Justice, Fairness, Bias

The Big Three
Justice, Fairness and Bias

• You will remember that Kant was emphasizing the importance of human dignity.

• Individuals expect to be treated fairly; the violation of human dignity leads to discrimination.

• Discrimination is the unjust treatment of people based on the groups or classes they belong to. Discrimination may stem from biases.

• We often talk about algorithmic fairness, since algorithms may amplify existing economic and societal bias.
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi1, Kai-Wei Chang2, James Zou1, Venkatesh Saligrama1,2, Adam Kalai2
1Boston University, 8 Saint Mary’s Street, Boston, MA
2Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kwk@wchang.net, jameszou@gmail.com, serv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.
Discrimination from the perspective of harms

• Allocative Harms
  • Harm is defined in terms of available resources (e.g., women being paid less, risks for people of color created by harmful algorithms)
  • Easier to spot, hard to mitigate
    We are trying to optimize utilities to allocate goods in a society full of inequalities

• Representational Harms
  • Harm is defined in terms of the representation of groups and individuals.
  • They affect the narrative (e.g., word embeddings such as secretary-woman, manager-man; search for CEO resulting in a bunch of white-male men)
  • More difficult to spot, hard to mitigate

The Trouble with Bias by Kate Crawford: https://www.youtube.com/watch?v=fMym_BKWQzk
Hidden Bias: Gendering AI Technologies

• Unesco Report – I'd Blush if I could
• The report offers guidance for education and steps to address the issues to push for equality.

"The more that culture teaches people to equate women with assistants, the more real women will be seen as assistants — and penalized for not being assistant-like."

https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1
Bias in AI Systems

- Awareness
  - Understanding bias in AI systems
- Detection
  - Measuring bias in AI systems
- Mitigation
  - Fixing bias in AI systems
2. Check bias metrics

Dataset: German credit scoring
Mitigation: none

Protected Attribute: Sex
Privileged Group: Male, Unprivileged Group: Female
Accuracy with no mitigation applied is 75%
With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

Protected Attribute: Age
Privileged Group: Old, Unprivileged Group: Young
Accuracy with no mitigation applied is 75%
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics
2. Check bias metrics

Dataset: German credit scoring
Mitigation: none

Protected Attribute: Sex
Privileged Group: Male, Unprivileged Group: Female
Accuracy with no mitigation applied is 75%
With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

Protected Attribute: Age
Privileged Group: Old, Unprivileged Group: Young
Accuracy with no mitigation applied is 75%
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics
A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Harini Suresh
John Guttag
hsuresh@mit.edu
guttag@mit.edu

ABSTRACT
As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle. In this paper, we provide a framework that identifies seven distinct potential sources of downstream harm in machine learning, spanning data collection, development, and deployment. In doing so, we aim to facilitate more productive and precise communication around these issues, as well as more direct, application-grounded ways to mitigate them.

necessarily because the statement “data is biased” is false, but because it treats data as a static artifact divorced from the process that produced it. This process is long and complex, grounded in historical context and driven by human choices and norms. Understanding the implications of each stage in the data generation process can reveal more direct and meaningful ways to prevent or address harmful downstream consequences that overly broad terms like “biased data” can mask.

Moreover, it is important to acknowledge that not all problems should be blamed on the data. The ML pipeline involves a series of choices and practices, from model definition to user interfaces used upon deployment. Each stage involves decisions that can lead
Historical Bias

• It arises if the world as it is or was leads to a model that produces harmful outcomes.

• Garg et al. show that:
  • Word embeddings reflect real-world biases about women and ethnic minorities;
  • Specific adjectives and occupations become more closely associated with certain populations over time.

Word embeddings quantify 100 years of gender and ethnic stereotypes. Nikhil Garg, Londa Schiebinger, Dan Jurafsky, James Zou
Proceedings of the National Academy of Sciences Apr 2018, 115 (16) E3635-E3644; DOI: 10.1073/pnas.1720347115
Representation Bias

- Target population **does not reflect** the use population
  - Model is trained on population X and applied to population Y
  - Model is trained on the same population in different time frames
- Target population contains **under-represented groups**
  - For example, some age groups may not be represented well in the data
- **Sampling method is limited** (**sampling bias**)
  - Target population is set to X, but the data available is only a small subset of X
• Buolamwini and Gebru analyze two benchmarks to report gender and skin type distribution.

Measurement Bias

• It is difficult to choose correct proxies to measure constructs, which can be quite complex. What features measure construct X the best?
• The method of measurement can be different. (e.g., COVID tests)
• The accuracy of measurement can be different.
  • For example, systematically higher rates of misdiagnosis/underdiagnosis are reported for certain groups.
Depression or schizophrenia? Black patients more likely to be misdiagnosed

01 April 2019

Researchers examined the medical records of more than 1,600 people at a community behavioral clinic in the US.

An analysis of mental health care in the US has found that African Americans are more likely to be misdiagnosed with schizophrenia than white patients.

The Rutgers University study builds on years of evidence that clinicians' racial biases — whether conscious or unconscious — affect the types of mental-health diagnoses ethnic minority patients receive.

The latest research found that about 20 percent of African Americans who had been diagnosed with schizophrenia also screened positive for major depression — nearly six times the percentage ratio for white patients.

https://www.mentalhealthtoday.co.uk/news/awareness/depression-or-schizophrenia-black-patients-more-likely-to-be-misdiagnosed
Aggregation Bias

• Aggregate level findings may not also be observed at an individual level.
• Exploratory data analysis is important to find out any group-based trends.
• One should be very careful to conduct any analysis based on aggregated data which does not capture true relationships (e.g., correlations between variables).
Learning Bias

- **Learning bias** happens when modeling choices amplify performance disparities.
Disparate Impact on Model Accuracy

- **Differential privacy (DP)** comes with a cost, which is a reduction in the model's accuracy.
- Bagdasaryan *et al.* show that accuracy of models, trained with DP stochastic gradient descent, drops much more for the underrepresented classes and subgroups.
- This gap is **bigger** in the DP model than in the non-DP model.
- The results are reported from the sentiment analysis of text and image classification.

Evaluation Bias

What is Evaluation Bias?

• Evaluation bias occurs when the benchmark datasets (e.g., ImageNet) do not represent the use population.

• The choice of metrics can also result in evaluation bias (e.g., aggregate results, reporting one type of metric)

Gender Shades

• Buolamwini and Gebru analyze two benchmarks to report gender and skin type distribution.

Gender Shades (Evaluation/Learning Bias example)

• They use their dataset (PPB) to evaluate three commercial gender classification systems (Microsoft, IBM, Face++):

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Metric</th>
<th>All</th>
<th>F</th>
<th>M</th>
<th>Darker</th>
<th>Lighter</th>
<th>DF</th>
<th>DM</th>
<th>LF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSFT</strong></td>
<td>PPV(%)</td>
<td>93.7</td>
<td>89.3</td>
<td>97.4</td>
<td>87.1</td>
<td>99.3</td>
<td>79.2</td>
<td>94.0</td>
<td>98.3</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>6.3</td>
<td>10.7</td>
<td>2.6</td>
<td>12.9</td>
<td>0.7</td>
<td>20.8</td>
<td>6.0</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>93.7</td>
<td>96.5</td>
<td>91.7</td>
<td>87.1</td>
<td>99.3</td>
<td>92.1</td>
<td>83.7</td>
<td>100</td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>6.3</td>
<td>3.5</td>
<td>12.9</td>
<td>0.7</td>
<td>16.3</td>
<td>7.9</td>
<td>1.3</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Face++</strong></td>
<td>PPV(%)</td>
<td>90.0</td>
<td>78.7</td>
<td>99.3</td>
<td>83.5</td>
<td>95.3</td>
<td>65.5</td>
<td>99.3</td>
<td>94.0</td>
<td>99.2</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>10.0</td>
<td>21.3</td>
<td>0.7</td>
<td>16.5</td>
<td>4.7</td>
<td>34.5</td>
<td>0.7</td>
<td>6.0</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>90.0</td>
<td>98.9</td>
<td>85.1</td>
<td>83.5</td>
<td>95.3</td>
<td>98.8</td>
<td>76.6</td>
<td>98.9</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>10.0</td>
<td>14.9</td>
<td>1.1</td>
<td>16.5</td>
<td>4.7</td>
<td>23.4</td>
<td>1.2</td>
<td>7.1</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>IBM</strong></td>
<td>PPV(%)</td>
<td>87.9</td>
<td>79.7</td>
<td>94.4</td>
<td>77.6</td>
<td>96.8</td>
<td>65.3</td>
<td>88.0</td>
<td>92.9</td>
<td>99.7</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>12.1</td>
<td>20.3</td>
<td>5.6</td>
<td>22.4</td>
<td>3.2</td>
<td>34.7</td>
<td>12.0</td>
<td>7.1</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>87.9</td>
<td>92.1</td>
<td>85.2</td>
<td>77.6</td>
<td>96.8</td>
<td>82.3</td>
<td>74.8</td>
<td>99.6</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>12.1</td>
<td>14.8</td>
<td>7.9</td>
<td>22.4</td>
<td>3.2</td>
<td>25.2</td>
<td>17.7</td>
<td>5.20</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Deployment Bias

• Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.

• One good example is framing trap in the sociotechnical context.

Other: Label Bias

• **Labels reflect interpretations about data.**
• Labelling can be done by humans:
  • The use of crowdsourcing platforms in order to get a labelled dataset.
  • Domain experts annotating data (e.g., medical domain)
• Weak-ML techniques may be used to annotate data automatically.
  • Based on defined functions, heuristics etc.
Other: Human Bias / Team Bias

https://sites.psu.edu/ethicsofdatamanagement/
1. In the Image Recognition context, why do you think we often observe representation/evaluation bias?

2. How could data practitioners recognize historical bias?

3. Which bias types are difficult to detect? Why?
Removing Bias

• Anti-classification
  • Removal of protected attributes (gender, ethnicity...) and their proxies from data

• Resampling
  • Goal: Having a good distribution of examples across groups
  • If the dataset is large, a small portion of the dataset could be used to achieve this.
  • If the dataset is small, removing examples from data is expensive. One approach is to increase the size of samples from underrepresented groups in the training set.