School of CIT –
Social Computing
Research Group

Technical
University
of Munich

Advanced Natural Language Processing
CIT4230002

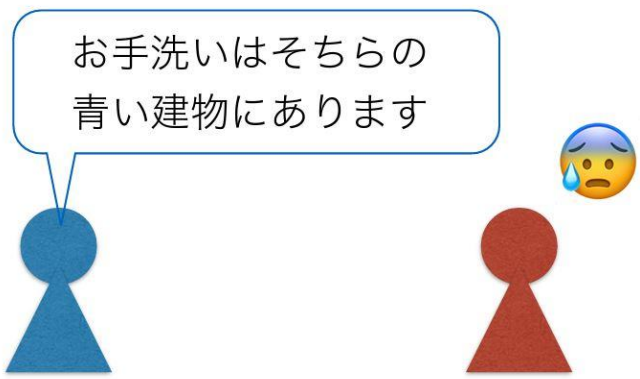
Prof. Dr. Georg Groh
Tobias Eder, M.A. M.Sc.

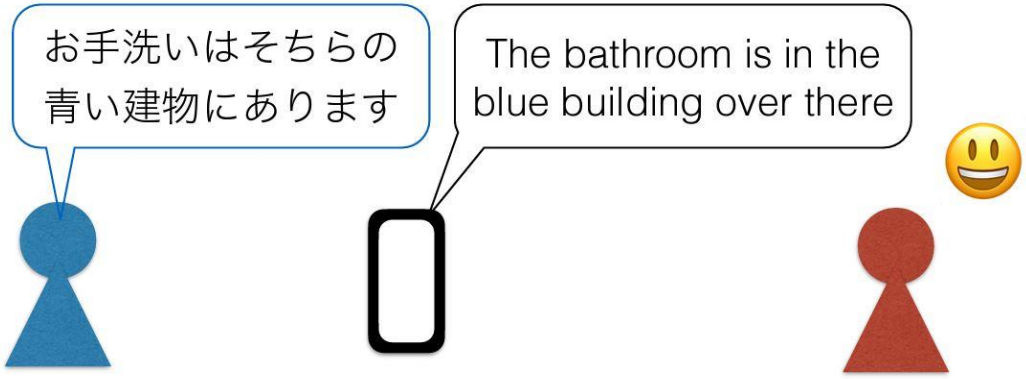
Lecture 7
Bias and Ethical Issues

Outline

- **Fairness and ethics**
- Detecting bias
- Mitigating bias
- NLP, ethics and impact

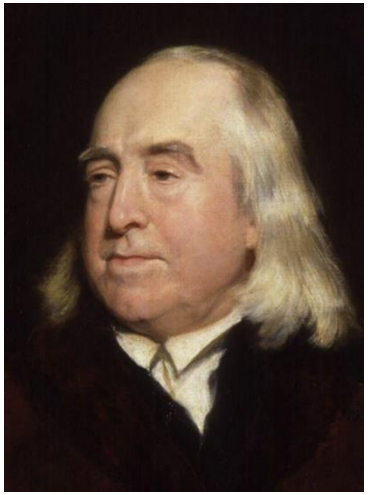
Fairness and ethics | Why we build NLP systems





- Technical systems aimed to help people perform a task better
- “Make the world a better place”
- How do we quantify better?

- "The greatest happiness of the greatest number is the foundation of morals and legislation"



- Maximizing utility seems to be a promising approach
- Follow-up questions:
  - Is it okay to decrease someone's utility for an overall gain in utility?
  - Is it okay if only some have a gain in utility?
- Utility alone is not the only value to consider

Fairness and ethics | Causing Harm

- “Don’t be evil”

- Decisions made by systems might have positive or negative impact
- For most cases, the negatives might be unintended
- This disparate impact can often be an issue of fairness and discrimination
- In general, we are talking about biases when systems show unintended disparaging outcomes

Fairness and ethics | Assigning Responsibility

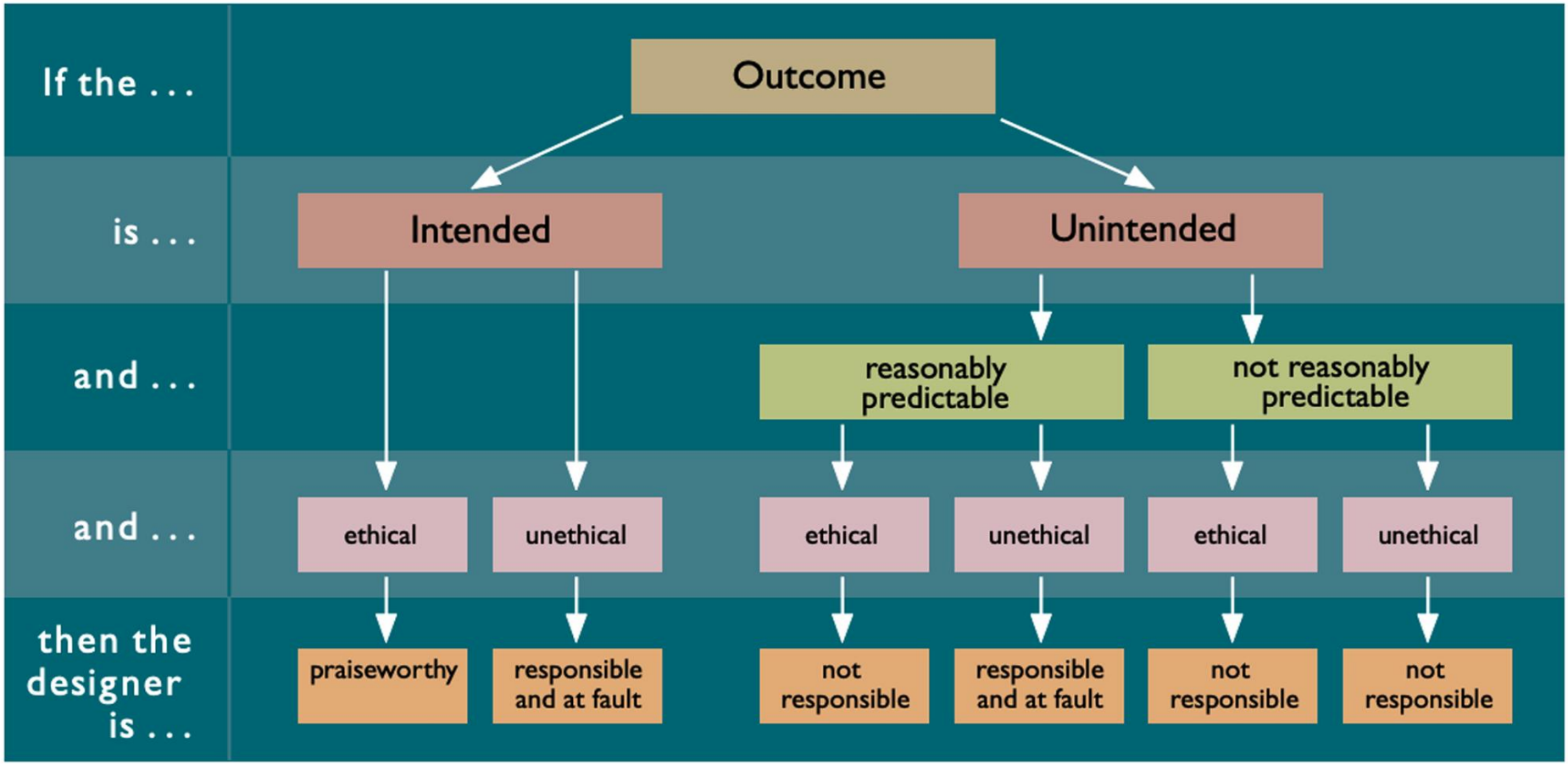


Figure 5. Flow chart clarifying the levels of ethical responsibility associated with predictable and unpredictable intended and unintended consequences.

- Berdichevsky and Neuenschwander (1999) on responsibility in tech

Fairness and ethics | Types of bias

- There are various ways in which bias can enter a system
  - Data collection
  - Annotation
  - Algorithmic choice
  - Training objective
  - Feedback loops in deployment

- Often it is not a question of whether a system will exhibit bias but under what circumstances

Outline

- Fairness and ethics
- **Detecting bias**
- Mitigating bias
- NLP, ethics and impact

Detecting bias in data

- Some bias might be inherent to the data
- Also called the "biased world" issue

Word embeddings quantify 100 years of gender and ethnic stereotypes

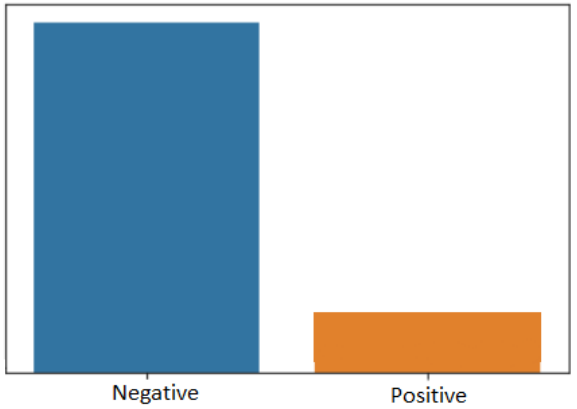
Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

- Types of bias in data:
  - Co-occurrence bias
  - Framing bias
  - Epistemological bias

Detecting bias in data | Datasets and annotation

- Other problems can stem from the way data was collected or labeled
- In the trivial case this comes down to a class imbalance



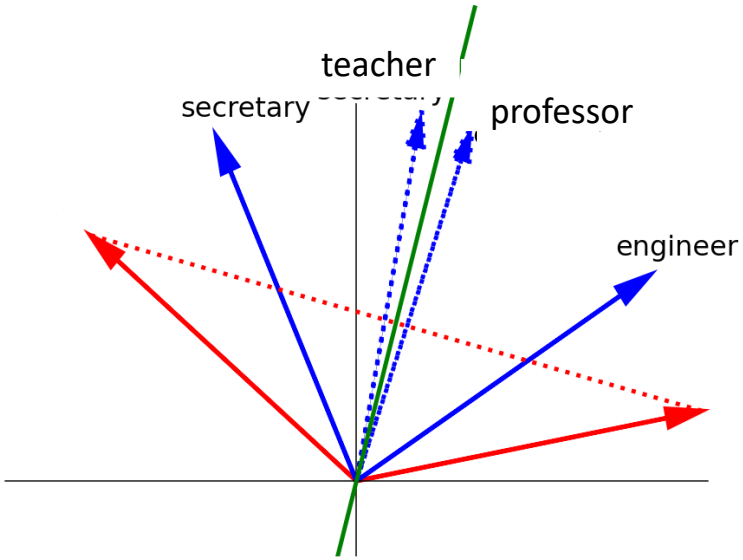
- Rarely is it obvious where bias exists
  → We need to know what we are trying to measure
- This means defining protected attributes to check against
- Critically examining where your data comes from already helps

Detecting bias in systems

- Detecting bias through a system can be easier than trying to sift through the data
- Embeddings and representation learning gives us the opportunity to compare across different dimensions
  → Association tests
- The output is also the natural step to evaluate for bias in the first place
  → Error rate analysis
- Purposefully constructing test cases to detect differences
  → Counterfactual evaluation

Detecting bias in systems | Association tests

- Representations are themselves learned from co-occurrence



- Base assumption:
  - Two sets of target words (e.g., professor, engineer, …, teacher, secretary, …)
  - Two sets of attribute words (e.g., man, male, …, woman, female, …)
  - Null hypothesis: No difference between the target words with respect to the different attribute sets

Detecting bias in systems | Association tests

- We define our measurement of bias s with inputs X, Y the target words and A, B the attribute words:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad \text{where}$$

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

- s close to 0: Null hypothesis wins
- s tending negative: targets X more strongly correlated with attributes A
- s tending positive: targets Y more strongly correlated with attributes B

Detecting bias in systems | Association tests

- Example application on simple word embeddings (Caliskan et al., 2017)
- The resulting embeddings naturally mirror the distribution in the data

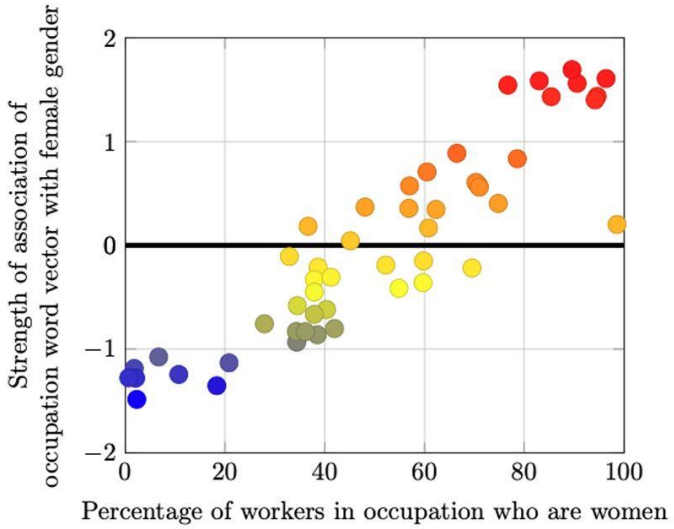


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

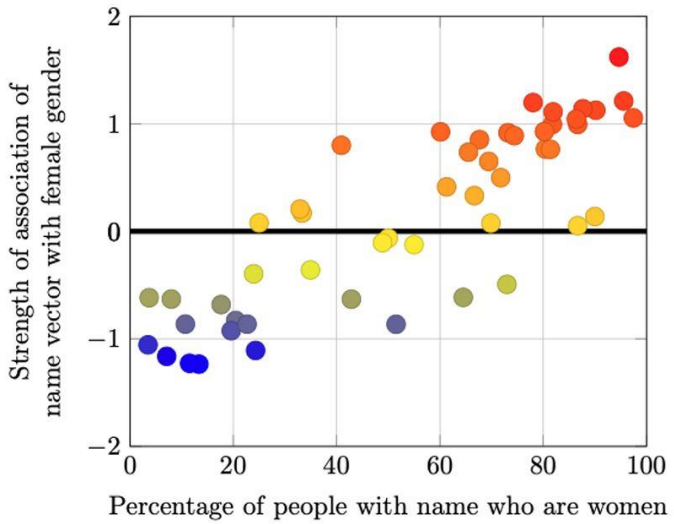


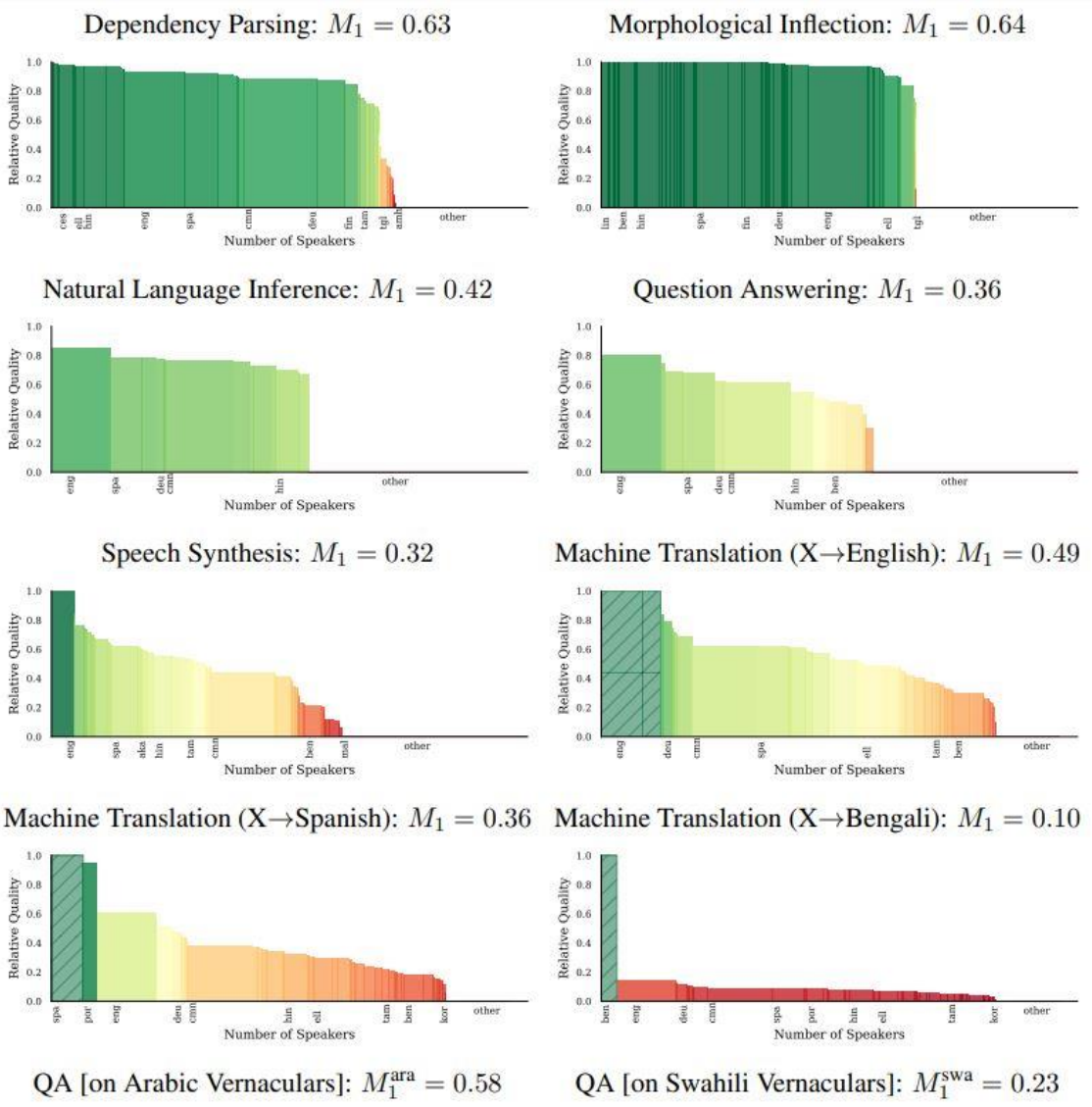
Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Detecting bias in systems | Error analysis

- A lot of the time we can measure bias directly from the performance on test data
- This is largely model agnostic but requires additional information
  - Split your data into subgroups
  - Evaluate on each of them separately
- Highly variant performances are indicative of impact disparity

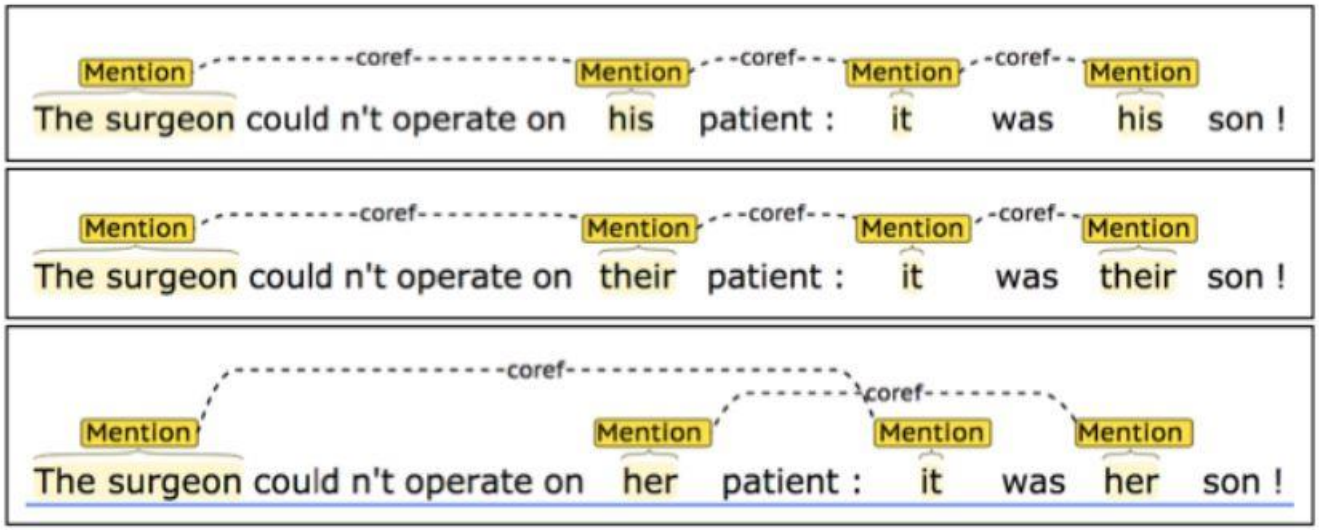
Detecting bias in systems | Error analysis

- An obvious example:
- Meta-analysis of task performance on NLP tasks for state-of-the-art systems in different Languages (Blast et al. 2021)



Detecting bias in systems | Counterfactual evaluation

- Manipulating your test cases to change part of the input without changing the target label
- E.g., Gender Bias in Coreference Resolution (Rudinger et al. 2018)
- Minimal sentence pairs that only differ by pronouns used.



Outline

- Fairness and ethics
- Detecting bias
- **Mitigating bias**
- NLP, ethics and impact

Mitigating bias | Embedding debiasing

- Biases mean vector differences across unwanted dimensions
- Idea: Identify a direction that captures a specific bias
- Then either:
  - Neutralize – Zero out the subspace
  - Equalize – Establish equidistance to an equality set
  - Soften – Shift while maintaining similarity
- Experiments by Bolukbasi et al. 2016 on hard and soft debiasing:

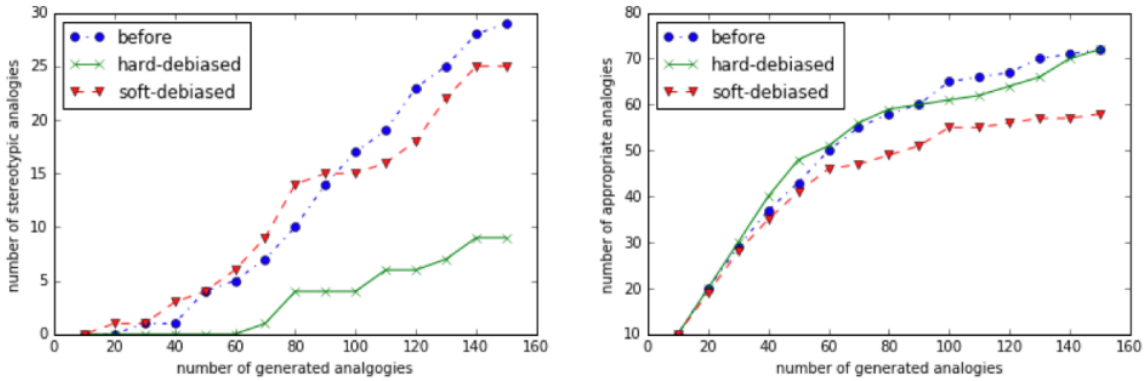
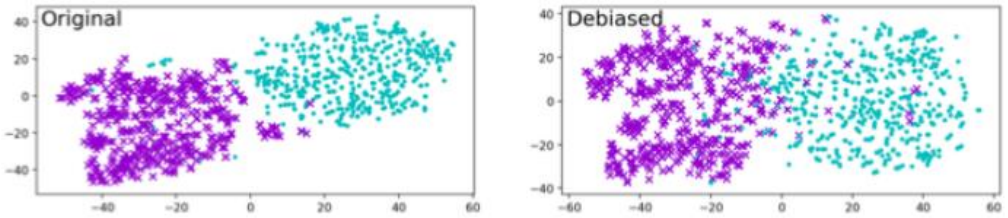


Figure 8: Number of stereotypical (Left) and appropriate (Right) analogies generated by wordembeddings before and after debiasing.

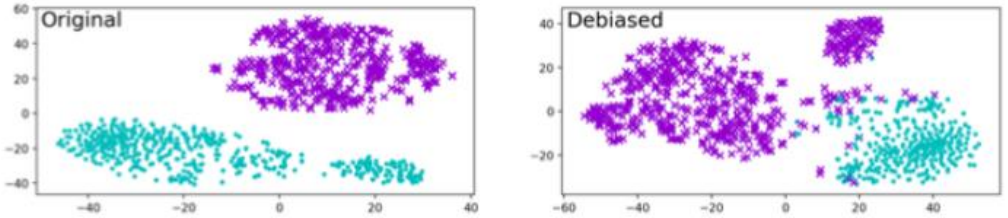
Mitigating bias | Embedding debiasing

- Problem: We seem to cover up biases without really making them disappear (Gonen and Goldberg 2019)

- Example of gender: Male and female biases still cluster after debiasing
- Gender is still encoded although no longer explicitly



(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.

Mitigating bias | Feature invariant methods

- Another idea: We change the training objective
- Explicitly penalize being able to identify protected variables (Zemel et al. 2013)

$$L = \sum_k CrossEntropy(y^{(k)}, \hat{y}^{(k)}) + \alpha \sum_k \left| x^{(k)} - \hat{x}^{(k)} \right| + \beta \left| \frac{1}{|X+|} \sum_{X+} z_i^{(k)} - \frac{1}{|X-|} \sum_{X-} z_i^{(k)} \right|$$

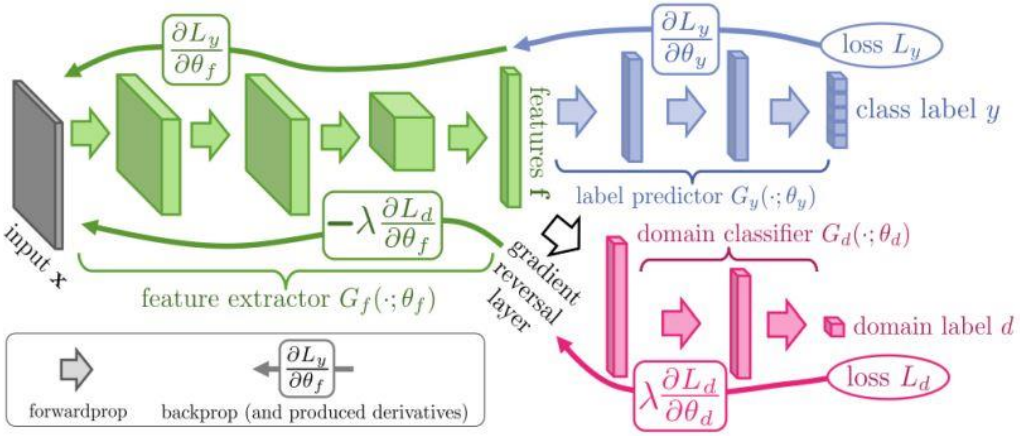
Classifications should be good

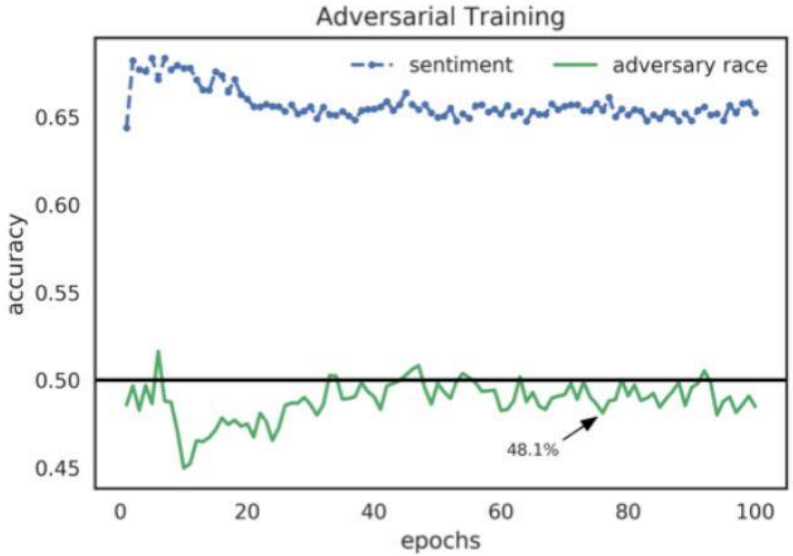
Reconstructions should be good

Intermediate Representations should be indistinguishable across values of the protected variable

Mitigating bias | Feature invariant methods

- In a practical example by Ganin and Lempitsky 2015





Mitigating bias | Data augmentation

- How about we tackle the problem on the input side?
- Idea: Alter text to invert specific biases and combine with original data
- Counterfactual data augmentation (Lu et al. 2018)
- Albeit simple this can have at least some effects at softening the bias

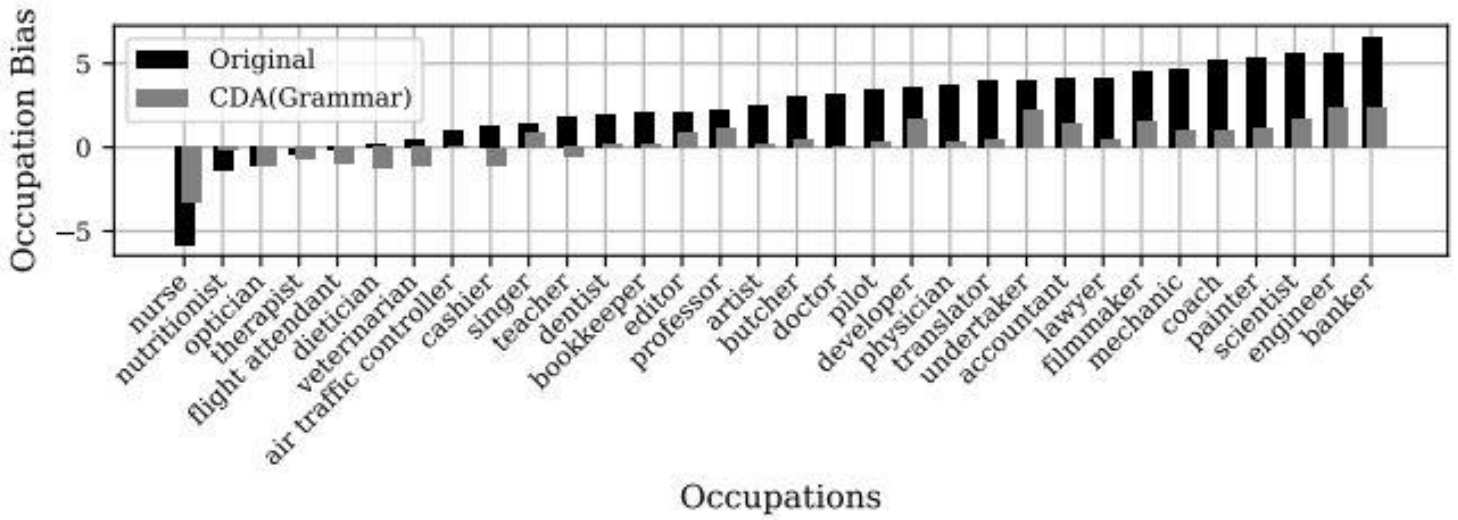


Figure 2: Model 1.1 & 1.2: Bias for Occupations in Original & CDA Model

Mitigating bias | Data augmentation

- What about doing it manually?
- Human in the loop approaches to generating more balanced data (Kaushik et al. 2021)
- Advantages:
  - More robust in different domains
  - Less generic patterns that come with rule-based approaches
- Huge disadvantage:
  - Cost and time

Outline

- Fairness and ethics
- Detecting bias
- Mitigating bias
- **NLP, ethics and impact**

NLP, ethics and impact | Bias research problems

- Conceptual problems with solving the issue
- Survey by Blodgett et al. 2020:
- Research in the field is often:
  - Vague as to who is harmed and why
  - Inconsistent definitions of bias
  - Mismatch between constituted problem and proposed solutions
  - Almost no engagement with literature from outside core NLP

NLP, ethics and impact | Bias research problems

NLP task	Stated motivation	Categories: Motivations	Categories: Techniques
Language modeling (Bordia and Bowman, 2019)	*"Existing biases in data can be amplified by models and the resulting output consumed by the public* ***can influence them***, ***encourage and reinforce harmful stereotypes***, *or* ***distort the truth***. *Automated systems that depend on these models can* ***take problematic actions*** *based on biased profiling of individuals."*	Allocational harms, stereotyping	Questionable correlations
Sentiment analysis (Kiritchenko and Mohammad, 2018)	*"Other biases can be inappropriate and result in* ***negative experiences*** *for some groups of people. Examples include,* ***loan eligibility and crime recidivism prediction systems***...*and* ***resumé sorting systems*** *that believe that men are more qualified to be programmers than women (Bolukbasi et al., 2016). Similarly, sentiment and emotion analysis systems can also perpetuate and accentuate inappropriate human biases, e.g., systems that consider* ***utterances from one race or gender to be less positive*** *simply because of their race or gender, or customer support systems that* ***prioritize a call from an angry male over a call from the equally angry female***."	Allocational harms, other representational harms (system performance differences w.r.t. text written *by* different social groups)	Questionable correlations (differences in sentiment intensity scores w.r.t. text *about* different social groups)
Machine translation (Cho et al., 2019)	*"[MT training] may incur an association of gender-specified pronouns (in the target) and gender-neutral ones (in the source) for lexicon pairs that frequently collocate in the corpora. We claim that this kind of phenomenon seriously threatens the fairness of a translation system, in the sense that it* ***lacks generality*** *and* ***inserts social bias*** *to the inference. Moreover, the input is* ***not fully correct*** *(considering gender-neutrality) and* ***might offend the users*** *who expect fairer representations."*	Questionable correlations, other representational harms	Questionable correlations
Machine translation (Stanovsky et al., 2019)	*"Learned models exhibit social bias when their training data* ***encode stereotypes not relevant for the task, but the correlations are picked up anyway***."	Stereotyping, questionable correlations	Stereotyping, other representational harms (system performance differences), questionable correlations

Final Takeaways

- NLP and research in general should have clearly formulated goals and users
- In designing systems beware of disparaging outcomes and the dual-use problem
- Biases and unfair treatment in existing systems are
  - Easy to detect …
  - … but very hard to eliminate
- However: This does not mean we should stop building systems
- But while building them we should be clear about
  - Goals and supported use-cases
  - Limitations and blind-spots

Study Approach

Minimal

- Work with the slides

Standard

- Minimal approach + read reference 3

In-Depth

- = standard approach + look at references 5 and 9

See you next time!

Resources

(1) Graham Neubig: CMU Advanced NLP 2022

(2) Berdichevsky and Neuenschwander: Towards an Ethics of Persuasive Technology 1999

(3) Garg et al.: Word Embeddings Quantify 100 years of Gender and Ethnic Stereotypes 2017

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

(5) Blasi et al.: Systematic Inequalities in Language Technology Performance across the World's Languages 2021

(6) Rudinger et al.: Gender Bias in Coreference Resolution 2018

(7) Bolukbasi et al: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings 2016

(8) Hila Gonen and Yoav Goldberg: Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them 2019

(9) Blodgett et al: Language (Technology) is Power: A Critical Survey of “Bias” in NLP 2020