Ethics, Fairness, and Bias in Machine Learning

17 October 2023

Zeerak Talat
MBZUAI
z@zeerak.org | @zeeraktalat
Discussion-based lecture

Today:
- Natural Language Processing
- Computer Vision
- Machine Learning
- Cross-polination between NLP/CV
- LLMs
- Your Individual responsibility
- Discussing case studies
At the end of this lecture you should have

- An idea of how to think about ethical issues
- An understanding of your role in research and development of AI tools
- A way to question what you think is acceptable work
- An understanding of how (AI) technologies influence the world around us

*Crucially: How your perspective can change AI to make it less harmful.*
Why should you care?

Created a dataset for hate speech detection and got this table of features.

Q: What problem arises when you look at these features?

<table>
<thead>
<tr>
<th>Feature (sexism)</th>
<th>Feature (racism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'xist'</td>
<td>'sl'</td>
</tr>
<tr>
<td>'sexi'</td>
<td>'sla'</td>
</tr>
<tr>
<td>'ka'</td>
<td>'slam'</td>
</tr>
<tr>
<td>'sex'</td>
<td>'isla'</td>
</tr>
<tr>
<td>'kat'</td>
<td>'l'</td>
</tr>
<tr>
<td>'exis'</td>
<td>'a'</td>
</tr>
<tr>
<td>'xis'</td>
<td>'isl'</td>
</tr>
<tr>
<td>'exi'</td>
<td>'lam'</td>
</tr>
<tr>
<td>'xi'</td>
<td>'i'</td>
</tr>
<tr>
<td>'bite'</td>
<td>'e'</td>
</tr>
<tr>
<td>'ist'</td>
<td>'mu'</td>
</tr>
<tr>
<td>'bit'</td>
<td>'s'</td>
</tr>
<tr>
<td>'itch'</td>
<td>'am'</td>
</tr>
<tr>
<td>'ite'</td>
<td>'m'</td>
</tr>
<tr>
<td>'fem'</td>
<td>'la'</td>
</tr>
<tr>
<td>'ex'</td>
<td>'ls'</td>
</tr>
<tr>
<td>'bi'</td>
<td>'slim'</td>
</tr>
<tr>
<td>'irl'</td>
<td>'musl'</td>
</tr>
<tr>
<td>'wom'</td>
<td>'usli'</td>
</tr>
<tr>
<td>'girl'</td>
<td>'lim'</td>
</tr>
</tbody>
</table>

Table 5: Most indicative character n-gram features for hate-speech detection
Content moderation is the process of determining what is acceptable and what is not. Models reproduce what is available to them in their datasets.

Q: What issues arise with this approach?
Q: How can we address such issues, without changing data or model?
Q: What are the limits of those approaches?
ML & Statistics

A little history: Francis Galton and Eugenics

Goal: Classify good/bad human traits

Method

Average intra-group diffs
Highly inter-group diffs

A Galtonian Composite as shown by Alan Sekula: The Body and the Archive (1986). October. MIT Press
The Distributional Hypothesis

The Distributional Hypothesis describes a frequentist approach to salience: What frequently co-occurs should be treated related.

For NLP Tokens frequently co-occurring with the same tokens ➞ Similar semantically
For ML Frequently co-occurring patterns ➞ Highly salient patterns

Used to draw decision boundaries between classes & cluster information within classes.
The Distributional Hypothesis

Infrequent information at the edge of the vector space → incorrect classification / Infrequent generation

Full breadth of data impossible to collect

Q: Would this change with a full sample?

Q: When does it matter that we have a narrow sample?
# Computer Vision: Face Recognition

<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><strong>100</strong></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><strong>20.8</strong></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><strong>100</strong></td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>6.3</td>
<td>8.3</td>
<td>3.5</td>
<td>12.9</td>
<td>0.7</td>
<td><strong>16.3</strong></td>
<td>7.9</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><strong>99.3</strong></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><strong>34.5</strong></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><strong>98.9</strong></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><strong>23.4</strong></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><strong>99.7</strong></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><strong>34.7</strong></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><strong>99.6</strong></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><strong>25.2</strong></td>
<td>17.7</td>
<td>5.20</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).
Computer Vision & NLP: ImageNet

(a) Class-wise counts of the offensive classes

(b) Samples from the class labelled n*****r

Figure 1: Results from the 80 Million Tiny Images dataset exemplifying the toxicities of it’s label space

Prahbu and Birhane
Attempted (technical) Solutions: ML

Protected Attribute: $A$ (e.g., Male/Female person)
Predicted Class: $O$ (An outcome e.g., gets admitted to MBZ)
Predictive attribute: $Y$ (e.g., Variable that indicates degree attainment)

Demographic Parity:

$$
P(O = 1 | A = 0) = P(O = 1 | A = 1)
$$

Equalized Odds:

$$
FNR = P(O = 0, A = 1, Y = 1) = P(O = 0 | A = 1, Y = 0)$$
$$
FPR = P(O = 1, A = 1, Y = 1) = P(O = 1 | A = 1, Y = 0)
$$
Attempted (technical) Solutions: CV
Attempted (technical) Solutions: NLP

Measurement (e.g., Bolukbasi et al., 2016; Nangia et al., 2020)
Debiasing vector representations (e.g., Bolukbasi et al., 2016; Kaneko and Bollegala, 2019)
Counter-factuals / Invariance (e.g., Liu and Avci, 2019)
Value alignment (e.g., Solaiman and Dennison, 2021)
Attempted (technical) Solutions: NLP

Bolukbasi et al. 2016
### Attempted (technical) Solutions: NLP

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>I am gay</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td>I am straight</td>
<td>0.085</td>
</tr>
<tr>
<td>Our Method</td>
<td>I am gay</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>I am straight</td>
<td>0.144</td>
</tr>
</tbody>
</table>

Table 1: Toxicity probabilities for samples of a baseline CNN model and our proposed method. Words are shaded based on their attribution and italicized if attribution is $> 0$.

Liu and Avci, 2019.
Attempted (technical) Solutions: NLP

Value Alignment
- Just prompt engineering and penalizing models for bad completions
- Also what is done using RLHF
Evaluation Paradigms

Intrinsic - Fixing model representations (i.e., gender bias in model representations)
Extrinsic - Evaluating on a downstream task (i.e., discriminatory classifications)

Q: Which evaluation paradigm would you prefer? Why?
Generative AI

Alignment with human values
  Done through fine-tuning on datasets
  Through RLHF
Blocklists
Partial Views and Subjective Knowledge

A particularly starry night in August
Partial Views and Subjective Knowledge

A particularly starry night in August
Partial Views and Subjective Knowledge

Our knowledges and experiences provide the background for how we view the world.

E.g., Face Recognition example

Partial views are okay — important thing is to critically examine what we might be missing.
Summary

Discussed different ethical issues
  Content moderation
  Face Detection
  The distributional hypothesis / Frequency
Generative AI and its issues
Different approaches to addressing harms
How we as researchers impact technology
References


Case I: Autonomous Weaponry

Case II: Advertisement

Case III: Combining Disparate Sources

Case IV: Automatic Speech Recognition

**TECH**

Prisons are using Amazon Transcribe and AI to monitor inmates’ phone calls

A new report sheds light on companies like LEO Technologies, whose AI-scanning audio software employs Amazon speech-to-text recognition.

Andrew Paul. Prisons are using Amazon Transcribe and AI to monitor inmates’ phone calls, 2021