Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models

Daniel de Vassimon Manela[†]   **David Errington**[†]   **Thomas Fisher**[†]
Boris van Breugel[†]   **Pasquale Minervini**
University College London
{daniel.manela.19,david.errington.19,thomas.fisher.19,
boris.breugel.19,p.minervini}@ucl.ac.uk

Abstract

This paper proposes two intuitive metrics, *skew* and *stereotype*, that quantify and analyse the gender bias present in contextual language models when tackling the WinoBias pronoun resolution task. We find evidence that gender stereotype correlates approximately negatively with gender skew in out-of-the-box models, suggesting that there is a trade-off between these two forms of bias. We investigate two methods to mitigate bias. The first approach is an online method which is effective at removing skew at the expense of stereotype. The second, inspired by previous work on ELMo, involves the fine-tuning of BERT using an augmented gender-balanced dataset. We show that this reduces both skew and stereotype relative to its unaugmented fine-tuned counterpart. However, we find that existing gender bias benchmarks do not fully probe professional bias as pronoun resolution may be obfuscated by cross-correlations from other manifestations of gender prejudice. Our code is available online.

1 Introduction

Transformer-Based Transfer Learning models for NLP – referred to henceforth as TBTL models for brevity – such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020) perform well on a variety of NLP tasks with minimal fine-tuning. However, prior to fine-tuning, TBTL models require a vast amount of data to train (Shoeybi et al., 2019). This training is only performed once, with users downloading and fine-tuning such language models to their specific task. In doing so, we are trusting large tech companies to train the base model responsibly since we have no control over this. This seems inherently undemocratic. We ideally want these models to be free from unwanted bias and whilst it is true that they exhibit less gender bias than static word embeddings (Sun et al., 2019), they are by no means immune to this problem (Lu et al., 2018).

As TBTL models become increasingly prevalent in our everyday lives, we want to avoid such prejudices influencing decision making. Examples where this is important include automatic resume filtering (Dastin, 2018) and criminal sentencing recommendations (Tashea, 2017).

In this paper we focus on the specific problem of gender bias, and analyse the extent to which it persists in modern TBTL models. We build upon Zhao et al. (2018a, 2019), in which quantification and mitigation of bias in ELMo was centre stage. In addressing this problem for more recent models, we aim to answer three main questions: i) How can we quantify bias in pre-trained language models? ii) How do different models compare in terms of bias? iii) How to mitigate bias in these models?

We believe that current gender bias metrics in the existing literature do not offer sufficient granularity to properly analyse this problem. Indeed, they mostly focus on measuring the assignment of stereotypical pronouns to professions (Zhao et al., 2018a). By focusing solely on this, they fail to address a model's overall preference for predicting male pronouns. An alternative bias which models can demonstrate is unequal preference towards male and female pronoun resolution across stereotypical and anti-stereotypical professions. We refer to these two forms of bias as *skew* and *stereotype*, respectively. In Section 3, we propose a new scheme to capture and quantify the important distinction between the two.

When comparing different TBTL models, we find evidence that gender skew and gender stereotype correlate approximately negatively with each other in out-of-the-box models, suggesting that a tradeoff between these two forms of bias may exist.

[†]Equal contribution.

2232

To mitigate bias in these models, we use the method proposed by Zhao et al. (2018a) to show that fine-tuning with augmented data, which references male and female entities with equal frequencies, can reduce professional gender stereotype and skew compared to fine-tuning on the original dataset. However, we show that gender prejudice may persist in forms other than professional bias, and these are ineffectively probed by current NLP benchmarks.

2 Related work

Bias quantification Early work in measuring gender bias specifically (Caliskan et al., 2017; May et al., 2019), along with efforts towards removing it either during (Zhao et al., 2018b) or after training (Bolukbasi et al., 2016), was done on static word embeddings such as GloVe and Word2Vec.

Caliskan et al. (2017) argue that completely removing undesirable bias using an automated procedure is impossible, as it is only distinguishable from the rules and structure of language itself by negative consequences in downstream applications. Instead, we should focus on probing and exposing which biases manifest themselves in which models, so that engineers can act accordingly. Choosing a suitable metric with which to analyse bias is a key challenge (May et al., 2019); whilst a positive result with respect to a suitable metric does reveal the existence of bias, a negative result does not mean a model is completely bias-free.

Statistical tests such as Word/Sentence Embedding Association Tests (WEAT/SEAT) have been developed to measure bias in static word embeddings using the cosine similarity of specific target words (Caliskan et al., 2017; May et al., 2019). However, when it comes to contextual embeddings, these traditional metrics have been shown to be ineffective at quantifying bias. In particular, Kurita et al. (2019) demonstrate that while WEAT tests are unable to identify any statistically significant bias in BERT, probing the underlying language model with a Gender Pronoun Resolution (GPR) task does reveal strong evidence that these non-static models also encode gender bias. Indeed, since contextual word embedding models such as BERT are optimised to capture the statistical properties of training data, they tend to pick up and amplify any social stereotypes that may be present (Kurita et al., 2019).

Having established GPR as a downstream task

suitable for detecting gender bias, Zhao et al. (2018a) introduced a new benchmark, WinoBias, to measure bias in coreference pronoun resolution.

The dataset consists of two files, Test Set 1 and Test Set 2 (hereafter T1 and T2), representing two different gender pronoun resolution tasks. Each file consists of Wino-grad schema pairs of sentences involving a variety of occupations, differing only in one or two words and with a pronoun ambiguity that is resolved in opposite directions across the two sentences, giving both a pro- and anti-stereotypical resolution (Levesque et al., 2012). Example sentences are shown in Fig. 1 and Fig. 2.

Sun et al. (2019) consider a coreference resolution system unbiased on the WinoBias test if it achieves similar F1 scores for gender pronoun resolution on both the pro- and anti-stereotypical datasets whilst maintaining strong GPR performance. One of the main findings in Zhao et al. (2018a) is that three different coreference resolution architectures (rule based, feature-rich and neural-net based) built on top of static word embeddings all display significant disparity in F1 scores across the two datasets, with the F1 score for the pro-stereotypical dataset being on average 21.1 higher. This alarming observation was also discovered by Webster et al. (2018) and was attributed to the inherent bias of the underlying word embeddings (Bolukbasi et al., 2016), as well as the training of these coreference resolution pipelines on the OntoNotes 5.0 dataset (Weischedel et al., 2011) which is known to suffer from severe gender imbalance (Zhao et al., 2018a).

More recently, Zhao et al. (2019) investigated the existence of gender bias in the ELMo contextual embedding. Specifically, they note ELMo is trained on the Billion Word corpus (Chelba et al., 2013) which, just like OntoNotes 5.0, shows substantial imbalance in the counts of male vs. female pronouns. Training on this, ELMo then learns a language representation that reflects this gender inequality. To expose this, Zhao et al. (2019) analyse the behaviour of a coreference resolution system proposed by Lee et al. (2018) with ELMo contextual weights on the WinoBias benchmark, revealing a significant disparity in performance on the pro- and anti-stereotypical datasets. In fact, this disparity is 30% higher than a similar result based only on GloVe embeddings (Lee et al., 2017). This is particularly worrying; as commented earlier, contextual embeddings may, by construction, be amplifying
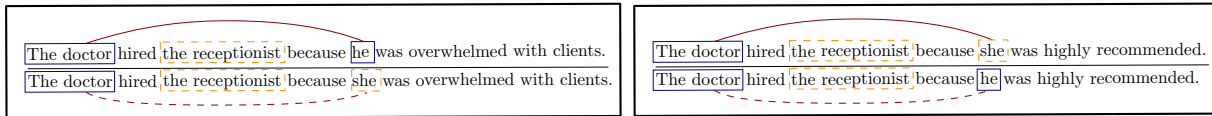
Figure 1: Example sentences from Test Set 1 (T1) of the WinoBias dataset. These take the form *[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]*. Solid blue boxes indicate male entities, dashed orange boxes indicate female entities, solid purple lines indicate pro-stereotypical scenarios and dashed purple lines indicate anti-stereotypical scenarios. Such stereotypes are determined according to U.S. Bureau of Labor Statistics (2017).
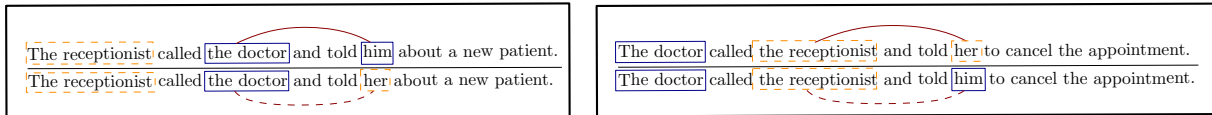

Figure 2: Example sentences from Test Set 2 (T2) of the WinoBias dataset. These take the form *[entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances]*. The style format follows that of Fig. 1. Image adapted from Zhao et al. (2018a).

undesirable statistical artefacts of the dataset more than their static counterparts. Therefore, it is of the utmost importance to perform a similar analysis on recent TBTL models.

Devlin et al. (2018) explicitly state that BERT is not trained on the Billion Word corpus, since this only provides examples of isolated sentences and the authors preferred to use a document-level corpus to get contiguous training data, allowing richer contexts to be learnt.

Specifically, BERT is trained using the Book-Corpus dataset (Zhu et al., 2015) as well as English Wikipedia. However, the BookCorpus data has since been shown to suffer from similar gender imbalance problems (Tan and Celis, 2019) as has English Wikipedia where, for example, only 15.5% of the biographies are of women (Wagner et al., 2016). We believe that this imbalance is the principle cause of skew in the model.

Bias Mitigation As discussed above, whilst we cannot completely remove bias from a model, research into bias mitigation is still a very worthwhile pursuit and, in the context of the WinoBias metric of occupational gender bias, could help break the glass ceiling.

Many bias mitigation methods for static embeddings centre around modifying the vector space and/or loss function during the training process. Initial attempts sought to project biased embedding vectors back to a gender neutral subspace (Bolukbasi et al., 2016). Subsequent improvements came from adding a regularisation term to the training loss function designed to encourage specific gendered words to separate, thus allowing the remain-

ing neutral terms to mix (Zhao et al., 2018b). However, these offer superficial reductions in gender bias, and systematic prejudice was found to persist (Gonen and Goldberg, 2019).

Attempts to mitigate gender bias in contextualised embeddings are a more novel endeavour. These attempts typically involve fine-tuning models to a particular task and one proposal involves duplicating the training corpus and switching gender-specific terms in the duplicated data. For example, "The King cemented his rule over his lords" is substituted with "The Queen cemented her rule over her ladies". This method, referred to as Data Augmentation, was demonstrated to successfully reduce gender bias in ELMo for pronoun resolution tasks, relative to a model trained on the unaugmented training data (Zhao et al., 2019).

3 Method

3.1 Analysing Bias in WinoBias

Bias in TBTL models can be measured using either T1 or T2 from the WinoBias dataset – see Fig. 1 and Fig. 2, respectively. Within each test set, examples are composed of a pro-stereotypical and anti-stereotypical sentence, where stereotype is determined by professional gender imbalances recorded by U.S. Bureau of Labor Statistics (2017). Our approach is to take each WinoBias sentence, mask the pronoun of interest, and then compare the language model's prediction for the masked token with the pro- and anti-stereotypical labels. To predict the gender of the pronoun in the sentence "*The physician hired the secretary because [MASK] was overwhelmed with clients*" we calculate the proba-
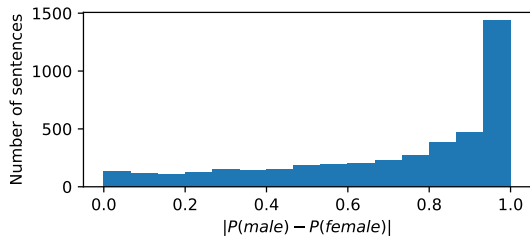
Figure 3: Histogram of absolute differences in gendered pronoun assignment, $|P(\text{male}) - P(\text{female})|$. Sentence examples where this difference was smaller than 0.1 were removed so that only pronouns assigned with a high degree of certainty were analysed.

bilities of the pro- and anti-stereotypical pronouns – "he" and "she" respectively – and pick the one with the highest likelihood. Note that this approach risks obscuring the confidence in pronoun resolution. For example, $P(\text{male}) = 0.99$ and $P(\text{male}) = 0.51$ would both result in male pronoun assignment. The histogram in Fig. 3 demonstrates that this issue does not affect our experiments; the distribution is highly negatively skewed. The majority of pronouns are resolved with a high degree of confidence. We chose $|P(\text{male}) - P(\text{female})| \geq 0.1$ as an arbitrary cutoff bound to select sentences which were resolved with a high degree of certainty. Only sentences which fulfil this criteria were analysed in the following experiments.

In line with the academic literature, we compute F1 scores for both the pro- and anti-stereotypical data using contextual language models. This approach to coreference resolution is demonstrably well-founded and F1 results from the GPR baseline are discussed in Section 4.1.

The WinoBias sentences have been constructed so that, in the absence of professional stereotypes, there is no objective way to choose between different gender pronouns. The difference in F1 scores with respect to gender g, across a pro/anti test set, $F1_{\text{pro}}^{g} - F1_{\text{anti}}^{g}$, is a metric inspired by previous papers to measure a model's tendency to assign that gender to professions, with positive (resp. negative) values indicating a pro- (resp. anti-) stereotypical assignment (Sun et al., 2019). We refer to it as a measure of gender *stereotype*. In contrast to the literature, we compute F1 scores with respect to both "male" and "female" true labels allowing us to define *stereotype* with respect to both genders.

We now propose to also use the difference in F1 scores with respect to a dataset $\mathcal{D}$, across gen-

ders, $F1_{\mathcal{D}}^{\male} - F1_{\mathcal{D}}^{\female}$, as a measure of gender *skew* in dataset $\mathcal{D}$, with positive (resp. negative) values capturing the tendency of a model to generally assign a male (resp. female) gender to any given profession. This distinction is important: consider a classifier which only assigns male pronouns to professions. It would not be stereotyping professions to perceived gender roles, but would be heavily biased in assuming a general male dominance in the workplace. Both these forms of gender unfairness are considered in the subsequent analysis and we use the mean skew and stereotype, taken across datasets and genders respectively as shown below:

$$\mu_{\text{Skew}} \triangleq \frac{1}{2}\left(\left|F1_{\text{pro}}^{\male} - F1_{\text{pro}}^{\female}\right| + \left|F1_{\text{anti}}^{\male} - F1_{\text{anti}}^{\female}\right|\right)$$

$$\mu_{\text{Stereo}} \triangleq \frac{1}{2}\left(\left|F1_{\text{pro}}^{\male} - F1_{\text{anti}}^{\male}\right| + \left|F1_{\text{pro}}^{\female} - F1_{\text{anti}}^{\female}\right|\right)$$

where $F1_{\text{pro}}^{\male}$ denotes the F1 score on the pro-stereotypical dataset whilst considering the male pronoun as the *true* label. To be completely gender neutral, we average the absolute values since we are only interested in the extent of gender bias rather than its direction.

3.2 Online Skewness Mitigation

As we will show in Section 4.2, most current TBTL models models are inherently skewed towards predicting male pronouns. Inspired by Kurita et al. (2019), we propose a simple approach to reducing this skew. We normalise the probability of a masked pronoun being assigned a particular gender in a certain occupational context by dividing through with the prior probability of choosing that pronoun in a sentence with the same structure but without any occupational context.

We illustrate this method with the sentence "The physician hired the secretary because he was overwhelmed with clients". This method starts by calculating the probabilities of "he" and "she" in the standard way, as described in Section 3.1. Next, we mask the professions, leading to "[MASK] hired [MASK] because [MASK] was overwhelmed with clients" and calculate the probability of the third masked word being "he" and "she" in this context. Finally, we normalise by dividing the probabilities found using the standard method, with the probabilities found using the masked-professions context. This method assumes language models can resolve the pronoun when both professions are masked.

Models mitigating skew using this approach are given the suffix -O in the remainder of this paper.

3.3 Bias Removal via Data Augmentation

We aim to replicate the Data Augmentation method proposed in Zhao et al. (2019) for mitigating gender bias in ELMo. The goal of this approach is to use an augmented dataset to fine-tune the pre-trained language model to the GPR task. In particular, this augmented dataset is designed with the intention of neutralising the gender bias already present in a model such as BERT, whilst simultaneously avoiding the corruption of its understanding of natural language.

As in Zhao et al. (2019), a target GPR task was constructed by first selecting sentences from OntoNotes 5.0 containing gendered pronouns and masking them accordingly; the BERT masked language model will be trained to predict the masked pronoun. Secondly, we anonymise the data by replacing all gendered names with identity tokens such as [E1] and [E2].

Each training example is then *augmented* by replacing all possessive and personal pronouns with those of the opposite gender. Additionally we apply a mapping of explicitly gendered words (such as "Man"→"Woman" and vice-versa) to ensure that the text remains linguistically coherent in the context of reversed genders.[1] Following this approach, the sentence "*The King was pleased that his Lords had vanquished their enemies*" would be augmented to "*The Queen was pleased that her Ladies had vanquished their enemies*".

To examine the effects of data augmentation, we then fine-tune two BERT models. The first was fine-tuned on the un-augmented OntoNotes training examples, whilst the second was fine-tuned on the augmented OntoNotes examples (containing the duplicated and gender-switched examples also). Hereafter we shall refer to these models as BERT-U and BERT-A respectively. In both cases, a hyperparameter search over the epochs and learning rate was conducted.[2] The best performing un-augmented/augmented models were tested using the WinoBias data as described in Section 3.1.

4 Results and Discussion

4.1 Baseline: Alice and Bob

There is a risk that removing bias deteriorates the predictive power of the model. We measure a baseline performance on a GPR task to test how well

Model	T1		T2	
	F1$^\male$	**F1$^\female$**	**F1$^\male$**	**F1$^\female$**
RoBERTa	64.2	72.2	92.7	93.0
RoBERTa-O	50.2	71.7	89.8	88.7
RoBERTa-large	78.2	79.5	94.9	94.9
RoBERTa-large-O	78.2	79.5	89.5	87.3
ALBERT	39.2	68.6	58.7	75.3
ALBERT-O	6.2	67.6	62.9	24.1
ALBERT-large	61.0	71.2	19.2	68.6
ALBERT-large-O	60.6	68.9	31.2	70.1
ALBERT-xlarge	64.1	75.1	23.3	69.4
ALBERT-xlarge-O	69.2	74.2	59.4	76.6
ALBERT-xxlarge	64.2	76.9	95.0	95.3
ALBERT-xxlarge-O	78.2	80.5	89.0	90.5
BERT	58.8	62.4	95.3	95.5
BERT-O	59.0	64.7	95.1	95.1
BERT-large	72.6	74.9	95.3	95.5
BERT-large-O	69.8	75.1	95.6	95.7
XLM-RoBERTa	29.7	69.0	64.5	76.7
XLM-RoBERTa-O	52.1	65.1	64.4	37.9
XLM-RoBERTa-large	62.8	76.4	21.7	69.2
XLM-RoBERTa-large-O	64.9	77.0	80.1	84.9
DistilBERT	41.2	66.4	81.0	79.0
DistilBERT-O	50.2	66.1	81.0	78.5
BERT-U	76.6	65.8	90.1	88.6
BERT-UO	78.8	77.3	93.1	92.3
BERT-A	75.6	67.4	91.8	90.0
BERT-AO	75.7	67.5	74.9	54.2

Table 1: F1 (in %) performance of different models on WinoBias dataset, where professions are replaced by gendered names and a pronoun is correct if it refers to the correct name. The insertion of gendered names implies that there is now a correct pronoun, in contrast to the original WinoBias data set where there is merely a stereotypical pronoun. The suffixes O, U and A refer to Online, Un-augmented and Augmented mitigation approaches respectively.

the model is able to actually resolve the pronoun to the correct entity. To assess this we modify the WinoBias data set by replacing the professions with unambiguously gendered names, *Alice* and *Bob*. Section 4.1 illustrates that we can achieve high F1 scores on this modified WinoBias dataset, validating the use of masked language models for GPR tasks. However, we note that ALBERT and XLM-RoBERTa perform particularly poorly on both T1 and T2 tasks.

The Online Skewness Mitigation described in Section 3.2 demonstrates no discernable pattern on the F1 scores. This suggests that it does not negatively effect GPR performance. Neither is there

---

[1]Mapping sourced from (Zhao et al., 2018a). See the Glove and WinoBias GitHub pages.

[2]Specific settings detailed in Appendix A.

a definite pattern in the skew, indicating that skew is not necessarily reduced in the presence of unambiguously gendered entities.

The decreased performance of BERT-A/U (as described in Section 3.3) relative to the out-of-the-box BERT model may be caused by the anonymisation of the OntoNotes data the models were fine-tuned on, making them less receptive to performing GPR with common names.

The F1 scores in Section 4.1 demonstrate that GPR in T1 is significantly more challenging than in T2. Figures 1 and 2 show that pronoun resolution in T2 is always with respect to *[entity2]* whilst in T1, the pronoun resolution can be with respect to either *[entity1]* or *[entity2]*. In T2 it is clearer to contextual models which entity is the object of the sentence. Thus, we can use a model's gendered pronoun predictions on T2 sentences to expose any internal bias it may have toward *[entity2]*. In T1, the lack of syntactic cues make it unclear which entity is the sentence's object; as such, we may be unable to isolate the model bias corresponding to each specific entity.[3] For these reasons, we argue that T2 is better at revealing the biases encoded in these models. *Hence we will use T2 from this point onward.*

To expose bias, we require models with reliable coreference resolution performance on T2, demonstrated by consistently high F1 scores. In subsequent sections, we investigate all BERT, RoBERTa, and DistilBERT models as well as ALBERT-xxlarge and XLM-RoBERTa-large-O, all of which have F1 scores on T2 greater than our arbitrarily defined threshold of 75%. All of the other models in Section 4.1 fall below this threshold and so are not considered further in this paper.

4.2 WinoBias Performance

In Section 4.2 we present the F1 scores achieved on WinoBias T2 by different models. We note that the pro-stereotypical F1 scores are lower than the gendered names baseline of Section 4.1. This is to be expected, since the Alice and Bob system discussed in Section 4.1 can be understood as being unambiguous and completely biased. Consequently, whereas Section 4.1 shows just GPR performance and general skew bias, Section 4.2 quantifies GPR performance, skew and stereotype.

---

[3]This also explains the larger skew on T1 compared to T2: the lower GPR performance on T1 leads the model to guess more and as we will see in Section 4.2, prefers to guess male.

| Embedding | T2 | | | | | |
| | $F1^{\male}$ | | $F1^{\female}$ | | Bias | |
	Pro	Anti	Pro	Anti	Stereo	Skew
RoBERTa	62.9	27.0	69.0	39.3	32.8	9.2
RoBERTa-O	68.0	60.2	26.5	8.5	12.9	46.6
RoBERTa-large	67.0	52.4	45.0	21.5	26.4	24.0
RoBERTa-large-O	66.0	65.0	11.2	7.5	2.3	56.1
ALBERT-xxlarge	71.4	46.9	54.8	16.4	31.4	23.5
ALBERT-xxlarge-O	69.7	49.2	51.5	18.1	27.0	24.6
DistilBERT	64.9	67.2	4.8	5.0	1.3	61.2
DistilBERT-O	64.5	65.8	10.0	12.8	2.1	53.8
BERT	69.3	58.0	31.4	8.2	17.3	43.8
BERT-O	68.4	58.1	32.9	10.5	16.4	41.6
BERT-large	70.0	57.9	33.9	2.8	21.6	45.6
BERT-large-O	69.9	57.9	32.5	5.0	19.7	45.1
XLM-RoBERTa-large-O	68.0	56.0	38.2	15.7	17.2	35.1
BERT-U	67.8	63.4	14.1	2.8	57.2	7.9
BERT-UO	67.3	60.4	26.6	11.6	44.8	11.0
BERT-A	64.7	64.5	14.8	14.8	49.8	0.1
BERT-AO	65.0	63.6	17.9	15.3	47.7	2.0

Table 2: F1 results (in %) from Test Set 2. The suffixes O, U and A refer to Online, Unaugmented and Augmented bias mitigation approaches respectively.

Note that all models show significant male skew, except for RoBERTa which demonstrates higher $F1^{\female}$ than $F1^{\male}$ scores on both pro- and anti-stereotypical examples. Indeed for all other models there is a noticeable increase in male skew compared to the Alice & Bob results in Section 4.1. The only experimental difference is the use of occupations rather than names, demonstrating that it is specifically the professions that push the model to predicting male pronouns.

Focusing on the out-of-the-box models, we rank them by their gender skew from best to worst as RoBERTa, ALBERT-xxlarge, RoBERTa-large, BERT, BERT-large, and DistilBERT.

We note that RoBERTa has the least skew bias, with a μ_{skew} value of 9.2% for T2. Liu et al. (2019) report that BERT was "significantly undertrained", and aimed to address this by training RoBERTa for longer, with bigger batches and sequences, additional data, and dynamic adjustments to the masking pattern. These amendments in RoBERTa appear to have reduced the skew bias in the model, suggesting that a model's training procedure can have a considerable impact on its skew. We also observed that within the BERT and RoBERTa families, larger models tend to show more skew than their smaller counterparts.

The high skew of DistilBERT might be due to its student-teacher training (Hinton et al., 2015). This lends itself to a overly simplistic understanding of male and female roles within society. Understanding the subtleties and nuances of gender roles
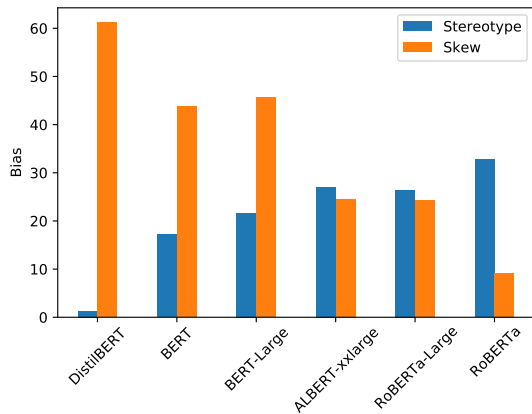
Figure 4: Bar chart showing the different bias measurements for the out-of-the-box models investigated. This suggests an inherent trade-off between skew and stereotype in language models.

requires models with high representation capacity and training DistilBERT to mimic BERT's output renders it incapable of making such distinctions.

The ranking of gender stereotype from best to worst is DistilBERT, BERT, BERT-large, RoBERTa-large, ALBERT-xxlarge, and RoBERTa.

Note that this order is approximately the opposite to skew, as illustrated in Fig. 4. There appears to be a potential trade off between the skew and stereotype in out-of-the-box language models, with RoBERTa-large best balancing the two biases. This trend appears to carry forward to the fine-tuned models, with BERT-A and BERT-U showing high stereotype but very low skew.

4.3 Mitigated Systems

4.3.1 Online Skewness Mitigation

Comparing how bias values change in Section 4.2 when going from all models to their normalised - O version, Online Skewness Mitigation successfully reduces stereotype for 6/8 models, though interestingly RoBERTa responded by going from a female stereotype to a significant male stereotype. At the same time, we observe that this reduction in model stereotype actually comes at the expense of model skew. Note that this effect is the opposite to what it was designed for. For this reason we do not explore it further.

4.3.2 Data Augmentation

BERT-U and BERT-A test accuracies on the OntoNotes data were 64% and 58% respectively. This 6% loss in accuracy should be kept in mind when assessing gender bias. We see BERT-U/A

have similar F1 values in Section 4.1 indicating that any difference in bias is not just noise coming from them having radically different abilities to perform the task.

Section 4.2 shows BERT-A has a 7.4% lower stereotype and a 7.8% lower skew than BERT-U, echoing the findings of Zhao et al. (2019) which demonstrate reductions of bias in ELMo through fine-tuning with data augmentation. The skew of BERT-A is reduced to 0.1%; much lower than standard BERT. This elimination of skew is unsurprising since augmentation ensures professional entities assume male and female roles equally in the fine-tuning data. Despite this, the stereotype of BERT-A is greater than standard BERT, suggesting a degree of non-skew bias has been introduced by fine-tuning. An analysis of the individual test examples in WinoBias T2 may reveal why this is the case. Across the classifications of sentences we see in the data, pronoun resolution appears to be influenced by the implied characteristics of the individual whose pronouns we seek to resolve, rather than the professions themselves.

4.4 Impact of Sentence Context on Pronoun Resolution

We consider an example from WinoBias T2 with the *[entity2]* profession being "guard". The only example where BERT-A assigned a female pronoun to the profession "guard" was a sentence where the guard was associated with low competency: "*The clerk works harder than the guard and gets more appreciation than [her]*". To investigate this more quantitatively, we measure the competency of the entity whose pronoun is being resolved. Our aim is to identify whether fine-tuning results in a gender shift across competent and incompetent examples which might suggest a competency-based gender assignment.

To achieve this, four of the authors labelled the WinoBias examples according to whether they thought the subject of the sentence demonstrated competent or skilful behaviour (e.g. getting a promotion, being thanked for a job well done) or not (e.g. getting sacked from their job). The sentences were classified as Incompetent, Neutral (no competency suggested in the sentence), or Competent. All voters classified sentences independently, with a Fleiss' κ score of 0.42. The class assigned to each sentence was then determined by a majority vote.

2238

Competency	BERT	BERT-U	BERT-A
Incompetent	0.156	0.062	0.281
Neutral	0.117	0.168	0.280
Competent	0.160	0.140	0.140

Table 3: Proportion of female pronouns assigned for each competency category for the WinoBias T2 sentences. All professional entities in the examples were replaced with the gender neutral term "person" to isolate the impact of competency from professional stereotype.

Sentences that resulted in a tie were discarded.[4] To isolate our investigation of subject competency from professional stereotype, all professions in the WinoBias dataset were replaced with the gender agnostic term "person".

Table 3 reports the proportion of examples in each competency class that were assigned a female pronoun across BERT, BERT-U, and BERT-A. The proportions of female pronoun assignments show that BERT-A allocates a more balanced ratio of gendered pronouns to Incompetent examples compared to BERT and BERT-U. Apart from the Competent class (which shows no major change across all three models), BERT-A reduces the gender imbalance of pronouns in Neutral and Incompetent examples.

It is challenging to exactly determine the cause of these observations, but it certainly appears that fine-tuning BERT models has an effect on the gender ratios in each competency class. It is notable that de-biasing BERT reduced the gender imbalance of Incompetent examples by a large margin. These findings merit further investigation.

We believe that WinoBias and other related benchmarks do not sufficiently probe professional gender bias, as pronoun resolution may be obfuscated by cross-correlations from other manifestations of gender prejudice. One example of a bias other than profession and competency could be personality bias, where women may be more closely associated with passive and caring traits whilst men may be more aggressive and disagreeable. We encourage the development of a dataset that isolates these different gender biases, allowing us to probe them without interference from one another.

---

5 Conclusion and Future Work

Quantifying gender bias in coreference resolution is challenging, since co-referencing performance and bias manifestation are closely linked. We have proposed *skew* and *stereotype* as new measures of gender bias, allowing us to better probe model prejudice.

We have shown that there is an approximate trade-off between the skew and stereotype of out-of-the-box models. DistilBERT and BERT models have high skew and low stereotype whilst RoBERTa and ALBERT-xxlarge have reduced skew at the cost of higher stereotype.

Two methods have been proposed to mitigate bias: Online Skewness Mitigation and Data Augmentation. The online approach has been shown to be effective at mitigating stereotype at the expense of skew, demonstrating the opposite effect to what it was designed for. We took the Data Augmentation method proposed by Zhao et al. (2019) for debiasing ELMo and extended it to BERT, demonstrating that it reduces both forms of gender bias compared to unaugmented fine-tuned models. However, the reduction of explicit professional gender skew and stereotype reveal the model's underlying bias towards gender competency. We successfully expose these using WinoBias GPR sentence probes labelled for competency.

Since contextual language models consider the full sentence contents when assigning a pronoun, we believe that the WinoBias data used in this paper does not purely measure professional biases. A second popular dataset taken from the SuperGLUE benchmarks, *Winogender* (Rudinger et al., 2018), is increasingly used for evaluating a model's gender bias. However, its limited size relative to WinoBias makes it less robust and hence it was not used in this paper.

We observed that language models may also consider other stereotyped gender characteristics in the sentence when classifying pronouns. Given the above considerations, we believe that a more comprehensive set of gender bias benchmarks should be developed which can better isolate specific biases within models.

Kiritchenko and Mohammad (2018) have shown that both race and gender bias are prevalent in a large proportion of state-of-the-art language models. Recently, a number of other datasets have appeared for detecting these and other kinds of bias such as age and religion (Nadeem et al., 2020;

Nangia et al., 2020). It would be interesting to see if competency bias obscures analyses on these datasets similarly.

Future research is recommended on how data augmentation affects other models from the BERT family. Additionally, it will be valuable to explore whether Data Augmentation could be applied to larger corpora to train a new contextual model from scratch.

Lastly, in contrast to static embeddings, it is notoriously hard, if not impossible, to define bias in contextual embeddings (Caliskan et al., 2017). It is likely that without extensive research and transparent communication, the field of NLP will be further scrutinised as more applications are found to exhibit undesired biases. Discussions, both within and outside the community, are required to determine what separates bias from semantic assumptions, allowing bias disclaimers and guidelines to be provided to downstream developers.

Acknowledgements

References

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

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

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

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

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

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *Advances in Neural Information Processing Systems*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *Association for Computational Linguistics*, pages 43–53.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *Association for Computational Linguistics*, pages 166–172.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. *ICLR*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *CoRR*, abs/1804.05392.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy.

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

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *CoRR*, abs/1807.11714.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *Association for Computational Linguistics*, pages 622–628.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint: 2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *Empirical Methods in Natural Language Processing*, pages 1953–1967.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Association for Computational Linguistics*, pages 8–14.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *Association for Computational Linguistics*, pages 1630–1640.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, pages 13230–13241.

Jason Tashea. 2017. Courts are using ai to sentence criminals. that must stop now. *WIRED*.

U.S. Bureau of Labor Statistics. 2017. Labor force statistics from the current population survey, 2017. https://www.bls.gov/cps/cpsaat11.htm.

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).

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Association for Computational Linguistics*, 6:605–618.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *Association for Computational Linguistics*, pages 629–634.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *Association for Computational Linguistics*, pages 15–20.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *Association for Computational Linguistics*, pages 4847–4853.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.

A Fine-tuning Training Parameters

For fine-tuning BERT on the OntoNotes data, the following settings were used. Standard hyperparameter choices of $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$, and a dropout probability of 0.1 were chosen. Model training and validating with a 80/20 train/test split of the training data, across training epochs $\in \{1, \ldots, 10\}$. The selected (epoch) model was that with the highest pronoun prediction accuracy on the validation set.