

CS11-747 Neural Networks for NLP

Bias in NLP models

Divyansh Kaushik



Carnegie Mellon University

Language Technologies Institute

Site

<https://phontron.com/class/nn4nlp2021/>

Bias/Artifacts/Spurious Associations

- ML systems exploit mutual information b/w features and labels to make predictions.
- Growing concern that models rely on the wrong features: ***artifacts, bias, superficial/spurious associations.***
- However, these terms hold no formal meaning in standard ML framework.

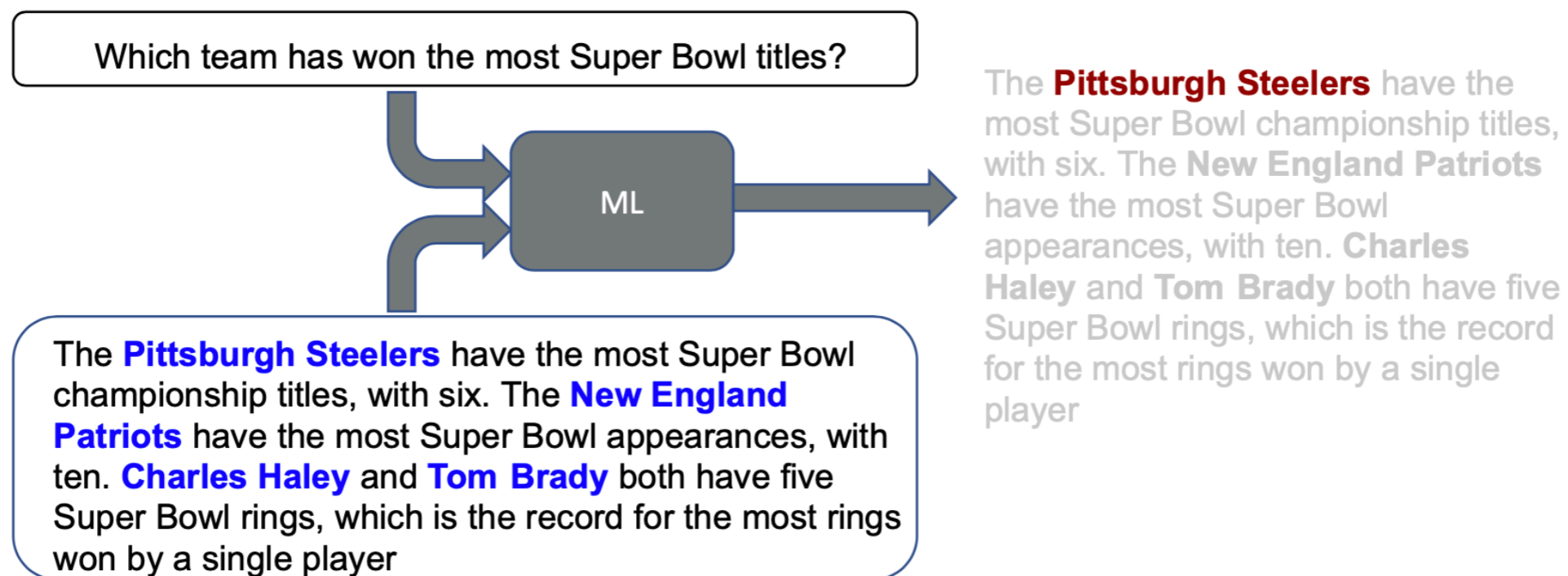
Why Are These Important?

- **Procedural Fairness:** Decisions should be based on qualifications, not on distant proxies that are spuriously associated with the outcome of interest.
- **Distribution Shift:** Expect models to not fail under unseen distributions

What Kinds Of Issues
Do We Observe?

Curated Training Task Fail To Represent Reality

- E.g., how much *reading* does reading comprehension require? (Kaushik et al., 2018)
- Models can predict correct answers by ignoring the questions altogether.



Models Are Often Vulnerable To Small, Irrelevant Perturbations To Inputs

- E.g., adversarial examples for evaluating reading comprehension systems (Jia and Liang, 2017)
- Just adding a distractor phrase at the end of the paragraph elicits an incorrect prediction from QA models.

Article: Super Bowl 50

Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”*

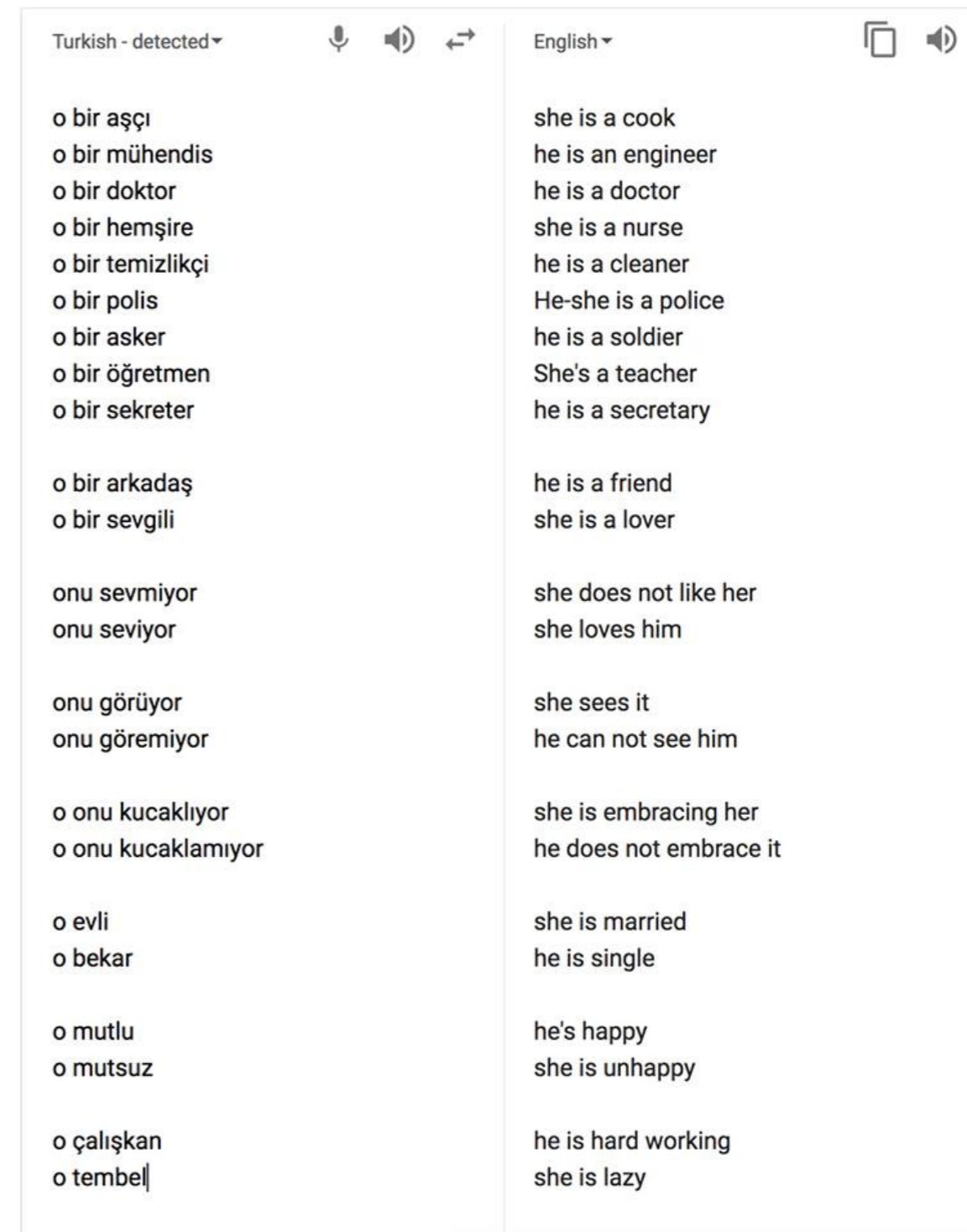
Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Bias Towards Protected Attributes

- E.g., when translating gender neutral Turkish sentences into English, Google associates he/she pronouns with stereotypically male/female dominated jobs, etc.



The screenshot shows a Google Translate interface with Turkish on the left and English on the right. The Turkish text is gender-neutral, but the English translations are biased towards male or female roles.

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover
onu sevmiyor	she does not like her
onu seviyor	she loves him
onu görüyor	she sees it
onu göremiyor	he can not see him
o onu kucaklıyor	she is embracing her
o onu kucaklamıyor	he does not embrace it
o evli	she is married
o bekar	he is single
o mutlu	he's happy
o mutsuz	she is unhappy
o çalışkan	he is hard working
o tembel	she is lazy

Bias In Human Annotation

- For e.g., Toxicity classification datasets are biased against LGBTQ community (Dixon et al., 2017).
- Can arise from a combination of (possibly) underspecified annotations guidelines and the positionality of annotators themselves.
 - Different cultural and social norms. See Byrne (2016) and Fazelpour (2020).

Detecting Biases In NLP Systems

Commonly Employed Techniques

- Association tests
- Analyzing performance measures across groups
- Counterfactual evaluations

Word Embedding Association Test (WEAT)

- Embeddings learn relationships derived from co-occurrence statistics (e.g., king - man + woman = queen)
- But what if your words also keep company with unsavoury stereotypes and biases? (e.g., doctor - man + woman = nurse)
- Consider **two sets of target words** (e.g., programmer, engineer, ... and nurse, teacher, ...) and **two sets of attribute words** (e.g., man, male, ... and woman, female ...).
- **Null Hypothesis:** There is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words.
- **Permutation test:** measures the (un)likelihood of the null hypothesis by computing the probability that a random permutation of the attribute words would produce the observed (or greater) difference in sample means.

Mathematical Formulation

- Let X and Y be two sets of target words of equal size, and A, B the two sets of attribute words.
- The test statistic is:

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

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

- $s(w, A, B)$: association of w with the attribute.
- $s(X, Y, A, B)$: differential association of the two sets of target words with the attribute.
- Let $\{(X_i, Y_i)\}_i$ denote all the partitions of $X \cup Y$ into two sets of equal size. The one-sided p-value of the permutation test is $\text{Pr}_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$.

Associative Biases In Word Embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017)

- Use WEAT to show that word embeddings exhibit human like social biases.

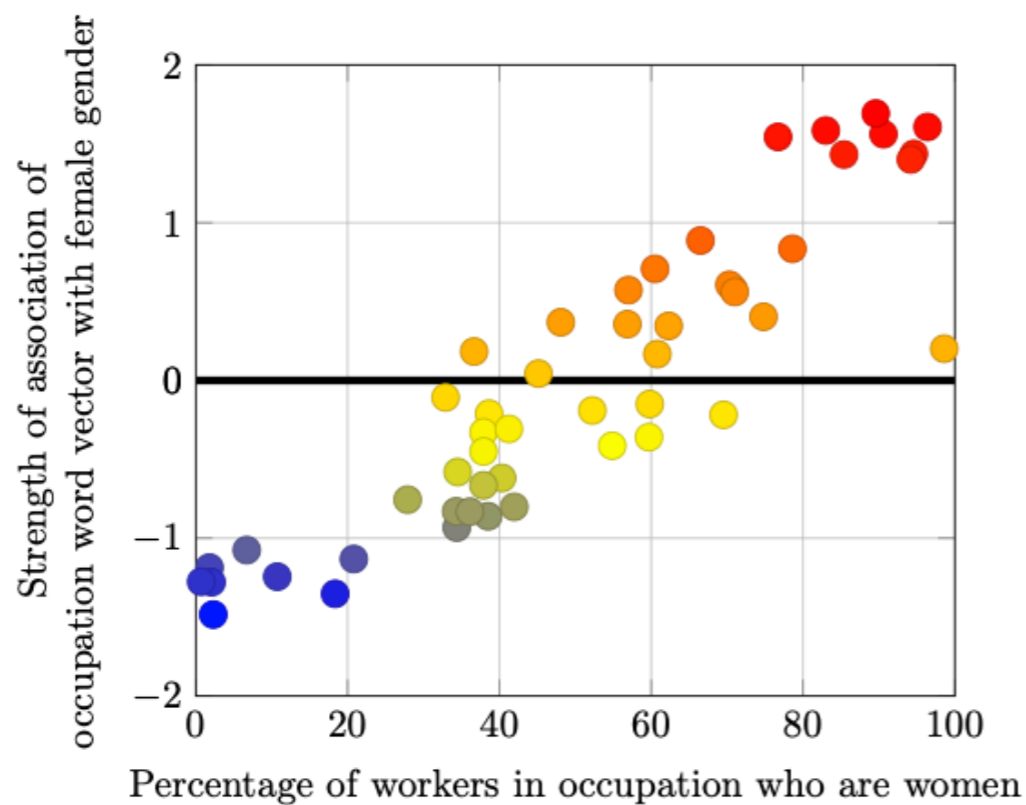


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p -value $< 10^{-18}$.

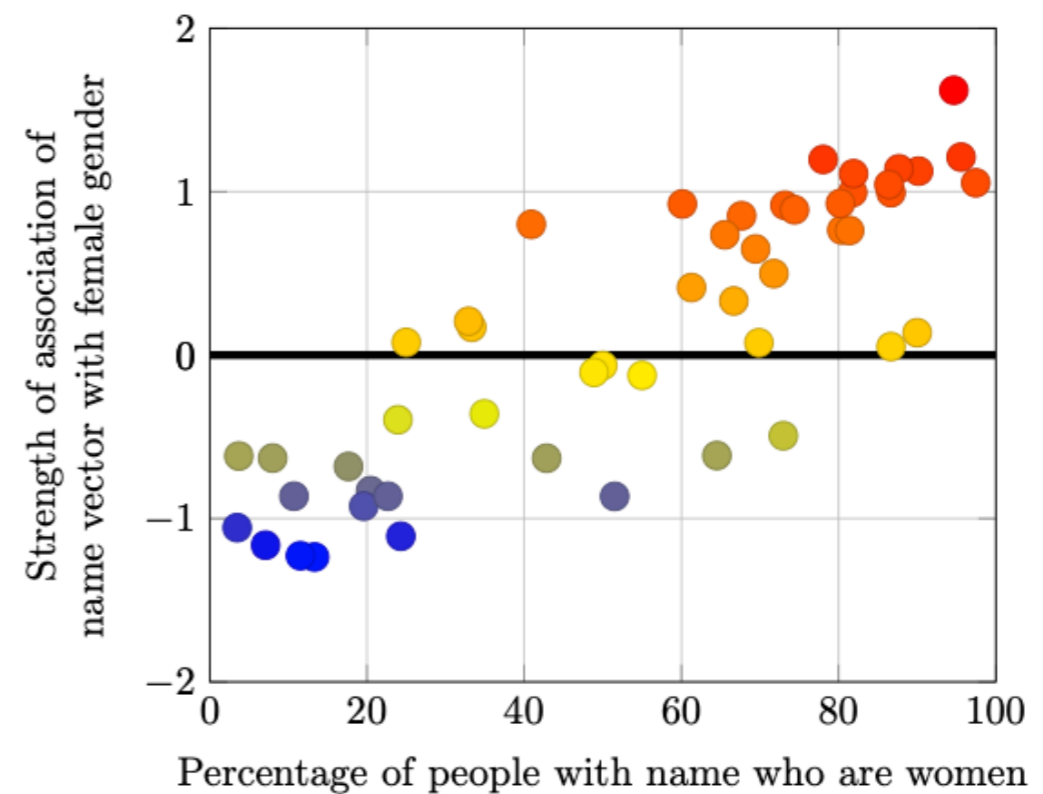


Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p -value $< 10^{-13}$.

Extending Embedding Association Test To Sentences (May et al., 2019)

- Extend WEAT to measure bias in sentence encoders (Sentence Encoder Association Test; SEAT).
- Slot words into each of several semantically bleached sentence templates such as “This is <word>.”, “<word> is here.”
- Templates are designed to convey little specific meaning beyond that of the terms inserted into them.
- ELMo and BERT display less evidence of association bias compared to older (context free) methods.

Social And Intersectional Biases In Contextualized Word Representations (Tan and Celis, 2019)

- Sentence templates may not be as semantically bleached
- Lack of evidence of bias should not be taken as a lack of bias.
- Solution: Instead of using the sentence encoding, use the contextual word representation of the token of interest.
- Avoids confounding contextual effects at the sentence level, which can obscure bias.

Social And Intersectional Biases In Contextualized Word Representations (Tan and Celis, 2019)

- Racial bias is strongly encoded in contextual word models.
- Bias effects for intersectional minorities are exacerbated beyond their constituent minority identities.
 - Intersectionality (Crenshaw, 1989): Interconnected nature of social categorizations such as race, class, and gender, regarded as creating overlapping and interdependent systems of discrimination or disadvantage.
- BERT exhibits the highest proportion of bias on both race and intersectional tests, and the highest proportion overall among contextual word models.

Issues w/ Association Tests

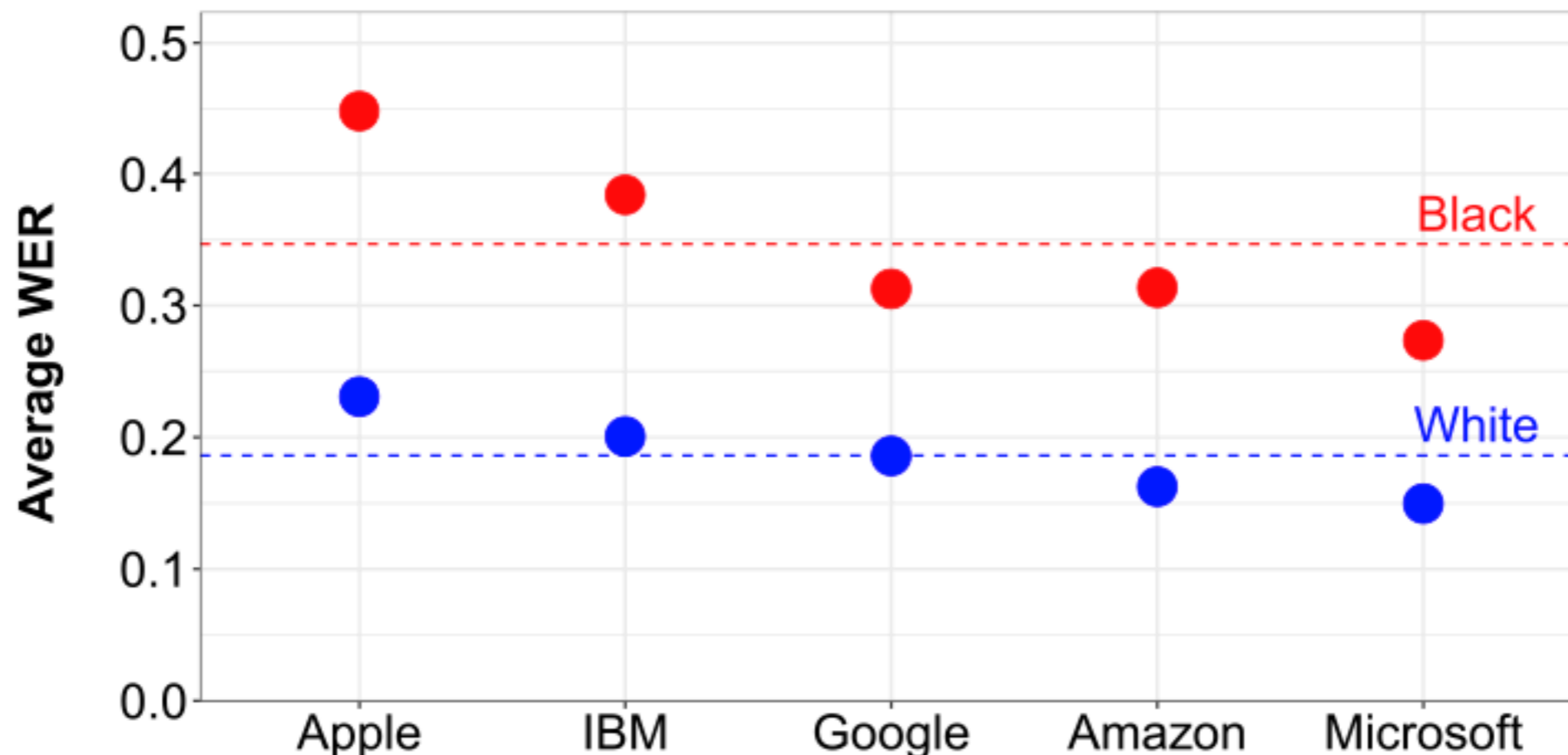
- **Positive predictive ability:** It can detect presence of bias, but not it's absence.
- Representations are trained without explicit bias control mechanisms on naturally occurring text. **A lack of evidence of bias is not a lack of bias.**

Analysis Over Error Rates

- Background: In U.S. Labor Law disparate impact is when practices adversely affect one group of people of a protected characteristic more than other (even unintentionally).
- Loosely speaking, algorithms exhibit *impact disparity* when outcomes differ across subgroups.
- One way to identify this disparity in NLP systems is by comparing performance measures (e.g., error rates, false positives, false negatives, etc.) across groups.

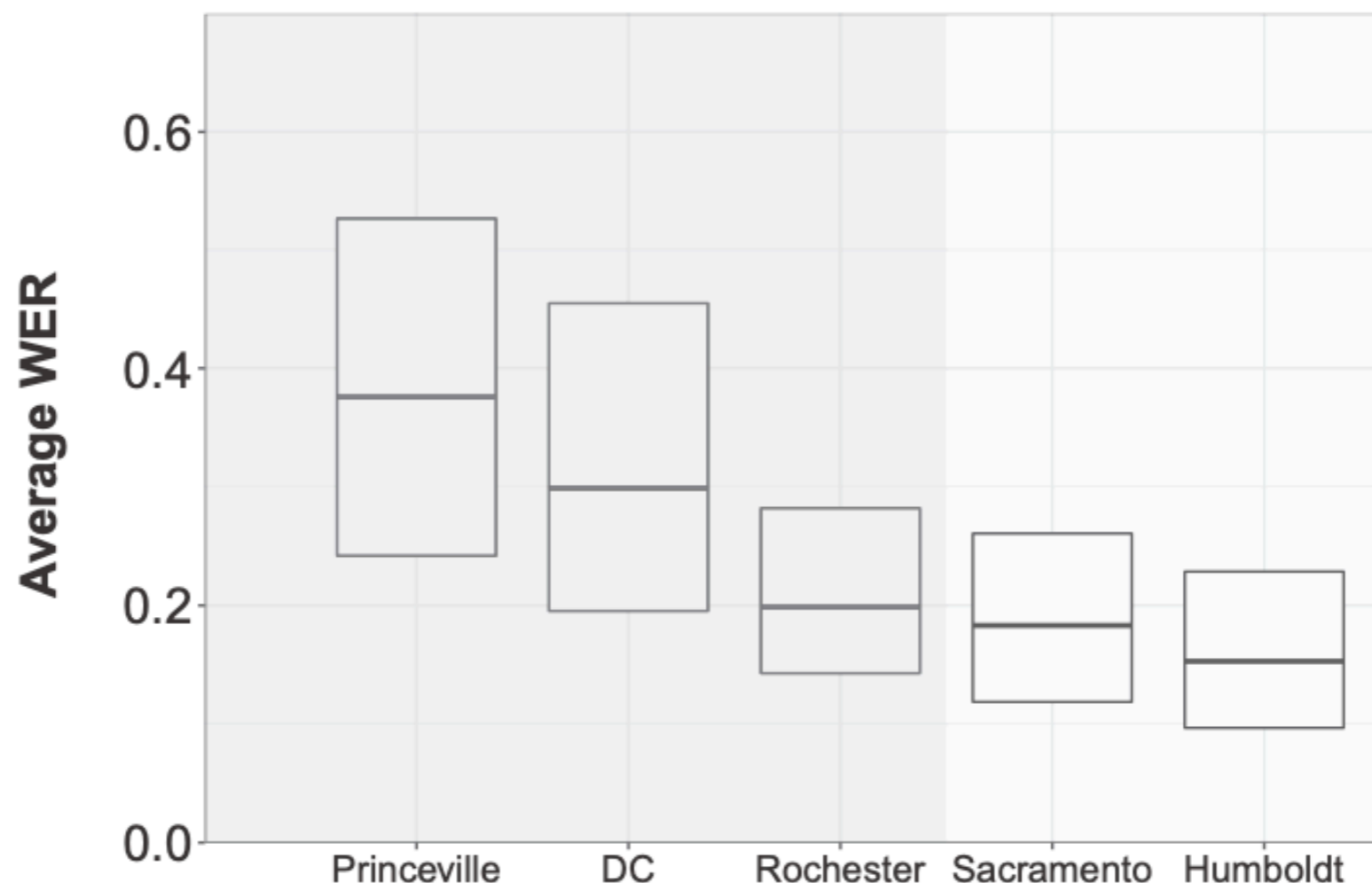
Racial Disparities In Automated Speech Recognition (Koenecke et al. 2020)

- Examined five ASR systems by Amazon, Apple, Google, IBM, and Microsoft.
- 42 white speakers and 73 black speakers; average word error rate (WER) for black speakers was 0.35 compared to 0.19 for white speakers.



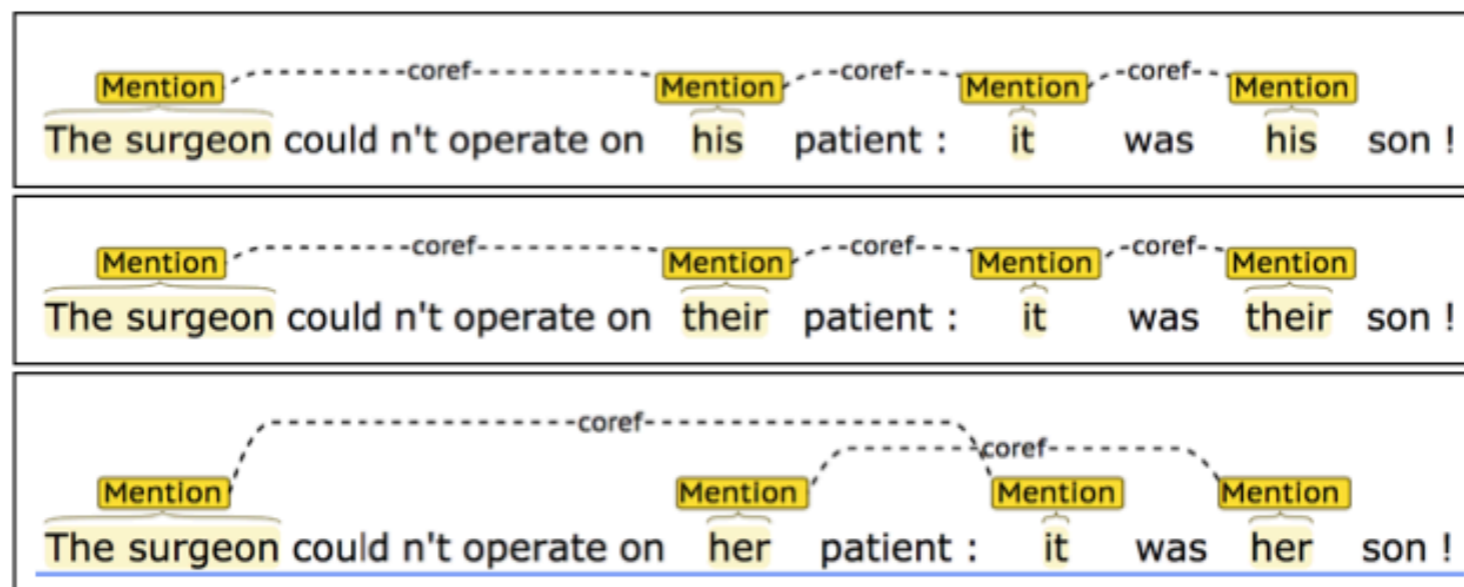
Racial Disparities In Automated Speech Recognition (Koenecke et al. 2020)

- Similar disparities were observed between predominantly African American cities (in grey) and predominantly White cities (in white).



Counterfactual Evaluation

- Modify text by flipping protected attributes (gender, race, etc.) and observe differences in model performance.
- For e.g., Gender Bias in Coreference Resolution (Rudinger et al., 2018).
- Introduce a set of minimal pair sentences that differ only by pronoun gender.



Mitigating(?) Biases

(Imperfect) Ways To Mitigate

- **Automatic mitigation**
- **Careful data creation/augmentation:** balancing groups, diversifying data, etc.
- **Humans in the loop:** counterfactually augmented data, feature feedback, etc.

Common Automatic Mitigation Techniques

- Feature invariant learning
- Debiasing embeddings
- Null space projection

Feature Invariant Learning

- Learn representations that produce accurate classifications while not being good at identifying protected variables (Zemel et al., 2013).

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

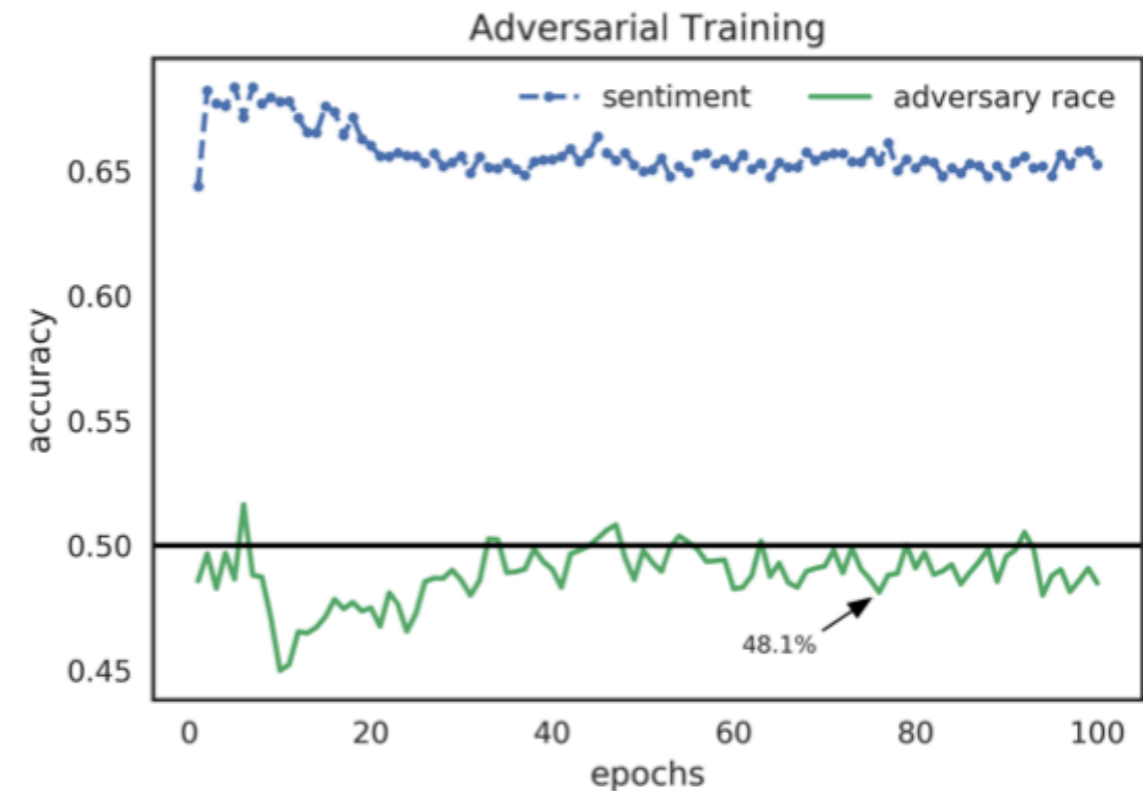
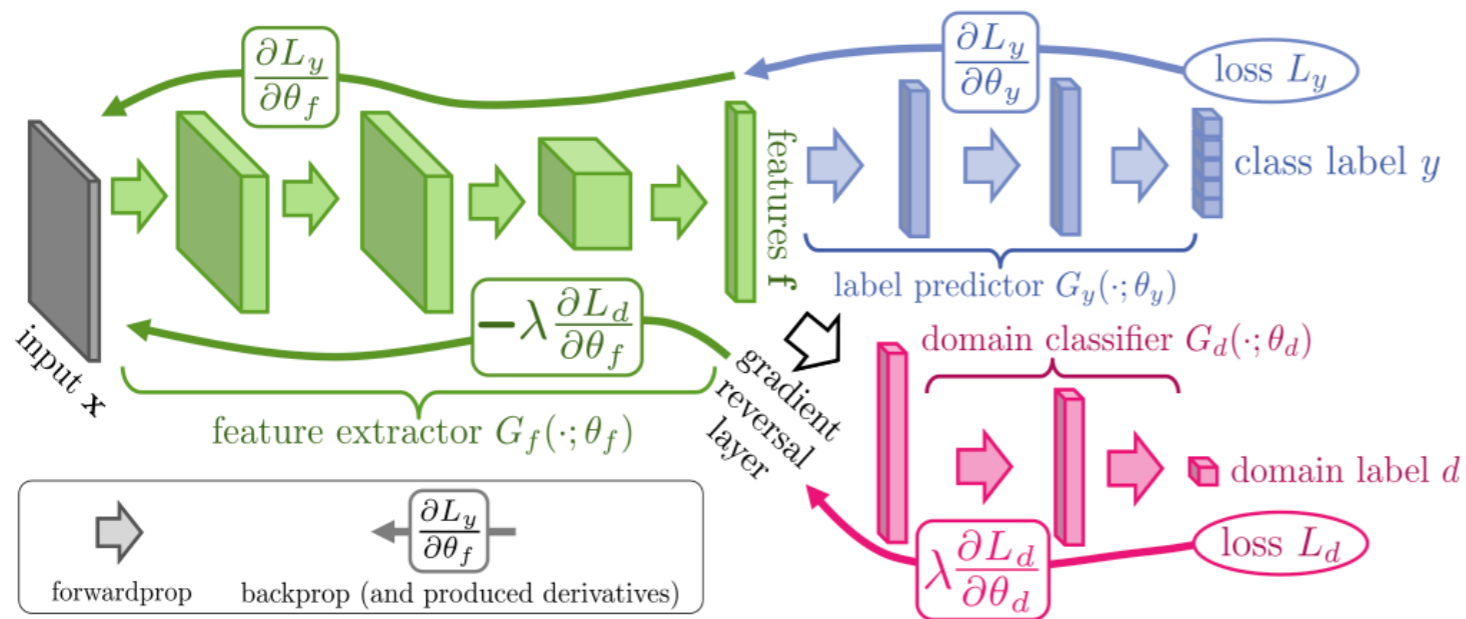
Classifications should be good

Reconstructions should be good

Intermediate Representations should be indistinguishable across values of the protected variable

Feature Invariant Learning

- Adversarial training (Ganin and Lempitsky, 2015): Learn representations invariant to protected attributes (for e.g., race).



Issues w/ Adversarial Removal

- Demographic information can be recovered even after adversarial training (Elazar and Goldberg, 2018).

Data	Task	Protected Attribute	Task Acc	Leakage	Δ
DIAL	Sentiment	Race	64.7	56.0	5.0
	Mention	Race	81.5	63.1	9.2
PAN16	Mention	Gender	75.6	58.5	8.0
	Mention	Age	72.5	57.3	6.9

Debiasing Word Embeddings (Bolukbasi et al., 2016)

- Identify a direction of the embedding that captures the bias.
- Then: Neutralize and Equalize or Soften.
 - Neutralize: gender neutral words are “zero” in the gender subspace.
 - Equalize: Any neutral word is equidistant to all words in each *equality set*. **Neutralize and equalize is referred as hard-debiasing.**
 - Soften: Reduces the differences between equality sets while maintaining as much similarity to the original embedding as possible. **Neutralize and soften is referred as soft-debiasing.**

Debiasing Word Embeddings (Bolukbasi et al., 2016)

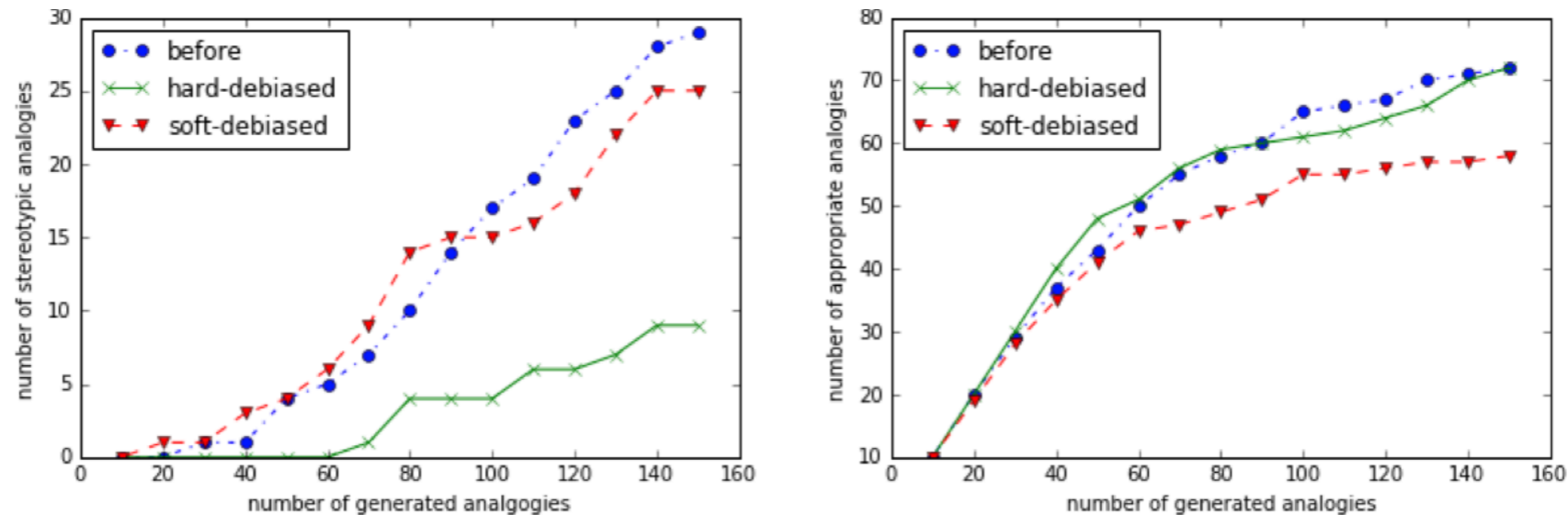


Figure 8: Number of stereotypical (Left) and appropriate (Right) analogies generated by wordembeddings before and after debiasing.

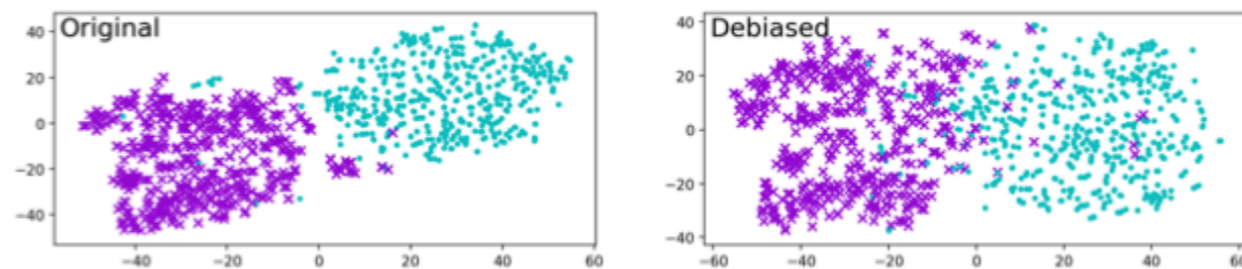
- Trouble in paradise: Consider {grandmother, grandfather} and {guy, gal}
- Babysit would become equidistant to both words in each set
- What about the sentence *Grandfather a regulation?* Should this be equally probable as *Grandmother a regulation?*

Debiasing Methods Cover Up Systematic Gender Biases (Gonen and Goldberg, 2019)

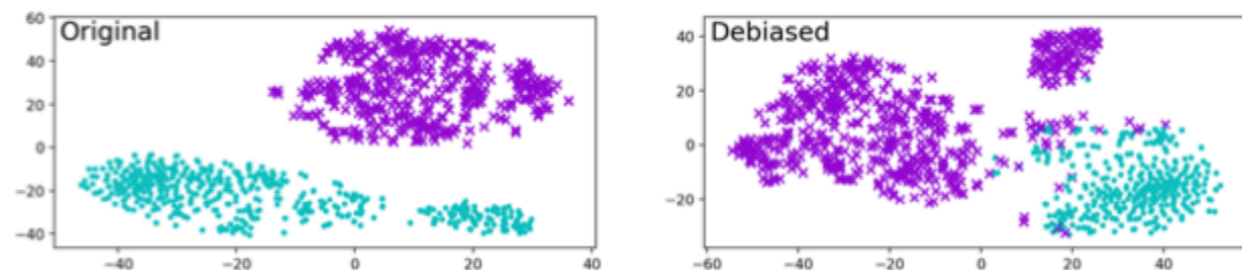
- Male- and female-biased words cluster together.
- Embedding clusters align with gender 85% of the time.
- Conclusion: Gender bias is still embedded in the representation after de-biasing.

Debiasing Methods Cover Up Systematic Gender Biases (Gonen and Goldberg, 2019)

- Cannot directly “observe” the bias for a word.
- But word is still close to *socially-marked* feminine words.
- For e.g., “nurse” will no longer be closer to explicitly marked feminine words but will be close to “receptionist”, “caregiver” and “teacher”.



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



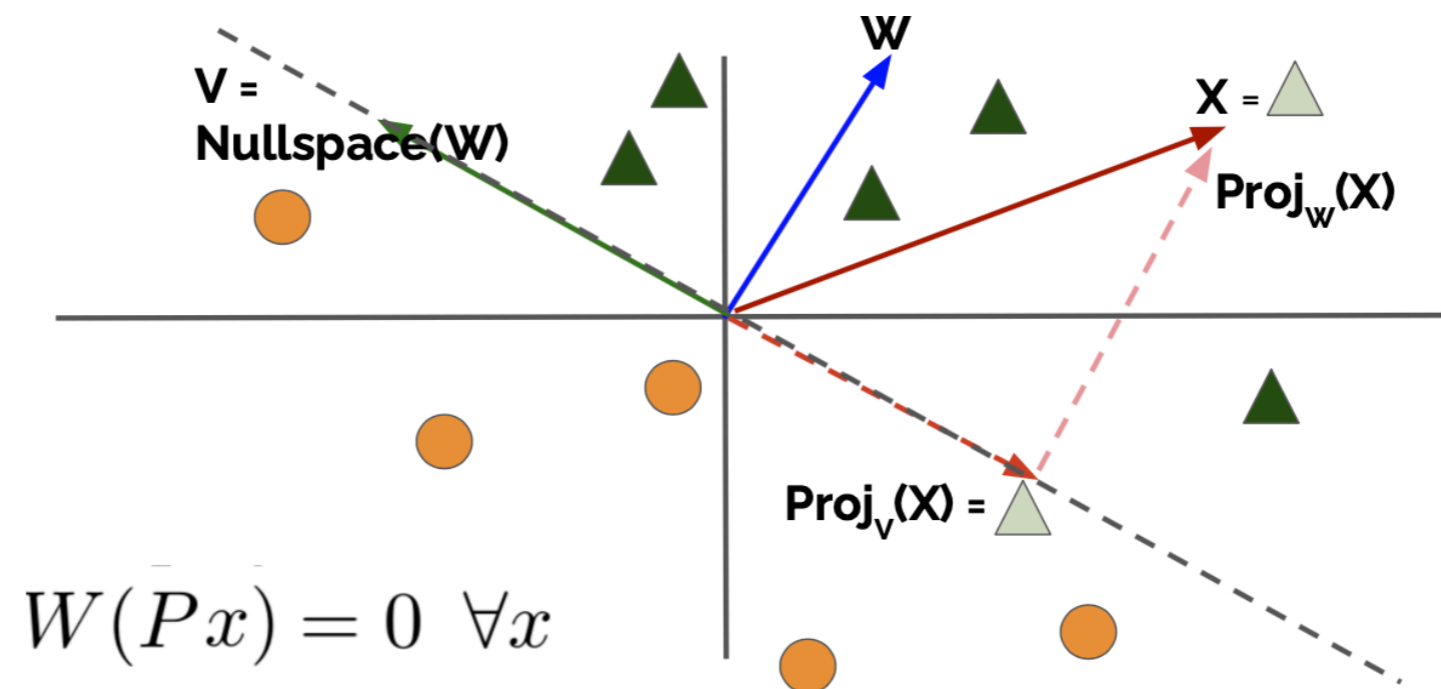
(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Iterative Nullspace Projection (Ravgofel et al., 2020)

- Learn a transformation $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that z_i (label of protected category) cannot be predicted from $g(x_i)$.
- No linear classifier $w(\cdot)$ can predict z_i from $g(x_i)$ with an accuracy greater than majority class baseline.
- Also wish for $g(x_i)$ to stay informative: Want $g(x)$ to have as minimal influence as possible on the end task performance.

Iterative Nullspace Projection (Ravgofel et al., 2020)

- Let c be a trained linear classifier, parameterized by a matrix $W \in \mathbb{R}^{k \times d}$, that predicts a property z (race, gender, etc.) with some accuracy.
- Find a projection matrix P , which projects into the nullspace $N(W) = \{x | Wx = 0\}$



Iterative Nullspace Projection (Ravgofel et al., 2020)

		BoW	FastText	BERT
Accuracy (profession)	Original	78.2	78.1	80.9
	+INLP	80.1	73.0	75.2
$GAP_{male}^{TPR,RMS}$	Original	0.203	0.184	0.184
	+INLP	0.124	0.089	0.095

Table 2: Fair classification on the Biographies corpus.

- Each projection only removes a single direction.
- Repeat this process until convergence

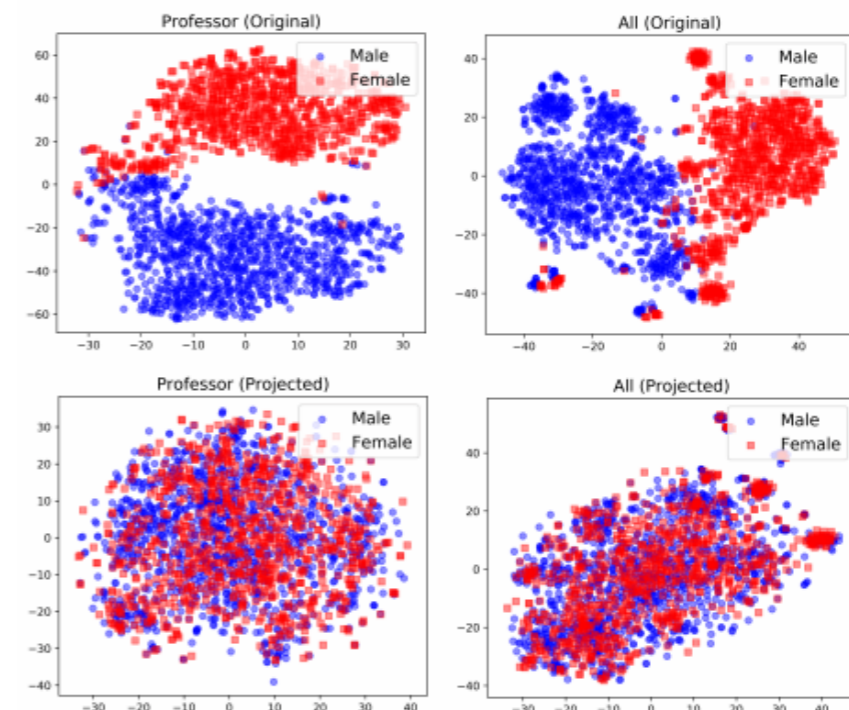


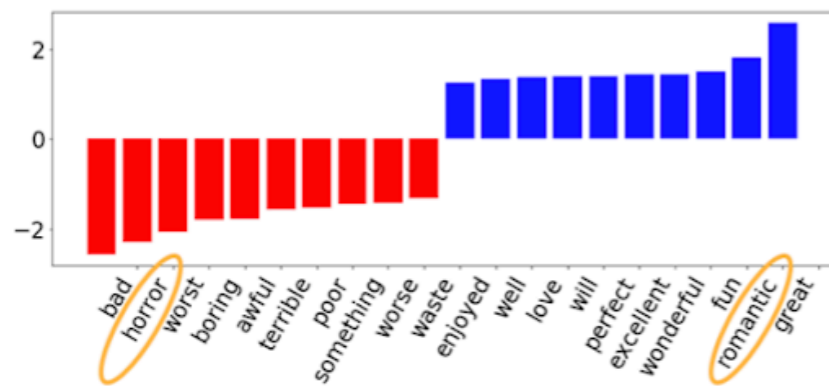
Figure 3: t-SNE projection of BERT representations for the profession “professor” (left) and for a random sample of all professions (right), before and after the projection.

Automatic Data Augmentation

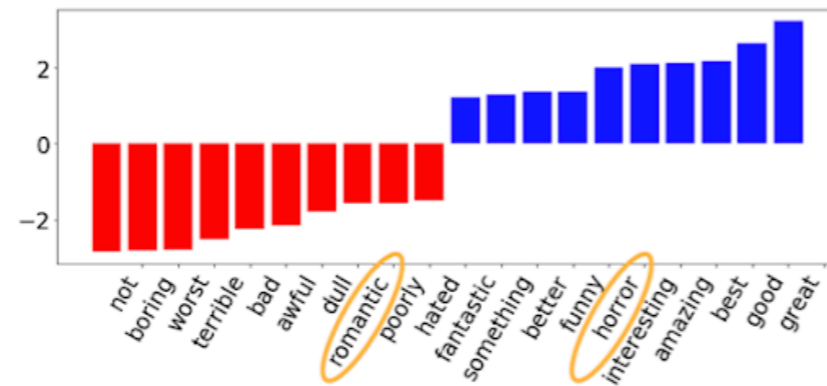
- Lu et al. (2018): programmatically alter text to invert gender bias. Combine the original and manipulated data.
 - For example, *the doctor ran because he is late* becomes *the doctor ran because she is late*.
 - Con: No substitutions even if names co-refer to a gendered pronoun.
- Zmigrod et al. (2019): Use a Markov random field to infer how the sentence must be modified while altering the grammatical gender of particular nouns to preserve morpho-syntactic agreement.

Mitigation With Humans In The Loop

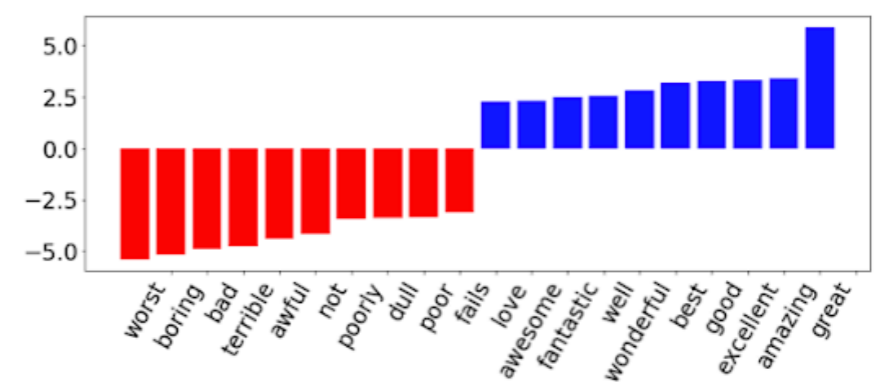
- Kaushik et al. (2020; 2021) employ humans to edit documents to make a counterfactual label applicable.
- Models trained on augmented data are more robust out-of-domain and tend to rely less on spurious patterns.



(a) Trained on the original dataset



(b) Trained on the revised dataset



(c) Trained on combined dataset

Can Model Biases Be Exploited To Understand Society?

- Using NLP to quantify gender bias in sports journalism (Fu et al., 2016).
- Using NLP to quantitatively study the ways in which the language used to describe men and women is different (Hoyle et al., 2019).
- Using NLP to study racial bias in sports commentary (Merullo et al., 2019).

What Are We Doing Wrong?

Critiques Of “Bias” In NLP (Lin Blodgett et al., 2020)

- Survey 146 papers analyzing “bias” in NLP systems
- **Found motivations as often vague, inconsistent, and lacking in normative reasoning.**
- Mismatch between motivations and proposed quantitative techniques for measuring or mitigating “bias”
- Papers do not engage with the relevant literature outside of NLP.

Critiques Of “Bias” In NLP (Lin Blodgett et al., 2020)

- Recommendations on how to conduct work analyzing “bias” in NLP
 - Ground work in relevant literature outside of NLP.
 - Provide explicit statements of why the system behaviors that are described as “bias” are harmful, in what ways, and to whom.
 - Engage with the lived experiences of members of communities affected by NLP systems.

Cis-normativity In Published NLP Papers (Trista Cao and Daume 2020)

- Took a sample of ~150 papers from the ACL anthology that mention the word *gender* and coded them according to some questions:
 - Does the paper discuss coreference resolution?
 - Does the paper study English?
 - Does the paper deal with linguistic gender (grammatical gender or gendered pronouns)?
 - Does the paper deal with social gender?
 - Does the paper distinguish linguistic from social gender?, etc.

Cis-normativity In Published NLP Papers (Trista Cao and Daume 2020)

- 22 papers looked at coreference but...
- Only 5.5% distinguish social from linguistic gender (despite it being relevant)
- Only 5.6% explicitly model gender as inclusive of non-binary identities
- No papers treat gender as anything other than completely immutable
- Only one paper (!) considers neopronouns and/or specific singular THEY.

Well-Intentioned Works Can Have Dual Impacts

- Advanced grammar analysis: improve search and educational NLP, but also reinforce prescriptive linguistic norms.
- Stylometric analysis: help discover provenance of historical documents, but also unmask anonymous political dissenters.
- Text classification and IR: help identify information of interest, but also aid censors.
- NLP can be used to identify fake reviews and news, and also to generate them.

These types of problems are difficult to solve, but important to think about, acknowledge and discuss.

Additional Resources

- Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem (Saunders and Byrne, 2020)
- Towards Controllable Biases In Language Generation (Sheng et al., 2020)
- Gender as a Variable in Natural-Language Processing: Ethical Considerations (Larson, 2017)
- Do Artifacts Have Politics? (Winner, 1980)
- The Trouble With Bias (Crawford, 2017)
- Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview (Shah et al., 2020)
- Moving beyond “algorithmic bias is a data problem” (Hooker, 2021)
- Fairness and Machine Learning. (Barocas et al., 2019)

Questions?