UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

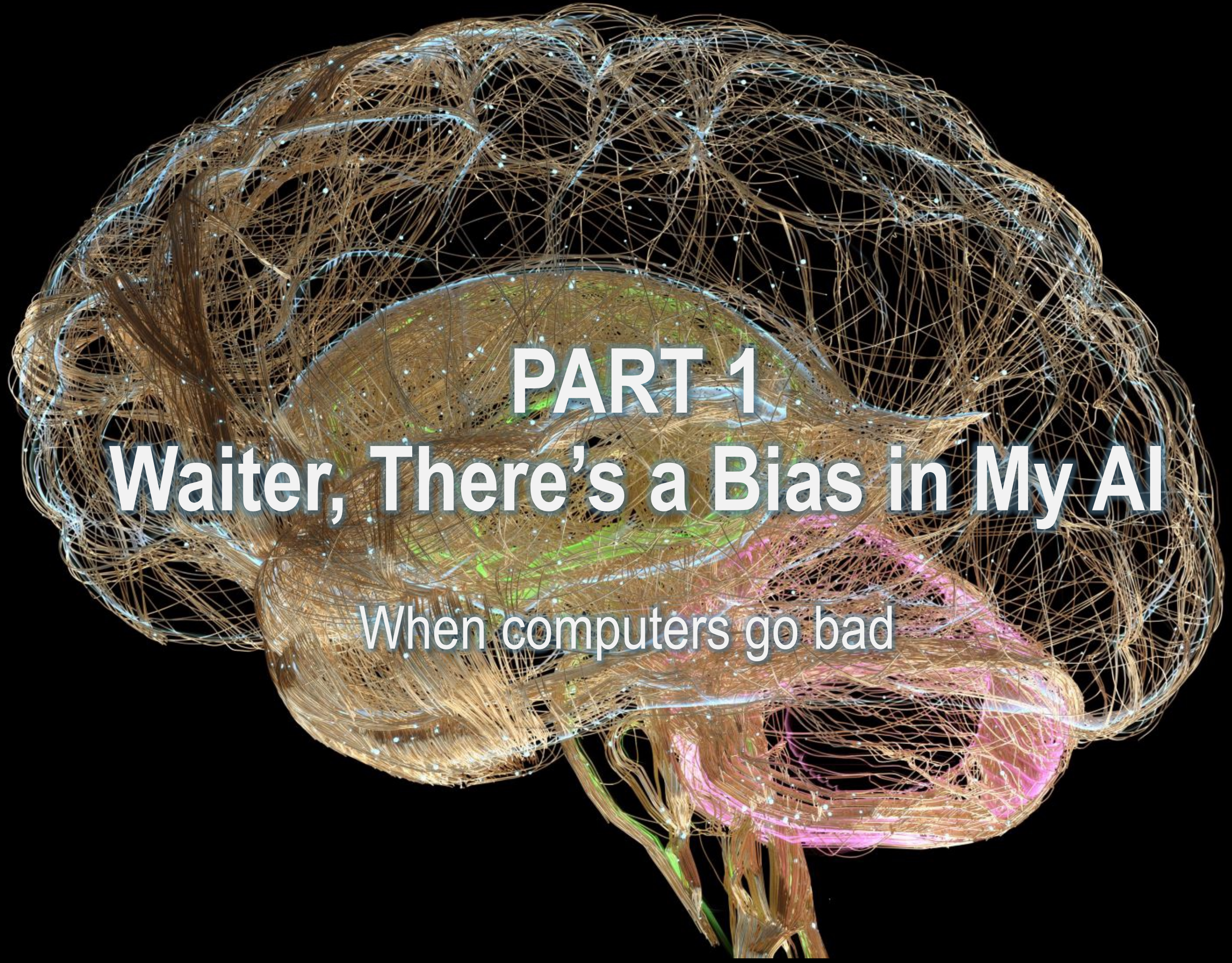
Intrinsic Bias Metrics Do Not Correlate with Application Bias

Ricardo Muñoz Sánchez

Based on work done with:
Seraphina Goldfarb-Tarrant, Rebecca Marchant, Mugdha Pandya, and Adam Lopez

Overview

- Biases in AI
- How do we measure them?
- Can they be removed?
- Even if we can, does it do *anything* at all?

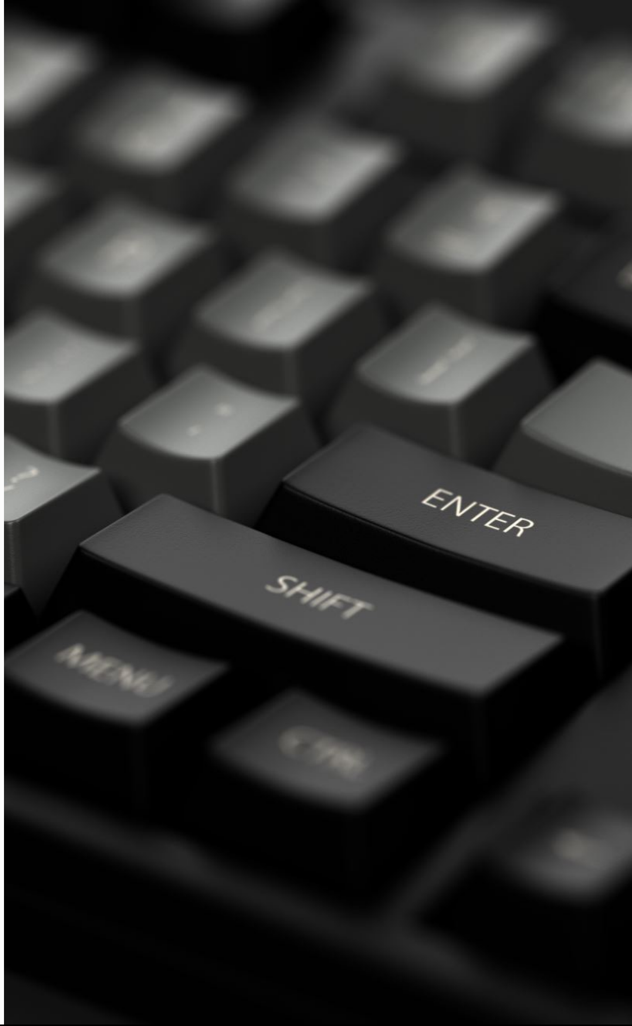


PART 1
Waiter, There's a Bias in My AI

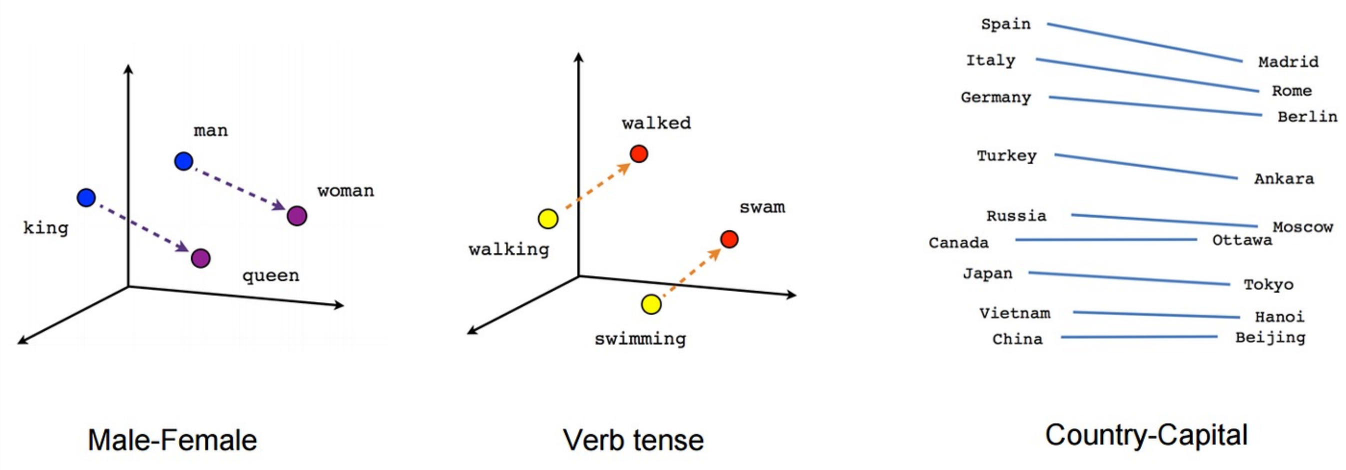
When computers go bad

Word Embeddings

- Ways to represent words in a computer-interpretable manner
- They encode both semantics and syntax of words
- Have nice geometric properties



Word Embeddings – Geometry



Word Embeddings – Analogies

- *Man* is to *king* as *woman* is to… *queen*
- *Walk* is to *swim* as *walking* is to… *swimming*
- *Spain* is to *Madrid* as *Italy* is to… *Rome*

Word Embeddings – Analogies

- *Man* is to *king* as *woman* is to… *queen*
- *Walk* is to *swim* as *walking* is to… *swimming*
- *Spain* is to *Madrid* as *Italy* is to… *Rome*
- *Man* is to *programmer* as *woman* is to…

Word Embeddings – Analogies

- *Man* is to *king* as *woman* is to… *queen*
- *Walk* is to *swim* as *walking* is to… *swimming*
- *Spain* is to *Madrid* as *Italy* is to… *Rome*
- *Man* is to *programmer* as *woman* is to… *homemaker*

Wait, what?

Encoding Biases

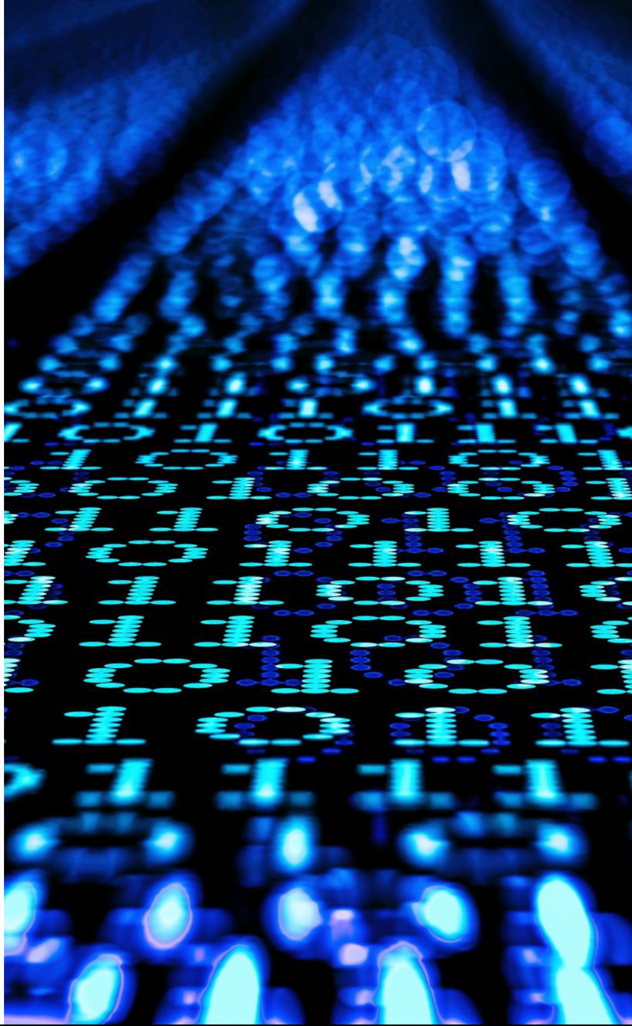
- As with all AI models, embeddings find and exploit patterns in the data

- However, stereotypes are patterns in the data

- Our systems can and will pick these patterns and perpetuate unwanted biases, such as sexism and racism

Encoding Biases

- As with all AI models, embeddings find and exploit patterns in the data
- However, humans are biased and this is reflected in the data we produce
- Our models can and will pick these patterns and perpetuate unwanted biases



Correference Resolution

- The doctor hired a nurse because he was busy (Correct)

Correference Resolution

- The doctor hired a nurse because he was busy (Correct)
- The doctor hired a nurse because she was busy (Wrong)

Correference Resolution

- The doctor hired a nurse because he was busy (Correct)
- The doctor hired a nurse because she was busy (Wrong)
- The doctor hired a nurse because she was busy (Correct)

PART 2
Measuring Bias

Fixing the World by Fixing our Models

Bias Metrics

Intrinsic

- Measure unwanted associations in language models and embeddings
- Examples are WEAT, CEAT, and DisCo
- Also called bias metrics

Extrinsic

- Measure the disparity of performance in downstream applications
- Examples are demographic parity, equality of opportunity, and predictive rate parity
- Also called fairness metrics

Intrinsic Metrics - WEAT

Based on the Implicit Association Test (IAT)

We are given a set of target words and two lists of characteristics (stereotypical and anti-stereotypical)

Are the targets’ representations closer to their stereotypes’?

Intrinsic Metrics - WEAT

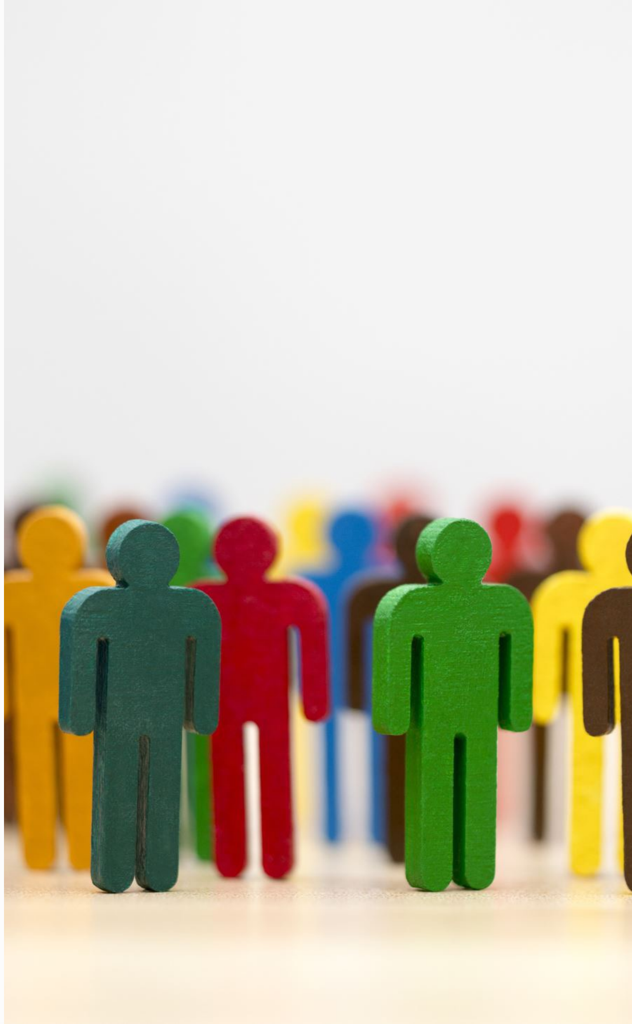
Test	Target Set #1	Target Set #2	Attribute Set #1	Attribute Set #2
T1	Flowers (e.g., *aster*, *tulip*)	Insects (e.g., *ant*, *flea*)	Pleasant (e.g., *health*, *love*)	Unpleasant (e.g., *abuse*)
T2	Instruments (e.g., *cello*, *guitar*)	Weapons (e.g., *gun*, *sword*)	Pleasant	Unpleasant
T3	Euro-American names (e.g., *Adam*)	Afro-American names (e.g., *Jamel*)	Pleasant (e.g., *caress*)	Unpleasant (e.g., *abuse*)
T4	Euro-American names (e.g., *Brad*)	Afro-American names (e.g., *Hakim*)	Pleasant	Unpleasant
T5	Euro-American names	Afro-American names	Pleasant (e.g., *joy*)	Unpleasant (e.g., *agony*)
T6	Male names (e.g., *John*)	Female names (e.g., *Lisa*)	Career (e.g. *management*)	Family (e.g., *children*)
T7	Math (e.g., *algebra*, *geometry*)	Arts (e.g., *poetry*, *dance*)	Male (e.g., *brother*, *son*)	Female (e.g., *woman*, *sister*)
T8	Science (e.g., *experiment*)	Arts	Male	Female
T9	Physical condition (e.g., *virus*)	Mental condition (e.g., *sad*)	Long-term (e.g., *always*)	Short-term (e.g., *occasional*)
T10	Older names (e.g., *Gertrude*)	Younger names (e.g., *Michelle*)	Pleasant	Unpleasant

Table 1: WEAT bias tests.

From "Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors" by Lauscher and Glavaš (2019) [Link]

Extrinsic Metrics – Group Fairness

- Unawareness
- Demographic Parity
- Equalized Odds
- Equality of Opportunity



Extrinsic Metrics – Equality of Opportunity

- Is only defined for binary classification
- Both classes have the same opportunity of being classified correctly
- It is measured as the difference between both classes' recalls

PART 3
Removing Biases

Fixing Our Models to Fix the World

Removing Biases

In general there are two philosophies

- Debiasing is reducing intrinsic bias metrics
  - Note that “debiasing” is a very loaded term!
- Fairness is reducing extrinsic bias metrics
  - We will not be focusing on these for this talk

Debiasing

- **Our assumption**: removing biases in language models removes biases in downstream applications

- Can either be attempted mathematically or by altering the data

Debiasing – Dataset Balancing

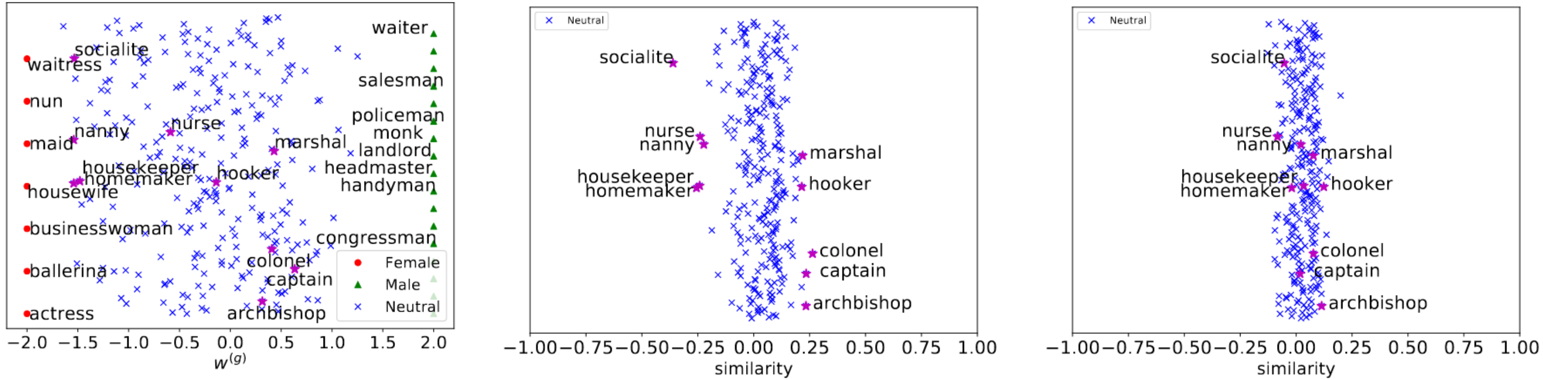
The task	Abusive comment detection on the Wikipedia comment section
The problem	All comments that contain the word “gay” are classified as abusive
The issue	All sentences in the dataset with the word “gay” use it as an insult
The solution	Add data where the word “gay” is used to represent a sexual orientation/preference

Debiasing – Removing the “Gender” Axis

- Identify the main dimension in which gender is represented in non-contextual embeddings

- Remove this dimension:
  - Completely removing it
  - Reducing the representation of non-gendered words

Debiasing – Removing the “Gender” Axis



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

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

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

Figure 1: Cosine similarity between the gender direction and the embeddings of gender-neutral words. In each figure, negative values represent a bias towards female, otherwise male.

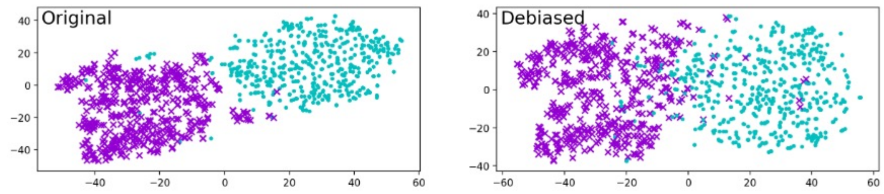
From “Learning Gender-Neutral Word Embeddings” by Zhao et al. (2018) [Link]

Does it Actually Work?

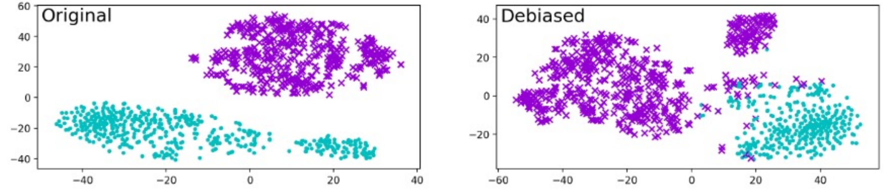
Does it Actually Work?

Not Always

Biases Might Stay Hidden



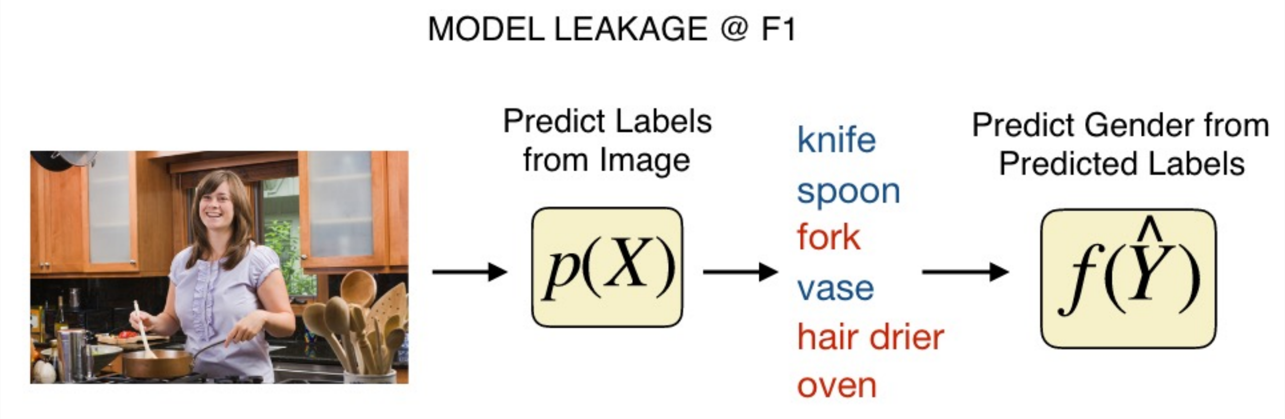
(a) Clustering for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.



(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

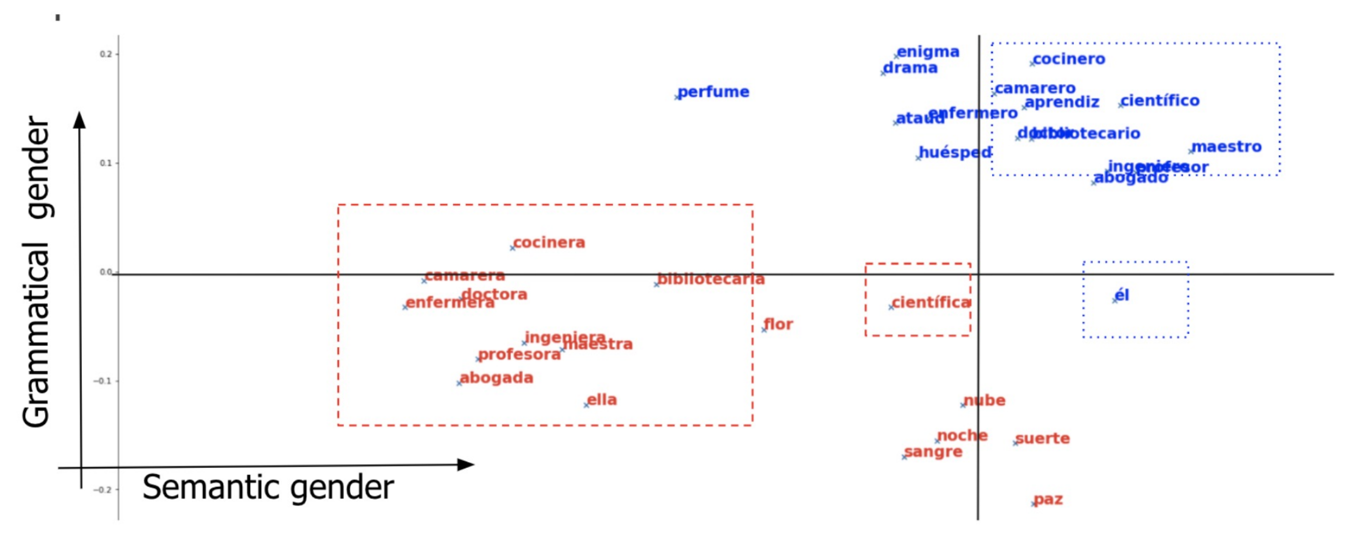
From “Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them” by Gonen and Goldberg (2019) [Link]

Biases are Complex



From “Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations” by Wang et al. (2019) [Link]

What About Other Languages?



From “Analyzing and Mitigating Gender Bias in Languages with Grammatical Gender and Bilingual Word Embeddings” by Zhou et al. (2019) [Link]

PART 4
But Does it Work?

Removing Biases in Downstream Applications

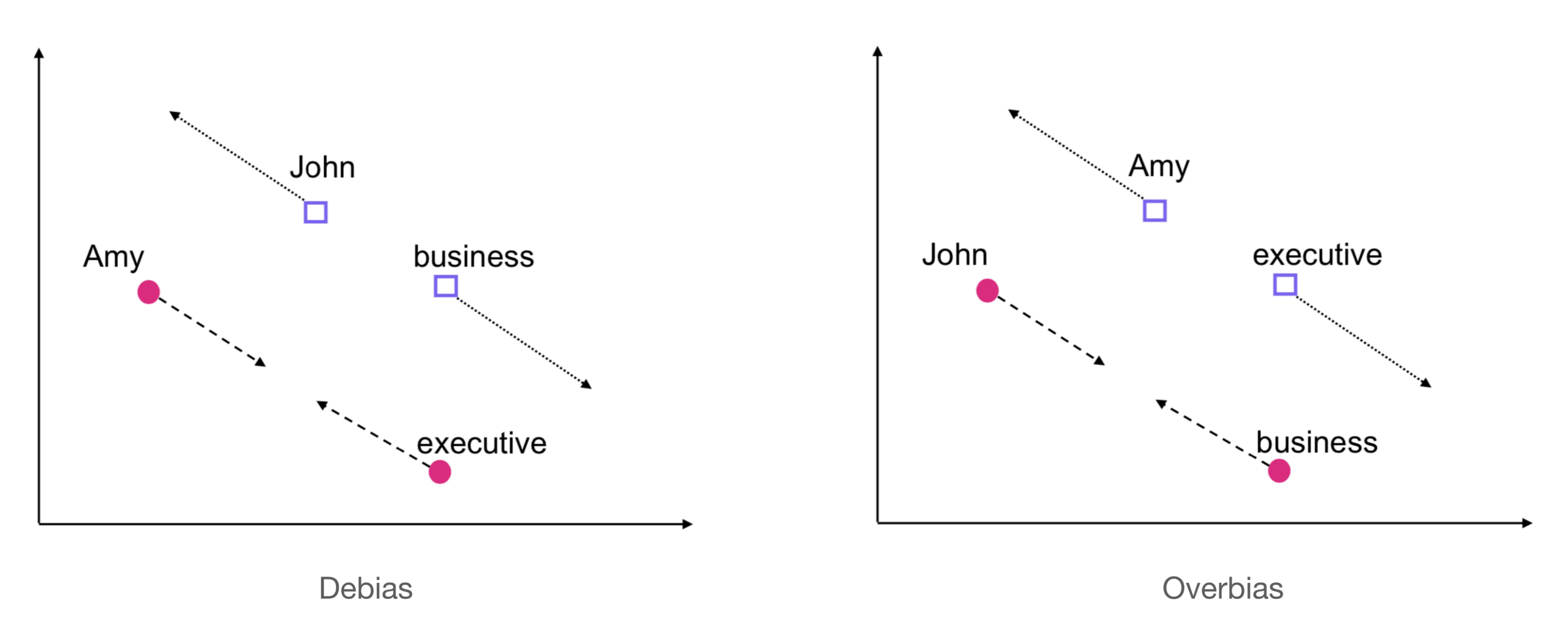
General Experimental Design

- Ways to Measure Bias
  - WEAT
  - Equality of Opportunity

- Two methods to reduce bias
  - Dataset balancing
  - Attract-Repel

- Two and a half downstream applications
  - Correference resolution in English
  - Hatespeech detection in English and in Spanish

Attract-Repel



About the Languages

English

- Has been used in most bias studies
- Only has semantic gender

Spanish

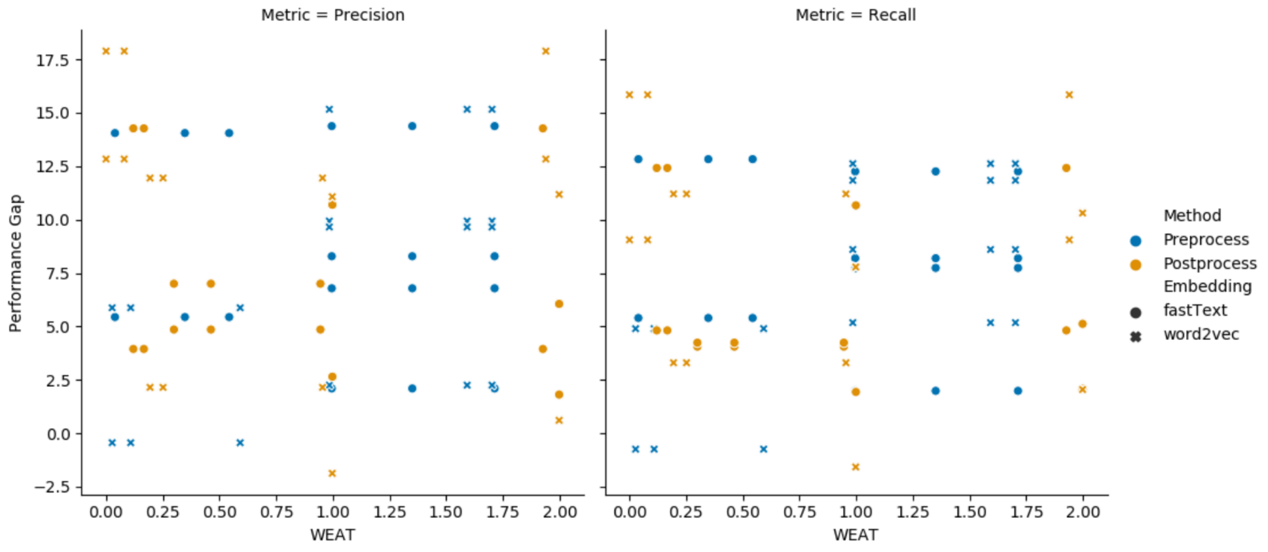
- Some studies have analysed biases in Spanish embedding spaces
- There is enough data available for most tasks
- Has both grammatical and semantic gender

Downstream Applications

- Correference resolution for gendered pronouns
  - A stereotypical task for bias assessment in English
  - However, it’s trivial in Spanish

- Hate speech classification
  - Allows us to compare the effects of semantic vs grammatical bias
  - Against women (English and Spanish)
  - Against migrants (Spanish)

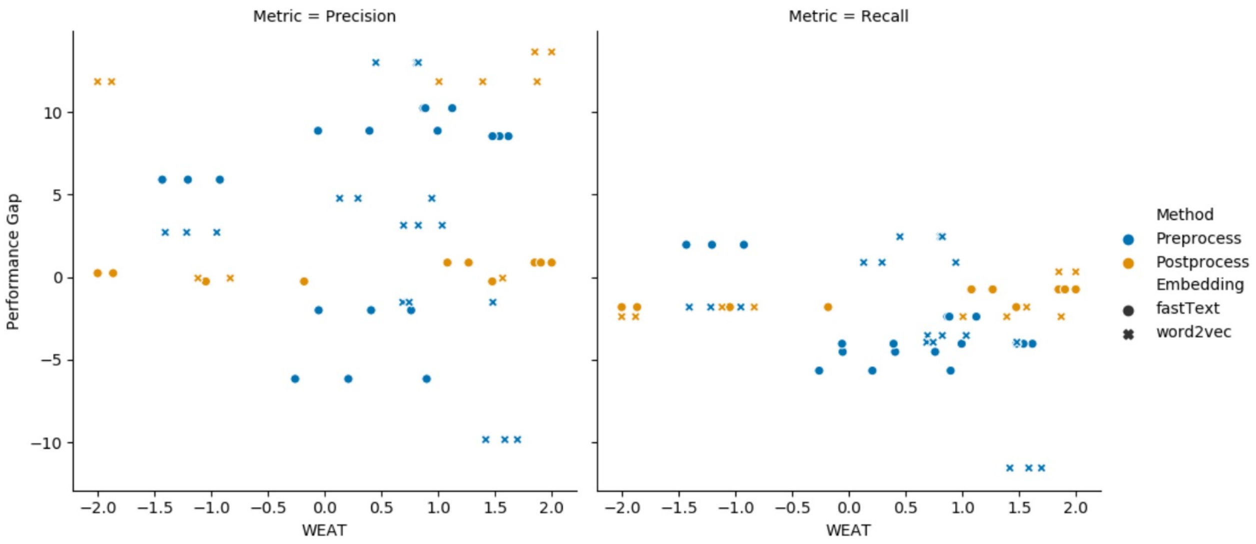
The Results – Correference (eng)



(a) Coreference (en) results broken out by bias modification method (pre- vs. post-processing).

From “Intrinsic Bias Metrics Do Not Correlate with Application Bias” by Goldfarb-Tarrant et al. (2021) [Link]

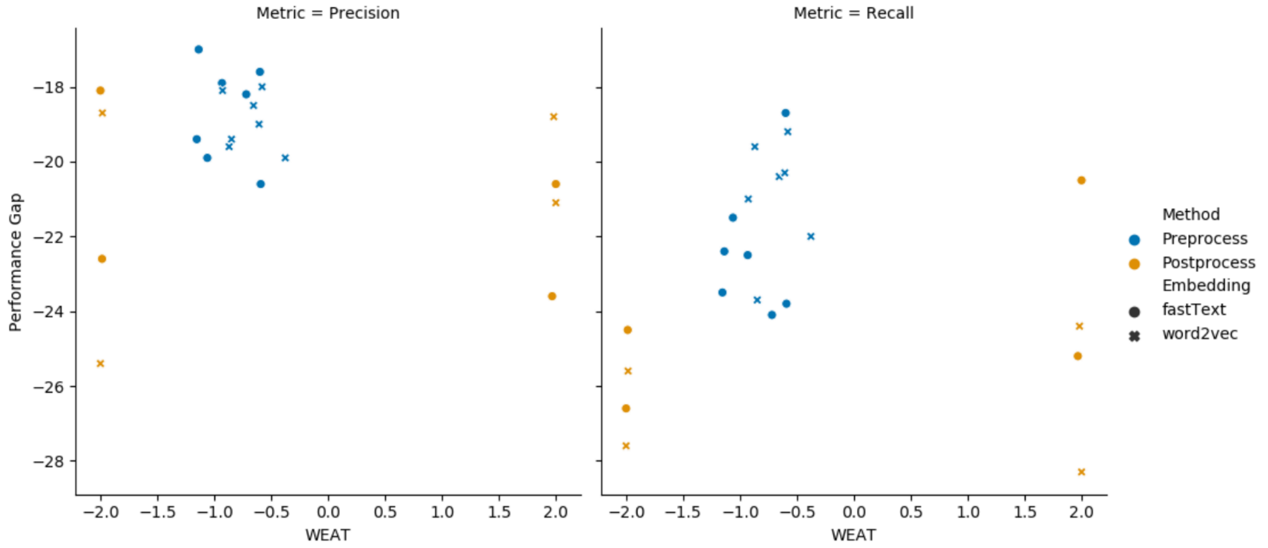
The Results – Hate Speech Detection (eng)



(b) Hatespeech (en) results broken out by bias modification method (pre- vs. post-processing).

From "Intrinsic Bias Metrics Do Not Correlate with Application Bias" by Goldfarb-Tarrant et al. (2021) [Link]

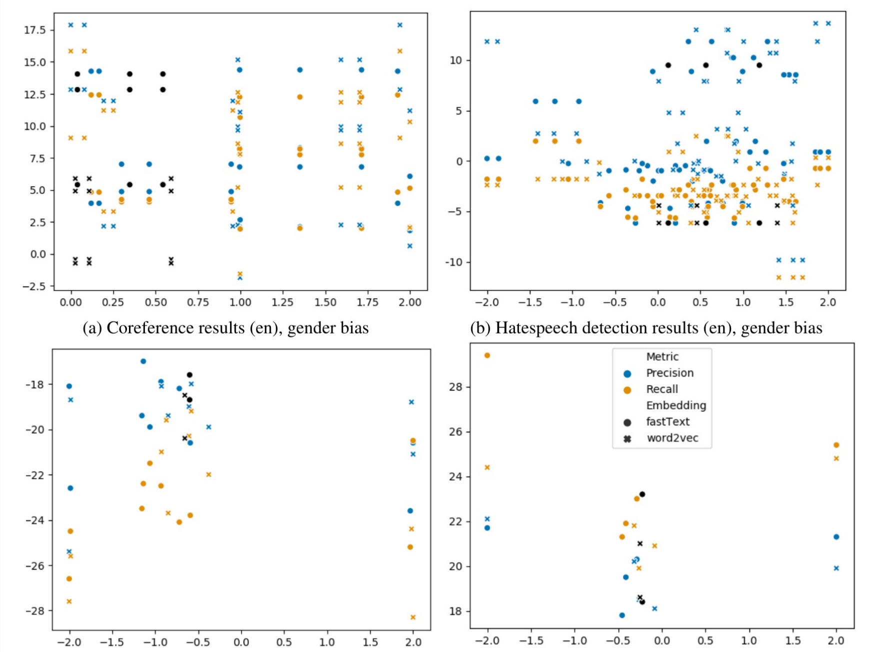
The Results – Hate Speech Detection (spa)



(c) Hatespeech (es) results for gender bias metrics broken out by bias modification method.

From “Intrinsic Bias Metrics Do Not Correlate with Application Bias” by Goldfarb-Tarrant et al. (2021) [Link]

The Results



(a) Coreference results (en), gender bias
(b) Hatespeech detection results (en), gender bias
(c) Hatespeech detection results (es), gender bias
(d) Hatespeech detection results (es), migrant bias

From “Intrinsic Bias Metrics Do Not Correlate with Application Bias” by Goldfarb-Tarrant et al. (2021) [Link]

Dorsa Amir
@DorsaAmir

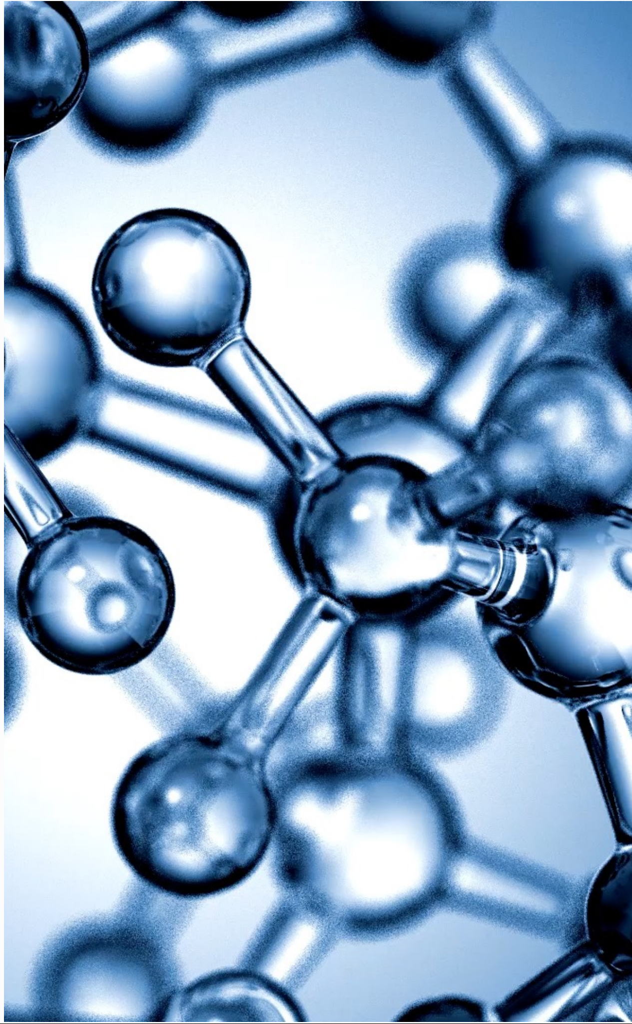
Follow

The (real) scientific method.

Our Insights

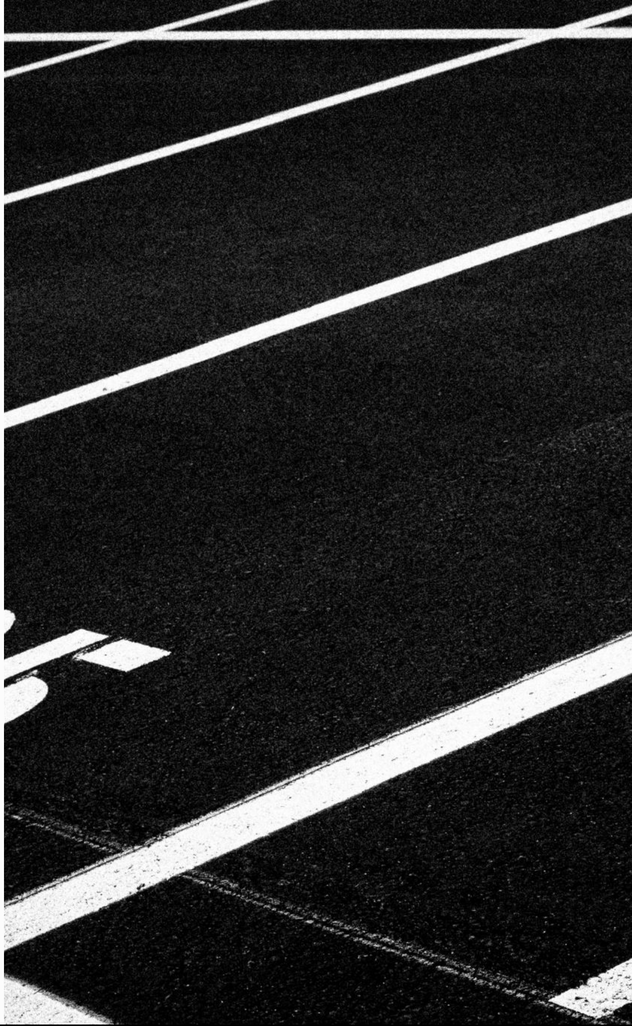
- Reducing bias in embedding spaces is unpredictable in terms of downstream application biases
  - It has been reproduced with more downstream applications
  - Language models also have these issues

- Spanish (X)WEAT is biased itself!
  - Almost all science words were grammatically male
  - Some issues with translations
  - No usual names from Spanish-speaking countries



Going Forward

- Most bias and fairness research focuses on
  - Gender as a binary (male/female)
  - Race in the United States as a binary (white/black)
- Biases are very diverse but the experiments ran more often than not aren’t
- Getting this research into the hands of those who need it is important



GÖTEBORGS UNIVERSITET

SPRÅKBANKEN TEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io