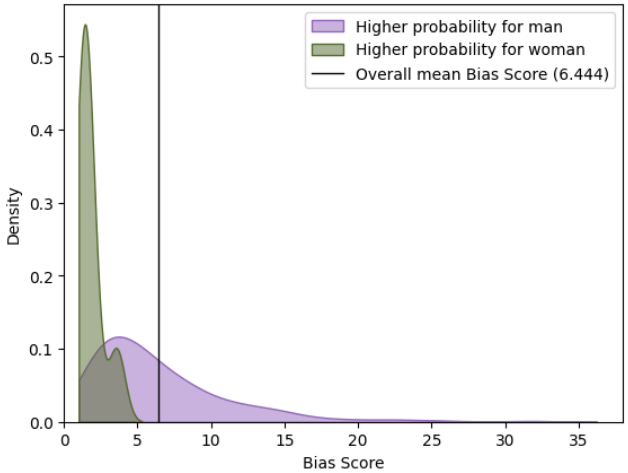
(a) Baseline BERTje model



(b) Adapted BERTje model



(c) Baseline BERTje model separated by prediction



(d) Adapted BERTje model separated by prediction

Figure 2: Bias Score distributions for the baseline and adapted BERTje models. Subfigures (a) and (b) are cropped at a Bias Score of 250 for clarity. Subfigures (c) and (d) show the distributions separately for sentences to which the model assigned a higher probability for man or woman.

Table 14: Dutch list of occupations used in MLM templates to evaluate bias.

aannemer accountant accountmanager acrobaat acteur activiteitenbegeleider administrateur adviseur advocaat advocaat-generaal afdelingsmanager afwasser agent akkerbouwer ambtenaar ambulancechauffeur ambulanceverpleegkundige analist anesthesioloog antropoloog apotheker apothekersassistent applicatiebeheerder arbeidsinspecteur archeoloog archiefmedewerker architect archivaris artiest assistent assurantieadviseur assurantiemakelaar astroloog astronoom au pair auteur automonteur autospuiter babysitter baggeraar baliemedewerker bankbediende banketbakker barbediende barkeeper bedrijfsarts bedrijfseconoom bedrijfsleider beeldhouwer beeldtechnicus begrafenisondernemer behanger beheerder belastingambtenaar belastinginspecteur belegger beleggingsadviseur beleidsadviseur beroepssporter beroepsvoetballer beroepswielrenner beurshandelaar beveiliger bewaker bezorger bibliothecaris bibliotheekmedewerker binnenhuisarchitect bioloog bioscoopoperateur bloembollenkweker bloemenkweker bloemist boekbinder boekhouder boomkweker boomverzorger boswachter bouwcoördinator bouwhistoricus bouwkundige brancardier brandwacht brandweerofficier broodbakker brugwachter burgemeester buschauffeur business analist butler buurtagent cabaretier cabinepersoneel callcentermedewerker cardioloog cargadoor cartograaf centralist chauffeur chef chef-kok chemicus chiropractor chirurg chocolatier choreograaf cipier ciseleur clown codeur collectiebeheerder columnist commentator componist conciërge conducteur conservator constructeur consulent consultant controleur copywriter corrector croupier dakdekker decaan decorontwerper demograaf depothouder dermatoloog deurwaarder dichter dierenarts dierenverzorger diëtist directeur dirigent docent documentalist doktersassistent dokwerker dominee douaneambtenaar douanier drogist duiker econoom edelsmid effectenhandelaar effectenmakelaar ehbo-instructeur elektricien elektronicus elektrotechnicus energieadviseur enquêteur entertainer eventmanager examinator expediteur exporteur exportmanager facturist fietsenmaker filiaalmanager filosoof fitter fotograaf fotojournalist fotomodel fruitplukker fysiotherapeut gasmonteur gastouder gedragsdeskundige geldwisselaar geluidstechnicus genealoog generaal geograaf gerechtsdeurwaarder gereedschapsmaker gevangenisbewaarder gids glasblazer glazenier glazenwasser goochelaar goudsmid grafdelver graficus grafoloog griffier grimeur groenteboer groentekweker grondwerker handelaar havenarbeider heier helikopterpiloot historicus hoefsmid hoogleraar horlogemaker houthakker hovenier huisarts huisschilder hygiënist hypnotiseur hypotheekadviseur ict-beheerder ict-coördinator ict-medewerker ijsverkoper illustrator imker importeur informant informaticus ingenieur inkoper inspecteur installateur instructeur interieurontwerper interviewer it-manager jager jongerenwerker journalist jurist juwelier kaakchirurg kaasboer kaasmaker kantonrechter kantoorbediende kapitein kapper kassamedewerker kelner keramist kernfysicus kinderoppas klantmanager kledingontwerper kleermaker klerk klokkenmaker koerier koetsier kok koster kraamverpleegkundige krantenverkoper kredietbeoordelaar kuiper kunsthistoricus kunstschilder kwaliteitscontroleur laborant landarbeider landschapsarchitect lasser leerkracht leerling leerlooier leerplichtambtenaar leidekker lijmer lithograaf lobbyist logopedist lokettist longarts loodgieter luchtverkeersleider maaltijdbezorger maatschappelijk werker machinist magazijnbediende makelaar manager mantelzorger marconist masseur matroos mediator melkboer melkveehouder mentor metaalbewerker meteoroloog metrobestuurder metselaar meubelmaker meubelontwerper mijnbouwkundige milieukundige minister modeontwerper mondhygiënist monteur museumgids

museummedewerker musicus muzikant natuurkundige netwerkbeheerder neuroloog nieuwslezer notaris notulist ober oesterkweker office manager officier van justitie omroeper oncoloog onderwijsassistent onderwijscoördinator onderzoeksjournalist ontwerper oogarts operator oppas opsporingsambtenaar opticien optometrist opzichter orthodontist pakketbezorger parketteur patholoog patissier pedagogisch medewerker pedagoog perronopzichter pianostemmer pikeur piloot pizzabakker pizzakoerier planoloog plugger poelier politicoloog politieagent pompbediende portier postbode pottenbakker presentator privédetective procesoperator producent productieleider productmanager productontwikkelaar proever professor programmaleider programmeur projectleider psychiater psycholoog psychotherapeut purser putjesschepper radiopresentator radioverslaggever rangeerder recensent receptionist rechercheur rechter reclasseringswerker recruiter rector redacteur regieassistent regisseur reisagent rekwisiteur restaurator retoucheur rietdekker salesmanager schaapherder schaatser scheepsbevrachter scheidsrechter schipper schoenmaker schoolbegeleider schoonheidsspecialist schoonmaker schoorsteenveger schrijver secretaris seinwachter selecteur sergeant skileraar slachter slager slijter slotenmaker sluiswachter smid softwareontwikkelaar soldaat sommelier sportjournalist sportverslaggever stationschef steenhouwer stenograaf steward straatveger strandwacht student studieadviseur stukadoor stuntwerker stuurman supervisor surveillant systeemanalist systeembeheerder systeemontwerper systeemontwikkelaar taartenbakker takelaar tandarts tandartsassistent tatoeëerder taxateur taxichauffeur teamleider technicus tegelzetter tekenaar telefonist telegrafist telemarketeer therapeut tolk toneelspeler topsporter touwslager trainer traiteur trambestuurder treindienstleider treinmachinist trouwambtenaar tutor tweede kamerlid typist typograaf uitbener uitgever uitsmijter vakbondsbestuurder vakkenvuller valkenier vastgoedadviseur vastgoedmanager veehandelaar verhuizer verkeersagent verkeersleider verloskundige verpleegkundige verslaggever vertaler verzekeringsadviseur verzekeringsagent vishandelaar visser voedingsdeskundige voeger voetbaltrainer vuilnisophaler waarzegger webdesigner webdeveloper wegbeheerder weger wegwerker werktuigkundige wetenschapper wijkagent winkelier wiskundige woordvoerder zorgmakelaar