Bias and Fairness in NLP

Margaret Mitchell
Google Brain

Kai-Wei Chang
UCLA

Vinodkumar Prabhakaran
Google Brain

Vicente Ordóñez Román
University of Virginia
Tutorial Outline

- Part 1: Cognitive Biases / Data Biases / Bias laundering
- Part 2: Bias in NLP and Mitigation Approaches
- Part 3: Building Fair and Robust Representations for Vision and Language
- Part 4: Conclusion and Discussion
What’s in this tutorial

- Motivation for Fairness research in NLP
- How and why NLP models may be unfair
- Various types of NLP fairness issues and mitigation approaches
- What can/should we do?
What’s **NOT** in this tutorial

- Definitive answers to fairness/ethical questions
- Prescriptive solutions to fix ML/NLP (un)fairness
What do you see?
What do you see?

- Bananas
What do you see?

- Bananas
- Stickers
What do you see?

- Bananas
- Stickers
- Dole Bananas
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

...We don’t tend to say **Yellow Bananas**
What do you see?

Green Bananas

Unripe Bananas
What do you see?

Ripe Bananas

Bananas with spots
What do you see?

Yellow Bananas

Yellow is prototypical for bananas
Prototype Theory

One purpose of categorization is to *reduce the infinite differences* among stimuli to behaviourally and *cognitively usable proportions*. There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975). May also store exemplars (Wu & Barsalou, 2009).
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
“Doctor”

“Female doctor”
The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University
### World learning from text

Gordon and Van Durme, 2013

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>“spoke”</td>
<td>11,577,917</td>
</tr>
<tr>
<td>“laughed”</td>
<td>3,904,519</td>
</tr>
<tr>
<td>“murdered”</td>
<td>2,834,529</td>
</tr>
<tr>
<td>“inhaled”</td>
<td>984,613</td>
</tr>
<tr>
<td>“breathed”</td>
<td>725,034</td>
</tr>
<tr>
<td>“hugged”</td>
<td>610,040</td>
</tr>
<tr>
<td>“blinked”</td>
<td>390,692</td>
</tr>
<tr>
<td>“exhale”</td>
<td>168,985</td>
</tr>
</tbody>
</table>
World learning from text

Gordon and Van Durme, 2013

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>“spoke”</td>
<td>11,577,917</td>
</tr>
<tr>
<td>“laughed”</td>
<td>3,904,519</td>
</tr>
<tr>
<td>“murdered”</td>
<td>2,834,529</td>
</tr>
<tr>
<td>“inhaled”</td>
<td>984,613</td>
</tr>
<tr>
<td>“breathed”</td>
<td>725,034</td>
</tr>
<tr>
<td>“hugged”</td>
<td>610,040</td>
</tr>
<tr>
<td>“blinked”</td>
<td>390,692</td>
</tr>
<tr>
<td>“exhale”</td>
<td>168,985</td>
</tr>
</tbody>
</table>
Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals.
Training data are collected and annotated
Training data are collected and annotated

Model is trained
Training data are collected and annotated.

Model is trained.

Media are filtered, ranked, aggregated, or generated.
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output
Training data are collected and annotated

**Human Biases in Data**

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect
Human Biases in Data

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Group attribution error
- Halo effect
- Prejudice

Train data are collected and annotated

Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter's bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity
**Reporting bias:** What people share is not a reflection of real-world frequencies

**Selection Bias:** Selection does not reflect a random sample

**Out-group homogeneity bias:** People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

**Confirmation bias:** The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

**Overgeneralization:** Coming to conclusion based on information that is too general and/or not specific enough

**Correlation fallacy:** Confusing correlation with causation

**Automation bias:** Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

More at: https://developers.google.com/machine-learning/glossary/
Biases in Data
Selection Bias: Selection does not reflect a random sample

World Englishes

Is the data we use to train our English NLP models representative of all the Englishes out there?
Biases in Data

Selection Bias: Selection does not reflect a random sample

- 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)
Biases in Data

Selection Bias: Selection does not reflect a random sample

Map of Amazon Mechanical Turk Workers

CREDIT
© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek
Biases in Data

Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members.
It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.
Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.

Biases in Interpretation
Biases in Interpretation

**Confirmation bias:** The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs.
Biases in Interpretation

**Overgeneralization:** Coming to conclusion based on information that is too general and/or not specific enough (related: overfitting)

---

CREDIT
Sidney Harris
Biases in Interpretation

Correlation fallacy: Confusing correlation with causation

Post Hoc Ergo Propter Hoc

Women were allowed to vote in the early 1900’s and then we had two world wars. Clearly giving them the vote was a bad idea.

CREDIT
© mollysdad - Slideshare - Introduction to Logical Fallacies
Biases in Interpretation

**Automation bias:** Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation.
<table>
<thead>
<tr>
<th>Human Biases in Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting bias</td>
</tr>
<tr>
<td>Selection bias</td>
</tr>
<tr>
<td>Overgeneralization</td>
</tr>
<tr>
<td>Out-group homogeneity bias</td>
</tr>
<tr>
<td>Stereotypical bias</td>
</tr>
<tr>
<td>Historical unfairness</td>
</tr>
<tr>
<td>Implicit associations</td>
</tr>
<tr>
<td>Implicit stereotypes</td>
</tr>
<tr>
<td>Prejudice</td>
</tr>
<tr>
<td>Group attribution error</td>
</tr>
<tr>
<td>Halo effect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human Biases in Collection and Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data are collected and annotated</td>
</tr>
<tr>
<td>Sampling error</td>
</tr>
<tr>
<td>Non-sampling error</td>
</tr>
<tr>
<td>Insensitivity to sample size</td>
</tr>
<tr>
<td>Correspondence bias</td>
</tr>
<tr>
<td>In-group bias</td>
</tr>
<tr>
<td>Bias blind spot</td>
</tr>
<tr>
<td>Confirmation bias</td>
</tr>
<tr>
<td>Subjective validation</td>
</tr>
<tr>
<td>Experimenter's bias</td>
</tr>
<tr>
<td>Choice-supportive bias</td>
</tr>
<tr>
<td>Neglect of probability</td>
</tr>
<tr>
<td>Anecdotal fallacy</td>
</tr>
<tr>
<td>Illusion of validity</td>
</tr>
</tbody>
</table>
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output

Human Bias
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output

Human Bias
Training data are collected and annotated. The Model is trained. Media are filtered, ranked, aggregated, or generated. People see output and act based on it. This process creates a Feedback Loop that introduces Human Bias at each step.
Human data perpetuates human biases.

As ML learns from human data, the result is a bias network effect

“Bias Laundering”
BIAS = BAD ??
“Bias” can be Good, Bad, Neutral

- **Bias in statistics and ML**
  - Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
  - The "bias" term $b$ (e.g., $y = mx + b$)

- **Cognitive biases**
  - Confirmation bias, Recency bias, Optimism bias

- **Algorithmic bias**
  - Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making
“Bias” can be Good, Bad, Neutral

- Bias in statistics and ML
  - Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
  - The "bias" term $b$ (e.g., $y = mx + b$)

- Cognitive biases
  - Confirmation bias, Recency bias, Optimism bias

- Algorithmic bias
  - Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making
“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
“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
Fairness in Machine Learning

A Few Case Studies
Language Identification
Language Identification

Most NLP models in practice has a Language Identification (LID) step
Language Identification

Most NLP models in practice have a Language Identification (LID) step.
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”

LID Usage Example: Public Health Monitoring

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

Selection Bias: Selection does not reflect a random sample

World Englishes

Is the data we use to train our English NLP models representative of all the Englishes out there?
How does this affect NLP models?

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 systematically excluded from this technology
Predicting Homosexuality
Predicting Homosexuality

- “Sexual orientation detector” using 35,326 images from public profiles on a US dating website.
- “Consistent with the prenatal hormone theory [PHT] of sexual orientation, gay men and women tended to have gender-atypical facial morphology.”
Predicting Homosexuality

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.

See Medium article: “Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?”
Predicting Criminality
Predicting Criminality

Israeli startup, Faception

“Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial image.”

Offering specialized engines for recognizing “High IQ”, “White-Collar Offender”, “Pedophile”, and “Terrorist” from a face image.

Main clients are in homeland security and public safety.
Predicting Criminality

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

1,856 closely cropped images of faces; Includes “wanted suspect” ID pictures from specific regions.

“[…] angle $\theta$ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals …”

See our longer piece on Medium, “Physiognomy’s New Clothes”
Predicting Toxicity in Text
Toxicity Classification

We asked the internet what they thought about:

Climate Change  Brexit  US Election

Showing 46 of 49 total comments based on toxicity:

- Climate change is happening and it's not changing in our favor. If you think differently you're an idiot.
- They're allowed to do that. But if they act like assholes about, I will block them.
- You're stupid, it's getting warmer, we should enjoy it while it lasts.
- I think those people are stupid and short-sighted
- I think its a farce and stinks like a bathroom after 26 beers
- Fools
- My thoughts are that people should stop being stupid and ignorant. Climate change is scientifically proven. It isn't a debate.
- They are uninformed or ignorant
- Their opinion, just don't force it down my throat.
Toxicity 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

Unintended biases towards certain identity terms:

<table>
<thead>
<tr>
<th>Comment</th>
<th>Toxicity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Gay and Lesbian Film Festival starts today.</td>
<td>0.82</td>
</tr>
<tr>
<td>Being transgender is independent of sexual orientation.</td>
<td>0.52</td>
</tr>
<tr>
<td>A Muslim is someone who follows or practices Islam</td>
<td>0.46</td>
</tr>
</tbody>
</table>

## Toxicity Classification

Unintended biases towards **named entities:**

<table>
<thead>
<tr>
<th>Comment</th>
<th>Toxicity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I hate Justin Timberlake.</td>
<td>0.90</td>
</tr>
<tr>
<td>I hate Rihanna.</td>
<td>0.69</td>
</tr>
</tbody>
</table>

- Prabhakaran et al. (2019). “Perturbation Sensitivity Analysis to Detect Unintended Model Biases” EMNLP 2019
# Toxicity Classification

Unintended biases towards **mentions of disabilities**:

<table>
<thead>
<tr>
<th>Comment</th>
<th>Toxicity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am a person.</td>
<td>0.08</td>
</tr>
<tr>
<td>I am a tall person.</td>
<td>0.03</td>
</tr>
</tbody>
</table>

## Unintended biases towards mentions of disabilities:

<table>
<thead>
<tr>
<th>Comment</th>
<th>Toxicity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am a person.</td>
<td>0.08</td>
</tr>
<tr>
<td>I am a tall person.</td>
<td>0.03</td>
</tr>
<tr>
<td>I am a blind person.</td>
<td>0.39</td>
</tr>
<tr>
<td>I am a deaf person.</td>
<td>0.44</td>
</tr>
</tbody>
</table>

## Toxicity Classification

Unintended biases towards mentions of disabilities:

<table>
<thead>
<tr>
<th>Comment</th>
<th>Toxicity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am a person.</td>
<td>0.08</td>
</tr>
<tr>
<td>I am a tall person.</td>
<td>0.03</td>
</tr>
<tr>
<td>I am a blind person.</td>
<td>0.39</td>
</tr>
<tr>
<td>I am a deaf person.</td>
<td>0.44</td>
</tr>
<tr>
<td>I am a person with mental illness.</td>
<td>0.62</td>
</tr>
</tbody>
</table>

NLP Research on Bias and Fairness
Fairness Research in NLP


## Fairness Research in NLP

<table>
<thead>
<tr>
<th></th>
<th>Authors</th>
<th>Title</th>
<th>Conference/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Social Disparities (and Stereotypes) → Word Embeddings?

He is... → She is...

But aren’t they just reflecting Society?
Gender bias in Occupations

Garg et al. (2018)
Gender bias in Adjectives over the decades

Garg et al. (2018)
“Asian bias” in Adjectives with “Outsider” words

Garg et al. (2018)
“Islam bias” in Adjectives with “Terrorist” words

Garg et al. (2018)
But aren’t they just reflecting Society?

Yup!
Word embeddings …

… get things normatively wrong precisely because they get things descriptively right!
Shouldn’t we then just leave them as is?
Shouldn’t we then just leave them as is?

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

Photo: Getty

Source: Gizmodo
What kind of harm?

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
Measuring Algorithmic Fairness/Bias
Evaluate for Fairness & Inclusion

Disaggregated Evaluation

Create for each (subgroup, prediction) pair. Compare across subgroups.
Evaluate for Fairness & Inclusion

Disaggregated Evaluation

Create for each (subgroup, prediction) pair. Compare across subgroups.

Example: women, face detection
         men, face detection
Evaluate for Fairness & Inclusion

Intersectional Evaluation

Create for each (subgroup1, subgroup2, prediction) pair. Compare across subgroups.

Example: black women, face detection
white men, face detection
Evaluate for Fairness & Inclusion: Confusion Matrix

Model Predictions
Evaluate for Fairness & Inclusion: Confusion Matrix
Evaluate for Fairness & Inclusion: Confusion Matrix

<table>
<thead>
<tr>
<th>References</th>
<th>Model Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Exists</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td>True Positives</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Doesn’t exist</td>
</tr>
<tr>
<td></td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td>True Negatives</td>
</tr>
</tbody>
</table>
Evaluate for Fairness & Inclusion: Confusion Matrix

<table>
<thead>
<tr>
<th>References</th>
<th>Model Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>Exists</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td>True Positives</td>
</tr>
<tr>
<td>Negative</td>
<td>Exists</td>
</tr>
<tr>
<td></td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td>False Negatives</td>
</tr>
<tr>
<td>Negative</td>
<td>Doesn’t exist</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td>False Positives</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doesn’t exist</td>
</tr>
<tr>
<td></td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td>True Negatives</td>
</tr>
</tbody>
</table>
Evaluate for Fairness & Inclusion: Confusion Matrix

- **Model Predictions**
  - **Positive**
    - Exists
    - Predicted
    - **True Positives**
  - **Negative**
    - Exists
    - Not predicted
    - **False Negatives**

- **Successful Predictions**
  - Positive: True Positives
  - Negative: True Negatives

- **Failed Predictions**
  - Positive: False Positives
  - Negative: False Negatives

- **References**
  - Positive: Exists, Predicted
  - Negative: Doesn’t exist, Predicted

- **Evaluation Metrics**
  - **Recall, False Negative Rate**
  - **False Positive Rate, Specificity**
  - **Precision, False Discovery Rate**
  - **Negative Predictive Value, False Omission Rate**
  - LR+, LR-
## Evaluate for Fairness & Inclusion

### Female Patient Results

<table>
<thead>
<tr>
<th>True Positives (TP)</th>
<th>10</th>
<th>False Positives (FP)</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Negatives (FN)</td>
<td>1</td>
<td>True Negatives (TN)</td>
<td>488</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909 \)

Recall = \( \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909 \)

### Male Patient Results

<table>
<thead>
<tr>
<th>True Positives (TP)</th>
<th>6</th>
<th>False Positives (FP)</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Negatives (FN)</td>
<td>5</td>
<td>True Negatives (TN)</td>
<td>48</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667 \)

Recall = \( \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545 \)
Evaluate for Fairness & Inclusion

<table>
<thead>
<tr>
<th></th>
<th>Female Patient Results</th>
<th>Male Patient Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives (TP)</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>False Positives (FP)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>False Negatives (FN)</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>True Negatives (TN)</td>
<td>488</td>
<td>48</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909 \)

Recall = \( \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909 \)

Precision = \( \frac{6}{6 + 3} = 0.667 \)

Recall = \( \frac{6}{6 + 5} = 0.545 \)

“Equality of Opportunity” fairness criterion: Recall is equal across subgroups
Evaluate for Fairness & Inclusion

<table>
<thead>
<tr>
<th>Female Patient Results</th>
<th>Male Patient Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives (TP) = 10</td>
<td>True Positives (TP) = 6</td>
</tr>
<tr>
<td>False Positives (FP) = 1</td>
<td>False Positives (FP) = 3</td>
</tr>
<tr>
<td>False Negatives (FN) = 1</td>
<td>False Negatives (FN) = 5</td>
</tr>
<tr>
<td>True Negatives (TN) = 488</td>
<td>True Negatives (TN) = 48</td>
</tr>
</tbody>
</table>

**Precision**

Female Patient Results: \[
\text{Precision} = \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909
\]

Male Patient Results: \[
\text{Precision} = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667
\]

**Recall**

Female Patient Results: \[
\text{Recall} = \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909
\]

Male Patient Results: \[
\text{Recall} = \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545
\]

“Predictive Parity” fairness criterion:

Precision is equal across subgroups
Choose your evaluation metrics in light of acceptable tradeoffs between False Positives and False Negatives.
False Positives Might be Better than False Negatives

Privacy in Images

**False Positive:** Something that doesn’t need to be blurred gets blurred.

Can be a bummer.

**False Negative:** Something that needs to be blurred is not blurred.

Identity theft.
False Negatives Might Be Better than False Positives

Spam Filtering

False Negative: Email that is SPAM is not caught, so you see it in your inbox. Usually just a bit annoying.

False Positive: Email flagged as SPAM is removed from your inbox. If it is an interview call?
AI Can Unintentionally Lead to Unjust Outcomes

- Lack of insight into sources of bias in the data and model
- Lack of insight into the feedback loops
- Lack of careful, *disaggregated* evaluation
- Human biases in interpreting and accepting results

So… What do we do?
Part 2: Bias in NLP and Mitigation Approaches (Kai-Wei)
Part 3: Building Fair and Robust Representations for Vision and Language (Vicente)
Part 4: Conclusion and Discussion (Vinod)
Data Really, Really Matters
Understand Your Data Skews

Facets: pair-code.github.io
Datasheets for Datasets

Timnit Gebru¹ Jamie Morgenstern² Briana Vecchio³ Jennifer Wortman Vaughan¹ Hanna Wallach¹ Hal Daumé III¹⁴ Kate Crawford¹⁵

### Motivation for Dataset Creation

**Why was the dataset created?** (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

**What (other) tasks could the dataset be used for?** Are there obvious tasks for which it should not be used?

**Has the dataset been used for any tasks already?** If so, where are the results so others can compare (e.g., links to published papers)?

**Who funded the creation of the dataset?** If there is an associated grant, provide the grant number.

**Any other comments?**

### Data Collection Process

**How was the data collected?** (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

**Who was involved in the data collection process?** (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

**Over what time-frame was the dataset collected?** Does the collection time-frame match the creation time-frame?

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/observed from other data (e.g., part of speech tags, model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

**Does the dataset contain all possible instances?** Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

**If the dataset is a sample, then what is the population?** What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

### Dataset Fact Sheet

<table>
<thead>
<tr>
<th><strong>Metadata</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Author</strong></td>
</tr>
<tr>
<td><strong>Email</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Keywords</strong></td>
</tr>
<tr>
<td><strong>Records</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Variables</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>priors_count</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Dependency Probability</strong></th>
<th>Pearson R</th>
</tr>
</thead>
</table>

### Probabilistic Modeling

Analysis

---
Release Your Models Responsibly
Transparency for Electronics Components

Slide by Timnit Gebru
“Operating Characteristics” of a component
Model Cards for Model Reporting

- Currently no common practice of reporting how well a model works when it is released

What It Does
A report that focuses on transparency in model performance to encourage responsible AI adoption and application.

How It Works
It is an easily discoverable and usable artifact presented at important steps of a user journey for a diverse set of users and public stakeholders.

Why It Matters
It keeps model developer accountable to release high quality and fair models.

## Intended Use, Factors and Subgroups

<table>
<thead>
<tr>
<th>Example Model Card - Toxicity in Text</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Details</strong></td>
</tr>
<tr>
<td>Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.</td>
</tr>
<tr>
<td><strong>Intended Use</strong></td>
</tr>
<tr>
<td>Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.</td>
</tr>
<tr>
<td><strong>Factors</strong></td>
</tr>
<tr>
<td>Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.</td>
</tr>
</tbody>
</table>

## Metrics and Data

<table>
<thead>
<tr>
<th><strong>Metrics</strong></th>
<th><em>Pinned AUC</em>, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation Data</strong></td>
<td>A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.</td>
</tr>
<tr>
<td><strong>Training Data</strong></td>
<td>Includes comments from a variety of online forums 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”.</td>
</tr>
</tbody>
</table>

### Considerations, Recommendations

<table>
<thead>
<tr>
<th>Ethical Considerations</th>
<th>A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caveats &amp; Recommendations</td>
<td>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.</td>
</tr>
</tbody>
</table>

Mitchell et al. [Model Cards for Model Reporting](https://modelcards.ai/). FAT*, 2019.
Disaggregated Intersectional Evaluation

### Toxicity @1

<table>
<thead>
<tr>
<th>Identity groups</th>
<th>Subgroup AUC</th>
<th>BPSN AUC</th>
<th>BNISP AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>lesbian</td>
<td>0.93</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td>gay</td>
<td>0.94</td>
<td>0.65</td>
<td>0.99</td>
</tr>
<tr>
<td>queer</td>
<td>0.98</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>straight</td>
<td>0.99</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>bisexual</td>
<td>0.96</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>homosexual</td>
<td>0.87</td>
<td>0.53</td>
<td>0.99</td>
</tr>
<tr>
<td>heterosexual</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>cis</td>
<td>0.99</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>trans</td>
<td>0.97</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>nonbinary</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>black</td>
<td>0.91</td>
<td>0.85</td>
<td>0.95</td>
</tr>
<tr>
<td>white</td>
<td>0.91</td>
<td>0.88</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### Pinned AUC Toxicity Scores @1

- black straight
- black queer
- black trans
- black bisexual
- black gay
- black lesbian

### Pinned AUC Toxicity Scores @5

- black straight
- black queer
- black trans
- black bisexual
- black gay
- black lesbian

---

[Logos: Jigsaw, The False Positive]
In Summary...

- Always **be mindful** of various sorts of biases in the NLP models and the data
- Explore “debiasing” techniques, but **be cautious**
- **Identify the biases that matter** for your problem and test for those biases
- Consider this an **iterative process**, than something that has a “done” state
- Be **transparent** about your model and its performance in different settings
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
Questions?
BACKUP Slides
Moving from majority representation...
Moving from majority representation...

...to diverse representation
Moving from majority representation...
...to diverse representation
...for ethical AI
Thanks!
margarmitchell@gmail.com
m-mitchell.com

Need MOAR?  ml-fairness.com
Measuring and Mitigating Unintended Bias in Text Classification

John Li
jetpack@google.com

Lucas Dixon
ldixon@google.com

Nithum Thain
nthain@google.com

Lucy Vasserman
lucyvasserman@google.com

Jeffrey Sorensen
sorensj@google.com

More free, hands-on tutorials on how to build more inclusive ML

ml-fairness.com
Get Involved

- Find free machine-learning tools open to anyone at ai.google/tools
- Check out Google’s ML Fairness codelab at ml-fairness.com
- Explore educational resources at ai.google/education
- Take a free, hands-on Machine Learning Crash Course at https://developers.google.com/machine-learning/crash-course/
- Share your feedback: acceleratewithgoogle@google.com

Build for everyone