NLP and Society:
Towards Socially Responsible NLP

Vinodkumar Prabhakaran
Research Scientist
Google

What’s in this talk...

- Motivation for Machine Learning (ML) Fairness research
- Why and how ML models may be unfair
- Fairness issues in ML-based Natural Language Processing
- What can/should we do?

What’s NOT in this talk...

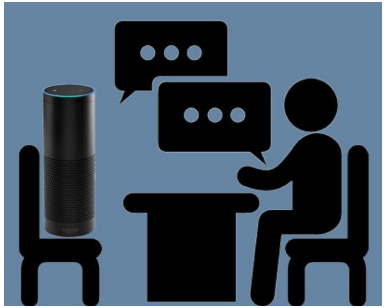
- Definitive answers to fairness/ethical questions
- Prescriptive solutions to fix ML/NLP (un)fairness
- Focus on research done by myself, my team, or Google.
- ...

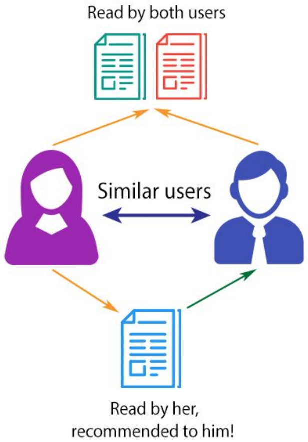
What’s also in this talk...

- Research done in academia and various industry labs
- Research from other disciplines, including Psychology, Philosophy, and Social Sciences in general …
- Uncomfortable impacts of technology on society

Machine Learning is Everywhere!!!

Machine Learning is Everywhere!!!





Slide credit: Yulia Tsvetkov

"It's true that they can follow instructions at superhuman speed, with superhuman fidelity and over unimaginable quantities of data. ***But these instructions don't come from nowhere****. Although neural networks might be said to write their own programs, they do so towards* ***goals set by humans, using data collected for human purposes****. If the data is skewed, even by accident, the computers will amplify injustice."*

— The Guardian

Slide credit: lovejoy@

"It's true that they can follow instructions at superhuman speed, with superhuman fidelity and over unimaginable quantities of data. ***But these instructions don't come from nowhere****. Although neural networks might be said to write their own programs, they do so towards g****oals set by humans, using data collected for human purposes****. If the data is skewed, even by accident, the computers will* ***amplify injustice****."*

— The Guardian

Slide credit: lovejoy@

Fairness in Machine Learning

A Few Case Studies

Photo captioning

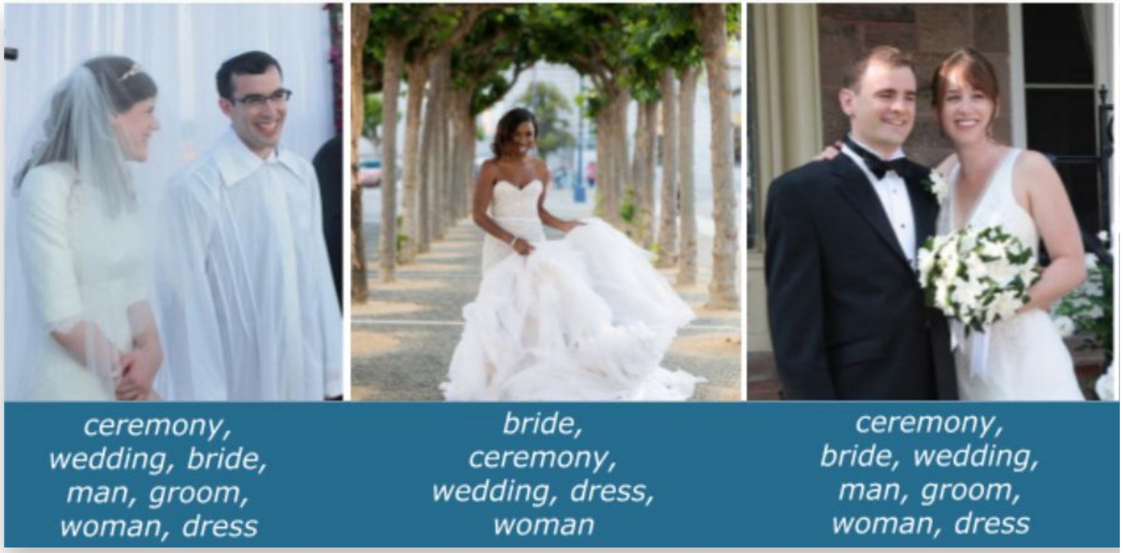
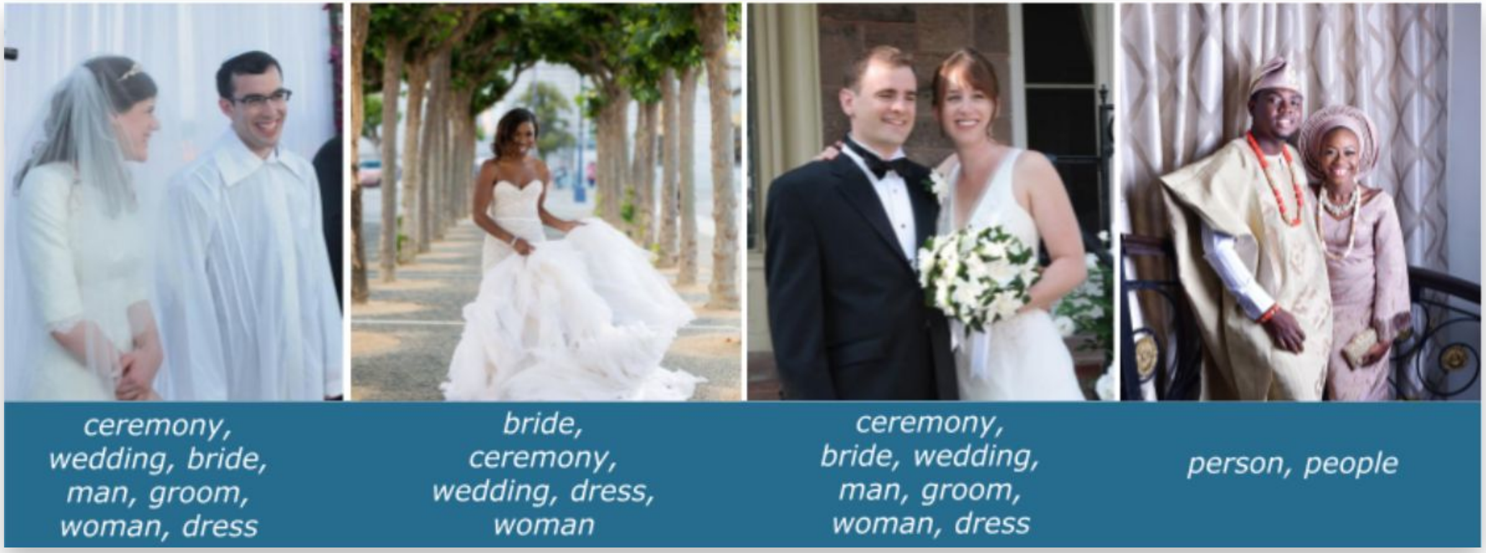
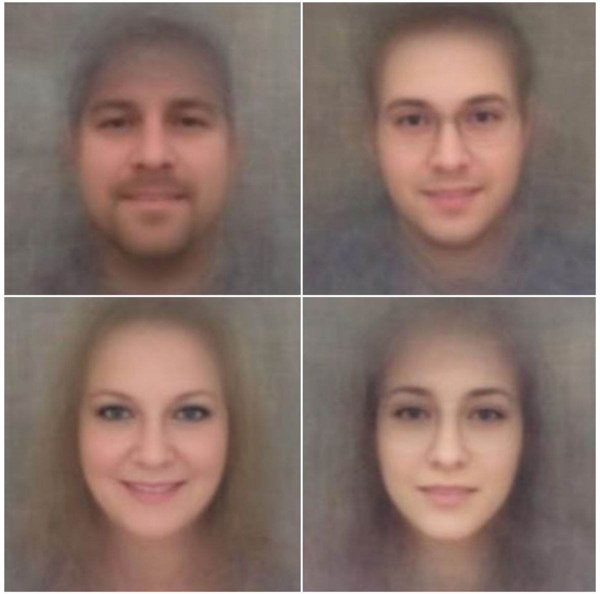


Photo captioning



Predicting Sexual Orientation



Original Paper: “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images” Wang and Kosinsky, 2017. PsyArXiv

Predicting Sexual Orientation

“Differences between lesbian or gay and straight faces in selfies relate to grooming, presentation, and lifestyle— **that is, differences in culture, not in facial structure.**”

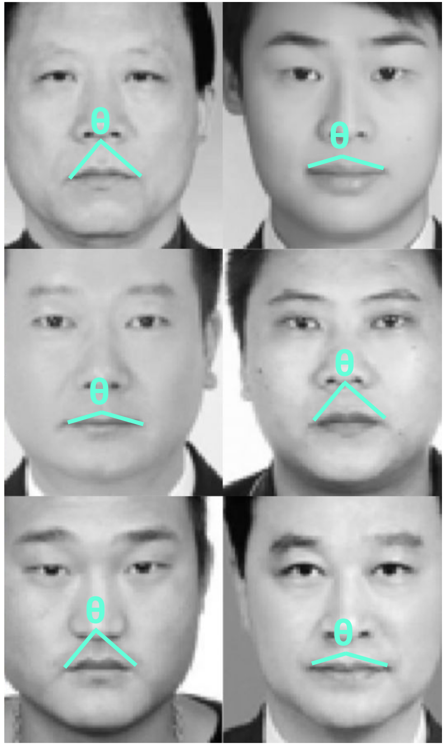
“Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?” Medium, Blaise Agüera y Arcas, Alexander Todorov and Margaret Mitchell

Predicting criminality

“Automated Inference on Criminality using Face Images”
Wu and Zhang, 2016. arXiv

Predicting criminality

“[...] angle **θ** from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ...”



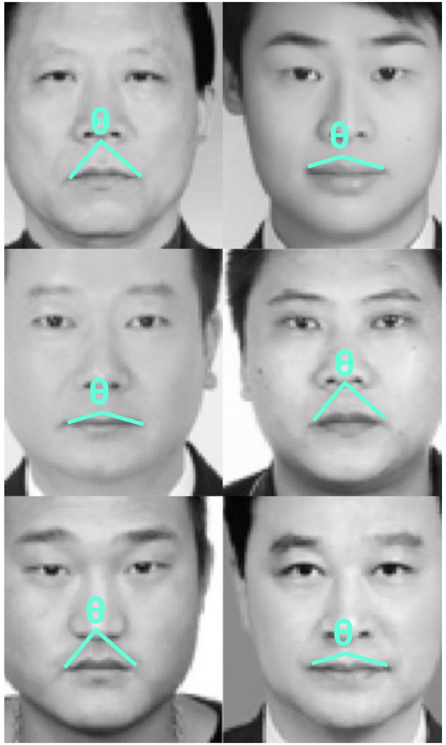
“Automated Inference on Criminality using Face Images”
Wu and Zhang, 2016. [arXiv]

Predicting criminality: physiognomy?

"[...] angle **θ** from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ..."

Physiognomy's New Clothes (Medium Blog Post) - by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov

> *"Deep learning based on superficial features is decidedly not a tool that should be deployed to "accelerate" criminal justice; attempts to do so will instead perpetuate injustice."*



"Automated Inference on Criminality using Face Images"
Wu and Zhang, 2016. arXiv

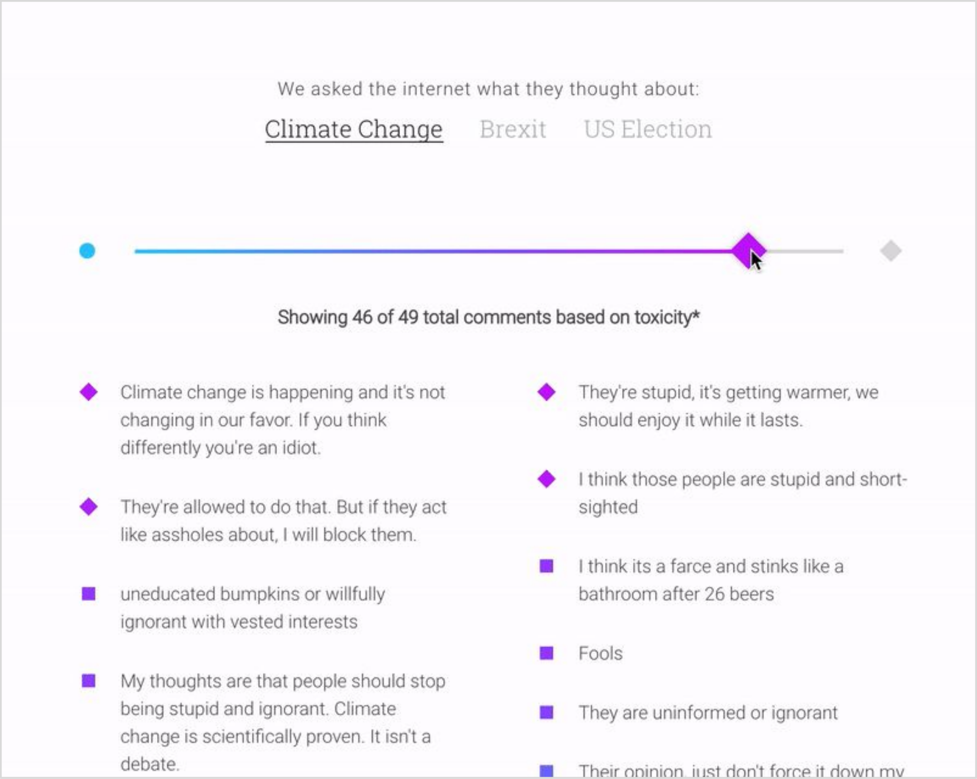
Toxicity Classification

Jigsaw

theguardian

WIKIPEDIA

The Economist

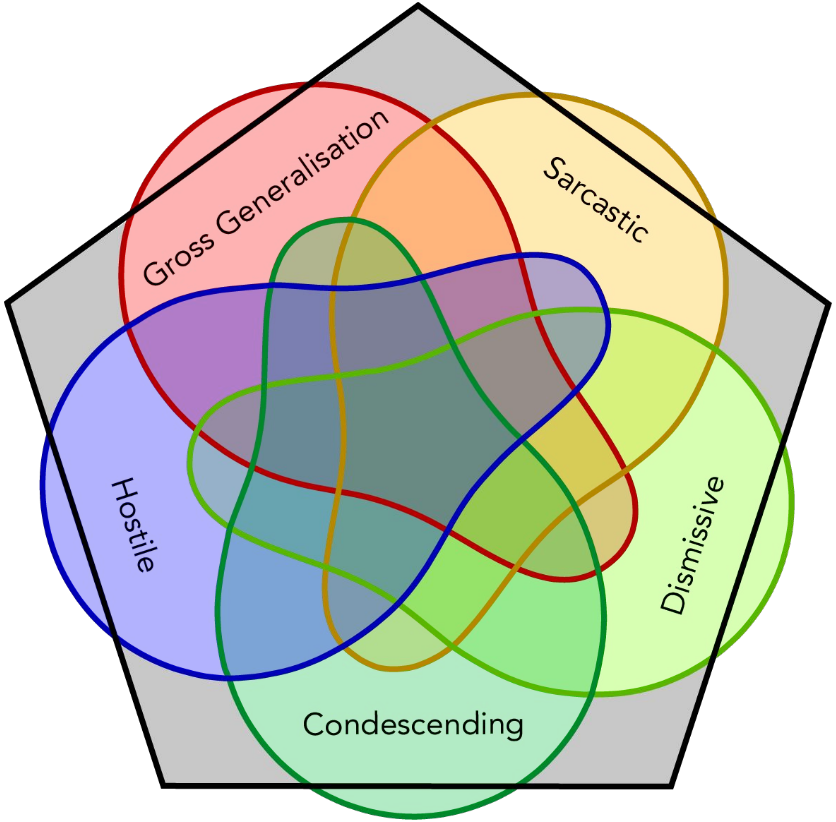


Source
perspectiveapi.com

Toxicity Classification

The False Positive

Toxic is defined as... "**a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.**"



Source
"The Challenge of Identifying Subtle Forms of Toxicity Online" - Jigsaw
https://medium.com/the-false-positive/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

Toxicity Classification

A naively trained model will have **strong unintended biases** as illustrated by these examples:

Comment	Toxicity Score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam	0.46

- "The Challenge of Identifying Subtle Forms of Toxicity Online". Jigsaw. The False Positive (2018).

Toxicity Classification

A naively trained model will have **strong unintended biases** as illustrated by these examples:

Comment	Toxicity Score
I hate Justin Timberlake.	0.90
I hate Rihanna.	0.69

- Prabhakaran et al. (2019). "Perturbation Sensitivity Analysis to Detect Unintended Model Biases" EMNLP 2019

Toxicity Classification

A naively trained model will have **strong unintended biases** as illustrated by these examples:

Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Toxicity Classification

A naively trained model will have **strong unintended biases** as illustrated by these examples:

Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Toxicity Classification

A naively trained model will have **strong unintended biases** as illustrated by these examples:

Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44
I am a person with mental illness.	0.62

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Allocative Harm

"when a system allocates or withholds a certain opportunity or resource"

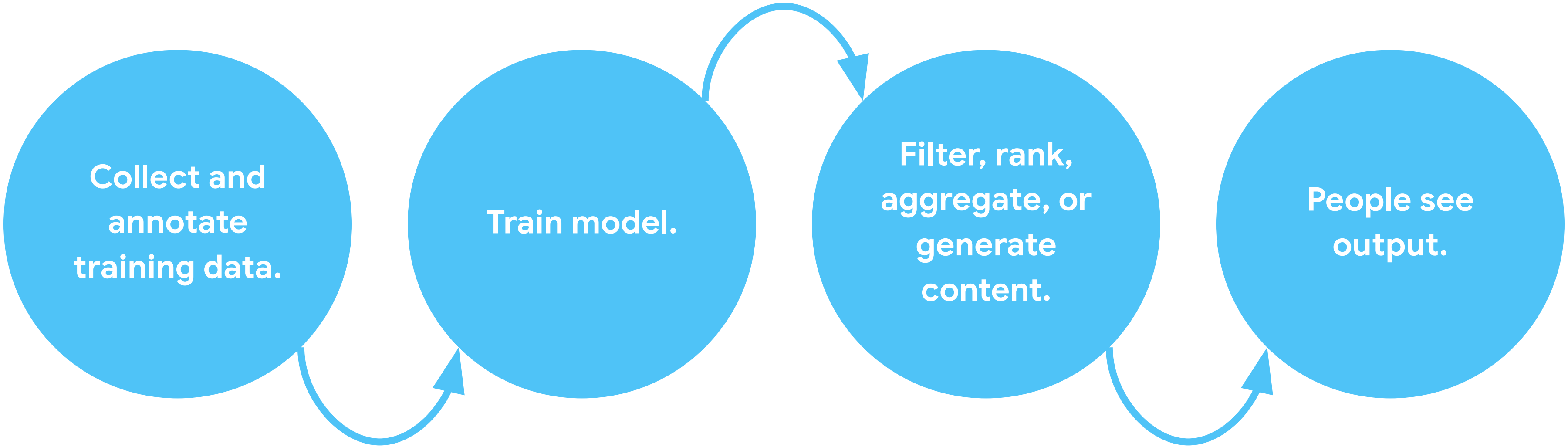
Associative Harm

"when systems reinforce the subordination of some groups along the lines of identity"

Source: Kate Crawford, The Trouble with Bias, NIPS 2017

Why do these things happen?

Machine Learning “sequence”



Google

Potential biases

Collect and annotate training data.

Human Biases in Data

Reporting bias

Selection bias

Overgeneralization

Out-group homogeneity bias

Unconscious bias from "the world" that we might reflect in ML when using some of the world's data

Human Biases in Collection and Annotation

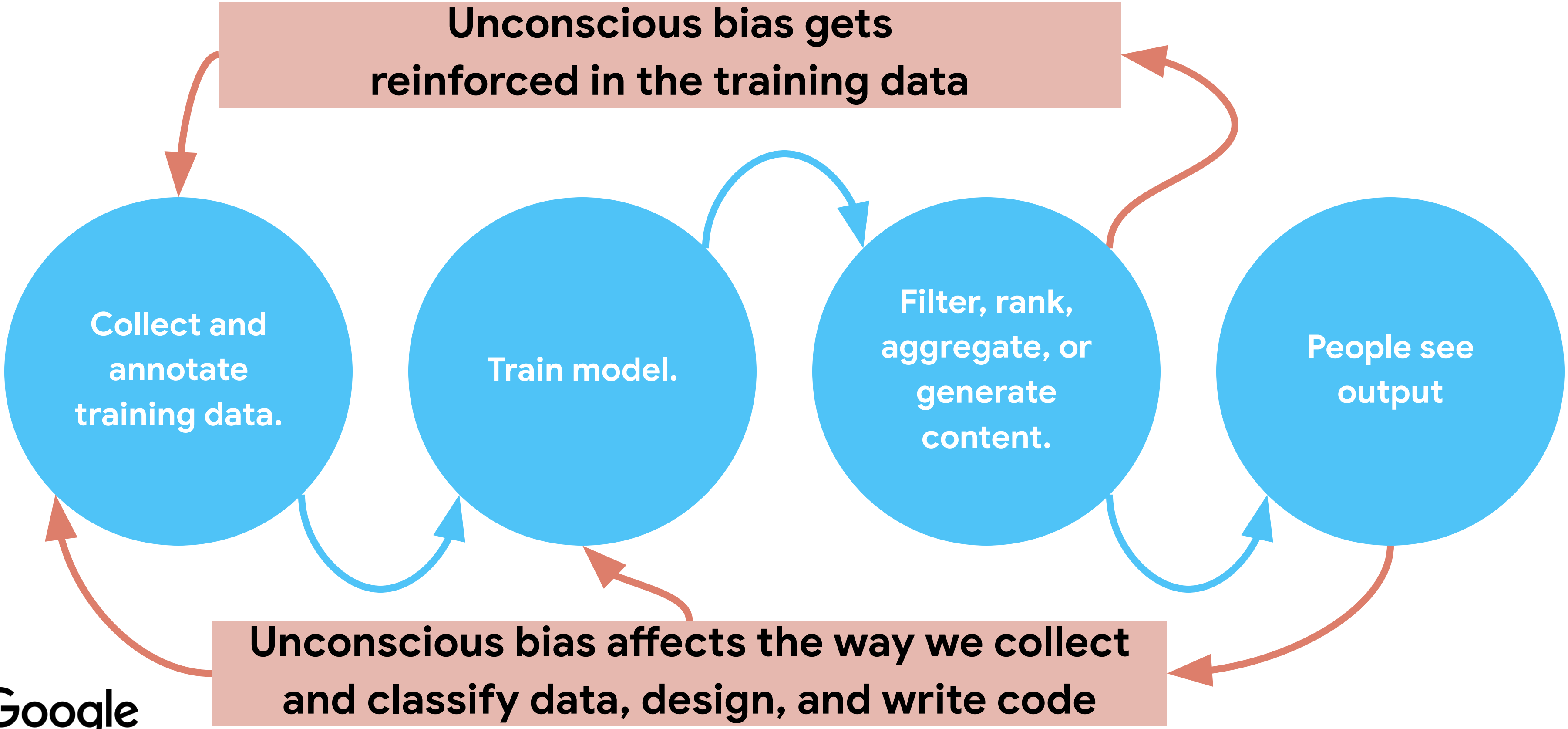
Confidence bias / Overconfidence effect

Confirmation bias

Experimenter's bias

Unconscious bias in our procedures that we might reflect in our ML

Unconscious bias interferes



Google

Fairness in Natural Language Processing

A Deeper Dive

The common misconception is that language has to do with **words** and what they mean.

It doesn’t.

It has to do with people and what they mean.

Herbert H. Clark & Michael F. Schober, 1992

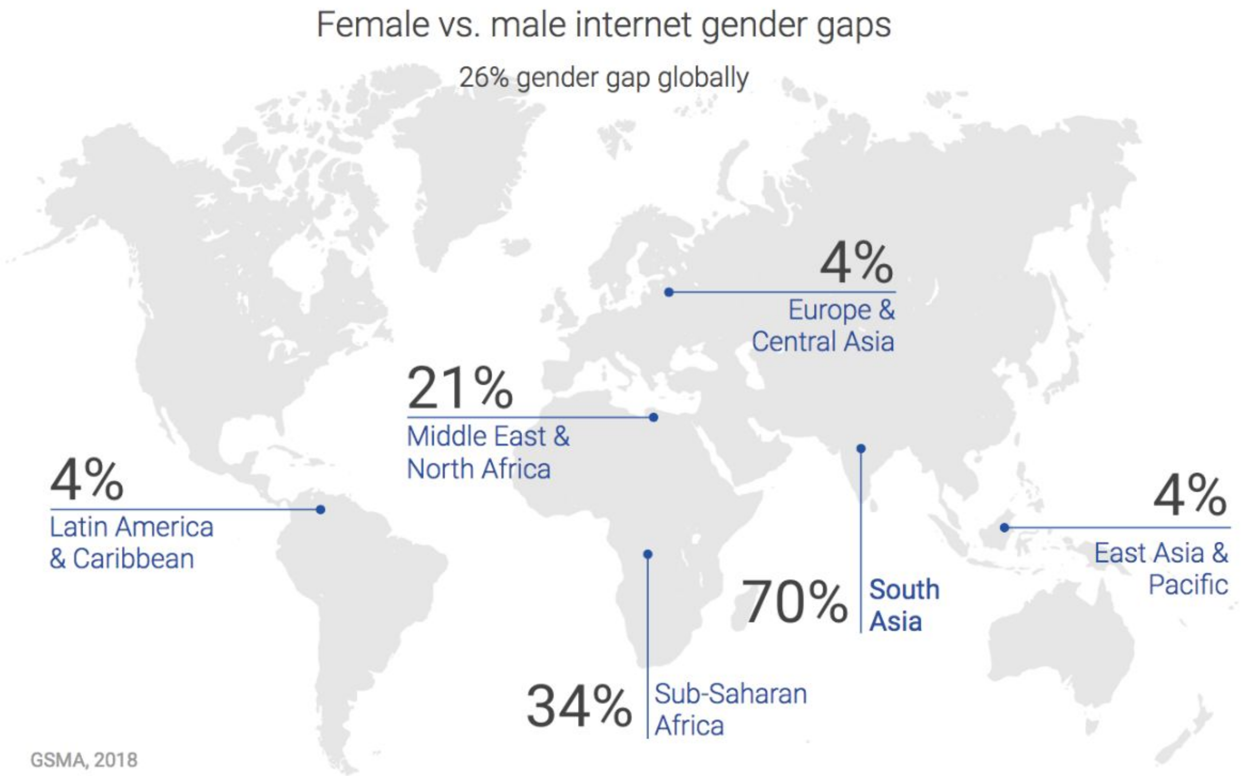
Fairness in Natural Language Processing

A Deeper Dive

- Is my data biased?

Selection Bias: World Englishes

Selection Bias: Gender Equity



Selection Bias: Gender Equity

- Men are over-represented in web-based news articles

(Jia, Lansdall-Welfare, and Cristianini 2015)

- Men are over-represented in twitter conversations

(Garcia, Weber, and Garimella 2014)

- Gender bias in Wikipedia and Britannica

(Reagle & Rhuee 2011)

A case study:
Language Identification

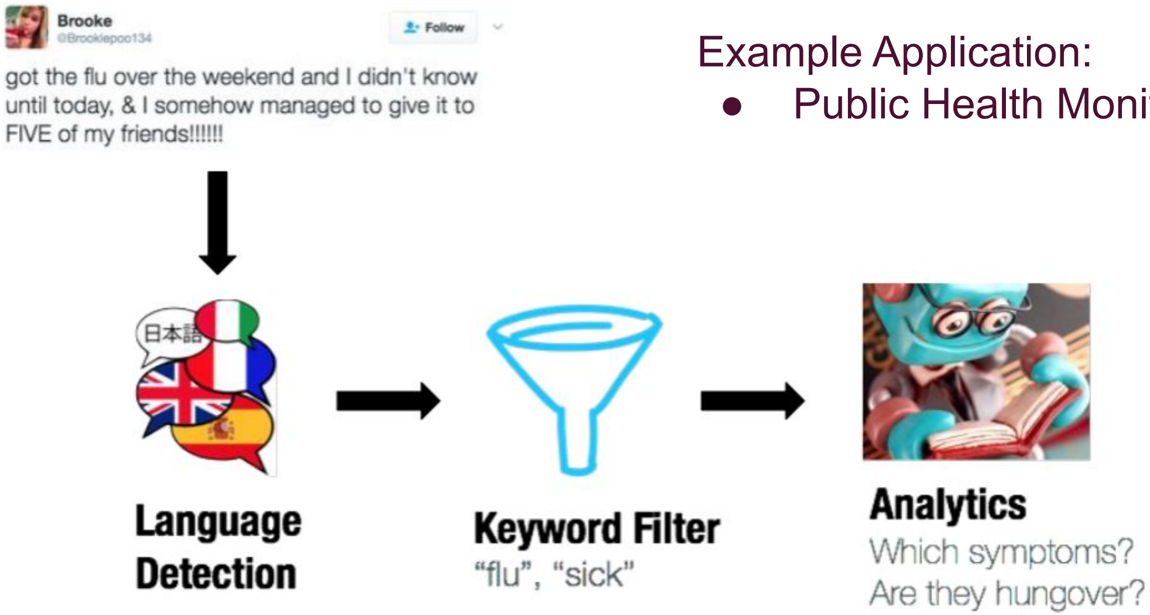
Sampling Bias in Language Identification (LID)

- Most NLP applications employ off-the-shelf LID systems as the first step

Slide credit: David Jurgens
(Jurgens et al. ACL'17)

Sampling Bias in Language Identification (LID)

- Most NLP applications employ off-the-shelf LID systems as the first step



Example Application:

- Public Health Monitoring

Slide credit: David Jurgens

(Jurgens et al. ACL'17)

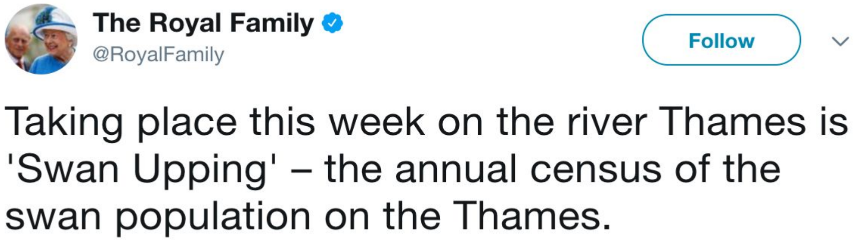
How well do LID systems do?

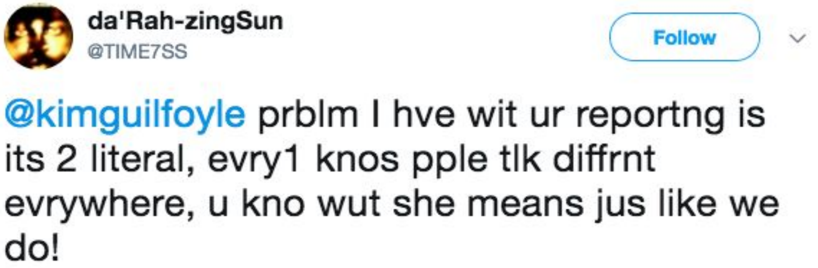
"This paper describes […] how even the most simple of these methods **using data obtained from the World Wide Web achieve accuracy approaching 100%** on a test suite comprised of ten European languages"

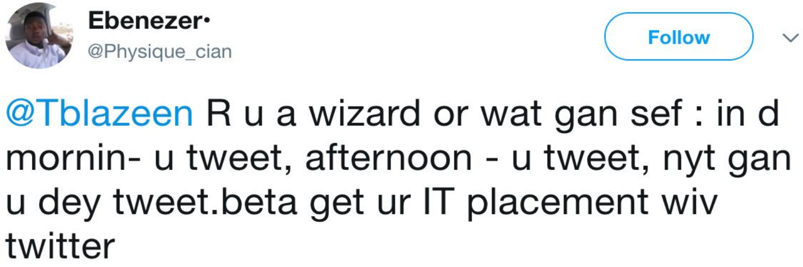
McNamee, P., "Language identification: *a solved problem* suitable for undergraduate instruction" Journal of Computing Sciences in Colleges 20(3) 2005.

World Englishes

World Englishes







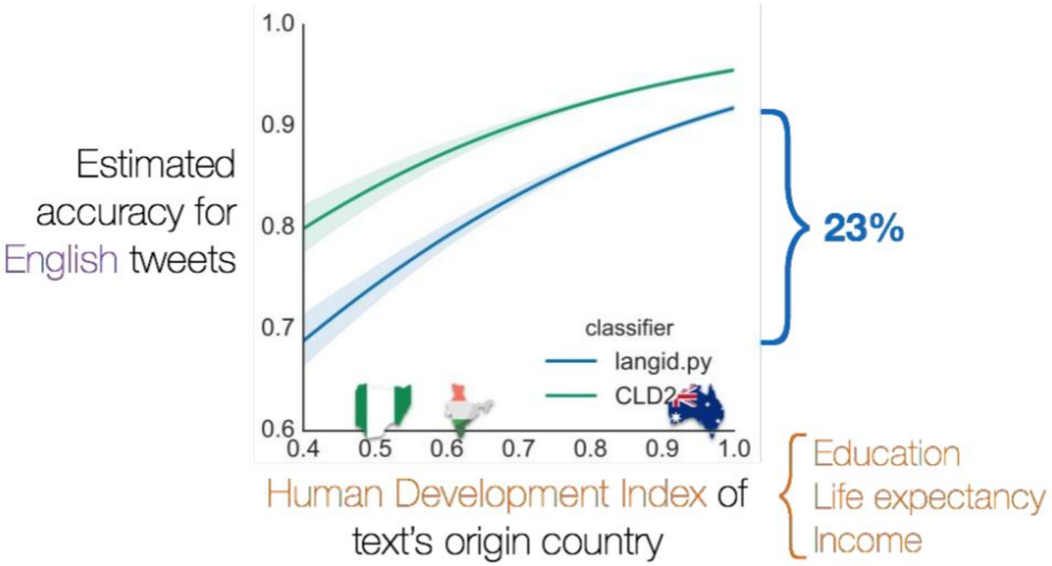
- Language identification degrades significantly on African American Vernacular English (Blodgett et al. 2016)

LID Usage Example: Public Health Monitoring

Slide credit: David Jurgens
(Jurgens et al. ACL’17)

Socioeconomic Bias in Language Identification

- Off-the-shelf LID systems under-represent populations in less-developed countries



1M geo-tagged Tweets with any of 385 English terms from established lexicons for *influenza*, *psychological well-being*, and *social health*

Slide credit: David Jurgens
(Jurgens et al. ACL'17)

i.e.

people who are the most marginalized,
people who'd benefit the most from such technology,
are also the ones who are more likely to be
systemically excluded from this technology

Better Social Representation through Network-based Sampling

- Re-sampling from strategically-diverse corpora

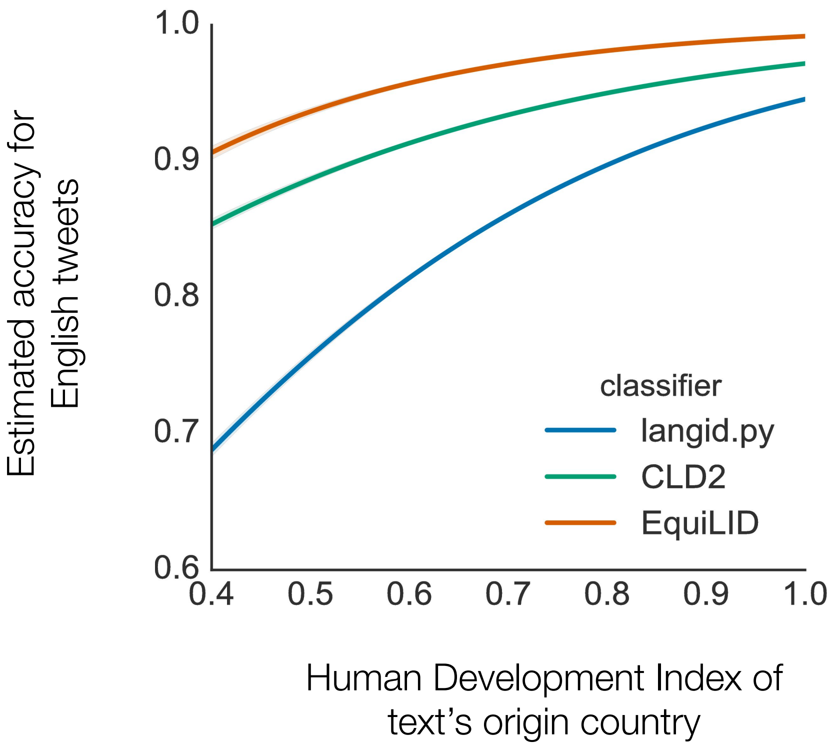
Topical

Geographic

Social

Multilingual

Slide credit: David Jurgens
(Jurgens et al. ACL’17)



Slide credit: David Jurgens

(Jurgens et al. ACL'17)

Fairness in Natural Language Processing

A Deeper Dive

- Is my data biased?
- Is my model biased?

Bias in NLP Mo…

Slide from SRNLP Tutorial at NAACL 2018

1. Bolukbasi T… …aligrama V., Kalai A. (2016) **Man is to … …to Homemaker? Debiasing Word Em…**

2. …liskan, …, Brys…, J. J. and Narayanan, A. (2017) **Semantics derived automatically from language corpora contain human-like biases.** *Science*

3. Nikhil Garg, Londa Schiebinger, Dan Jurafsky, James Zou. (2018) **Word embeddings quantify 100 years of gender and ethnic stereotypes.** *PNAS.*

Bias in NLP Models

1. Bolukbasi et al. **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS* (2016)
2. Caliskan, et al. **Semantics derived automatically from language corpora contain human-like biases.** *Science* (2017)
3. Garg et al. **Word embeddings quantify 100 years of gender and ethnic stereotypes.** *PNAS.* (2018)
4. Zhao, Jieyu, et al. **Men also like shopping: Reducing gender bias amplification using corpus-level constraints.** *arXiv* (2017)

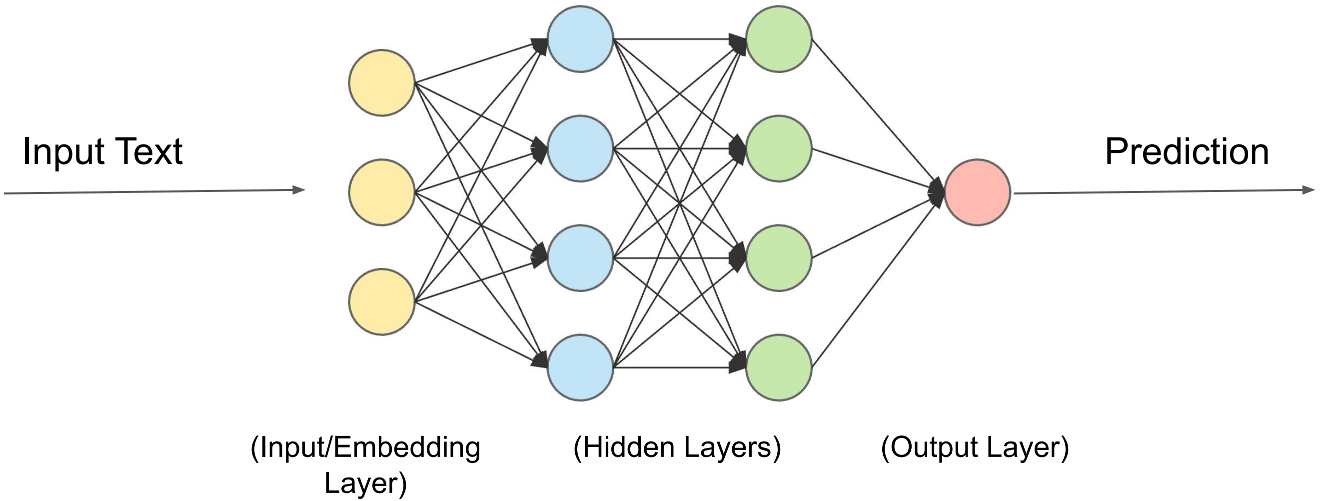
2018

5. Zhao, Jieyu, et al. **Gender bias in coreference resolution: Evaluation and debiasing methods.** *arXiv* (2018)
6. Zhang, et al. **Mitigating unwanted biases with adversarial learning.** *AIES*, 2018
7. Webster, Kellie, et al. **Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns.** *TACL* (2018)
8. Svetlana and Mohammad. **Examining gender and race bias in two hundred sentiment analysis systems.** *arXiv* (2018)
9. Díaz, et al. **Addressing age-related bias in sentiment analysis.** *CHI Conference on Human Factors in Computing Systems*. (2018)
10. Dixon, et al. **Measuring and mitigating unintended bias in text classification.** *AIES.* (2018)
11. Prates, et al. **Assessing gender bias in machine translation: a case study with Google Translate.** *Neural Computing and Applications* (2018)
12. Park, et al. **Reducing gender bias in abusive language detection.** *arXiv* (2018)
13. Zhao, Jieyu, et al. **Learning gender-neutral word embeddings.** *arXiv* (2018)
14. Anne Hendricks, et al. **Women also snowboard: Overcoming bias in captioning models.** *ECCV*. (2018)
15. Elazar and Goldberg. **Adversarial removal of demographic attributes from text data.** *arXiv* (2018)
16. Hu and Strout. **Exploring Stereotypes and Biased Data with the Crowd.** *arXiv* (2018)

2019

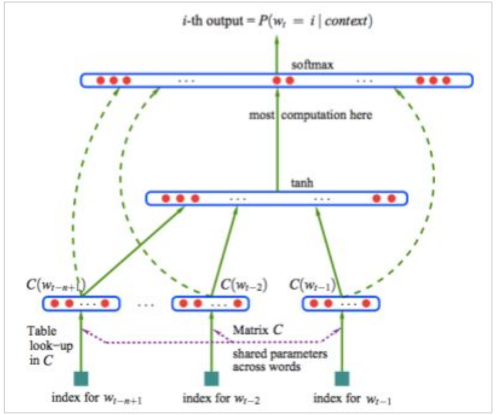
17. Swinger, De-Arteaga, et al. **What are the biases in my word embedding?** *AIES* (2019)
18. De-Arteaga et al. **Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting.** *FAT** (2019)
19. Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.** NAACL (2019).
20. Manzini et al. **Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings.** NAACL (2019).
21. Sap et al. **The Risk of Racial Bias in Hate Speech Detection**. ACL (2019)
22. Stanovsky et al. **Evaluating Gender Bias in Machine Translation**. ACL (2019)
23. Garimella et al. **Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing**. ACL (2019)
24. …

Where to look for biases?

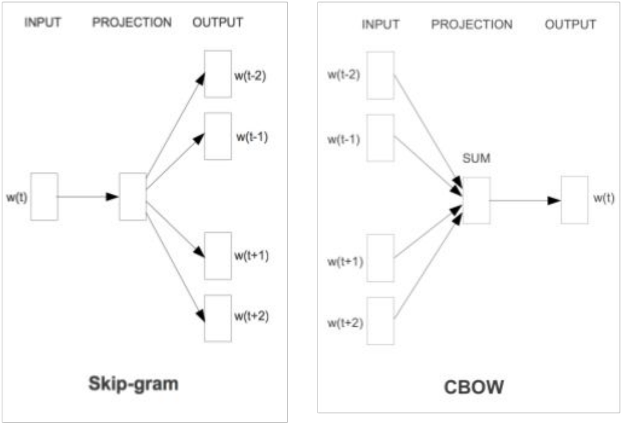


Bias in Input Representations?

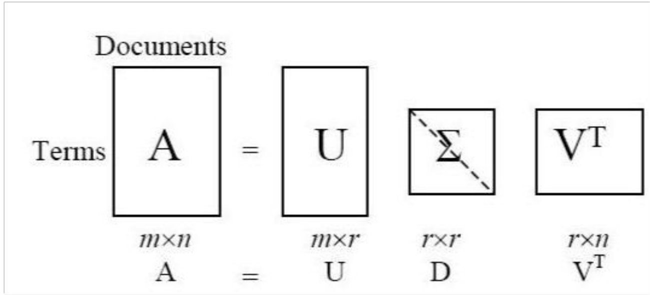
Input Representation: Word Embeddings



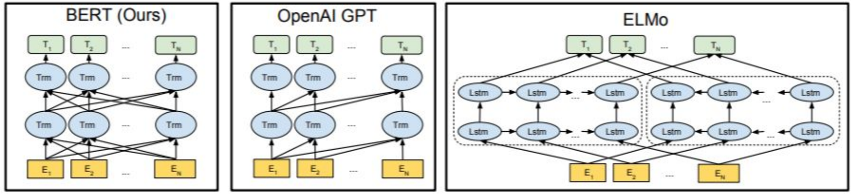
Neural Language Model (Bengio et al, `03)



word2vec (Mikolov et al, `03)



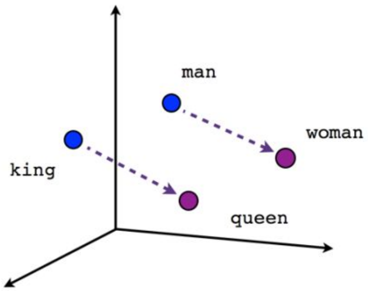
Latent Semantic Analysis
(Deerwester et al, `90, Turney & Pantel `10)



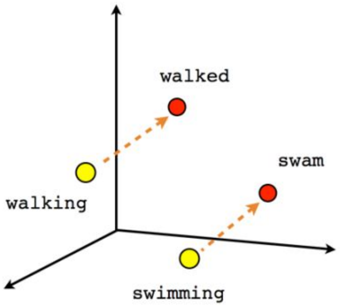
BERT, GPT/GPT-2, ELMo
(Devlin et al. ‘19, Radford et al. ‘18, Peters et al. ‘18)

Word Analogy Tasks

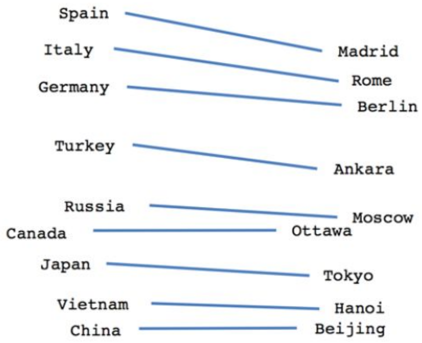
- Mikolov et al. ‘13



Male-Female



Verb tense



Country-Capital

$$\min cos(\overrightarrow{man} - \overrightarrow{woman}, \overrightarrow{king} - x) \ s.t. \ ||king - x||_2 < \delta$$

Social Stereotypes → Word Embeddings?

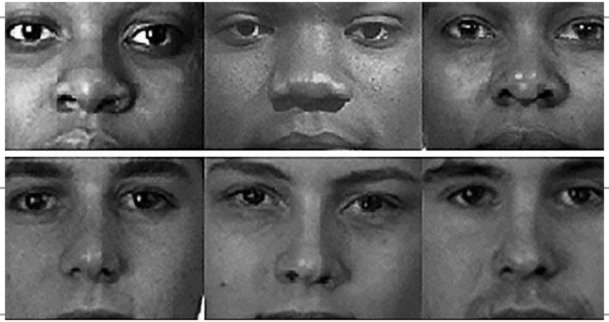
Biases in NLP Representations

- Bolukbasi et al. **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS* (2016)
- Caliskan, et al. **Semantics derived automatically from language corpora contain human-like biases.** *Science* (2017)
- Garg et al. **Word embeddings quantify 100 years of gender and ethnic stereotypes.** *PNAS.* (2018)
- Swinger, De-Arteaga, et al. **What are the biases in my word embedding?** *AIES* (2019)
- Manzini et al. **Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings.** NAACL (2019).
- ...

Implicit bias in humans?

Implicit Association Test - Greenwald et al. 1998

Category	Items
Good	Spectacular, Appealing, Love, Triumph, Joyous, Fabulous, Excitement, Excellent
Bad	Angry, Disgust, Rotten, Selfish, Abuse, Dirty, Hatred, Ugly
African Americans	
European Americans	



Implicit Association Test

The IAT involves making repeated judgments (by pressing a key on a keyboard) to label words or images that pertain to one of two categories presented simultaneously (e.g., categorizing pictures of African American or European American and categorizing positive/negative adjectives).

The test compares response times when different pairs of categories share a response key on keyboard
(e.g., African American + GOOD vs African American + BAD vs European American + GOOD vs European American + BAD)

IAT - Societal groups⇔Stereotype words

Disability IAT — *Disability* **('Disabled - Abled' IAT).** This IAT requires the ability to recognize symbols representing abled and disabled individuals.

Asian IAT — *Asian American* **('Asian - European American' IAT).** This IAT requires the ability to recognize White and Asian-American faces, and images of places that are either American or Foreign in origin.

Sexuality IAT — *Sexuality* **('Gay - Straight' IAT).** This IAT requires the ability to distinguish words and symbols representing gay and straight people. It often reveals an automatic preference for straight relative to gay people.

Arab-Muslim IAT — *Arab-Muslim* **('Arab Muslim - Other People' IAT).** This IAT requires the ability to distinguish names that are likely to belong to Arab-Muslims versus people of other nationalities or religions.

Age IAT — *Age* **('Young - Old' IAT).** This IAT requires the ability to distinguish old from young faces. This test often indicates that Americans have automatic preference for young over old.

Skin-tone IAT — *Skin-tone* **('Light Skin - Dark Skin' IAT).** This IAT requires the ability to recog skinned faces. It often reveals an automatic preference for light-skin relative to da

Race IAT — *Race* **('Black - White' IAT).** This IAT requires the ability to distinguish faces of African origin. It indicates that most Americans have an automatic preference for

https://implicit.harvard.edu/implicit/selectatest.html

Greenwald et al. 1998

Religion IAT — *Religion* **('Religions' IAT).** This IAT requires some familiarity with religious terms from various world religions.

Native IAT — *Native American* **('Native - White American' IAT).** This IAT requires the ability to recognize White and Native American faces in either classic or modern dress, and the names of places that are either American or Foreign in origin.

Gender-Science IAT — *Gender - Science***.** This IAT often reveals a relative link between liberal arts and females and between science and males.

Gender-Career IAT — *Gender - Career***.** This IAT often reveals a relative link between family and females and between career and males.

Presidents IAT — *Presidents* **('Presidential Popularity' IAT).** This IAT requires the ability to recognize photos of Donald Trump and one or more previous presidents.

Weight IAT — *Weight* **('Fat - Thin' IAT).** This IAT requires the ability to distinguish faces of people who are obese and people who are thin. It often reveals an automatic preference for thin people relative to fat people.

Weapons IAT — *Weapons* **('Weapons - Harmless Objects' IAT).** This IAT requires the ability to recognize White and Black faces, and images of weapons or harmless objects.

Can we apply this to NLP models?

Caliskan et al. (2017)

IAT for Word Embeddings

- Word Embedding Association Test (WEAT)
  - Latency ⇔ Cosine similarity
  - Target words
    - X = {*programmer, engineer, scientist, …*}
    - Y = {*nurse, teacher, librarian, …*}
  - Attribute words
    - A = {*man, male, …* }
    - B = {*woman, female, …*}

Caliskan et al. (2017)

Word Embedding Association Test

- Target words
  - X = {*programmer, engineer, scientist, …*}
  - Y = {*nurse, teacher, librarian, …*}
- Attribute words
  - A = {*man, male, …* }
  - B = {*woman, female, …*}

Association of a word w with an attribute: $$s(w, A, B) = \text{mean}_{a \in A}\cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B}\cos(\vec{w}, \vec{b})$$

Association of two sets of target words with an attribute: $$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

The effect size of bias: $$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$$

Additional statistical tests to measure how separated are two distributions and statistical significance

Caliskan et al. (2017)

Word Embedding Association Test

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

- **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
- **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.
- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

Caliskan et al. (2017)

Word Embedding Association Test: Results

IAT | WEAT

Target words	Attrib. words	Original Finding				Our Finding			
		Ref	N	d	p	N_T	N_A	d	p
Flowers vs insects	Pleasant vs unpleasant	(5)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}

Caliskan et al. (2017)

Word Embedding Association Test

- **European American names**: Adam, *Chip*, Harry, Josh, Roger, Alan, Frank, *Ian*, Justin, Ryan, Andrew, *Fred*, Jack, Matthew, Stephen, Brad, Greg, *Jed*, Paul, *Todd*, *Brandon*, *Hank*, Jonathan, Peter, *Wilbur*, Amanda, Courtney, Heather, Melanie, *Sara*, *Amber*, *Crystal*, Katie, *Meredith*, *Shannon*, Betsy, *Donna*, Kristin, Nancy, Stephanie, *Bobbie-Sue*, Ellen, Lauren, *Peggy*, *Sue-Ellen*, Colleen, Emily, Megan, Rachel, *Wendy* (deleted names in italics).
- **African American names**: Alonzo, Jamel, *Lerone*, *Percell*, Theo, Alphonse, Jerome, Leroy, *Rasaan*, Torrance, Darnell, Lamar, Lionel, *Rashaun*, Tvree, Deion, Lamont, Malik, Terrence, Tyrone, *Everol*, Lavon, Marcellus, *Terryl*, Wardell, *Aiesha*, *Lashelle*, Nichelle, Shereen, *Temeka*, Ebony, Latisha, Shaniqua, *Tameisha*, *Teretha*, Jasmine, *Latonya*, *Shanise*, Tanisha, Tia, Lakisha, Latoya, *Sharise*, *Tashika*, Yolanda, *Lashandra*, Malika, *Shavonn*, *Tawanda*, Yvette (deleted names in italics).

- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

Caliskan et al. (2017)

Word Embedding Association Test: Results

		IAT				WEAT			
Target words	**Attrib. words**	**Original Finding**				**Our Finding**			
		Ref	**N**	**d**	**p**	$\mathbf{N_T}$	$\mathbf{N_A}$	**d**	**p**
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(5)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}

WEAT finds similar biases in Word Embeddings as IAT did for humans

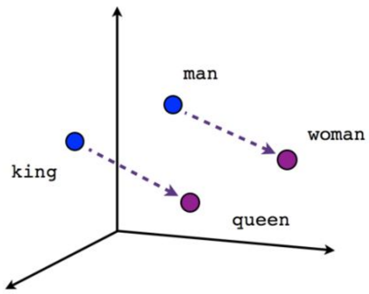
Other ways to detect biases?

Bolukbasi et al. (2016)

Gender Bias in Word Embeddings

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

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$



Male-Female

surgeon vs. nurse

architect vs. interior designer

shopkeeper vs. housewife

superstar vs. diva

....

Manzini et al. (2019)

Beyond Gender & Race/Ethnicity Bias

Gender Biased Analogies	
man → doctor	woman → nurse
woman → receptionist	man → supervisor
woman → secretary	man → principal
Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Biases in word embeddings trained on the Reddit data from US users.

Social Stereotypes → Word Embeddings?

Yes, they do!

But aren’t they just reflecting Society?

Garg et al. (2018)

Gender bias in Occupations

Garg et al. (2018)

Gender bias in Adjectives over the decades

	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.71	0.68	0.71	0.65	0.67	0.54	0.45	0.45
1920	0.71	1.00	0.71	0.68	0.65	0.66	0.55	0.51	0.46
1930	0.68	0.71	1.00	0.73	0.71	0.70	0.60	0.52	0.53
1940	0.71	0.68	0.73	1.00	0.74	0.69	0.56	0.53	0.51
1950	0.65	0.65	0.71	0.74	1.00	0.71	0.58	0.51	0.49
1960	0.67	0.66	0.70	0.69	0.71	1.00	0.62	0.54	0.50
1970	0.54	0.55	0.60	0.56	0.58	0.62	1.00	0.63	0.56
1980	0.45	0.51	0.52	0.53	0.51	0.54	0.63	1.00	0.62
1990	0.45	0.46	0.53	0.51	0.49	0.50	0.56	0.62	1.00

Height of women's movements in 1960s-70s

But aren't they just reflecting Society?

Yup!

Word embeddings …

… get things **normatively wrong** *precisely because* they get things **descriptively right**!

Oisin Deery & Katherine Bailey
Ethics in NLP workshop. NAACL ‘18

Shouldn’t we then just leave them as is?

Shouldn’t we then just leave them as is?

1. Would that harm certain groups of people?

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

Rhett Jones

Yesterday 10:32am • Filed to: ALGORITHMS

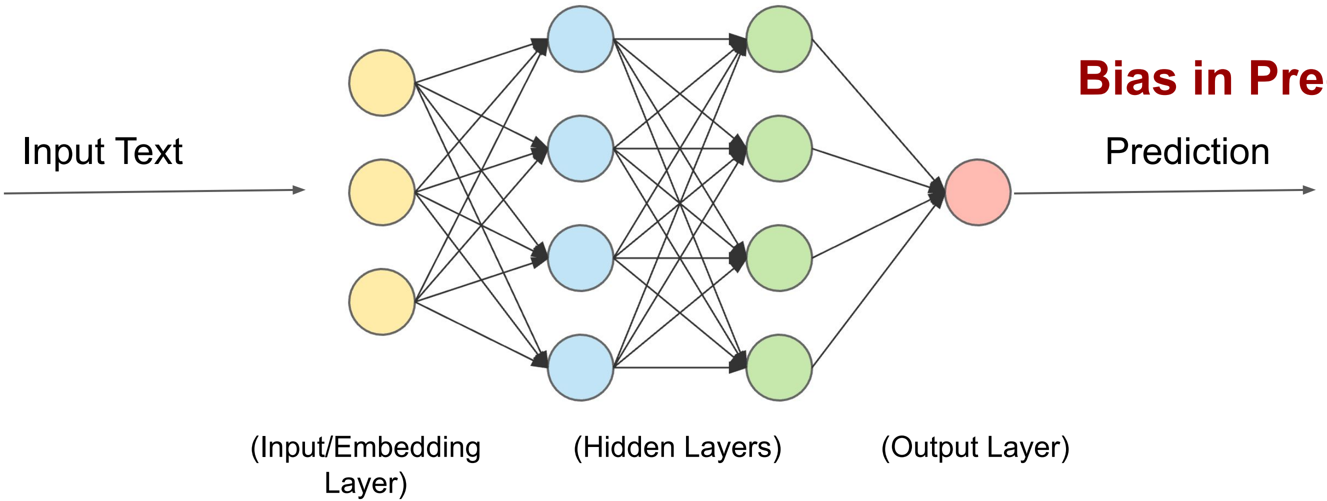
22.3K 96 2

Photo: Getty

Source: [Gizmodo]

Where to look for biases?

Bias in Predictions?



Bias in Input Representations?

Biases in NLP Classifiers/Taggers

- Gender Bias in Part of speech tagging and Dependency parsing
  - Garimella et al. **Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing**. ACL (2019)
- Gender Bias in Coreference resolution
  - Zhao, Jieyu, et al. **Gender bias in coreference resolution: Evaluation and debiasing methods.** *arXiv* (2018)
  - Webster, Kellie, et al. **Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns.** *TACL* (2018)
- Gender, Race, and Age Bias in Sentiment Analysis
  - Svetlana and Mohammad. **Examining gender and race bias in two hundred sentiment analysis systems**. arXiv (2018)
  - Díaz, et al. **Addressing age-related bias in sentiment analysis.** CHI Conference on Human Factors in Comp. Systems. (2018)
- LGBTQ identitiy terms bias in Toxicity classification
  - Dixon, et al. **Measuring and mitigating unintended bias in text classification.** AIES. (2018)
  - Sap, et al. **The Risk of Racial Bias in Hate Speech Detection.** ACL. (2019)
- Gender Bias in Occupation Classification
  - De-Arteaga et al. **Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting.** FAT* (2019)
- Gender bias in Machine Translation
  - Prates, et al. **Assessing gender bias in machine translation: a case study with Google Translate.** Neural Computing and Applications (2018)

Shouldn’t we then just leave them as is?

1. Would that harm certain groups of people?
2. Would that make things worse?

Bias Amplification

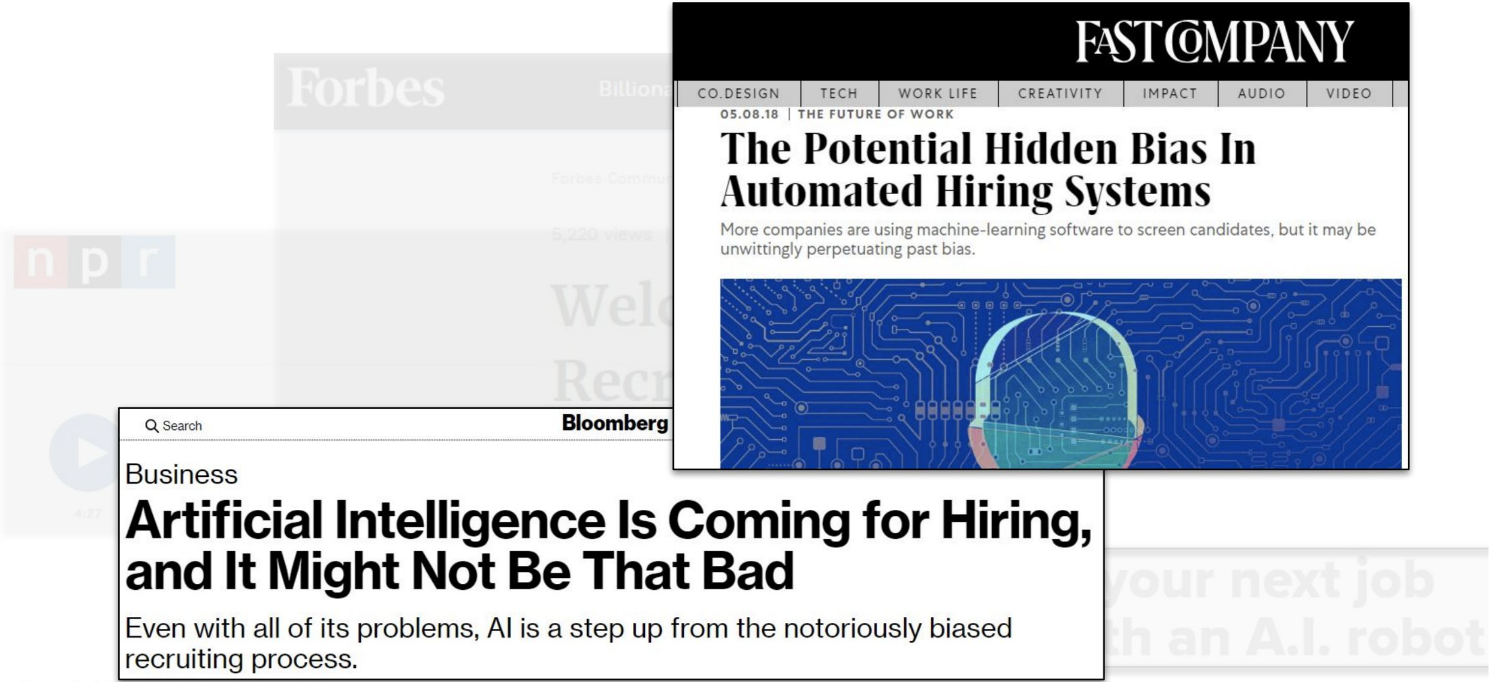
- Zhao et al. **Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint.** *EMNLP* (2017)

- *De-Arteaga et al.* **Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting.** *FAT* (2019)*

Swinger et al. (2019)

Examples of Harm from NLP Bias

An artificially intelligent headhunter?



Swinger et al. (2019)

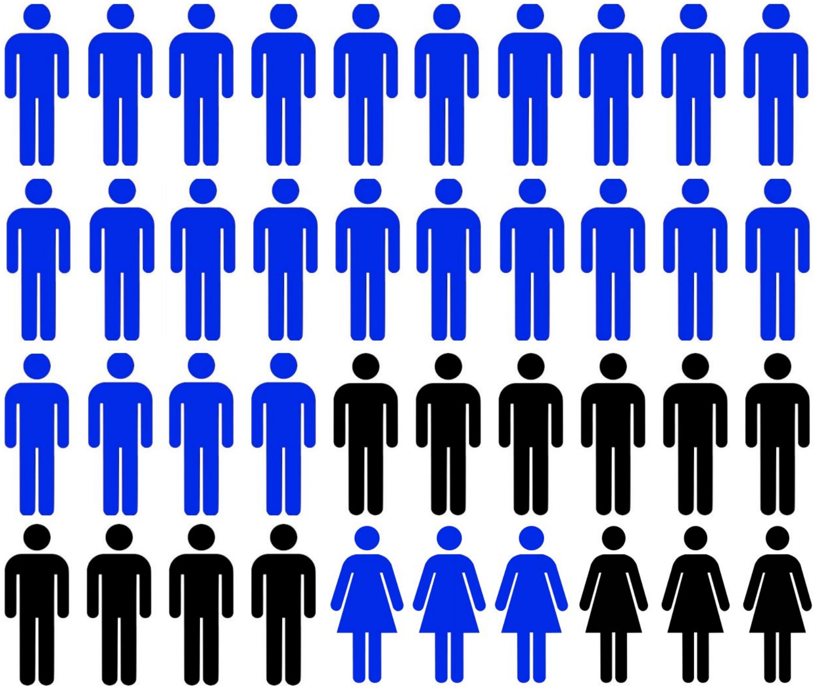
Examples of Harm from NLP Bias

Compounding imbalances

Surgeons

females in data:
14.6%

females in true positives:
11.6%



Males:
71% recall

Females:
54% recall

Slide credit: Maria De-Arteaga

Ok, How do we make NLP models fair?

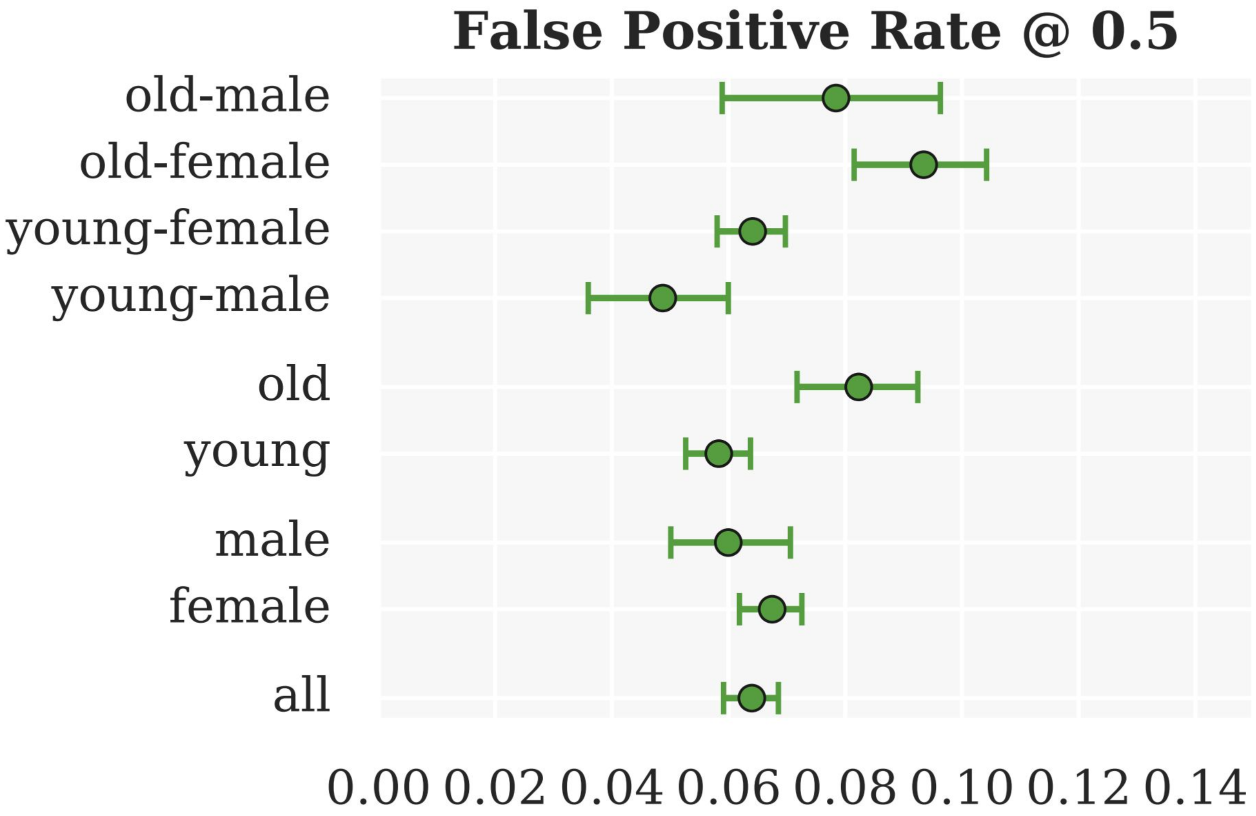
What does it mean to be Fair?

Different Types of Fairness

- Group Fairness
  - “treat different groups equally”
  - E.g., demographic parity across groups (along age, gender, race, etc.)

- Individual Fairness
  - “treat similar examples similarly”
  - E.g., counterfactual fairness (if we switch the gender, does the prediction change?)

Group Fairness



Individual Fairness

```
text_to_sentiment("My name is Emily")
```

```
2.2286179364745311
```

```
text_to_sentiment("My name is Heather")
```

```
1.3976291151079159
```

```
text_to_sentiment("My name is Yvette")
```

```
0.9846380213298556
```

```
text_to_sentiment("My name is Shaniqua")
```

```
-0.47048131775890656
```

http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

Can we computationally remove undesirable biases?

- Debiasing Meaning Representations

Methods to “de-bias” NLP models

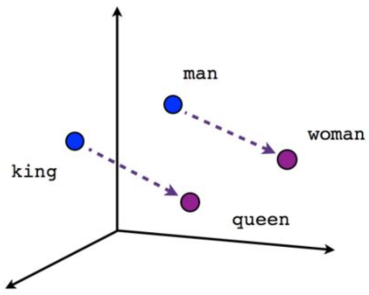
- Gender De-Biasing
  - Bolukbasi et al. **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS* (2016)
  - Zhao, Jieyu, et al. **Men also like shopping: Reducing gender bias amplification using corpus-level constraints.** arXiv (2017)
  - Park, et al. **Reducing gender bias in abusive language detection.** arXiv (2018)
  - Zhao, Jieyu, et al. **Learning gender-neutral word embeddings.** arXiv (2018)
  - Anne Hendricks, et al. **Women also snowboard: Overcoming bias in captioning models.** ECCV. (2018)
- General De-Biasing
  - Beutel et al. **Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations.** FATML (2017)
  - Zhang, et al. **Mitigating unwanted biases with adversarial learning.** AIES, 2018
  - Elazar and Goldberg. **Adversarial removal of demographic attributes from text data.** arXiv (2018)
  - Hu and Strout. **Exploring Stereotypes and Biased Data with the Crowd.** arXiv (2018)

Bolukbasi et al. (2016)

Gender Bias in Word Embeddings

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

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$



Male-Female

surgeon vs. nurse

architect vs. interior designer

shopkeeper vs. housewife

superstar vs. diva

....

Bolukbasi et al. (2016)

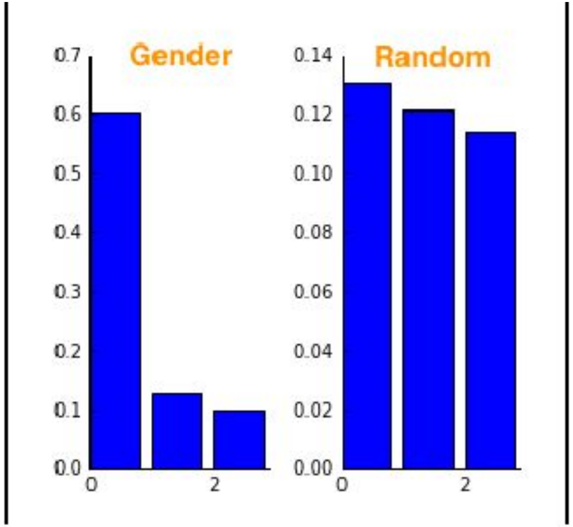
Towards Debiasing

1. Identify gender subspace: B

Bolukbasi et al. (2016)

Gender Subspace

$$\overrightarrow{\text{she}} - \overrightarrow{\text{he}}$$
$$\overrightarrow{\text{her}} - \overrightarrow{\text{his}}$$
$$\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$$
$$\overrightarrow{\text{Mary}} - \overrightarrow{\text{John}}$$
$$\overrightarrow{\text{herself}} - \overrightarrow{\text{himself}}$$
$$\overrightarrow{\text{daughter}} - \overrightarrow{\text{son}}$$
$$\overrightarrow{\text{mother}} - \overrightarrow{\text{father}}$$
$$\overrightarrow{\text{gal}} - \overrightarrow{\text{guy}}$$
$$\overrightarrow{\text{girl}} - \overrightarrow{\text{boy}}$$
$$\overrightarrow{\text{female}} - \overrightarrow{\text{male}}$$



The top PC captures the gender subspace

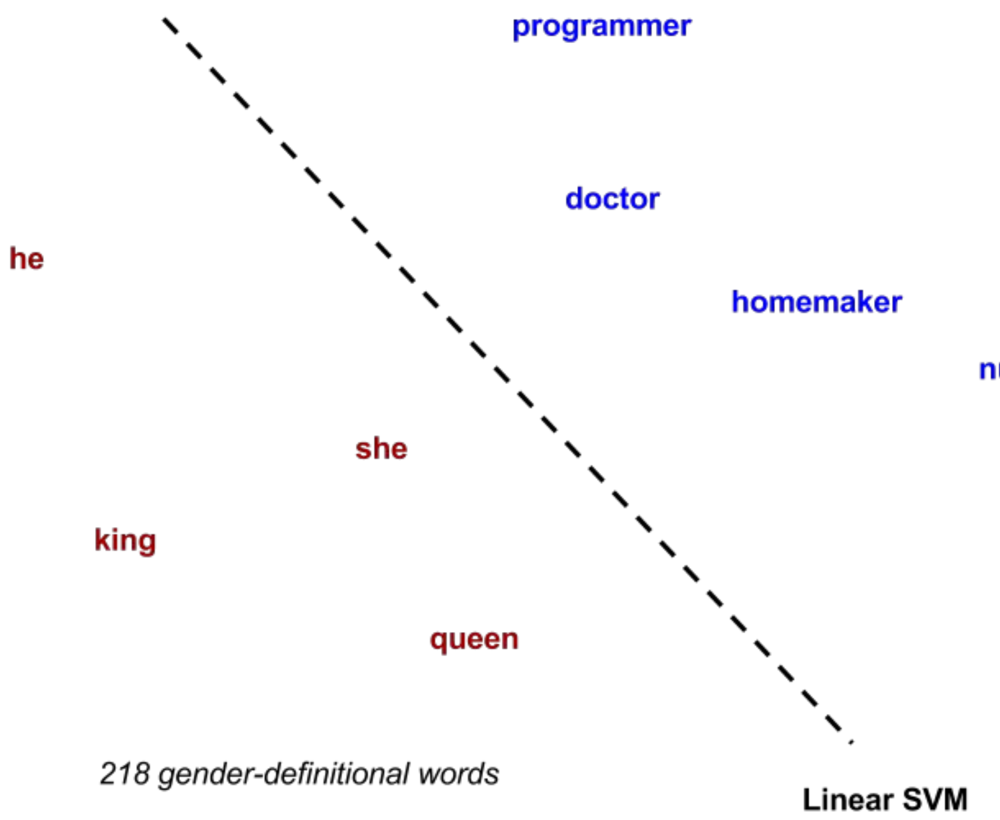
Bolukbasi et al. (2016)

Towards Debiasing

1. Identify gender subspace: B
2. **Identify gender-definitional (S) and gender-neutral words (N)**

Bolukbasi et al. (2016)

Gender-definitional vs. Gender-neutral Words



Plus
Bootstrapping

Bolukbasi et al. (2016)

Towards Gender Debiasing

1. Identify gender subspace: B
2. Identify gender-definitional (S) and gender-neutral words (N)

Bolukbasi et al. (2016)

Towards Gender Debiasing

1. Identify gender subspace: B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply transform matrix (T) to the embedding matrix (W) such that
   a. Project away the gender subspace B from the gender-neutral words N
   b. But, ensure the transformation doesn't change the embeddings too much

$$min_T \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{Don't modify embeddings too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{Minimize gender component}}$$

T - the desired debiasing transformation
W - embedding matrix

B - biased space
N - embedding matrix of gender neutral words

Can we computationally remove undesirable biases?

- Debiasing Meaning Representations
- Debiasing Model Predictions

Debiasing using Adversarial Learning

Beutel et al. (2017)
Zhang et al. (2018)

Bias Mitigation

- Handling biased predictions
- Removing signal for problematic variables
  - Stereotyping
  - Sexism, Racism, *-ism

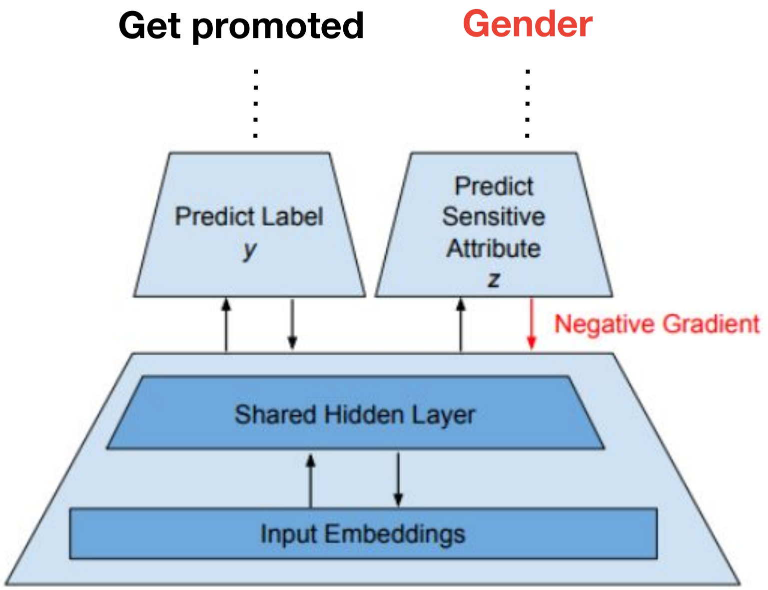
Debiasing using Adversarial Learning

Beutel et al. (2017)
Zhang et al. (2018)

Bias Mitigation

- Handling biased predictions
- Removing signal for problematic variables
  - Stereotyping
  - Sexism, Racism, *-ism

Adversarial Multi-task Learning



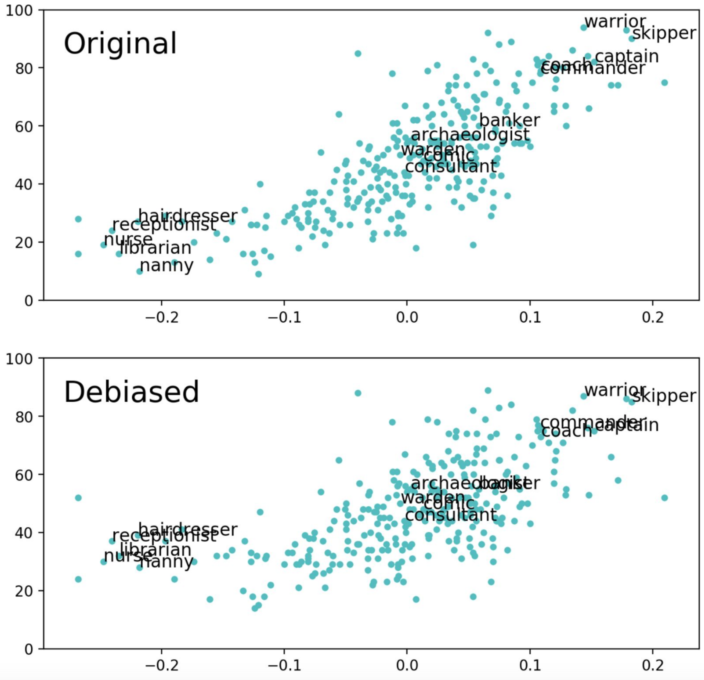
Can we computationally remove undesirable biases?

YES!

Are we done?

Issues with relying entirely on ‘debiasing’

- Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.** NAACL (2019).



So…

What should we do?

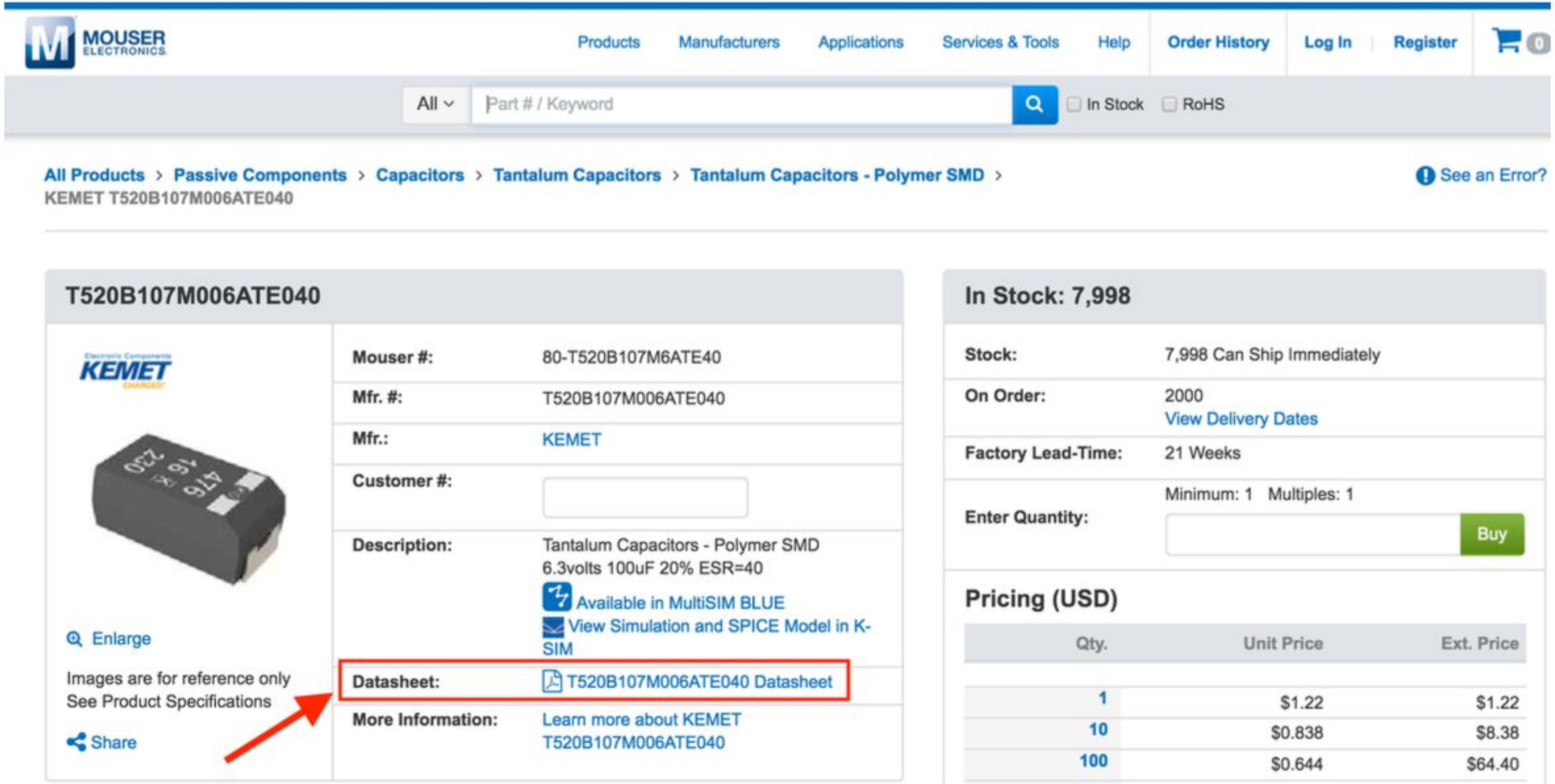
Can we computationally remove undesirable biases?

Recommendations

- Always **be mindful** of various sorts of biases in the NLP models and the data
- Explore “debiasing” techniques, **but be cautious**
- Think about the biases that matter for your problem and **test for those biases**
- Be **transparent** about the models you release to the world

Speaking of Transparency...

Transparency for Electronics Components



Slide by Timnit Gebru

Transparency for Electronics Components

XICON

Miniature Aluminum Electrolytic Capacitors — XRL Series

■ FEATURES

- Low impedance characteristics
- Case sizes are smaller than conventional general-purpose capacitors, with very high performance
- Can size larger than 9mm diameter has safety vents on rubber end seal
- RoHS Compliant

RoHS Compliant

■ CHARACTERISTICS

Item			Characteristics											
Operating Temperature Range	-40°C ~ +85°C													
Capacitance Tolerance	±20% at 120Hz, 20°C													
Leakage Current	≤100V	I = 0.01CWV or 3µA whichever is greater after 2 minutes of applied rated DC working voltage at 20°C Where: C = rated capacitance in µF; WV = rated DC working voltage												
	>100V	CWV ≤ 1000 µF: I= 0.03 CWV + 15uA; C= rated capacitance in uF CWV ≥ 1000 µF: I= 0.02 CWV + 25uA; WV= rated DC working voltage in V												
Dissipation Factor (Tan δ, at 20°C 120Hz)	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Tan δ		0.23	0.20	0.16	0.14	0.12	0.10	0.09	0.08	0.12	0.17	0.20	0.25
	For capacitors whose capacitance exceeds 1,000µF, the specification of tan δ is increased by 0.02 for every addition of 1,000µF													
Surge Voltage	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Surge voltage (SV)		8	13	20	32	44	63	79	125	200	300	400	500
Low Temperature Characteristics (Imp. ratio @ 120Hz)	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Z(-25°C)/Z(+20°C)	øD<16	6	4	3	3	2	2	2	2	3	8	12	16
		øD≥16	8	6	4	4	3	3	3	3	3	8	12	16
	Z(-40°C)/Z(+20°C)	øD<16	10	8	6	6	4	3	3	3	4	10	16	20
		øD≥16	18	16	12	10	8	8	6	6	4	10	16	20
Load Test	When returned to +20°C after 2,000 hours application of working voltage at +85°C, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value													
Shelf Life Test	When returned to +20°C after 1,000 hours at +85°C with no voltage applied, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value													

■ PART NUMBERING SYSTEM

1 4 0	–	X R L	1 6 V	1 0 0	–	R C
Prefix		Series	Voltage Actual Value	Capacitance (µF) Actual Value		Suffix RoHS Compliant

■ RIPPLE CURRENT AND FREQUENCY MULTIPLIERS

Capacitance (µF)	Frequency (Hz)				
	60 (50)	120	500	1K	≥10K
<100	0.70	1.0	1.30	1.40	1.50
100 ~ 1000	0.75	1.0	1.20	1.30	1.35
>1000	0.80	1.0	1.10	1.12	1.15

■ RIPPLE CURRENT AND TEMPERATURE MULTIPLIERS

Temperature (°C)	≤50	70	85
Multiplier	1.78	1.4	1.0

XICON

XICON PASSIVE COMPONENTS • (800) 628-0544

XC-600178 Date Revised: 1/8/07

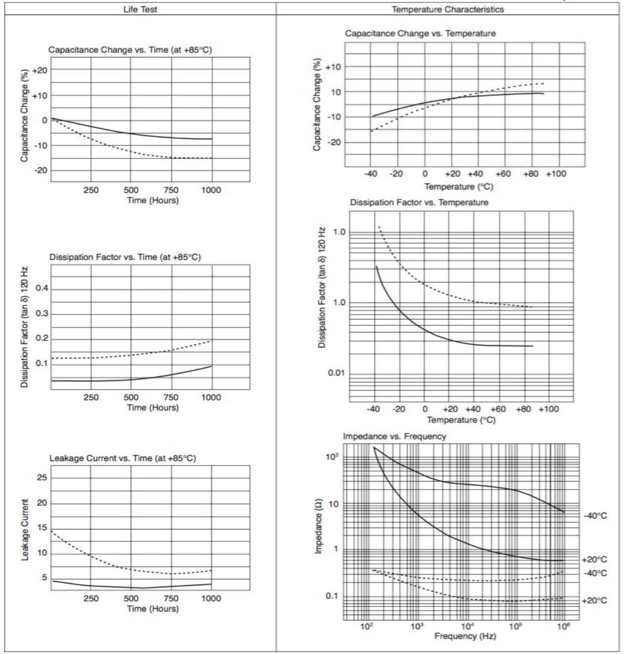
Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

XICON

Miniature Aluminum Electrolytic Capacitors — XRL Series

■ TYPICAL PERFORMANCE CHARACTERISTICS

- - - - - - 1000µF 16V
——— 1µF 50V



XICON

XICON PASSIVE COMPONENTS • (800) 628-0544

XC-600178 Date Revised: 1/8/07

Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

Slide by Timnit Gebru

Speaking of Transparency...

- Data Sheets for Datasets

Datasheets for Datasets

- Gebru et al. (2019)
  - https://arxiv.org/pdf/1803.09010.pdf
- Key questions for each stage:
  - Motivation
  - Composition
  - Collection Process
  - Preprocessing/cleaning/labeling
  - Uses
  - Distribution
  - Maintenance
- For dataset creators:
  - Encourage reflection on the process and assumptions
- For dataset consumers:
  - Provide information for making informed decisions

Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity: given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The polarity rating is binary {positive,negative}. An example instance is shown in Figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1400 instances in total in the original (v1.x versions) and 2000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is (presumably) intended to be a random sample of instances of movie revies from newsgroup postings. No tests were run to determine representativeness.

[1] Information in this datasheet is taken from one of five sources; any errors that were introduced are our fault. http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file neg/cv452_tok-18656.txt.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and alter fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient con-

Speaking of Transparency...

- Data Sheets for Datasets
- Model Cards for model reporting

Model Card for Toxicity Model

Mitchell et al. (2019)
FAT *

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

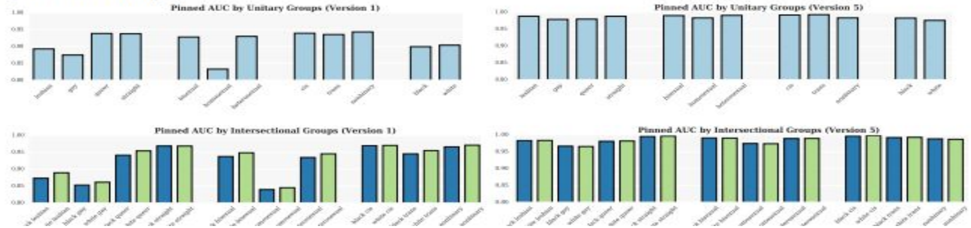
Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses



Closing Note

"Fairness and justice are properties of social and legal systems"

"To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore [...] an abstraction error"

Selbst et al., Fairness and Abstraction in Sociotechnical Systems. FAT* 2018

Thank You!

Acknowledgments:

Team





Internal


External