Gender Bias in Word Embeddings

Hila Gonen
UW, Meta AI

NLP Course, UW, Prof. Noah A. Smith

March 2022

Quick Question

A doctor is walking down the street with a boy.

The boy is the doctor’s son, but the doctor is not the boy’s father.

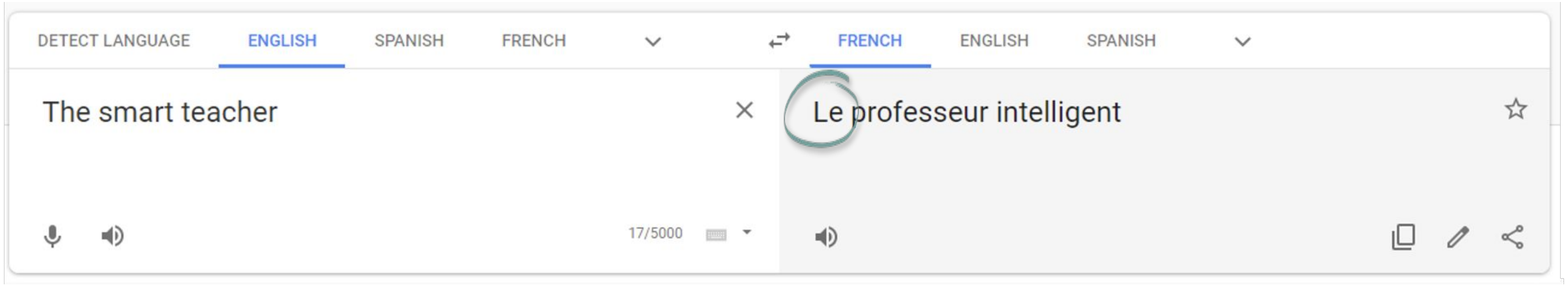
How is that possible?

Simple Answer

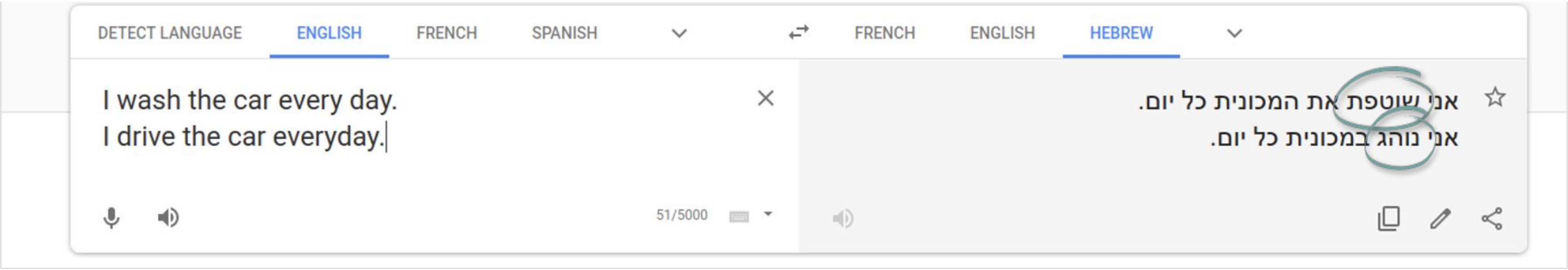
The doctor is the boy’s mother…

What do we mean by gender bias?

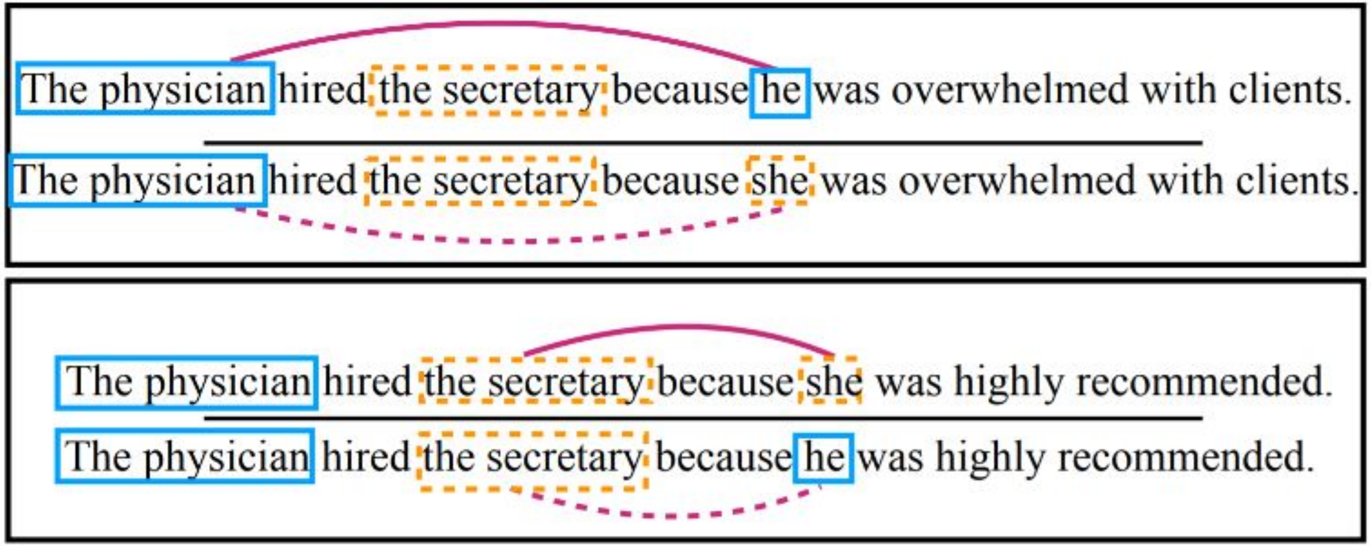
Example – Gender Bias in Translation



Example – Gender Bias in Translation



Example – Gender Bias in Coreference



Zhao et al., NAACL 2018

Example – Stereotyped Analogies

Generate analogies using word embeddings:

he* to *x* is as *she* to *y

he to DOCTOR is as she to NURSE

he to KING is as she to QUEEN

Bolukbasi et al., 2016

Word Embeddings

Word Embeddings

Word embeddings are successfully used for various NLP applications:
Semantic similarity, Word sense Disambiguation, Named entity Recognition, Summarization, etc.

Each word in the vocabulary is represented by a low dimensional vector (~300d)

All words are embedded into the same space
Similar words have similar vectors (= close to each other in the vector space)

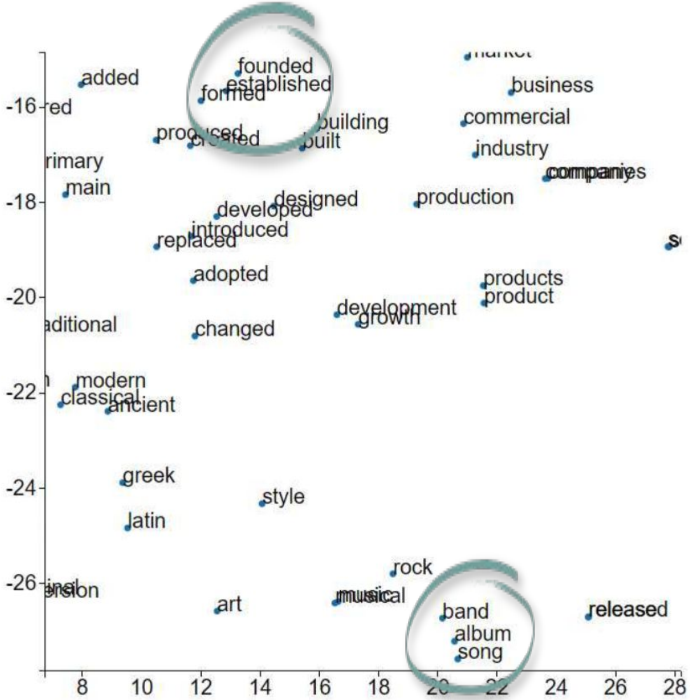
Word Embeddings

Trained with raw text

The Distributional Hypothesis:

- Words that occur in the same contexts tend to have similar meanings (Harris, 1954)
- “You shall know a word by the company it keeps” (Firth, 1957)

Word Embeddings



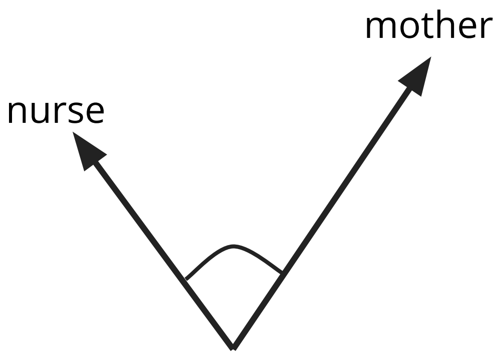
Gender bias in word embeddings

Word2Vec

nurse

nearest neighbors (cosine-similarity):

Word	Cosine distance
midwives	0.597824
nurses	0.523600
nursing	0.522353
midwife	0.505857
obstetrics	0.497042
mother	0.494208
hospital	0.486670
midwifery	0.446893
elsie	0.430787
child	0.428072
veterinarian	0.425949
care	0.420312
housekeeper	0.415515
wife	0.414742
aunt	0.414349
orphaned	0.410652
menopause	0.409759
orphanage	0.406390
orphan	0.403061
widower	0.401952
gynecology	0.400221

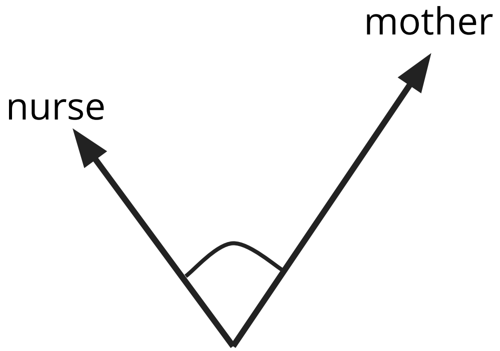


Word2Vec

nurse

nearest neighbors (cosine-similarity):

Word	Cosine distance
midwives	0.597824
nurses	0.523600
nursing	0.522353
midwife	0.505857
obstetrics	0.497042
mother	0.494208
hospital	0.486670
midwifery	0.446893
elsie	0.430787
child	0.428072
veterinarian	0.425949
care	0.420312
housekeeper	0.415515
wife	0.414742
aunt	0.414349
orphaned	0.410652
menopause	0.409759
orphanage	0.406390
orphan	0.403061
widower	0.401952
gynecology	0.400221



Bias in Word Embeddings (Caliskan et al.)

Caliskan et al. (2017) replicate a spectrum of known biases from the literature using word embeddings

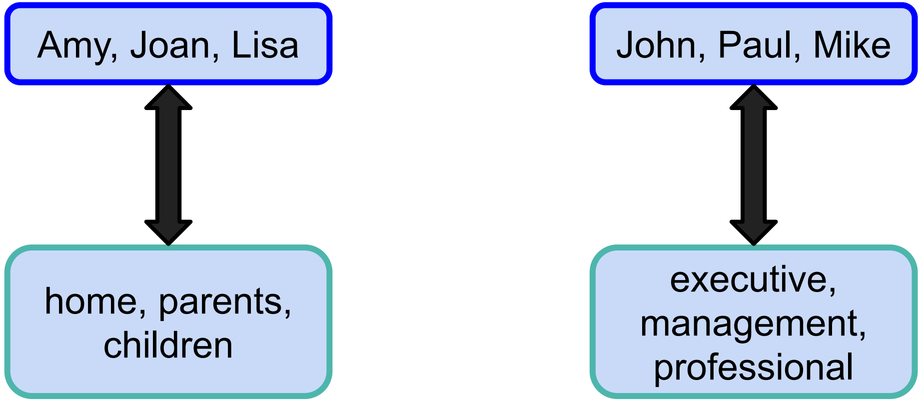
Show that text corpora contain several types of biases:
gender and racial biases, among others

Bias in Word Embeddings (Caliskan et al.)

They use a permutation test:

X, Y: sets of **target** words (e.g. male names vs. female names)

A, B: sets of **attribute** words (e.g. career terms vs. family terms)



Bias in Word Embeddings (Caliskan et al.)

They use a permutation test:

> ***X, Y***: sets of **target** words (e.g. male names vs. female names)
>
> ***A, B***: sets of **attribute** words (e.g. career terms vs. family terms)

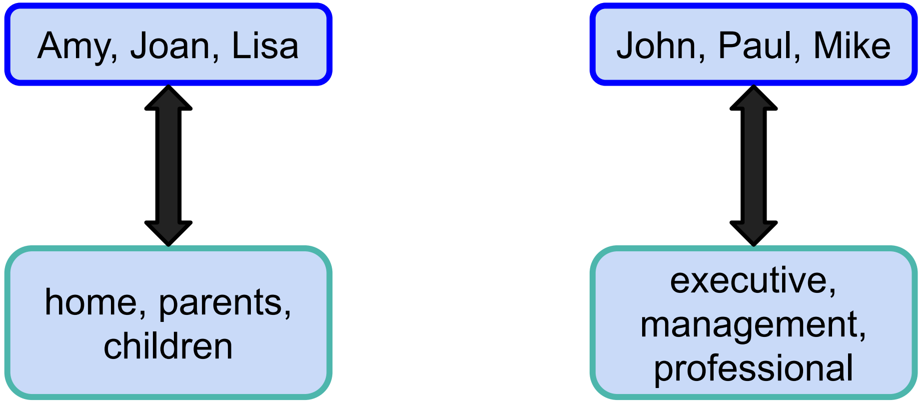
Null hypothesis:
no difference between the two sets of target words
in their relative similarity to the attribute

Bias in Word Embeddings (Caliskan et al.)

They use a permutation test:

X, Y: sets of **target** words (e.g. male names vs. female names)

A, B: sets of **attribute** words (e.g. career terms vs. family terms)



Bias in Word Embeddings (Caliskan et al.)

Examples:

X	Y	A	B
Flowers: buttercup, daisy, lily	**Insects:** ant, caterpillar, flea	**Pleasant:** freedom, health, love	**Unpleasant:** abuse, crash, filth
European American names: Brad, Brendan	**African American names:** Darnell, Lakisha	**Pleasant:** joy, love, peace	**Unpleasant:** agony, terrible
Male terms: male, man, boy	**Female terms:** female, woman, girl	**Math words:** math, algebra, geometry	**Arts Words:** poetry, art, dance

Bias in Word Embeddings (Caliskan et al.)

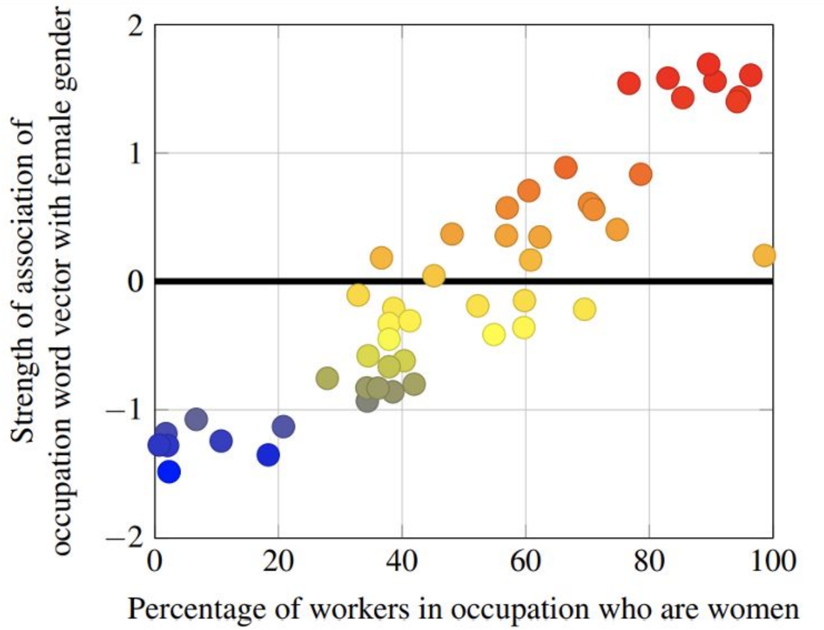


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

Bias in Word Embeddings

Bias in our world translates to bias in our representations

How do we define
gender bias in word embeddings?

Definition of Gender Bias in Word Embeddings

Work by Bolukbasi et al. (2016)

Check how similar a word is to “he” and “she” (cosine similarity)

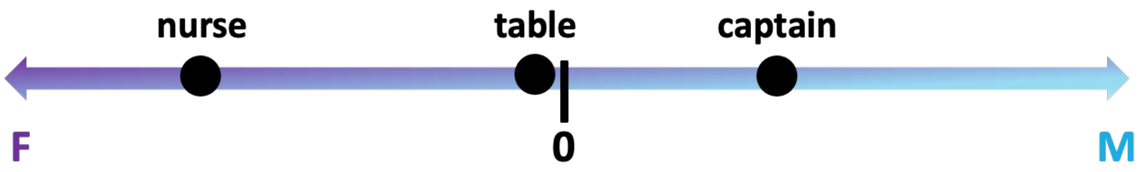
Note that we care about the **difference** between the two

This is the **projection on the direction** of “he – she”:

$$\mathrm{bias}(w) = \overrightarrow{w} \cdot \overrightarrow{he} - \overrightarrow{w} \cdot \overrightarrow{she} = \overrightarrow{w} \cdot \underline{(\overrightarrow{he} - \overrightarrow{she})}$$

Definition of Gender Bias in Word Embeddings

- bias(nurse) = -0.2471 negative (F)
- bias(captain) = 0.1521 positive (M)
- bias(table) = -0.0003 neutral



Existing debiasing methods

Debiasing in post-processing

Bolukbasi et al. (2016) suggest to remove bias in post-processing:

- Define a **gender direction**:

  The principal component of 10 gender pair difference vectors

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

- Define **inherently neutral** words using dictionary definitions: E.g. mother, aunt, chairman, girlfriend, prince

Debiasing in post-processing

Bolukbasi et al. (2016) suggest to remove bias in post-processing:

- **Zero the projection** of all neutral words on the gender direction:

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|$$

$\vec{w}_B$ – Projection of w on the gender direction

Debiasing in post-processing

Bolukbasi et al. (2016) suggest to remove bias in post-processing:

- **Zero the projection** of all neutral words on the gender direction:

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|$$

$\vec{w}_B -$ Projection of w on the gender direction

- The bias of all neutral words is now **zero by definition**

Debiasing in post-processing

Bolukbasi et al. (2016) suggest to remove bias in post-processing:

- **Zero the projection** of all neutral words on the gender direction:

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|$$

$\vec{w}_B -$ Projection of w on the gender direction

- The bias of all neutral words is now **zero by definition**

We will address these embeddings as **HARD-DEBIASED**

Debiasing during Training

Zhao et al. (2018) suggest to reduce bias during training:

- Train word embeddings using GloVe (Pennington et al., 2014)
- Alter the loss to encourage the gender information **to concentrate in the last coordinate.**

To ignore gender information – simply remove the last coordinate

Debiasing during Training

Zhao et al. (2018) suggest to reduce bias during training:

- How to push gender information **to the last coordinate?**
    - Use two groups of male/female seed words, and encourage words from different groups to differ in their last coordinate.
    - Encourage the representation of gender-neutral words (excluding the last coordinate) to be orthogonal to the gender direction.

Debiasing during Training

Zhao et al. (2018) suggest to reduce bias during training:

- How to push gender information **to the last coordinate?**
  - Use two groups of male/female seed words, and encourage words from different groups to differ in their last coordinate.
  - Encourage the representation of gender-neutral words (excluding the last coordinate) to be orthogonal to the gender direction.

We will address these embeddings as **GN-GLOVE**

These methods work

Compelling results of bias reduction without hurting standard tasks

HARD-DEBIASED:

- Bias of all inherently-neutral words is zero by definition
- Generated analogies are less stereotyped

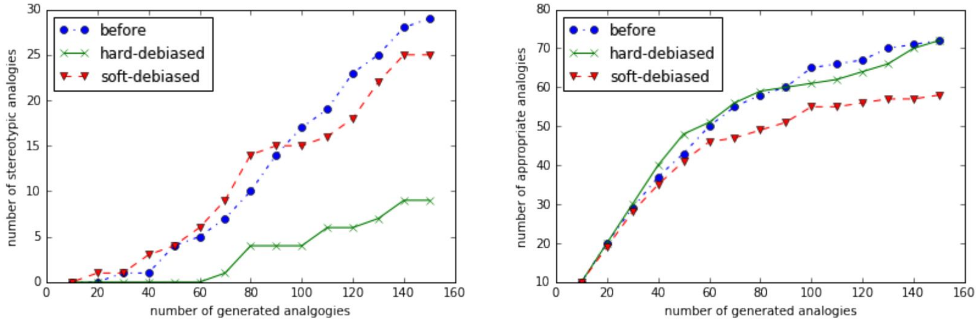


Figure 4: Number of stereotypical (Left) and appropriate (Right) analogies generated by word embeddings before and after debiasing.

These methods work

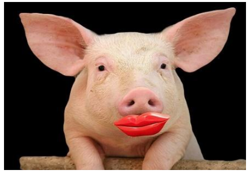
Compelling results of bias reduction without hurting standard tasks

GN-GLOVE: Decreases bias in coreference resolution

Embeddings	OntoNotes-test	PRO	ANTI	Avg	Diff
GloVe	66.5	76.2	46.0	61.1	30.2
Hard-Glove	66.2	70.6	54.9	62.8	15.7
GN-GloVe	66.2	72.4	51.9	62.2	20.5
GN-GloVe(w_a)	65.9	70.0	53.9	62.0	16.1

Table 3: F1 score (%) on the coreference system.

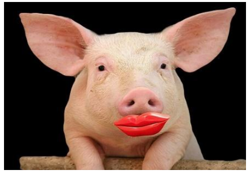
And they are popular - Bolukbasi et al. with over 1700 citations!



Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen, Yoav Goldberg

NAACL 2019

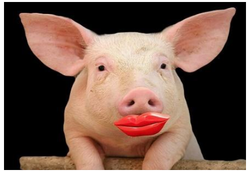


Do they really work?

Both methods and their results rely on the **gender direction**

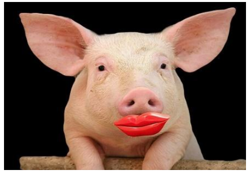
Bias is much more **profound** and systematic

We will now present a series of experiments showing that most of the **bias information is still recoverable**



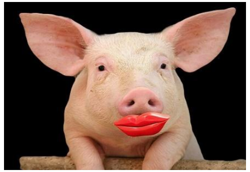
Demonstrating the remaining bias

Clustering male- and female- biased words



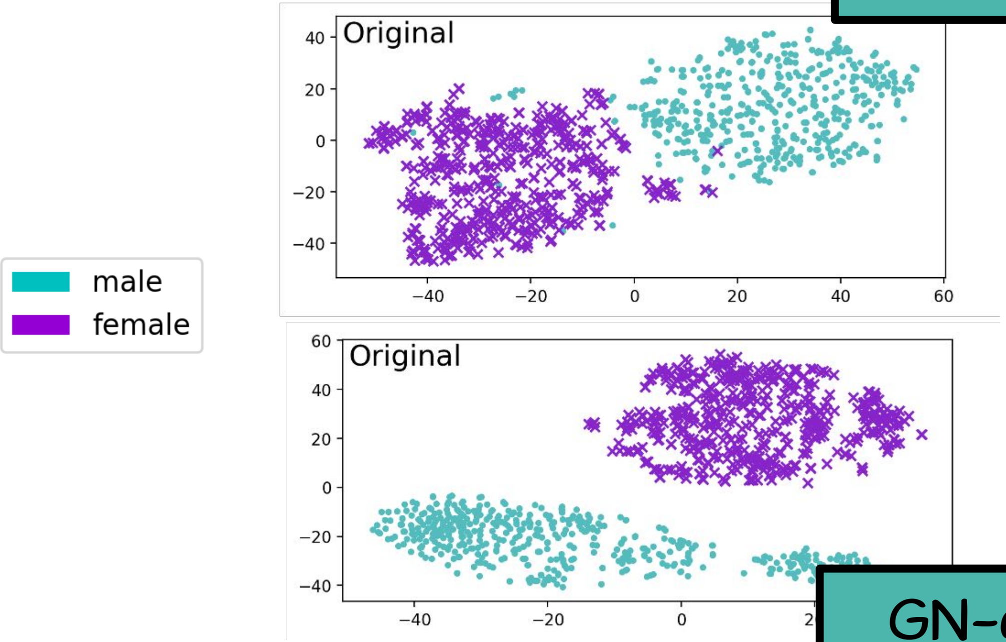
- We take the most biased words in the vocabulary according to the original bias (500 male, 500 female)

- We cluster them into two clusters using K-means



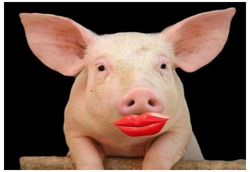
Clustering male- and female- biased words

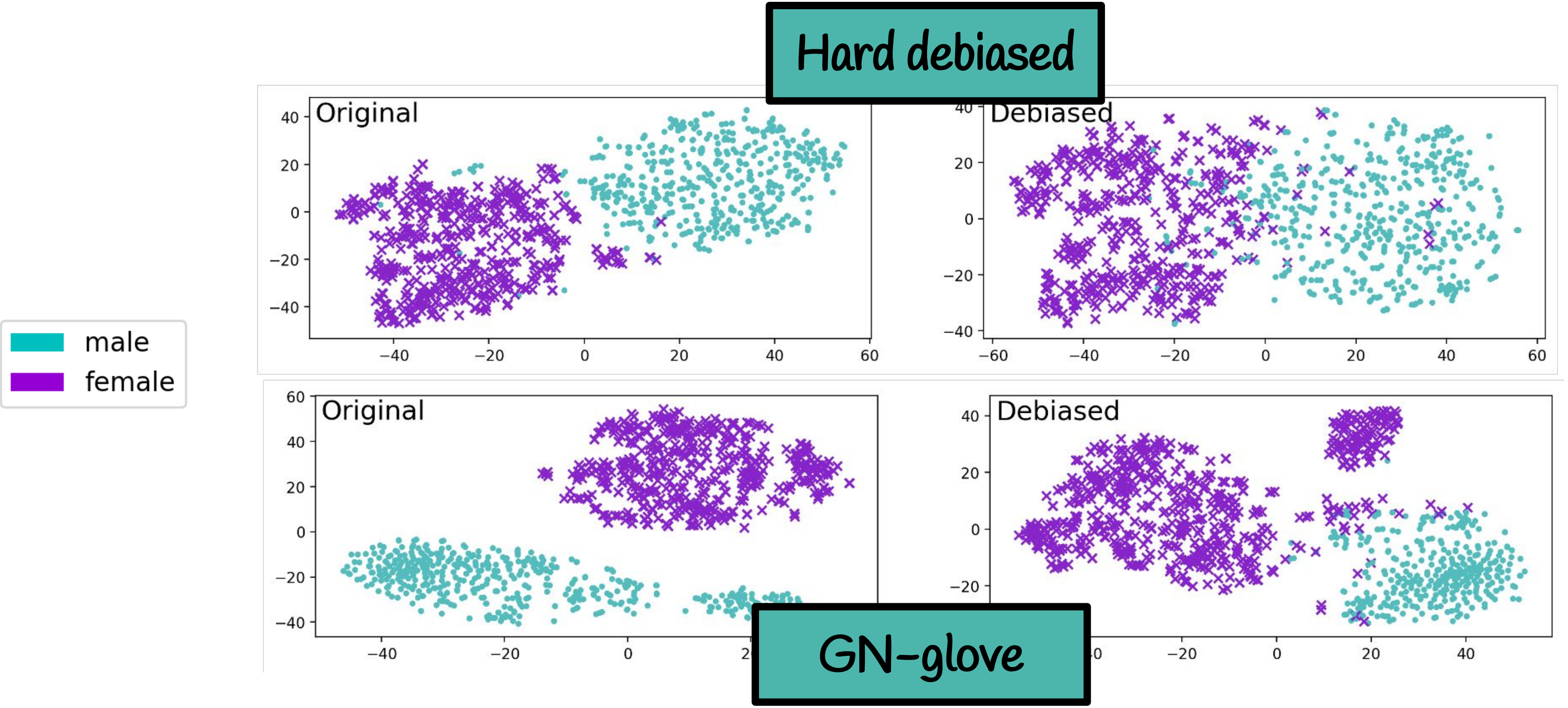
Hard debiased



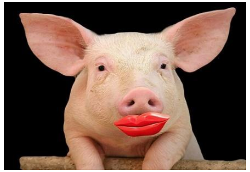
GN-glove

Clustering male- and female- biased words





Clustering male- and female- biased words

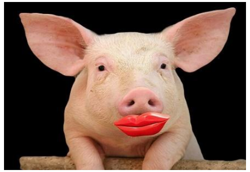


- We take the most biased words in the vocabulary according to the original bias (500 male, 500 female)

- We cluster them into two clusters using K-means

- **The clusters align with gender with accuracy of:**
  - **92.5% compared to 99.99% (HARD-DEBIASED)**
  - **85.6% compared to 100% (GN-GLOVE)**

Bias by neighbors



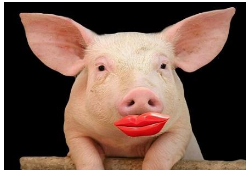
Bias is still manifested in similarities between words

An alternative mechanism for measuring bias:

- The **percentage of male/female** socially-biased words among the **k-nearest neighbors** of the target word

Pearson correlation with bias-by-projection:

- **0.69** compared to 0.74 (HARD-DEBIASED)
- **0.74** compared to 0.77 (GN-GLOVE)

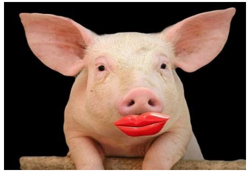


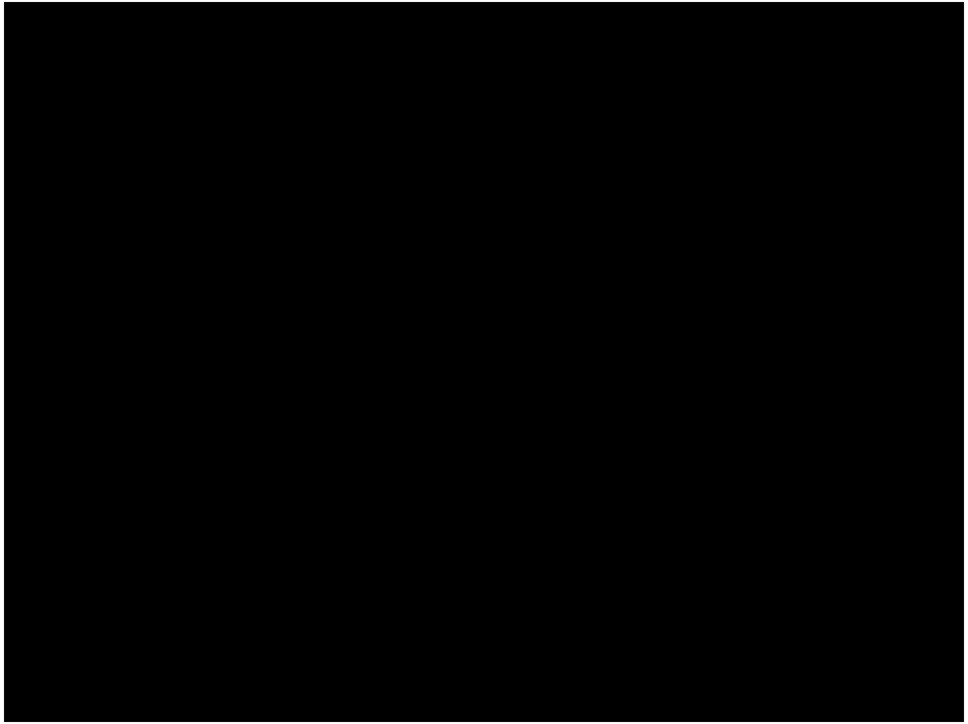
Professions

We take a predefined list of professions

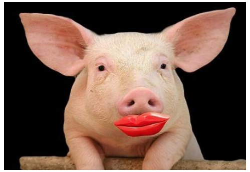
We show correlation between the **bias-by-projection** and **bias-by-neighbors**, before and after debiasing

Professions

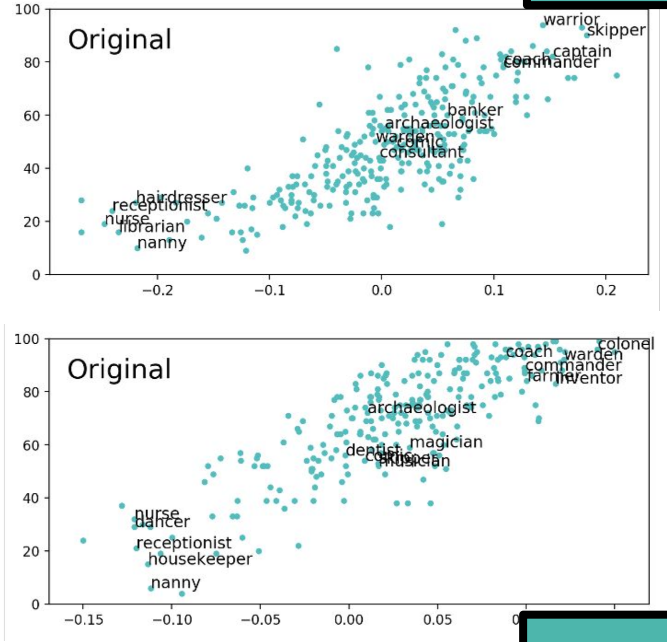




Professions



Hard debiased

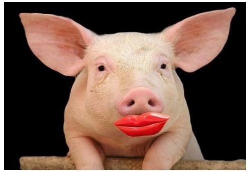


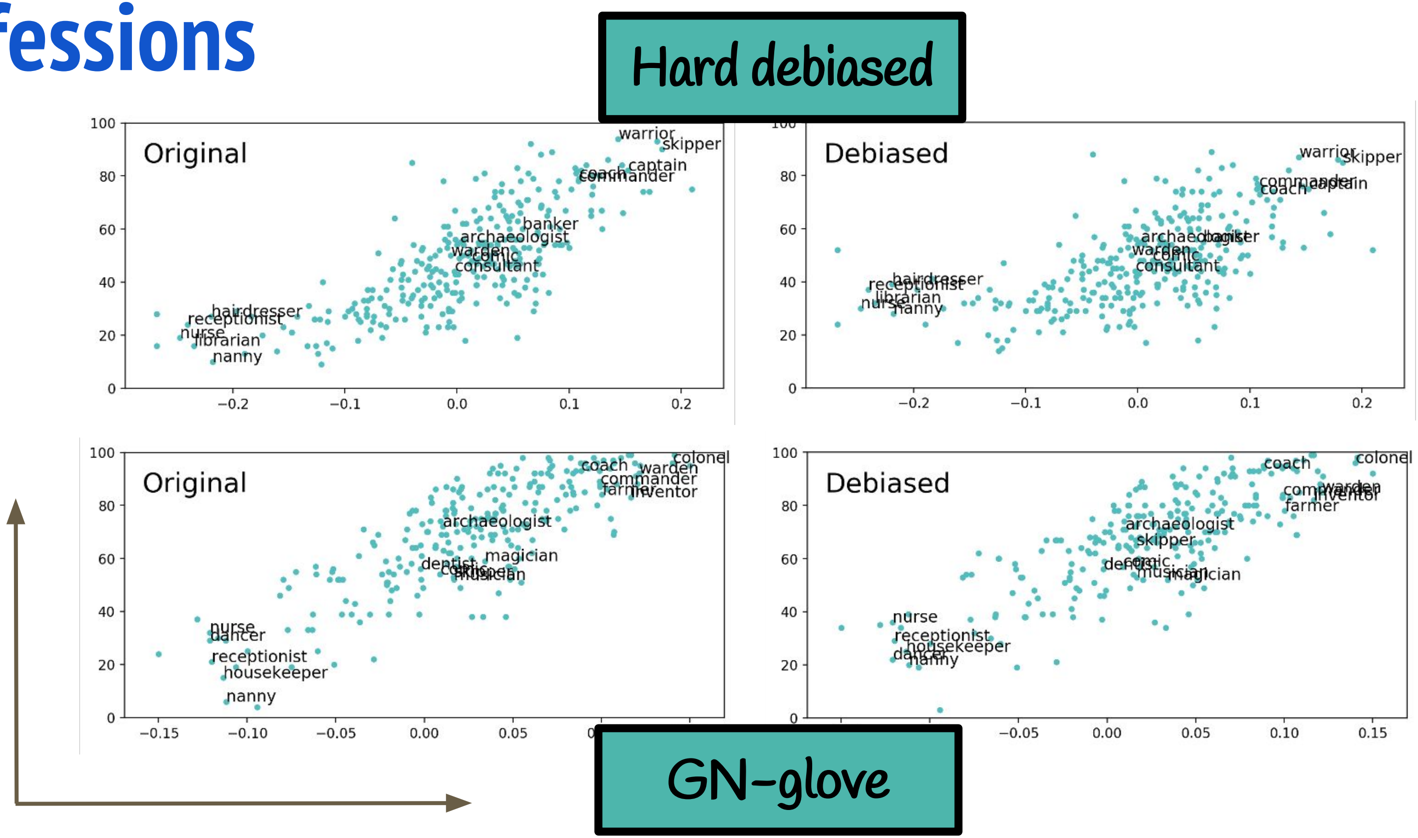
Bias by neighbors

GN-glove

Original bias by projection (reference)

Professions

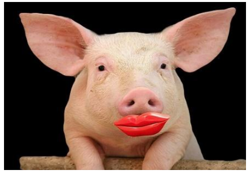




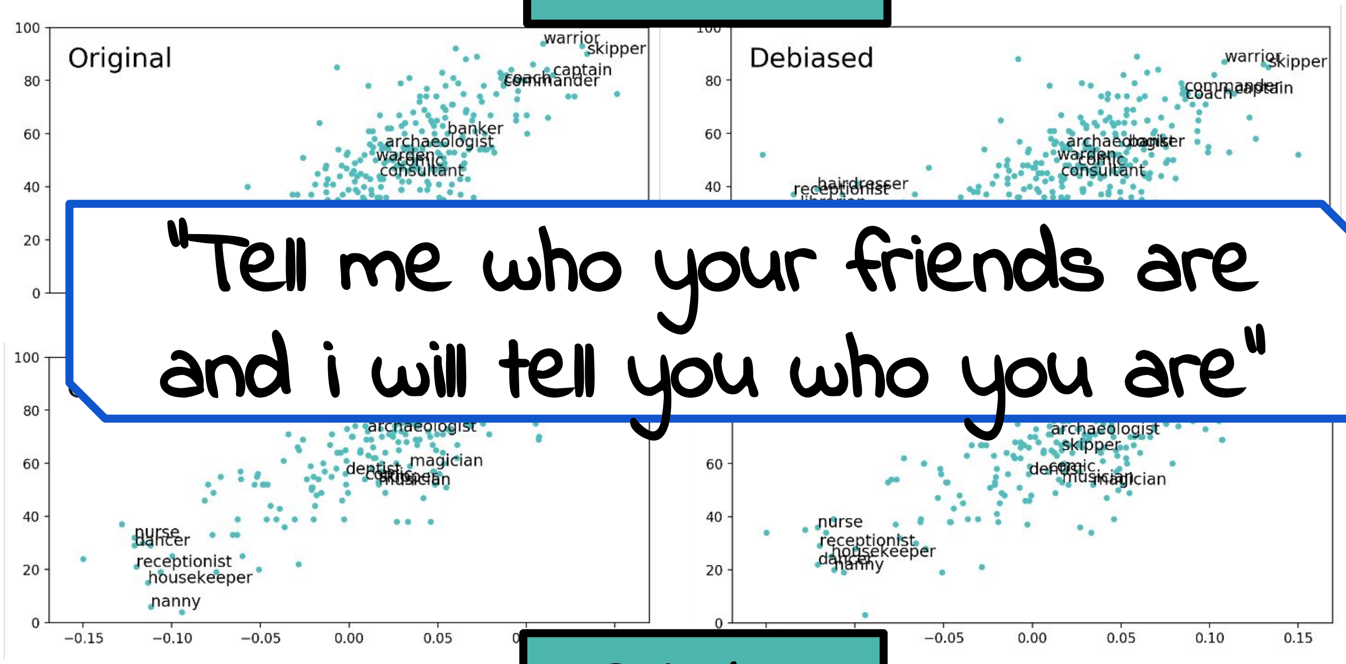
Bias by neighbors

Original bias by projection (reference)

Professions



Hard debiased

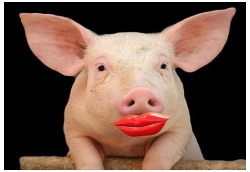


"Tell me who your friends are and i will tell you who you are"

GN-glove

Bias by neighbors

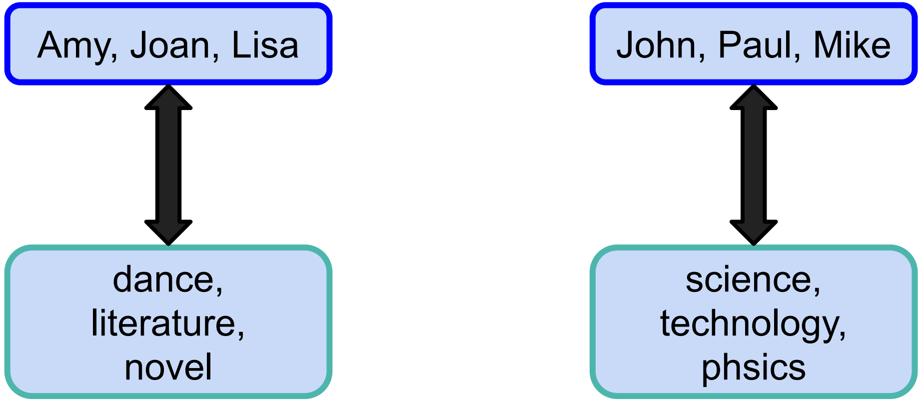
Original bias by projection (reference)



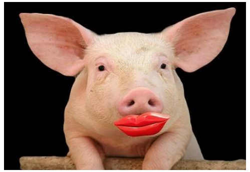
Association with stereotypes

We reproduce the experiments from Caliskan et al.

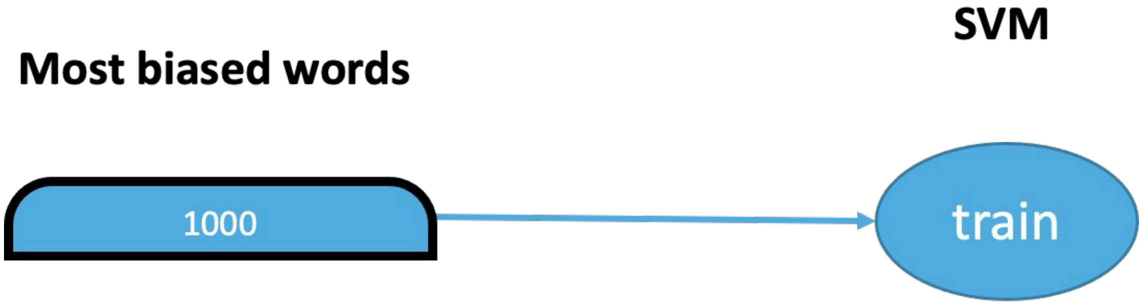
All associations are significant with $p < 0.0005$ also after debiasing



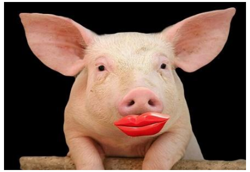
Classifying to gender



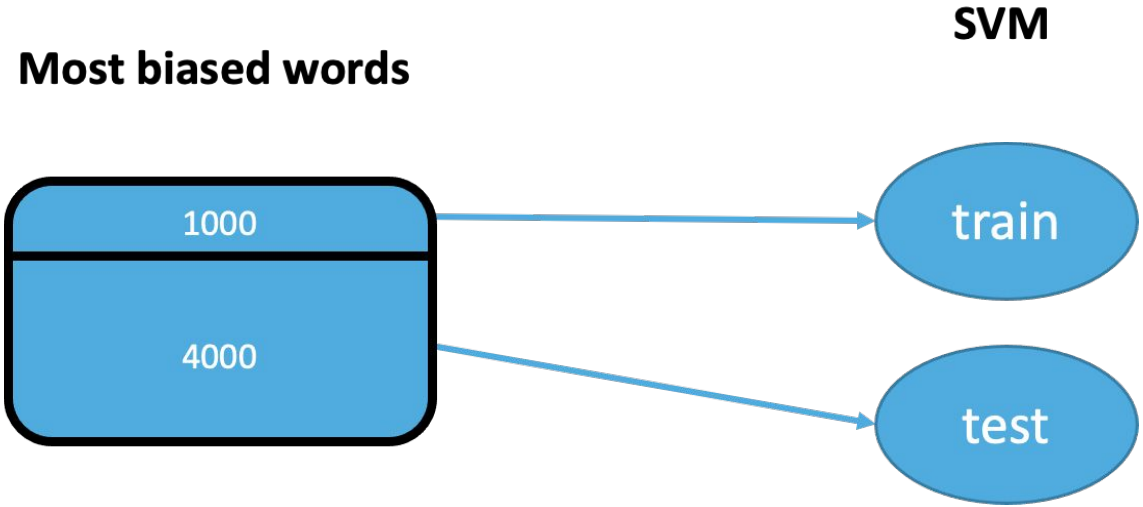
Can we train a **classifier** to predict gender based on the vectors?



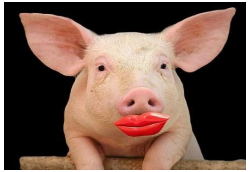
Classifying to gender

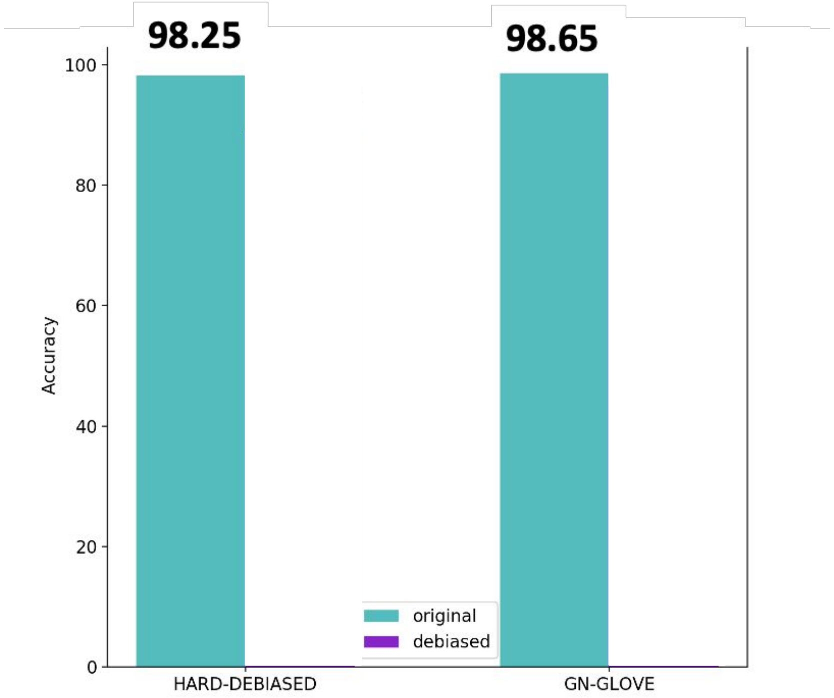


Can we train a **classifier** to predict gender based on the vectors?

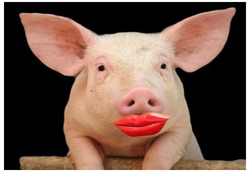


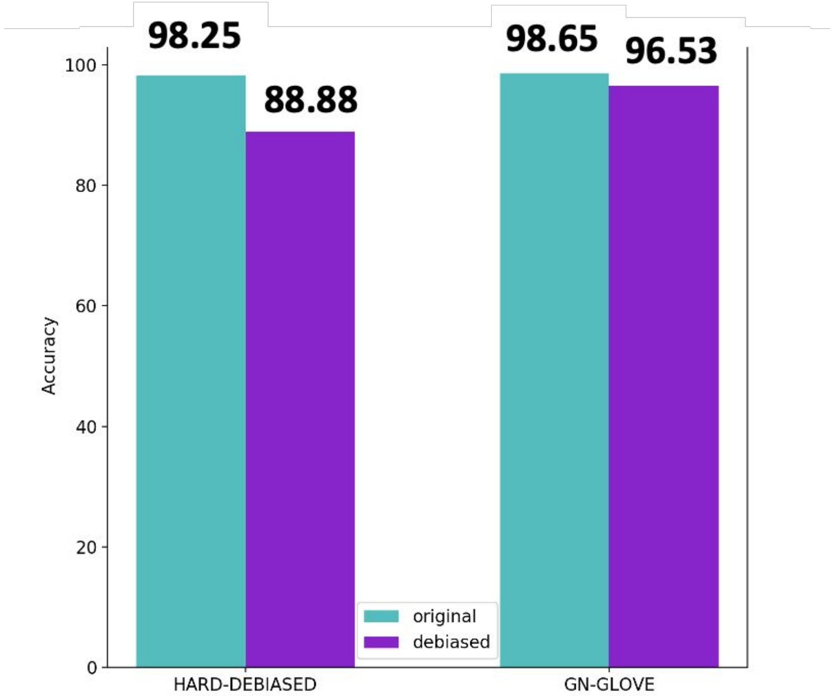
Classifying to gender



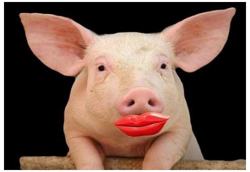


Classifying to gender





What have we seen?



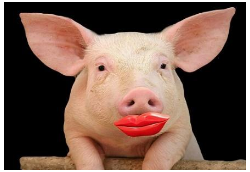
The **embedding space** stays largely the same

Stereotyped words still tend to group together

Clustering of representations reveals gender bias, even when not measured directly (using projection)

Gender of words with strong previous bias is **easy to predict** based on their vectors alone

What does that mean?



Debiasing based on the projection on the gender direction is **mostly superficial**

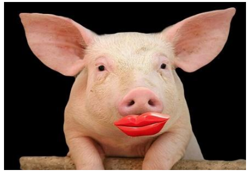
The societal bias is **deeply ingrained** in the embeddings

Gender-direction provides a way to measure the bias.
harder to measure after removing, but bias is still there

Gender bias definition is not reliable and **should be revisited**

Evaluation!

Conclusion

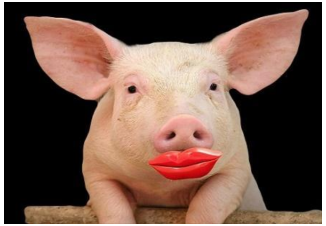


- Word embeddings **exhibit gender bias**
  - Societal gender bias is picked up from the data by the models

- **Debiasing is hard!**
  - A lot of the bias information is still recoverable when debiasing based on the gender direction

- Debiasing should be done **carefully**, while revising definitions and evaluations alike

Thanks! Questions?



What *can* we do about it then?

Two types of interventions in follow up works:

1. On the data level:

   It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution

2. On the representation level:

   Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

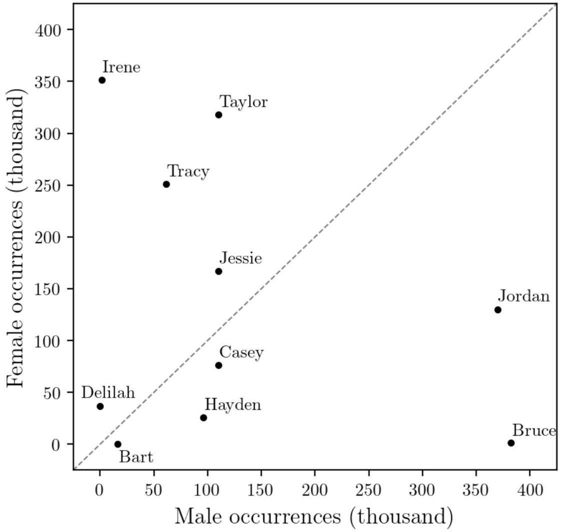
It’s All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution

Rowan Hall Maudslay[1] **Hila Gonen**[2] **Ryan Cotterell**[1] **Simone Teufel**[1]
[1] Department of Computer Science and Technology, University of Cambridge
[2] Department of Computer Science, Bar-Ilan University
{rh635,rdc42,sht25}@cam.ac.uk hilagnn@gmail.com

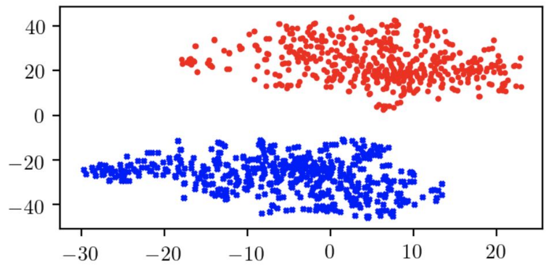
What can we do about it then?

Counterfactual Data Substitution:

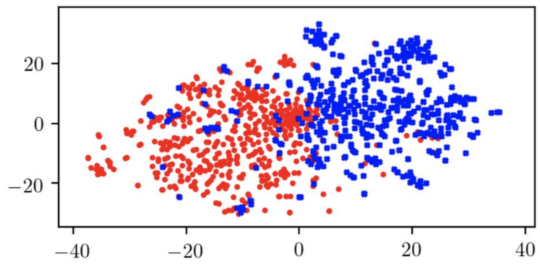
1. Swap gendered words in 50% percent of the documents
2. Names intervention while considering:
    a. Name frequency
    b. Gender specificity



before



after



We intervene on the data-level

INLP

INLP is a method for removing information from neural representations (Ravfogel et al. 2020):

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel[1,2] **Yanai Elazar**[1,2] **Hila Gonen**[1] **Michael Twiton**[3] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Independent researcher
{shauli.ravfogel, yanaiela, hilagnn, mtwito101, yoav.goldberg}@gmail.com

INLP

INLP is a method for removing information from neural representations (Ravfogel et al. 2020):

1. Train a linear **classifier** that predicts a certain property to remove
2. Project the representations on the classifier's null-space
3. **Repeat**

The classifiers become oblivious to the target property
hard to linearly separate the data according to it

INLP

INLP is a method for removing information from neural representations (Ravfogel et al. 2020):

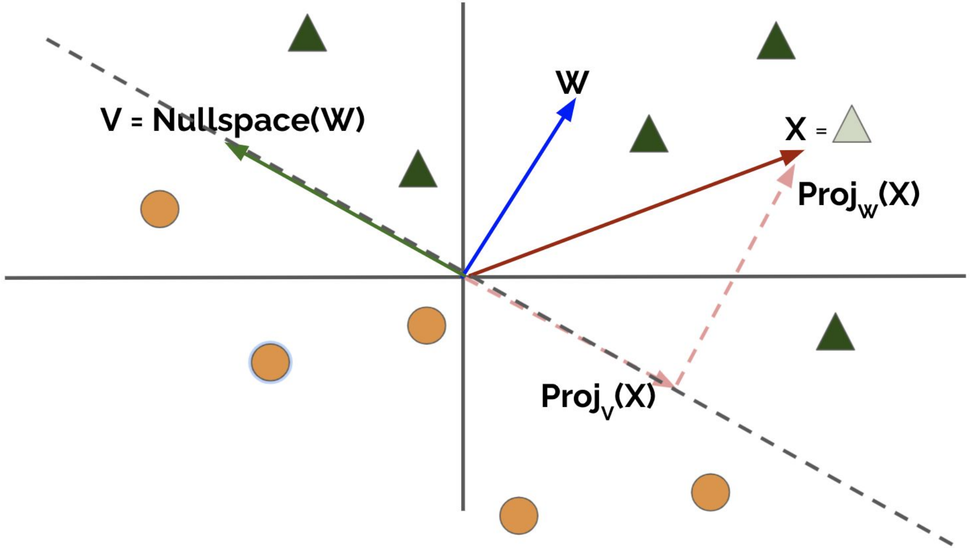
1. T
2. F
3. **Repeat**

We intervene on the representation-level

The classifiers become oblivious to the target property hard to linearly separate the data according to it

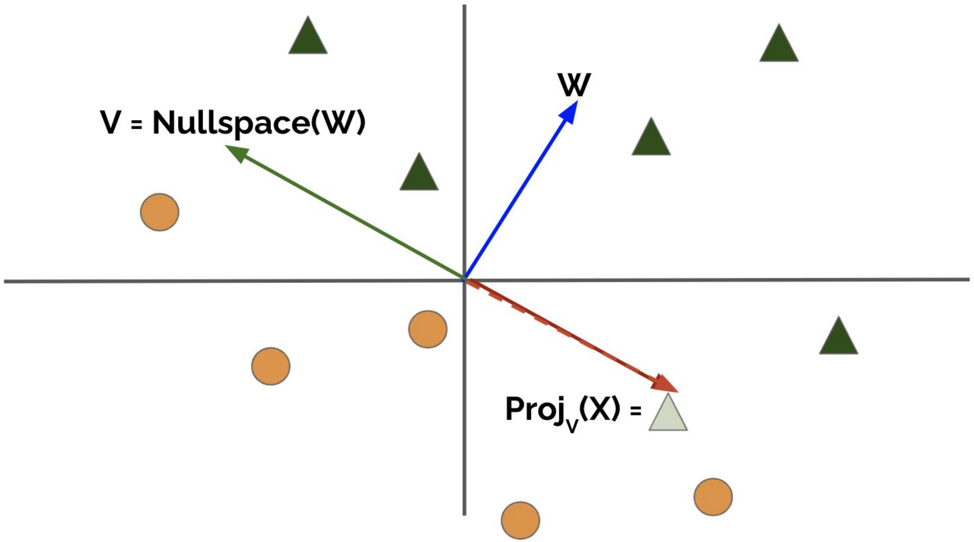
INLP

A single iteration:

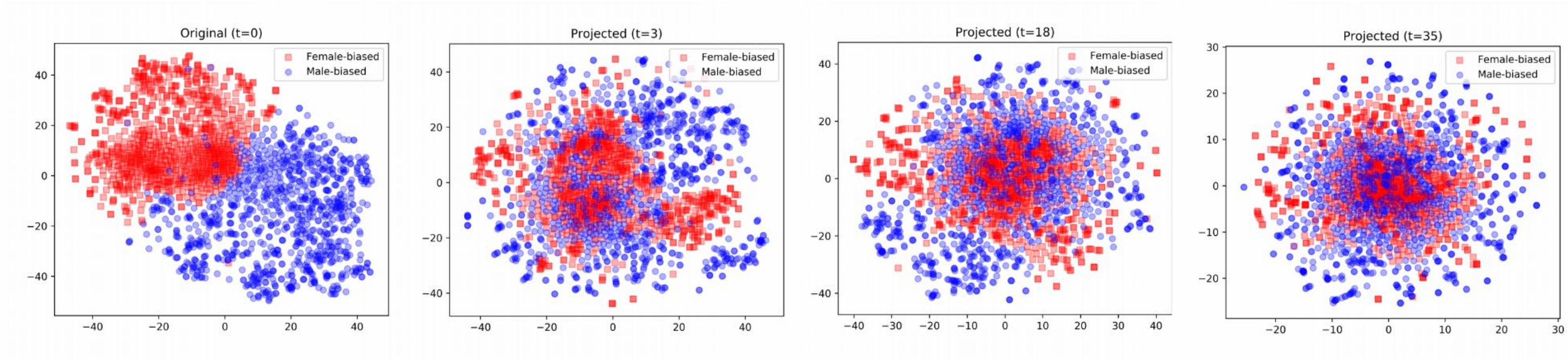


INLP

A single iteration:



INLP



We show that it works substantially better at removing bias!