Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Nico Hertel, Seth Siriya, Thomas Decker, Uzair Akbar, Zhenchen Liao

Chair for Data Processing
Department of Electrical and Computer Engineering
Technical University of Munich

Abstract

By learning from real-world data, machine learning is doomed to adopt social bias such as sexism. This paper analyzes gender bias in the context of the natrual language processing technique, word embeddings. The paper proposes a method to determine the underlying bias in a dataset and an algorithm to eliminate this bias while preserving the ability to cluster words. This is shown using a public dataset of news articles to reduce gender stereotypes in analogy tasks.

Introduction and Preliminary

A *word embedding* represents each word w as a d-dimensional word vector $\vec{w} \in \mathbb{R}^d$ with two imortant properties: similar words have similar vectors, and the difference between two word vectors has been shown to represent the difference between the corresponding words. These arithmetic properties can be used to solve analogy tasks, for example, 'man is to brother as woman is to X?'. The difference between the word vectors of 'man' and 'woman' should be similar to 'brother' and X (e.g. 'sister'): $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{brother} - \overrightarrow{sister}$.

To measure bias, a gender-neutral word like 'nurse' is compared to two gender-specific words like 'man' and 'woman'. If the distance between the neutral word and one specific word is smaller than between the other, this suggests bias. Figure 1 shows a sample of words ordered by their distance to 'he' and 'she' (x-axis) and their bias (y-axis). For this paper, the public word2vec embedding is used, trained on a set of Google News Articles, cosisting of 3 million english words.

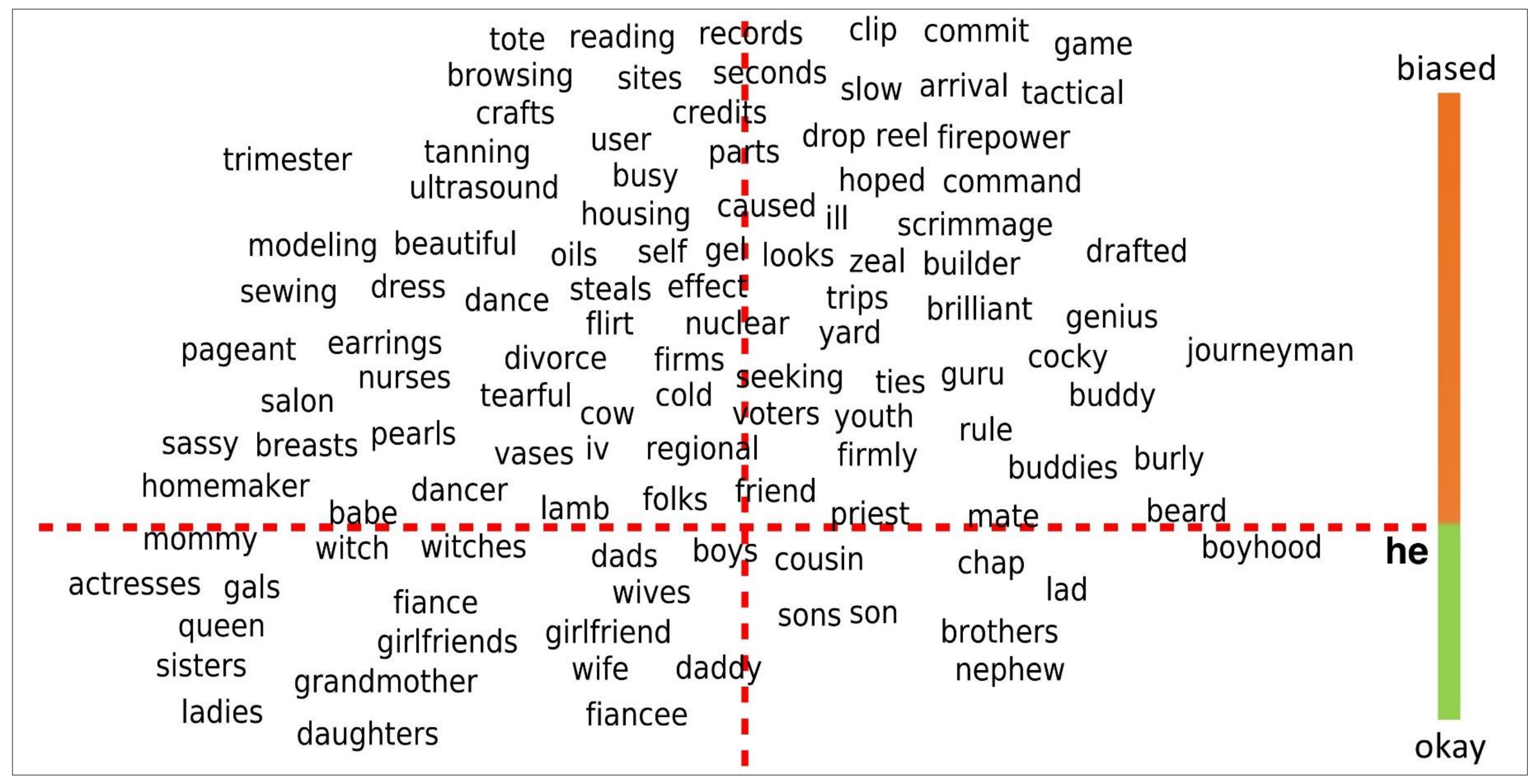
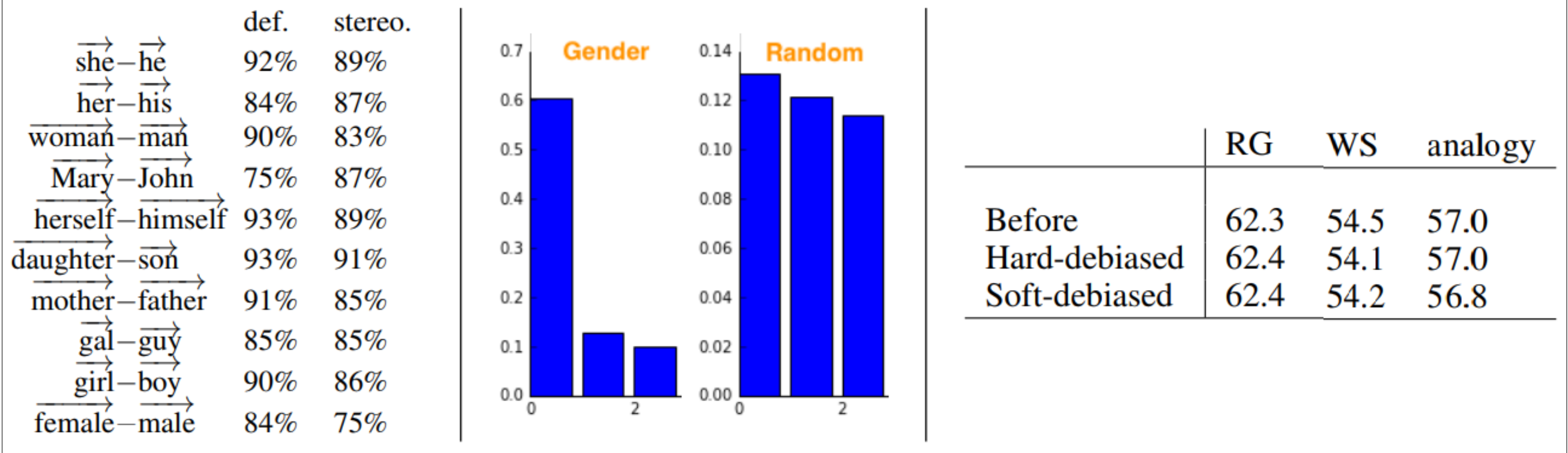


Figure 1: Selected words projected along gender (x) and neutrality (y) axis.



	def.	stereo.
$\overrightarrow{she}-\overrightarrow{he}$	92%	89%
$\overrightarrow{her}-\overrightarrow{his}$	84%	87%
$\overrightarrow{woman}-\overrightarrow{man}$	90%	83%
$\overrightarrow{Mary}-\overrightarrow{John}$	75%	87%
$\overrightarrow{herself}-\overrightarrow{himself}$	93%	89%
$\overrightarrow{daughter}-\overrightarrow{son}$	93%	91%
$\overrightarrow{mother}-\overrightarrow{father}$	91%	85%
$\overrightarrow{gal}-\overrightarrow{guy}$	85%	85%
$\overrightarrow{girl}-\overrightarrow{boy}$	90%	86%
$\overrightarrow{female}-\overrightarrow{male}$	84%	75%

	RG	WS	analogy
Before	62.3	54.5	57.0
Hard-debiased	62.4	54.1	57.0
Soft-debiased	62.4	54.2	56.8

Figure 2: Left: Ten word pairs used to define gender, and agreement with definition and stereotype sets from crowd.
Center: Variance explained by PCA for 10 pairs of gender words and random words.
Right: Performance of w2vNEWS embedding before and after de-biasing on coherence & analogy-solving metrics.

Bias in Word Embeddings

- **Occupational Stereotypes:** Crowdwork evaluated occupational stereotypes strongly correlated with *she-he* axis projections ($\rho = 0.51$).
- **Stereotypical Analogies:** Scored *she-he* analogies are rated via crowdwork as (a) gender-appropriate (b) stereotypic. Scoring metric: $S(x,y) = \cos\left(\overrightarrow{she} - \overrightarrow{he}, \vec{x} - \vec{y}\right)$, $\|\vec{x} - \vec{y}\| \le \delta$.
- **Direct Bias:** $\frac{1}{|N|}\sum_{w \in N} |\cos(\vec{w}, \boldsymbol{g})|^c$, for gender direction/ basis $\boldsymbol{g}$.

Debiasing Algorithm

Some definitions:

- Word set $W = N \cup S$, N is gender neutral, S is gender specific set.
- Bias subspace $B \subset \mathbb{R}^d$ with k basis $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k$.
- Projection $\boldsymbol{v}_B$ of vector $\boldsymbol{v}$ onto B.
- Bias subspace defining sets $D_1, \cdots, D_n \subset S$.
- Equality sets $E_1, \cdots, E_m \subset S$.
- Soft de-biasing transformation T.

The algorithm:

Identify Bias Subspace

$$\boldsymbol{\mu}_i := \sum_{w \in D_i} \vec{w}/|D_i| \quad , \forall \ i \in [1, n]$$

$$C := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \boldsymbol{\mu}_i)^T (\vec{w} - \boldsymbol{\mu}_i)/|D_i|$$

$$B := \text{first } k \text{ rows of } \mathbb{SVD}(C)$$

Hard de-biasing

Neutralize

$$\vec{w} := (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\| \quad ,\forall \ w \in N$$

Equalize

$$\boldsymbol{\mu}_i := \sum_{w \in E_i} \vec{w}/|E_i| \quad , \boldsymbol{v}_i := \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i_B} \quad ,i \in [1, m]$$

$$\vec{w} := \boldsymbol{v}_i + \sqrt{1 - \|\boldsymbol{v}_i\|^2} \frac{\vec{w}_B - \boldsymbol{\mu}_{i_B}}{\|\vec{w}_B - \boldsymbol{\mu}_{i_B}\|} \quad ,i \in [1, m], \forall\, w \in E_i$$

Soft de-biasing

$$T^* := arg\min_T \| \overbrace{(TW)^T(TW)}^{\text{new par-wise inner products}} - \overbrace{W^TW}^{\text{old par-wise inner products}} \|_F^2 + \lambda \| \underbrace{(TN)^T(TB)}_{\text{Projection of } N \text{ onto } B} \|_F^2$$

$$\widehat{W} := \{T^*\vec{w}/\|T^*\vec{w}\|_2 \, , w \in W\}$$

Debiasing Results

An *SVM* classifier was trained on a subset of the Google News data called *w2vNEWS* to identify gender-neutral words, then generalized to the rest of the dataset. 10-fold cross validation resulted in an *F-Score* of 0.627. Debiasing was also measured by having a crowd evaluate whether analogies generated from the embedding are appropriate or reflect gender stereotypes, with the hard debiased embedding having least stereotypical analogies (Figure 3) and most approved analogies (Figure 4).

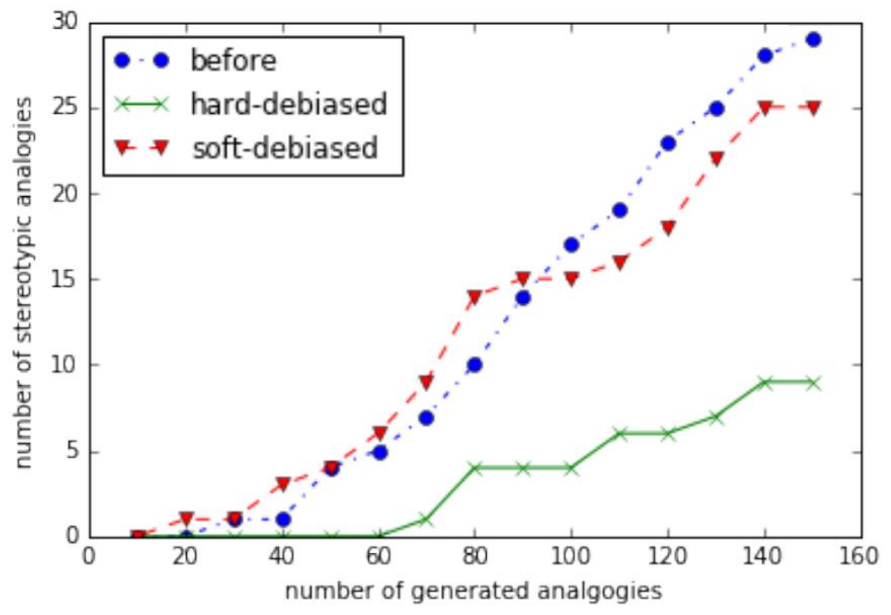


Figure 3: Number of stereotypical analogies generated by word embedding before and after de-biasing.

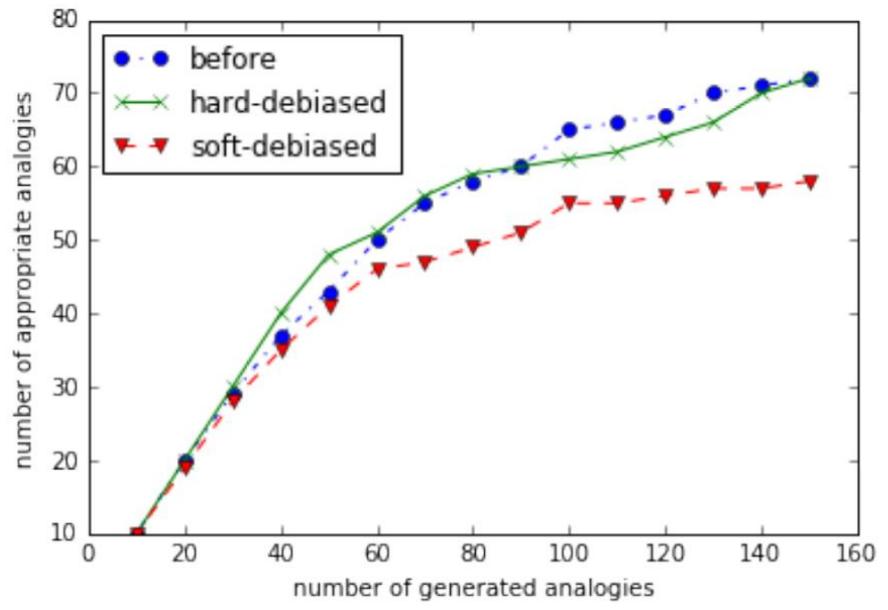
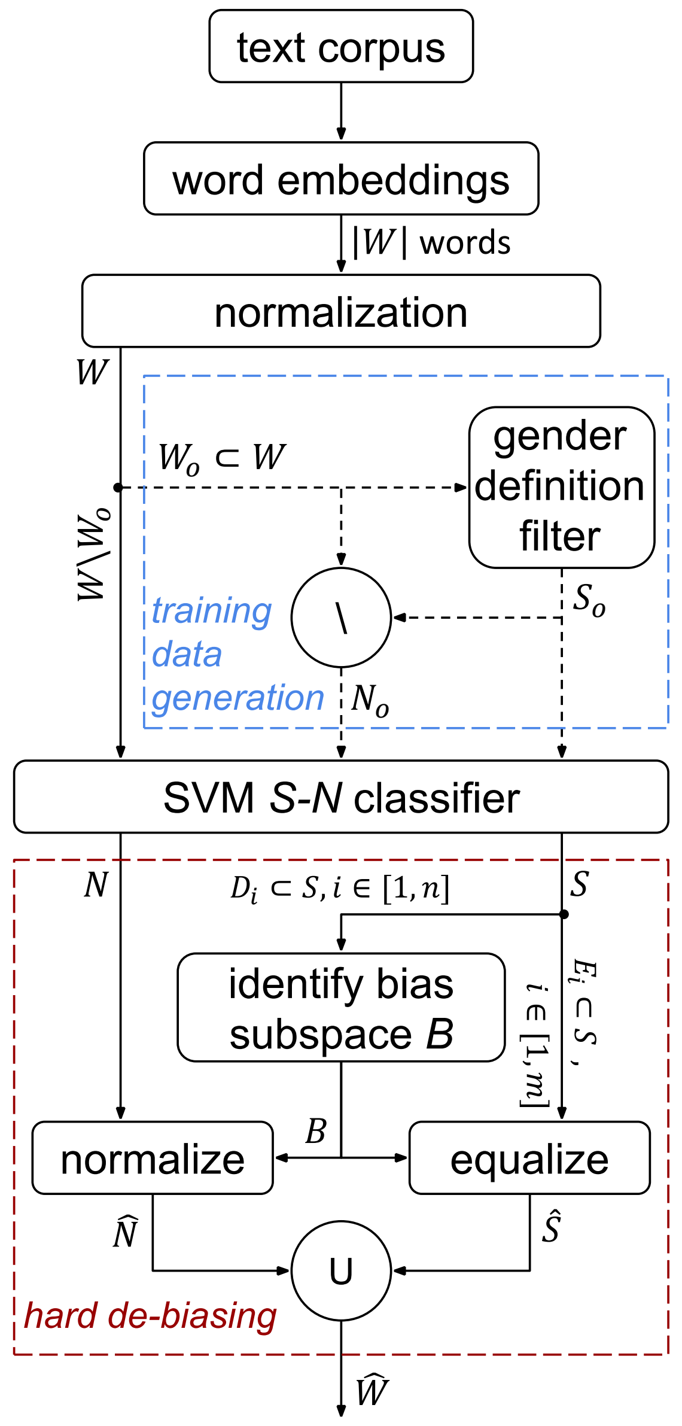


Figure 4: Number of appropriate analogies generated by word embedding before and after de-biasing.



text corpus

word embeddings

$|W|$ words

normalization

W

$W \setminus W_o$

$W_o \subset W$

gender definition filter

S_o

training data generation

$\setminus$

N_o

SVM *S-N* classifier

N

$D_i \subset S, i \in [1, n]$

S

identify bias subspace *B*

$E_i \subset S$, $i \in [1, m]$

B

normalize

equalize

$\widehat{N}$

$\hat{S}$

U

hard de-biasing

$\widehat{W}$

Figure 5: Flow chart of process from text corpus to hard de-biased word embedding.

Discussion

Overall, a single direction was able to capture gender information. Hard debiasing was the most effective on the embedding, where placing neutral words orthogonal to the gender direction and paired word sets equidistant from neutralised words resulted in reduction of gender bias whilst still capturing appropriate analogies.

Disclaimer: This poster presentation is based off of the following work:
T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," arXiv:1607.06520 [cs, stat], Jul. 2016.
We do not claim any credit for the aforementioned work and are only presenting this work as part of our 'Information Retrieval in High Dimensional Data' course-work (Summer Semester 2018) at the Technical University of Munich (TUM).