# Towards Debiasing Sentence Representations

**Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim,**
**Ruslan Salakhutdinov, Louis-Philippe Morency**
Machine Learning Department and Language Technologies Institute
Carnegie Mellon University
`pliang@cs.cmu.edu`

## Abstract

As natural language processing methods are increasingly deployed in real-world scenarios such as healthcare, legal systems, and social science, it becomes necessary to recognize the role they potentially play in shaping social biases and stereotypes. Previous work has revealed the presence of *social biases* in widely used word embeddings involving gender, race, religion, and other social constructs. While some methods were proposed to debias these word-level embeddings, there is a need to perform debiasing at the sentence-level given the recent shift towards new contextualized sentence representations such as ELMo and BERT. In this paper, we investigate the presence of social biases in sentence-level representations and propose a new method, SENT-DEBIAS, to reduce these biases. We show that SENT-DEBIAS is effective in removing biases, and at the same time, preserves performance on sentence-level downstream tasks such as sentiment analysis, linguistic acceptability, and natural language understanding. We hope that our work will inspire future research on characterizing and removing social biases from widely adopted sentence representations for fairer NLP.

## 1 Introduction

Machine learning tools for learning from language 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). Key to the success of these models are powerful *embedding* layers which learn continuous representations of input information such as words, sentences, and documents from large amounts of data (Devlin et al., 2019; Mikolov et al., 2013). Although word-level embeddings (Pennington et al., 2014; Mikolov et al., 2013) are highly informative features useful for a variety of tasks in Natural Language Processing (NLP), recent work has shown that word-level embeddings reflect and propagate *social biases* present in training corpora (Lauscher and Glavaš, 2019; Caliskan et al., 2017; Swinger et al., 2019; Bolukbasi et al., 2016). Machine learning systems that incorporate these word embeddings can further amplify biases (Sun et al., 2019b; Zhao et al., 2017; Barocas and Selbst, 2016) and unfairly discriminate against users, particularly those from disadvantaged social groups. Fortunately, researchers working on fairness and ethics in NLP have devised methods towards *debiasing* these word representations for both binary (Bolukbasi et al., 2016) and multiclass (Manzini et al., 2019) bias attributes such as gender, race, and religion.

More recently, sentence-level representations such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT (Radford et al., 2019) have become the preferred choice for text sequence encoding. When compared to word-level representations, these models have achieved better performance on multiple tasks in NLP (Wu and Dredze, 2019), multimodal learning (Zellers et al., 2019; Sun et al., 2019a), and grounded language learning (Urbanek et al., 2019). As their usage proliferates across various real-world applications (Huang et al., 2019; Alsentzer et al., 2019), it becomes necessary to recognize the role they play in shaping social biases and stereotypes.

Debiasing sentence representations is difficult for two reasons. Firstly, it is usually *unfeasible* to fully retrain many of the state-of-the-art sentence-based embedding models. In contrast with conventional word-level embeddings such as GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) which can be retrained on a single machine within a few hours, the best sentence encoders such as BERT (Devlin et al., 2019), and GPT (Radford et al., 2019) 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 sentence encoder whenever a new source of bias is uncovered from data. We therefore focus on *post-hoc* debiasing techniques which add a post-training debiasing step to these sentence representations before they are used in downstream tasks (Bolukbasi et al., 2016; Manzini et al., 2019). Secondly, sentences display large variety in how they are composed from individual words. This variety is driven by many factors such as topics, individuals, settings, and even differences between spoken and written text. As a result, it is difficult to scale traditional word-level debiasing approaches (which involve bias-attribute words such as *man, woman*) (Bolukbasi et al., 2016) to sentences.

**Related Work:** Although there has been some recent work in measuring the presence of bias in sentence representations (May et al., 2019; Basta et al., 2019), none of them have been able to successfully *remove* bias from pretrained sentence representations. In particular, Zhao et al. (2019), Park et al. (2018), and Garg et al. (2019) are not able to perform post-hoc debiasing and require changing the data or underlying word embeddings and retraining which is costly. Bordia and Bowman (2019) only study word-level language models and also requires re-training. Finally, Kurita et al. (2019) only measure bias on BERT by extending the word-level Word Embedding Association Test (WEAT) (Caliskan et al., 2017) metric in a manner similar to May et al. (2019).

In this paper, as a compelling step towards generalizing debiasing methods to sentence representations, we capture the various ways in which bias-attribute words can be used in natural sentences. This is performed by *contextualizing* bias-attribute words using a *diverse* set of sentence templates from various text corpora into bias-attribute sentences. We propose SENT-DEBIAS, an extension of the HARD-DEBIAS method (Bolukbasi et al., 2016), to debias sentences for both binary[1] and multiclass bias attributes spanning gender and religion. Key to our approach is the *contextualization* step in which bias-attribute words are converted into bias-attribute sentences by using a *diverse* set

---

| Binary Gender | Multiclass Religion |
|---|---|
| man, woman | jewish, christian, muslim |
| he, she | torah, bible, quran |
| father, mother | synagogue, church, mosque |
| son, daughter | rabbi, priest, imam |

Table 1: Examples of word pairs to estimate the binary gender bias subspace and the 3-class religion bias subspace in our experiments.

of sentence templates from text corpora. Our experimental results demonstrate the importance of using a large number of diverse sentence templates when estimating bias subspaces of sentence representations. Our experiments are performed on two widely popular sentence encoders BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), showing that our approach reduces the bias while preserving performance on downstream sequence tasks. We end with a discussion about possible shortcomings and present some directions for future work towards accurately characterizing and removing social biases from sentence representations for fairer NLP.

## 2 Debiasing Sentence Representations

Our proposed method for debiasing sentence representations, SENT-DEBIAS, consists of four steps: 1) *defining* the words which exhibit bias attributes, 2) *contextualizing* these words into bias attribute sentences and subsequently their sentence representations, 3) *estimating* the sentence representation bias subspace, and finally 4) *debiasing* general sentences by removing the projection onto this bias subspace. We summarize these steps in Algorithm 1 and describe the algorithmic details in the following subsections.

**1) Defining Bias Attributes:** The first step involves *identifying* the bias attributes and defining a set of *bias attribute words* that are indicative of these attributes. For example, when characterizing bias across the male and female genders, we use the word pairs (*man*, *woman*), (*boy*, *girl*) that are indicative of gender. When estimating the 3-class religion subspace across the Jewish, Christian, and Muslim religions, we use the tuples (*Judaism*, *Christianity*, *Islam*), (*Synagogue*, *Church*, *Mosque*). Each tuple should consist of words that have an equivalent meaning except for the bias attribute. In general, for $d$-class bias attributes, the set of words forms a dataset $\mathcal{D} = \{(w_1^{(i)}, ..., w_d^{(i)})\}_{i=1}^m$ of $m$ entries where each entry $(w_1, ..., w_d)$ is a $d$-tuple of words that are each representative of a particular

**Algorithm 1** SENT-DEBIAS: a debiasing algorithm for sentence representations.

---

**SENT-DEBIAS:**

1: Initialize (usually pretrained) sentence encoder $M_\theta$.
2: Define bias attributes (e.g. binary gender $g_m$ and $g_f$).
3: Obtain words $\mathcal{D} = \{(w_1^{(i)}, ..., w_d^{(i)})\}_{i=1}^m$ indicative of bias attributes (e.g. Table 1).
4: $\mathcal{S} = \bigcup_{i=1}^m \text{CONTEXTUALIZE}(w_1^{(i)}, ..., w_d^{(i)}) = \{(s_1^{(i)}, ..., s_d^{(i)})\}_{i=1}^n$     // words into sentences
5: **for** $j \in [d]$ **do**
6:     $\mathcal{R}_j = \{M_\theta(s_j^{(i)})\}_{i=1}^n$                                        // get sentence representations
7: **end for**
8: $\mathbf{V} = \mathbf{PCA}_k\left(\bigcup_{j=1}^d \bigcup_{\mathbf{w} \in \mathcal{R}_j} (\mathbf{w} - \boldsymbol{\mu}_i)\right)$     // compute bias subspace
9: **for** each new sentence representation $\mathbf{h}$ **do**
10:     $\mathbf{h_V} = \sum_{j=1}^k \langle \mathbf{h}, \mathbf{v}_j \rangle \mathbf{v}_j$                              // project onto bias subspace
11:     $\hat{\mathbf{h}} = \mathbf{h} - \mathbf{h_V}$                                      // subtract projection
12: **end for**

---

bias attribute (we drop the superscript $(i)$ when it is clear from the context). Table 1 shows some bias attribute words that we use to estimate the bias subspaces for binary gender and multiclass religious attributes (full pairs and triplets in appendix).

Existing methods that investigate biases tend to operate at the word-level which simplifies the problem since the set of tokens is bounded by the vocabulary size (Bolukbasi et al., 2016). This simple approach has the advantage of identifying the presence of biases using predefined sets of word associations, and estimate the bias subspace using the predefined bias word pairs. On the other hand, the potential number of sentences are unbounded which makes it harder to precisely characterize the sentences in which bias is present or absent. Therefore, it is not trivial to directly convert these words to sentences to obtain a representation from pretrained sentence encoders. In the subsection below, we describe our solution to this problem.

**2) Contextualizing Words into Sentences:** A core step in our SENT-DEBIAS approach involves *contextualizing* the predefined sets of bias attribute words to sentences so that sentence encoders can be applied to obtain sentence representations. One option is to use a simple template-based design to simplify the contextual associations a sentence encoder makes with a given term, similar to how May et al. (2019) proposed to measure (but not remove) bias in sentence representations. For example, each word can be slotted into templates such as "This is <word>.", "I am a <word>.". We take an alternative perspective and hypothesize that for a given bias attribute (e.g. gender), a single bias subspace

exists across all possible sentence representations. For example, the bias subspace should be the same in the sentences "The boy is coding.", "The girl is coding.", "The boys at the playground.", and "The girls at the playground.". In order to estimate this bias subspace accurately, it becomes important to use sentence templates that are as *diverse* as possible to account for all occurrences of that word in surrounding contexts. In our experiments, we empirically demonstrate that estimating the bias subspace using a large and diverse set of templates from text corpora leads to improved bias reduction as compared to using simple templates.

To capture the variety in syntax across sentences, we use large text corpora to find naturally occurring sentences. These naturally occurring sentences therefore become our sentence "templates". To use these templates to generate new sentences, we replace words representing a single class with another. For example, a sentence containing a male term "*he*" is used to generate a new sentence but replacing it with the corresponding female term "*she*". This contextualization process is repeated for all word tuples in the bias attribute word dataset $\mathcal{D}$, eventually contextualizing the given set of bias attribute words into *bias attribute sentences*. Since there are multiple templates which a bias attribute word can map to, the contextualization process results in a *bias attribute sentence dataset $\mathcal{S}$* which is substantially larger in size:

$$\mathcal{S} = \bigcup_{i=1}^m \text{CONTEXTUALIZE}(w_1^{(i)}, ..., w_d^{(i)}) \quad (1)$$

$$= \{(s_1^{(i)}, ..., s_d^{(i)})\}_{i=1}^n, \ |\mathcal{S}| > |\mathcal{D}| \quad (2)$$

| Dataset | Type | Topics | Formality | Length | Samples |
|---|---|---|---|---|---|
| WikiText-2 | written | everything | formal | 24.0 | "the mailing contained information about their history and advised people to read several books, which primarily focused on {*jewish/christian/muslim*} history" |
| SST | written | movie reviews | informal | 19.2 | "{*his/her*} fans walked out muttering words like horrible and terrible, but had so much fun dissing the film that they didn't mind the ticket cost." |
| Reddit | written | politics, electronics, relationships | informal | 13.6 | "roommate cut my hair without my consent, ended up cutting {*himself/herself*} and is threatening to call the police on me" |
| MELD | spoken | comedy TV-series | informal | 8.1 | "that's the kind of strength that I want in the {*man/woman*} I love!" |
| POM | spoken | opinion videos | informal | 16.0 | "and {*his/her*} family is, like, incredibly confused" |

Table 2: Comparison of the various datasets used to find natural sentence templates. Length represents the average length measured by the number of words in a sentence. Words in italics indicate the words used to estimating the binary gender or multiclass religion subspaces, e.g. (*man*, *woman*), (*jewish*, *christian*, *muslim*). This demonstrates the variety in our naturally occurring sentence templates in terms of topics, formality, and spoken/written text.

where $\text{CONTEXTUALIZE}(w_1, ..., w_d)$ is a function which returns a set of sentences obtained by matching words with naturally-occurring sentence templates from text corpora.

Our text corpora originate from the following five sources: 1) **WikiText-2** (Merity et al., 2017a), a dataset of formally written Wikipedia articles (we only use the first 10% of WikiText-2 which we found to be sufficient to capture formally written text), 2) **Stanford Sentiment Treebank** (Socher et al., 2013), a collection of 10000 polarized written movie reviews, 3) **Reddit** data collected from discussion forums related to politics, electronics, and relationships, 4) **MELD** (Poria et al., 2019), a large-scale multimodal multi-party emotional dialog dataset collected from the TV-series Friends, and 5) **POM** (Park et al., 2014), a dataset of spoken review videos collected across 1,000 individuals spanning multiple topics. These datasets have been the subject of recent research in language understanding (Merity et al., 2017b; Liu et al., 2019; Wang et al., 2019) and multimodal human language (Liang et al., 2018, 2019). Table 2 summarizes these datasets. We also give some examples of the diverse templates that occur naturally across various individuals, settings, and in both written and spoken text.

**3) Estimating the Bias Subspace:** Now that we have contextualized all $m$ word $d$-tuples in $\mathcal{D}$ into $n$ sentence $d$-tuples $\mathcal{S}$, we pass these sentences through a pre-trained sentence encoder (e.g. BERT, ELMo) to obtain sentence representations. Suppose we have a pre-trained encoder $M_\theta$ with parameters $\theta$. Define $\mathcal{R}_j, j \in [d]$ as sets that collect all sentence representations of the $j$-th entry in the $d$-tuple, $\mathcal{R}_j = \{M_\theta(s_j^{(i)})\}_{i=1}^n$. Each of these sets $\mathcal{R}_j$ defines a vector space in which a specific bias attribute is present across its contexts. For example, when dealing with binary gender bias, $\mathcal{R}_1$ (likewise $\mathcal{R}_2$) defines the space of sentence representations with a male (likewise female) context. The only difference between the representations in $\mathcal{R}_1$ versus $\mathcal{R}_2$ should be the specific bias attribute present. Define the mean of set $j$ as $\boldsymbol{\mu}_j = \frac{1}{|\mathcal{R}_j|} \sum_{\mathbf{w} \in \mathcal{R}_j} \mathbf{w}$. The bias subspace $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_k\}$ is given by the first $k$ components of principal component analysis (PCA) (Abdi and Williams, 2010):

$$\mathbf{V} = \mathbf{PCA}_k \left( \bigcup_{j=1}^{d} \bigcup_{\mathbf{w} \in \mathcal{R}_j} (\mathbf{w} - \boldsymbol{\mu}_j) \right). \quad (3)$$

$k$ is a hyperparameter in our experiments which determines the dimension of the bias subspace. Intuitively, $\mathbf{V}$ represents the top-$k$ orthogonal directions which most represent the bias subspace.

**4) Debiasing:** Given the estimated bias subspace $\mathbf{V}$, we apply a partial version of the HARD-DEBIAS algorithm (Bolukbasi et al., 2016) to remove bias from new sentence representations. Taking the example of binary gender bias, the HARD-DEBIAS algorithm consists of two steps:

**4a) Neutralize:** Bias components are removed from sentences that are not gendered and should not contain gender bias (e.g., *I am a doctor.*, *That nurse is taking care of the patient.*) by removing the projection onto the bias subspace. More formally, given a representation $\mathbf{h}$ of a sentence and the previously estimated gender subspace $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_k\}$, the debiased representation $\hat{\mathbf{h}}$ is given by first obtaining $\mathbf{h}_\mathbf{V}$, the projection of $\mathbf{h}$ onto the bias subspace $\mathbf{V}$ before subtracting $\mathbf{h}_\mathbf{V}$ from $\mathbf{h}$. This results in a vector that is *orthogonal* to the bias subspace

**V** and therefore contains no bias:

$$\mathbf{h_V} = \sum_{j=1}^{k} \langle \mathbf{h}, \mathbf{v}_j \rangle \mathbf{v}_j, \qquad (4)$$

$$\hat{\mathbf{h}} = \mathbf{h} - \mathbf{h_V}. \qquad (5)$$

**4b) Equalize:** Gendered representations are centered and their bias components are equalized (e.g. *man* and *woman* should have bias components in opposite directions, but of the same magnitude). This ensures that any neutral words are equidistant to biased words with respect to the bias subspace. In our implementation, we skip this Equalize step because it is hard to identify all or even the majority of sentence pairs to be equalized due to the complexity of natural sentences. For example, we can never find all the sentences that *man* and *woman* appear in to equalize them appropriately. Note that even if the magnitudes of sentence representations are not normalized, the debiased representations are still pointing in directions orthogonal to the bias subspace. Therefore, skipping the equalize step still results in debiased sentence representations as measured by our definition of bias.

## 3 Experiments

We test the effectiveness of SENT-DEBIAS at removing biases and retaining performance on downstream tasks. All experiments are conducted on English terms and downstream tasks. We acknowledge that biases can manifest differently across different languages, in particular gendered languages (Zhou et al., 2019), and emphasize the need for future extensions in these directions. Experimental details are in the appendix and code is released at `https://github.com/pliang279/sent_debias`.

### 3.1 Evaluating Biases

Biases are traditionally measured using the Word Embedding Association Test (WEAT) (Caliskan et al., 2017). WEAT measures bias in word embeddings by comparing two sets of target words to two sets of attribute words. For example, to measure social bias surrounding genders with respect to careers, one could use the target words *programmer, engineer, scientist*, and *nurse, teacher, librarian*, and the attribute words *man, male*, and *woman, female*. Unbiased word representations should display no difference between the two target words in terms of their relative similarity to the two sets of attribute words. The relative similarity as measured by WEAT is commonly known as the *effect*

*size*. An effect size with absolute value closer to 0 represents lower bias.

To measure the bias present in sentence representations, we use the method as described in May et al. (2019) which extended WEAT to the Sentence Encoder Association Test (SEAT). For a given set of words for a particular test, words are converted into sentences using a template-based method. The WEAT metric can then be applied for fixed-length, pre-trained sentence representations. To measure bias over multiple classes, we use the Mean Average Cosine similarity (MAC) metric which extends SEAT to a multiclass setting (Manzini et al., 2019). For the binary gender setting, we use words from the Caliskan Tests (Caliskan et al., 2017) which measure biases in common stereotypes surrounding gendered names with respect to careers, math, and science (Greenwald et al., 2009). To evaluate biases in the multiclass religion setting, we modify the Caliskan Tests used in May et al. (2019) with lexicons used by Manzini et al. (2019).

### 3.2 Debiasing Setup

We first describe the details of applying SENT-DEBIAS on two widely-used sentence encoders: BERT[2] (Devlin et al., 2019) and ELMo (Peters et al., 2018). Note that the pre-trained BERT encoder must be fine-tuned on task-specific data. This implies that the final BERT encoder used during debiasing changes from task to task. To account for these differences, we report two sets of metrics: 1) **BERT**: simply debiasing the pre-trained BERT encoder, and 2) **BERT post task**: first fine-tuning BERT and post-processing (i.e. normalization) on a specific task before the final BERT representations are debiased. We apply SENT-DEBIAS on BERT fine-tuned on two single sentence datasets, Stanford Sentiment Treebank (**SST-2**) sentiment classification (Socher et al., 2013) and Corpus of Linguistic Acceptability (**CoLA**) grammatical acceptability judgment (Warstadt et al., 2018). It is also possible to apply BERT (Devlin et al., 2019) on downstream tasks that involve *two* sentences. The output sentence pair representation can also be debiased (after fine-tuning and normalization). We test the effect of SENT-DEBIAS on Question Natural Language Inference (**QNLI**) (Wang et al., 2018) which converts the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) into a binary classification task. These results are

---

[2]We used uncased BERT-Base throughout all experiments.

| Test | BERT | BERT post SST-2 | BERT post CoLA | BERT post QNLI | ELMo |
|---|---|---|---|---|---|
| C6: M/F Names, Career/Family | $+0.477 \rightarrow \mathbf{-0.096}$ | $+\mathbf{0.036} \rightarrow -0.109$ | $\mathbf{-0.009} \rightarrow +0.149$ | $-0.261 \rightarrow \mathbf{-0.054}$ | $-0.380 \rightarrow \mathbf{-0.298}$ |
| C6b: M/F Terms, Career/Family | $+\mathbf{0.108} \rightarrow -0.437$ | $+\mathbf{0.010} \rightarrow -0.057$ | $+0.199 \rightarrow +\mathbf{0.186}$ | $-0.155 \rightarrow \mathbf{-0.004}$ | $-0.345 \rightarrow \mathbf{-0.327}$ |
| C7: M/F Terms, Math/Arts | $+0.253 \rightarrow +\mathbf{0.194}$ | $-0.219 \rightarrow -\mathbf{0.221}$ | $+\mathbf{0.268} \rightarrow +0.311$ | $-0.584 \rightarrow \mathbf{-0.083}$ | $-0.479 \rightarrow -0.487$ |
| C7b: M/F Names, Math/Arts | $+0.254 \rightarrow +\mathbf{0.194}$ | $+1.153 \rightarrow -\mathbf{0.755}$ | $+\mathbf{0.150} \rightarrow +0.308$ | $\mathbf{-0.581} \rightarrow -0.629$ | $+0.016 \rightarrow \mathbf{-0.013}$ |
| C8: M/F Terms, Science/Arts | $+0.399 \rightarrow \mathbf{-0.075}$ | $+0.103 \rightarrow +\mathbf{0.081}$ | $+0.425 \rightarrow +\mathbf{0.163}$ | $-0.087 \rightarrow +0.716$ | $\mathbf{-0.296} \rightarrow -0.327$ |
| C8b: M/F Names, Science/Arts | $+0.636 \rightarrow +\mathbf{0.540}$ | $-0.222 \rightarrow -\mathbf{0.047}$ | $+\mathbf{0.032} \rightarrow -0.192$ | $-0.521 \rightarrow \mathbf{-0.443}$ | $+0.554 \rightarrow +\mathbf{0.548}$ |
| Multiclass Caliskan | $+0.035 \rightarrow +\mathbf{0.379}$ | $+1.200 \rightarrow +\mathbf{1.000}$ | $+0.243 \rightarrow +\mathbf{0.757}$ | $-$ | $-$ |

Table 3: Debiasing results on BERT and ELMo sentence representations. First six rows measure binary SEAT effect sizes for sentence-level tests, adapted from Caliskan tests. SEAT scores closer to 0 represent lower bias. C$N$: test from Caliskan et al. (2017) row $N$. The last row measures bias in a multiclass religion setting using MAC (Manzini et al., 2019) before and after debiasing. MAC score ranges from 0 to 2 and closer to 1 represents lower bias. Results are reported as $x_1 \rightarrow x_2$ where $x_1$ represents score before debiasing and $x_2$ after, with lower bias score in bold. Our method reduces bias of BERT and ELMo for the majority of binary and multiclass tests.

reported as **BERT post SST-2**, **BERT post CoLA**, and **BERT post QNLI** respectively.

For ELMo, the encoder stays the same for downstream tasks (no fine-tuning on different tasks) so we just debias the ELMo sentence encoder. We report this result as **ELMo**.

### 3.3 Debiasing Results

We present these debiasing results in Table 3, and see that for both binary gender bias and multiclass religion bias, our proposed method reduces the amount of bias as measured by the given tests and metrics. The reduction in bias is most pronounced when debiasing the pre-trained BERT encoder. We also observe that simply fine-tuning the BERT encoder for specific tasks also reduces the biases present as measured by the Caliskan tests, to some extent. However, fine-tuning does not lead to consistent decreases in bias and cannot be used as a standalone debiasing method. Furthermore, fine-tuning does not give us control over which type of bias to control for and may even amplify bias if the task data is skewed towards particular biases. For example, while the bias effect size as measured by Caliskan test C7 decreases from +0.542 to −0.033 and +0.288 after fine-tuning on SST-2 and CoLA respectively, the effect size as measured by the multiclass Caliskan test increases from +0.035 to +1.200 and +0.243 after fine-tuning on SST-2 and CoLA respectively.

### 3.4 Comparison with Baselines

We compare to three baseline methods for debiasing: 1) **FastText** derives debiased sentence embeddings using an average of debiased FastText word embeddings (Bojanowski et al., 2016) using word-level debiasing methods (Bolukbasi et al., 2016), 2)

| Debiasing Method | Ave. Abs. Effect Size |
|---|---|
| BERT original (Devlin et al., 2019) | $+0.354$ |
| FastText (Bojanowski et al., 2016) | $+0.565$ |
| BERT word (Bolukbasi et al., 2016) | $+0.861$ |
| BERT simple (May et al., 2019) | $+0.298$ |
| SENT-DEBIAS BERT (ours) | $+\mathbf{0.256}$ |

Table 4: Comparison of various debiasing methods on sentence embeddings. FastText (Bojanowski et al., 2016) (and BERT word) derives debiased sentence embeddings with an average of debiased FastText (and BERT) word embeddings using word-level debiasing methods (Bolukbasi et al., 2016). BERT simple adapts May et al. (2019) by using simple templates to debias BERT representations. SENT-DEBIAS BERT represents our method using diverse templates. We report the average absolute effect size across all Caliskan tests. Average scores closer to 0 represent lower bias.

**BERT word** obtains a debiased sentence representation from average debiased BERT word representations, again debiased using word-level debiasing methods (Bolukbasi et al., 2016), and 3) **BERT simple** adapts May et al. (2019) by using simple templates to debias BERT sentence representations. From Table 4, SENT-DEBIAS achieves a lower average absolute effect size and outperforms the baselines based on debiasing at the word-level and averaging across all words. This indicates that it is not sufficient to debias words only and that biases in a sentence could arise from their debiased word constituents. In comparison with BERT simple, we observe that using diverse sentence templates obtained from naturally occurring written and spoken text makes a difference on how well we can remove biases from sentence representations. This supports our hypothesis that using increasingly diverse templates estimates a bias subspace that generalizes to different words in their context.
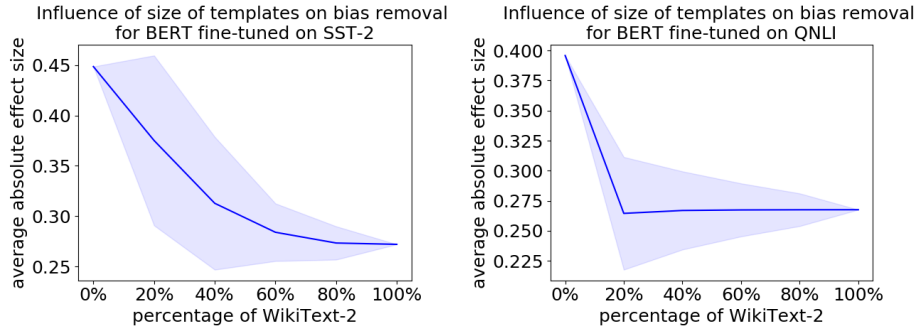
Figure 1: Influence of the number of templates on the effectiveness of bias removal on BERT fine-tuned on SST-2 (left) and BERT fine-tuned on QNLI (right). All templates are from WikiText-2. The solid line represents the mean over different combinations of domains and the shaded area represents the standard deviation. As increasing subsets of data are used, we observe a *decreasing trend* and *lower variance* in average absolute effect size.

## 3.5 Effect of Templates

We further test the importance of sentence templates through two experiments.

**1) Same Domain, More Quantity:** Firstly, we ask: *how does the number of sentence templates impact debiasing performance*? To answer this, we begin with the largest domain WikiText-2 (13750 templates) and divide it into 5 partitions each of size 2750. We collect sentence templates using all possible combinations of the 5 partitions and apply these sentence templates in the contextualization step of SENT-DEBIAS. We then estimate the corresponding bias subspace, debias, and measure the average absolute values of all 6 SEAT effect sizes. Since different combinations of the 5 partitions result in a set of sentence templates of different sizes (20%, 40%, 60%, 80%, 100%), this allows us to see the relationship between size and debiasing performance. Combinations with the same percentage of data are grouped together and for each group we compute the mean and standard deviation of the average absolute effect sizes. We perform the above steps to debias BERT fine-tuned on SST-2 and QNLI and plot these results in Figure 1. Please refer to the appendix for experiments with BERT fine-tuned on CoLA, which show similar results.

For BERT fine-tuned on SST-2, we observe a *decreasing trend* in the effect size as increasing subsets of the data is used. For BERT fine-tuned on QNLI, there is a decreasing trend that quickly tapers off. However, using a larger number of templates *reduces the variance* in average absolute effect size and *improves the stability* of the SENT-DEBIAS algorithm. These observations allow us to conclude the importance of using a large *number* of templates from naturally occurring text corpora.

**2) Same Quantity, More Domains:** *How does the number of domains that sentence templates are extracted from impact debiasing performance?* We fix the total number of sentence templates to be 1080 and vary the number of domains these templates are drawn from. Given a target number $k$, we first choose $k$ domains from our Reddit, SST, POM, WikiText-2 datasets and randomly sample $1080/k$ templates from each of the $k$ selected domains. We construct 1080 templates using all possible subsets of $k$ domains and apply them in the contextualization step of SENT-DEBIAS. We estimate the corresponding bias subspace, debias and measure the average absolute SEAT effect sizes. To see the relationship between the number of domains $k$ and debiasing performance, we group combinations with the same number of domains ($k$) and for each group compute the mean and standard deviation of the average absolute effect sizes. This experiment is also performed for BERT fine-tuned on SST-2 and QNLI datasets. Results are plotted in Figure 2.

We draw similar observations: there is a *decreasing trend* in effect size as templates are drawn from more domains. For BERT fine-tuned on QNLI, using a larger number of domains *reduces the variance* in effect size and *improves stability* of the algorithm. Therefore, it is important to use a large *variety* of templates across different domains.

## 3.6 Visualization

As a qualitative analysis of the debiasing process, we visualize how the distances between sentence representations shift after the debiasing process is performed. We average the sentence representations of a concept (e.g. *man, woman, science, art*) across its contexts (sentence templates) and plot the t-SNE (van der Maaten and Hinton, 2008)
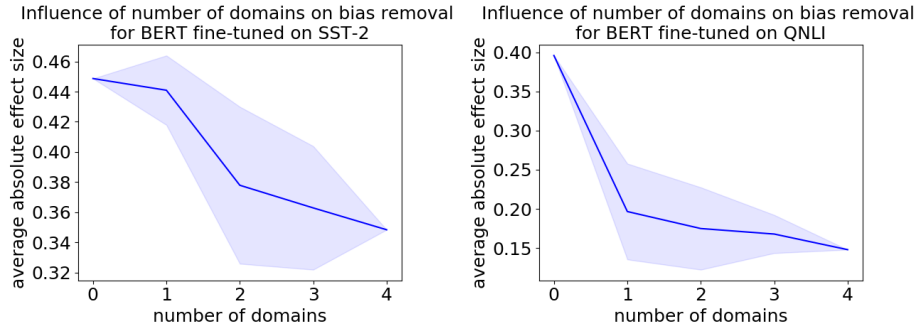
5508

Figure 2: Influence of the number of template domains on the effectiveness of bias removal on BERT fine-tuned on SST-2 (left) and BERT fine-tuned on QNLI (right). The domains span the Reddit, SST, POM, WikiText-2 datasets. The solid line is the mean over different combinations of domains and the shaded area is the standard deviation. As more domains are used, we observe a *decreasing trend* and *lower variance* in average absolute effect size.
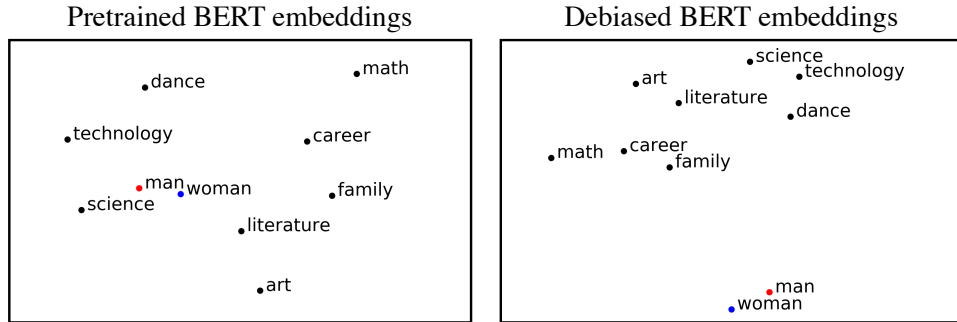


Figure 3: t-SNE plots of average sentence representations of a word across its sentence templates before (left) and after (right) debiasing. After debiasing, non gender-specific concepts (in black) are more equidistant to genders.

embeddings of these points in 2D space. From Figure 3, we observe that BERT average representations of *science* and *technology* start off closer to *man* while *literature* and *art* are closer to *woman*. After debiasing, non gender-specific concepts (e.g *science, art*) become more equidistant to both *man* and *woman* average concepts.

### 3.7 Performance on Downstream Tasks

To ensure that debiasing does not hurt the performance on downstream tasks, we report the performance of our debiased BERT and ELMo on SST-2 and CoLA by training a linear classifier on top of debiased BERT sentence representations. From Table 5, we observe that downstream task performance show a small decrease ranging from $1-3\%$ after the debiasing process. However, the performance of ELMo on SST-2 increases slightly from 89.6 to 90.0. We hypothesize that these differences in performance are due to the fact that CoLA tests for linguistic acceptability so it is more concerned with low-level syntactic structure such as verb usage, grammar, and tenses. As a result, changes in sentence representations across bias directions may impact its performance more. For example,

sentence representations after the gender debiasing steps may display a mismatch between gendered pronouns and the sentence context. For SST, it has been shown that sentiment analysis datasets have labels that correlate with gender information and therefore contain gender bias (Kiritchenko and Mohammad, 2018). As a result, we do expect possible decreases in accuracy after debiasing. Finally, we test the effect of SENT-DEBIAS on QNLI by training a classifier on top of debiased BERT sentence pair representations. We observe little impact on task performance: our debiased BERT fine-tuned on QNLI achieves $90.6\%$ performance as compared to the $91.3\%$ we obtained without debiasing.

## 4 Discussion and Future Work

Firstly, we would like to emphasize that both the WEAT, SEAT, and MAC metrics are not perfect since they only have positive predictive ability: they can be used to detect the presence of biases but not their absence (Gonen and Goldberg, 2019). This calls for new metrics that evaluate biases and can scale to the various types of sentences appearing across different individuals, topics, and in both

| Test | BERT | debiased BERT | ELMo | debiased ELMo |
|------|------|---------------|------|---------------|
| SST-2 | 92.7 | 89.1 | 89.6 | 90.0 |
| CoLA | 57.6 | 55.4 | 39.1 | 37.1 |
| QNLI | 91.3 | 90.6 | - | - |

Table 5: We test the effect of SENT-DEBIAS on both single sentence (BERT and ELMo on SST-2, CoLA) and paired sentence (BERT on QNLI) downstream tasks. The performance (higher is better) of debiased BERT and ELMo sentence representations on downstream tasks is not hurt by the debiasing step.

spoken and written text. We believe that our positive results regarding contextualizing words into sentences implies that future work can build on our algorithms and tailor them for new metrics.

Secondly, a particular bias should only be removed from words and sentences that are neutral to that attribute. For example, gender bias should not be removed from the word "grandmother" or the sentence "she gave birth to me". Previous work on debiasing word representations tackled this issue by listing all attribute specific words based on dictionary definitions and only debiasing the remaining words. However, given the complexity of natural sentences, it is extremely hard to identify the set of neutral sentences and its complement. Thus, in downstream tasks, we removed bias from all sentences which could possibly harm downstream task performance if the dataset contains a significant number of non-neutral sentences.

Finally, a fundamental challenge lies in the fact that these representations are trained without explicit bias control mechanisms on large amounts of naturally occurring text. Given that it becomes infeasible (in standard settings) to completely retrain these large sentence encoders for debiasing (Zhao et al., 2018; Zhang et al., 2018), future work should focus on developing better post-hoc debiasing techniques. In our experiments, we need to re-estimate the bias subspace and perform debiasing whenever the BERT encoder was fine-tuned. It remains to be seen whether there are debiasing methods which are invariant to fine-tuning, or can be efficiently re-estimated as the encoders are fine-tuned.

## 5 Conclusion

This paper investigated the post-hoc removal of social biases from pretrained sentence representations. We proposed the SENT-DEBIAS method that accurately captures the bias subspace of sentence representations by using a diverse set of templates from naturally occurring text corpora. Our experiments show that we can remove biases that occur in

BERT and ELMo while preserving performance on downstream tasks. We also demonstrate the importance of using a large number of diverse sentence templates when estimating bias subspaces. Leveraging these developments will allow researchers to further characterize and remove social biases from sentence representations for fairer NLP.

## References

Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *WIREs Comput. Stat.*, 2(4):433–459.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

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

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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 *Proc. of NIPS*, pages 4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

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.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

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. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony Greenwald, T Poehlman, Eric Luis Uhlmann, and Mahzarin Banaji. 2009. Understanding and using the implicit association test: Iii. meta-analysis of predictive validity.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

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

Brian Larson. 2017. 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. Association for Computational Linguistics.

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

Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Strong and simple baselines for multimodal utterance embeddings. In *NAACL-HLT*, pages 2599–2609, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

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)*. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017a. Pointer sentinel mixture

models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017b. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Matthew 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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. 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. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Chen Sun, Austin Oliver Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019a. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019b. Mitigating gender bias in natural language processing: Literature review. In *ACL*.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA. Association for Computing Machinery.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktaschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *CoRR*, abs/1903.03094.

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. 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11 – 19.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP*.

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

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*.

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

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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. 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 5276–5284, Hong Kong, China. Association for Computational Linguistics.

## A   Debiasing Details

We provide some details on estimating the bias subspaces and debiasing steps.

**Bias Attribute Words:** Table 6 shows the bias attribute words we used to estimate the bias subspaces for binary gender bias and multiclass religious biases.

**Datasets:** We provide some details on dataset downloading below:

1. WikiText-2 was downloaded from `https://github.com/pytorch/examples/tree/master/word_language_model`. We took the first 10% of WikiText-2 sentences as naturally occurring templates representative of highly formal text.

2. Hacker News and Reddit Subreddit data collected from news and discussion forums related to topics ranging from politics to electronics was downloaded from `https://github.com/minimaxir/textgenrnn/`.

## B   Experimental Details

### B.1   BERT

All three variants of BERT (BERT, BERT post SST, BERT post CoLA) are uncased base model with hyper-parameters described in Table 7.

For all three models, the second output "pooled_output" of BERT is treated as the sentence embedding. The variant BERT is the pretrained model with weights downloaded from `https://s3.amazonaws.com/models.huggingface.co/bert/bert-base-uncased.tar.gz`. The variant BERT post SST is BERT after being fine-tuned on the Stanford Sentiment Treebank(SST-2) task, a binary single-sentence classification task (Socher et al., 2013). During fine-tuning, we first normalize the sentence embedding and then feed it into a linear layer for classification. The variant BERT post CoLA is BERT fine-tuned on the Corpus of Linguistic Acceptability (CoLA) task, a binary single-sentence classification task. Normalization and classification are done exactly the same as BERT post SST. All BERT models are fine-tuned for 3 epochs which is the default hyper-parameter in the huggingface transformers repository. Debiasing for BERT models that are fine-tuned is done just before the classification layer.

| Binary Gender |
| --- |
| man, woman |
| boy, girl |
| he, she |
| father, mother |
| son, daughter |
| guy, gal |
| male, female |
| his, her |
| himself, herself |
| John, Mary |

| Multiclass Religion |
| --- |
| jewish, christian, muslim |
| jews, christians, muslims |
| torah, bible, quran |
| synagogue, church, mosque |
| rabbi, priest, imam |
| judaism, christianity, islam |

Table 6: Word pairs to estimate the binary gender bias subspace (left) and the 3-class religion bias subspace (right).

| Hyper-parameter | Value |
| --- | --- |
| attention_probs_dropout_prob | 0.1 |
| hidden_act | gelu |
| hidden_dropout_prob | 0.1 |
| hidden_size | 768 |
| initializer_range | 0.02 |
| intermediate_size | 3072 |
| max_position_embeddings | 512 |
| num_attention_heads | 12 |
| num_hidden_layers | 12 |
| type_vocab_size | 2 |
| vocab_size | 30522 |

Table 7: Configuration of BERT models, including BERT, BERT$\rightarrow$ SST, and BERT$\rightarrow$ CoLA.

### B.2   ELMo

We use the ElmoEmbedder from allennlp.commands.elmo. We perform summation over the aggregated layer outputs. The resulting sentence representation is a time sequence vector with data dimension 1024. When computing gender direction, we perform mean pooling over the time dimension to obtain a 1024-dimensional vector for each definitional sentence. In debiasing, we remove the gender direction from each time step of each sentence representation. We then feed the debiased representation into an LSTM with hidden size 512. Finally, the last hidden state of the LSTM goes through a fully connected layer to make predictions.

## C   Additional Results

We also studied the effect of templates on BERT fine-tuned on CoLA as well. Steps taken are exactly the same as described in **Effect of Templates: Same Domain, More Quantity** and **Effect of Templates: Same Quantity, More Domains**. Results are plotted in Figure 4. It shows
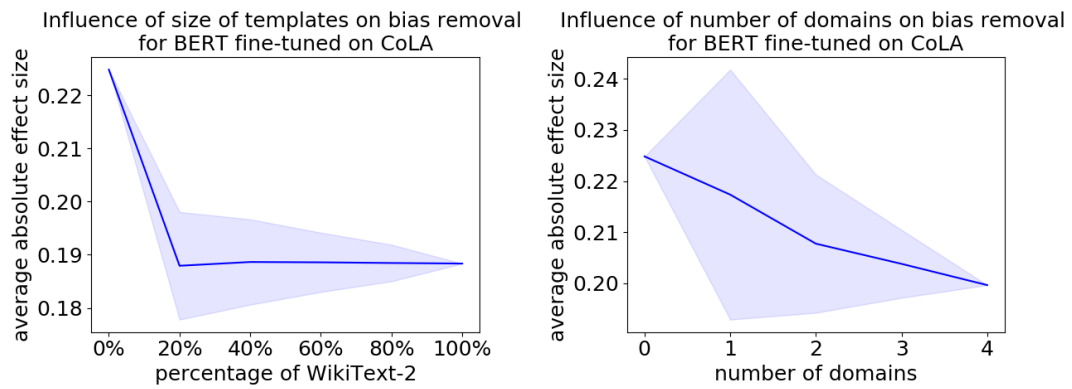
Figure 4: Evaluation of Bias Removal on BERT fine-tuned on CoLA with varying percentage of data from a single domain (left) and varying number of domains with fixed total size (right).

that debiasing performance improves and stabilizes with the number of sentence templates as well as the number of domains.