Gender Discrimination in Natural Language Processing

A focus on embedding biases and possible solutions

Davide Angelani
Eric Rossetto
Giuseppe Murro
Salvatore Pisciotta
Xiaowei Wen

Alma Mater Studiorum University of Bologna
Ethics in Artificial Intelligence 2021/22 course

Table of contents

Intro

Introduction

Many *Natural Language Processing* (NLP) systems affect actual people:

- systems that interact with people (conversational agents)
- perform some reasoning over people (e.g. recommendation systems, targeted ads)
- make decisions about people's lives (e.g., employment, immigration, parole decision - form of early release of prisoners in Canada)

Questions of **ethics** arise in all of these applications!

Trustworthy AI

From **The Ethics Guidelines for Trustworthy Artificial Intelligence (AI)**:

”Data sets used by AI systems (both for training and operation) may suffer from the inclusion of **inadvertent historic bias**, incompleteness and bad governance models.
The continuation of such biases could lead to **unintended (in)direct prejudice and discrimination** against certain groups or people, potentially exacerbating prejudice and marginalisation.[...]
The way in which AI systems are developed (e.g. algorithms’ programming) may also suffer from unfair bias.”

Example of ethical issue

Definition of lexico.com (powered by Oxford)

Sexism is defined as ”prejudice, stereotyping, or discrimination, typically against women, on the basis of sex.”

GeBNLP 2022

Association for Computational Linguistics

4th Workshop on Gender Bias in Natural Language Processing at ACL conference

At NAACL in Seattle, USA, during July 15, 2022

”Gender bias, among other demographic biases (e.g. race, nationality, religion), in machine-learned models is of increasing interest to the scientific community and industry.There is a growing body of research into improved representations of gender in NLP models.[...] In order to make progress as a field, we **need to create widespread awareness of bias and a consensus on how to work against it**, for instance by developing standard tasks and metrics.[...]”

Gender bias in Word Embeddings

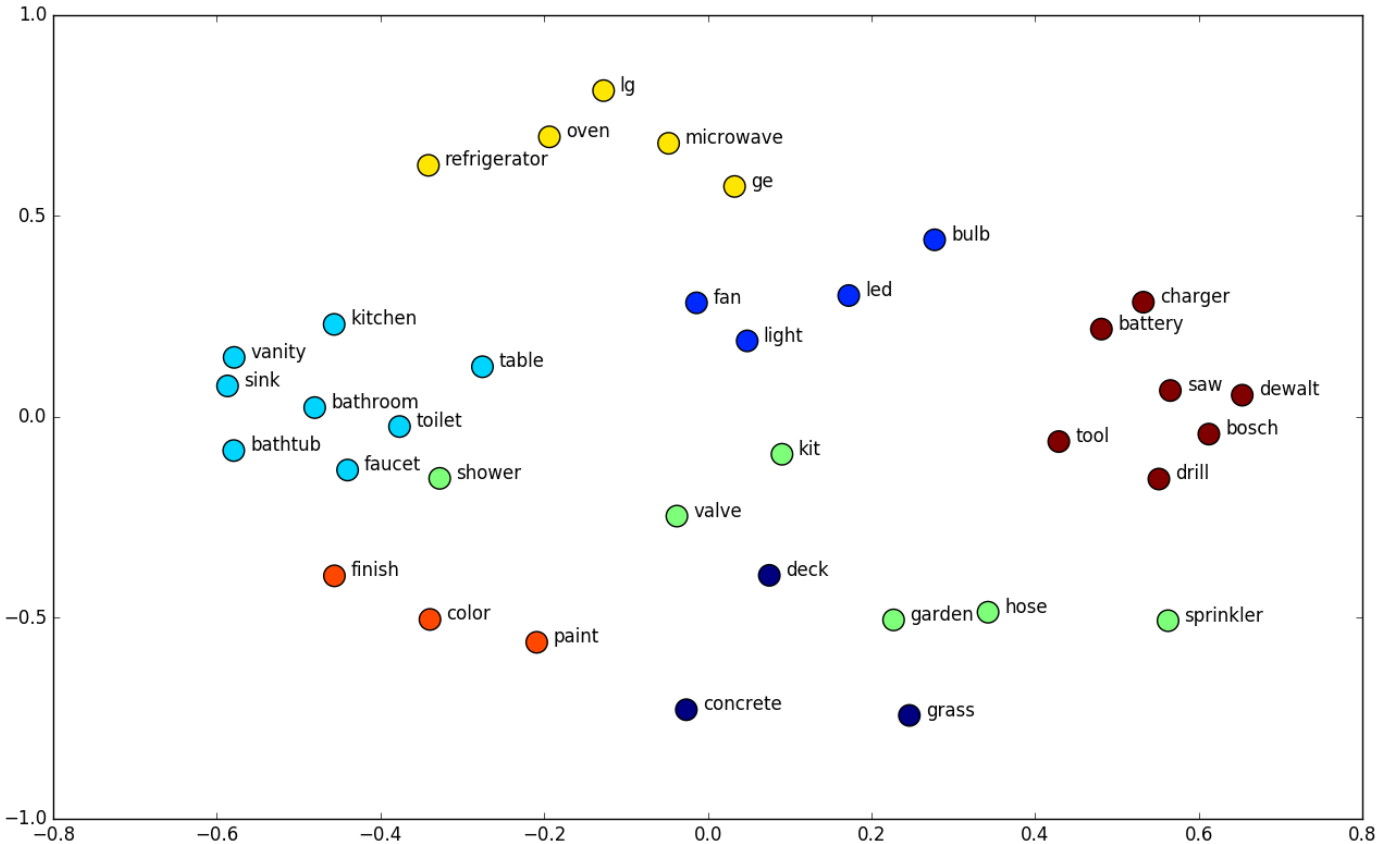
Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

One of the first focus on this problem was made by Bolukbasi et al. (2016) in the paper **"Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings"**

Word Embeddings

Word embeddings are a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Each word is mapped to one vector and the vector values are **learned**

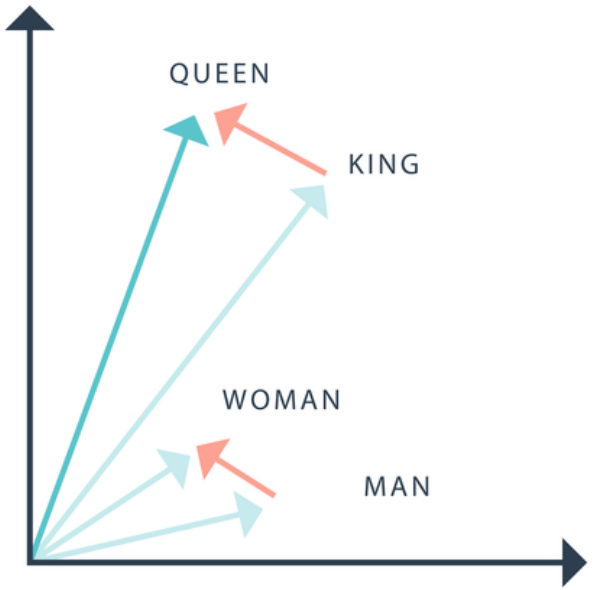


Metrics

How is possible to calculate bias?

The word embedding vectors are expected to reflect word relations by their geometrical relations and so it is possible to see the biases in them as the imperfection in this vector representation where words are close or far from other words.

Based on this definition it is possible to measure bias in terms of geometric relations.



Different types of bias

In order to investigate bias is important to distinguish between two kind of biases:

- **Skew**: describes the effect that all words from a set are (on average) biased towards the same attribute.
- **Stereotype**: shows the presence of groups stronger associated with different bias attributes than other groups.

Where attribute indicate the set of words related to the same semantic meaning.
For example man, son, husband,etc. for male gender and woman, daughter, wife,etc. for female gender.

Metrics requirements

According to Schröder et al. (2022) good metric has to be:

- **Stereotype and skew sensitive**
- **Comparability**: be sensitive towards the relation of neutral words towards a bias direction and invariant to other embedding properties (i.e. relations of attributes towards each other, embedding dimensions, average vector lengths in the embedding space)
- **Trustworthy**: always report a bias if there is at least one biased word in the set (i.e. the word is not equidistant to all attribute sets). On the other hand if all words are unbiased (equidistant to all attribute sets), then the bias score must report no bias.

SAME metric

The metric used in order to quantify the bias in our experiments is the SAME metric that, always according to Schröder et al. (2022), satisfy all the previous requirements in a better way with respect to other presented in the literature like WEAT, MAC and Direct Bias.

In its general form it is expressed like:

$$b(W, A_i, A_j) = \frac{1}{|W|} \cdot \sum_{\mathbf{w} \in W} |b(\mathbf{w}, A_i, A_j)|$$

Skew form :

$$b_{skew}(\mathbf{W}, A_i, A_j) = \frac{1}{|W|} \cdot \sum_{\mathbf{w} \in W} b(\mathbf{w}, A_i, A_j)$$

Stereotype form:

$$b_{stereo}(\mathbf{W}, A_i) = \frac{1}{|W|} \cdot \sqrt{\sum_{\mathbf{w} \in W} (b(\mathbf{w}, A_i, A_j) - b_{skew}(\mathbf{w}, A_i, A_j))^2}$$

with

$$b(\mathbf{W}, A_i) = \cos(\mathbf{w}, \hat{a}_i - \hat{a}_j) \quad and \quad \hat{a}_i = \sum_{\mathbf{a} \in A_i} \frac{1}{|A_i|} \cdot a_i$$

Case of study

Analyze bias in Italian Word Embeddings

Goal: Develop a proof of concept about how to address the gender discrimination in NLP

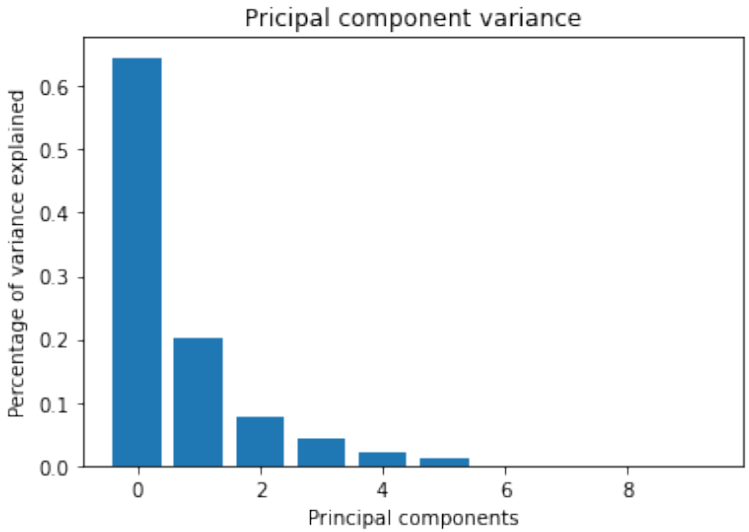
- The stereotypes encoded in word embeddings and depend on a linguistic and cultural context therefore they must be considered for different languages
- Just few papers are been published about adapting established approaches that assess gender bias in English WEs for **Italian WEs** (Biasion et al., 2020)
- Focus on **Italian Twitter embeddings word embeddings** created from 46.935.207 tweets, with 128 dimension and generated with word2vec models. Available at *ItaliaNLP Lab* website.

Gender direction

To identify a vectorial subspace which encodes information about **gender**, we follow Bolukbasi et al. (2016) by building a list of gender definitional pairs

Male/female definitionals

LUI (HE), LEI (SHE)
UOMO (MAN), DONNA (WOMAN)
PADRE (FATHER), MADRE (MOTHER)
MARITO (HUSBAND), MOGLIE (WIFE)
FRATELLO (BROTHER), SORELLA (SISTER)
MASCHIO (MALE), FEMMINA (FEMALE)



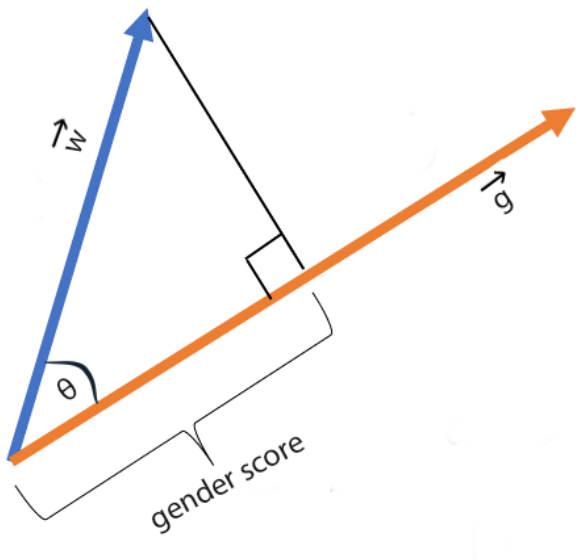
We perform a Principal Component Analysis (PCA) on the six vector differences resulting from each gender definitional pair. The first principal component explain 64% of variance. We normalize the first PC and consider it the main **gender direction**.

Gender score

Bolukbasi et al. (2016) define **gender score** $s_g(w)$ as the magnitude of the projection onto the gender direction g of a word embedding w representing a gender-neutral word:

$$s_g(w) = \|w\| \cos\theta = \frac{w \cdot g}{\|g\|}$$

- high value means that w is closer to the male terms
- strongly low value entails the opposite



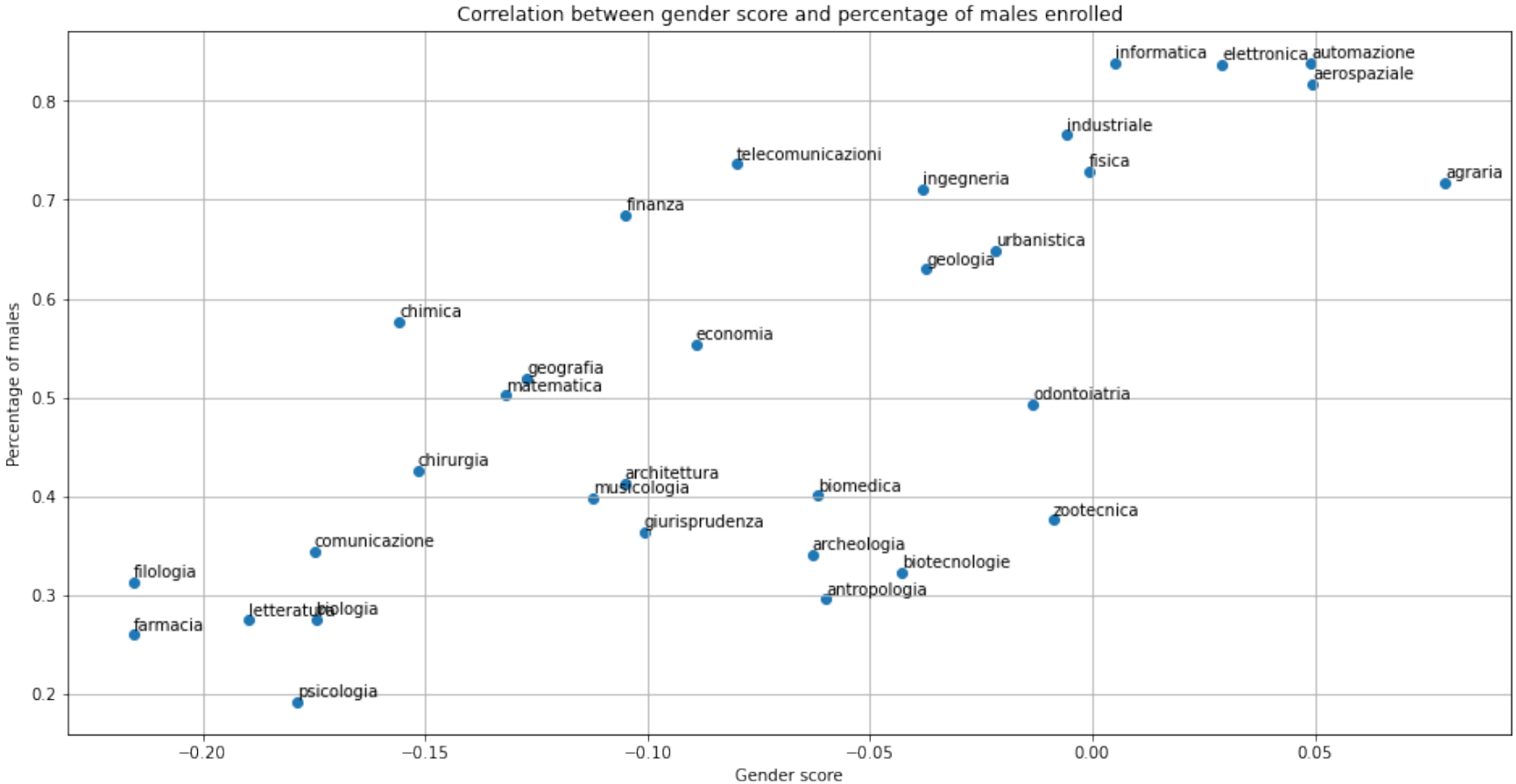
Gender score and university stereotypes

First attempt: demonstrate that also *Italian word embeddings* contain biases in their geometry that reflect gender stereotypes present in broader society

- We need to select gender-neutral words, regardless of *grammatical gender* assigned to all nouns in the Italian language
- Consider a list of 32 names of **degree classes** in the Italian university system
- Compute the *gender score* for the embedding of each word in the list
- Collect statistics about percentage of males enrolled in each degree class during the academic year 2020/21 (provided by the MIUR opendata) website

Gender score and university stereotypes

Results: actually the gender scores are strongly correlated with the percentage of males enrolled in each degree class (Pearson correlation of 0.73)



Hard Debiasing - Algorithm

Debiasing: our goal is to reduce biases in the word embedding preserving its useful properties. So we want to:

- reduce bias keeping gender neutral word equidistant between gender pairs (he-she)
- maintain embedding utility

The hard debiasing algorithm is composed by the following steps:

- Neutralize
- Equalize

Hard Debiasing - Neutralize

Neutralize: ensures that gender neutral words are zero in the gender subspace.

- Compute the projection of degree class embeddings onto the gender direction g

$$proj_g(w) = \frac{w \cdot g}{\|g\|^2} g$$

- Re-embedding every degree class name through an orthogonal projection

$$w' = \frac{w - proj_g(w)}{\|w - proj_g(w)\|}$$

- By construction, in this new embedding space, grammatical gender should have a lower influence on the geometry of word vectors

Hard Debiasing - Equalize

Equalize: ensures sets of words outside the gender subspace are equal and enforces the property that any neutral word is equidistant to all words in each equality pair.

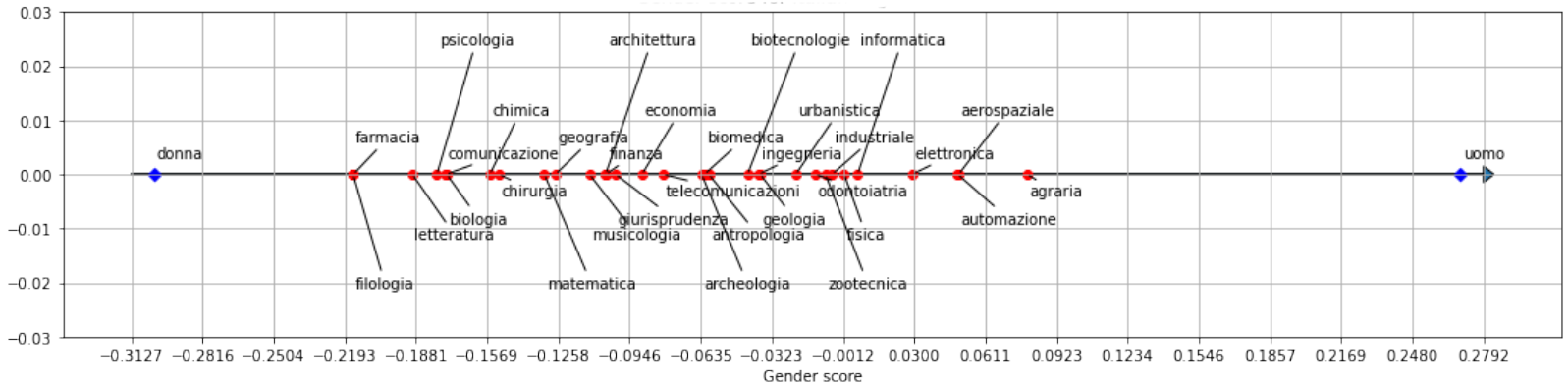
- Make use of family of equality pairs $\varepsilon = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$.
- Recenter the words into the equality pairs with respect to the gender direction

Equality pairs

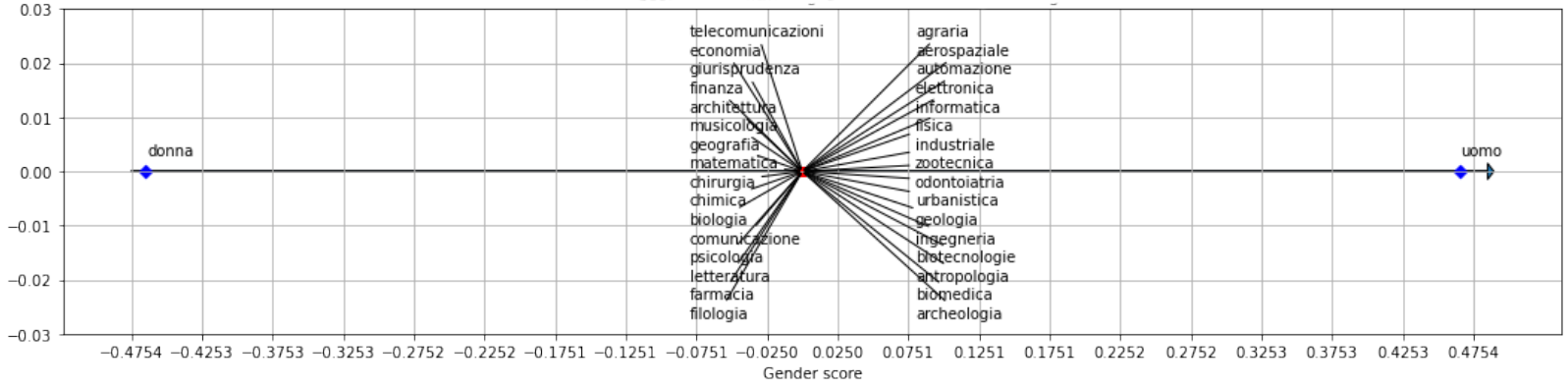
ragazzo (boy), ragazza (girl)
zio (uncle), zia (aunt)
figlio (son), figlia (daughter)
...
maestro (teacher), maestra (teacher)
cuoco (cook), cuoca (cook)

Hard Debiasing - Results

Gender scores before hard-debiasing:



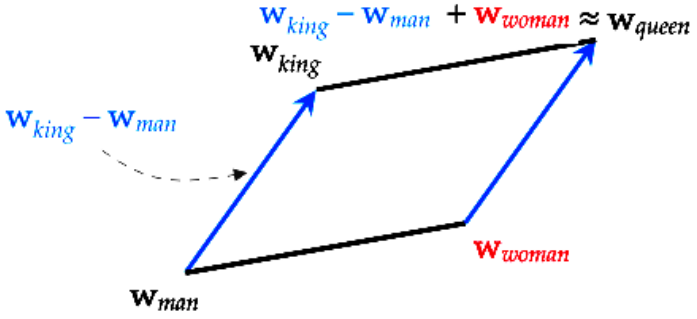
Gender scores after hard-debiasing:



Metrics	Pre-debiasing	Post-debiasing
SAME	0.089	1.21e-08
Skew	-0.075	6.63e-09
Stereotype	0.015	2.44e-09

Hard Debiasing - Evaluation

Analogies are a useful way to both evaluate the quality of a word embedding and also its stereotypes. They allow to make sure that our *debiasing algorithm* preserves the desirable properties of the original embedding while reducing the gender bias.



Analogies	Pre-debiasing	Post-debiasing
Fair	*uomo : re = donna : regina* *milan : italia = chelsea : inghilterra*	*uomo : re = donna : regina* *milan : italia = chelsea : inghilterra*
Biased	*uomo : informatica = donna : psicologia* *uomo : anatomia = donna : psicologia* *studente : matematica = studentessa : biologia* *lui : microeconomia = lei : filologia*	*uomo : informatica = donna : programmatrice* *uomo : psicologo = donna : psicologia* *studente : matematica = studentessa : algebra* *lui : letteratura = lei : filologia*

Problem of hard debiasing method

Zhao et al. (2018a) criticizes the hard-debiasing model by saying that it is essentially a pipeline approach and it strongly relies on the *pre-defined gender-neutral words list*.

This raises two problems:

- the re-embedding of words can **affect the performance** of the whole model
- the method **removes gender information** and this can be undesirable in domains such as social science and medicine

GN-GloVe

Goal: Develop an algorithm which trains word embeddings with protected attributed.

Due to the importance of word embeddings, (Zhao et al., 2018a,b) found the necessity to mitigate also the gender bias in such representations.

In particular, (Zhao et al., 2018b) proposed a GloVe-based method called **GN-GloVe** which learns the word embeddings while keeping protected attributes in certain dimensions while neutralizing the others during training.

GloVe

GloVe is a model proposed by (Pennington et al., 2014) for obtaining word embeddings:

- it is **unsupervised** algorithm
- the model trains on global word-word **co-occurrence counts** X and tries to capture the word proximity
- the context of a word is defined by a window of a predefined size which spans words near the target one.

GN-GloVe model

In *GN-GloVe* they defined:

- the word vector w which consists of two parts $w = [w^{(a)}; w^{(g)}]$ where $w^{(a)}$ and $w^{(g)}$ stand for neutralized and gendered components respectively,
- three predefined sets Ω_M, Ω_F and Ω_N in which all words are categorized and represents the male-definition, female-definition and gender-neutral respectively.

GN-GloVe model

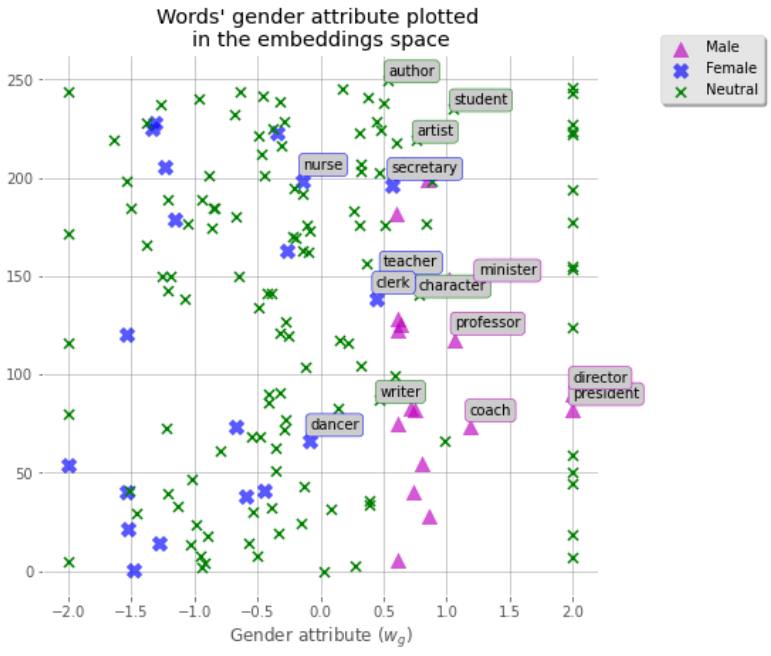
The model works by defining a new loss function

$$J = J_G + \lambda_d J_D + \lambda_e J_E$$

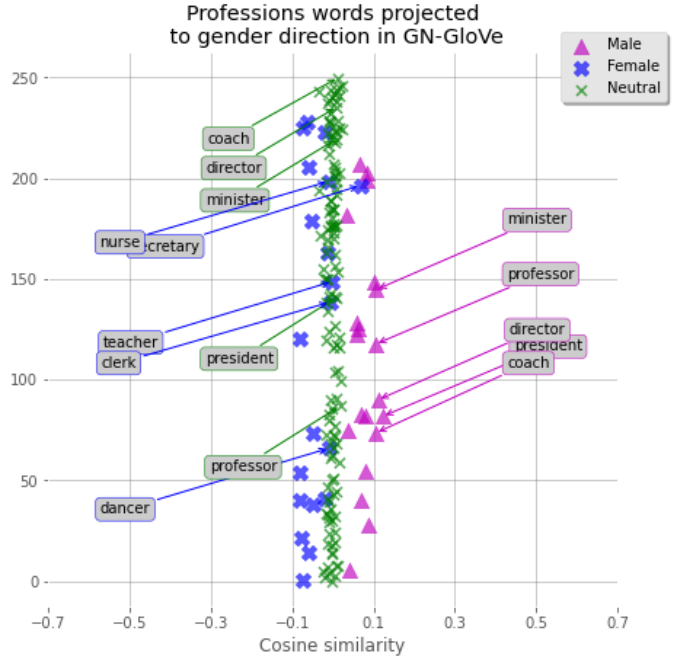
which is composed of three components:

- J_G is the GloVe loss function
- J_D is aimed to restrict gender information in $w^{(g)}$
- J_E is aimed to retain $w^{(a)}$ in the null space of the gender direction v_g

GN-GloVe results



(a) $w^{(g)}$ dimension for all the professions



(b) Gender-neutral profession words projected to gender direction in GN-GloVe

Other approaches

Debiasing Methods Overview

Besides the **hard debiasing** method, there exists several other approaches for debiasing gender stereotypes in NLP which mainly work on two tangents (Sun et al., 2019):

1. text corpus and their representation, i.e., their embeddings,
2. prediction algorithms

A further distinction could be made accordingly to how this methods affect the model:

- *retraining* methods,
- *inference* methods, in which the hard debiasing method falls.

Debiasing Methods by Adjusting Algorithms

Such approaches that act by debiasing predictions found by models are called **algorithm adjustment methods**:

- by using a constrained conditional model to *constrain predictions* which do not amplify biases present in the training set (Zhao et al., 2017),
- by exploiting the capabilities of *adversarial networks*: a generator that prevents the discriminator from identifying the gender in a task (Zhang et al., 2018).

Debiasing Methods via Data Manipulation

Performing **Data Augmentation** on the text corpora proved to be an easy and effective way to counteract an uneven distribution of occurrences of words specific to a particular gender (Zhao et al., 2018a).

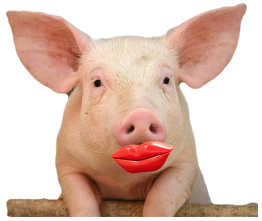
Concretely:

1. **create** an augmented corpus identical to the original one but biased toward the opposite gender by performing *gender-swapping*, e.g.,

   "*He* works in engineering" $\longrightarrow$ "*She* works in engineering",

2. **train** the model on the union between the original corpus and the augmented one.

Lipstick on a Pig

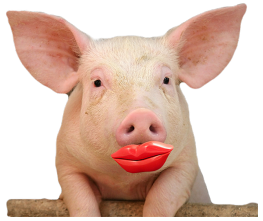


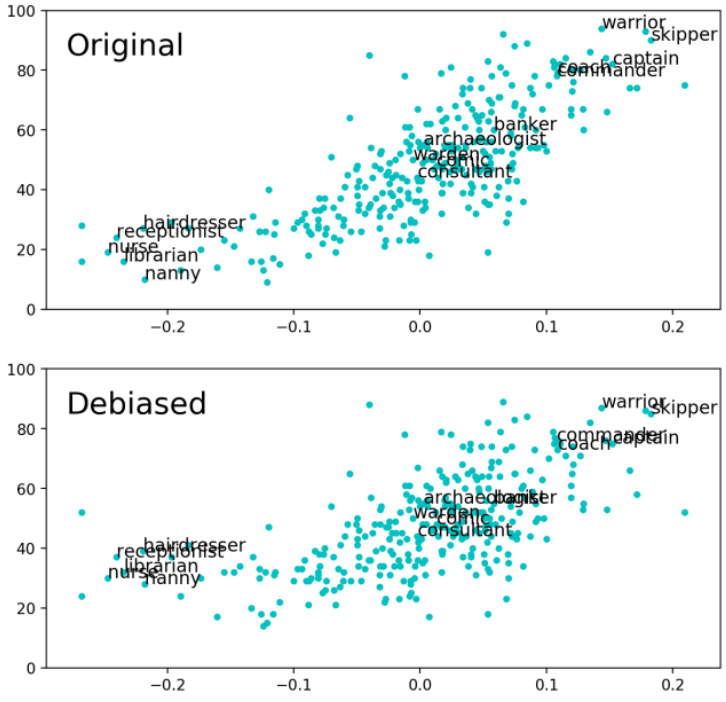
In Gonen and Goldberg (2019), they experimentally proved that exists a systematic bias in the embeddings, which is **independent** of the gender direction.

Despite debiasing methods work well at removing the gender direction, the debiasing is still **superficial** $\rightarrow$ they are ”party tricks”:

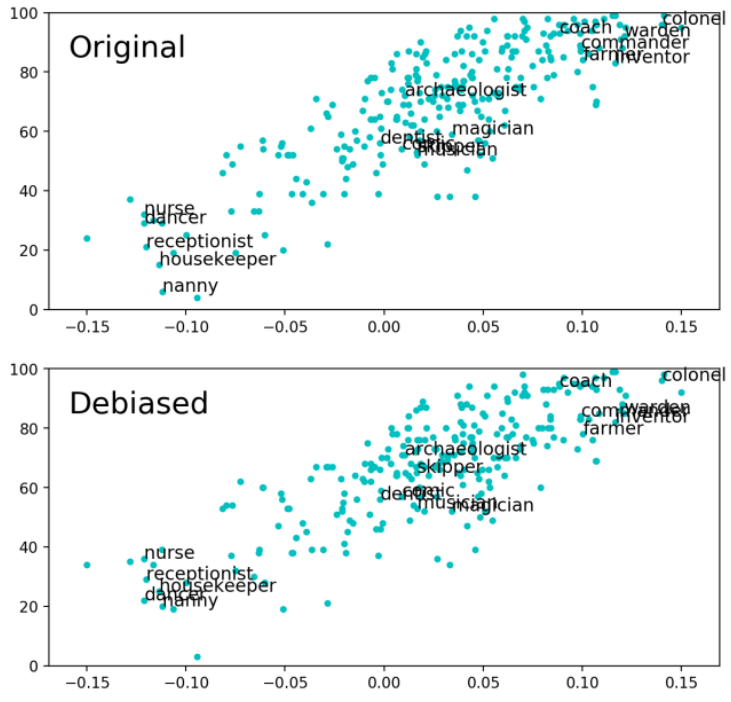
- there is the need to associate *implicitly* gendered terms with other implicitly gendered terms, or,
- look for gender-specific regularities in the corpus by learning to condition on gender-biased words and *generalizing* to other biased words.

Lipstick on a Pig



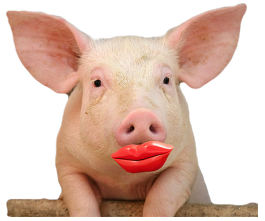


(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.



(b) The plots for GN-GLOVE embedding, before (top) and after (bottom) debiasing.

Lipstick on a Pig



Crucial point (Gonen and Goldberg, 2019)

Gender-direction provides a way to **measure** the gender-association of a word, but **does not determine it**.

The methods discussed so far are for the most part solely **hiding the gender bias** and not removing it.

The definitions provided for quantify and removing bias are thus **insufficient**, and we need to search for other aspects that could help us in our debiasing journey.

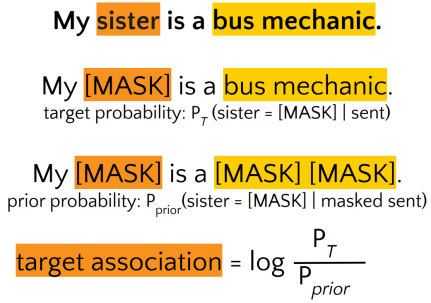
Future Developments

Impact on the latest technologies

Contextual word embeddings

Contextualized word embeddings have been replacing standard embeddings in NLP. Nowadays one of the most used is *BERT*, that learns contextualized word representations using a masked language modelling objective.

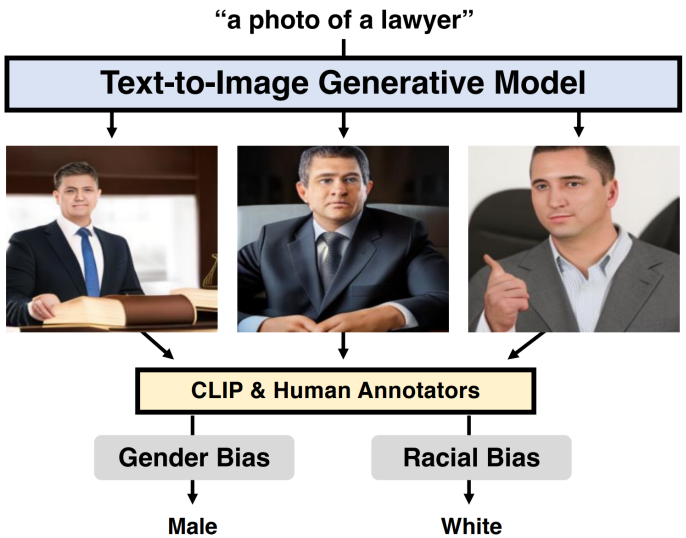
- **Measuring gender bias** - Make use of sentence templates to measure the association of *target* (male or female person) and *attribute* (profession or emotion) in a sentence. (Bartl et al., 2020)



- **Mitigating gender bias** - Apply *Counterfactual Data Substitution*, in which the gender of words denoting persons in a training corpus is swapped in place in order to counterbalance bias and then the model is fine-tuned

Impact on the latest technologies

DALL-E: text-to-image generation DALL-E is a transformer language model, based on GPT-3, trained to generate images from text descriptions, using a dataset of text–image pairs. (Ramesh et al., 2021)



- Cho et al. (2022) demostrated that these pretrained models learn specific **gender/racial biases** from web image-text pairs, showing gender bias towards male and racial bias towards white.
- current text-to-image generation models have several avenues for future research on understanding social biases

New challenges

- **Mitigating Gender Bias in Languages Beyond English** - Language-specific solutions are required, since gender is expressed in different ways across languages, so future work can look to apply existing methods or devise new techniques towards *mitigating gender bias in other languages* as well (Sun et al., 2019)
- **Non-Binary Gender Bias** - Most research on debiasing in NLP treats gender as a binary variable neglecting its fluidity and continuity. (Richards et al., 2016)

Conclusions

- Most of the newly developed algorithms do **not test their models** for bias and disregard possible ethical considerations of their work
- The propagation of gender bias in NLP algorithms poses the danger of **reinforcing damaging stereotypes** in downstream applications
- Real-world consequences have been raised about automatic **resume filtering** systems giving preference to male applicants
- We should investigate how **data can be better collected** avoiding bias in the first instance
- Perhaps we should attempt to **debias society** rather than word embeddings

Thanks for the attention!

Get the source code of this project and the generated results from
`github.com/WenXiaowei/GenderDiscriminationNLP`

References

M. Bartl, M. Nissim, and A. Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias, 2020. URL *https://arxiv.org/abs/2010.14534*.

D. Biasion, A. Fabris, G. Silvello, and G. A. Susto. Gender bias in italian word embeddings. In *CLiC-it*, 2020.

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, 2016. URL *https://arxiv.org/abs/1607.06520*.

J. Cho, A. Zala, and M. Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. 2022.

H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, 2019. URL *https://arxiv.org/abs/1903.03862*.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL *http://www.aclweb.org/anthology/D14-1162*.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation, 2021. URL *https://arxiv.org/abs/2102.12092*.

C. Richards, W. P. Bouman, L. Seal, M. J. Barker, T. O. Nieder, and G. T'Sjoen. Non-binary or genderqueer genders. *International Review of Psychiatry*, 28(1):95–102, 2016. URL *https://doi.org/10.3109/09540261.2015.1106446*.

S. Schröder, A. Schulz, P. Kenneweg, R. Feldhans, F. Hinder, and B. Hammer. The same score: Improved cosine based bias score for word embeddings, 2022. URL *https://arxiv.org/abs/2203.14603*.

T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review, 2019. URL *https://arxiv.org/abs/1906.08976*.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning, 2018. URL *https://arxiv.org/abs/1801.07593*.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, 2017. URL *https://arxiv.org/abs/1707.09457*.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics, jun 2018a. doi: 10.18653/v1/N18-2003. URL *https://aclanthology.org/N18-2003*.

J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, Oct.-Nov. 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL *https://aclanthology.org/D18-1521*.