Fair NLP

May 6, 2020

Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

Recap

- Basic Data Preprocessing Techniques for Fairness
- The Expected Joint Distribution Under $Y \perp\!\!\!\perp A$

$$P_{exp}(Y = y, A = a) = P(Y = y) \cdot P(A = a)$$
$$= \frac{|\{x \in \mathcal{D} | x_Y = y\}|}{|\mathcal{D}|} \cdot \frac{|\{x \in \mathcal{D} | x_A = a\}|}{|\mathcal{D}|}$$

- Our Observed Joint Distribution

$$P_{obs}(Y = y, A = a) = \frac{|\{x \in \mathcal{D} | x_Y = y, x_A = a\}|}{|\mathcal{D}|}$$

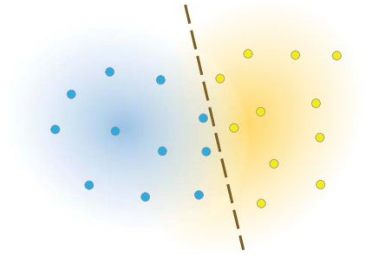
Resample/Reweigh Data to Match Expected Distribution

Recap

- Reweighting

$$W(x) = \frac{P_{exp}(Y = x_y, A = x_a)}{P_{obs}(Y = x_y, A = x_a)}$$

- Resampling
  - Universal Sampling
    - Sample uniformly
  - Preferential Sampling
    - Sample based on model uncertainty

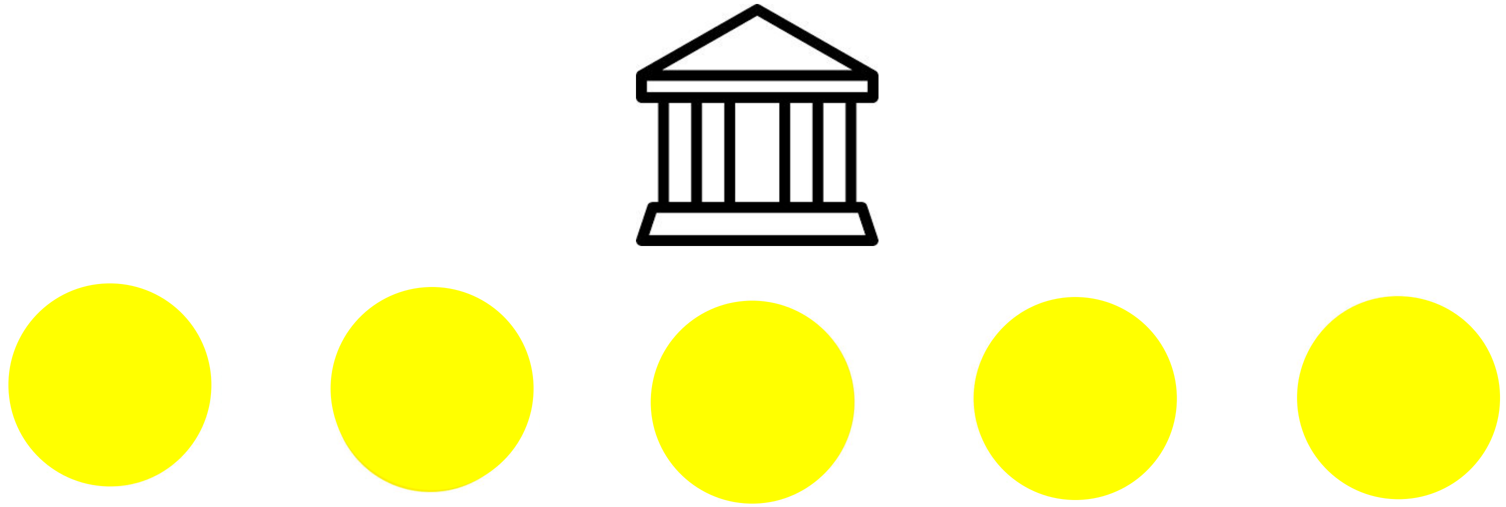


[Kamiran et al, 2012]

Outline

- Fairness Through Data/Prediction Manipulations
  - Individual Fairness
  - Optimized Pre-processing
  - Learning to Defer
- Fair NLP
  - Biases in NLP Models
  - Data Augmentation
  - Debiasing Word Embedding
  - Adversarial Learning

Individual Fairness



Income = $50k Credit Score = 690	Income = $43k Credit Score = 650	Income = $50k Credit Score = 690	Income = $70k Credit Score = 740	Income = $100k Credit Score = 750
Accepted	Accepted	Denied ???	Accepted	Accepted

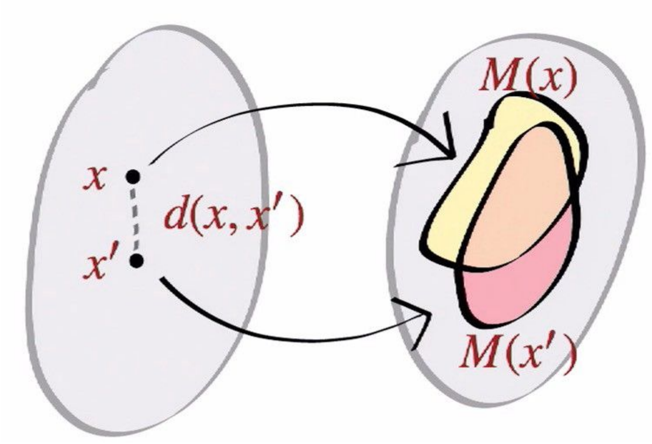
group 1

group 2

Individual Fairness

- A predictor M achieves individual fairness under a distance metric d iff
    - Similar Samples are treated similarly, in other words

$$M(x_i) \approx M(x_j) | d(x_i, x_j) \approx 0$$



Individual Fairness

Individual

$$M(x_i) \approx M(x_j) | d(x_i, x_j) \approx 0$$

Group 1

Income = $19k
Credit Score = 690

Income = $23k
Credit Score = 720

Income = $60k
Credit Score = 800

Group 2

Income = $20k
Credit Score = 680

Income = $27k
Credit Score = 700

Income = $65k
Credit Score = 810

Fairness Criteria

Individual Treatment	Group Treatment
Fairness Through Unawareness Excludes Sensitive Information A from the predictor	**Demographic Parity** $P(\hat{Y}=1 \mid A=1) = P(\hat{Y}=1 \mid A=0)$
Individual Fairness $M(x_i) \approx M(x_j) \mid d(x_i, x_j) \approx 0$	**Equal Opportunity/Odds** $P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=1)$ $P(\hat{Y}=1 \mid A=0, Y) = P(\hat{Y}=1 \mid A=1, Y)$

Outline

- Fairness Through Data/Prediction Manipulations
  - Individual Fairness
  - Optimized Pre-processing
  - Learning to Defer
- Fair NLP
  - Biases in NLP Models
  - Data Augmentation
  - Debiasing Word Embedding
  - Adversarial Learning

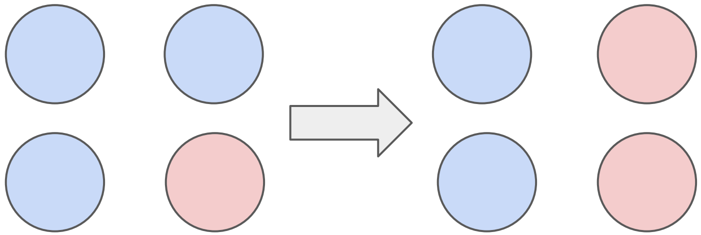
Optimized Pre-Processing for Fairness

- Can We Automate the Resampling Process?
  - Turn the the manual process into an optimization based approach
  - Include more criteria than Demographic Fairness
  - Allow transformations of data
- Optimized Pre-Processing
  - Given sensitive feature D, learn a probabilistic mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ that transfers
  - Satisfies three constraints

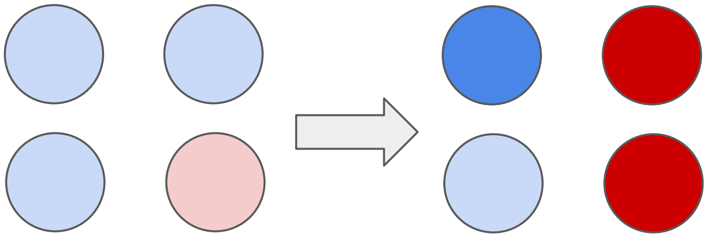
$$\{(D_i, X_i, Y_i)\}_{i=1}^{n} \xrightarrow{p_{\hat{X},\hat{Y}|X,Y,D}} \{(D_i, \hat{X}_i, \hat{Y}_i)\}_{i=1}^{n}$$

Calmon el al, 2017

Resampling and Transforming



Resampling



Transforming

Constraint 1: Utility Preservations

- A Utility Function to Preserve the Joint Probability
  - e.g. KL Divergence

$$p_{\hat{X},\hat{Y}} \Longleftrightarrow p_{X,Y}$$

transformed data | original data

Calmon el al, 2017

Constraint 2: Discrimination Control

- Constrain the dependency of the target variable y given sensitive feature d to march target $p_{Y_T}(y)$
  - J - distance measure $J(p,q) = \left|\frac{p}{q} - 1\right|$
  - $\epsilon_{y,d}$ - a small number used as our tolerance

$$J\left(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) \leq \epsilon_{y,d} \; \forall \, d \in \mathcal{D}, y \in \{0,1\}$$

When $p_{\hat{Y}|D}(y|d) = p_{Y_T}(y)$, we achieve Demographic Parity

Calmon el al, 2017

Constraint 3: Distortion Control

- An Implementation of the Individual Fairness

$$M(x_i) \approx M(x_j) | d(x_i, x_j) \approx 0$$

- The Mapped Sample $\hat{X}, \hat{Y}$ Has to Stay Close to the Original Sample x, y
  - $c_{d,x,y}$ - tolerance
  - δ - a similarity function
    - 1 - very different
    - 0 - very similar

$$\Pr\left(\delta((x,y),(\hat{X},\hat{Y})) = 1 \mid D = d, X = x, Y = y\right) \leq c_{d,x,y}$$

Calmon el al, 2017

Putting Things Together

Utility

Discrimination control
group fairness

$$\min_{p_{\hat{X},\hat{Y}|X,Y,D}} \quad \Delta\left(p_{\hat{X},\hat{Y}}, p_{X,Y}\right)$$

$$\text{s.t.} \quad J\left(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) \leq \epsilon_{y,d} \ \text{ and}$$

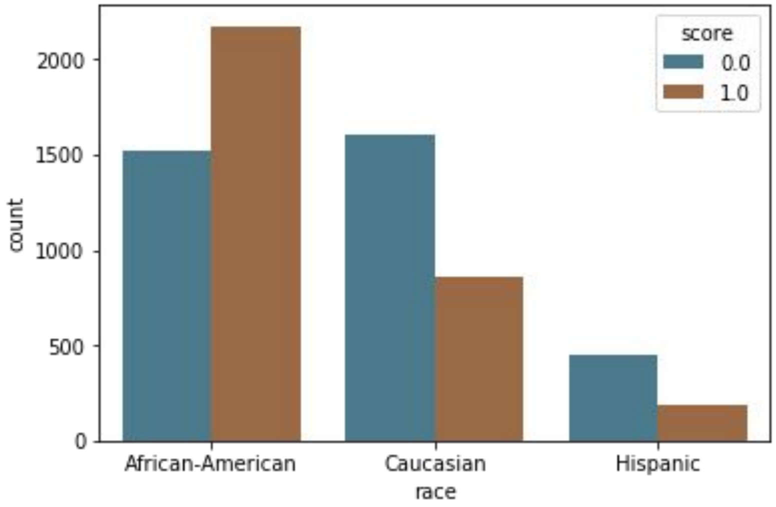
$$\mathbb{E}\left[\delta((x,y),(\hat{X},\hat{Y})) \mid D=d, X=x, Y=y\right] \leq c_{d,x,y}$$

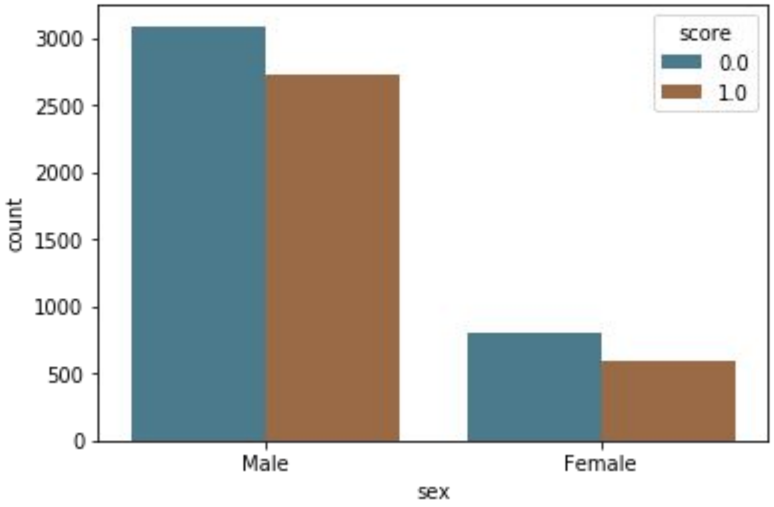
Distortion Control
Individual fairness

[Calmon el al, 2017]

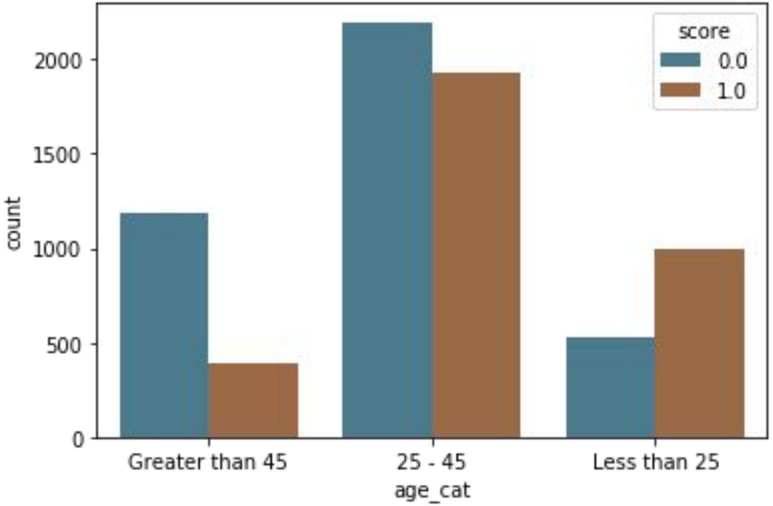
COMPAS Dataset

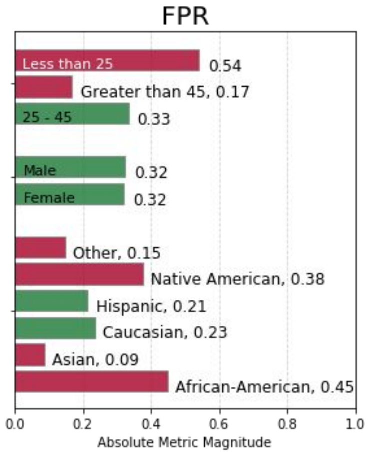
COMPAS Dataset



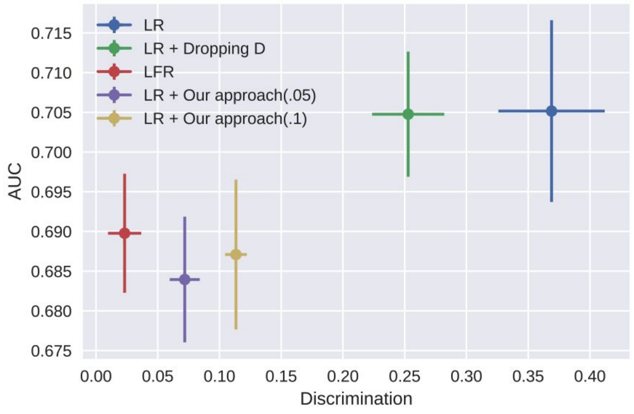


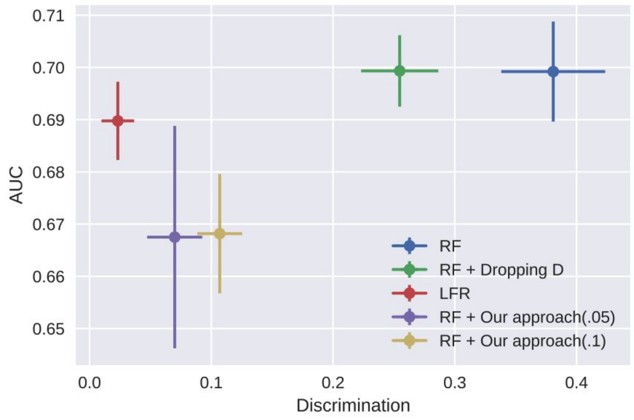
COMPAS Dataset





Results on COMPAS dataset





Logistic Regression

Random Forest

LFR - Learning Fair Representations ([Zemel et al, 2013]())

[Calmon el al, 2017]()

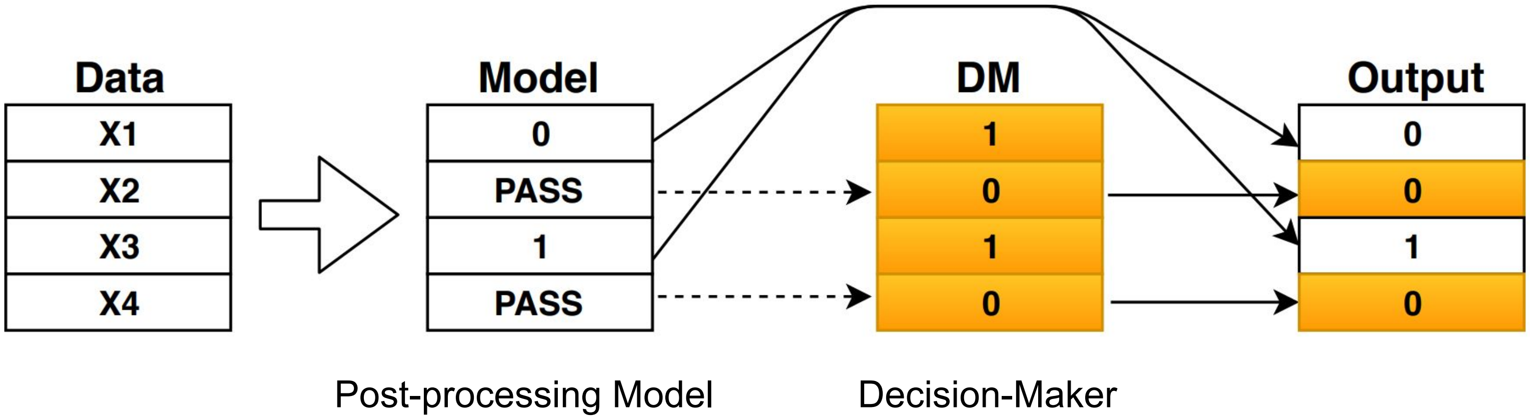
Outline

- Fairness Through Data/Prediction Manipulations
    - Individual Fairness
    - Optimized Pre-processing
    - Learning to Defer
- Fair NLP
    - Biases in NLP Models
    - Data Augmentation
    - Debiasing Word Embedding
    - Adversarial Learning

Post-Processing Methods for Fairness

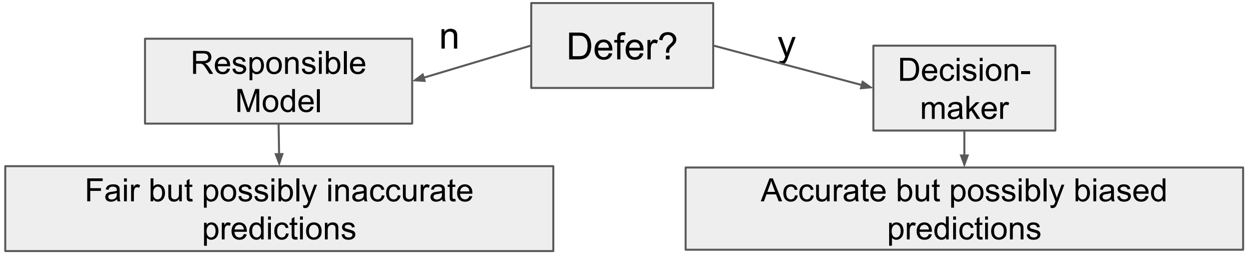
- Why Post-Processing?
  - Flexibility: No need to fine-tune the ML model
  - Model Agnostic: Can be applied across a wide range of models

- Learning to Defer



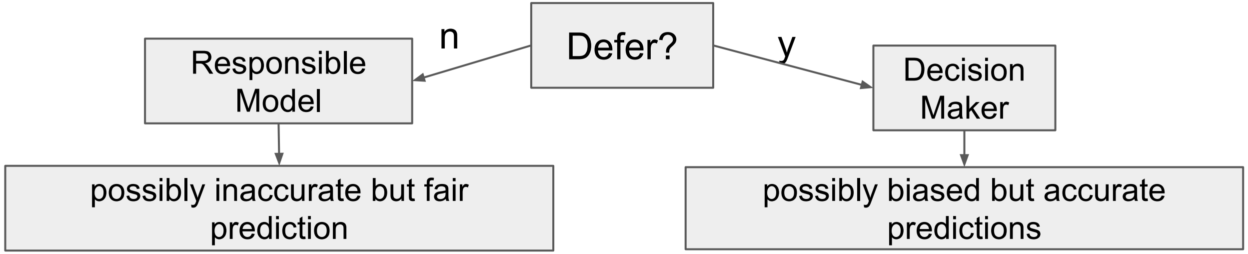
Learning to Defer

- Working Together with A Black-box Decision-maker Model
    - Decision-maker models (e.g. human) have access to important information that our model does not has
    - Decision-maker models might be biased
- Performance and Fairness Trade-offs
    - Fix the unfair predictions of the decision-maker model
    - Defer to the decision-maker the model is uncertain

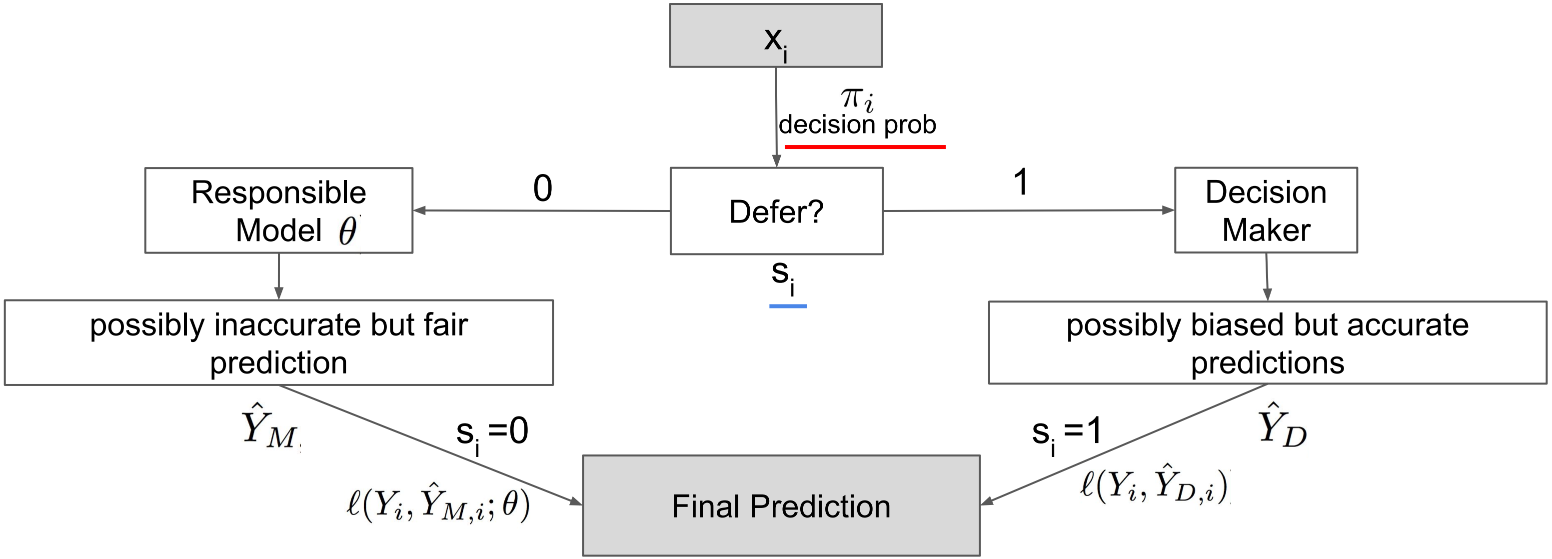


Learning to Defer

- Decision-maker Model
  - Considered as a black-box model
  - No fine-tuning, no access to its training data
- Responsible Model
  - Have access to additional data
  - Stick to fairness constraints



Training the Defer Model



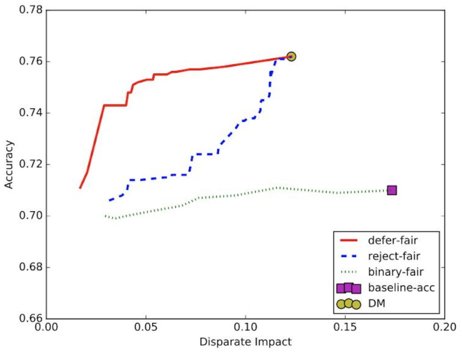
$$\mathcal{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, \pi; \theta) = \sum_i \mathbb{E}_{s_i \sim Ber(\pi_i)}[(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}; \theta) + s_i \ell(Y_i, \hat{Y}_{D,i})] + \alpha_{fair}\mathcal{R}(Y, \hat{Y}_M, \hat{Y}_D, s)$$

Fair regularizer

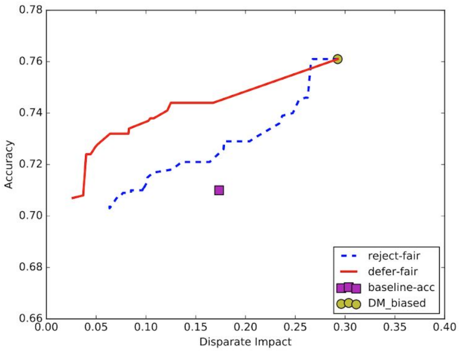
Madras et al, 2018

Results on COMPAS

- DM Model
  - High-Accuracy - DM has more data, Highly-Biased - DM is extremely biased



COMPAS, High-Accuracy DM



COMPAS, Highly-Biased DM

- DM - Decision-maker model
- Defer - Fair - Learning to Defer
- Reject- Fair - Only reject or accept DM

- Baseline - Model trained only to optimize accuracy, no DM
- Binary - Fair - Baseline optimized with fairness

[Madras et al, 2018]

Outline

- Fairness Through Data/Prediction Manipulations
  - Individual Fairness
  - Optimized Pre-processing
  - Learning to Defer
- Fair NLP
  - Biases in NLP Models
  - Data Augmentation
  - Debiasing Word Embedding
  - Adversarial Learning

Biases of NLP Models

- Denigration
  - The use of culturally or historically derogatory terms
- Under-representation
  - The disproportionately low representation of a specific group
  - e.g., a classifier's performance is adversely affected due to sampling biases of the minority protected group
- Stereotyping
  - An over-generalized belief about a particular category of people
  - e.g., a classifier attributes man to computers more than woman
- Recognition
  - Algorithms perform different for protected groups because of their inherent characteristics
  - e.g., a voice recognition algorithm has better capabilities in recognizing voices in low frequency

Biases of NLP Models

Task	Example of Representation Bias in the Context of Gender	S
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)	✓
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).	✓
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).	✗
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).	✓
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).	✓

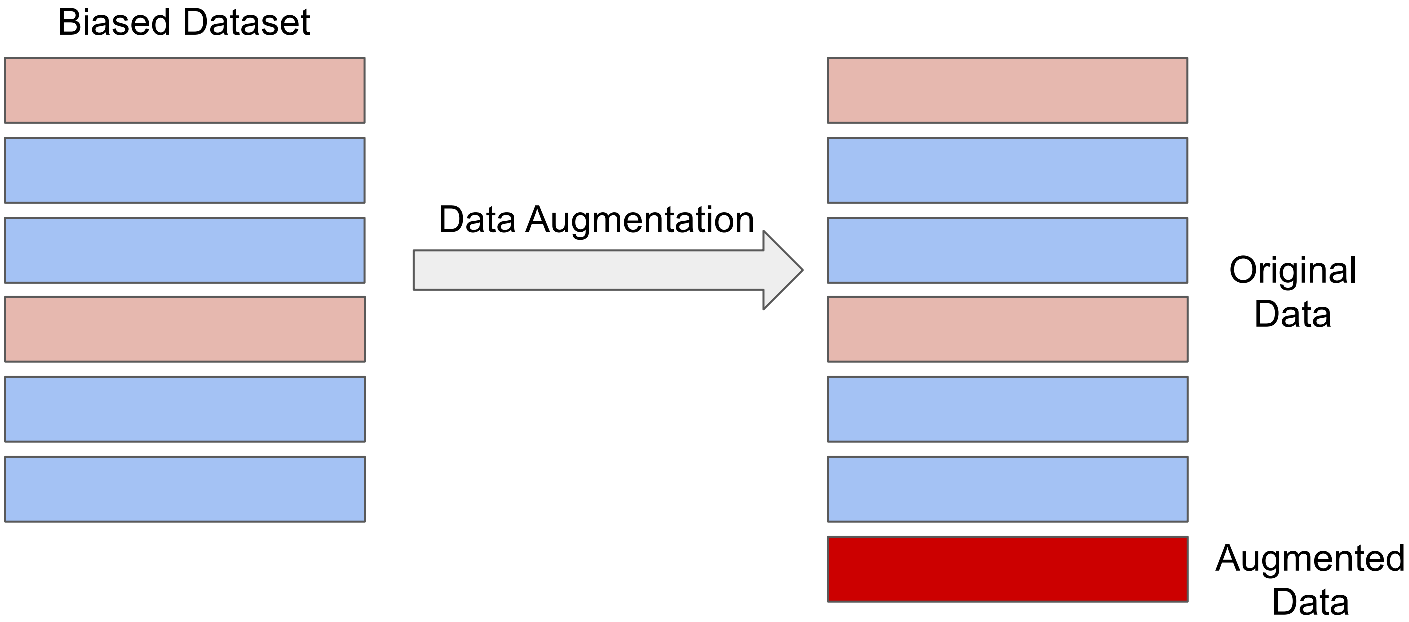
(S)tereotyping, (D)enigration, (R)ecognition, (U)nder-representation

Sun et al, 2019

Outline

- Fairness Through Data/Prediction Manipulations
    - Individual Fairness
    - Optimized Pre-processing
    - Learning to Defer
- Fair NLP
    - Biases in NLP Models
    - Data Augmentation
    - Debiasing Word Embedding
    - Adversarial Learning

Data Augmentation



Coreference Resolution

A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, “I can’t operate on this boy, he’s my son!

Does this paragraph make sense to you?

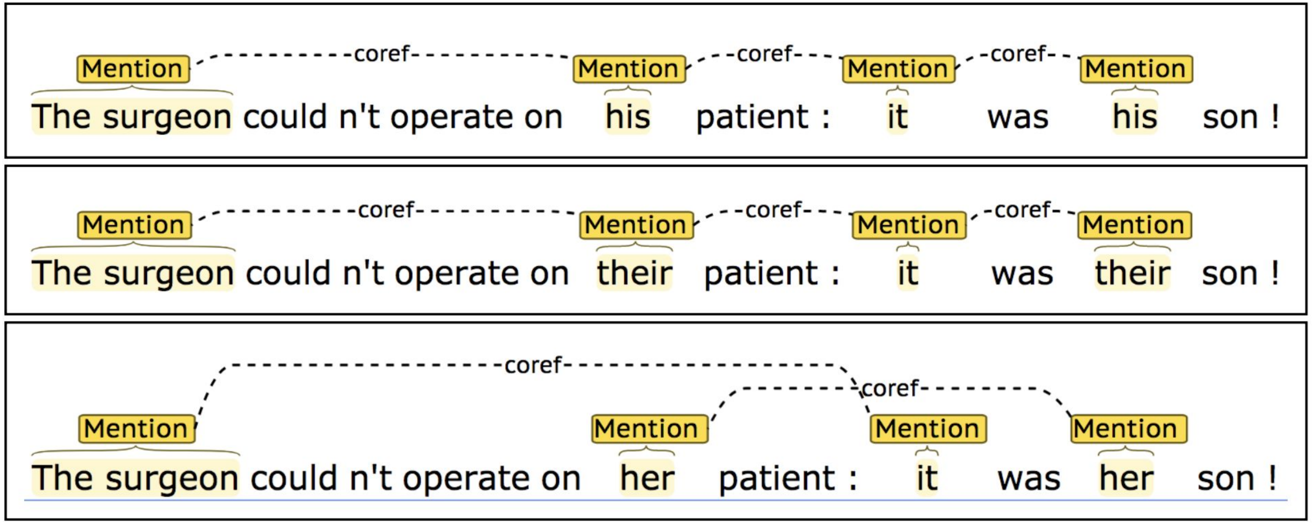
[Rudinger et al, 2018]

Gender Swapping in Coreference Resolution

Original sample

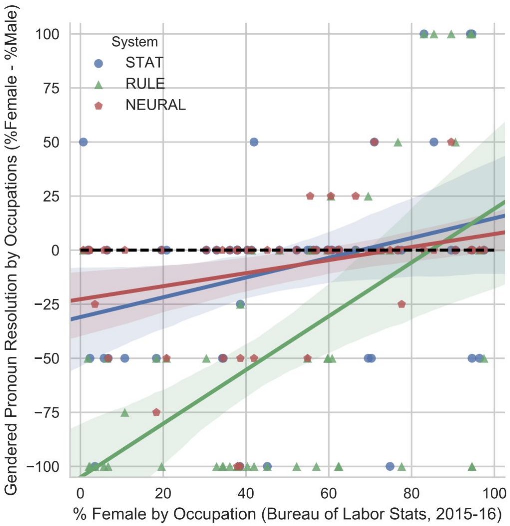
Gender swap

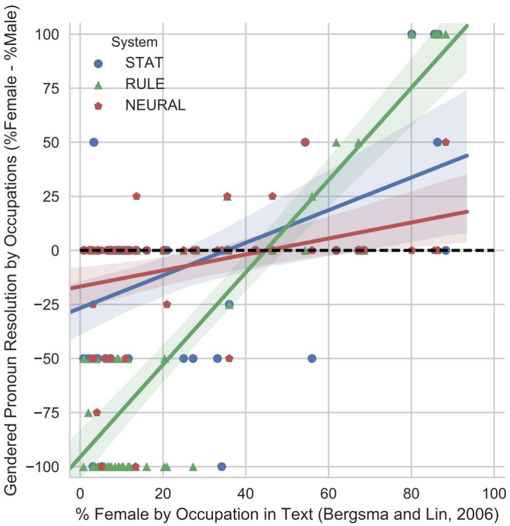
Gender swap



Rudinger et al, 2018

Results





STAT- Statistical Model (Durrett et al, 2013)
RULE - Rule Based Model (Lee et al, 2011)
NEURAL - Neural Based Model (Clark et al, 2016)

Rudinger et al, 2018

Results

Method	Anon.	Resour.	Aug.	OntoNotes	T1-p	T1-a	Avg	| Diff |	T2-p	T2-a	Avg	| Diff |
E2E	✓	✓		66.5	67.2	59.3	63.2	7.9*	81.4	82.3	81.9	0.9
E2E	✓	✓	✓	66.3	63.9	**62.8**	**63.4**	**1.1**	81.3	83.4	82.4	**2.1**
Feature	✓	✓		61.2	61.8	**62.0**	**61.9**	**0.2**	67.1	63.5	65.3	3.6
Feature	✓	✓	✓	61.0	62.3	60.4	61.4	1.9*	71.1	**68.6**	**69.9**	**2.5**

E2E (Lee et al, 2011)
Feature (Durrett et al, 2013)
Diff - Difference between pro/anti

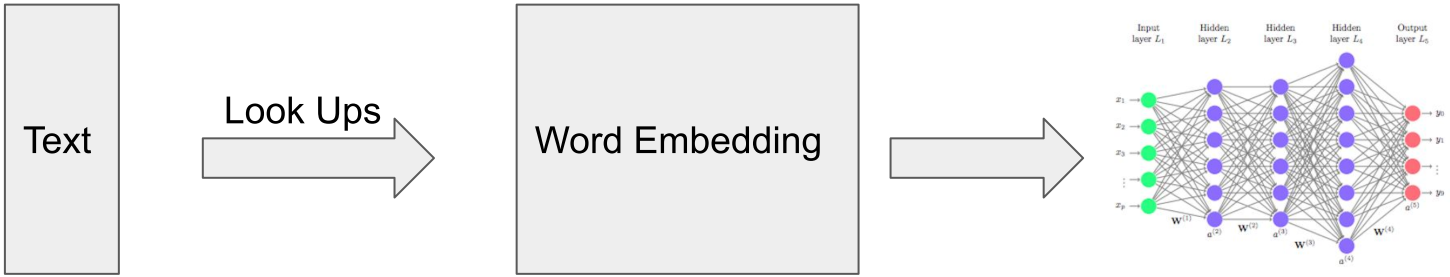
Zhao et al, 2018

Outline

- Fairness Through Data/Prediction Manipulations
  - Individual Fairness
  - Optimized Pre-processing
  - Learning to Defer
- Fair NLP
  - Biases in NLP Models
  - Data Augmentation
  - Debiasing Word Embedding
  - Adversarial Learning

Word Embeddings

- An Essential Part of Deep NLP Models
  - Classifications (e.g., Sentiment Analysis)
  - Text Generation (e.g., translation, summarization)
  - Text Retrieval (e.g., Question Answering)
  - Visual-Language Representations (e.g., Image Captioning)



Discrete Space

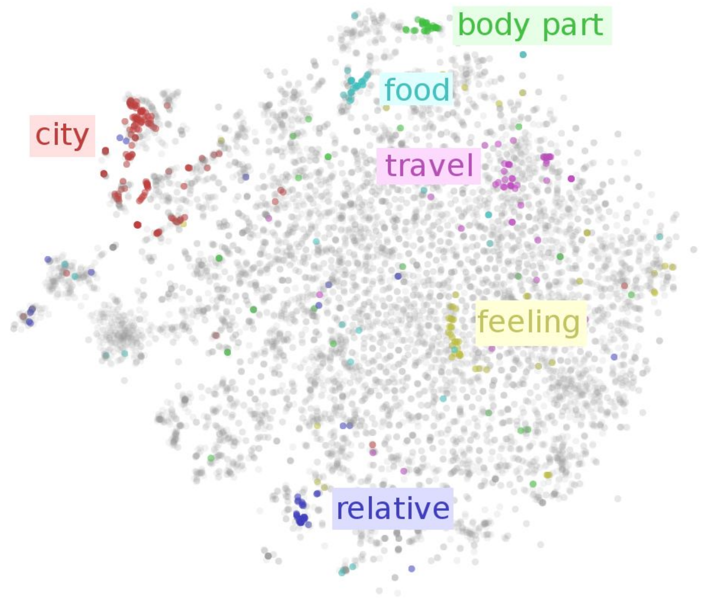
Continuous Space

Neural Networks

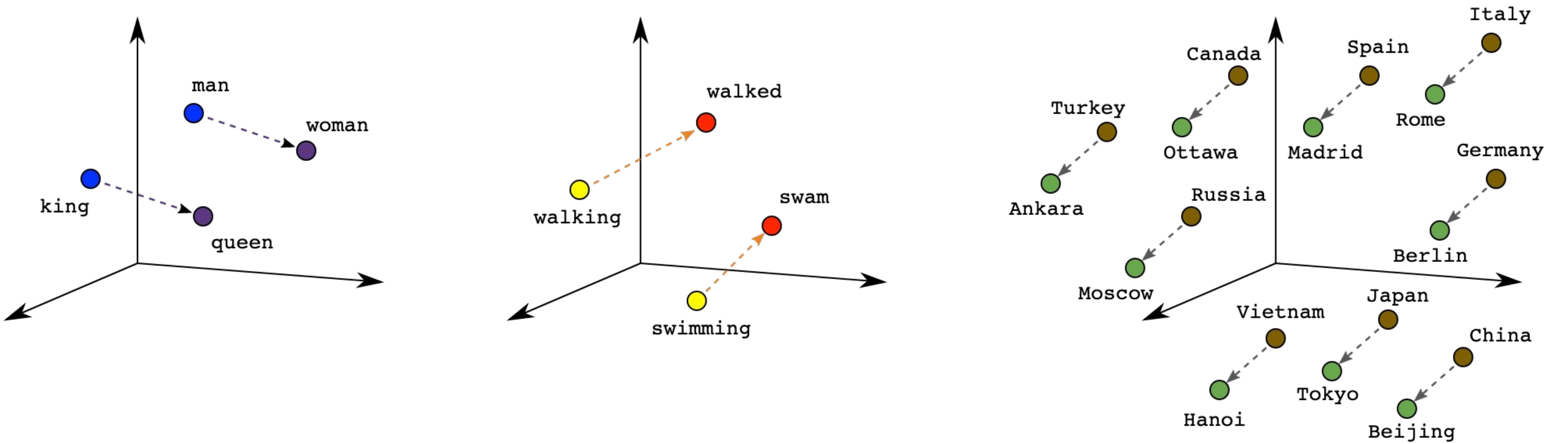
Word Embeddings

- Embedding Techniques
  - GloVe (Pennington et al, 2014)
  - Word2Vec (Rong et al, 2014)

- Trained Through A Proxy Task
  - Word proximity (GloVe)
  - SkipGram (Word2Vec)

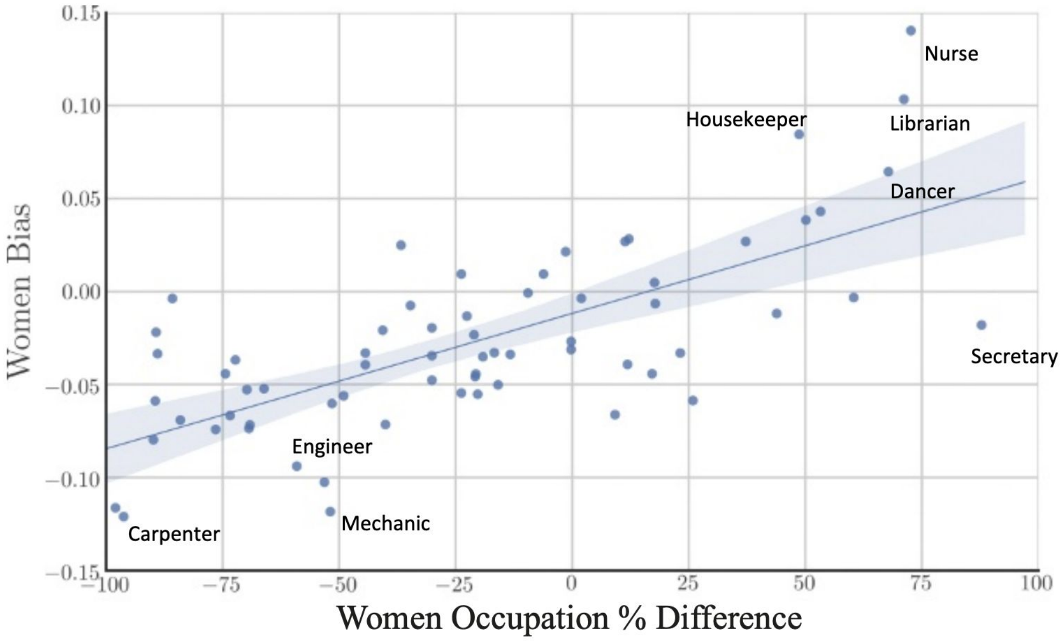


Geometric Properties of Word Embeddings



Male-Female Verb Tense Country-Capital

Can Word Embedding Be Biased?



[Garga et al, 2017](#)

Types of Gender Associations

- Definitional Gender Associations

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

- Stereotypical Gender Associations

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

Bolukbasi et al, 2016

Definitional and Stereotypical Associations

Gender stereotype *she*-*he* analogies.

sewing-carpentry	register-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	hairdresser-barber

Gender appropriate *she*-*he* analogies.

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

Bolukbasi et al, 2016

Gender Subspace

$$\overrightarrow{\text{grandmother}} - \overrightarrow{\text{grandfather}} = \overrightarrow{\text{gal}} - \overrightarrow{\text{guy}} = g$$

$$\overrightarrow{\text{she}} - \overrightarrow{\text{he}}$$
$$\overrightarrow{\text{her}} - \overrightarrow{\text{his}}$$
$$\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$$
$$\overrightarrow{\text{Mary}} - \overrightarrow{\text{John}}$$
$$\overrightarrow{\text{herself}} - \overrightarrow{\text{himself}}$$

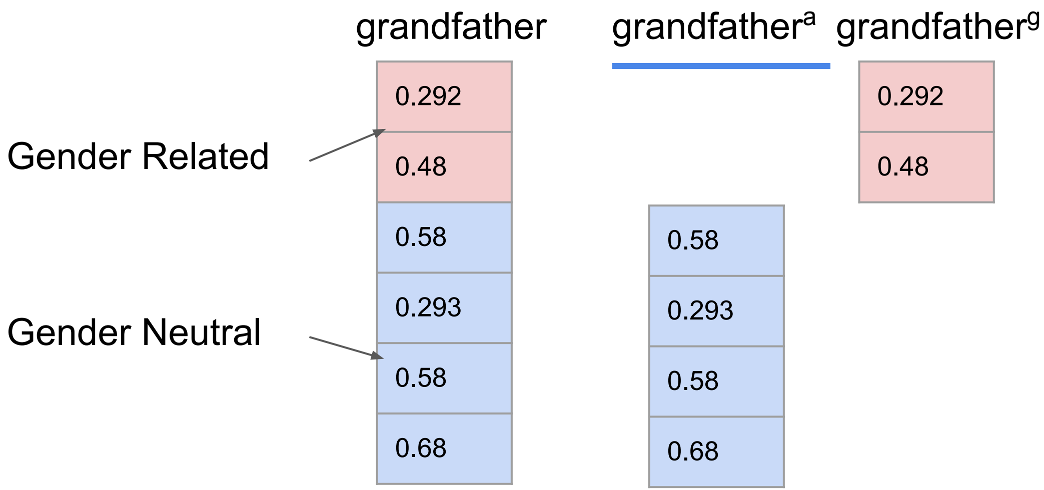
$$\overrightarrow{\text{daughter}} - \overrightarrow{\text{son}}$$
$$\overrightarrow{\text{mother}} - \overrightarrow{\text{father}}$$
$$\overrightarrow{\text{gal}} - \overrightarrow{\text{guy}}$$
$$\overrightarrow{\text{girl}} - \overrightarrow{\text{boy}}$$
$$\overrightarrow{\text{female}} - \overrightarrow{\text{male}}$$

Bolukbasi et al, 2016

Gender-Neutral Word Embeddings

- Decompose Word Embeddings Into Gender-Related and Gender-Neural Parts

$$w = [w^{(a)}; w^{(g)}]$$



[Zhao et al, 2018]

Gender-Neutral Word Embeddings

- Fine-tuning Word Embeddings Using Debiasing Regularizers

$$J = \underbrace{J_G}_{\text{Glove Loss Function}} + \underbrace{\lambda_d J_D}_{\text{Regulate Gender-related Words}} + \underbrace{\lambda_e J_E}_{\text{Regulate All Other Words}}$$

Ω_F Female Seed Words

Ω_N All Other Words

Ω_M Male Seed Words

Zhao et al, 2018

Gender-Neutral Word Embeddings

- Fine-tuning Word Embeddings Using Debiasing Regularizers

$$J = J_G + \lambda_d \underline{J_D} + \lambda_e J_E$$

Regulate
Gender-related
Words

Ω_F
Female Seed Word

Push Toward Extremes
On Gender Dimensions

Ω_M
Male Seed Word

$w^{(g)}$ - Gender-related Components
$w^{(a)}$ - Gender-neutral Components

$$J_D = -\left\| \sum_{w \in \Omega_M} w^{(g)} - \sum_{w \in \Omega_F} w^{(g)} \right\|_1 + \sum_{w \in \Omega_M} \left\| 1 - w^{(g)} \right\|_2^2 + \sum_{w \in \Omega_F} \left\| -1 - w^{(g)} \right\|_2^2$$

[Zhao et al, 2018]()

Gender-Neutral Word Embeddings

- Fine-tuning Word Embeddings Using Debiasing Regularizers

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

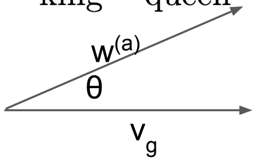
Regulate All Other Words

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}}$$

$$\overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$v_g = \frac{1}{|\Omega'|} \sum_{(w_m, w_f) \in \Omega_M, \Omega_F} \left(w_m^{(a)} - w_f^{(a)}\right)$$

Gender Subspace

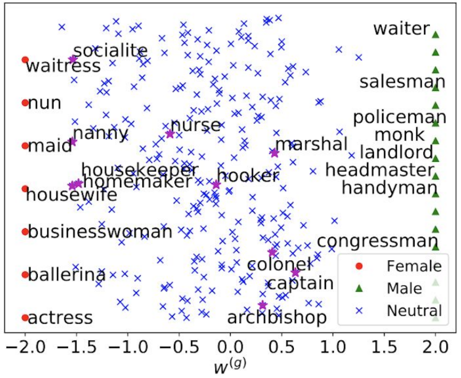


$$J_E = \sum_{w \in \Omega_N} \left(v_g^T w^{(a)}\right)^2$$

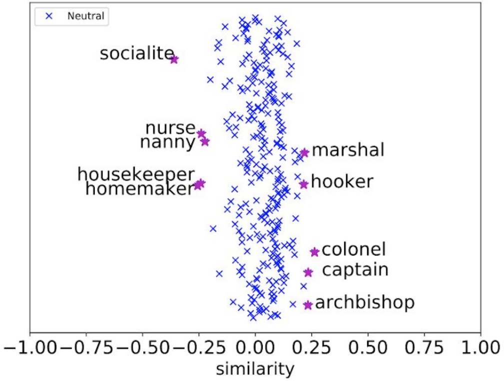
$w^{(g)}$ - Gender-related Components
$w^{(a)}$ - Gender-neutral Components

[Zhao et al, 2018]

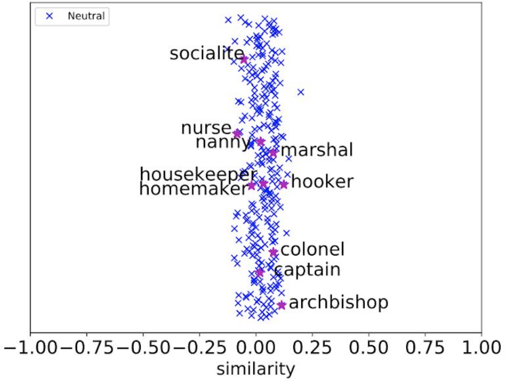
Gender Attribute Separation



w$^{(g)}$ of All Occupations



w$^{(a)}$ of GloVe for Gender Neutral Occupations



w$^{(a)}$ of Gender-Neutral GloVE for Gender Neutral Occupations

w$^{(g)}$ - Gender-related Components
w$^{(a)}$ - Gender-neutral Components

Gender Relational Analogy

Question 1: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. What relation best describes these X:Y word pairs?

(1) "X worships/reveres Y"
(2) "X seeks/desires/aims for Y"
(3) "X harms/destroys Y"
(4) "X uses/exploits/employs Y"

Dataset	Embeddings	Definition	Stereotype	None
SemBias	GloVe	80.2	10.9	8.9
	GN-GloVe	97.7	1.4	0.9
SemBias (subset)	GloVe	57.5	20	22.5
	GN-GloVe	75	15	10

Jurgens et al , 2012

Coreference Resolution

Embeddings	OntoNotes-test	PRO	ANTI	Avg	Diff
GloVe	66.5	76.2	46.0	61.1	30.2
GN-GloVe	66.2	72.4	51.9	62.2	20.5
GN-GloVe(w_a)	65.9	70.0	53.9	62.0	16.1

$w^{(a)}$ - Gender-neutral Components

Jurgens et al , 2012

Outline

- Fairness Through Data/Prediction Manipulations
  - Individual Fairness
  - Optimized Pre-processing
  - Learning to Defer
- Fair NLP
  - Biases in NLP Models
  - Data Augmentation
  - Debiasing Word Embedding
  - Adversarial Learning

Summary

- Optimized Pre-processing for Fairness
    - Optimizes several fairness criteria (Demographic Parity, Individual Fairness) at the same time
    - Transform data to meet criteria
- Post-processing Techniques for Fairness
    - Learning to Defer
    - Fix biased predictions from the decision-maker
    - Take advantage of high performance of the decision-maker model
- Word Debiasing
    - Separate gender specific and gender neutral embeddings
- Data Augmentation
    - Gender Swapping
- Adversarial Learning

Reading Assignments

- Gonen, Hila, and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them, NAACL 2019
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Contextualized Word Embeddings, NAACL 2019
- Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the Origins of Bias in Word Embeddings, ICML 2019
- Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation, EMNLP 2019
- Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection, ACL 2019

Next Lecture

Fair Visual Representations