PS/HS Bias: Overview & Intro

Katja Markert

Uni Heidelberg

WS 2020

Allgemeines Gleichbehandlungsgesetz Paragraph 1

Ziel des Gesetzes

Ziel des Gesetzes ist, Benachteiligungen aus Gründen der Rasse oder wegen der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität zu verhindern oder zu beseitigen.

Discrimination

Illegitimate/illegal treatment differences of individuals or groups on the basis of one or more criteria.

Protected attributes

legitimate vs. nonlegitimate treatment differences

intended vs. unintended discrimination

Method or process of discrimination open

Further discussion of definitions next week and throughout the course

Algorithmische Entscheidungen Beispiele

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions): assesses the likelihood of a defendent becoming a recidivist
- Targeting patients for high-risk care management programs
- Automatic employment decisions
- Word Embeddings: Decide what occupations are similar to men and which to women
- Language identification algorithms decide which twitter comments are in English and therefore displayed in a search

System Bias: many definitions

Bias Definition I

Inconsistent behaviour of a system towards input from different demographic groups

(adapted from Hardt et al 2016

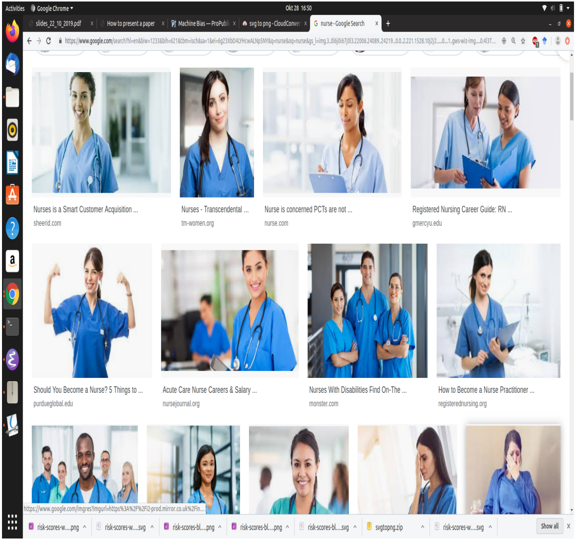
Bias Definition II

Model is biased if it learns inappropriate stereotypical correlations of concepts

Both definitions are relevant for us!

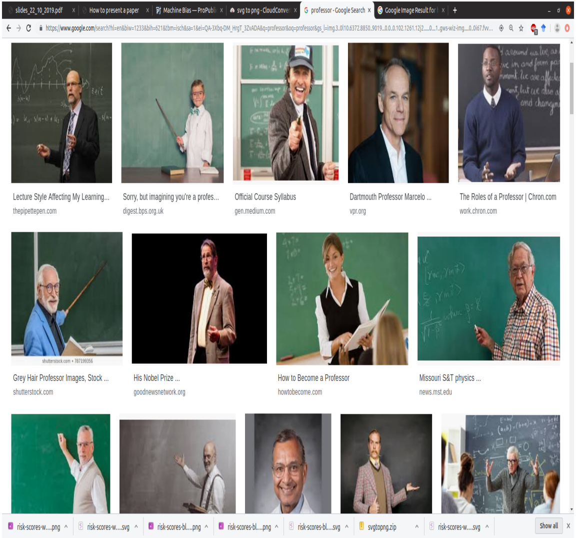
Example case of “stereotype” bias I

Google image search for *nurses*:



Example case of "stereotype" bias I

Google image search for *professors*:



Example case of “stereotype” bias II

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Aus Bolukbasi et al (2016)

Why does it matter?

Embeddings used in almost all current systems as building blocks. Examples:

- Coreference resolution: *Donald Trump* . . . *Hilary Clinton* . . . *the president*.
- Text classification: Present text via word embeddings instead of words → topic classification, sentiment classification . . .

Example cases of ML behaviour bias: COMPAS

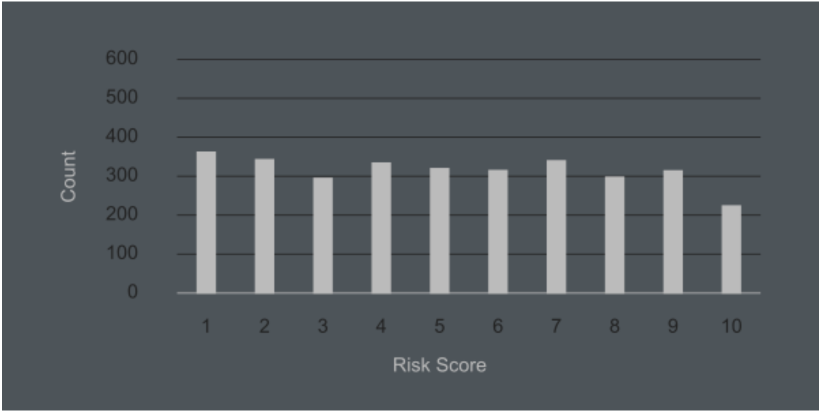
COMPAS: To automatically assess risk of recidivism (used for bail, sometimes sentenceing etc)

Two shoplifting arrests:

This and follow-on graphics on next two slides are from ProPublica Report on COMPAS. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

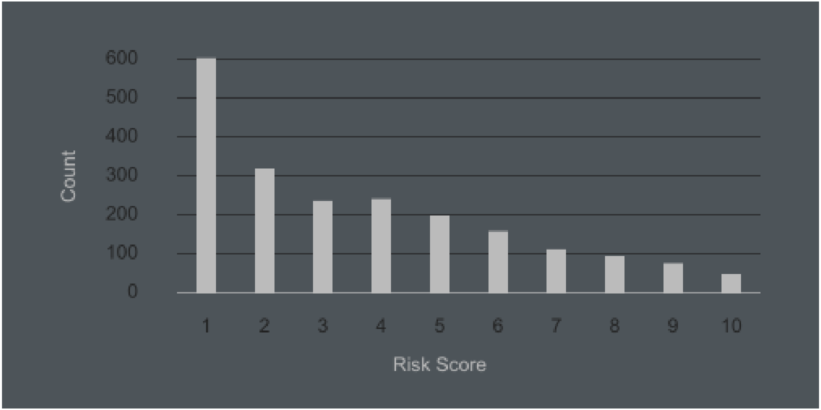
Example case of ML behaviour bias: COMPAS

Risk scores of blacks:



Example case of ML behaviour bias: COMPAS

Risk scores of whites:



Example case of ML behaviour bias: COMPAS

	White	African American
Labeled higher risk, but did not re-offend	23.5%	44.9%
labeled lower risk, yet did re-offend	47.7%	28%

How did this particular bias come about?

Rsik score depends on 137 factors, including

- Arrests of parents
- Arrests of friends
- Do you have a job?
- Direct value for race was not included.

Impact of AI/ML/NLP

White House Podesta Report 2014

big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.

What is this course not about?

- Sentiment recognition: Detection of biased (positive or negative) opinions explicitly expressed in text
- Learning biases of specific algorithms, such as Occam's razor

Overview

Topics

- Part I: Sessions on Bias in Embeddings
- Part II: Sessions on biased (and unbiased) Corpora
- Part III: Sessions on algorithmic bias in supervised ML classification
- Interspersed: Applications and Case Studies
  - Coreference
  - Bias in Dialect Processing
  - MT
  - Hate Speech Classification
  - Visual Semantic Role Labeling and Image Retrieval

In the course

We will learn

- how one can define and operationalise bias

In the course

We will learn

- how one can define and operationalise bias
- how to measure bias

In the course

We will learn

- how one can define and operationalise bias
- how to measure bias
- how to mitigate bias

In the course

We will learn

- how one can define and operationalise bias
- how to measure bias
- how to mitigate bias

We will look at

- "stereotype" bias in word embeddings

In the course

We will learn

- how one can define and operationalise bias
- how to measure bias
- how to mitigate bias

We will look at

- “stereotype” bias in word embeddings
- algorithmic bias in supervised ML classification

In the course

We will learn

- how one can define and operationalise bias
- how to measure bias
- how to mitigate bias

We will look at

- "stereotype" bias in word embeddings
- algorithmic bias in supervised ML classification
- impact on applications

In the course

We will learn

- how one can define and operationalise bias
- how to measure bias
- how to mitigate bias

We will look at

- "stereotype" bias in word embeddings
- algorithmic bias in supervised ML classification
- impact on applications
- bias for different groups with emphasis on gender and race

Definitions of bias and causes of bias

- direct or indirect discrimination?
- explainable and unexplainable discrimination?
- statistical discrimination
- historical bias: reflecting reality? (see nurse/professor image search)
- representation bias: sample most images from western countries
- ...

Operationalisations of fairness

When is a machine classifier fair?

Demographic/Statistical Parity (equal positive rates)

$$p(\tilde{y} = 1 | A = 0) = p(\tilde{y} = 1 | A = 1)$$

vs.

Equal opportunity (equal true positive rates)

$$p(\tilde{y} = 1 | A = 0, y = 1) = p(\tilde{y} = 1 | A = 1, y = 1)$$

vs.

Fairness through unawareness

An algorithm is fair as long as any protected attributes A are not explicitly used in decision-making process.

Topic I.1: Measuring word embeddings bias

Caliskan et al (2017): Semantics derived automatically from language corpora contain human-like biases. *Science 2017*

- African-American names (*Leroy, Shaniqua*) had a higher similarity with unpleasant words (*abuse, stink, ugly*)
- European American names (*Brad, Greg, Courtney*) had a higher cosine with pleasant words (*love, peace, miracle)*

Uses psychological association tests (WEAT) to measure bias in word embeddings outcomes

Topic I.1: Measuring word embeddings bias

Garg et al (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of sciences*

Changes in embeddings track demographic and occupational shifts over time.

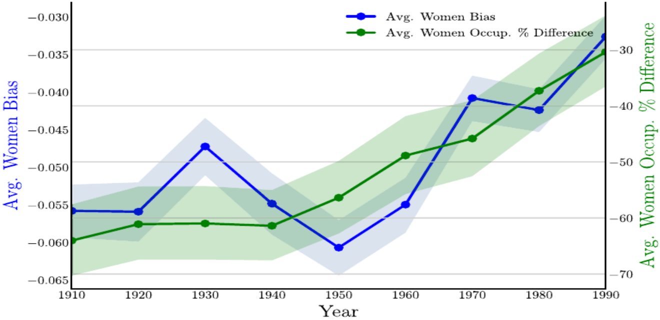


Fig. 2. Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.

Topic I.2: Mitigating word embeddings bias via algorithm change

Main Idea: gender subspace hypothesis

There exists a linear subspace $B \subset R^d$ that contains (most of) the gender bias in the space of word embeddings.

Topic I.2: Mitigating WE bias via algorithm change

Bolukbasi et al (2016): Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proc of NIPS*

- Postprocessing approach
- Identify gender-subspace B (single dimension) using linear algebra methods and gender-definitional pairs (*he-she*)
- Represent vector of a word as $v = v_B + v_{\perp B}$
- Neutralise: Remove gender bias from not-explicitly gendered words (found in separate classifier)
- Equalise: Make pairs of explicitly gendered words *mother-father* equidistant to all not explicitly gendered words

Extension by Manzini et al (2019) to more than two classes

Topic I.2: Mitigating WE bias via algorithm change

Zhao et al (2018): Learning Gender Neutral word embeddings

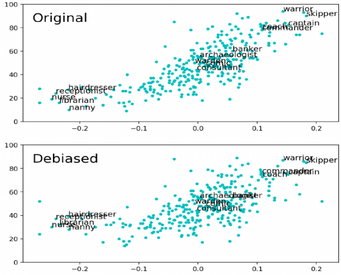
- in-processing: GN-GLOVE
- represents protected attributes in certain dimensions
- Enhances Glove optimization objective to restrict gender information to certain dimensions

Follow on paper Zhao et al (2019) looks at bias in contextual word embeddings (Elmo)

Topic I.3: Do these mitigation techniques really work?

Gonen and Goldberg (2019): Lipstick on a pig: Debiasing methods cover up systematic gender bias in word embeddings but do not remove them.

- Are we really non-biased if each non-explicitly gendered word is in equal distance to both elements of all explicitly gendered pairs?
- The structure/similarities between non-explicitly gendered words remains!



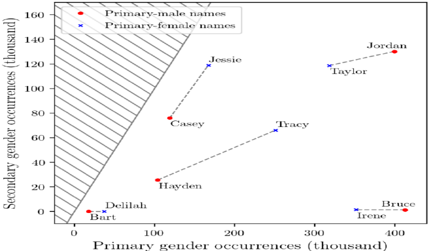
(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.

Topic I.3: Mitigating WE bias via data modification

Maudslay et al (2019): It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution.

Change data instead of algorithm

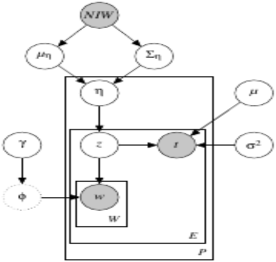
- *The woman cleaned the kitchen* → *The man cleaned the kitchen* (with a certain probability)
- *Elizabeth ... She cleaned the kitchen* ???
- Leaving overall individual word distributions the same while treating many biased associations



Topic II.1: Selection bias: Bias in Wikipedia

A wide variety of papers looks at bias in encyclopedias and knowledge bases.

- Wagner et al (2015,2016) look at different linguistic and topical as well as network positions for men and women
- Bamman and Smith (2014) discover abstract event classes in biographies, based on a probabilistic latent/variable model. Find bias in women's charcterization per event



(a) FULL.

Topic II.2: Unbiased evaluation corpora

Example: Webster et al (Tacl 2018) present GAP a balanced corpus of gender-ambiguous pronouns

Type	Pattern	Example
FINALPRO	(Name, Name, Pronoun)	*Preckwinkle* criticizes *Berrios'* nepotism: [...] County's ethics rules don't apply to **him**.
MEDIALPRO	(Name, Pronoun, Name)	*McFerran*'s horse farm was named Glen View. After **his** death in 1885, *John E. Green* acquired the farm.
INITIALPRO	(Pronoun, Name, Name)	Judging that **he** is suitable to join the team, *Butcher* injects *Hughie* with a specially formulated mix.

Table 1: Extraction patterns and example contexts for each.

- Performance of existing coreference tools such as Lee et al (2017): 67.2 on male pronouns, 62.2 on female pronouns
- Proposes the application of transformer models for the task

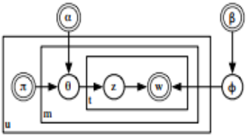
Topic III Example: Bias in dialect processing

Racial Disparity in language identification in Blodgett et al (2016)

	AAE	White-Aligned
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Table 3: Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. (§4.1)

- Corpus collection: Distant supervision of US Census Data combined with language model



$$\theta_m \sim Dir(\alpha\pi_u), \ \ \phi \sim Dir(\beta/V)$$
$$z_t \sim \theta_m, \ \ w_z \sim \phi_{z_t}$$

- Ensemble classifier for language identification

Topic IV.1: Classic and seminal papers in ML classification

- Range from simple less biased NB classifiers to sophisticated ML models (see literature list)
- Most papers in machine learning and NIPS conferences
- Relatively mathematical

Topic IV.1 Example: Hardt et al 2016

- Proposes the *equality of odds* fairness criterion
- Special case:

Equal opportunity (equal true positive rates)

$$p(\tilde{y} = 1 | A = 0, y = 1) = p(\tilde{y} = 1 | A = 1, y = 1)$$

- How to optimally adjust any learned predictor so as to reove discrimination according to definition
- Uses linear programming to derive this predictor

Topic IV.2: Case Studies II - Bias in textclassification

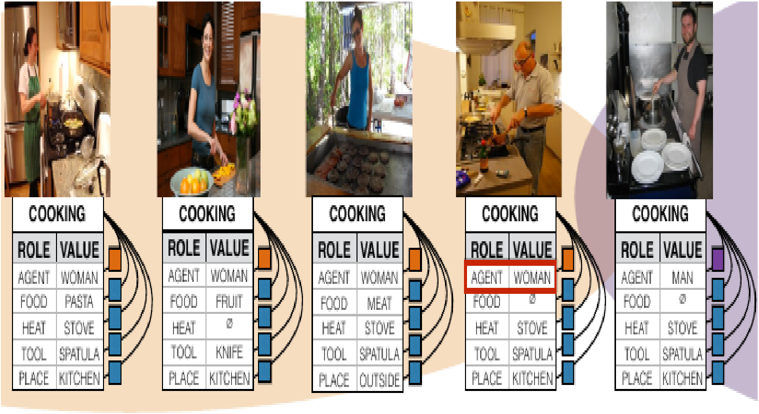
For NLP, text classification one of the most important supervised classification tasks.

Kiritchenko and Mohammad (2018): Examining gender and race bias in two hundred sentiment analysis systems

- *The conversation with my dad/mom was heartbreaking*
- *The conversation with Ebony/Amanda was heartbreaking.*
- Most sentiment systems show higher scores for sadness when used with females, higher scores with feat when used with males
- Higher scores for anger,fear,sadness for African American names. Higher score for joy and positive affect for European American names

Topic VI.1 Bias in vision

Zhao et al (2017): Men also like shopping: Reducing gender bias amplification using corpus-level constraints



Calibrate a structured prediction model to avoid amplifying bias