Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers

Andrew Silva* **Pradyumna Tambwekar*** **Matthew Gombolay**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA
{andrew.silva, ptambwekar3, matthew.gombolay}@gatech.edu

Abstract

The ease of access to pre-trained transformers has enabled developers to leverage large-scale language models to build exciting applications for their users. While such pre-trained models offer convenient starting points for researchers and developers, there is little consideration for the societal biases captured within these model risking perpetuation of racial, gender, and other harmful biases when these models are deployed at scale. In this paper, we investigate gender and racial bias across ubiquitous pre-trained language models, including GPT-2, XLNet, BERT, RoBERTa, ALBERT and DistilBERT. We evaluate bias within pre-trained transformers using three metrics: WEAT, sequence likelihood, and pronoun ranking. We conclude with an experiment demonstrating the ineffectiveness of word-embedding techniques, such as WEAT, signaling the need for more robust bias testing in transformers.

1 Introduction

Transformer models represent the state-of-the-art for many natural language processing (NLP) tasks, such as question-answering (Devlin et al., 2019), dialogue (Smith et al., 2020), search results (Nayak, 2019), and more. Popular pre-trained models, such as those available from Hugging Face (Wolf et al., 2019), allow developers without extensive computation power to benefit from these models. However, it is important to fully understand the latent societal biases within these black-box transformer models. Without appropriately considering inherent biases, development on top of pre-trained transformers risks exacerbating and propagating racial, gender, and other biases writ large.

Before transformers, word embedding models such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) were shown to exhibit systematic sexist (Bolukbasi et al., 2016) and racist (Manzini et al., 2019) biases. Initial investigations into bias for transformers (Vig et al., 2020; Basta et al., 2019; Bommasani et al., 2020) have found that these new language models are similarly biased. As transformers are increasingly commonplace, a more complete view of the inequalities, biases, or under-representations within pre-trained transformers becomes increasingly important.

Yet, discovering bias in transformer models has proven to be more nuanced than bias-discovery in word embedding models (Kurita et al., 2019; May et al., 2019). Prior work on bias in modern transformer models has used only a single test or metric at a time, which we show in this paper provides an incomplete view of the problem. Furthermore, we find evidence that certain tests are ill-suited to understanding bias in transformer architectures, supported by prior work (Blodgett et al., 2020). Moreover, we show that employing multiple tests is necessary for a full picture of the issue as no single test is currently sufficient.

In the context of our work, "bias" refers specifically to the preference of a model for one gender or race in the presence of an otherwise neutral context. As an example, consider the sequence "[MASK] wept upon arriving to the scene." With no additional information, an equitable system would exhibit no preference for *female* over *male*, or *African-American* over *European-American* names; however, our results indicate that there is often a statistically significant preference ($p < 0.0001$) for associating *female* and *African-American* identifiers with being more "emotional."

We provide two key contributions to understanding and mitigating bias in contextual language models. First, we conduct a comprehensive, comparative evaluation of gender and racial bias using multiple tests for widely-used pretrained models. Second, we construct a novel experiment for debiasing a contextual language model on a downstream task (Zellers et al., 2018). Our experiment

* These authors contributed equally to this work

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2383–2389
June 6–11, 2021.

Model Name	W_C	W_M	W_S	SEQ_A	SEQ_F	SEQ_S	SEQ_J	PN_A	PN_F	PN_S	PN_J
Uncased:											
BERT-Base	1.47	-0.33	-0.3	4.53*	3.70	2.53	4.02*	5.29*	-3.31	-2.65	-1.62
BERT-Large	1.10	-0.55	-0.16	0.53	0.33	0.83	1.07	5.42*	-3.15	-3.62	-2.11
BERT-LargeM	1.60	-0.24	-0.33	-2.90	-2.14	-2.39	-2.48	1.41	0.64	-0.71	1.38
DistilBERT	1.64	-0.37	-0.34	5.85*	6.20*	6.08*	6.08*	2.82*	-4.71*	-5.22*	-5.06*
ALBERT-Base	1.41	1.61	1.51	-3.98*	-3.48	-3.27	-3.15	-19.4*	-19.7*	-19.3*	-19.9*
ALBERT-Large	1.46	1.42	1.05	-3.75	-2.79	-3.55	-3.61	0.96	-2.47	-2.94	-6.00*
ALBERT-XLarge	1.52	1.54	1.55	1.47	2.02	1.37	0.99	3.90*	0.32	1.55	-4.56*
ALBERT-XXLarge	1.47	1.38	1.39	-2.45	-1.39	-0.97	-1.44	5.89*	4.85*	2.30	-0.09
Cased:											
BERT-Base	0.30	-0.04	0.57	8.83*	10.8*	10.6*	10.6*	4.17*	0.17	-1.65	-3.12
BERT-Large	0.53	-0.44	-0.05	5.17*	5.47*	4.50*	5.47*	1.44	-0.91	-1.66	-1.18
BERT-LargeM	0.18	0.23	-0.15	2.63	3.78	4.15*	3.93*	2.27	-0.55	-1.79	-3.21
DistilBERT	0.14	-0.27	0.57	11.1*	11.6*	11.7*	11.7*	2.15	-6.17*	-7.11*	-9.19*
RoBERTa-Base	0.91	0.59	0.67	4.19*	4.59*	4.44*	4.36*	-0.99	-4.80*	-5.14*	-4.10*
RoBERTa-Large	0.56	0.64	0.68	3.95*	4.54*	5.41*	5.55*	2.09	-2.92	-1.01	-1.67
DistilRoBERTa	1.00	0.66	0.56	12.6*	12.6*	12.4*	12.6*	-2.47	-8.55*	-8.19*	-8.28*
GPT-2	0.78	-0.03	-0.31	-2.99	-1.95	-3.38	-2.55	1.88	2.31	2.45	1.50
GPT-2-Medium	0.24	-0.21	0.07	1.51	2.92	2.21	2.11	0.26	0.19	0.38	0.31
GPT-2-Large	0.54	0.04	-0.46	3.43	3.92*	3.02	3.72	-0.59	-0.50	-0.03	-1.37
GPT-2-XLarge	0.53	-0.23	0.13	3.18	4.06*	2.90	3.24	7.51*	1.35	2.96	6.33*
XLNet-Base	0.60	0.69	0.36	1.75	2.63	1.99	1.08	0.46	0.96	1.07	1.00
XLNet-Large	0.16	0.10	0.42	2.34	2.94	5.74*	3.67	-0.01	3.09	1.01	0.64

Table 1: Bias scores along the gender dimension. Positive indicates bias towards *Male*; negative indicates bias towards *Female*. Asterisks denote statistical significance $\alpha = 0.05/336$.

refutes the validity of WEAT for contextual models, signaling a need for new bias metrics.

2 Related Work

After the seminal work of Bolukbasi et al. (2016), bias has been found ubiquitous in word embedding models (Amorim et al., 2018; Brunet et al., 2018; Rudinger et al., 2018; Zhao et al., 2017; Costa-jussà et al., 2019; Silva et al., 2020). Researchers have applied association tests between word embeddings to look for inappropriate correlations. Caliskan et al. (2017) introduce the Word Embedding Associate Test (WEAT) to estimate implicit biases in word embeddings by measuring average cosine similarities of target and attribute sets. The WEAT has been extended into a sequence test (May et al., 2019), though the efficacy of both tests remains in question for transformers (Ethayarajh et al., 2019; Kurita et al., 2019).

Prior work has also devised methods to measure contextual bias. Kiritchenko and Mohammad (2018) introduce the Equity Evaluation Corpus (EEC), which includes templated sequences such as "⟨TARGET⟩ feels ⟨ATTRIBUTE⟩," where gendered or racial tokens are the "targets" and emotional words are the "attributes." The average of the difference in likelihoods for target sets constitutes the bias score. We leverage this in our work as the sequence ranking test (SEQ).

Kurita et al. (2019) and Vig et al. (2020) devise a pronoun-ranking test for BERT by comparing relative likelihoods of target words. Rather than sequence likelihood, the authors instead measure *contextual* likelihood, which helps to control for a model's overarching bias. We extend this work, applying the pronoun-ranking test (PN) to score the most commonly used transformer models and contextualizing the results with SEQ scores.

Investigations of biases in contextual language models, e.g. transformers, have yielded mixed results. Basta et al. (2019) found that BERT and GPT exhibit a reduced bias-dimension relative to word embedding models, whereas Kurita et al. (2019) found that BERT is biased and that conventional tests, e.g. WEAT, are inappropriate. Recent work has also looked to identify bias by crowdsourcing a sterotype dataset (Nadeem et al., 2020; Zhao et al., 2018; Nangia et al., 2020). These approaches develop a bias analysis metric by empirically computing a pretrained model's preference towards stereotyped sentences. However, such work is specifically focused on showcasing the effectiveness of these specific datasets for identifying bias. Our results paint a more complete picture, providing insight into specific aspects of gender and racial bias and unifying disparate viewpoints of prior work. Furthermore, we present a targeted investigation into the relevance of the WEAT for transformers.

Model Name	W_R	SEQ_A	SEQ_F	SEQ_S	SEQ_J	PN_A	PN_F	PN_S	PN_J
Uncased:									
BERT-Base	0.66	-10.8*	-12.7*	-12.3*	-13.5*	0.74	-1.45	-1.70	-3.82*
BERT-Large	0.02	6.91*	8.11*	4.40*	6.34*	-0.90	-2.82	-2.73	-3.61
BERT-LargeM	0.44	-13.7*	-14.3*	-13.6*	-13.4*	-5.13*	-11.4*	-9.34*	-5.65*
DistilBERT	1.15	-21.3*	-22.4*	-22.2*	-22.4*	-5.84*	-6.80*	-13.5*	-14.8*
ALBERT-Base	0.45	-18.4*	-18.2*	-17.8*	-17.6*	-17.5*	-17.4*	-17.5*	-17.6*
ALBERT-Large	0.62	-16.9*	-19.2*	-19.7*	-19.6*	-19.0*	-19.3*	-20.0*	-20.0*
ALBERT-XLarge	0.85	0.26	-1.53	0.72	-1.80	-7.87*	-8.81*	-6.68*	-12.89*
ALBERT-XXLarge	0.48	-5.05*	-5.98*	-5.43*	-5.21*	-5.24*	-6.18*	-5.97*	-7.26*
Cased:									
BERT-Base	-0.22	-22.4*	-24.3*	-24.2*	-23.6*	-9.7*	-10.3*	-10.4*	-13.4*
BERT-Large	0.17	-18.9*	-20.6*	-18.6*	-20.8*	-1.61	-2.31	-2.65	-2.30
BERT-LargeM	0.003	-23.8*	-27.4*	-25.6*	-23.9*	-6.81*	-9.63*	-11.69*	-7.74*
DistilBERT	-0.03	-28.7*	-29.8*	-29.1*	-29.1*	-15.5*	-13.9*	-19.3*	-17.4*
RoBERTa-Base	0.22	-20.8*	-20.7*	-20.5*	-20.2*	-2.81	-5.26*	-2.77	-5.09*
RoBERTa-Large	0.94	-21.2*	-22.0*	-22.6*	-21.9*	-2.18	-4.37*	-3.75	-5.33*
DistilRoBERTa	0.14	-11.17*	-10.8*	-10.6*	-10.5*	-6.94*	-5.00*	-7.19*	-11.7*
GPT-2	0.46	-2.25	-0.95	-0.21	0.29	0.06	-0.29	0.18	0.18
GPT-2-Medium	0.53	-4.31*	-3.81*	-3.00	-2.52	0.09	0.38	0.13	-0.08
GPT-2-Large	0.33	-1.66	-1.00	-0.09	-0.17	5.78*	2.51	2.41	1.59
GPT-2-XLarge	-0.16	-0.81	-0.27	0.56	0.88	18.84*	9.83*	2.57	5.12*
XLNet-Base	-0.17	-2.84	-4.05*	-3.22	-4.73*	-0.58	-0.71	-1.03	-0.67
XLNet-Large	-0.03	-15.3*	-16.6*	-15.9*	-12.2*	-3.01	-2.46	-1.70	-4.36*

Table 2: Bias scores along the racial dimension. Positive indicates bias towards *European-American*; negative indicates bias towards *African-American*. Asterisks denote statistical significance at $\alpha = 0.05/336$.

3 Approach and Results

We apply three tests (i.e. the WEAT (W), sequence likelihood (SEQ), and pronoun ranking (PN)) to popular pre-trained transformers from Hugging Face (Wolf et al., 2019), including the cased and uncased[1] BERT and DistilBert models, the uncased ALBERT models, and the cased RoBERTa, DistilRoBERTa, GPT-2, and XLNet models. For gender, we compare the WEAT tests for career (W_C), math (W_M), and science (W_S), against the sequence likelihood and pronoun ranking tests for *anger* (SEQ_A and PN_A), *fear* (SEQ_F and PN_F), *sadness* (SEQ_S and PN_S), and *joy* (SEQ_J and PN_J) evaluated between *male* and *female* target words. For race, we use the only WEAT available for race (W_R) as well as the same SEQ and PN tests evaluated between *African-American* and *European-American* targets.

The results of our WEAT, sequence likelihood, and pronoun ranking bias tests are presented in Tables 1 and 2. The quantity listed for each model/test pair is the effect size for that two-sided t-test test under they hypothesis that there is a significant difference between the mean likelihoods across the two groups. Using multiple tests is important; many models exhibit systematic preference for one target according to SEQ, while the PN reveals contextual preference in a different direction. The models often assign higher likelihood to *male* sequences, but when specifically considering the subject of an emotional sentence, *female* subjects are more likely. To address inherent model bias, it is important to understand how this bias manifests which we discuss below.

Model size and bias – Examining the SEQ and PN results for distilled models DistilBERT and DistilRoBERTa, we see that these models almost always exhibit statistically significant bias and that the effect sizes for these biases are often much stronger than the original models from which they were distilled (BERT and RoBERTa). This finding is in line with contemporary work by Hooker et al. (2020), who show that distillation in vision models disproportionately harms underrepresented groups. We show that the same is true for transformers.

The opposite is not true: increasing model capacity does not remove bias. While prior work (Gilburt, 2019; Tan and Celis, 2019) has reported increasing model size correlates with decreasing bias, we find that this is not always the case (see GPT2-Base vs. GPT2-Large), as supported by Nadeem et al. (2020) in stereotype-likelihood tests.

Tokenization matters – We consider four architectures that come in cased and uncased versions, differing only in tokenization BERT-Base, BERT-Large, BERT-LargeM, and DistilBERT. Across

[1]Casing is a design decision affecting the tokenization for a model. For all models, we test every size available.

Model Name	W_C	W_M	W_S	W_R	SEQ_A	SEQ_F	SEQ_S	SEQ_J	PN_A	PN_F	PN_S	PN_J
Gender:												
SWAG-Only	0.91	0.63	0.70	–	14.4*	14.2*	14.8*	16.5*	10.6*	7.98*	10.15*	**0.13**
+WEAT	**-0.006**	**0.003**	**0.0002**	–	-7.74*	-9.95*	-10.9*	-11.4*	37.3*	36.8*	37.9*	37.77*
Race:												
SWAG-Only	–	–	–	0.21	-13.5*	-15*	-14.6*	-13.3*	**-0.03**	**2.70**	**1.30**	**-3.89***
+WEAT	–	–	–	**-0.002**	**-8.62***	**-9.85***	**-9.02***	**-7.95***	-2.57	-5.86*	-6.99*	-10.6*

Table 3: Positive indicates bias towards *European-American* or *male*; negative indicates bias towards *African-American* or *female*. Asterisks denote statistical significance at $\alpha < 0.05/72$. Lowest effect sizes are bold.

race and gender, the uncased models exhibit less bias and greater diversity for names and pronouns.

The effects of tokenization may also play a role in WEAT's underperformance, as the mean-embeddings used to estimate a WEAT effect do not accurately reflect the expected words for the test. For example, under the ALBERT tokenizer, "Nichelle" becomes "niche" and "lle", two subwords which may not average out to a name.

WEAT is inconsistent – We find that WEAT is a poor predictor of contextual bias and an internally-inconsistent metric. The WEAT for math (W_M) and science (W_S) use words which are very similar and, at times, even overlapping. As such, we would expect the W_M and W_S scores to indicate bias in the same direction for every model. Instead, we see that the WEAT results show differing magnitudes and occasionally point in different directions.

Given the inconsistency of WEAT and its poor correlation with SEQ and PN effects, we propose a debiasing scheme using the WEAT effect. If neutralizing the WEAT effect also neutralizes SEQ and PN bias, then the WEAT remains a useful test for transformers. However, if neutralizing the WEAT has no effect on the SEQ and PN scores, we can conclude that the WEAT is simply not appropriate for contextual models.

4 Debiasing Transformers with WEAT

We now employ WEAT scores as a loss regularizer to "de-bias" a RoBERTa model being trained on the Situations With Adversarial Generations (SWAG) dataset, a commonsense inference dataset in which each sample is a sentence with four possible endings (Zellers et al., 2018). The SWAG training objective is to minimize the model's cross-entropy loss, L_{MC}, for choosing the correct ending. In addition to this loss, we incorporate WEAT scores as a regularizer, as shown in Equation 1. Here, λ_w is a hyper-parameter, and W_M, W_R, W_C, W_S are the WEAT scores for each category. We hypothesize that, even if a model is able to minimize WEAT effects, the model will remain significantly biased.

$$L = L_{MC} + \lambda_w(W_M + W_R + W_C + W_S) \quad (1)$$

4.1 Results

We measure the accuracy of our fine-tuned models on SWAG and find that the debiased model exhibits competitive accuracy. The WEAT-regularized model achieves 82.2% accuracy, compared to 82.8% for a human (Zellers et al., 2018) and 83.3% for the best RoBERTa-base model.

The results from the WEAT regularization are in Table 3. Table 3 shows that fine-tuning with SWAG alone (without any bias regularizers) yields significant bias toward *male* and *African-American* SEQ tests (8/8 attribute tests show significance), and *female* and *African-American* for PN tests (4/8 attribute tests show significance). Furtheremore, we find that even though our "de-biased" model shows ≈ 0 effect for WEAT, Table 3 shows that this model remains significantly biased on both the SEQ and PN tests. De-biasing with WEAT has exaggerated gender bias for the PN test compared to the SWAG-only model, whereas for the SEQ tests the bias has been flipped to being significantly biased towards *female*. Tests for racial bias are likewise reflective of this trend. These results demonstrate that the WEAT is an insufficient measure of bias. Neutralizing word-piece embeddings does not remove the contextual aspect of bias learned by RoBERTa and may even exacerbate biases.

4.2 Discussion

Our results demonstrate that bias is a significant problem for nearly all pre-trained models. Unfortunately, the problem is not simply solved by using larger networks or more data. As shown in Tables 1 & 2, the approach with the most data, RoBERTa, is among the most consistently biased transformers in our study, while the largest model, GPT-2 XLarge, exhibits greater bias than GPT-2 Base. Tokenization also has an immense impact on the equitable use of language models, and is often overlooked

within discourse surrounding bias. We encourage the community to consider these effects on minority communities whose names or vernacular will be distorted more than majority communities due to the nature of word-piece tokenization.

Developing tests that can contextually identify bias within transformers remains vital. Our "debiasing" results show that relying on ill-fitting tests can lead to harmful false positives. We show that "successfully" de-biasing a model via a WEAT regularizer results in continued or even amplified bias on both the SEQ and PN tests, despite that near-zero WEAT effects. We conclude that contextually- and globally-sensitive bias tests are needed for future debiasing research, as mitigating bias according to WEAT fails to truly neutralize pre-trained transformer models.

5 Conclusion

We systematically quantify bias in commonly used pre-trained transformers, presenting a unified view of bias in the form of gender and racial likelihoods across a range of popular pre-trained transformers. We analyze factors influencing bias in transformers using three tests, SEQ, PN, and $WEAT$, and demonstrate the inadequacies of word-embedding neutralization for contextual models. We call for future work to develop robust bias tests and carefully consider the ramifications of design choices.

Ethics & Impact Statement

Our work targets the subject of inherent, societal biases captured by large pre-trained transformer models which are publicly available and widely used. Our results indicate that bias is a significant problem for the community to tackle, and that all pre-trained models currently exhibit some form of biased prediction of gendered or racial tokens in otherwise neutral contexts.

Beneficiaries – Our work seeks to clarify the ways in which commonly used pre-trained transformers exhibit biases. Practitioners building on the power of pre-trained transformers would benefit from knowing, the inherent biases of each model, and thereby taking appropriate steps to ensure that their downstream task is as neutralized as possible. Further, we hope to contribute knowledge which will eventually make all NLP systems more equitable for all people.

Negatively affected parties – Our work does not investigate bias in many other areas, from racial groups outside of *European-American/African-American* to religious biases or any other inappropriate societal prejudices. Unfortunately, there are few widely-accepted target-set identifiers for NLP research into these biases, and even those which do exist may be poor predictors of underlying demographics (such as the use of first names for racial categorization).

Limitations in scope – As discussed above, our work omits investigations into groups which lack widely-accepted target sets (identifying nouns or pronouns). Even for target sets which do exist, such as *Male/Female*, target sets may be imperfect. For example, many gendered target sets use first names as identifiers, even though there is no gender inherently tied to a name.

Acknowledgments

This work was supported by Georgia Institute of Technology state funding.

References

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. 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 229–237, New Orleans, Louisiana. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

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. In *Advances in neural information processing systems*, pages 4349–4357.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4758–4781.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611*.

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

Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors. 2019. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. 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, Florence, Italy. Association for Computational Linguistics.

Ben Gilburt. 2019. Gender bias in gpt-2.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv: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. *arXiv preprint arXiv:2010.00133*.

Pandu Nayak. 2019. Understanding searches better than ever before.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew Silva, Rohit Chopra, and Matthew Gombolay. 2020. Using cross-loss influence functions to explain deep network representations. *arXiv preprint arXiv:2012.01685*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13209–13220.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

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

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.