The Birth of Bias: A case study on the evolution of gender bias in an English language model

Oskar van der Wal, Jaap Jumelet, Katrin Schulz, Willem Zuidema
Institute for Logic, Language and Computation, University of Amsterdam
{o.d.vanderwal,j.w.d.jumelet,k.schulz,w.h.zuidema}@uva.nl

Abstract

Detecting and mitigating harmful biases in modern language models are widely recognized as crucial, open problems. In this paper, we take a step back and investigate how language models come to be biased in the first place. We use a relatively small language model, using the LSTM architecture trained on an English Wikipedia corpus. With full access to the data and to the model parameters as they change during every step while training, we can map in detail how the representation of gender develops, what patterns in the dataset drive this, and how the model's internal state relates to the bias in a downstream task (semantic textual similarity). We find that the representation of gender is dynamic and identify different phases during training. Furthermore, we show that gender information is represented increasingly locally in the input embeddings of the model and that, as a consequence, debiasing these can be effective in reducing the downstream bias. Monitoring the training dynamics, allows us to detect an asymmetry in how the female and male gender are represented in the input embeddings. This is important, as it may cause naive mitigation strategies to introduce new undesirable biases. We discuss the relevance of the findings for mitigation strategies more generally and the prospects of generalizing our methods to larger language models, the Transformer architecture, other languages and other undesirable biases.

1 Introduction

Large Language Models (LLMs), such as BERT (Tenney et al., 2019) and GPT-3 (Brown et al., 2020), have become crucial building blocks of many AI systems (Bommasani et al., 2021). As these models are used in ever more real world applications, it has become increasingly important to monitor, understand and mitigate the harmful behaviours they may exhibit. In particular, many of those LLMs have been shown to learn undesirable biases towards certain social groups (Bender et al., 2021; Weidinger et al., 2021). These biases pose a serious threat for the usefulness of the technology, as they may unfairly influence the decisions, recommendations or texts that AI systems building on those LLMs generate. If we want to keep exploring the immense potential of the technology, we need to find ways to avoid or at least mitigate unwanted biases in language models.

However, detecting, mitigating and even defining undesirable biases have proven to be extremely challenging tasks. One key difficulty is deciding on where in the language modelling pipeline to measure and to intervene: in the data used for training, in the internal representations of the models, or only in the applications that are built on top of the language models (the *downstream applications*)? Many recent papers have proposed methods that work at one or two of these loci, for example, by focusing on the dataset (Dixon et al., 2018; Hall Maudslay et al., 2019; Lu et al.), the training procedure (Zhang et al., 2018; Zhao et al., 2018b; Liu and Avci, 2019), or on measuring and fixing biases in word embeddings or internal states of language models (Bolukbasi et al., 2016; Ethayarajh et al., 2019; Wang et al., 2020; Basta et al., 2019; May et al., 2019; Kurita et al., 2019; Tan et al., 2021).

In this paper, we do not choose one of these loci, but rather aim to reach an understanding of how they all three relate to each other: how do patterns in the dataset yield a particular structure in the internal states of the language model, and how does this internal structure, in turn, lead to biased behaviour in a downstream task? To answer these difficult questions, we constrain our work quite radically. First, we work with an LSTM language model and dataset that, although still involving $\sim$90 million words, is small compared to some recent, high-profile LLMs. By doing so we have full control of the training of the model, and full access to the dataset and the internal states of the model at many

intermediate points (*checkpoints*) during training. Second, we limit ourselves to only a single, heavily studied bias: gender bias (measured along a female-to-male gender axis) in English . This allows us to make use of many tools already developed for this task, including measures for bias applicable to each of the components of the language modelling pipeline and a method for debiasing.

With this setup, we study how strongly bias measurements in the various stages of the pipeline correlate, how the representations and correlations evolve over training time, and establish a causal link between the identified representation of gender and downstream bias. This provides a uniquely detailed view on the birth of one type of bias, as well as some key lessons that we expect to be useful for detecting and mitigating other biases, in other language models and other languages as well.

2 Approach

In our experiments, we study the evolution of gender bias in different representations of an English LSTM language model.[1] We explain how we define gender bias in this particular context and motivate our approach in relation to understanding the source of downstream representational harms, but how we operationalize gender bias is explained in Sections 3 and 4 when discussing the experiments.

Gender bias We understand bias as a systematic deviation in behaviour from a norm. As our focus is on gender bias in language models, the relevant behaviour we are measuring is how strongly certain words or concepts (in our case occupation terms such as *nurse* or *carpenter*) are associated by the model with one gender instead of another. This strength of association can be measured in different ways and at different points in the language modelling pipeline. In particular, we will look at bias in internal representations of the model and in its output behaviour. Ideally, the strength of association should be equal for different genders. If the model deviates from this norm, we say that the model exhibits gender bias.[2]

Whether bias in a language model causes harm, depends on the downstream application of the model and what constitutes fair and just behaviour in this particular context, but we believe that a detailed understanding of how bias is learned by and represented in these models can facilitate the development of methods to counteract bias that are tailored to a particular application and the potential harm bias can cause in that context.

With this broad scope, we hope to learn how the language model represents gender stereotypes of occupations in earlier representations of the pipeline, and how these may help explain representational harms (Blodgett et al., 2020) downstream. For instance, if a language model with gender stereotypes for occupations is used in a translation system, it may propagate the undesirable world-view of all doctors being male and all nurses being female. Understanding how these stereotypical representations come about can help in developing new detection and mitigation strategies for these and other stereotypes in AI systems building on language models.

The LSTM language model In this paper, we study the gender bias of an LSTM language model (Hochreiter and Schmidhuber, 1997). We follow the setup from Gulordava et al. (2018), and train the model on their training set of ~90M tokens, with a vocabulary of 50,000 (full-word) tokens, extracted from the English Wikipedia corpus. Following Gulordava et al., we lower the learning rate at epoch 20 using a plateau scheduler. Our training regime differs in one aspect: we use weight-tying for the encoder and decoder (Press and Wolf, 2017). We make this adjustment to simplify our analysis, as it leads to a smaller model size with comparable performance and limits the available static word vectors to one embedding space instead of two.

We train three language models with different random seeds for 40 epochs, where an epoch is defined as one full pass through all the training data. During training, we save intermediate checkpoints of the LSTM in order to examine how its behaviour develops over time. Because model behaviour changes most drastically in the first epoch, we save checkpoints with a higher granularity for that phase.

In the rest of this paper, we investigate the representation of gender (bias) in three components of the language modelling pipeline: (i) the dataset, (ii) the input embeddings (provided by the

[1] Our code can be found at https://github.com/bias-barometer/birth-of-bias.

[2] Because we heavily rely on existing tools from previous works for measuring gender bias, we restrict ourselves to representing gender along a female-to-male axis. We recognize that this is an unfortunate simplification (e.g. West and Zimmerman, 1987; Richards et al., 2016) and hope to overcome this limitation in future work.

encoder), and (iii) the downstream behaviour (a semantic textual similarity task).

3 The evolution of gender representation in the input embeddings

In order for a model to acquire undesirable gender biases, it first needs to build up a representation of the concept of gender. In fact, gender bias can be seen as an extension of this concept to words to which we do not want to assign gender. Understanding the process of how a model develops a representation of gender is, therefore, an important part of understanding the evolution of gender bias. For this reason, we start in this section with an investigation of the learning dynamics of gender in general. We will focus our analysis on gender representations within input embeddings.

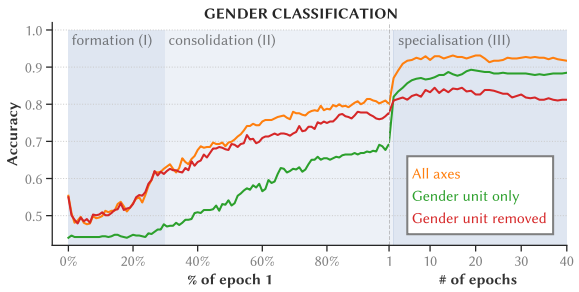


Figure 1: Classification accuracy of gender using three different classifiers, that use only the dominant gender unit (green), all other units (red), or all units (orange). Curves show results over training time, averaged across seeds.

Method Previous work has shown that gender (and resulting biases) are encoded by only a small number of units (Vig et al., 2020b; De Cao et al., 2021). We build on this work and examine what stages a model undergoes in order to obtain such a local representation of gender. For all the saved checkpoints of our LSTM models we train a **linear classifier** based on the word embeddings of 82 gendered word pairs (e.g. *he-she*, *son-daughter*, the full list is shown in Appendix C.). The classifier is trained with L2 regularisation on an 80/20 train/test split. We utilise the distance to the resulting decision boundary as a proxy for the gender subspace of the model. With the resulting set of classifiers, we conducted several experiments to gain insights into *how* gender is represented.

3.1 How localised is the representation of gender?

The results of the gender classification task are shown in Figure 1. The performance on the test corpus (orange curve) can be seen to be increasing gradually over time, already reaching around 85% at the end of the first epoch, and settling at around 93% after 3 epochs of training. Furthermore, and in line with results from other studies, we find that the representation of gender is very localised: a single unit in the embeddings dominates the representation of gender, which we call the **gender unit**.

To quantify how well this unit captures gender, and how this quantity changes over time, we train a new classifier that uses *solely* the gender unit in the embeddings. Results for this experiment are shown as the green curve in Figure 1. We find that in the initial stages, this classifier performs at chance level. After this stage, a surprisingly gradual increase in accuracy takes place, and after around 4 epochs of training the model settles at a local gender representation with an accuracy of around 90%.

The single gender unit is thus able to capture gender almost as well as the classifier that had access to the full embedding. To investigate to what extent this unit is special in capturing gender compared to the other embedding axes we also train classifiers in which the gender unit has been *removed* (red curve). It can be seen that in the initial stages this classifier performs on par with the full classifier. However, along the course of epoch 1 it slowly starts to deteriorate; in epoch 2, it is even being surpassed by the single gender unit classifier. These three curves show that the model has concentrated the majority of gender information into a single unit, but that part of it is still distributed over the remaining axes.

To see how the gender unit develops over time we compute whether or not the dominant gender unit is the same at different time points (see Figure 7b in the appendix). After ~30% of epoch 1 the model has settled on what the main unit is going to be to represent gender on. Prior to that point the model undergoes a phase in which it alternates between several gender units, none of which are equal to the final gender unit. Even though the model has already settled on the final gender unit at an early point, it still takes more than a full epoch of training before it has arranged its word embed-

ding space in such a way that gender is captured optimally by that unit.

We utilise these findings to define three distinct phases that a model undergoes to form its representation of gender: i) the **formation phase**, in which the model is exploring a suitable gender representation; ii) the **consolidation phase**, after around 30% of epoch 1, in which the model gradually restructures its space around the newly found gender representation; iii) the **specialisation phase**, after around 3 epochs, in which the model amplifies the gender signals that have been formed in the previous phase.

3.2 Which words drive the organisation of the gender representations?

Next, we examine which tokens play a vital role in the shaping of a model's gender representation. Soon after the start of training, certain embeddings start to reflect (linguistic) features such as gender. Slowly, the model forms a more general notion of gender, aligning other (gendered) tokens with the initial set of gendered tokens that drove the learning process. We utilise the decision boundary distance to examine which tokens play an early role in the development of gender. We do this for two types of classifiers: (i) the single gender unit classifier that has been explored in the previous experiments, and (ii) the classifier that utilises all but the gender unit.

The result for this procedure is shown in Figure 2. We see a striking pattern emerging here: the development of the dominant gender unit is strongly driven by female tokens, whereas male tokens dominate the development of gender information that is distributed across all other dimensions. This is in line with earlier work that showed that masculinity acts as the default gender class for a language model (Jumelet et al., 2019). A model will only prefer the prediction of a feminine token once it has encountered explicit evidence for it, and it is able to do so by channelling this information through a localised dimension.

4 The evolution of gender bias

Building on the last section, we now turn our attention to gender bias, i.e. the association of gender with words that are not explicitly gendered. Specifically, inspired by previous work (Caliskan et al., 2017; Rudinger et al., 2018; Zhao et al., 2018a; Webster et al., 2020) we consider the gender bias for 54 occupation terms (see Table 3(c) in the appendix).

4.1 From gender representation to gender bias

We follow Ravfogel et al. (2020) and use a *support vector machine* to find the optimal linear decision boundary between 18 unambiguously feminine and masculine words (also used by previous work (e.g. Bolukbasi et al., 2016; Ethayarajh et al., 2019; Ravfogel et al., 2020) [3]), of which the orthogonal axis serves as the primary gender subspace, $\vec{g}$. Given this subspace $\vec{g}$, gender bias (w.r.t. the gender-neutral norm) can be defined using the scalar projection of every input embedding, $\vec{w}$, onto the subspace, see Equation 1.[4]

$$\text{bias}_{\text{IE}}(w) = \langle \vec{g}, \vec{w} \rangle \tag{1}$$

The resulting scalar value quantifies the strength of the bias, while the sign indicates the direction on the female-to-male axis. In the rest of the paper, we refer to this bias as the input embedding (IE) bias.

When studying the average input embedding bias for the non-gendered occupation terms, we observe a steady (absolute) increase over the course of training, with the strongest growth in the first half of epoch 1, and a levelling off in the last 20 epochs (we refer to Figure 9 in the appendix).

Does this spreading out of occupation terms along the gender dimension correlate with a bias in downstream behaviour? For the purpose of this paper, we use the *semantic textual similarity* task adapted for gender bias (STS-B, Webster et al., 2020), which, as is common in the literature, measures bias on a carefully created collection of sentences (a 'challenge set'). This task contains 276 template sentences $t \in T$, where for each occupation o that sentence either starts with that occupation, "man", or "woman", resulting in a triplet

[3] We leave out the word pair ('guy', 'gal'), as we have noticed better results without the word pair. Ethayarajh et al. (2019) and Du et al. (2021) warn that including low-frequency words can negatively impact the bias measure, which we suspect is the case here.

[4] Please note that the gender subspace we define here is closely related to the approach in Section 3 for identifying the *gender unit*. Even though the classifier for finding the *gender subspace* is only trained on a subset of the gendered word-list used in the previous section—which is done to match previous work more closely (e.g. Bolukbasi et al., 2016; Ethayarajh et al., 2019; Ravfogel et al., 2020)—we find that the decision boundaries of both approaches correlate very strongly and that the observations on the locality of gender information are relevant here as well.

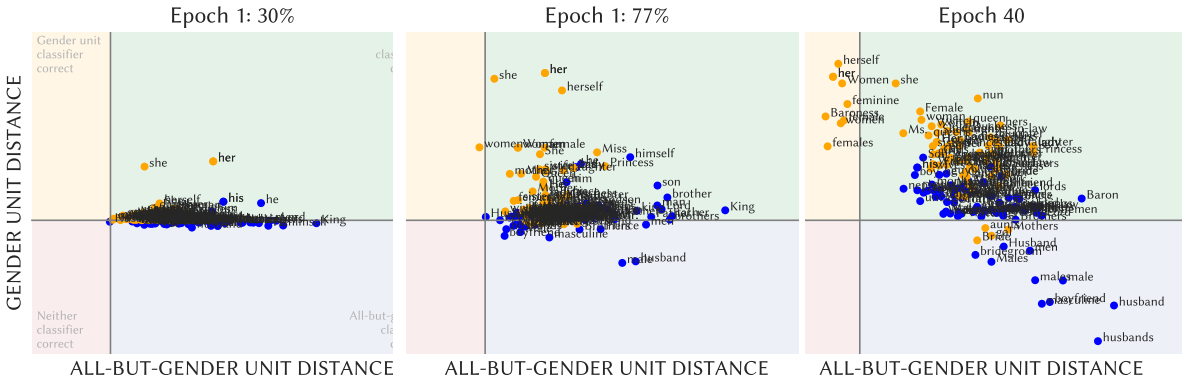


Figure 2: Gender information encoded on the dominant gender unit (plotted vertically) mainly serves to distinguish female words from other words; gender information encoded in all other units (plotted horizontally) mainly serves to distinguish male words. Shown are distances from the decision boundary for the gender-unit-only classifier and for the gender-unit-removed classifier for each word, at three different time steps during training.

$(t(o), t(\text{“man”}), t(\text{“woman”}))$. One of the sentence triplets is, for example, *“A man/woman/janitor is playing the guitar”*. The gender bias for occupation o is calculated as the average difference in similarity with the sentence starting with "man" compared to the sentence starting with "woman", see Equation 2. We use the cosine similarity of the last hidden states of our LSTM model as a proxy for the semantic similarity, to avoid training an additional semantic similarity classifier and making the relationship to the earlier stages of the language modelling pipeline less interpretable.

$$\text{bias}_{\text{STS-B}}(o) = \frac{1}{|T|}\sum_{t\in T}\text{similarity}(t(o), t(\text{“man”})) - \text{similarity}(t(o), t(\text{“woman”}) \quad (2)$$

With this measure for downstream bias in our hands, we can return to the questions whether there is a relation between the dynamic behaviour of bias we observed in the input embeddings and the bias in downstream behaviour. The answer is a qualified yes. We find that the progression of bias in the STS-B task grows very rapidly in the first few training batches, is extremely variable during epoch 1, and does grow to a level of around 0.3 in the second half of epoch 1 (see Figure 9 in the appendix). It then remains around that point for the remaining 39 epochs.

Moreover, while the *change* in the metrics is clearly no longer correlated from halfway epoch 1, at each time slice we do find a fairly strong correlation between the two measures across the vocabulary of interest. E.g., at epoch 40 we find a correlation of 60%, indicating that the input embedding bias scores for *nurse*, *receptionist*, *engineer*, *architect*, *mechanic*, etc. are fairly predictive of the downstream bias scores for STS-B sentences containing these words.

A similar observation can be made when looking at individual occupation words (Figure 3). Here as well do we find that the input embedding and STS-B bias are correlated. For instance, both bias measures broadly capture a strong male bias for “engineer”, while “nurse” and “receptionist” have a strong female association for both representations. Complementary to this, we find that for both the input embeddings and STS-B, some words show these biases much sooner than other words. For the word “nurse”, for example, a female bias can be found earlier during training than for “receptionist’, even though both have a strong female bias after training. We hypothesize that this reflects the differences in their dataset statistics. For instance, we find that “nurse” occurs 783 times, while “receptionist” only 66 times. On top of that, “nurse” also has a higher PMI association with female gendered words (we explain the PMI association in more detail in Section 4.2).

However, the fact that the correlation between the two metrics is not higher than 0.6 highlights that there still are some important differences. First, we notice that the STS-B bias is noisier than its input embedding counterpart in the first epoch, which is not a surprise given that the language modelling relies on contextual information and is measured on a relatively small set of examples. More impor-

tantly, however, we observe in Figure 3 that the gender bias is heavily skewed towards a female bias for the input embeddings, but this asymmetric pattern is not as apparent for the STS-B task. It seems that this asymmetry gets masked at the level of the downstream task, but the underlying cause is still asymmetric, which is relevant when considering countermeasures. We will come back to the asymmetry in gender bias in Section 5.

4.2 Relating gender bias back to dataset statistics

So far, we have seen that the way gender is represented in gendered words helps us understand how gender bias is represented in the input embeddings of non-gendered words, and how these representations change over time. Moreover, we have seen that the used bias metric at the level of these input embeddings is fairly predictive of the downstream bias measured through STS-B. We now turn our attention to the question of how and why non-gendered words get mapped to the emergent gender axes of the language model. For this, we examine how well the model biases correlate with dataset features and external U.S. labour statistics with the ratio of male and female workers for each occupation (see Appendix A). We will not be able to give a firm answer to this question, as neural models are capable of learning from more sophisticated, and perhaps implicit, features of the dataset than we consider, but there are still some interesting patterns we can observe.

Following others (Zhao et al., 2019; Tan and Celis, 2019; Fast et al., 2016; Gao et al., 2020), we examine the word-count statistics for the dataset in our experiments. We consider two statistics, namely (i) the word counts and (ii) the pointwise mutual information (PMI) with a set of 18 gendered words (see Table 3(a,b) in Appendix C). The PMI statistic is defined as given in Equation 3, where $p(x)$ indicates the probability of word x, which we estimate by the word count $c(x)$. The joint probability $p(x, y)$ is estimated with the co-occurrence count for words x and y, for which we use a sliding window of 35 tokens that is equal to the BPTT window of our LSTM models. In our case, x is an occupation and y the set of gendered words (either female or male words, indicated by subscript ♀ and ♂, respectively). We also combine the PMI statistics for the two genders to capture an aggregate association, where $PMI_{♂-♀} = PMI_{♂} - PMI_{♀}$.

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{c(x, y)}{c(x)c(y)} \quad (3)$$

We first check how these statistics are correlated with each other, looking only at the dataset (independently of the language model). We find that $PMI_{♀}$ (-0.23) correlates fairly well with the labour statistics, and more strongly than $PMI_{♂}$ (0.12). In other words, female gendered words ("she", "her", "woman") in the vicinity of an occupation term, are weakly predictive of the percentage of female or male workers in that occupation, while male gendered words reveal less[5]. The highest correlation, however, is obtained with an aggregate of the two PMI measures, $PMI_{♂-♀}$ (0.33).

Partitioning the training period in the three phases for gender representations that we identified in the previous section, we see an interesting pattern of results for the correlation with the input embedding bias (see Figure 4). In the *formation* phase, all correlations are low, except for the correlation with word count; i.e., high bias scores are best predicted by simple frequency of the term.

In the *consolidation* phase, word count starts losing its predictive power for bias, and the correlation with the labour statistics starts building up, reaching approximately 40% by the end of the *consolidation* phase and remaining there throughout the *specialization* phase. Note that the labour statistic is external; the language model only has access to statistical patterns that are reflected in the input text. We do not know which text statistic mediate the formation of this correlation, but it is interesting that the steepest growth of the labour statistic correlation, coincides with the aggregate PMI-measure $PMI_{♂-♀}$, taking dominance over both female and male specific measures.

4.3 Summary

The projection on the gender subspace from Section 3 finds plausible gender associations with different occupations, and we observe that the input embedding bias measure is predictive of the bias that we measured in the STS-B task. However, the correlation is far from perfect, and there are some interesting differences between both measures with respect to gender asymmetry. We also saw that the

[5]This is in line with the often observed *male-as-norm* phenomenon in language: the male category is used more generally, while female gendered words are more specific for indicating that particular gender (Danesi, 2014).

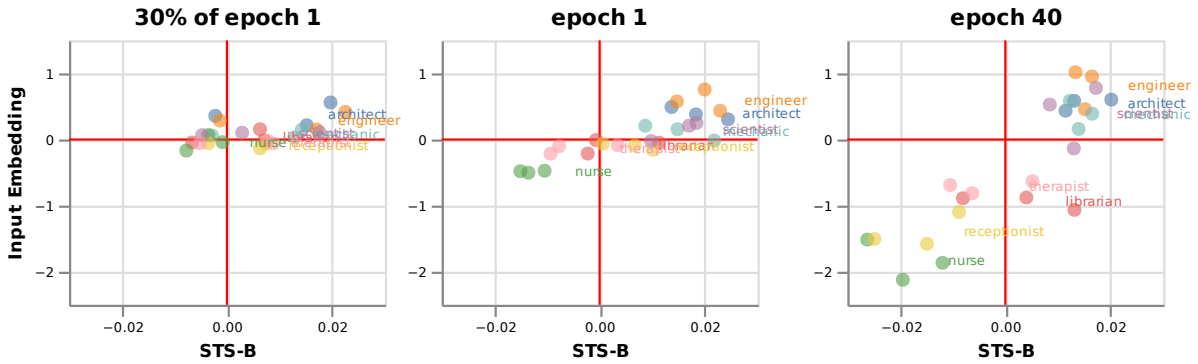


Figure 3: For three different points in time, we show the input embedding bias and STS-B bias for a selected few occupation words, averaged over the three different random seeds. The occupation terms visualised in this figure are "receptionist", "nurse", "librarian", "therapist", "mechanic", "engineer", "scientist", and "architect", which we have found to display strong biases for both bias metrics.

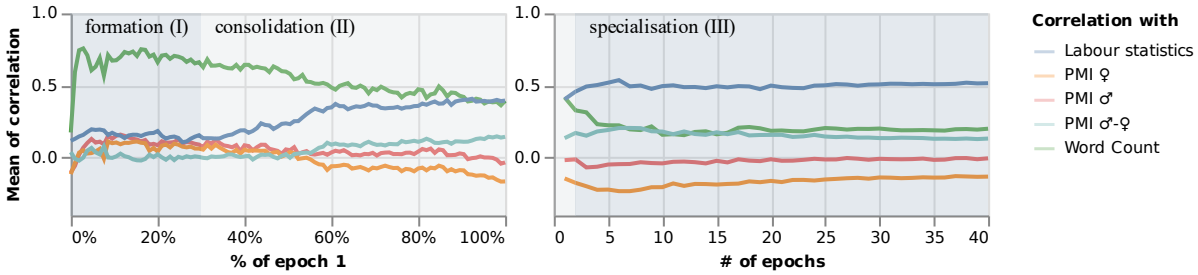


Figure 4: Pearson correlation of the input embedding bias scores for the occupation terms with the different dataset features (word count and PMI) and labour statistics, and how these change during training time.

input embedding bias can be to some extent related to statistics in the dataset, although the language model clearly picks up on many more sources of information on gender association than can be captured by measures like PMI.

Finally, each of the correlations we studied shows interesting dynamics over training time. We see that our measures for input embedding bias and downstream bias grow together during the formation phase, but decouple during the consolidation phase; that word count is dominant in the formation phase, but becomes a progressively less important data statistic in later phases; and that the aggregate PMI measure gives better correlations than separate $PMI_{\male}$ and $PMI_{\female}$ about halfway the first epoch.

5 Diagnostic intervention: changing downstream bias by changing embeddings

So-far, our analyses have all been correlational. In this section, we aim at establishing a causal role for the representations of gender and gender bias that we have described in the previous two sections. We do so by *intervening* on the input embeddings, using the debiasing method **Iterative Null-space Projection** of Ravfogel et al. (2020). In each debiasing step, a gender subspace is identified (as discussed in the previous section), after which all word vectors are projected on its null-space to remove this gender information. The authors show that performing a null-space projection once is not sufficient for removing bias completely. However, repeating this procedure multiple times turns out to be an effective mitigation strategy, without an overall decay of the embeddings (Ravfogel et al., 2020). In our experiments, we denote the number of null-space projections as k.

We apply this method to the input embeddings, and measure the effects on the downstream behaviour, again using the STS-B task. Our goal is not, in the first place, practical (i.e. to end up with an unbiased language model), but rather diagnostic: shedding light on the nature of the representation of gender and gender bias, and the way they influence the behaviour of the model. Ultimately, we hope that our analysis allows us to draw conclusions about the conditions under which debiasing input embeddings (using this particular method)

might be an attractive strategy to mitigate bias in contextual word embeddings.

5.1 Comparing the effect of debiasing across training time

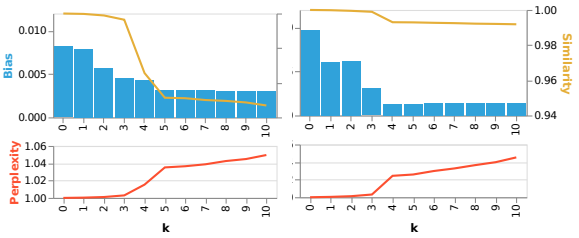


Figure 5: The average STS-B bias, RSA similarity with the original embeddings ($k = 0$), and perplexity values after k debiasing steps for two points in time: epoch 1 (*left*) and 40 (*right*). The perplexity is normalised with respect to the original language model before debiasing. Please note that the starting perplexities are different for epoch 1 and 40.

We perform the Iterative Null-space Projection 10 times on the input embeddings of our language models, at two different points in training: after the first and last epoch. The representation of gender is likely to comprise multiple linear components (Ravfogel et al., 2020), with the most dominant one being the *gender unit* from Section 3. By repeating the debiasing procedure multiple times, we can learn more about this underlying representation as well as explore how these change during training time. The results for this experiment are shown in Figure 5. We measure the downstream bias using STS-B for the original embeddings, as well as for each of the ten iterations of the debiasing algorithm. Moreover, we also measure the quality of the language modelling using the standard perplexity metric. Finally, we measure qualitative changes in the topological organisation of the semantic space of the occupation and gendered words[6], by measuring Representational Similarity (Kriegeskorte et al., 2008) between the original model and the debiased models.

Figure 5 shows a number of important effects. First, we see that there is a visible decrease of the measured bias after debiasing the input embeddings. And, importantly, both perplexity and Representational Similarity show only minor changes up to three debiasing iterations. Performance of the language model starts to diminish at four debiasing steps, and decrease further at five steps and more.

[6]See the word-lists in Appendix C.

These results are in line with our earlier findings about how gender is represented: mostly along the dominant gender unit (Section 3), but with gender information also encoded in the rest of the embedding space, and mostly encoded in such a way that it can be decoded using linear classifiers.

Strikingly, debiasing is much more effective at epoch 40 (the end of training, and the end of the *specialisation* phase), than at epoch 1 (the end of the *consolidation* phase). At epoch 40, the average bias of the model is worse before debiasing, but much better after debiasing, reaching a bias score of just above 0.01. These results agree with our earlier observations that the gender information is encoded more locally during training, which would be easier to remove effectively and selectively. For our specific setup, three debiasing iterations seems to be a sweet-spot, where the perplexity increase is still minimal and the debiasing effect is strong.

5.2 Asymmetry in debiasing female and male bias

To get a more fine-grained picture of how debiasing the input embeddings affects downstream bias, we also consider the effect on the female and male bias separately, as we expect some asymmetry from our earlier observations in Sections 3 and 4. Figure 6 displays the bias scores for a set of female and male biased occupation words for the fully trained language model after k debiasing steps. For this figure, we measure bias both on the input embeddings and in the STS-B task. We can see that a single debiasing step already has a visible effect on both bias measures.

Interestingly, when we consider the input embeddings, we see a strong reduction of the female bias. In contrast, we even observe an increase of the average male bias after one debiasing step. Only after another few steps do we see that both the male and female bias get reduced more significantly. These results are related to our earlier observations in Section 3 about gender asymmetry: the dominant gender unit is used primarily to encode the feminine feature of words, while masculine word information is more distributed over the rest of the input embedding space. We also observe a slight increase in bias after $k > 6$, which we attribute to the bias metric being sensitive to noise in the absence of an actual linear gender representation.[7]

[7]In actual applications of Iterative Null-space Projection this is less of a problem, since you typically stop debiasing if the accuracy of classifier is close to random.

The STS-B bias, however, shows a different behaviour. Debiasing the input embeddings clearly has an effect on the downstream behaviour, but debiasing once has a larger effect on the *male bias* instead. It takes more than two debiasing steps before both the female and male bias is reduced. Interestingly, we found earlier that three debiasing iterations is a sweet spot, but we have no satisfying explanation for why especially the male bias is reduced in the first iteration.

5.3 Summary

We conclude that there is a causal effect of the gender representation in the input embeddings on the downstream bias. First, we find that the Iterative Null-space Projection is surprisingly effective and that three debiasing steps result in a bias reduction with minimal harm to the perplexity of the language model and topological representation of the embedding space. This reflects our earlier finding that gender information is encoded very locally, but also suggests that the model relies a lot on this *linearly decodable* gender representation. Secondly, we find that observing the effect on male and female biased occupation terms separately shows an asymmetry for both the input embedding and STS-B bias. While the asymmetry towards female bias in the input embeddings can be explained by our earlier observations in Section 3, we are not sure why removing this information affects especially the male bias in the contextual embeddings. More work is needed to explore possible explanations for this incongruity as it can have important consequences for certain types of mitigation strategies.

6 Discussion

Although in our experiments we have restricted ourselves to gender bias in English, we believe our results have relevance for the broader study of bias in language models. More concretely, this paper contributes to the ongoing research on bias in language models in three ways: we shed light on the question of how the internal representation of the model relates to its downstream bias, we show that studying the dynamic nature of bias can be illuminating, and we point out that there are potential asymmetries in the underlying bias representation that researchers should be aware of.

6.1 Relationship internal representations and downstream bias

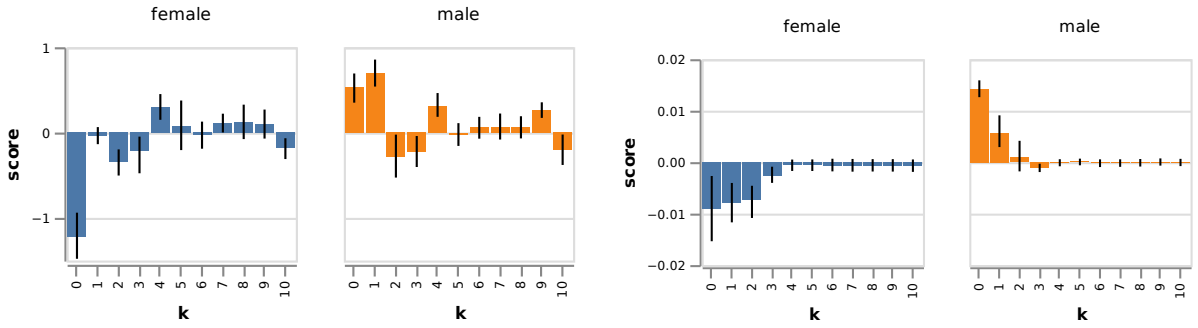
When deciding on a bias mitigation strategy, it is crucial to understand the relation between the internal representations of the language model and the bias in downstream tasks. This is because successful debiasing of the internal representations will likely generalise over many downstream tasks, but this strategy is only viable if the internal representation that is manipulated is causally linked to the downstream behaviour of the model. Whether this causal connection exists, however, might differ from case to case.

In a study looking at static word embeddings (not language models, as we do here) and a number of different downstream tasks, Goldfarb-Tarrant et al. (2021) find no correlation between the bias in the embeddings and in the downstream tasks. In constrast, Ravfogel et al. (2020) find that debiasing embeddings can be effective in reducing racial bias in a sentiment classifier. More closely related to our setup of investigating gender bias in a language model, Vig et al. (2020a) and De Cao et al. (2021) actually show that gender information can be stored in or mediated by a small part of a language model by selectively changing neuron activations and analysing the effect on the output. De Cao et al. even study the parameters of the same LSTM architecture that we consider in this paper.

In line with these findings, we also observe that gender information is represented very locally in the input embeddings of the LSTM language model. Furthermore, we find that manipulating bias in the input embeddings does indeed affect downstream bias, adding evidence for a causal relation between this particular level of internal representation and downstream behaviour of the model. However, we should add here that whether one finds such a connection might strongly depend on the choice of bias metric, model architecture, and downstream task, and furthermore depends on the particular learning phase a language model is in with respect to a particular type of bias.

6.2 Different phases in the evolution of gender

Based on our finding we can distinguish three phases in the evolution of gender representation and gender bias: (i) formation, (ii) consolidation, and (iii) specialisation. We saw that our measures for bias and the method for bias mitigation behave differently in these different phases, which appears



(a) Input embedding bias

(b) STS-B bias

Figure 6: Effect of debiasing the language model at epoch 40 on the bias scores for a list of male- and female biased occupations. Based on the bias scores for IE and STS-B bias, we chose "receptionist", "nurse", "librarian", and "therapist" for the female words and "mechanic", "engineer", "scientist", and "architect" for the male word-list.

to be connected to how locally gender information is represented in the internal representations of the model. Only if the relevant information is concentrated in a particular part of the model and linearly decodable, can we reliably and selectively remove gender information without hurting the overall language model performance.

This observation might not be so important when thinking about gender bias in current large language models, as the sheer scale of the datasets that these models are trained on and the high frequency of gendered words makes it very likely that they have progressed far into the 'specialisation phase' with respect to their representation of gender. However, it could matter when considering other types of biases, where the words and phrases driving the birth of these biases may be much less frequent. Hence, even in large language models trained on several orders of magnitude more data than the language model we used in this study, the relevant representations for other biases might very well still be in something equivalent to our 'formation' or 'consolidation phase'. Indeed, work on studying the effect of fine-tuning has shown that the manifestation of bias can still change significantly in pre-trained models (Choenni et al., 2021; Webster et al., 2020).

6.3 Asymmetry in the gender representation

Gender asymmetries are regularly observed in word frequencies and co-occurrences in datasets (e.g. Zhao et al., 2019; Tan and Celis, 2019; Wagner et al., 2016) and in language use in general (e.g. the "male-as-norm bias", Danesi, 2014). Interestingly, we also observed a strong asymmetry in how gender bias is represented in the input embeddings, but we did not see the same asymmetry in the downstream task. This could have consequences for how mitigation strategies should be evaluated. When debiasing the model while being unaware of the underlying representation, one could disproportionately harm one group more than another. This could lead to the introduction of a new form of bias. In developing and evaluating mitigation strategies, it is therefore important to do a thorough analysis of the representation of bias present in the NLP system and how certain social groups could be affected disproportionally if not accounted for.

7 Conclusion

While there is a lot of important work on detecting and mitigating undesirable biases in language models, we still lack a good understanding of the mechanisms underlying the biased behaviour. The goal of this study was to take a step back and analyse the birth of bias in language models. To this end, we present a temporal investigation of how an English LSTM language model learns a representation of gender in the input embeddings and how this affects downstream biased behaviour.

There are many interesting directions for future research. An important open question is, for instance, how intrinsic representations of bias relate to other downstream tasks that may be closer to real-world systems where the representational and allocative harms to social groups are more clear (Blodgett et al., 2020). For future work, we also plan to do further investigations on how our training dynamics analysis may generalise to other undesirable social biases, model architectures, training corpora, and downstream tasks, as well as other possible representations in the internal states of the

language model that are useful for understanding bias. Furthermore, the robustness of our results with respect to different random initialisations of the language model should be checked (Webster et al., 2020; D'Amour et al., 2020).

In this paper, we take a step towards a more thorough understanding of the evolution of bias in language models across the different stages of the language modelling pipeline. Hopefully, it will inspire more work on the dynamic behaviour of language models, with respect to bias, but also other still poorly understood features of these models.

Acknowledgements

This publication is part of the project "The biased reality of online media - Using stereotypes to make media manipulation visible" (with project number 406.DI.19.059) of the research programme Open Competition Digitalisation-SSH, which is financed by the Dutch Research Council (NWO).

We want to thank Maartje ter Hoeve for her feedback on an earlier version of the paper.

References

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*. ArXiv: 2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

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

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1477–1491. Association for Computational Linguistics.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina

Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv preprint arXiv:2011.03395*.

Marcel Danesi. 2014. *Dictionary of media and communications*. Routledge.

Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. 2021. Sparse interventions in language models with differentiable masking. *arXiv preprint arXiv:2112.06837*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73. Association for Computing Machinery.

Yupei Du, Qixiang Fang, and Dong Nguyen 0002. 2021. Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10012–10034. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705. Association for Computational Linguistics.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Tenth International AAAI Conference on Web and Social Media*, pages 112–120. AAAI Press.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1926–1940. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. 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)*, pages 5267–5275. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jaap Jumelet, Willem H. Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 1–11. Association for Computational Linguistics.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172. Association for Computational Linguistics.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender Bias in Neural Natural Language Processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, Lecture Notes in Computer Science, pages 189–202. Springer International Publishing.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

Christina Richards, Walter Pierre Bouman, Leighton Seal, Meg John Barker, Timo O Nieder, and Guy T'Sjoen. 2016. Non-binary or genderqueer genders. *International Review of Psychiatry*, 28(1):95–102.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

Samson Tan, Shafiq R. Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4153–4169. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020a. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv:2004.12265 [cs]*. ArXiv: 2004.12265.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020b. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):1–24.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5443–5453. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and Reducing Gendered Correlations in Pre-trained Models.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

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

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853. Association for Computational Linguistics.

A Labour statistics

In this work, we use the US Bureau of Labor statistics on the percentage of female workers (Caliskan et al., 2017) for comparison with the gender bias in the language modelling pipeline (see Table 1).[8] Please note that the ordering of this list can be reversed when computing correlations.

B Dataset statistics

In Section 4, we score the occupation words with various dataset features and rank these with the labour statistics. The results can be found in Table 2.

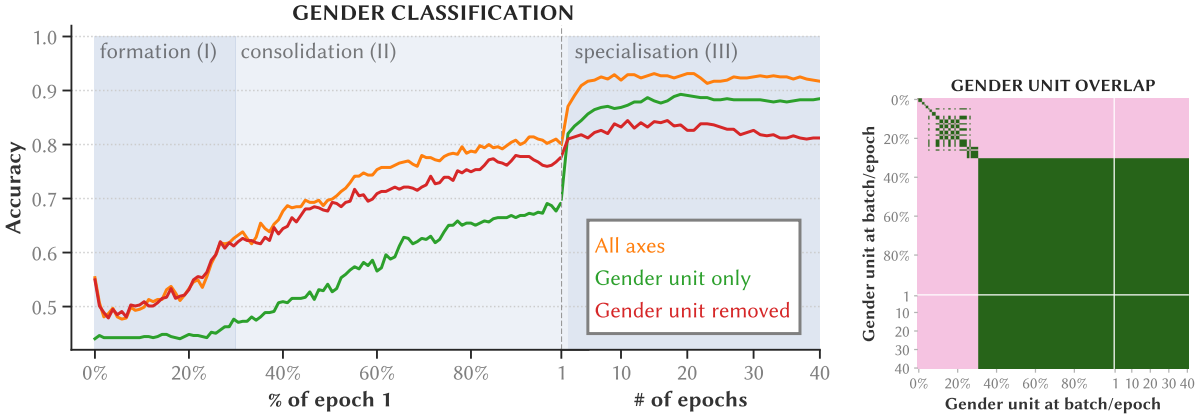
C Wordlists

We use two sets of word-lists in the experiments of Sections 3, 4, and 5. The first word-list used in Section 3, contains a list of 82 gendered word-pairs (also considering capitalised and pluralised versions in the model vocabulary), as shown in Table 4. Then, in Sections 4 and 5, we use a subset of the previous gendered word-pairs that is more similar to what is used in previous work for finding a *gender subspace* (e.g. Bolukbasi et al., 2016; Ethayarajh et al., 2019; Ravfogel et al., 2020) and can be found in Table C. This last table also contains a list of 54 occupation words for studying gender bias, which corresponds to the list in Table 1. We have indicated the overlap between the two word-lists in bold for reference.

D Extra figures

Figures 7 and 8 support Section 3, while we refer to Figure 9 in Section 4.

[8] https://github.com/rudinger/winogender-schemas/blob/master/data/occupations-stats.tsv



(a) (b)

Figure 7: Classification accuracy of gender using three different classifiers, that use only the dominant gender unit (green), all other units (red), or all units (orange). Curves show results over training time, averaged across seeds. Gender unit overlap (*right*) shows the equality of the principal gender units across time, with green indicating units being equal.

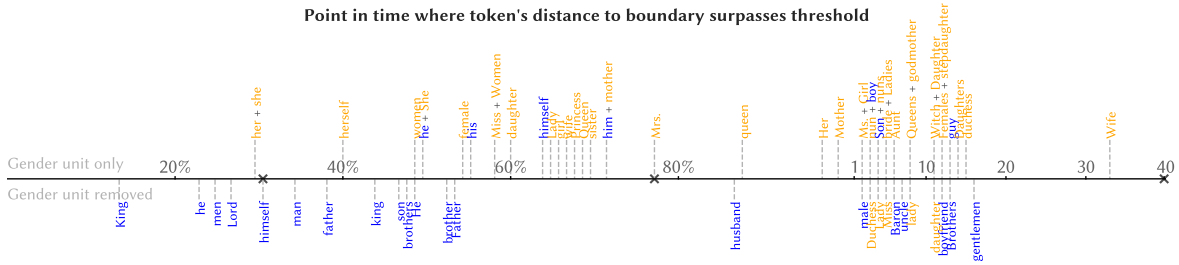


Figure 8: Point in time where a token's distance to the gender decision boundary surpassed a threshold. Tokens at the top are based on the single gender unit classifiers; tokens at the bottom are bottom are based on the classifier containing all but the gender unit.

Figure 9: Average absolute bias scores over the occupation terms for the input embeddings and downstream STS-B task.

occupation	% female	occupation	% female
pathologist	97.50	scientist	41.94
secretary	94.60	specialist	41.35
hairdresser	94.20	technician	40.34
receptionist	90.60	supervisor	38.64
nurse	89.58	manager	38.51
librarian	83.00	worker	37.92
planner	77.60	doctor	37.90
therapist	76.70	advisor	37.90
practitioner	74.79	physician	37.90
cashier	72.50	surgeon	37.90
teacher	71.00	chemist	36.10
educator	70.80	lawyer	34.50
psychologist	70.30	janitor	34.30
clerk	69.53	paramedic	32.90
counselor	66.48	officer	30.42
examiner	62.46	architect	20.81
instructor	62.30	chef	19.60
baker	60.80	programmer	18.35
veterinarian	60.50	engineer	10.72
bartender	59.80	machinist	6.70
auditor	59.70	inspector	6.40
accountant	59.70	painter	5.70
pharmacist	57.00	firefighter	3.50
dispatcher	56.30	electrician	2.30
broker	55.50	carpenter	2.07
administrator	54.86	mechanic	1.80
investigator	45.15	plumber	0.70

Table 1: US Labour Statistics with the percentage of female workers for the occupations we consider in gender bias analysis of the LSTM language model. Any differences with previous work are due to some occupations being left out, as these do not occur in the model vocabulary.

Dataset feature	Pearson correlation
Word Count	0.108785
$\text{PMI}_{♂}$	0.116814
$\text{PMI}_{♀}$	-0.234644
$\text{PMI}_{♂-♀}$	0.333687

Table 2: Pearson correlation with labour statistics for occupation words.

Type	Words
Male	man, boy, he, father, son, male, his, himself, John
Female	woman, girl, she, mother, daughter, female, her, herself, Mary
Occupations	technician, accountant, supervisor, engineer, worker, educator, clerk, counselor, inspector, mechanic, manager, therapist, administrator, receptionist, librarian, advisor, pharmacist, janitor, psychologist, physician, carpenter, nurse, investigator, bartender, specialist, electrician, officer, pathologist, teacher, lawyer, planner, practitioner, plumber, instructor, surgeon, veterinarian, paramedic, examiner, chemist, machinist, architect, hairdresser, baker, programmer, scientist, dispatcher, cashier, auditor, painter, broker, chef, doctor, firefighter, secretary

Table 3: Word-lists considered for finding the *gender subspace* in the input embeddings of the language model. This subset of gendered words is also used for finding the PMI associations. The occupation words are the same as in Table 1.

Male	Female
man	**woman**
boy	**girl**
guy	gal
gentleman	lady
lord	lady
Mister	Miss
Mr.	Ms.
Mr.	Mrs.
male	**female**
masculine	feminine

Male	Female
king	queen
prince	princess
Baron	Baroness
duke	duchess
monk	nun
wizard	witch
landlord	landlady

Male	Female
he	**she**
him	her
himself	**herself**
his	**her**
his	hers

Male	Female
father	**mother**
dad	mum
brother	sister
nephew	niece
uncle	aunt
grandfather	grandmother
son	**daughter**
grandson	granddaughter
son-in-law	daughter-in-law
stepfather	stepmother
stepson	stepdaughter
father-in-law	mother-in-law
bridegroom	bride
groom	bride
husband	wife
boyfriend	girlfriend
godfather	godmother

Table 4: List of gendered word-pairs used in Section 3. The word-pairs in bold are also used for finding the gender subspace and computing the PMI associations in Section 4. We enrich this list by also incorporating the capitalised and pluralised versions of the pairs that are present in the model vocabulary.