

Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence

Navid Rekasaz,¹ Robert West,² James Henderson,³ Allan Hanbury⁴

¹Johannes Kepler University, Linz Institute of Technology, AI Lab,
²EPFL,

³Idiap Research Institute,

⁴TU Wien, Complexity Science Hub Vienna

navid.rekasaz@jku.at, robert.west@epfl.ch, james.henderson@idiap.ch, allan.hanbury@tuwien.ac.at

Abstract

Text corpora are widely used resources for measuring societal biases and stereotypes. The common approach to measuring such biases using a corpus is by calculating the similarities between the embedding vector of a word (like *nurse*) and the vectors of the representative words of the concepts of interest (such as genders). In this study, we show that, depending on what one aims to quantify as bias, this commonly-used approach can introduce non-relevant concepts into bias measurement. We propose an alternative approach to bias measurement utilizing the smoothed first-order co-occurrence relations between the word and the representative concept words, which we derive by reconstructing the co-occurrence estimates inherent in word embedding models. We compare these approaches by conducting several experiments on the scenario of measuring gender bias of occupational words, according to an English Wikipedia corpus. Our experiments show higher correlations of the measured gender bias with the actual gender bias statistics of the U.S. job market – on two collections and with a variety of word embedding models – using the first-order approach in comparison with the vector similarity-based approaches. The first-order approach also suggests a more severe bias towards female in a few specific occupations than the other approaches.

Introduction

Text data has been widely utilized for studying and monitoring societal phenomena – such as biases and stereotypes – commonly by exploiting co-occurrence statistics of words in text. In these approaches, a societal bias construct (an unobservable abstraction that we aim to characterize) is quantified using measures of words association. A word such as *nurse* is considered to be stereotypically biased towards the female concept, when a significant imbalance is observed between the associations of *nurse* to female versus male concept. Each of the concepts is commonly defined by a group of words, referred to as *concept-representative words*. The focus of the present work is on the computational methods for measuring biases from text corpora – a particularly essential component in various social studies.

The common approach to calculate words associations for bias measurement is by adopting word embedding mod-

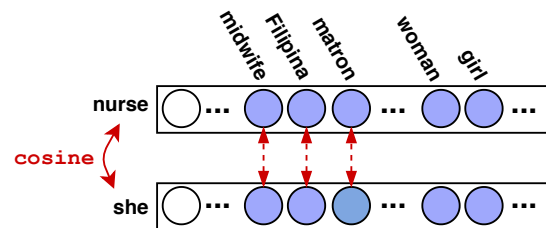


Figure 1: Word vectors with explicit dimensions. Estimating bias with similarity between the vectors also counts some (arguably) irrelevant context-words to the female concept (red dashed lines). The example is based on the results in Table 2.

els trained on text corpora as in preceding studies (Lenton, Sedikides, and Bruder 2009; Hoyle et al. 2019; Zhou et al. 2019; Chang and McKeown 2019; Zhao et al. 2019; Garg et al. 2018; Caliskan, Bryson, and Narayanan 2017; Bolukbasi et al. 2016). In these studies, the associations of words to concepts are measured based on some form of vector similarity, for instance by using the cosine metric. The present study sheds light on and discusses what is captured as bias by these vector similarity-based approaches, and proposes a complementary bias measurement approach based on a smoothed variant of direct (first-order) co-occurrences. Let us first have a closer look at what is measured by vector similarity or more generally by similarity metrics applied to distributional representations.

In distributional representations, two vectors are more similar if the corresponding words both frequently co-occur with a set of context-words (second-order co-occurrence). Figure 1 elaborates this using a toy example. In the example, the association of the word *nurse* to the female concept, represented by the word *she*, is calculated using cosine similarity between their vectors. The word vectors in the example are high-dimensional representations, such that each dimension of a vector is defined explicitly with a specific context-word, and each value of the vector represents the first-order co-occurrence relation between the word and the corresponding context-word. We refer to such vectors as *explicit representations*. As shown, *nurse* and *she* are simi-

lar since they co-occur with several common context-words, depicted with the blue circles in both vectors.

Let us assume that in this example our objective of gender bias measurement can be formulated as “to quantify the extent to which *nurse* is perceived as female versus male”. Considering this objective, we observe that the *nurse-to-she* similarity is influenced by the common context-words such as *woman* and *girl*, which are typically considered as good representatives of the female concept (Garg et al. 2018; Bolukbasi et al. 2016). However, this similarity is also affected by several other context-words, such as *midwife* and *matron*. According to the literal definitions of these words (as defined in a dictionary), *midwife* is gender-neutral, and *matron* is a mixture of the female concept and the concept of “being in charge of medical arrangements”. Considering this, one can argue that such common context-words can introduce partially or completely irrelevant concepts to female into the measured association, and hence into the calculated gender bias.

While we discuss this issue on explicit representations, it is also present in low-dimensional embeddings, although such an explicit dividing of dimensions into groups is not per se possible. Another difference between explicit representations and embeddings is that, since word embeddings are defined in low dimensions, the similarity of embedding vectors does not only capture second-order co-occurrences, but other orders of co-occurrence, such as first-order as well as higher orders.¹ We therefore refer to the approaches that use word embeddings for quantifying bias as *high-order bias measurements*.

We approach the discussed issue in high-order bias measurement by revisiting the utilization of first-order co-occurrences for measuring bias. Our proposed approach, referred to as *first-order bias measurement*, estimates the association of a concept and a word by averaging the first-order co-occurrences between the word and the concept-representative context-words. The first-order bias measurement introduces an alternative to high-order approach, and has the advantage of only taking into account the context-words which are strongly related to the concept of interest.

First-order co-occurrence of words has been widely used to calculate societal phenomena particularly through counting and weighting words (Monroe, Colaresi, and Quinn 2008; Kirchler 1992; Rekabsaz et al. 2017). Among various metrics, the ones based on Pointwise Mutual Information (PMI) (Church and Hanks 1990) are commonly used to measure words co-occurrence in local contexts. As mentioned by Lenton, Sedikides, and Bruder (2009), a draw back of such count-based co-occurrence metrics is the high sparsity of their resulting vectors, as many related words never appear in the same local context.

We address the issue of sparsity in count-based metrics by proposing two novel explicit representations, created from pre-trained word2vec Skip-Gram (Mikolov et al. 2013), and GloVe (Pennington, Socher, and Manning 2014) models.

¹For instance, Kontostathis and Pottenger (2006) found that Latent Semantic Analysis (Deerwester et al. 1990) can take into account up to fifth-order co-occurrences.

The proposed explicit variants exploit the word and context embedding vectors of word2vec and GloVe to estimate forms of the co-occurrence relations. Such co-occurrence relations, achieved from the reconstruction of explicit vectors from low-dimensional embeddings, provide smoothed variants of the count-based co-occurrence estimations.

We use the discussed word representations, trained on an English Wikipedia corpus to study the characteristics of the first- and high-order bias measurement approaches. We conduct several experiments on the gender bias of occupations. We first revisit the experiments conducted in previous studies (Garg et al. 2018; Caliskan, Bryson, and Narayanan 2017) on the correlations of the gender bias of some occupations, measured using the discussed methods, to the actual statistics of gender bias in the U.S. job market. We use two collections, provided by Zhao et al. (2018a) and Garg et al. (2018). We observe that, in all studied word representation models and the two collections, the results of our proposed first-order bias measurement shows higher correlations in comparison with the high-order approaches.

Next, we analyze the measured gender bias of around 500 occupations using first- and high-order approaches, observing several cases of the influence of non-relevant context-words on the results of the high-order bias measurement method. Overall, our results suggest the existence of a more severe degree of bias towards female in the underlying corpus, previously undetected by high-order approaches.

Finally, we study how each bias measurement approach reacts to (hypothetical) changes in the corpus, particularly when the corpus moves towards a more balanced representation of genders. To this end, we manipulate the corpus using the Counterfactual Data Augmentation (CDA) method (Zhao et al. 2018a; Lu et al. 2018; Zmigrod et al. 2019), such that the genders are represented in the augmented corpora in a more balanced way. Our observations show that, while both bias measurement methods report a decrease of gender bias, the first-order bias measurement is more sensitive (reacts faster) to the changes in the corpus.

Limitations of the study. Gender is treated in this study as a binary construct, and the definition of gender bias is limited to the disparity between female and male. We acknowledge that this choice neglects the broad meaning of gender, but the decision is necessary for taking practical steps. Our study is also limited to the English language. Our introduced method is however generic and can be applied to other languages as well as other forms of societal biases such as related to race, age and ethnicity.

Outline of the paper. We first discussed related work, followed by the relevant previous methods. Our bias measurement approach is introduced next. Finally, the gender bias experiments are described, and whose results are presented.

Related Work

Various aspects of word embeddings and distributional representations in areas such as social sciences and psychology are studied in previous work. Lenton, Sedikides, and Bruder (2009) discuss the use of Latent Semantic Analysis (Deerwester et al. 1990) for measuring gender bias, highlighting

the sparsity issue of the first-order method based on count-based metrics. The present work complements this study by exploring the benefits of a smoothed first-order bias measurement approach as an alternative to previous approaches. In a recent study, Günther, Rinaldi, and Marelli (2019) discuss the applications and common misconceptions of distributional semantic models in psychology.

Several pieces of work exploit word embeddings to study societal aspects. Garg et al. (2018) investigate the changes in gender- and race-related stereotypes over decades using historical text data. Caliskan, Bryson, and Narayanan (2017) and more recently Chang and McKeown (2019) study the patterns of language use, indicating accurate imprints of historical biases. Bolukbasi et al. (2016) show the reflection of gender stereotypes in word analogies derived from word embeddings. Zhou et al. (2019) propose methods to measure societal biases in languages with grammatical gender. Our work directly contributes to these studies by proposing a more accurate approach for measuring bias in corpora. In this line, Hoyle et al. (2019) propose a method to measure the differences between descriptions of men and women. In contrast to our bag-of-words-based method, they measure bias using a parsed corpus.

Gender bias is also studied in various downstream tasks, such as sentiment analysis (Kiritchenko and Mohammad 2018), visual semantic role labeling (Zhao et al. 2017), coreference resolution (Zhao et al. 2018a; Rudinger et al. 2018), information retrieval (Rekabsaz and Schedl 2020) text, and classification (Dixon et al. 2018; Barrett et al. 2019; Elazar and Goldberg 2018; De-Arteaga et al. 2019), as well as in language generation models (Sheng et al. 2019).

Mitigating the existence of societal biases in data and models has been the topic of several studies. Some pieces of work propose the debiasing of word embeddings by identifying and removing gender subspace (Ethayarajh, Duvenaud, and Hirst 2019b; Bolukbasi et al. 2016; Kaneko and Bollegala 2019; Zhao et al. 2018b). As pointed out by Gonen and Goldberg (2019), these methods successfully remove the *explicit bias*, while the *implicit bias*, examined through the ability of a classifier or a clustering algorithm to retrieve the gender of vectors after debiasing, still remains. Recently, Lauscher et al. (2020) propose a general framework for mitigating both forms of bias.

Another approach to address bias, and related to this paper, is bias reduction in corpus via data augmentation. Counterfactual Data Augmentation (CDA) is a common method, which, in its basic form, extends a corpus by adding new sentences, achieved from swapping the indicative words of the concept of interest in the corpus. Zhao et al. (2018a) use CDA in the context of coreference resolution, Lu et al. (2018) show the effectiveness of combination of CDA and embedding debiasing, Zmigrod et al. (2019) and later Maudslay et al. (2019) extend CDA to address bias in morphologically rich languages, as well as names. In this work, we use the basic CDA method to study the reaction of the bias measurement methods to the changes in the corpus.

High-Order Bias Measurements

We define bias as the discrepancy between the associations of a concept and its counterpart concept to a word. High-order bias measurement methods use vector similarity to quantify the associations. We define the concept Z (and similarly its counterpart concept Z') as a set, containing a group of representative words. In general, three approaches to high-order bias measurement are proposed in the literature, explained in what follows.

DIRECTIONAL: In this method, first a matrix of directional vectors D is created using a set of word pairs $\mathbb{P}_{Z,Z'} = \{(x, x') | x \in Z, x' \in Z'\}$, such that $D = \{v_x - v_{x'} | (x, x') \in \mathbb{P}_{Z,Z'}\}$, where v_x is the vector of the word x . Using D , the first principle component of the directional vectors is then calculated, which we refer to as v_d .

Bolukbasi et al. (2016) define the bias of the word w using the cosine similarity of v_w and v_d . Ethayarajh, Duvenaud, and Hirst (2019b) propose to normalize only v_d to avoid the overestimation of the association degree, resulting in the following bias measurement:

$$\psi(w) = \frac{v_d v_w}{\|v_d\|} \quad (1)$$

where $\psi(w)$ denotes the degree of bias of the word w , and its sign defines whether the word is biased towards the concept Z or Z' . This definition is applicable to all the bias measurements, discussed through the paper.

CENTROID: This method, used in several studies such as Garg et al. (2018) and Dev and Phillips (2019), first defines the representative vector v_Z as the mean of the embeddings of the representative words:

$$v_Z = \sum_{x \in Z} \frac{v_x}{|Z|} \quad (2)$$

The association of the concept Z to the word w is then defined using the cosine metric of v_Z and v_w . Finally, the bias is calculated as follows:

$$\psi(w) = \text{cosine}(v_Z, v_w) - \text{cosine}(v_{Z'}, v_w) \quad (3)$$

AVERAGE_{HIGH}: Caliskan, Bryson, and Narayanan (2017) introduce Word Embedding Association Test (WEAT), a statistical test to examine the existence of bias using vector similarity. Ethayarajh, Duvenaud, and Hirst (2019b) criticize WEAT, showing that the conclusion of the test can be manipulated by swapping gender-related concept words. We therefore only study the method used by Caliskan, Bryson, and Narayanan (2017) to measure the associations of words to concepts. This method calculates the average of the cosine similarities between the vector of the target word and the vectors of the concept words. We refer to this method as AVERAGE_{HIGH}, formulated as follows:

$$\text{AVERAGE}_{\text{HIGH}}(w, Z) = \frac{1}{|Z|} \sum_{x \in Z} \text{cosine}(v_x, v_w) \quad (4)$$

The bias using AVERAGE_{HIGH} is calculated as follows:

$$\psi(w) = \text{AVERAGE}_{\text{HIGH}}(w, Z) - \text{AVERAGE}_{\text{HIGH}}(w, Z')$$

Novel Bias Measurement

As discussed in Introduction, the first-order bias measurement requires the estimation of co-occurrence relations, which we provide using explicit representations. A well-known method to create explicit representations is based on the Point Mutual Information (PMI) metric. The PMI representation uses the count-based probabilities, where the co-occurrence relation between a word and a context-word in the PMI representation is calculated by $\log(p(w, c)/p(w)p(c))$. Positive PMI (PPMI) is a commonly-used variation, where negative values are replaced with zero. Levy and Goldberg (2014) draw the relation between word2vec SkipGram (SG) embeddings and PMI representations, showing that SG can be seen as a factorization of the PMI matrix shifted by $\log k$. Based on this idea, they propose the Shifted Positive PMI (SPPMI) representation by subtracting $\log k$ from PMI vector representations and setting the negative values to zero.

In the following of this section, we first explain our approach to creating the explicit variations of the SG and GloVe vectors, referred to as *explicit Skip-Gram (eSG)* and *explicit GloVe (eGloVe)*. Our approach reconstructs explicit representations from embedding vectors, and is related to previous studies such as Ethayarajh, Duvenaud, and Hirst (2019a), and Levy and Goldberg (2014). We then describe our first-order bias measurement method, defined based on any explicit vector.

Smoothed Explicit Representations

explicit Skip-Gram (eSG) The original SG model consists of two parameter matrices: word (\mathbf{V}) and context (\mathbf{U}) matrices, both of size $|\mathbb{V}| \times d$, where \mathbb{V} is the set of words in the collection and d is the embedding dimensionality. Given the word c , appearing in a context of word w , the model calculates $p(y = 1|w, c) = \sigma(\mathbf{v}_w \mathbf{u}_c^\top)$, where \mathbf{v}_w is the vector representation of w , \mathbf{u}_c context-vector of c , and σ denotes the sigmoid function. The SG model is optimized by maximizing the difference between $p(y = 1|w, c)$ and $p(y = 1|w, \tilde{c})$ for k negative samples \tilde{c} , randomly drawn from a noisy distribution \mathcal{N} . The \mathcal{N} distribution is set to the unigram distribution of the corpus, while downsampled by the context distribution smoothing parameter.

The $p(y = 1|w, c)$ term in the SG model measures the probability that the co-occurrence of two words w and c comes from the genuine co-occurrence distribution, derived from the training corpus. The model uses this probability to learn the embedding vectors, by separating these genuine co-occurrence relations from the sampled negative ones. We therefore use this estimation of the co-occurrence relations to define the vectors of the eSG representation, resulting to the following definition of eSG vector:

$$e_{w:c} = \sigma(\mathbf{v}_w \mathbf{u}_c^\top), \quad e_w = \sigma(\mathbf{v}_w \mathbf{U}^\top) \in \mathbb{R}^{|\mathbb{V}|} \quad (5)$$

where $e_{w:c}$ denotes the value of the corresponding dimension of the vector of w to the context-word c , and e_w in $|\mathbb{V}|$ dimensions is the explicit variation of the SG vector of word

w .²

We should note that the eSG representation is considerably different from the shifted PMI representation (Levy and Goldberg 2014): shifted PMI assumes very high embedding dimensions during the model training, while eSG draws the co-occurrence relations after the model is trained on low-dimensional embeddings. In fact, as SG is an implicit factorization of shifted PMI (Levy and Goldberg 2014; Ethayarajh, Duvenaud, and Hirst 2019a), eSG provides a smoothed variation of shifted PMI.

explicit GloVe (eGloVe) The GloVe model first defines an explicit matrix (size $|\mathbb{V}| \times |\mathbb{V}|$), where the corresponding co-occurrence value of each word and context-word is set to $\log p(w|c)$. This log probability is calculated based on the number of co-occurrences (denoted by $\#(\cdot, \cdot)$), such that $\log p(w|c) = \log \#(w, c) - \log \#(\cdot, c)$. We refer to this sparse explicit matrix as **initGlove**.

The GloVe model then implicitly factorizes the initGlove matrix, achieving two low-dimensional matrices of size $|\mathbb{V}| \times d$, as well as two bias vectors of size $|\mathbb{V}|$, where one assign a bias value to each word, and the other to each context-word. Using the same notation as SG, the factorization is done such that the dot products of the vectors of the matrices \mathbf{V} and \mathbf{U} plus the corresponding bias values estimate the logarithm of the co-occurrence values, as defined in the following:

$$\mathbf{v}_w \mathbf{u}_c^\top + b_w + \tilde{b}_c \approx \log \#(w, c) \quad (6)$$

where b_w and \tilde{b}_c denote the bias value of word w and context-word c , respectively.

Similar to eSG, the eGloVe representation estimates the co-occurrence relations using the word and context vectors, after training the GloVe model. Considering Eq. 6, we define the co-occurrence relations of eGloVe as the dot product of the word and context vectors,³ shown as follows:

$$e_{w:c} = \mathbf{v}_w \mathbf{u}_c^\top, \quad e_w = \mathbf{v}_w \mathbf{U}^\top \in \mathbb{R}^{|\mathbb{V}|} \quad (7)$$

The eGloVe representation in fact reconstructs initGlove, providing a smoothed variation.

First-Order Bias Measurement

The main difference between the first-order bias measurement method and the high-order approaches is in the estimation of the associations of a word to the concepts. We define our bias measurement method based on the AVERAGE_{HIGH} method, by replacing the cosine metric with co-occurrence relations, and therefore refer to it as AVERAGE_{FIRST}. Given an explicit vector denoted as e , AVERAGE_{FIRST} is defined as

²An alternative formulation of eSG is to normalize it by dividing its values with the square root of the expectations of the co-occurrence relations for each word and context-word, namely $\mathbb{E}_{c' \sim \mathcal{N}} \sigma(\mathbf{v}_w \mathbf{u}_{c'}^\top)$ and $\mathbb{E}_{w' \sim \mathcal{N}} \sigma(\mathbf{v}_{w'} \mathbf{u}_c^\top)$. We study this variation in pilot experiments, observing similar results to the introduced variation. We therefore stay with the less complex formulation (Eq. 5).

³As in eSG, eGloVe does not need extra normalization, since the bias terms b_w and \tilde{b}_c in Eq. 6, learned during training, act as normalizers to the co-occurrence estimation $\log \#(w, c)$.

the mean of the co-occurrence values of w with the representative words Z , formulated as follows:

$$\text{AVERAGE}_{\text{FIRST}}(w, Z) = \frac{1}{|Z|} \sum_{c \in Z} e_{w:c} \quad (8)$$

The bias toward Z is then defined as the differences between the associations of w to Z and Z' :

$$\psi(w) = \text{AVERAGE}_{\text{FIRST}}(w, Z) - \text{AVERAGE}_{\text{FIRST}}(w, Z')$$

As shown, $\text{AVERAGE}_{\text{FIRST}}$ only considers the context-words related to the Z and Z' concepts. This avoids the influence of other non-relevant concepts as in the high-order bias measurements.

Let us review the required calculations for $\text{AVERAGE}_{\text{FIRST}}(w, Z)$ when using the introduced smoothed explicit representations, namely eSG or eGloVe. In this setting, the main computation of the bias measure is the dot products of the vector of w to the context-vectors of the words in Z . We should note that the computational complexity of this calculation is the same as the one of $\text{AVERAGE}_{\text{HIGH}}(w, Z)$ on SG/GloVe. In fact, considering the dot product of two embedding vectors as computation unit, the complexity of both bias measurement methods is $\mathcal{O}(|Z|)$.

Finally, a practical consideration in calculating eSG and eGloVe is that this computation requires the context vectors in addition to word vectors. These context vectors are commonly stored in the libraries used for SG and GloVe embeddings alongside the word vectors, mainly for the purpose of continuous training. In eSG/eGloVe, these context vectors are exploited to estimate smoothed first-order relations.

Gender Bias Experiment Design

In this section, we explain the design of our experiment and the resources. Our source code together with all resources, including the lists of occupational and gender-representative words are publicly available.⁴

Word Representations. We conduct our experiments on the PMI, PPMI, SPPMI, SG, eSG, GloVe, and eGlove representations. In addition, we create the low-dimensional vectors of the the PMI-based representations using Singular Value Decomposition (SVD), referred to as PMI-SVD, PPMI-SVD, and SPPMI-SVD. These word representation models are created on the English Wikipedia corpus of August 2017. We project all characters to lower case, and remove numbers and punctuation marks. For all models, we use the window size of 5, and filter the words with frequencies lower than 200, resulting in 197,549 unique words. The number of dimensions of the embeddings are set to 300. The rest of the parameters are set using the default parameter setting of the word2vec Skip-Gram model in the Gensim library (Rehurek and Sojka 2010), and the GloVe model in the provided tool by its authors. As suggested by (Levy, Goldberg, and Dagan 2015), we apply subsampling and context distribution smoothing on all PMI-based models with the same parameter values as the SG model.

⁴<https://github.com/navid-rekabsaz/SmoothedFirstOrderBias>

Gender-Representative Words. In all bias measurements, the concepts Z and Z' are assumed as female and male, respectively. Therefore, a positive bias value indicates the inclination towards female, and a negative one towards male. The concepts are defined using two sets, each with 28 words, containing words like *she, her, woman* for the female and *he, his, man* for the male concept. These sets are taken from previous studies (Bolukbasi et al. 2016; Garg et al. 2018). From the same sets, we form the gender pairs, used in the CDA method, and listed in Supplemental Material.

Occupational Words. We provide a list of 496 occupations, among which 17 are female-specific (e.g. *congresswoman*), 9 male-specific (e.g. *congressman*), and the rest are gender-neutral (e.g. *nurse* and *dancer*). The set and assigned genders are listed in Supplemental Material.

Job market statistics. We study the correlation of the calculated gender bias values with the statistics of the gender bias of a set of occupations, obtained from the U.S. job market. These statistics are provided by two collections, where in each the bias for an occupation is the percent of people in the occupation who are reported as female (e.g. 90% of nurses are women). The first collection uses the data provided by Zhao et al. (2018a). The collection contains the statistics of 40 occupations, gathered from the U.S. Department of Labor. We refer to the collection as Labor Data. The second is provided by Garg et al. (2018) using the U.S. census data. From the provided data, we use the gender bias statistics of the year 2015 – the most recent year in the collection, resulting in a list of 96 occupations. We refer to this as Census Data. We use Spearman ρ and Pearson’s r correlations.

Experiments and Results

In this section, we first present the results of correlation studies with job market statistics, followed by analysing the results of the high-order bias measurement and discussing the characteristics of the two approaches. We then visualize and compare the gender bias of occupational words when measured with the first-order versus high-order approach, and finally study the reaction of the bias measurement approaches to bias reduction in the corpus.

Correlation Results with Job Market Statistics

We calculate the gender bias of the occupations in the Labor Data and Census Data collections, using the high-order bias measurements (DIRECTIONAL , CENTROID , and $\text{AVERAGE}_{\text{HIGH}}$) on both low-dimensional and explicit representations, and the first-order bias measurement ($\text{AVERAGE}_{\text{FIRST}}$) on explicit vectors.

Table 1 (next page) shows the results of the correlation between the calculated gender bias, and the gender bias statistics, provided by the Labor Data and Census Data collections. Each section of the table is assigned to a family of the models, namely PMI, PPMI, SPPMI, GloVe, and SG. The highest correlation results of each section are shown in bold, and the highest overall correlations are indicated with underlines.

In all families of representations, the first-order bias measurement shows higher correlations than the high-order

Order	Representation	Method	Labor Data		Census Data	
			Spearman ρ	Pearson's r	Spearman ρ	Pearson's r
High-Order	PMI	DIRECTIONAL	0.28	0.07	0.18	0.02
		CENTROID	0.14	0.21	0.35	0.40
		AVERAGE _{HIGH}	0.33	0.24	0.27	0.19
	PMI-SVD	DIRECTIONAL	0.05	0.07	0.00	0.00
		CENTROID	0.41	0.47	0.46	0.53
		AVERAGE _{HIGH}	0.41	0.49	0.49	0.56
First-Order	PMI	AVERAGE _{FIRST}	0.53	0.51	0.57	0.62
High-Order	PPMI	DIRECTIONAL	0.45	0.49	0.39	0.47
		CENTROID	0.43	0.46	0.45	0.50
		AVERAGE _{HIGH}	0.43	0.46	0.45	0.52
	PPMI-SVD	DIRECTIONAL	0.05	0.07	0.00	0.00
		CENTROID	0.41	0.47	0.46	0.53
		AVERAGE _{HIGH}	0.41	0.49	0.49	0.56
First-Order	PPMI	AVERAGE _{FIRST}	0.59	0.58	0.64	0.64
High-Order	SPPMI	DIRECTIONAL	0.26	0.37	0.26	0.28
		CENTROID	0.39	0.45	0.45	0.48
		AVERAGE _{HIGH}	0.32	0.40	0.44	0.48
	SPPMI-SVD	DIRECTIONAL	0.17	0.29	0.11	0.03
		CENTROID	0.28	0.35	0.39	0.43
		AVERAGE _{HIGH}	0.26	0.38	0.36	0.46
First-Order	SPPMI	AVERAGE _{FIRST}	0.57	0.49	0.52	0.48
High-Order	GloVe	DIRECTIONAL	0.53	0.56	0.34	0.46
		CENTROID	0.58	0.60	0.39	0.51
		AVERAGE _{HIGH}	0.60	0.60	0.39	0.51
First-Order	initGlove eGloVe	AVERAGE _{FIRST}	0.38	0.42	0.40	0.51
		AVERAGE _{FIRST}	0.56	0.57	0.42	0.52
High-Order	SG	DIRECTIONAL	0.50	0.54	0.58	0.64
		CENTROID	0.55	0.57	0.60	0.65
		AVERAGE _{HIGH}	0.55	0.57	0.59	0.65
First-Order	eSG	AVERAGE _{FIRST}	0.66	0.61	0.67	0.70

Table 1: Spearman ρ and Pearson's r correlation results of the gender bias values, calculated with word representations, to the statistics of the portion of women in occupations

methods, on both collections and correlation metrics (with the exception of GloVe on Labor Data). Overall, eSG with AVERAGE_{FIRST} shows the highest correlations across the combinations. Interestingly, PPMI, as a simple count-based method, when combined with AVERAGE_{FIRST} shows higher correlations in comparison with any high-order bias measurement combined with SG or GloVe, on Census Data collection.

The results indicate that the calculated bias values using the first-order method more closely follows the distributions of the gender bias indicators for occupations. This emphasizes the importance of exploiting the first-order bias measurement approach, as an alternative to the high-order approaches. The results of these experiments are particularly applicable to the studies, which explore the correla-

tions of the text-based quantities with various external indicators (such as job market statistics) Caliskan, Bryson, and Narayanan (2017); Garg et al. (2018), as in these work the quantities are only calculated based on the high-order bias measurements.

Comparing the correlation results of high-order methods, overall, AVERAGE_{HIGH} and CENTROID show similar results, which are slightly higher than the ones of DIRECTIONAL. Considering the results of this study, in the following, we focus on analyzing the eSG and SG representations, and AVERAGE_{HIGH} among the three high-order measurements.

Analysis and Discussion

In this section, we first diagnose the results of the high-order bias measurement, and then discuss and compare the charac-

manicurist:	businesswoman, <u>nurse</u> , Filipina, seamstress, matron
midwife:	<u>midwife</u> , <u>nurse</u> , <u>feminist</u> , matron, suffragist
nurse:	<u>midwife</u> , <u>nurse</u> , matron, nursing, Filipina
socialite:	businesswoman, Filipina, <u>suffragist</u> , <u>feminist</u> , hostess
housekeeper:	matron, <u>midwife</u> , <u>nurse</u> , maid, governess

captain:	<u>commanded</u> , <u>capt</u> , <u>quartermaster</u> , <u>enlisted</u> , Hugh
colonel:	<u>commanded</u> , Hugh, Ernest, <u>guards</u> , <u>quartermaster</u>
mechanician:	<u>apprenticed</u> , Cyril, Ernest, Messrs, surveyor
lieutenant:	<u>commanded</u> , Ernest, Hugh, <u>enlisted</u> , <u>quartermaster</u>
engineer:	Jagmal, <u>surveyor</u> , <u>apprenticed</u> , draughtsman, <u>engineer</u>

Table 2: Context-words with the highest effects on the calculated gender bias with the AVERAGE_{HIGH} method. Gender-neutral context-words are shown with underlines.

teristics of the two bias measurement approaches.

We start by exploring which context-words in practice contribute the most to the calculated gender bias of the occupational words with AVERAGE_{HIGH}. To this end, we examine the measured gender bias of the occupational words using AVERAGE_{HIGH}, applied on the eSG vectors (instead of the SG vectors). We use eSG since its explicit vectors enable the diagnosis of the results, particularly by looking at the context-words with the highest contributions.

To diagnose the measured bias of AVERAGE_{HIGH}, we first create the explicit variations of each female- and male-representative word vector, denoted as e_z and $e_{z'}$ (see section Smoothed Explicit Representations). Given the occupation w with explicit vector e_w , we calculate the high-order bias with AVERAGE_{HIGH}, and provide its element-wise results by removing the summation of the cosine function, formulated as follows:

$$e_\psi = \frac{1}{|Z|} \sum_{z \in Z} \frac{e_z \odot e_w}{\|e_z\| \|e_w\|} - \frac{1}{|Z'|} \sum_{z' \in Z'} \frac{e_{z'} \odot e_w}{\|e_{z'}\| \|e_w\|}$$

where \odot denotes the element-wise product, and $e_\psi \in \mathbb{R}^{|V|}$ is the gender bias results of w for all context-words, each correspond to one dimension.

To select a representative set of words for the analysis, we first calculate the gender bias of all target words using AVERAGE_{HIGH} on SG vectors. We then select the top 5 gender-neutral occupations with the highest bias towards female, and the same towards male. These top female and male words are shown in the top and bottom part of Table 2, respectively. For each word, we show the top 5 context-words with the largest influence on the results of the gender bias measurement, namely the highest absolute values of the dimensions in e_ψ .

In these representative samples, we can observe the existence of various groups of context-words: several gender-neutral context-words like *nurse*, *commanded*, and *apprenticed* shown with underlines; several gender-specific words, which also contain other concepts such as an occupation like *matron* and *actress*; and finally person names such as *Ernest* and *Cyril*.

As discussed in Introduction, some of these context-words are arguably irrelevant to what we aim to characterize, namely the extent to which a word is perceived as female/male. For instance, the association of *manicurist* with female is influenced by context-words such as *nurse* and *businesswoman*, or in the association of *captain* with male is affected by *commanded* and *enlisted*. This issue is avoided in AVERAGE_{FIRST}, as the method calculates the associations by only considering a pre-defined set of highly-relevant context-words.

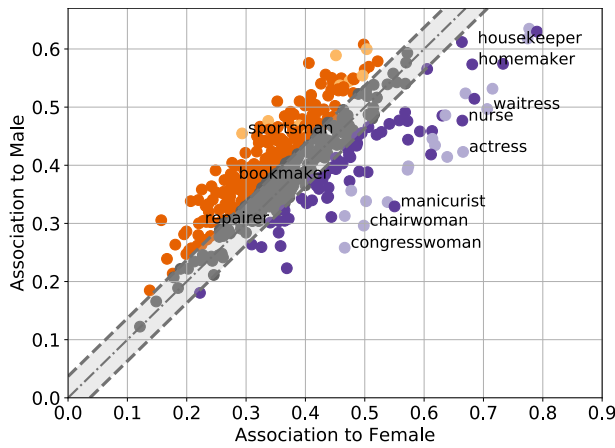
Finally, let us put forward the question of *which approach should be applied to bias measurement using text corpora?* Indeed answering this question, foremost, requires a clear definition of what is aimed to be captured as bias. Our study proposes AVERAGE_{FIRST}, as a complementary to the existing high-order approaches, which exploits smoothed first-order co-occurrence relations of the highly-relevant context-words. This characteristic makes AVERAGE_{FIRST} particularly appropriate for bias measurement scenarios, in which capturing the direct relations between the words and the concept of interest is favorable. However, we should note that, depending on the objective of bias measurement, utilizing second- or higher-order co-occurrences can be also highly important. For instance, if the aim of a study is to measure the stereotypical bias of *girls* and *boys* towards the color *pink*, it might be favorable to take into account the second-order co-occurrences, for example the ones through *cloth*. However, even in this scenario, we should still be aware that utilizing the discussed high-order bias measurements might yet not be an ideal approach, as these methods are still affected by other non-relevant second- or higher-order context-words. An optimal bias measurement method for such a scenario still remains an open question, and studying it is a potential direction for future work.

Visualization of Gender Bias Results

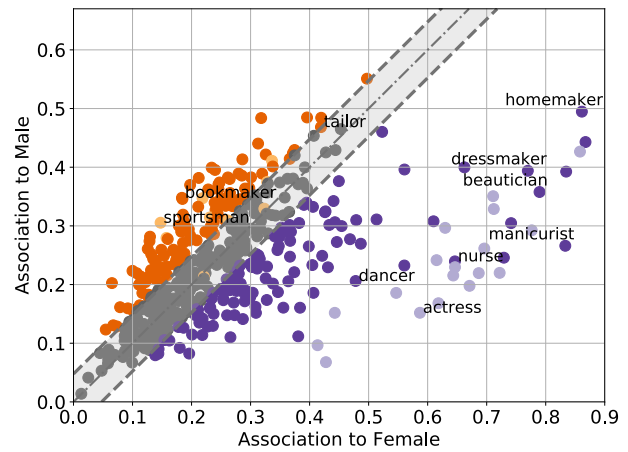
In this section, we continue our analyses by visualizing the gender bias of all occupations, measured using AVERAGE_{HIGH} on SG, and AVERAGE_{FIRST} on eSG.

The results of the high- and first-order methods are shown in Figures 2a and 2b, respectively. To make the results visually comparable, we apply Min-Max normalization to the measured associations of each approach, where the min/max values are calculated over the measured associations of all words in vocabulary to male and female.

In both bias measurement methods, we are interested in distinguishing between the words with significantly higher bias values and any random word with low bias values. To do this, we define a threshold for each plot, below which the words are considered as unbiased. To find the thresholds for distinguishing between the words with significantly higher bias values and any random word with low bias values, since



(a) High-order measurement ($AVERAGE_{HIGH}$)



(b) First-order measurement ($AVERAGE_{FIRST}$)

Figure 2: The associations of the occupations to the female and male concepts, indicating their inclinations towards the genders. The ones in the gray area are considered as unbiased. The occupations, inclined towards female and male are shown in purple and orange, respectively, where among them the gender-neutral ones, namely the occupations with stereotyped biases, have darker colors.

the number of biased words to a concept are limited, we assume that there is a high probability that any randomly sampled word is unbiased. We therefore define the threshold for each bias measurement method as the mean of the absolute bias values of all words in vocabulary. These thresholds for $AVERAGE_{HIGH}$ with SG and $AVERAGE_{FIRST}$ with eSG are 0.036 and 0.047, respectively.

In each plot, the area where the distance from the diagonal is less than the corresponding threshold value, is referred to as the unbiased area, and shown in gray. The occupations located in the unbiased areas are considered as unbiased (no significant stereotypical bias is observed).

A gender-neutral occupation is considered to be stereotypically biased to either female or male, when it is inclined towards the female/male associations, namely when it is located below/above the unbiased area, shown in dark purple and dark orange, respectively. The plot also depicts the gender-specific occupations (e.g. *actress*, *sportman*), while we should consider that these occupations are not stereotypically biased, but are expected to be inclined towards their respective genders. We show these gender-specific occupations in light colours, namely light purple for female, and light orange for male. As expected, all gender-specific occupation words are on the correct side of the diagonal for both measures.

Both figures show the existence of significant gender bias in several occupations. However, Figure 2a and 2b provide considerably different perspectives on the extents of the associations of the occupations to genders. In particular, the $AVERAGE_{FIRST}$ method shows relatively larger degrees of bias towards female, specially for some gender-neutral occupations such as *nurse*, and *housekeeper*.

To have a better view on the distribution of the bias values, Figure 3 shows the histogram of the occupations over the range of the bias values, measured with $AVERAGE_{HIGH}$ and

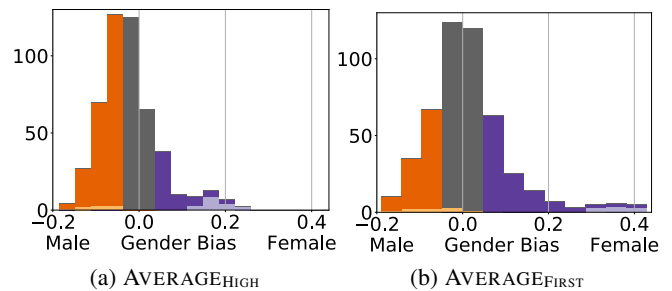


Figure 3: Histograms of the gender bias of the occupations, measured using $AVERAGE_{HIGH}$ and $AVERAGE_{FIRST}$.

$AVERAGE_{FIRST}$. Similar to Figure 2, the gray color shows the number of unbiased occupations, and the purple and orange colors indicate the number of biased ones towards female and male in each bin, among which the gender-specific ones are shown with light colors.

In both measurement methods, a larger number of occupations are biased to male. However, the first-order bias measurement captures a larger degree of bias towards female, suggesting the existence of a more severe degree of female bias, previously undetected by the high-order approach.

Sensitivity to Changes in Corpus

How do the bias measurement methods react to (hypothetical) future changes, especially when the provided corpora move towards more balanced representations of genders? To explore this question, we use the basic form of the CDA method (see Related Work) to expand the corpus with synthetic augmented data. Using the list of gender words pairs, for every sentence in the corpus, we swap each female word with its counter-part male word, and vice versa. We create

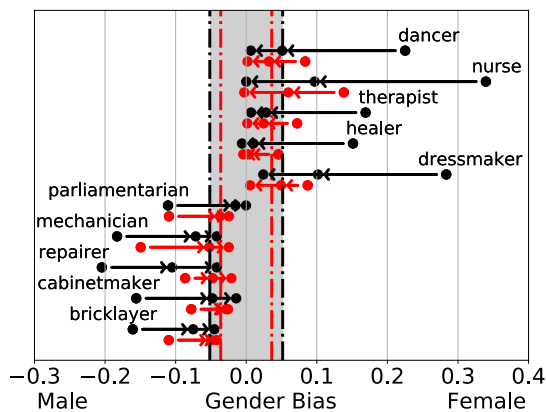


Figure 4: Changes in the gender bias values, measured on the original corpus and two augmented ones with the CDA method. The black and red arrows indicate the results of $AVERAGE_{FIRST}$ and $AVERAGE_{HIGH}$, respectively.

two new augmented corpora: the first one expands the original corpus with only one half of the new sentences (selected randomly), while the second one adds all the new sentences to the original corpus.

We create the SG models, and consecutively the eSG representations, based on these two new corpora. We then measure gender bias of the corpora using SG with $AVERAGE_{HIGH}$, and eSG vectors with $AVERAGE_{FIRST}$. Figure 4 shows the results of the changes in the bias values of 5 randomly-selected gender-neutral occupations for each gender, which are perceived as biased in the original corpus by both bias measurements. Each arrow shows the direction of the changes, where the starting point indicates the measured bias in the original corpus, and the middle and final points are calculated based on the first and second augmented corpora. The black and red arrows show the results of $AVERAGE_{FIRST}$ and $AVERAGE_{HIGH}$, respectively. The gray areas, bordered with the black and red vertical dashed lines show the unbiased areas of the results of $AVERAGE_{FIRST}$ and $AVERAGE_{HIGH}$, respectively. As mentioned in previous subsection, since in each method the values are normalized over all words, the bias values across the methods are reasonably comparable.

As shown, in both methods the gender bias of all occupations consistently decrease (move towards zero). The gender bias values measured on the second augmented corpus, for all the occupations, reach the corresponding unbiased area of each method, indicating that these occupations in this corpus are considered as unbiased. Comparing the bias measurement methods, for these occupations, the $AVERAGE_{FIRST}$ approach indicates the existence of a higher degree of bias in the initial corpus in comparison with $AVERAGE_{HIGH}$, and also demonstrates a larger degree of bias reduction in each step. The aggregated degrees of changes over all gender-neutral occupations are reported Table 3. As shown, in both steps the $AVERAGE_{FIRST}$ demonstrate larger changes in comparison with $AVERAGE_{HIGH}$.

These larger changes in the measured bias indicate the

Corpus Augmentation	$AVERAGE_{HIGH}$ with SG	$AVERAGE_{FIRST}$ with eSG
First Step	0.037	0.044
Second Step	0.012	0.016

Table 3: Mean of the absolute values of bias changes in each step of corpus augmentation experiments, discussed

higher sensitivity of the $AVERAGE_{FIRST}$ approach to the changes in the corpus in comparison with $AVERAGE_{HIGH}$, which is a valuable characteristic of the first-order method particularly in the study of social changes using text data.

Conclusion

We highlight a potential issue of the commonly-used high-order bias measurement approaches, and propose a novel approach based on the smoothed first-order co-occurrence relations between words and their context words. To estimate these relations, we reconstruct explicit vectors from pre-trained word2vec Skip-Gram, and GloVe embeddings, proposing smoothed first-order bias measurement method.

Let us visit again the question of “*which approach should we use for measuring bias (or any societal phenomena) using text corpora?*” To answer this question, we recommend the reader to first contemplate: “*what construct/quality is aimed to be captured as bias*”. As an example, measuring “*to what extent different gender-neutral occupations are perceived differently*” – as pursued in our experiments – arguably looks for a different construct/realization of bias in comparison with measuring “*what the stereotypical relation of color pink is towards girls and boys*”. The smoothed first-order approach offers an alternative to high-order approaches, and particularly has the advantage of narrowing down the measurement to the effect of only a specific set of representative words, appearing in the vicinity of the word under study.

To have a better understanding of the differences of these two approaches, our experiments examine and compare the measured gender bias of a set of occupational words, calculated on a Wikipedia text corpus. We observe that the gender bias calculated with the proposed first-order approach correlates higher with the gender-related statistics of the job market in comparison to the high-order methods. Our analyses provide showcases of the existence of non-relevant context-words when using high-order measurements, and suggests the existence of a more severe degree of female-bias regarding some jobs. While our experiments are conducted on gender bias, the introduced method is generic and can be extended to study other forms of societal biases such as related to race, age and ethnicity.

Supplemental Material

Gender Definitional Words

Female: *girl, girls, sister, sisters, mom, moms, mother, mothers, fiancée, grandmother, grandma, granddaughter, granddaughters,*

she, her, herself, hers, gal, gals, female, females, woman, women, madam, daughter, daughters, stepmother, stepdaughter

Male: boy, boys, brother, brothers, dad, dads, father, fathers, fiance, grandfather, grandpa, grandson, grandsons, he, him, himself, his, lad, lads, male, males, man, men, sir, son, sons, stepfather, stepson

Pairs: (boy, girl), (boys, girls), (brother, sister), (brothers, sisters), (dad, mom), (dads, moms), (father, mother), (fathers, mothers), (fiance, fiancée), (grandfather, grandmother), (grandpa, grandma), (grandson, granddaughter), (grandsons, granddaughters), (he, she), (him, her), (himself, herself), (his, hers), (lad, gal), (lads, gals), (male, female), (males, females), (man, woman), (men, women), (sir, madam), (son, daughter), (sons, daughter), (stepfather, stepmother), (stepson, stepdaughter)

Target Words

Female-specific occupations: actress, ballerina, barmaid, businesswoman, chairwoman, chambermaid, congresswoman, forewoman, housewife, landlady, maid, masseuse, matron, mistress, policewoman, sportswoman, stewardess, stuntwoman, usherette, waitress

Male-specific occupations: barman, businessman, foreman, landlord, policeman, salesman, serviceman, sportsman, waiter

Gender-neutral occupations: accountant, actor, administrator, adventurer, adviser, advocate, agent, aide, ambassador, analyst, animator, announcer, anthropologist, apprentice, archaeologist, archeologist, architect, archivist, artist, artiste, assassin, assessor, assistant, astrologer, astronaut, astronomer, athlete, attendant, attorney, auctioneer, auditor, author, bailiff, baker, ballplayer, banker, barber, bargee, baron, barrister, bartender, basketmaker, beautician, beekeeper, bibliographer, biochemist, biologist, biotechnologist, blacksmith, boatman, bodyguard, boilerfitter, boilermaker, bookbinder, bookkeeper, bookmaker, bootmaker, boss, botanist, boxer, breeder, brewer, bricklayer, broadcaster, broker, bureaucrat, butcher, butler, buyer, cabbie, cabinetmaker, cameraman, campaigner, captain, cardiologist, caretaker, carpenter, cartographer, cartoonist, cashier, cellist, ceo, ceramicist, chairman, chancellor, chaplain, character, chef, chemist, chief, chiropractor, choreographer, cinematographer, citizen, cleaner, clergy, cleric, clerical, clerk, coach, collector, colonel, columnist, comedian, comic, commander, commentator, commissioner, composer, compositor, concreter, conductor, confectioner, confessor, congressman, conservationist, conservator, constable, consultant, cook, cooper, cop, coremaker, correspondent, councilman, councilor, counsellor, counselor, courtier, critic, crooner, croupier, crusader, crusher, curator, custodian, cutler, dancer, dean, decorator, dentist, deputy, dermatologist, designer, detective, developer, dietician, dietitian, digger, diplomat, director, dispatcher, diver, doctor, doorkeeper, draughtsman, draughtsperson, dresser, dressmaker, drilller, driver, drummer, drycleaner, dyer, ecologist, economist, editor, educator, electrician, embroiderer, employee, engineer, engraver, entertainer, entrepreneur, environmentalist, envoy, epidemiologist, ergonomist, ethnographer, evangelist, expert, farmer, farrier, filmmaker, financier, firebrand, firefighter, fisherman, fitter, footballer, furrier, gangster, gardener, geneticist, geographer, geologist, geophysicist, gilder, glassmaker, glazier, goalkeeper, goatherd, goldsmith, gravedigger, grinder, guard, guide, guitarist, gunsmith, hairdresser, handler, handyman,

hardener, harpooner, hatter, headmaster, healer, herbalist, historian, homemaker, hooker, housekeeper, hydrologist, illustrator, industrialist, infielder, inspector, instructor, insulator, interpreter, inventor, investigator, janitor, jeweler, jeweller, joiner, journalist, judge, jurist, keeper, knitter, laborer, labourer, lacemaker, lawmaker, lawyer, lecturer, legislator, librarian, lieutenant, life-guard, lithographer, lumberjack, lyricist, machinist, maestro, magician, magistrate, maltster, manager, manicurist, marksman, marshal, mason, masseur, master, mathematician, mechanic, mechanician, mediator, medic, melter, merchandiser, metallurgist, metalworker, meteorologist, metrologist, microbiologist, midfielder, midwife, miller, millwright, miner, minister, missionary, mobster, model, modeller, mover, musician, musicologist, nanny, narrator, naturalist, negotiator, neurologist, neurosurgeon, novelist, nurse, nutritionist, observer, obstetrician, officer, official, operator, optician, optometrist, organiser, organist, orthotist, owner, packer, paediatrician, painter, palmists, paperhanger, paralegal, paramedic, parishioner, parliamentarian, pastor, pathologist, patrolman, patternmaker, paver, pawnbroker, pediatrician, pedicurist, performer, pharmacist, philanthropist, philosopher, photographer, photojournalist, physician, physicist, physiotherapist, pianist, pilot, planner, plasterer, playwright, plumber, poet, police, politician, pollster, porter, potter, poulterer, preacher, president, priest, principal, prisoner, producer, professor, programmer, projectionist, promoter, prompter, proprietor, prosecutor, prosthetist, protagonist, protege, protester, provost, psychiatrist, psychologist, psychotherapist, publicist, publisher, pundit, radiographer, radiologist, radiotherapist, ranger, realtor, receptionist, referee, refiner, registrar, repairer, reporter, representative, rescuer, researcher, restaurateur, retoucher, rigger, roaster, roofer, sailor, saint, sales, salesperson, sausagemaker, saxophonist, scaffolder, scholar, scientist, screenwriter, scriptwriter, sculptor, seaman, secretary, senator, student, servant, setter, sewer, shepherd, sheriff, shoemaker, shopkeeper, shunter, singer, skipper, smith, socialite, sociologist, soldier, solicitor, soloist, songwriter, specialist, spinner, sportswriter, staff, statesman, statistician, steelworker, steeplejack, steward, stockbroker, stonecutter, stonemason, storekeeper, strategist, student, stuntman, stylist, substitute, superintendent, supervisor, surgeon, surveyor, sweep, swimmer, tailor, tamer, tanner, tannery, teacher, technician, technologist, telecaster, teller, therapist, tinsmith, toolmaker, tracklayer, trader, trainer, trainman, translator, traveller, treasurer, trooper, trucker, trumpeter, tuner, turner, tycoon, typesetter, tyre fitter, undersecretary, understudy, upholsterer, usher, valedictorian, valuer, varnisher, vendor, veterinarian, viticulturist, violinist, vocalist, warden, warrior, washer, weaver, weigher, welder, whaler, wigmaker, worker, wrestler, writer, zookeeper

References

- Barrett, M.; Kementchedjheva, Y.; Elazar, Y.; Elliott, D.; and Søggaard, A. 2019. Adversarial Removal of Demographic Attributes Revisited. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* .

- Chang, S.; and McKeown, K. 2019. Automatically Inferring Gender Associations from Language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Church, K. W.; and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*.
- De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.
- Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*.
- Dev, S.; and Phillips, J. 2019. Attenuating Bias in Word vectors. In *The International Conference on Artificial Intelligence and Statistics*.
- Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Elazar, Y.; and Goldberg, Y. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Ethayarajh, K.; Duvenaud, D.; and Hirst, G. 2019a. Towards Understanding Linear Word Analogies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Ethayarajh, K.; Duvenaud, D.; and Hirst, G. 2019b. Understanding Undesirable Word Embedding Associations. In *Proceedings of the Conference of the Association for Computational Linguistics*.
- Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*.
- Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *Proceedings of the North American Chapter of the Association for Computational Linguistics*.
- Günther, F.; Rinaldi, L.; and Marelli, M. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*.
- Hoyle, A.; Wallach, H.; Augenstein, I.; and Cotterell, R. 2019. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. In *Proceedings of the Conference of the Association for Computational Linguistics*.
- Kaneko, M.; and Bollegala, D. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Kirchler, E. 1992. Adorable woman, expert man: Changing gender images of women and men in management. *European Journal of Social Psychology*.
- Kiritchenko, S.; and Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*.
- Kontostathis, A.; and Pottenger, W. M. 2006. A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*.
- Lauscher, A.; Glavaš, G.; Ponzetto, S. P.; and Vulić, I. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Lenton, A. P.; Sedikides, C.; and Bruder, M. 2009. A latent semantic analysis of gender stereotype-consistency and narrowness in American English. *Sex Roles*.
- Levy, O.; and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the International Conference on Neural Information Processing Systems*.
- Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*.
- Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; and Datta, A. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.
- Maudslay, R. H.; Gonen, H.; Cotterell, R.; and Teufel, S. 2019. Its All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems*.
- Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*.
- Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Rehurek, R.; and Sojka, P. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*.

Rekabsaz, N.; Lupu, M.; Baklanov, A.; Dür, A.; Andersson, L.; and Hanbury, A. 2017. Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Rekabsaz, N.; and Schedl, M. 2020. Do Neural Ranking Models Intensify Gender Bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2065–2068.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of North American Chapter of the Association for Computational Linguistics*.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhou, P.; Shi, W.; Zhao, J.; Huang, K.-H.; Chen, M.; Cotterell, R.; and Chang, K.-W. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the Conference of the Association for Computational Linguistics*.