DETECTING AND MITIGATING INDIRECT STEREOTYPES IN WORD EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Societal biases in the usage of words, including harmful stereotypes, are frequently learned by common word embedding methods. These biases manifest not only between a word and an explicit marker of its stereotype, but also between words that share related stereotypes. This latter phenomenon, sometimes called "indirect bias," has resisted prior attempts at debiasing. In this paper, we propose a novel method to mitigate indirect bias in distributional word embeddings by modifying biased relationships between words before embeddings are learned. This is done by considering how the co-occurrence probability of a given pair of words changes in the presence of words marking an attribute of bias, and using this to average out the effect of a bias attribute. To evaluate this method, we perform a series of common tests and demonstrate that the semantic quality of the word embeddings is retained while measures of bias in the embeddings are reduced. In addition, we conduct novel tests for measuring indirect stereotypes by extending the Word Embedding Association Test (WEAT) with new test sets for indirect binary gender stereotypes. With these tests, we demonstrate that this method can reduce the presence of more subtle stereotypes not properly addressed by previous work.

1 INTRODUCTION

Distributional word embeddings, such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014), are computer representations of words as vectors in semantic space. These embeddings are popular because the geometry of the vectors corresponds to semantic and syntactic structure (Mikolov et al., 2013b). Unfortunately, societal stereotypes, such as those pertaining to race, gender, national origin, or sexuality, are typically reflected in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Papakyriakopoulos et al., 2020). These stereotypes are so pervasive that they have proved resistant to many existing debiasing techniques (Gonen & Goldberg, 2019).

Techniques attempting to remove or mitigate bias in word vectors are common in the literature. The typical case study for bias mitigation methods in the literature is binary[1]gender. Subspace methods, such as hard debiasing from Bolukbasi et al. (2016) and GN-GloVe from Zhao et al. (2018b), attempt to identify or create a vector subspace of gender-related information (typically a "gender direction") and drop this subspace. Counterfactual Data Substitution from Maudslay et al. (2019), based on Counterfactual Data Augmentation from Lu et al. (2020), swaps explicitly gendered words to counter stereotyped associations. James & Alvarez-Melis (2019) and Qian et al. (2019) both propose methods to reduce bias towards binary gender by encouraging learned conditional probabilities of words appearing with "he" and with "she" to be equal.

Gonen & Goldberg (2019) showed that common "debiasing" methods failed to meaningfully reduce bias in word embeddings. They describe how bias can manifest not only as undesirable association

[1]We use the phrase "binary gender" to refer to the common yet unrealistic simplification of gender as just "male" or "female", which we take to be the main source of bias of study in this work. This is a limitation of this work.

between stereotyped words and words marking a bias attribute[2], but also between stereotyped words themselves. These manifestations are sometimes called *direct bias* and *indirect bias*[3], using the terminology introduced by Bolukbasi et al. (2016). An example of this second manifestation of bias is that the word "doctor" might be associated more strongly with stereotypically masculine[4] words than with stereotypically feminine words. At the time, bias mitigation algorithms commonly attempted to address direct bias while leaving indirect bias mostly present.

A common trend in the study of the indirect bias is the departure from *stereotypes* as the object of study in favor of *clustering*. While the measures introduced by Bolukbasi et al. (2016); Caliskan et al. (2017) attempt to quantify the existence of commonly understood stereotypes, work on indirect bias typically uses the measures introduced by Gonen & Goldberg (2019) which merely attempt to measure how well proposed bias mitigation methods disperse words with similar relationships to the bias attribute in the embedding space. These clustering measures, while useful at capturing some forms of indirect bias, are limited. In particular, it is unclear how dispersed stereotyped words *should* be in the embedding space, given that the stereotype of a word is not entirely arbitrary and can potentially be estimated based on its semantic, non-stereotypical, meaning.

These new bias measures have inspired countless new bias mitigation methods. Nearest neighbor bias mitigation from James & Alvarez-Melis (2019) attempts to equalize each word's association with its masculine (defined by the original undebiased embeddings) neighbors and its feminine neighbors. Double hard debias from Wang et al. (2020) projects off the direction defined by the "most gender-biased words" (again, based on alignment in the original embedding's "gender direction") in addition to the standard gender-related subspace. Bordia & Bowman (2019) modify the loss function when learning word embeddings to penalize neutral words having large components in the gender-related subspace which can then be dropped off. Kumar et al. (2020) propose RAN-Debias which attempts to disperse words in the embedding space that share similar binary gender biases (defined again by the original word embeddings) while preserving the original geometry as much as possible. Lauscher et al. (2020) describe multiple bias mitigation methods: the standard projection method, averaging original word vectors with an orthogonal transformation that attempts to swap the bias attribute, and a neural method that uses a loss function to group together words exhibiting a bias attribute away from neutral words. These methods, similarly to the bias measures of Gonen & Goldberg (2019), focus on the clustering and dispersion of words in relation to the bias attribute.

In current state-of-the-art models, word embeddings have largely been replaced by contextualized embeddings from transformer models such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018). However, word embeddings remain a popular object of study when quantifying bias in NLP, in part due to their simplicity and theoretical results that make them easier to reason about. As advances in the understanding of bias and stereotypes in word embeddings have been adapted for these newer models (Liang et al., 2020; May et al., 2019), novel techniques to measure and mitigate bias in word embeddings remain relevant.

2 BACKGROUND

2.1 WORD EMBEDDINGS

Many word embedding algorithms use the empirical probability that two given words appear near each other in the corpus (Levy & Goldberg, 2014; Pennington et al., 2014). This empirical probability is computed by counting how many times one word appears in the context of another as a *word–context* pair. A word–context pair is defined as a pair of words from the corpus that appear within a certain fixed distance from each other, the *window size*, and within the same sentence. A word–context pair designates one word as appearing in the context of another; in this paper, we will refer to a word–context pair as (a, b) where a is a the word appearing in the context of the

[2]We use the phrase "bias attribute" to refer to an attribute associated with stereotypes that are to be removed. The reader can assume "bias attribute" in this work will always refer to binary gender without loss of comprehension. We prefer "bias attribute" for generality.

[3]Or "explicit" and "implicit" bias. In this work we will refer to these as "direct bias" and "indirect bias" for clarity, even though they are actually just two different manifestations of the same phenomenon.

[4]Following the recommendations of Devinney et al. (2022), we prefer the terms "masculine" and "feminine" over "male" and "female" in this work.

word b. Contexts can be unidirectional or bidirectional. In the unidirectional case, in a word–context pair (a, b), a always occurs before b in the corpus (or alternatively, always after). In the bidirectional case, for any pair of nearby words a and b, there are two word–contexts pairs: (a, b) and (b, a).

From these corpus statistics, word embedding algorithms learn vectors for each word in a way that the word co-occurrence statistics can be derived from the geometry of the vectors. The exact details for how this is done is dependent on the exact word embedding algorithm used and is not important for this work.

2.2 Word Embedding Association Test

The Word Embedding Association Test (WEAT), introduced by Caliskan et al. (2017), is a common test used to quantify the presence of specific stereotypes in word embeddings. Given two sets of target words $\mathbb{X}$ and $\mathbb{Y}$ of equal size and two sets of attribute words $\mathbb{A}$ and $\mathbb{B}$, WEAT measures the association between the targets and the attributes. The sets are chosen so that $\mathbb{X}$ and $\mathbb{A}$ are stereotypically linked with each other, and similarly for $\mathbb{Y}$ and $\mathbb{B}$.

The association of a word $w \in \mathbb{X} \cup \mathbb{Y}$ with the attributes is

$$s(w, \mathbb{A}, \mathbb{B}) = \frac{1}{|\mathbb{A}|} \sum_{a \in \mathbb{A}} \cos(\vec{w}, \vec{a}) - \frac{1}{|\mathbb{B}|} \sum_{b \in \mathbb{B}} \cos(\vec{w}, \vec{b})$$

where $\vec{u}$ denotes the word vector corresponding to the word u and

$$\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}|| \, ||\vec{v}||}$$

is the cosine similarity between $\vec{u}$ and $\vec{v}$.

The outputs of WEAT are the test statistic, effect size, and p-value (of a permutation test), defined by the following equations respectively:

$$s(\mathbb{X}, \mathbb{Y}, \mathbb{A}, \mathbb{B}) = \sum_{x \in \mathbb{X}} s(x, \mathbb{A}, \mathbb{B}) - \sum_{y \in \mathbb{Y}} s(x, \mathbb{A}, \mathbb{B})$$

$$\text{EffectSize}(\mathbb{X}, \mathbb{Y}, \mathbb{A}, \mathbb{B}) = \frac{\frac{1}{|\mathbb{X}|} \sum_{x \in \mathbb{X}} s(x, \mathbb{A}, \mathbb{B}) - \frac{1}{|\mathbb{Y}|} \sum_{y \in \mathbb{Y}} s(x, \mathbb{A}, \mathbb{B})}{\text{stddev}_{w \in \mathbb{X} \cup \mathbb{Y}} \, s(w, \mathbb{A}, \mathbb{B})}$$

$$p = P(s(\mathbb{X}_i, \mathbb{Y}_i, \mathbb{A}, \mathbb{B}) > s(\mathbb{X}, \mathbb{Y}, \mathbb{A}, \mathbb{B}))$$

where $(\mathbb{X}_i, \mathbb{Y}_i)$ is a random partition of $\mathbb{X} \cup \mathbb{Y}$ into two sets of equal size.

Of these three measures, the effect size is most commonly used in the literature. In this work, we report the effect size along with the p-value. The effect size retains its original meaning from Caliskan et al. (2017) of measuring the strength and direction of the tested stereotype. We interpret the p-value not as a measure of statistical significance, as originally conceived. Instead, we interpret it as another measure of the tested stereotype, indicating how easily words in $\mathbb{X}$ can be separated from words in $\mathbb{Y}$ according to their association with $\mathbb{A}$ and $\mathbb{B}$: p-values of 0, 0.5, and 1 correspond to perfect separation according to the stereotype, no separation, and perfect separation according to the opposite stereotype, respectively. We will interpret these as relative measures: a WEAT effect size closer to zero and a WEAT p-value closer to 0.5 both suggest reduced presence of the stereotype in the embeddings.

3 Indirect Stereotypes

Previous work on measuring stereotypes has typically focused on measuring associations between stereotyped words and explicit markers of the stereotype, such as names[5] and semantically gendered words. In turn, previous work on mitigating bias have used these same markers as input for bias mitigation methods. By looking at the same sets of words for quantifying and mitigating bias, it

[5]While names do not inherently mark any bias attribute, aggregate groups of names typically isolate a bias attribute more directly than other groups of words can and, more importantly for this work, are commonly used as reference points in bias mitigation algorithms.

is easy to overestimate the effect of bias mitigation. This is what led Gonen & Goldberg (2019) to propose their new bias measures. These are based around measuring how clustered previously biased words are, but it is unclear what result should be desired from these measures. To address this issue, we demonstrate that we can capture some forms of the remaining bias as stereotypes, which we refer to as *indirect stereotypes*.

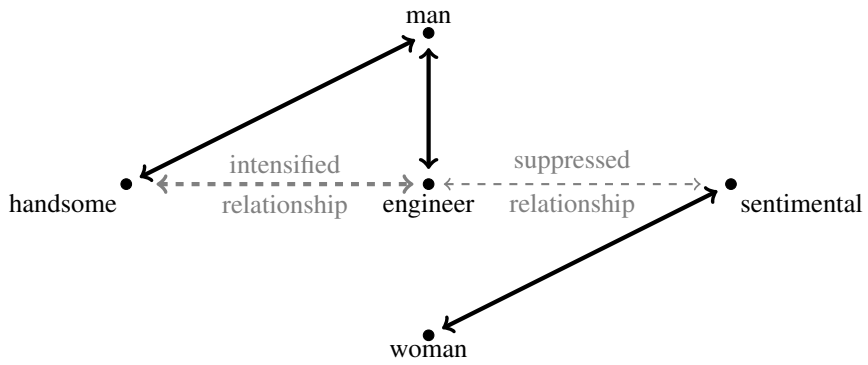

Figure 1: An example showing indirect stereotypes.

An example of an indirect stereotype is shown in Figure 1. There are three words here exhibiting a stereotype in regards to binary gender: "handsome" and "engineer," which are both masculine and not feminine, and "sentimental," which is feminine and not masculine. In a corpus with binary gender stereotypes, we would expect a quantifiably stronger association between "handsome" and "engineer" than between "sensitive" and "engineer" to exist purely because they are both stereotypically used to refer to men,[6] even if there is not a stereotype that engineers are more handsome than they are sensitive.

These stereotypes can come about as a result of sentences that exhibit multiple stereotypes at the same time. For example, the sentence "he is an a handsome engineer" exhibits the masculine stereotypes for "handsome" and "engineer". In a corpus that exhibits both of these stereotypes, this sentence would be more likely to occur than the following related sentences:

She is a handsome engineer.

He is a sensitive engineer.

She is a sensitive engineer.

This will result in "handsome" occurring with "engineer" more frequently than "sensitive" does. However, when word embeddings are "debiased" with respect to binary gender, the goal in previous work is typically to obtain word vectors that would predict each pair of the previous examples that differ only in choice of pronoun as equally likely to occur. This is not enough to correct the associations between "handsome," "engineer," and "sensitive." Furthermore, the association between "engineer" and "handsome" should exist even in sentences without explicit reference to binary gender, as in those cases "engineer" is more likely to refer to a man than a woman. More details on this analysis can be found in Appendix A.

To test indirect stereotypes, we use WEAT with the categories that are stereotyped according to binary gender (these are science/art, math/art, and career/family), as well a test set of professions and adjectives that are stereotypically masculine and feminine. These new test sets and the process used to generate them are detailed in Appendix B.

4 Bias Mitigation Method

Instead of only looking at the co-occurrence probability of two words, we consider the probability a word co-occurs with another and one of the two also occurs near a word marking the bias attribute. By using the word marking a bias attribute as a proxy for the bias attribute itself, we can determine

[6] Assuming there is no other relationship between "engineer" and "handsome" or "sensitive".

how the relationship between the two words varies as the bias attribute varies. Then, we can approximate what the relationship between two words would be if one of them had no association with the bias attribute.

As an example, suppose we would like to mitigate one of the stereotypes in Figure 1, perhaps that "engineer" and "handsome" are more closely associated than they should be based on non-stereotypical semantics. In the corpus, it might be that an engineer is described as handsome three times as often when "engineer" refers to a man as opposed to when it refers to a woman. It might also be the case that "engineer" refers to a man twice as often as it refers to a woman. If "engineer" is adjusted so that it refers to men just as often as women, then the probability that an engineer is referred to as handsome should also be adjusted to be

$$\frac{\frac{1}{2} \cdot P(\text{“handsome”}|\text{man}) + \frac{1}{2} \cdot P(\text{“handsome”}|\text{woman})}{\frac{2}{3} \cdot P(\text{“handsome”}|\text{man}) + \frac{1}{3} \cdot P(\text{“handsome”}|\text{woman})} = \frac{6}{7}$$

of what it was before, at least when just considering occurrences where "engineer" refers to a man or woman. Whether "engineer" refers to a man or woman is not known to us a priori, but can be estimated based on whether "engineer" is near a word marking binary gender. Occurrences where "engineer" is not near such a word could be left as is or also be adjusted with this factor based on the reasoning in Section 3. In this work, we take the latter approach. We note that word embeddings do not attempt to quantify whether or not "handsome" describes an engineer, but rather if "handsome" occurs near "engineer". However, the same principle still applies.

The general method is as follows. Suppose a word–context pair in the corpus is selected uniformly at random. Consider two words a and b where at least one of a or b is a word to be neutralized with respect to this bias. Additionally, let $\mathbb{X}$ and $\mathbb{Y}$ be two sets of words that are similar in usage but differ in terms of the bias attribute that is being mitigated. For example, in the case of binary gender, $\mathbb{X}$ could be the set {"he", "man", ...} and $\mathbb{Y}$ could be the set {"she", "woman", ...}.

Define the following events for a random word–context pair (c, d):

$$\begin{aligned}
&\text{A} : c = a \\
&\text{B} : d = b \\
&\text{X} : \text{the pair appears near a word in } \mathbb{X} \\
&\text{Y} : \text{the pair appears near a word in } \mathbb{Y}.
\end{aligned}$$

A word–context pair is defined as being near a word in $\mathbb{X}$ or in $\mathbb{Y}$ if this word is within a certain distance from c or from d. This distance can be taken to be the window size used for determining word–context pairs, for simplicity. In this work, we weight occurrences where a word in $\mathbb{X}$ or $\mathbb{Y}$ appears near just one of c or d half as much as if the word appeared near both of c and d, but we could just as easily weight these occurrences the same.

The probability the word–context pair (c, d) is (a, b) can be decomposed as

$$\begin{aligned}
P(\text{A} \cap \text{B}) = \frac{1}{P(\text{X} \cup \text{Y}|\text{A} \cap \text{B})} \big[& P(\text{B} \cap \text{X}) P(\text{A}|\text{B} \cap \text{X}) \\
& + P(\text{B} \cap \text{Y}) P(\text{A}|\text{B} \cap \text{Y}) \\
& - P(\text{A} \cap \text{B} \cap \text{X} \cap \text{Y}) \big],
\end{aligned}$$

using the definition of conditional probability and inclusion-exclusion. This representation isolates the contribution of the bias attribute on $P(\text{A} \cap \text{B})$. The terms $P(\text{B} \cap \text{X})$ and $P(\text{B} \cap \text{Y})$ indicate the bias b has, if any, while the terms $P(\text{A}|\text{B} \cap \text{X})$ and $P(\text{A}|\text{B} \cap \text{Y})$ describe how the relationship between a and b is influenced by the bias attribute.

Now if b is replaced with a hypothetical neutral word b' that has the same meaning except with no association to the bias attribute, the probability of b' occurring with a can be approximated. Denote by B' the event that a randomly selected word–context pair (c, d) satisfies $d = b'$.

The following exact equalities hold in this setting:

$$\begin{aligned}
P((\text{X} \cup \text{Y}) \cap \text{B}) &= P((\text{X} \cup \text{Y}) \cap \text{B}') \\
P(\text{X} \cup \text{Y}|\text{A} \cap \text{B}) &= P(\text{X} \cup \text{Y}|\text{A} \cap \text{B}') \\
P(\text{X} \cap \text{B}') &= P(\text{Y} \cap \text{B}').
\end{aligned}$$

The first two identities hold because b and b' have the same meaning except for the bias. The last identity holds because b' has no bias.

In addition, the following relations are approximately true:

$$P(\mathrm{A}|\mathrm{B} \cap \mathrm{X}) \approx P(\mathrm{A}|\mathrm{B}' \cap \mathrm{X})$$
$$P(\mathrm{A}|\mathrm{B} \cap \mathrm{Y}) \approx P(\mathrm{A}|\mathrm{B}' \cap \mathrm{Y})$$
$$P(\mathrm{B} \cap \mathrm{X} \cap \mathrm{Y}) \approx P(\mathrm{B}' \cap \mathrm{X} \cap \mathrm{Y})$$
$$P(\mathrm{A} \cap \mathrm{B} \cap \mathrm{X} \cap \mathrm{Y}) \approx P(\mathrm{A} \cap \mathrm{B}' \cap \mathrm{X} \cap \mathrm{Y})$$

In the first two lines, the explicit presence of the bias attribute of interest is likely to overpower the bias b has. In the last two lines, the approximate relationship is likely to hold as all the probabilities are small: the events are the intersection of three or more events and are subsets of $\mathrm{X} \cap \mathrm{Y}$.

These relations imply

$$P(\mathrm{X} \cap \mathrm{B}') = P(\mathrm{Y} \cap \mathrm{B}') \approx \frac{P(\mathrm{X} \cap \mathrm{B}) + P(\mathrm{Y} \cap \mathrm{B})}{2}$$

and

$$\begin{aligned} P(\mathrm{A} \cap \mathrm{B}') \approx & \frac{1}{P(\mathrm{X} \cup \mathrm{Y}|\mathrm{A} \cap \mathrm{B})} \big[(P(\mathrm{A}|\mathrm{B} \cap \mathrm{X}) + P(\mathrm{A}|\mathrm{B} \cap \mathrm{Y})) \cdot \tfrac{1}{2}(P(\mathrm{B} \cap \mathrm{X}) + P(\mathrm{B} \cap \mathrm{Y})) \\ & - P(\mathrm{A} \cap \mathrm{B} \cap \mathrm{X} \cap \mathrm{Y}) \big]. \end{aligned} \tag{1}$$

Since b' is neutral, if a is replaced with a neutral word a' and A replaced with a corresponding event A', it holds that $P(\mathrm{A}' \cap \mathrm{B}') \approx P(\mathrm{A} \cap \mathrm{B}')$. A similar argument shows equation 1 holds when the role of a and b are switched. These two approximations can be averaged to yield a better approximation for $P(\mathrm{A}' \cap \mathrm{B}')$:

$$\begin{aligned} P(\mathrm{A}' \cap \mathrm{B}') \approx & \frac{1}{P(\mathrm{X} \cup \mathrm{Y}|\mathrm{A} \cap \mathrm{B})} \big[\tfrac{1}{4}(P(\mathrm{B} \cap \mathrm{X}) + P(\mathrm{B} \cap \mathrm{Y}))(P(\mathrm{A}|\mathrm{B} \cap \mathrm{X}) + P(\mathrm{A}|\mathrm{B} \cap \mathrm{Y})) \\ & + \tfrac{1}{4}(P(\mathrm{A} \cap \mathrm{X}) + P(\mathrm{A} \cap \mathrm{Y}))(P(\mathrm{B}|\mathrm{A} \cap \mathrm{X}) + P(\mathrm{B}|\mathrm{A} \cap \mathrm{Y})) \\ & - P(\mathrm{A} \cap \mathrm{B} \cap \mathrm{X} \cap \mathrm{Y}) \big]. \end{aligned} \tag{2}$$

This approximation is the basis of our proposed bias mitigation method. We propose the right hand side of equation 2 can replace the probability of a word–context pair being (a, b) whenever this quantity would be used, where a or b is a word that is wanted to be neutralized. This can be done easily with word embedding methods that directly use the probability or counts of a specific word–context pair occurring, such as GloVe or PPMI factoring, but in principle could be adapted even to methods that only indirectly use these probabilities.

For all the terms in equation 2 to be defined, at least one occurrence each of $\mathrm{A} \cap \mathrm{X}$, $\mathrm{A} \cap \mathrm{Y}$, $\mathrm{B} \cap \mathrm{X}$, and $\mathrm{B} \cap \mathrm{Y}$ must be present in the corpus. Therefore, the vocabulary must be limited so that all words in the corpus occur in a word–context pair satisfying X and in a pair satisfying Y. While this may be undesirable, in practice this is not too strong of a restriction on the vocabulary.

5 Results

We conduct experiments by training GloVe embeddings on the UMBC webbase corpus (Han et al., 2013) and then attempting to mitigate the presence of bias from binary gender. Due to computational constraints, we only train on 30 million sentences from the corpus. We compare the proposed method from Section 4 with GloVe, with and without Counterfactual Data Subsitution (Maudslay et al., 2019). These are labeled was "Proposed Method", "Original", and "CDS" respectively in the following tables. The parameters we use to train the models for these experiments are in Appendix C. In all tables, the best results for each row are shown in bold.

As a first test, we evaluate all three word embeddings on a series of common semantic tests (Jastrzebski et al., 2017). The relative change in the results of these tests after the two tested bias mitigation methods are shown in Table 1. For all tests, a higher score indicates a better representation of the

Table 1: Semantic measures

Test	Proposed Method	CDS
MEN	$+\mathbf{0.54}\%$	-0.86%
WS353	$+\mathbf{0.92}\%$	-2.18%
WS353R	$+\mathbf{2.91}\%$	-2.97%
WS353S	$-\mathbf{1.01}\%$	-2.43%
SimLex999	$+\mathbf{0.59}\%$	-2.48%
RW	-9.49%	$-\mathbf{1.97}\%$
RG65	-13.82%	$-\mathbf{1.25}\%$
MTurk	-0.39%	$-\mathbf{0.15}\%$
AP	-9.05%	$-\mathbf{5.91}\%$
BLESS	-13.04%	$+\mathbf{0.62}\%$
Battig	-6.61%	$-\mathbf{2.56}\%$
ESSLI_2c	$\pm\mathbf{0}\%$	$\pm\mathbf{0}\%$
ESSLI_2b	$\pm\mathbf{0}\%$	$\pm\mathbf{0}\%$
ESSLI_1a	$\pm\mathbf{0}\%$	$\pm\mathbf{0}\%$
SemEval2012_2	-20.93%	$-\mathbf{2.51}\%$
Google	-23.93%	$+\mathbf{1.36}\%$
MSR	-17.25%	$-\mathbf{1.96}\%$

Table 2: p-values for WEAT for unrelated biases

Test	Original	Proposed Method	CDS
Flower/Insect–Pleasant/Unpleasant	**3.4e-5**	1.2e-4	5.0e-5
Music Instruments/Weapons–Pleasant/Unpleasant	1e-6	**3.3e-7**	1e-6

tested semantics in the word embeddings. The tests are grouped into three categories; from top to bottom, these are word similarity tasks, categorization tasks, and analogy tasks. Overall, our proposed method typically results in worse degradation of semantics than CDS does. This is most pronounced in the analogy tasks, where there is an approximately 20% decrease in accuracy.

As another test for semantics, we compare all the word embeddings on two tests for WEAT that do not capture stereotypes related to binary gender. These tests exhibit the bias that music instruments are more pleasant than weapons, and flowers are more pleasant than insects. The results are shown in Tables 2 and 3 for the p-value and effect sizes, respectively. Since these biases encode information that is not being attempted to remove, a small p-values and a large effect sizes are desirable. We see that all three word embeddings exhibit strong representations of these biases with only minor differences between them.

To test the standard direct stereotypes towards binary gender, we again use WEAT with three binary gender stereotypes used in the original paper. The p-value and effect sizes for these tests are in Tables 4 and 5. All three of these stereotypes are present in the original GloVe embedding, with the association between traditionally masculine and feminine names with career and home words being the strongest. The presence of all three of these stereotypes are reduced as a result of both bias mitigation methods, although the career/home stereotype is not substantially reduced by either method.

Table 3: Effect sizes for WEAT for unrelated biases

Test	Original	Proposed Method	CDS
Flower/Insect–Pleasant/Unpleasant	1.062	**1.066**	1.055
Music Instruments/Weapons–Pleasant/Unpleasant	1.256	**1.416**	1.260

Table 4: p-values for WEAT for direct stereotypes

Test	Original	Proposed Method	CDS
Math/Art–Masc./Fem. Words	0.028	0.134	**0.600**
Science/Art–Masc./Fem. Words	0.073	**0.624**	0.308
Masc./Fem. Names–Career/Home	7.8e-5	**1.55e-4**	**1.55e-4**

Table 5: Effect sizes for WEAT for direct stereotypes

Test	Original	Proposed Method	CDS
Math/Art–Masc./Fem. Words	0.971	0.586	**-0.138**
Science/Art–Masc./Fem. Words	0.771	**-0.170**	0.268
Masc./Fem. Names–Career/Home	1.658	1.560	**1.546**

Table 6: p-values for WEAT for indirect stereotypes

Test	Original	Proposed Method	CDS
Professions–Adjectives	0.0347	**0.0820**	0.0294
Math/Art–Adjectives	1.55e-4	**0.456**	0.0521
Science/Art–Adjectives	6.22e-4	**2.95e-3**	1.24e-4
Career/Home–Adjectives	7.8e-5	**1.55e-4**	7.8e-5

Table 7: Effect sizes for WEAT for indirect stereotypes

Test	Original	Proposed Method	CDS
Professions–Adjectives	0.831	**0.647**	0.861
Math/Art–Adjectives	1.591	**0.059**	0.834
Science/Art–Adjectives	1.544	**1.359**	1.414
Career/Home–Adjectives	1.823	**1.730**	1.800

Lastly, we investigate the presence of indirect stereotypes as described in Section 3. We use a list of stereotypically masculine and feminine adjectives and look at its association with stereotypically masculine and feminine professions, as well as the same math/art, science/art, and career/home categories from the standard WEAT tests. The results of these experiments are in Tables 6 and 7. From this, it can be seen that all these stereotypes do exist in the original word embeddings. Counterfactual Data Substitution does not succeed at significantly reducing any of these stereotypes, although it does have modest success at reducing the stereotype with math and art. Our proposed method reduces the presence of all these stereotypes to a greater extent than CDS, although it is not uniformly successful. In particular, it has modest success in reducing the presence of the indirect stereotype with professions and is able to significantly reduce the presence of the math and art indirect stereotype.

These results are modest but suggestive. Even by our own measures, we are not able to substantively mitigate all tested stereotypes present in the word embeddings. Furthermore, with recent results showing that intrinsic bias measures (including WEAT) do not necessarily generalize to downstream tasks (Goldfarb-Tarrant et al., 2021; Orgad & Belinkov, 2022), the performed tests alone cannot guarantee word embeddings lack any bias. However, by considering stereotypes as a measure of indirect bias, these tests can be extended to downstream bias measures such as WinoBias (Zhao et al., 2018a) and WinoGender (Rudinger et al., 2018) more readily than the previously used dispersion measures.

6 Discussion

In this paper, we discuss how indirect bias in word embeddings can manifest as stereotypes. Using the standard Word Embedding Association Test with additional test sets, we demonstrate that these "indirect stereotypes" have a substantial presence in word embeddings and are not removed by current debiasing methods. Furthermore, we propose a new method that attempts to directly mitigate these indirect stereotypes and demonstrate that this method can have some success in practice, albeit with trade-offs on the semantic quality of the resulting word vectors.

REFERENCES

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. *Advances in Neural Information Processing Systems*, 29, 2016.

Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

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

Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in NLP bias research. *arXiv preprint arXiv:2205.02526*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115 (16):E3635–E3644, 2018.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1926–1940, 2021.

Hila Gonen and Yoav Goldberg. 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)*, pp. 609–614, 2019.

Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 44–52, 2013.

Megumi Hosoda and Dianna L Stone. Current gender stereotypes and their evaluative content. *Perceptual and motor skills*, 90(3_suppl):1283–1294, 2000.

Hailey James and David Alvarez-Melis. Probabilistic bias mitigation in word embeddings. *arXiv preprint arXiv:1910.14497*, 2019.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *arXiv preprint arXiv:1702.02170*, 2017.

Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503, 2020.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8131–8138, 2020.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27, 2014.

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*, pp. 5502–5515, 2020.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pp. 189–202. Springer, 2020.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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)*, pp. 5267–5275, 2019.

Chandler May, Alex Wang, Shikha Bordia, Samuel 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)*, pp. 622–628, 2019.

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

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pp. 746–751, 2013b.

Hadas Orgad and Yonatan Belinkov. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 151–167, 2022.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 446–457, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

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*, pp. 223–228, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14, 2018.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5443–5453, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, 2018a.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018b.

A Analysis of Indirect Stereotype Occurrence

Consider the following list of sentences:

He is a handsome engineer.
She is a handsome engineer.
He is a sensitive engineer.
She is a sensitive engineer.

The first sentence exhibits the indirect stereotype found in Figure 1 and is the most likely sentence to occur in a corpus that treats both "engineer" and "handsome" as masculine and non-feminine words and "sensitive" as a feminine and non-masculine word.

Let $P(A)$ for a list A of words be the probability that a sentence in the corpus contains the words in A, given that it is one of the four example sentences above. As long as $P(\text{“he”}) > \frac{1}{2}$ and $P(\text{“handsome”}|\text{“he”}) > P(\text{“handsome”}|\text{“she”})$, then

$$\begin{aligned} P(\text{“handsome”}) &= P(\text{“handsome”}|\text{“he”})P(\text{“he”}) + P(\text{“handsome”}|\text{“she”})P(\text{“she”}) \\ &> \frac{1}{2} \cdot P(\text{“handsome”}|\text{“he”}) + \frac{1}{2} \cdot P(\text{“handsome”}|\text{“she”}). \end{aligned}$$

The expression on the last line is what we would expect the probability of a sentence with "handsome" appearing would be if "engineer" had no bias towards binary gender.

B Indirect Stereotype Word Lists

For the list of adjectives, we use the "key" feminine and masculine traits from Hosoda & Stone (2000). We remove the words "feminine" and "masculine" from these lists as our goal is to quantify indirect stereotypes. We also remove "hard-headed" from the masculine list as it it outside of the model's vocabulary.

Feminine adjectives: affectionate, sensitive, appreciative, sentimental, sympathetic, nagging, fussy, emotional

Masculine adjectives: handsome, aggressive, tough, courageous, strong, forceful, arrogant, egotistical, boastful, dominant

For the list of professions, we use the list from Caliskan et al. (2017) and take the ten jobs with the highest proportion of women and of men according to the data that they derive from the Bureau of Labor Statistics.

Feminine professions: therapist, planner, librarian, paralegal, nurse, receptionist, hairdresser, nutritionist, hygienist, pathologist

Masculine professions: plumber, mechanic, carpenter, electrician, machinist, engineer, programmer, architect, officer, paramedic

C Experiment Details

For all methods, we train GloVe with an embedding dimension of 300. In determining the cooccurrence matrix, we use a window size of 15 and skip over words with less than 5 occurrences in the corpus. We also preprocess our corpus by removing capitalization, skipping sentences with less than five words, and using Stanford CoreNLP (Manning et al., 2014) to separate out certain lexemes from words (e.g., n't and 's).

For the proposed bias mitigation method, we use the following sets of words:

$$\mathbb{X} = \{\text{“he”, “him”, “his”, “himself”, “man”, “men”, “boy”, “boys”}\}$$
$$\mathbb{Y} = \{\text{“she”, “her”, “hers”, “herself”, “woman”, “women”, “girl”, “girls”}\}.$$