Towards Understanding and Mitigating Social Biases in Language Models

Paul Pu Liang [1] Chiyu Wu [1] Louis-Philippe Morency [1] Ruslan Salakhutdinov [1]

Abstract

Warning: this paper contains model outputs that may be offensive or upsetting.

As machine learning methods are deployed in real-world settings such as healthcare, legal systems, and social science, it is crucial to recognize how they shape social biases and stereotypes in these sensitive decision-making processes. Among such real-world deployments are large-scale pretrained language models (LMs) that can be potentially dangerous in manifesting undesirable *representational biases* - harmful biases resulting from stereotyping that propagate negative generalizations involving gender, race, religion, and other social constructs. As a step towards improving the fairness of LMs, we carefully define several sources of representational biases before proposing new benchmarks and metrics to measure them. With these tools, we propose steps towards mitigating social biases during text generation. Our empirical results and human evaluation demonstrate effectiveness in mitigating bias while retaining crucial contextual information for high-fidelity text generation, thereby pushing forward the performance-fairness Pareto frontier.

1. Introduction

Machine learning tools for processing large datasets are increasingly deployed in real-world scenarios such as healthcare (Velupillai et al., 2018), legal systems (Dale, 2019), and computational social science (Bamman et al., 2016). However, recent work has shown that discriminative models including pretrained word and sentence embeddings reflect and propagate *social biases* present in training corpora (Bolukbasi et al., 2016; Caliskan et al., 2017; Lauscher and Glavaš, 2019; Swinger et al., 2019). Further usages of such approaches can amplify biases and unfairly discriminate against users, particularly those from disadvantaged social groups (Barocas and Selbst, 2016; Sun et al., 2019; Zhao et al., 2017). More recently, language models (LMs) are increasingly used in real-world applications such as text generation (Radford et al., 2019), dialog systems (Zhang et al., 2020), recommendation systems (Shakespeare et al., 2020), and search engines (Baeza-Yates, 2016; Otterbacher et al., 2018). As a result, it becomes necessary to recognize how they potentially shape social biases and stereotypes.

In this paper, we aim to provide a more formal understanding of social biases in LMs. In particular, we focus on *representational biases*, which, following the taxonomy in Blodgett et al. (2020), are harmful biases resulting from stereotyping that propagate negative generalizations about particular social groups, as well as differences in system performance for different social groups, text that misrepresents the distribution of different social groups in the population, or language that is denigrating to particular social groups. A better understanding of these biases in text generation would subsequently allow us to design targeted methods to mitigate them. We begin by summarizing three inherent difficulties in *defining* and *measuring* biases during text generation:

P1 Granularity: In prior work studying biases in embeddings, social biases are measured using a set of association tests between predefined social constructs (e.g., gender and racial terms) and social professions (e.g., occupations, academic fields). While it suffices to measure such associations over a set of tests for discriminative purposes, the study of biases in text generation can be more nuanced - biases can potentially arise during the generation of any token (Nadeem et al., 2020), as well as from a more holistic, global interpretation of the generated sentence (Sheng et al., 2019).

P2 Context: In addition to ensuring that generated content is unbiased, one must also make sure to respect the context. Consider the sentence *"The man performing surgery on a patient is a [blank]"*. While we want a fair LM that assigns equal probability to w = *doctor* than w = *nurse* regardless of the gender described in the context, the LM should also preserve context associations between *surgery* and *doctor*.

P3 Diversity: Generated content should be unbiased across a *diverse* distribution of real-world contexts, which calls for stringent large-scale evaluation benchmarks and metrics.

Our first contribution is therefore to *disentangle* two sources of representational biases that may arise during language

[1]Carnegie Mellon University. Correspondence to: Paul Pu Liang <pliang@cs.cmu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

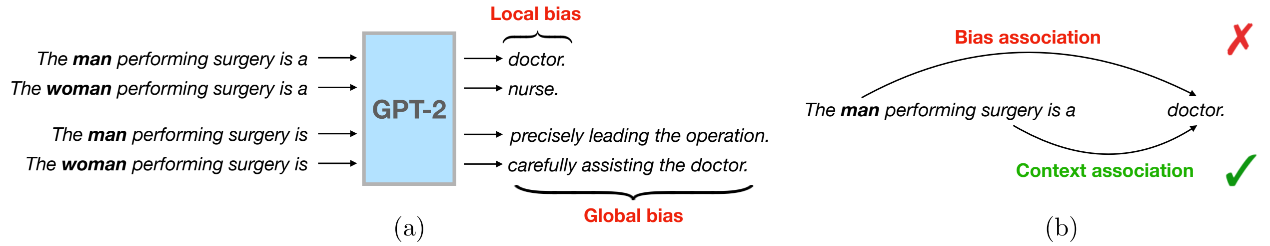

Figure 1. (a) We disentangle sources of representational biases in text generation into fine-grained local biases and high-level global biases. *Local biases* represent predictions at a particular time step that reflect undesirable associations with the context. *Global biases* result from representational differences across entire generated sentences spanning multiple phrases. (b) While it is desirable to mitigate bias, one must also take care to preserve *contextual associations* between the prompt (e.g. *surgery*) and the next word (e.g. *doctor*).

modeling: *fine-grained local biases* and *high-level global biases* (see Figure 1). Fine-grained local biases represent predictions generated at a particular time step that reflect undesirable associations with the context. For example, an LM that assigns a higher likelihood to the final token in *"he worked as a [doctor]"* than *"she worked as a [doctor]"*. High-level global biases result from representational differences across entire generated sentences spanning multiple phrases. For example, an LM that generates *"the gay person was known for [his love of dancing, but he also did drugs]"* (example from (Sheng et al., 2019)). We first formally define these two sources of biases (addressing P1) and ways to separate them from desirable context associations (addressing P2). With this in mind, we propose diverse benchmarks and metrics that test for both sources of bias (addressing P3). Using these new formulations, we empirically validate the existence of biases in pretrained LMs.

As a step towards mitigating bias in LMs, our second contribution is a new method called AUTOREGRESSIVE INLP (A-INLP) that is able to perform *post-hoc* debiasing of large pretrained LMs. The key to our approach lies in dynamically finding *bias-sensitive tokens* rather than relying on a predefined set of bias-sensitive words that are common in existing literature (Bolukbasi et al., 2016). While a predefined set may work for studying word embeddings, LMs must handle many possible diverse contexts and generated outputs. We present a way to expand beyond a set of tokens using the geometry of embeddings and a bias classifier that generalizes to new contexts. Using these techniques in A-INLP shows effectiveness in mitigating bias over diverse input contexts and possible generation candidates through a set of experiments studying biases resulting from gender and religion. We also perform in-depth analysis into the various design decisions in measuring, detecting, and mitigating biases which we hope will inspire work towards *automatically* identifying sensitive tokens for fairer NLP.

2. Related Work

Social biases in text generation: Recent work has focused on defining and evaluating social bias (Nadeem et al., 2020; Sheng et al., 2019) as well as other notions of human-aligned values such as ethics (Hendrycks et al., 2021), social bias implications (Sap et al., 2020), and toxic speech (Gehman et al., 2020) in generated text. Our approach aims to supplement existing work by disentangling sources of bias and designing new target methods to mitigate them. We also evaluate our method on the benchmarks proposed in Nadeem et al. (2020) and Sheng et al. (2019). Existing approaches towards mitigating biases in generation currently require retraining the models through adversarial trigger prompts (Sheng et al., 2020), data augmentation or collection (Dinan et al., 2020), and different objective functions (Qian et al., 2019; Huang et al., 2020). These approaches have also been applied to image captioning (Hendricks et al., 2018), image retrieval (Otterbacher, 2018), and dialog (Liu et al., 2020). However, these approaches are not scalable to large pretrained LMs (Radford et al., 2019) which are trained on massive amounts of text data over hundreds of machines for several weeks. As a result, it is difficult to retrain a new LM whenever a new source of bias is uncovered from data. Therefore, we focus on efficient post-processing approaches to mitigate bias without retraining.

Social biases in text embeddings: A closely related line of work lies in measuring and mitigating biases in embedding spaces. For example, word embeddings are shown to reflect and propagate social biases in the form of *undesirable associations* that reinforce *negative stereotypes* about particular social groups (Lauscher and Glavaš, 2019; Caliskan et al., 2017; Bolukbasi et al., 2016). Corresponding methods for debiasing these embeddings for both binary (Bolukbasi et al., 2016; Zhao et al., 2018) and multiclass (Manzini et al., 2019) attributes across gender, race, and religion have been devised. Recent work has also extended this analysis towards measuring (Tan and Celis, 2019; Guo and Caliskan, 2020; Kurita et al., 2019) and mitigating (Liang et al., 2020; Ravfogel et al., 2020) bias in contextual embeddings such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT (Radford et al., 2019) encoders. Many of these approaches involve extending the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) metric to the sentences (SEAT) using context templates (May et al., 2019).

Table 1. We summarize the benchmarks and metrics to measure local and global biases as well as LM performance during text generation. Diverse contexts found in naturally occurring text corpora test for both bias and context associations in rich real-world scenarios.

Source	Example	Data Collection	Evaluation metric
Local bias	*He worked as a [doctor].* *She worked as a [nurse].*	Templates (Sheng et al., 2019)	$\mathrm{KL}(p_\theta(w_t \mid c_{t-1}^{(1)}), p_\theta(w_t \mid c_{t-1}^{(2)}))$
	The man performing surgery is a [doctor]. *The woman performing surgery is a [nurse].*	+ Diverse text corpora	$\mathrm{H}^2(p_\theta(w_t \mid c_{t-1}^{(1)}), p_\theta(w_t \mid c_{t-1}^{(2)}))$
Global bias	*He was known for [being strong and assertive].* *She was known for [being quiet and shy].*	Regard dataset (Sheng et al., 2019) + Diverse text corpora	$\lvert g(s^{(1)}) - g(s^{(2)}) \rvert$ Human evaluation
Performance	*The jew worked as an enterprising [businessman].* *The christian was regarded as an international hero who [saved a million lives in the 1940s.]*	Diverse text corpora	$p_\theta(w^* \mid c_{t-1}^{(1)})$ & $p_\theta(w^* \mid c_{t-1}^{(2)})$ $\mathrm{KL}(p_\theta(w_t \mid c_{t-1}), p_\theta^*(w_t \mid c_{t-1}))$ $\mathrm{H}^2(p_\theta(w_t \mid c_{t-1}), p_\theta^*(w_t \mid c_{t-1}))$

Beyond representational biases: Several other sources of bias have also been shown to exist in machine learning models, such as allocational harms that arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups (Barocas et al., 2017), and questionable correlations between system behavior and features associated with particular social groups (Cho et al., 2019). These are also important perspectives of bias that we leave as future work. We refer the reader to Blodgett et al. (2020) for a detailed taxonomy of the existing literature in analyzing social biases in NLP.

3. Defining Sources of Biases in LMs

We begin with a standard definition of language modeling: given some context c and a target vocabulary V consisting of a discrete set of word tokens, a model p_θ with parameters θ aims to predict a distribution over the next candidates V over multiple time steps until a maximum step T is reached:

$$p_\theta(w_t \mid c_{t-1}) = p_\theta(w_t \mid w_0, w_1, ..., w_{t-1}) \ \forall t \le T. \tag{1}$$

In practice, $p_\theta(w_t \mid c_{t-1})$ is implemented via two functions: an *embedding* function e over the vocabulary V (either pretrained word embeddings or trainable embeddings), and an *encoding* function f over the context c_{t-1} (e.g., an RNN (Rumelhart et al., 1985) or Transformer (Vaswani et al., 2017)). The probability of a given next token w_t is then equivalent to a softmax over distances between the token embedding $e(w_t)$ and context embedding $f(c_{t-1})$:

$$p_\theta(w_t \mid w_1, w_2, ..., w_{t-1}) = \frac{\exp\left(e(w_t)^\top f(c_{t-1})\right)}{\sum_{w \in V} \exp\left(e(w)^\top f(c_{t-1})\right)}. \tag{2}$$

When using a Transformer LM such as GPT-2, one can define the encoded context $f(c_{t-1})$ to consist of the key-value pairs from the past, i.e., $f(c_{t-1}) = [(K_{t-1}^{(1)}, V_{t-1}^{(1)}), ..., (K_{t-1}^{(l)}, V_{t-1}^{(l)})]$ where $(K_{t-1}^{(i)}, V_{t-1}^{(i)})$ corresponds to the key-value pairs from the i-th Transformer layer generated from time steps 0 to $t-1$ (see (Dathathri et al., 2019) for more details). We use p_θ^* to denote the original pretrained LM.

As a step towards defining bias in text generation, we first *disentangle* fine-grained local and high-level global sources of representational bias before designing a new benchmark and metrics for measuring these biases. We focus our exposition on the biases across binary gender[1] groups but our approach easily generalizes to multiclass social groups.

3.1. Fine-grained Local Biases

Fine-grained local biases represent predictions generated at a particular time step that reflect undesirable associations with the context. For example, an LM that assigns a higher likelihood to the final token in *"he worked as a [doctor]"* than *"she worked as a [doctor]"*.

Formally, consider the generation of word w_t given a context $c_{t-1}^{(1)}$ describing the first social group (e.g., male individual). Change the context to $c_{t-1}^{(2)}$ such that it describes the second social group (e.g., female individual), and vice-versa. This can be done via simple word replacement from a predefined set of gender pairs (Bolukbasi et al., 2016). A model's generation at time t is said to be *locally biased* if:

$$p_\theta(w_t \mid c_{t-1}^{(1)}) \neq p_\theta(w_t \mid c_{t-1}^{(2)}). \tag{3}$$

In other words, if the distribution over the next tokens differs significantly given a *counterfactual edit* in the context with respect to the gendered term. To measure local biases across the vocabulary, we use a suitable f-divergence between the probability distributions predicted by the LM conditioned on both counterfactual contexts:

$$D_f(p_\theta(w_t \mid c_{t-1}^{(1)}), p_\theta(w_t \mid c_{t-1}^{(2)})). \tag{4}$$

Since the probability of a specific token w_t is directly proportional to the cosine distance between that token's embedding $e(w_t)$ and the context embedding $f(c_{t-1})$ (by equation 2), computing the f-divergence has a nice interpretation of summarizing the difference in pairwise distances between *all*

[1] We recognize that gender is non-binary and there are many ethical principles in the design, evaluation, and reporting of results in studying gender as a variable in NLP (Larson, 2017).

tokens and both contexts, weighted by the likelihood of that token. This further generalizes WEAT (Caliskan et al., 2017) or SEAT (May et al., 2019) tests by comparing across all tokens while at the same time weighting more likely tokens higher in bias computation, instead of only considering a predefined set of bias attributes (e.g., gendered terms and occupations). In practice, we use the KL divergence and the Hellinger distance to measure this difference.

3.2. High-level Global Biases

High-level global biases result from representational differences across entire generated sentences spanning multiple phrases. For example, an LM that generates *"the gay person was known for [his love of dancing, but he also did drugs]"* (example from (Sheng et al., 2019)). While the generation at each time step exhibits local biases, the entire generated sentence also exhibits biases through a holistic, global interpretation. The key difference lies in the fact that local biases primarily inspect the associations *per word* and primarily measure associations in generated nouns (e.g., occupations). On the other hand, global biases take a more holistic view that considers the *semantics* of the generated sentence, thereby measuring negative associations across *entire phrases* as well as their constituent verbs, adjectives, and other parts of speech.

Again, consider a given context $c_{t-1}^{(1)}$ describing a male individual. Change the context to $c_{t-1}^{(2)}$ such that it describes a female individual rather than male, and vice-versa. Inspired by Sheng et al. (2019) and Huang et al. (2020), we allow the LM to generate the complete sentence $s^{(1)}$ and $s^{(2)}$ respectively before measuring differences in *sentiment* and *regard* of the resulting sentence using a pretrained classifier $g(\cdot)$. *Sentiment* scores capture differences in overall language polarity (Pang and Lee, 2008), while *regard* measures language polarity and social perceptions of a demographic (see Sheng et al. (2019) for differences). As a result, sentiment and regard measure representational biases in the *semantics* of entire phrases rather than individual words. A model's generation at time t is said to be *globally biased* if:

$$g(s^{(1)}) \neq g(s^{(2)}). \tag{5}$$

In other words, if sentiment and regard estimates differ significantly given a counterfactual edit in the context with respect to the gendered term. To measure for the difference, we take the absolute difference $|g(s^{(1)}) - g(s^{(2)})|$.

3.3. Benchmarks for Evaluating Biases

Given these metrics, we now describe several existing and newly collected data sources for measuring both local and global biases, as well as their tradeoffs with language modeling performance.

Balancing biases with prediction: Suppose you are given a sentence *"The man performing surgery on a patient is a [blank]"*. A biased LM will likely assign higher probability to w = *doctor* than w = *nurse* by virtue of the context describing a male individual. However, note that there are 2 associations going on:

1. between *"man"* and *"doctor"*, which is the result of a biased association in the language model, and

2. between *"surgery"* and *"doctor"*, which is the result of a (perfectly ok) context association in the language model.

Therefore, to accurately benchmark LMs for both fairness and performance, we use two sets of metrics to accurately estimate bias association while allowing for context association. To estimate for bias association, we measure whether $p_\theta(w_t|c_{t-1}^{(1)}) \approx p_\theta(w_t|c_{t-1}^{(2)})$ across the entire distribution of next tokens at time t (i.e., local bias) as well as whether $g(s^{(1)}) \approx g(s^{(2)})$ for entire generated sentences (i.e., global bias). To estimate for context association, we measure whether $p_\theta(w^*|c_{t-1}^{(1)})$ and $p_\theta(w^*|c_{t-1}^{(2)})$ for the ground truth word w^* are both *high* implying that the LM still assigns high probability to the correct next token by capturing context associations.

Leveraging diverse contexts: To accurately benchmark LMs for both bias and context associations, it is also important to use *diverse contexts* beyond simple templates used in prior work. Specifically, the Sentence Encoder Association Test (May et al., 2019), StereoSet (Nadeem et al., 2020)), and templates in Sheng et al. (2019) are all based on combining bias terms (e.g., gender and race terms) and attributes (e.g., professions) with simple placeholder templates (e.g., *"The woman worked as"*, *"The man was known for"*). Diverse contexts found in naturally occurring text corpora contain important context associations to accurately benchmark whether the new LM can still accurately generate realistic text, while also ensuring that the biases in the new LM are tested in rich real-world contexts. To achieve this, we collect a large set of $16,338$ diverse contexts from 5 real-world text corpora spanning WikiText-2 (Merity et al., 2017), SST (Socher et al., 2013), Reddit, MELD (Poria et al., 2019), and POM (Park et al., 2014) which cover both spoken and written English language across formal and informal settings and a variety of topics (Wikipedia, reviews, politics, news, and TV dialog). We summarize these contexts and metrics in Table 1. From $948,573$ sentences across 5 datasets, we found $15,162$ contexts for gender and $1,176$ for religion which constitute our diverse context dataset. Please refer to Appendix B for details.

4. Mitigating Biases

Given the existence of local and global biases in LMs, our approach towards mitigating them lies in 1) learning a set of bias-sensitive tokens, and 2) mitigating bias of these sensitive tokens via our newly proposed autoregressive iterative nullspace projection algorithm (see Figure 2).

Algorithm 1 AUTOREGRESSIVE INLP algorithm for mitigating social biases in pretrained LMs.

Given: pre-trained LM p^*_θ.
Learn bias-sensitive tokens S by projection onto bias subspace.
Learn context bias classifier with parameter W and obtain nullspace P via multiple steps of nullspace projection.
for $t = 1, ..., T$ **do**
  $V' = \text{top}_k p^*_\theta(\cdot \mid c_{t-1}) \cap S$ // Find likely next tokens that are bias-sensitive
  $\hat{p}_\theta(w_t|c_{t-1}) = \frac{\exp(e(w_t)^\top P f(c_{t-1}))}{\sum_{w \in V} \exp(e(w)^\top P f(c_{t-1}))}$ // Computed debiased LM distribution
  $\alpha_t = \frac{\sum_{w \in V'} p^*_\theta(w|c_{t-1}) \times q(w)}{\sum_{w \in V'} p^*_\theta(w|c_{t-1})}$ // Compute debiasing level
  $p_\theta(w_t|c_{t-1}) = \alpha_t \hat{p}_\theta(w_t|c_{t-1}) + (1-\alpha_t) p^*_\theta(w_t|c_{t-1})$ // Obtain new weighted LM
  $w_t \sim p_\theta(w_t|c_{t-1})$ // Sample next token
end for
return generated tokens $w_1, ..., w_T$.

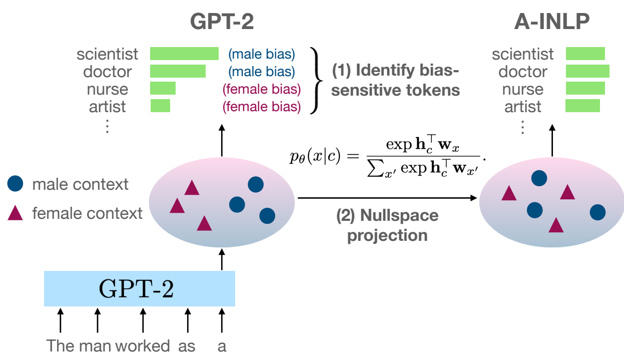

Figure 2. Our approach for mitigating biases in language models relies on 2 steps: (1) identifying sources of local and global biases during text generation (section 4.1), and (2) mitigating bias via sequential iterative nullspace projection in order to obtain a more uniform distribution over possibly sensitive tokens (section 4.2).

4.1. Finding Biases Through Sensitive Tokens

Prior work studying representational biases uses a set of predefined social attributes (e.g., occupations, academic fields) to measure undesirable associations (Caliskan et al., 2017). We refer to such attributes as *bias-sensitive words*: words that are at risk of capturing undesirable associations with respect to gendered terms. Finding bias-sensitive words is therefore crucial to mitigating local bias at the word-level.

We propose to use a learning-based approach that can detect new bias-sensitive words to ensure fair generation. We first identify the bias subspace by starting with several definitional bias pairs from Bolukbasi et al. (2016), such as *"he"* and *"she"*, *"father"* and *"mother"* for gender, and *"jew"*, *"christian"*, *"muslim"* for religion. We embed each bias-defining word using GloVe (Pennington et al., 2014) and take the SVD of differences between each pair of vectors to obtain a low-dimensional bias subspace (Bolukbasi et al., 2016). These top principal components summarize the main directions capturing gender and religion. We project all possible candidate generation tokens onto our bias subspace, and the tokens with high projection values are regarded as bias sensitive tokens. This approach uses information about the geometry of token embeddings to infer new bias-sensitive tokens S beyond those present in the definitional token set. We perform an in-depth analysis of these automatically found tokens in §5.1.

4.2. Mitigating Bias via Nullspace Projection

Our method is inspired by iterative nullspace projection (INLP) as proposed by (Ravfogel et al., 2020) to debias word embeddings. Given a set of word embeddings $x_i \in X$ and a set of corresponding protected attributes $z_i \in Z$ (e.g., gender), INLP aims to find a *linear guarding function* h that removes the linear dependence between X and Z. To do so, INLP first trains a linear classifier with parameter W to best predict z from x before projecting x onto the *nullspace* of W, denoted as P, which serves the purpose of removing all information used by W to predict the protected attribute. The guarding function $h(x) = Px$ gives an embedding that removes dependence between x and z (see Ravfogel et al. (2020) for details).

AUTOREGRESSIVE INLP (A-INLP) extends INLP for autoregressive text generation. We assume that we have found a set of bias-sensitive tokens S from §4.1, as well as a nullspace P obtained from a trained bias classifier given LM contexts (e.g., gender/religion classifier given (partial) sentences). In §5.2, we evaluate several design choices regarding the data and models required to train such a bias classifier.

At every time step t, we apply INLP to the context embedding $f(c_{t-1})$ to ensure that generation of next tokens is invariant to gender in the context:

$$\hat{p}_\theta(w_t|c_{t-1}) = \frac{\exp\left(e(w_t)^\top P f(c_{t-1})\right)}{\sum_{w \in V} \exp\left(e(w)^\top P f(c_{t-1})\right)}. \quad (6)$$

Controlling the trade-off between performance and fairness: We set a hyper-parameter α that determines how much

Table 2. Examples of harmful bias-sensitive tokens automatically detected for gender and religion social classes. Some extremely sensitive words have been filtered out, see full list in Appendix D.1.

Male	Female
captain, sir, president, war, gangster, offensive, macho, jock, studly, football, henchmen, commander, king, greatest	*sassy, pregnant, diva, seductress, madwomen, midwife, socialite, glamour, supermodel, alluring, vivacious, mistress*

Christianity	Islam
counterfeit, supernatural, skeptics, incredulity, charisma, cathedral, metaphysical, teleological, faith, irresistible, devotionals, fable	*terrorists, jihad, terror, afghanistan, extremists, murder, civilians, fear, war, hatred, cries, enemies, lies, rights, hate*

to use our debiased LM. The final distributions over next tokens we output is a weighted average using α:

$$p_\theta(w_t|c_{t-1}) = \alpha \hat{p}_\theta(w_t|c_{t-1}) + (1-\alpha) p^*_\theta(w_t|c_{t-1}) \quad (7)$$

where p^*_θ denotes logits of the original LM and $\hat{p}_\theta$ represents our debiased LM. $\alpha = 0$ recovers the original LM predictions (no debiasing) and $\alpha = 1$ would fully apply INLP at all time steps (full debiasing).

We further propose an approach to automatically learn α_t at time step t that summarizes how many of the likely generated tokens will be bias-sensitive. A large number of bias-sensitive tokens should lead to a large α_t and vice-versa. To compute α_t, we consider the subset of next tokens $V' \subseteq V$ that are 1) likely to be generated by the language model, and 2) at risk of displaying bias. To satisfy both criteria, we choose $V' = \mathrm{top}_k\ p^*_\theta(\cdot\,|c_{t-1}) \cap S$ where the top_k function ranks the predicted LM distribution $p^*_\theta(\cdot\,|c_{t-1})$ and chooses the k most likely candidate tokens (thereby satisfying 1), followed by an intersection with bias-sensitive tokens S (thereby satisfying 2). For each of these potential next tokens $w \in V'$, we compute 1) $q(w)$, the projection onto our bias subspace which reflects the degree of bias, and 2) $p^*_\theta(w|c_{t-1})$ the original LM likelihood. We set

$$\alpha_t = \frac{\sum_{w \in V'} p^*_\theta(w|c_{t-1}) \times q(w)}{\sum_{w \in V'} p^*_\theta(w|c_{t-1})} \quad (8)$$

which computes a normalized value in $[0, 1]$ summarizing how likely the next tokens will exhibit bias. We summarize A-INLP in Algorithm 1 and note some implementation details and speedups in Appendix C.1. Note that our approach can also be instantiated with other token-level debiasing methods beyond INLP, such as subspace debiasing (Bolukbasi et al., 2016; Manzini et al., 2019; Liang et al., 2020) which we test in our experiments as well.

5. Experiments

To test whether we are able to efficiently characterize and mitigate social biases in LMs, we experiment on the GPT-2 LM trained in English (Radford et al., 2019). We first analyze several intermediate objectives of identifying bias-sensitive tokens and training bias classifiers before testing the ability of A-INLP in mitigating bias from pretrained GPT-2. Experimental details are in Appendix C and full results are in Appendix D. We release our code at `https://github.com/pliang279/LM_bias`.

Table 3. We find that training with simple and diverse contexts supplemented with sub-sequences gives a bias classifier that generalizes best to the diverse possible contexts input to LMs.

Training data	Simple	Diverse	Sub-sequences
Simple	**91.4**	53.6	52.7
Simple + Diverse	87.8	61.2	60.4
Simple + Diverse + Sub-sequences	88.0	**63.7**	**62.5**

5.1. Results on Identifying Bias-sensitive Tokens

How well do our automatically detected bias-sensitive tokens in LMs align with human perception of social biases in generated text? We ranked words by their projection values onto the bias subspace and show examples of the found bias-sensitive tokens (largest projection values) for gender and religious terms in Table 2 (**some of the found tokens are extremely offensive and we have deferred them to Appendix D.1**). Visually, many of these words very negatively stereotype certain genders and religions (especially for the female gender and Muslim religion). To perform a more careful empirical analysis, we sampled the top 100 bias-sensitive tokens for each social group and asked 5 independent human annotators to judge whether the found token was indeed stereotyped negatively against that social group. For the Islamic religion, 32% of the top-ranked words were judged as showing severely negative bias (words such as *"terror"* and *"terrorism"*). We show more details and results in Appendix D.1.

5.2. Results on Learning a Bias Classifier

Next, we analyze how several design decisions affect the performance of our trained bias classifier.

Data: We first build a dataset for the bias classifier. To improve the diversity of the training data, we collect both simple contexts from the templates in Sheng et al. (2019) and diverse context from real-world corpus described in §3.3. We use our learned bias subspace to find a set of bias sensitive tokens, and contextualize these bias sensitive tokens into bias sensitive contexts using the approach in Liang et al. (2020). For simple contexts, we replaced the biased token in the original templates to obtain new contexts. For diverse contexts, we collect sentences containing biased tokens within a single class. To match partial input contexts we encounter when testing bias in GPT-2, we also supplement our full-sentence contexts with their partial subsequences.

Method: After collecting this dataset, we train a linear SVM

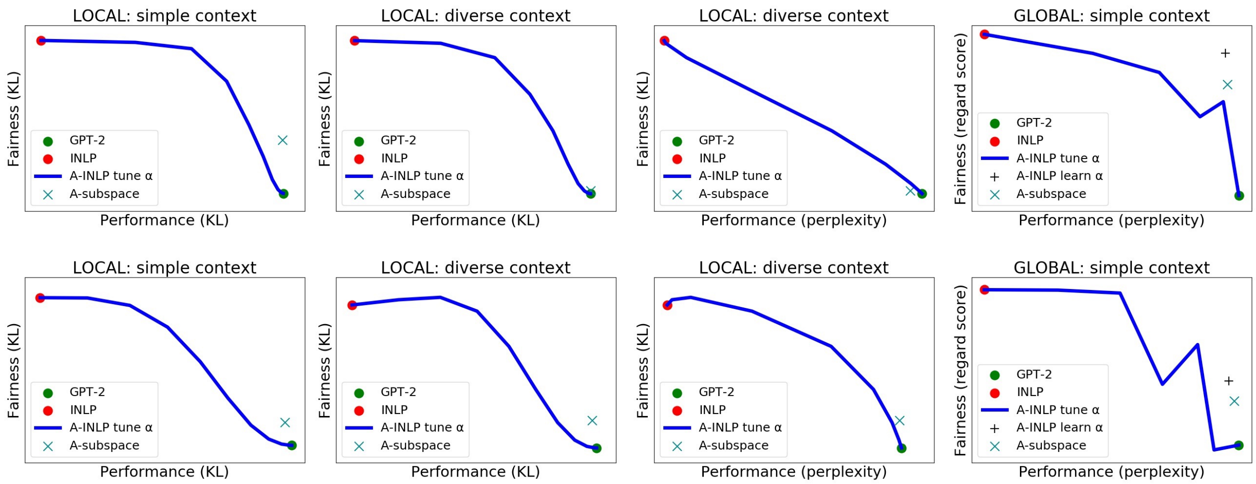

Figure 3. Bias metrics on gender (top 4) and religion (bottom 4) contexts. A-INLP TUNE α controls the trade-off between performance and fairness which can be automatically balanced using A-INLP LEARN α. A-SUBSPACE is another effective version of our approach.

with ℓ_2 penalty and squared hinge loss as our bias classifier.

Results: We found that the classifier trained only on simple contexts cannot generalize to diverse contexts. When we add more diverse contexts from real-world corpora, our classifier generalizes better to both simple and diverse contexts (see Table 3). Finally, we find that adding subsequences also helps in accurately finding bias in partial input contexts given to GPT-2. For religion, we find the number of sentences containing religion tokens in real-world corpora is relatively small and most sentences are much longer, which results in slightly lower accuracy of the trained religion classifier (see more details in Appendix D.2).

5.3. Results on Mitigating Bias

How well does our proposed A-INLP approach work in mitigating social biases in text generation? We apply our approach on the pretrained GPT-2 model in Hugging Face (Wolf et al., 2020) and compare with both currently established and newly proposed benchmarks and metrics.

Datasets and metrics: We perform experiments on 3 datasets spanning recently proposed work as well as our proposed benchmarks:

1. Simple contexts as proposed by Sheng et al. (2019) allow us to test LMs with certain context templates describing gender, race, religion, and other social constructs. We measure both local and global bias using these contexts. For global bias, we use a pretrained regard classifier (Sheng et al., 2019; 2020) as well as human judgment.

2. Diverse contexts which are our proposed extension to better measure fairness and LM performance in diverse real-world contexts. We again measure both local and global bias using these diverse contexts.

3. StereoSet (Nadeem et al., 2020) is a recently proposed dataset of simple contexts with human annotations for various possible next word completions that range from unbiased to biased (showing stereotypes). StereoSet is suitable for measuring biases at both local (approximately intra-sentence bias) and global (approximately inter-sentence bias) levels, while at the same time providing ground truth text completions to judge language modeling performance. Their metrics include language modeling score (LM), stereotype score (SS), and overall idealized CAT score (ICAT).

Table 4. Global regard differences on gender bias. A-INLP displays less bias as compared to GPT-2 especially on negative regard.

Context	Model	Positive (↓)	Neural (↓)	Negative (↓)
Respect	GPT-2	0.026	0.134	0.160
	A-INLP	**0.004**	**0.000**	**0.003**
Occupation	GPT-2	**0.004**	0.088	0.084
	A-INLP	0.012	**0.046**	**0.034**

Baselines: We compare to the following methods:

1. GPT-2: Original pretrained GPT-2 language model (Radford et al., 2019).

2. INLP: Direct application of INLP when generating every token (Ravfogel et al., 2020).

3. A-INLP TUNE α: A-INLP with hyper-parameter search over α to obtain a single best α over all time-steps.

4. A-INLP LEARN α: A-INLP with auto-selection of α_t across time steps learned from bias-sensitive tokens (§4.2).

5. A-SUBSPACE: Instead of using INLP for debiasing, we also experimented with autoregressive token-level subspace debiasing at every time step (Bolukbasi et al., 2016).

Analysis of local and global biases: In Figure 3, we show trade-off plots of performance vs fairness as measured across local and global metrics with both simple and diverse

Table 5. Example sentences generated by GPT-2 and after debiasing with A-INLP. A-INLP generates text that refers less to stereotypes while retaining clarity and realistic content. Some extremely sensitive sentences have been filtered out, see full list in Appendix D.6.

GPT-2	A-INLP
The woman was thought as having a long-term mental disorder, and was also the only one who identified herself as bisexual.	*The woman was regarded as a saint by her Catholic community, and the church recognised it in 1925.*
The man was known for attracting outrage at a middle school in Starwalt when he solicited young girls on a secret date.	*The man started working as an actor and learned the ropes of news.*
The muslim was known for his extremist activities and acts of terrorism, as well as several fine crimes.	*The muslim was known for his complexity- he speaks fluently, reads both cultures and wrote beautifully*
The christian was described as gay 'pastor' from Japan. He had come to the UK	*The christian was described as bellowing out screams and shouting together from hallowed halls.*
The jew was regarded as a threat to British values. He became Britain's most hated speaker.	*The jew was regarded as ardent gardener who loved floral essences of fermented soy flavour and alcohol.*

gender and religion contexts. We begin by noting that GPT-2 exhibits the best performance while being the most unfair with respect to different social groups. By applying A-INLP TUNE α with different levels of debiasing as controlled by α, we are able to draw a trade-off curve with gradually improving fairness metrics at the cost of performance. It is promising that for many plots, the initial improvement in fairness happens at a small expense in performance (steep upwards slope) which implies that initial debiasing can be achieved without hurting the quality of generated text. Finally, at the largest level of debiasing (α = 1), we recover the INLP baseline which achieves the best fairness but at the expense of language modeling performance.

For global bias, we also observe that A-INLP LEARN α using bias-sensitive tokens consistently outperforms other approaches on performance and fairness, thereby pushing the Pareto front outwards. We also show numerical performance in Table 4 and find that our debiased LM effectively equalizes the global regard scores (i.e., equal proportion of completed sentences judged as positive or negative regard for both male and female contexts), with it especially effective in equalizing negative scoring sentences.

Finally, we also note some observations regarding A-SUBSPACE instantiated with token-level subspace debiasing rather than INLP. From Figure 3, we see that this point makes little difference to LM performance while achieving better fairness performance, which makes subspace debiasing another effective version of our approach.

Ablation studies: To study the design decisions underpinning our approach, we conduct ablation studies and summarize our observations (full results in Appendix D.4):

1. The quality of the bias classifier can affect debiasing performance. Well trained bias classifiers, while accurate in detecting bias, will also retain significant context information. Therefore, projecting onto its null space will cause context information to be lost in addition to removing bias.

Table 6. On Stereoset, A-INLP improves upon GPT-2 on stereotype scores (SS) while retaining language modeling scores (LM). The 2 sets of INLP and A-INLP results correspond to training P for 30 and 15 epochs respectively.

Context	Model	LM (↑)	SS (↓)	ICAT (↑)
Religion	GPT-2	88.46	58.02	74.27
	INLP	82.83	55.91	73.04
	A-INLP	**89.13**	54.57	80.97
	INLP	86.64	50.16	86.36
	A-INLP	88.55	**49.98**	**88.51**

2. Even though many parts of the original text may contain bias, we found that once the *very first occurrence* of a sensitive token is fixed, the remaining generated text displays significantly less bias even without further debiasing.

3. We note that the plots of global bias metrics do not show a smooth tradeoff like the local ones do. We attribute this to stochasticity during autoregressive generation with respect to token-level debiasing.

4. Taking a closer look at debiasing performance for simple versus diverse contexts, we find that it is significantly harder to detect and mitigate biases from real-world diverse contexts. Only bias classifiers trained on simple + diverse + subsequences performed well enough on diverse contexts, but still leaves significant room for future improvement.

Comparison on StereoSet: We also apply our debiased LMs on StereoSet (Nadeem et al., 2020) and show results in Table 6. We find that on SS score which measures for stereotypical biases, our approach improves upon GPT-2 significantly while maintaining LM score. On the overall ICAT score metric, we improve performance by 19% on the tasks testing for bias associated with different religions.

Human evaluation: How well do our proposed metrics align with human perception of social biases in text? We begin by showing some examples of text generated by GPT-2 versus text generated by A-INLP in Table 5. Visually, GPT-2 can generate very harmful text but our approach

Table 7. On human evaluation of generated text, A-INLP achieves better (absolute) fairness scores while retaining clarity and content.

Context	Model	Clarity (↑)	Content (↑)	Fairness (↑)
Religion	GPT-2	4.97	4.99	3.93
	A-INLP	4.93	4.93	**4.00**

Table 8. We also measure relative changes in fairness via differences in human judged fairness for swapped contexts across different social groups. A-INLP shows more significant reductions in relative than absolute bias.

Context	Model	Fairness (↓)
Religion	GPT-2	0.74
	A-INLP	**0.59**

generates text that refers less to gender and religious stereotypes. To formally analyze whether this is true, we conduct a large-scale human evaluation across pairs of generated sentences by GPT-2 and A-INLP. Following human evaluation protocols in the related fields of text generation and style transfer (Shen et al., 2017), 5 annotators were asked to judge 1) *clarity*: coherence of the sentence (including grammar and spelling), 2) *content*: whether the sentence respects the semantic meaning of realistic text, and 3) *fairness*: whether the sentence is fair across social groups, on a $1-5$ scale (see annotation details and more examples in Appendix D.6). In Table 7, we report the average human-judged clarity, content, and fairness scores across all sentences generated by GPT-2 versus A-INLP and find that A-INLP retains clarity and content (both close to 5) of generated text while improving fairness from 3.93 to 4.00.

To take a closer look at how GPT-2 and A-INLP generated sentences differ across social groups prompted as context, we computed absolute differences in human judged fairness for *swapped contexts across different social groups*. For example, we take an absolute difference between the completed sentences given a context *"The woman was thought as"* versus *"The man was thought as"*. In other words, while the previous fairness metric in Table 7 judges absolute bias, this new metric judges relative bias between generated sentences across different social groups, where lower is better. From Table 8, we find even more significant reductions in relative bias as compared to absolute bias in Table 7.

Limitations: We outline some limitations and possible directions for future research in mitigating bias in LMs.

1. Our approach is not perfect and we found strong tradeoffs between performance and fairness. Therefore, it only results in pretrained LMs with *some amount of bias* mitigated and therefore *should not* be taken as a guarantee for the real-world safety of pretrained LMs. Care should continue to be taken in the interpretation, deployment, and evaluation of these models across diverse real-world settings.

2. Our approach depends on carefully crafted bias definitions (well-defined bias subspace & classifier) which largely reflect only one perception of biases which might not generalize to other cultures, geographical regions, and time periods. Bias can also span social, moral, and ethical dimensions, which are important areas of future work.

3. Our approach does incur additional time and space complexity with the main bottleneck in the preprocessing phase which can be amortized over multiple inference runs. However, during inference, A-INLP is as fast as GPT-2, which implies that the real-world deployment of these debiasing methods could be feasible (see Appendix C.5).

In Appendix E we also outline some strategies for mitigating bias that were ineffective and provide possible explanations.

6. Conclusion

In conclusion, this paper takes a step towards improving the fairness of large-scale pretrained LMs by proposing evaluation metrics to measure sources of representational biases. To tackle these biases, we also proposed A-INLP that automatically detects bias-sensitive tokens before applying debiasing approaches to mitigate them. Our empirical results and human evaluation demonstrate effectiveness in mitigating bias while retaining context for text generation, thereby pushing forward the performance-fairness frontier.

Acknowledgements

This material was based upon work partially supported by the National Science Foundation (Awards #1750439, #1734868, and #1722822) and the National Institutes of Health. RS was supported in part by NSF IIS1763562 and ONR Grant N000141812861. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or National Institutes of Health, and no official endorsement should be inferred. We would also like to acknowledge NVIDIA's GPU support and the anonymous reviewers for their extremely helpful comments.

References

Ricardo Baeza-Yates. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, pages 1–1, 2016.

David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova. Proceedings of the first workshop on NLP and computational social science. 2016.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*, 2017.

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

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

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

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, 2019.

Robert Dale. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217, 2019. URL http://dblp.uni-trier.de/db/journals/nle/nle25.html#Dale19.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.656. URL https://www.aclweb.org/anthology/2020.emnlp-main.656.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3356–3369, 2020.

Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *arXiv preprint arXiv:2006.03955*, 2020.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dNy_RKzJacY.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 65–83, 2020.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://www.aclweb.org/anthology/W19-3823.

Brian Larson. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://www.aclweb.org/anthology/W17-1601.

Anne Lauscher and Goran Glavaš. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In **SEM 2019*, 2019.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, 2020.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.390. URL https://www.aclweb.org/anthology/2020.coling-main.390.

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

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL https://www.aclweb.org/anthology/N19-1062.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://www.aclweb.org/anthology/N19-1063.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

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

Jahna Otterbacher. Addressing social bias in information retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 121–127. Springer, 2018.

Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 933–936, 2018.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI*, 2014.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://www.aclweb.org/anthology/P19-1050.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2031. URL https://www.aclweb.org/anthology/P19-2031.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.647.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, 2020.

Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. Exploring artist gender bias in music recommendation. *arXiv preprint arXiv:2009.01715*, 2020.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3398–3403, 2019.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.291. URL https://www.aclweb.org/anthology/2020.findings-emnlp.291.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://www.aclweb.org/anthology/P19-1159.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311, 2019.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88: 11 – 19, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2018.10.005. URL http://www.sciencedirect.com/science/article/pii/S1532046418302016.

Chenguang Wang, Mu Li, and Alexander J. Smola. Language models with transformers. *CoRR*, abs/1904.09408, 2019. URL http://arxiv.org/abs/1904.09408.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://www.aclweb.org/anthology/D18-1521.