Joint Multiclass Debiasing of Word Embeddings

Radomir Popović[1(✉)], Florian Lemmerich[1], and Markus Strohmaier[1,2]

[1] RWTH Aachen University, Aachen, Germany
`radomir.popovic@rwth-aachen.de`
[2] Gesis-Leibniz Institute for Social Sciences, Mannheim, Germany

Abstract. Bias in Word Embeddings has been a subject of recent interest, along with efforts for its reduction. Current approaches show promising progress towards debiasing single bias dimensions such as gender or race. In this paper, we present a joint multiclass debiasing approach that is capable of debiasing *multiple* bias dimensions simultaneously. In that direction, we present two approaches, HardWEAT and SoftWEAT, that aim to reduce biases by minimizing the scores of the Word Embeddings Association Test (WEAT). We demonstrate the viability of our methods by debiasing Word Embeddings on three classes of biases (religion, gender and race) in three different publicly available word embeddings and show that our concepts can both reduce or even completely eliminate bias, while maintaining meaningful relationships between vectors in word embeddings. Our work strengthens the foundation for more unbiased neural representations of textual data.

Keywords: Word embedding · Bias reduction · Joint debiasing · WEAT

1 Introduction

Word Embeddings, i.e., the vector representation of natural language words, are key components of many state-of-the art algorithms for a variety Natural Language Processing tasks, such as Sentiment Analysis or Part of Speech Tagging. Recent research established that popular embeddings are prone to substantial biases, e.g., with respect to gender or race [2,6], which demonstrated in results like *"Man is to Computer Programmer as Woman is to Homemaker"* [2] as results of basic analogy tasks. Since such biases can potentially have an effect on downstream tasks, several relevant approaches for debiasing existing word embedding have been developed. A common deficit of existing techniques is that debiasing is limited to a single bias dimension (such as gender). Thus, in this paper, we propose two new post-processing methods for joint/simultaneous multiclass debiasing, which differ in their trade-off between maintaining word relationships and decreasing bias levels: HardWEAT completely eliminates contained bias as measured by the established Word Embedding Association Test

[3]. SoftWEAT has a stronger and tunable emphasis on maintaining the original relationships between words in addition to bias removal. We demonstrate the effectiveness of our approach on the bias dimensions gender, race and religion on three conventional Word Embedding models: FastText, GloVe and Word2Vec.

2   Background and Related Work

In this section, we introduce key concepts formally and discuss related work on debiasing word embeddings.

We assume a *Word Embedding* with vocabulary size v that maps each word w to a vector representation $\vec{w} \in \mathbf{R}^d$. Protected *classes* c $\in$ C are entities on which bias can exist, e.g., race, religion, gender, or age. *Subclasses* S or *target set of words* refer to directions of a class, e.g., *male, female* for $c = gender$, and are associated with a set of definitional words S_c. The set of *neutral words* N contains all words that should not be associated with any $S_c \forall c \in C$. Finally, *attribute sets of words* A (A $\subset$ N) denote word sets that a target set of words can potentially be linked to, e.g., {*pleasant, nice, enjoyable*} or {*science, research, academic*}.

The Word Embedding Association Test. The state-of-the art way of measuring biases in embeddings is the Word Embedding Association Test (WEAT) [3]: It considers two sets of *attribute words* (A, B), e.g., family and career related words, and two target sets (X, Y), e.g., black and white names. The null hypothesis of this test states, that the relative association of target sets' words to both attribute sets' words are equally strong. Thus, rejecting the null hypothesis asserts bias. To examine this, a test statistic s quantifies how strongly X is associated to A in comparison to the association between Y and B. It is computed as $s(X, Y, A, B) = \sum_{\vec{x} \in X} h(\vec{x}, A, B) - \sum_{\vec{y} \in Y} h(\vec{y}, A, B)$, where $h(\vec{w}, A, B) = \text{mean}_{\vec{a} \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{\vec{b} \in B} \cos(\vec{w}, \vec{b})$ describes the relative association between a single target word $x \in X$ compared to the two attribute sets in a range $[-2, 2]$. Based on s, we assess the statistical significance via an one-sided permutation test through which we acquire p value; Additionally, an effect size d that quantifies the severity of the bias can be calculated as:

$$d(X, Y, A, B) = \frac{\text{mean}_{\vec{x} \in X} h(\vec{x}, A, B) - \text{mean}_{\vec{y} \in Y} h(\vec{y}, A, B)}{\text{std}_{w \in X \cup Y} h(\vec{w}, A, B)}$$

We will use this effect size in our novel debiasing approaches.

Existing Debiasing Techniques. To achieve reduction of bias, we will substantially extend the work of Bolukbasi et al. [2], which describes two ways of debiasing Word Embeddings in a post-processing step: *Hard and Soft Debiasing.* Both rely on identifying gender bias subspace B via Principal Component Analysis computed on gender word pairs differences, such as *he–she, and man–woman.* Hard debiasing employs a *neutralize* operation that removes, e.g., all non-gender related words N from a gender subspace by deducting from vectors their bias

subspace projection. Subsequently, an *equalize* operation positions opposing gender pair words (e.g., king, queen) to share the same angle with neutral words. By contrast, Soft Debiasing enables more gradual bias removal by utilizing a tuning parameter λ: An embedding $W \in \mathbb{R}^{d \times |vocab|}$ is being transformed by optimizing a transformation matrix $T \in \mathbb{R}^{d \times d}$:

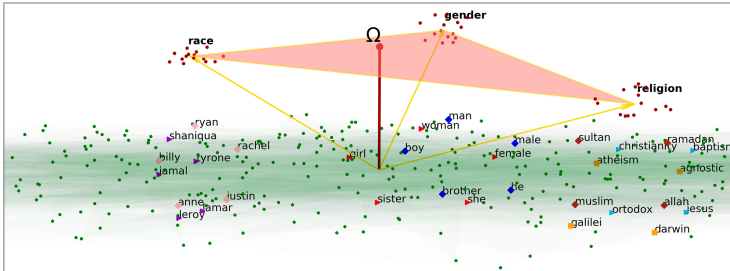$$\min_{T} \|(TW)^T(TW) - W^TW\|_F^2 + \lambda \|(TN)^T(TB)\|_F^2$$

Fig. 1. Visual interpretation of multiclass debiasing via HardWEAT.

Another recent approach by Manzini et al. [10] incorporates these ideas, but expands and evaluates results not only on gender, but separately also on race and religion. It suggests a bias subspace definition for non-binary class environment, which is formulated via PCA of mean shifted k-tuples (k = number of subclasses) of the definitional words. There are also other recently proposed gender bias post-processing [4,5] and pre-processing techniques [16].

In terms of existing joint multiclass debiasing techniques, *Conceptor Debiasing* [7], is based on applying Boolean-like logic operators using soft transformation matrix on a linear subspace where Word Embedding has the highest variance. We will use this technique for comparison in our experiments.

3   Approach

In this chapter, we present our novel debiasing techniques. Our methods substantially extend previous works of Bolukbasi et al. [2] and Caliskan et al. [3].

HardWEAT: To adapt the *neutralize* step from Bolubasi's work [2] jointly in a multiclass debiasing setting, we first define the concept of *class definitional vectors* $\overrightarrow{def_c}$ of classes (e.g., race, gender, . . .), which are computed as the top component of a PCA on the vector representations of definitional words $D_i \forall i \in \{1, ..., n\}$ for n subclasses (e.g., male, black, . . .), cf. [10]. Now, each particular class can vary in terms of bias amount according to WEAT tests with chosen sets of attribute and target words. Thus, we aggregate WEAT effect sizes d for each

class c into bias levels δ_c by averaging twice: First, we compute the mean value for all target/subclass pairs within a class. Second, we average those means for all results for class c. Then, we compute a *Centroid* Ω, as the average of the class definitional vectors weighted by their normalized bias levels: $\overrightarrow{\Omega} = \sum_{c \in C} \frac{(\delta_c \cdot \overrightarrow{def_c})}{\sum_{c \in C} \delta_c}$. We use this centroid to perform *neutralization*, i.e., we re-embed all neutral words such that they perpendicular to it, cf. [2].

To adapt the *equalize* step of Bolubasi's work, we move the definitional words of all subclasses in a way such that there is no relative preference of any attribute set towards a subclass. For that purpose, we generate equally spread out points $e_1, \ldots, e_n$ on some circle with radius r and center $\overrightarrow{o}$ as the new center points of the subclasses. In three or more dimensions, this circle can be defined by two vectors $(\overrightarrow{v_1}, \overrightarrow{v_2})$ perpendicular to $\overrightarrow{o}$: $\overrightarrow{e_i} = \overrightarrow{o} + r\cos(\frac{2\pi i}{n}) \cdot \overrightarrow{v_1} + r\sin(\frac{2\pi i}{n}) \cdot \overrightarrow{v_2}$.

Given this formula, we use *equidistancing* twice: First, we calculate new temporary central points for each class (e.g., gender) by neutralization (see above), then we determine new temporary central points $\overrightarrow{def_{S_i}}$ for each subclass (e.g., male) in a circle around that. Then, we compute new embeddings $\overrightarrow{w_i}$ for all definitional words D_i in a circle around this subclass vector. Note that r_c, r_{S_c} are randomly generated with the condition $r_{S_c} \gg r_c$, whereas $\overrightarrow{v_1}, \overrightarrow{v_2}$ are obtained via SVD. Along with an illustration (see Fig. 1), we present the full procedure in Algorithm 1.

For successful equidistancing, i.e., equal dispersion of words, we also require that the length of all subclass/target words within a particular class must be

---

Algorithm 1: HardWEAT algorithm.

Input: Word Embedding matrix $\boldsymbol{W}_{v \times d}$, Words W, set of classes C, set of subclasses S, Bias levels δ_c

1 **for** $c \in C$ **do**   // Generating class definitional vectors

2 $\quad \overrightarrow{def_c} := \text{PCA}\left(\bigcup_{i=1}^{n} \bigcup_{\vec{w} \in D_i} \vec{w} - \mu_i \right)$

3 $\overrightarrow{\Omega} = \sum_{c \in C} (\delta_c \cdot \overrightarrow{def_c})$ // Computing centroid

4 $N = W \setminus B$ // Defining neutral words

5 $\boldsymbol{N} := \boldsymbol{N} - \frac{N \cdot \overrightarrow{\Omega}}{||\overrightarrow{\Omega}||} \overrightarrow{\Omega}$ // Neutralizing

6 $\boldsymbol{W} = \frac{W}{||\boldsymbol{W}||}$ // Normalizing vectors

7 **for** $c \in C$ **do**   // Equidistancing

8 $\quad \overrightarrow{O_c} := \overrightarrow{def_c} - \frac{\overrightarrow{def_c} \cdot \overrightarrow{\Omega}}{||\overrightarrow{\Omega}||} \overrightarrow{\Omega}$

9 $\quad$ **for** $i \in \{1, ..., |S_c|\}$ **do**

10 $\quad\quad \overrightarrow{def_{S_i}\prime} = \overrightarrow{O_c} + r_c \cos(\frac{2\pi i}{|S_i|}) \cdot \overrightarrow{v_1} + r_c \sin(\frac{2\pi i}{|S_i|}) \cdot \overrightarrow{v_2}$

11 $\quad\quad$ **for** $w_j \in W_S$ **do**

12 $\quad\quad\quad \overrightarrow{w_j} = \overrightarrow{def_{S}\prime} + r_{S_c} \cos(\frac{2\pi j}{|W_S|}) \cdot \overrightarrow{v_1} + r_{S_c} \sin(\frac{2\pi j}{|W_S|}) \cdot \overrightarrow{v_2}$

13 $\boldsymbol{W} = \frac{W}{||\boldsymbol{W}||}$ // Normalizing vectors

equal. Violations of this requirement will result in inadequate WEAT scores. Furthermore, HardWEAT could in theory result in some equidistanced words becoming angle-close to random words: Thus, to counter this, we design this process in an iterative manner, modifying r_{S_c} until there are no neutral words having an angle greater than certain threshold (e.g., we use $45°$ as a default). We discuss factors of success and open issues in more detail in Sect. 4.4. Overall, the HardWEAT method ultimately aims at complete bias elimination, but randomizes the topological structure of subclass words.

SoftWEAT is a more gradual debiasing alternative with a greater focus on quality preservation. It provides the user with a choice on how to debias and to what extent. Let us assume that a particular target set of words S_c is closely related in terms of their angle to some attribute set of words A_i (e.g., A_1 is the set of *pleasant* words, A_2 is the set of *intellectual* words). Hypothetically, bias would be minimized if the subclass words S_c would be perpendicular to the attribute sets A_i. Thus, moving S_c towards such a perpendicular space/vector (null vector), noted as $\vec{n}$, results in bias reduction. Since full convergence towards a vector $\vec{n}$ may result in quality decrease we define a parameter $\lambda \in [0,1]$ as a *level of removal*. Setting $\lambda = 1$ results in maximal angle decrease between S_c and A_i (perpendicular vectors), while $\lambda = 0$ makes no transformation at all. Note, however, that placing vectors from some subclass words S_i to be perpendicular with attribute words A_i may not necessarily result in overall bias level reductions since WEAT tests are relative measures. E.g., they take the relationship between *male* and *intellectual* words and the relationship between *female* and *appearance*-related words into account at the same time. Additionally, higher λ also pose a greater risk of producing new bias. Moving the representation of subclass words away from attribute words A_1 may result in moving them closer to some other A_2 without prior intent. For example, moving *male* away from *science* can get them closer to *aggressive*, resulting in other WEAT test d increase. To address these issues, we choose a nullspace via SVD that minimizes the average WEAT effect size for all tests.

SoftWEAT executes the following steps: Given some S_c and its corresponding related attribute set of words $A_i, \ldots, A_n$, we generate a matrix $\boldsymbol{A_{n \times d}}$, where the n rows consist of the first principal components for each A_i respectively. For $\boldsymbol{A}$, we then find the nullspace vector $\vec{n}$ that decreases WEAT test scores the most. As a goal, we aim to translate our S_c to this space. Since initially, there might be only few target set words S_c, we expand it by adding for all of them the closest, frequently occurring neighbor words (e.g., the 20 closest neighbors with minimum frequency of 200 in the English Wikipedia [1]).) Afterwards, we calculate a transformation vector $\vec{\psi}$ for which it holds that: $\vec{\psi} = \vec{n} - \vec{m}$, where $\vec{m}$ is the mean of S_c words. This transformation vector is then multiplied with a parameter $\lambda \in [0,1]$ followed by conversion into linear translation matrix Ψ. By translating vectors from extended word sets, we preserve relationships of subclass words S_c to these words. Note that we only modify positions of words from the extended neighborhood of subclass words, but not all words from the vocabulary. Finally, as a last optional step, we normalize all vectors.

In SoftWEAT, the user can decide on the λ, but can also select target and attribute sets used for debiasing. By default, we will only take those into account, for which WEAT scores result in an aggregated effect size of $|d| > 0.6$. To compute this, we accumulate attribute sets per each target set that form matrix $\mathbf{A}$ per each S_c. In case of positive d, S_1 gets removed from A_1, and S_2 from A_2. In case of negative d, removal is done other way around. The algebraic operations are formalized as follows: Out of the new matrix $\mathbf{W'}$ with $|S_c|$ rows and $d+1$ columns, all but the last row are taken as a new vector representation for the given target set of words S_c.

$$\mathbf{A}\vec{n} = 0 \qquad \vec{\psi} = \lambda(\vec{n} - \vec{m})$$

$$\mathbf{W'}_{S_c} = \Psi \left[\frac{(W_{S_c})^T}{1} \right] = \begin{bmatrix} 1 & 0 & 0 & \dots & \psi_1 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ 0 & 0 & \dots & 1 & \psi_d \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} w_{11} & w_{21} & \dots & w_{n1} \\ \vdots & \vdots & \vdots & \ddots \\ w_{1d} & w_{2d} & \dots & w_{nd} \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

4 Experimental Evaluation

This section presents example results of our methods in practical settings and compares it with Conceptor Debiasing [7] as the current state-of-the-art approach. We measure the bias decrease along with deterioration of embedding quality and show the effects of biased/debiased embeddings in a Sentiment Analysis as a downstream task example. Due to limited space, we also provide an extended set of results in an accompanying online appendix[1].

4.1 Experimental Setup

Datasets: Extending the experimental design from [8], we apply debiasing simultaneously on following target sets/subclasses: *(male, female) - gender*, *(islam, christianity, atheism) - religion* and *(black and white names) - race* with seven distinct attribute set pairs. We collected target, attribute sets, and class definitional sets from literature [3,8,10,11,15,16], see our online appendix for a complete list. As in previous studies [7], evaluation was done on three pretrained Word Embedding models with vector dimension of 300: FastText[2] (English webcrawl and Wikipedia, 2 million words), GloVe[3] (Common Crawl, Wikipedia and Gigaword, 2.2 million words) and Word2Vec[4] (Trained on Google News, 3 million words). For the Sentiment Analysis task, we additionally employed a dataset of movie reviews [9].

---

[1] https://git.io/JvL10.
[2] https://fasttext.cc/docs/en/english-vectors.html.
[3] https://nlp.stanford.edu/projects/glove/.
[4] https://drive.google.com/uc?id=0B7XkCwpI5KDYNlNUTTlSS21pQmM.

Quality Assessment: First, we compared ranked lists of word pairs in terms of their vector similarity to human judgement via Spearman's Rank correlation coefficient [13], by using the collection taken from the Conceptor Debiasing evaluation [7], i.e., *RG65, WS, RW, MEN, MTurk, SimLex, SimVerb*. In addition, we also utilize Mikolov Analogy Test [12].

Methods: Regarding HardWEAT, we specified neutral words through set difference between all words and ones from priorly defined target sets. In terms of SoftWEAT we provide details such as target-attribute sets structure and number of changed words in section F of online appendix. We applied the OR operator in the Conceptor Debiasing, using the same word set of subclasses within the three above defined classes (subspaces in Conceptor Debiasing).

Table 1. Bias levels for gender, race, and religion before debiasing (REG) and after debiasing with Hard/SoftWEAT (HW/SW) or Conceptor (CPT)

		REG	HW	SW $\lambda = 0.2$	SW $\lambda = 0.4$	SW $\lambda = 0.6$	SW $\lambda = 0.8$	SW $\lambda = 1$	CPT
	gender	0.62	0.0**	0.47	0.3	0.28	0.15	0.09*	0.24
GloVe	race	0.71	0.0**	0.6	0.51	0.39	0.3	0.19*	0.43
	religion	0.77	0.0**	0.62	0.46	0.29	0.2	0.16*	0.28
	gender	0.52	0.0**	0.14*	0.17	0.32	0.31	0.32	0.46
FastText	race	0.36	0.0**	0.13*	0.16	0.25	0.3	0.31	0.31
	religion	0.6	0.0**	0.27	0.21*	0.29	0.35	0.42	0.54
	gender	0.63	0.0**	0.48	0.37	0.32	0.2	0.23	0.12*
Word2Vec	race	0.56	0.0**	0.32	0.28	0.2	0.16	0.16*	0.38
	religion	0.47	0.0**	0.31	0.19	0.11*	0.18	0.2	0.28

4.2 Bias Levels and Quality of Word Embeddings

First, we focus on overall bias levels, see Table 1, by measuring WEAT scores before and after debiasing. We observe that HardWEAT removes the measured bias completely as indicated by zero WEAT scores. SoftWEAT also substantially reduces the bias measurements to different degrees for different datasets. In comparison, for example with $\lambda = 1$, SoftWEAT still leads to a stronger reduction in bias compared to the state-of-the-art Conceptor algorithm in all but one measurement.

Regarding quality assessment, Table 2, shows the complete results for seven rank similarity datasets as well as the Mikolov analogy score on the example of the GloVe Word Embedding. As expected, a significant quality drop appears with HardWEAT, most notably on the RG65 dataset. With SoftWEAT, greater λ induce larger modifications of the original embeddings, which corresponds to

a greater drop in embedding quality. However, in three out of eight test settings, SoftWEAT with lower λ settings achieved a similar or higher score and also leads to competitive results compared to Conceptor, see Sect. 4.4 for an in-depth discussion. Extended results can be found in the online appendix.

Table 2. Measurements of quality tasks after debiasing for GloVe embeddings.

	REG	HW	SW $\lambda = 0.2$	SW $\lambda = 0.4$	SW $\lambda = 0.6$	SW $\lambda = 0.8$	SW $\lambda = 1$	CPT
RG65	**76.03***	63.42	75.68	75.39	74.34	73.25	71.42	68.73
WS	73.8	69.73	**74.0***	73.96	73.15	71.95	70.34	73.48
RW	46.15	46.09	46.15	46.13	46.08	46.02	45.94	**52.45***
MEN	80.13	77.52	80.34	**80.34***	80.1	79.57	78.74	79.99
MTurk	**71.51***	69.78	71.25	70.78	70.37	69.48	68.64	68.24
SimLex	40.88	40.44	41.46	42.0	42.07	42.21	42.18	**47.36***
SimVerb	28.74	28.74	28.9	29.15	29.27	29.54	29.78	**36.78***
Mikolov	0.65	0.64	**0.65***	0.65	0.65	0.64	0.64	0.63

4.3   Sentiment Analysis

To analyze the effects of debiasing in downstream tasks, we study a sentiment analysis task in the context of movie reviews, i.e., we investigate whether we observe significant differences in predicted sentiments when using biased and debiased Word Embeddings. Modifying setup from Packer et al. [14], we added to the end of each test sentences randomized word from opposing set pairs (e.g., a typical black name or a typical white name), and calculated the difference in the prediction, i.e., the *polarity score* according to a simple neural network that takes the respective pretrained word embeddings as pretrained embedding, which is followed by Flatten-operation and a sigmoid layer. The first two layers were not trainable.

For training, we used a combination of binary-cross entropy loss and Ada-Delta optimizer. We trained our model only on sentences with a maximum length of 50 that did not contain any target sets of words. Shuffling test and training set, we then calculated *polarity score* on 15 different model instances using 6 different kinds of Embeddings: Regular (not debiased), SoftWEAT with $\lambda = 0.1, 0.5, 1$, HardWEAT and Conceptor Debiasing with $\alpha = 2$.
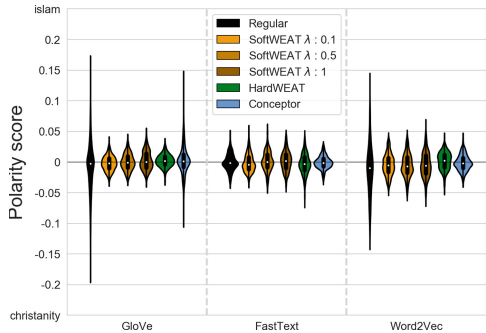

Fig. 2. Classifier Bias on (islam, christianity)

Results for *[christianity, islam]* are shown in Fig. 2. For the non debiased embeddings, we observe that our classifier has no clear trend such that the addition of a bias word in a sentence influences the polarity in a particular direction. However, we can see that bias words have a strong influence on the classifier, leading to a large variance of *polarity scores* in GloVe and Word2Vec-based models. When we apply our debiasing methods, we see that the respective variances are substantially decreased, e.g., around 5.45 times for SoftWEAT already with a small λ of 0.1 in GloVe. This is only the case with Conceptor Debiasing for the Word2Vec-based model, but not for the GloVe-based model.

4.4   Discussion

HardWEAT and SoftWEAT Success Factors: Based on our experiments, see also our online appendix, we could identify a variety of success factors for our methods. Regarding HardWEAT, higher embedding dimensionality and more dispersed vectors are more likely to output more desirable outcome. This is due to the possibility of random words appearing as close neighbors to a target words during the equidistancing procedure, which we counter iteratively as described above. Regarding centroid ($\overrightarrow{\Omega}$) neutralization, relevant factors of importance are number of classes, uniformity of bias levels and angles between all def_c: Having more classes with uniform bias levels and distant angles can output non-desirable results, e.g., too large distances between $\overrightarrow{\Omega}$ and $\overrightarrow{def_c}$ points. However, the user could also manually define an alternative centroid $\overrightarrow{\Omega}$. We acknowledge that *equidistancing* comes with a main drawback, which is the partial loss of relationship between target and non-target words, which is also reflected in a quality drop according to different metrics. Regarding SoftWEAT, decreased angles between target and attribute sets after translation may not necessarily result in lower bias levels, since WEAT is a relative measure. However, we take one nullspace which minimizes these values. Furthermore, we note that in our experiments we used all target and attribute set pairs within each of the WEAT tests, which can be further optimized: We may not want to debias something, which isn't priorly biased. E.g., removing *male* from *science* may be necessary, whereas doing the same for *female* from *art* may not, thus we could exclude this latter pair. Also, we should bear in mind, that attribute sets of words are often correlated (as also shown by our experiments in the online appendix). This implies that by moving subclass words towards a specific set of attribute words, we automatically change the associations also with other attribute sets. Thus, the user plays a crucial role in deciding which sets should be used for debiasing.

Comparison with Conceptor Debiasing: Given the experimental results, we conclude that neither Conceptor Debiasing nor SoftWEAT outperform each other. Yet, SoftWEAT exhibits some distinctive advantages: (i) With Soft-WEAT, relationships within the target set words remain completely the same, whereas in Conceptor Debiasing, overall angle distribution gets more narrow (See online appendix for more details). (ii) We argue that with SoftWEAT, user gets more freedom with choosing on which target/attribute set combination and to

which degree debiasing is applied. (iii) Given our method, there is no difference in word representation if one uses small subset of neutral words or complete vocabularies. Nevertheless, we acknowledge that Conceptor Debiasing does succeed in reducing bias equally well by applying it in more global behavior.

5 Conclusion

In this paper, we presented two novel approaches for multi-class debiasing of Word Embeddings. We demonstrated the general viability of these methods for reducing and/or eliminating biases while preserving meaningful relationships of the original vector representations. We also analyzed the effects of debiased representations in Sentiment Analysis as an example downstream task and find that debiasing leads to substantially decreased variance in the predicted polarity. Overall, our work contributes to ongoing efforts towards providing more unbiased neural representations of textual data.

References

1. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2017)
2. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 4349–4357 (2016)
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
4. Dev, S., Phillips, J.: Attenuating bias in word vectors. In: International Conference on Artificial Intelligence and Statistics, pp. 879–887 (2019)
5. Font, J.E., Costa-jussà, M.R.: Equalizing gender bias in neural machine translation with word embeddings techniques. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 147–154 (2019)
6. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Natl. Acad. Sci. **115**(16), E3635–E3644 (2018)
7. Karve, S., Ungar, L., Sedoc, J.: Conceptor debiasing of word representations evaluated on WEAT. In: Workshop on Gender Bias in NLP (2019)
8. Knoche, M., Popović, R., Lemmerich, F., Strohmaier, M.: Identifying biases in politically biased wikis through word embeddings. In: Conference on Hypertext and Social Media, pp. 253–257 (2019)
9. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Annual Meeting of the ACL: Human Language Technologies, pp. 142–150 (2011)
10. Manzini, T., Chong, L.Y., Black, A.W., Tsvetkov, Y.: Black is to criminal as caucasian is to police: detecting and removing multiclass bias in word embeddings. In: Conference of the North American Chapter of the ACL, pp. 615–621 (2019)

11. May, C., Wang, A., Bordia, S., Bowman, S., Rudinger, R.: On measuring social biases in sentence encoders. In: Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 622–628 (2019)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Myers, L., Sirois, M.J.: Spearman Correlation Coefficients, Differences Between. American Cancer Society (2006)
14. Packer, B., Mitchell, M., Guajardo-Céspedes, M., Halpern, Y.: Text embeddings contain bias. Here's why that matters. Technical report, Google (2018)
15. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: evaluation and debiasing methods. In: Conference of the North American Chapter of the ACL, pp. 15–20 (2018)
16. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning gender-neutral word embeddings. In: Conference on Empirical Methods in Natural Language Processing, pp. 4847–4853 (2018)