Lipstick on a Pig "Again": Analysis on Contextualised Embeddings

Hyoung Jo, Bhang and **Hong Seok, Kang** and **Kyung Wook, Nam**
School of Computing
KAIST (Korea Advanced Institute of Science and Technology)
{hyoungjo.bhang, ghdtjr0311, ruddnrdl}@kaist.ac.kr

Abstract

Lipstick on a Pig (Gonen and Goldberg, 2019) shows that existing gender bias removal methods in word embeddings are superficial, and should not be trusted. Meanwhile, recent advances in contextual models have been astonishing, and many researchers have proved the existence of biases and proposed different mitigation methods in those models. For this reason, we extend Gonen and Goldberg (2019) to discover gender biases in contextualised language models. As a result, we have identified the presence of *systematic* gender biases in pre-trained BERT and also in debiased BERT models. We hope that our work will inspire future research on analyzing gender biases of diverse debiasing methods.[1]

1 Introduction

Word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) were powerful tools for downstream tasks in NLP. Despite its broad use, the fact that word embeddings contain various social biases has been proven by many researchers. They not only demonstrated the existence of biases in language models and showed they can even amplify wrong biases, but some also tried to mitigate the biases. Typically, Bolukbasi et al. (2016) and Zhao et al. (2018) pointed out gender biases in Word2Vec and GloVe, respectively, and proposed debiasing methods. In this paper, we will refer to those as HARD-DEBIASED and GN-GloVe.

However, works suggesting otherwise followed. In Gonen and Goldberg (2019), the authors argued that bias mitigation methods (Bolukbasi et al., 2016; Zhao et al., 2018) cover up the biases rather than removing them. The paper especially focused on more profound *systematic* biases (which are independent of the gender direction, introduced in Gonen and Goldberg (2019)). To support that, the paper set up several experiments to show that the biases still remained. For instance, Figure 1 obviously demonstrates that gender-biased words still cluster after applying debiasing methods of HARD-DEBIASED and GN-GloVe.

Nonetheless, the recent emergence of BERT (Devlin et al., 2018) and other contextualised models such as ELMo (Peters et al., 2018) changed the trend. While the previous works only handled embeddings fixed for each word, not affected by its context, these language models now considered contextual information to understand the corpus and as a result, produced surprisingly increased performances. Yet again, just like word embeddings, researchers have proved that gender biases still exist in contextualised embeddings (Zhao et al., 2019; Basta et al., 2019) and are actively trying to measure and mitigate them.

Unfortunately, research on *systematic* biases has not been done. A related work, Basta et al. (2019), has revealed the presence of gender bias by using clustering and classification, following Gonen and Goldberg (2019) in ELMo. But it only presented biases in pre-trained ELMo itself. As an enhancement of Gonen and Goldberg (2019), our research examines biases in contextual embeddings from not only BERT but also compares it to the embeddings from BERT-debiased models. In that process, we put forward the next two research questions:

RQ1. How much gender bias do the contextualised BERT embeddings contain?

RQ2. Did the proposed debiased BERT models really resolve gender biases?

[1] Our code and data are available at https://github.com/colorsquare/lipstick-on-a-pig

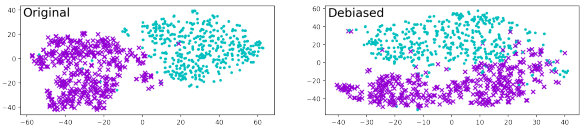

Figure 1: Visualization of Word2Vec word embeddings. Top 500 male- and female-biased words are used to show distinct clusters.

2 Approach

The main contributions of this research are two things. First, contextualised embeddings are used as our source. In Gonen and Goldberg (2019), static words embeddings from Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) are used. In contrast, a method that extracts embeddings that reflect contextual information from BERT is designed. Those static word embeddings will be replaced with newly generated contextualised embeddings. Second, as briefly mentioned, debiased models processed on top of BERT will be utilized. Three debiased BERT models are applied in total: (Liang et al., 2020; Kaneko and Bollegala, 2021; Bartl et al., 2020).

2.1 Word to Contextualised

Generating contextualised embeddings from BERT (Devlin et al., 2018) requires an additional process. Embeddings from multiple sentence templates are averaged to extract one representative embedding for the target words. Then, the embeddings are used to calculate the gender biases. Detailed will be elaborated in section 3.1.

Referring back to our second research question, we next tackled the effectiveness of debiasing methods. So for determining gender bias metric, we not only replicated Gonen and Goldberg (2019) as itself but adopted the concept of one of the major gender bias evaluation metrics in contextualised embeddings, *SEAT* (May et al., 2019).

2.2 Debiased Models

2.2.1 Sent-Debiased

Sent-Debiased (Liang et al., 2020) is a sentence-level extension of HARD-DEBIASED (Bolukbasi et al., 2016). It contextualises the pre-defined sets of *attribute* (potentially biased) words into sentences so that sentence encoders can be applied to obtain sentence representations. Using this, Sent-Debiased tries to mitigate gender biases by subtracting gender bias projection from the embeddings after the training (only leaving the orthogonal). This post-hoc method lowered the SEAT score which means this method works for mitigating the bias level in BERT and ELMo while preserving performance on downstream tasks.

2.2.2 Context-Debiased

Context-Debiased (Kaneko and Bollegala, 2021) adopted two losses in fine-tuning process. One minimizes the cosine similarity with the target gender word, and the other preserves useful information for fine-tuning performance. Context-Debiased showed the effectiveness of debiasing discriminative gender-related biases, while preserving useful semantic information in the pre-trained embeddings by SEAT (May et al., 2019) and MNLI task(Dev et al., 2020).

2.2.3 BERT-CDS

In order to mitigate *systematic* bias that (Gonen and Goldberg, 2019) mentioned, name-based CDS(counterfactual data substitution) was introduced. It swaps the gender of words that denotes all personnel in place for counterbalancing (Maudslay et al., 2019). Also, (Bartl et al., 2020) fine-tuned BERT with gender-swapped corpus and achieved statistically balanced gender associations on their own test procedure. This model will be denoted as BERT-CDS in the rest of the paper.

3 Data & Experiments

3.1 Experimental Framework

Below are the experimental setups conducted, to proceed with experiments based on (Gonen and Goldberg, 2019).

3.1.1 Sentence Data Generation

Following Gonen and Goldberg (2019), word pools used in (Bolukbasi et al., 2016; Zhao et al., 2018), are retained, and vocabularies are reduced equally: with most frequent 50,000 words or phrases, words with upper-case letters, digits, or punctuation and words longer than 20 characters are removed. Also, inherently gendered words (e.g. man, woman, son, ..), where natural bias is expected, are removed.

Using this, sentences for obtaining contextualised embedding are generated from semantically bleached templates. For instance, a sentence template looks like "This is a [blank]". And to compute the contextualised embedding of target word $word$, we insert $word$ into the blank of the template and then feed to the BERT.

Templates are chosen from May et al. (2019)[2], converting each noun word to 11 sentences (to 5 singular and 6 plural terms). In order to apply templates properly, only noun words are selected using NLTK Tag package. After the noun filtering, sentences using 18,445 words in Word2Vec and 39,385 words in GloVe are generated and fed to BERT.

3.1.2 Contextualised Embedding Extraction

Since BERT uses subword tokenization, words not found in BERT are tokenized into smaller components(e.g. 'babysitter' → 'baby', '##sit', and '##ter'). When extracting embedding in such a contextualised environment, it is important not only to retain enough contextualised information but also to maintain its characteristics or 'uniqueness' for performance in downstream tasks. According to Ács et al. (2021), which carried out a broad investigation on various pooling methods, the best performing approaches were embeddings created by adding another attention or LSTM layer. However, to examine the sheer bias of BERT embedding itself, we decided not to train another layer, but use the next best metric, adding the first and last token. Along with this metric, to preserve contextual information from other tokens as much as possible, concatenation of the last four hidden layers has been done (Devlin et al., 2018).

3.1.3 Definition of Gender Space

Contextual information is also needed to define the base direction for inherently gendered words. For that, each 80 male and female sentences are adopted from May et al. (2019). We define gender space as an average of those sentence encodings.

3.1.4 Debiased Model

For Sent-Debiased and Context-Debias, we use released pre-computed models by authors. To obtain CDS applied model, fine-tuning was applied on the CDS applied GAP coreference dataset (Webster et al., 2018). Following Maudslay et al. (2019), both gender pronouns and names were substituted. The sentences are pre-processed and attention masks are created. The model is trained for three epochs using an AdamW optimizer with a learning rate of 5×10^{-5} and a linear scheduler with a warm-up.

[2] https://github.com/W4ngatang/sent-bias

3.2 Experiments

3.2.1 Male- and female-biased words cluster together

500 most biased sentences (from each male and female) according to the original bias level in BERT are chosen. Then, those embedding vectors are clustered with K-Means (number of cluster = 2) and visualized by tSNE (Van der Maaten and Hinton, 2008). If a clear clustering is observed, it represents the bias level in embeddings. In addition, the accuracy between the cluster and the original biased words was also calculated.

3.2.2 Professions

Profession terms from Bolukbasi et al. (2016) and Zhao et al. (2018) are used. For each profession (e.g. accountant, entrepreneur, surgeon, ..), the number of male-biased neighbours is counted among the 100 closest neighbours as the Y-axis, and the X-axis is plotted using original bias accordingly. More words were distributed higher for greater male bias in the model, and the correlation with the original bias and the number of male neighbours are further observed.

3.2.3 Classifying previously female- and male-biased words

SVM classifiers are trained and tested if they can learn to distinguish gender-biased words. With 2,500 most biased embeddings from each gender, 20% were used for training, and 80% were used for evaluation. It is a rather simple but intuitive technique that yields strong results, with higher accuracy referring to greater bias in the embeddings themselves.

4 Results

4.1 Replication

Our replication of Gonen and Goldberg (2019) followed the specified steps from the authors' github repository.[3] Word embeddings before debiasing are from Word2Vec (Mikolov et al., 2013)[4] and GloVe (Pennington et al., 2014)[5], and the debiased word embeddings are from HARD-DEBIASED

[3] https://github.com/gonenhila/gender_bias_lipstick

[4] https://code.google.com/archive/p/word2vec/

[5] https://nlp.stanford.edu/projects/glove/

	Clustering (K-Means Accuracy)		Professions (Pearson Correlation)		SVM (Classification Accuracy)	
Models	**Word2Vec**	**GloVe**	**Word2Vec**	**GloVe**	**Word2Vec**	**GloVe**
Origin	99.9%	100%	0.747	0.820	99.97%	94.17%
Debiased	92.5%	85.6%	0.606	0.792	100.00%	99.98%

Table 1: Table shows the accuracy and correlation values for each experiment for replication

(Bolukbasi et al., 2016)[6] and GN-GloVe (Zhao et al., 2018)[7].

Table 1, Figure 2, Figure 3 show the replication results. It exactly replicates the results from the paper, but the classification where random selection of training data can differ. As stated in the paper, the results argue that the debiasing methods used in both Bolukbasi et al. (2016) and Zhao et al. (2018) are not really effective, but only cover up biases.

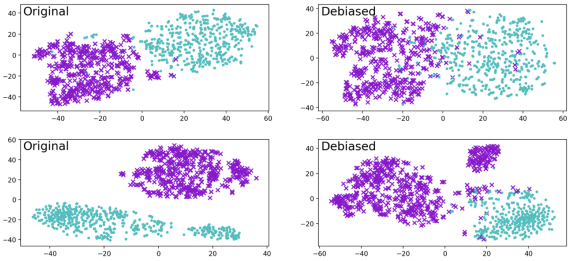

Figure 2: Clustering for Word2Vec (above) and GloVe (below) embeddings, before (left-hand side) and after (right-hand side) debiasing

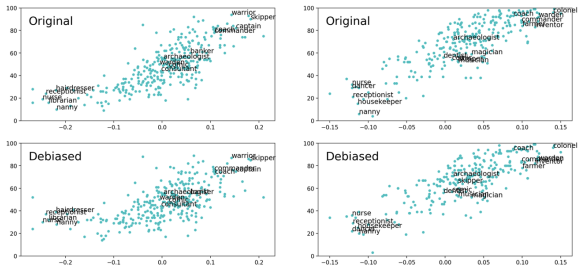

Figure 3: The plot for Word2Vec (left) and GloVe (right) embeddings, before (above) and after (below) debiasing

4.2 BERT Baseline

This section relates to our **RQ1.** *How much gender bias do the contextualised BERT embeddings contain?* As explained in section 3.1 in detail, contextualised BERT embeddings are extracted from the BERT model. Then, with the biases calculated by projections, experiments in section 3.2 are conducted.

Table 2, Figure 4, Figure 5 are the results. Clustering is more clear with Zhao et al. (2018) vocabulary, but Bolukbasi et al. (2016) also exhibits clear division. In Profession Figure 5, we can observe that the neighbors (denoted as Y-axis) are male-oriented for most professions. Classification accuracies in Table 2, also prove that the SVM classifier distinguishes male and female embeddings with great precision. In short, contextualised BERT embeddings contain a substantial amount of biases, quantified by the figures and scores.

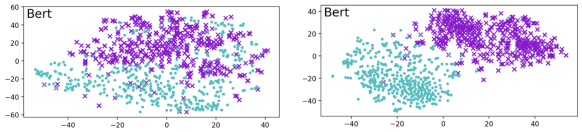

Figure 4: Clustering for contextualised BERT embeddings, Word2Vec (left-hand side) and GloVe (right-hand side)

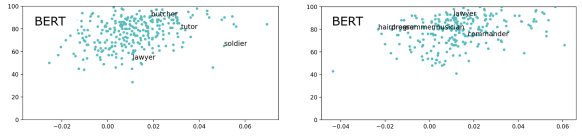

Figure 5: The plot for contextualised BERT embeddings, Word2Vec (left-hand side) and GloVe (right-hand side)

4.3 Debiased BERT Models

Debiased models using BERT from section 2.2 are tested: Sent-Debiased (Liang et al., 2020), Context-Debiased (Kaneko and Bollegala, 2021), BERT-CDS Bartl et al. (2020).

This section answers **RQ2.** *Did the proposed debiased BERT models really resolve gender biases?* To give a straightforward answer, all debiased models were not found to be effective enough. Although there are signs of mitigation of biases, most still remained after debiasing, and some models were even debiased further, creating biases towards the opposite gender.

[6] https://github.com/tolga-b/debiaswe
[7] https://github.com/uclanlp/gn_glove

	Clustering (K-Means Accuracy)		Professions (Pearson Correlation)		SVM (Classification Accuracy)	
Models	**Word2Vec**	**GloVe**	**Word2Vec**	**GloVe**	**Word2Vec**	**GloVe**
BERT	94%	100%	0.294	0.299	99.4%	99.9%
Sent-Debiased	*79.1%	99.7%	0.214	0.338	98.7%	100%
Context-Debiased	*51.5%	*51.2%	0.162	0.460	94.2%	98.5%
BERT-CDS	*54.3%	*53.3%	0.128	0.252	96.8%	98.8%

Table 2: Table shows the accuracy and correlation values for each experiment on pre-trained BERT and debiased models. Prefix * denotes accuracy score does not capture 'clustering' in figures, and underlined numbers mean classification accuracy drop on debiased models

In the following subsections, we revisit each experiment in section 3.2 to evaluate the bias level of the debiased models, and the details can be found in each.

4.3.1 Male- and female-biased words cluster together

Without a doubt, the embeddings from debiased models in Figure 6 are still clustered. Although it displays a slightly more 'mixed' compared to Figure 4 which represents the BERT baseline, regional clusters are evident in all six figures. Also, figures using Zhao et al. (2018) vocabulary show a more apparent separation than Bolukbasi et al. (2016) figures.

In contrast, K-Means Accuracy in Table 2 suggests otherwise. K-Means clustering sometimes captures a 'wrong cluster'. This will be further discussed in section 5.1.

4.3.2 Professions

Professions experiment demonstrated compelling results. Most profession embeddings have male-oriented neighbors in Figure 5 and in Figure 7. However, *female*-orientation was observed in Context-Debiased and BERT-CDS embeddings.

According to section 2.2, Sent-Debiased only removes the projection to the gender-bias direction. In comparison, Context-Debiased minimizes cosine similarity with the target gender word, and BERT-CDS switches variables from male to female. Then, both models, Context-Debiased and BERT-CDS, initiate a training process that emphasizes 'female' in embeddings against 'male'. Therefore, we believe a more careful approach should be designed as debiasing methods to prevent such "reversed bias", leaning towards female embeddings.

4.3.3 Classifying previously female- and male-biased words

All models regardless of the vocabulary used presented high accuracy. In the meanwhile, most debiased models showed decreased accuracy in Table 2 compared to the BERT baseline numbers, which indicates partial debiasing has taken place.

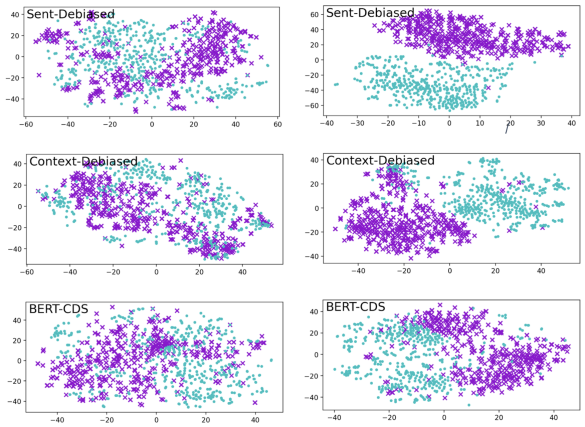

Figure 6: Clustering for Debiased Models (Sent-Debiased, Context-Debiased, BERT-CDS) embeddings from the top. The same words are used as in BERT.

5 Discussions

5.1 K-Means Algorithm

K-Means clustering sometimes captures the 'wrong cluster'. Simply thinking, not all male- or female-biased embeddings are located in one corner of the entire vector space. Instead, those embeddings can be spread in different parts of the vector space, with little bias encoded to each vector. In that case, it is obvious simple K-Means Algorithm with only two clusters will not be able to capture the entire picture of bias.

Even more, the shape of the distribution of vectors varies across different models. The number of clusters required to efficiently distinguish gen-

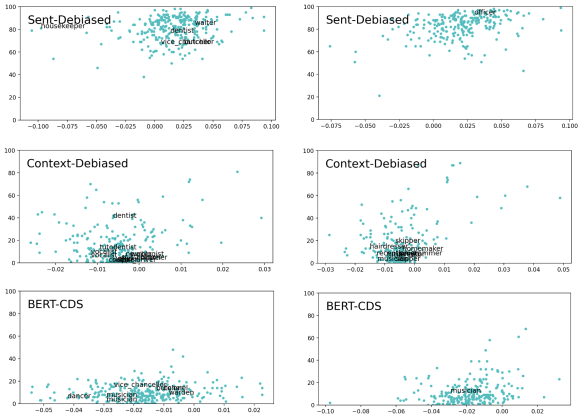

Figure 7: The plot for Debiased Models (Sent-Debiased, Context-Debiased, BERT-CDS) embeddings from the top. The same words are used as in BERT.

der bias vectors can be different, and an algorithm that can group similar clusters into one will also be necessary. We believe solving this problem will lead to huge enhancements since it enables a proper calculation of accuracy in "clustering" experiment of section 4.3.1. Thus, it will lead to a more precise analysis of debiasing models by quantifying the degree of clustering.

5.2 Validity of Pearson Correlation

In Gonen and Goldberg (2019), the number of male neighbors in the Y-axis and bias level in the X-axis showed a clear correlation, which justified the metric of 'the number of male neighbors'. However, from Figure 5, a contextualised environment, no clear correlation between those two values could be obtained. Therefore, a comparison between the correlation value before and after debiasing in our experiment seems not to be valid. Consequently, in Professions experiment (section 4.3.2), we only retrieve information that embeddings are either male- or female-oriented.

5.3 Future Works

One limitation of our study is that our pre-defined sentence templates used for extracting embedding do not contain real-world sentences. Also, to simplify the experiments, we only utilized nouns as attributes in sentence generation. To fully obtain the contextualised embedding of specific words, more complicated sentence templates from real corpora, and a diversified part-of-speech as attribute words will be helpful.

Next, it will be meaningful to analyze the presence of strong bias on gender vectors in BERT. Following Basta et al. (2019), principal components were computed from gender vectors and compared with random vectors. In their experimental setup, the first eigenvalue of gender vectors is significantly larger than others, which denotes a single direction. Although further research is necessary, integration of gender vector space when calculating cosine similarity might help us build a more concrete definition of gender bias.

Lastly, we present an interesting point compared with Basta et al. (2019). Despite the fact that they used a different language model (ELMo) from us (BERT), their clustering and classification accuracy were much lower. Compared to our BERT Baseline clustering and classification in section 4.2 which is close to 100%, their average clustering and classification accuracy are 70.1% and 85.56%, respectively. Also, they claim that 'making embeddings contextual' tends to mitigate gender bias. It is quite different from what we have found as well. Exploration towards analyzing which language model and experimental settings make the change will be very interesting.

With these future improvements, we look forward to a concrete methodology that evaluates systematic gender biases of different natural language models and the debiased models.

References

Judit Ács, Ákos Kádár, and András Kornai. 2021. Subword pooling makes a difference. *arXiv preprint arXiv:2102.10864*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*.

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

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*, 29.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

bidirectional transformers for language understanding. *arXiv preprint arXiv: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. *arXiv preprint arXiv:1903.03862*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

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

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

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

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.