A Module for Introducing Ethics in AI: Detecting Bias in Language Models

Professor Ameet Soni
Professor Krista Thomason
Swarthmore College

EAAI 2020, New York City, 02/08/2020
Course Context: Ethics and Technology

- Co-taught between Comp. Science and Philosophy
- Non-major course for CS
- First-Year Seminar
- 6 students with CS background, 6 with none

Course Syllabus available: https://works.swarthmore.edu/dev-dhgrants/28
Motivation: Bias in Algorithms

Machine Bias
There’s software used across the country to predict future criminals. And it’s biased against blacks.
What this assignment wants to achieve

• Inspect a real-world algorithm/machine learning model
  • Not a toy example
  • Use an algorithm with clear benefits

• Understand the difficulty of identifying bias

• Get students thinking about ethics as part of the design process

• Identify that responsibilities lie with many actors (clients, coders, project managers, society, government, etc.)
Assignment Setup


- (Optional) Introduce word embedding models
  - What is natural language processing?
  - Notion of using *co-occurrence* to understand meaning
  - *Representation* and *data* source are design choices

https://www.tensorflow.org/tutorials/representation/word2vec
Lab Practicum Part 1: Learning Embeddings

• Given: learned word embeddings
  • GloVe algorithm [https://nlp.stanford.edu/projects/glove/](https://nlp.stanford.edu/projects/glove/)
  • Training corpus: twitter, wikipedia, web

• Goal: validate the usefulness of word embeddings

```
$ ./findSimilarWords.py web aaai 5
Printing 5 most similar words to aaai

<table>
<thead>
<tr>
<th>word</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ijcai</td>
<td>0.485</td>
</tr>
<tr>
<td>usenix</td>
<td>0.422</td>
</tr>
<tr>
<td>ecai</td>
<td>0.408</td>
</tr>
<tr>
<td>sigmod</td>
<td>0.395</td>
</tr>
<tr>
<td>acm</td>
<td>0.387</td>
</tr>
</tbody>
</table>

$ ./findSimilarWords.py twitter nyc 5
Printing 5 most similar words to nyc

<table>
<thead>
<tr>
<th>word</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>chicago</td>
<td>0.900</td>
</tr>
<tr>
<td>toronto</td>
<td>0.887</td>
</tr>
<tr>
<td>downtown</td>
<td>0.876</td>
</tr>
<tr>
<td>vegas</td>
<td>0.874</td>
</tr>
<tr>
<td>nashville</td>
<td>0.869</td>
</tr>
</tbody>
</table>
```
Real-World Example: Gender Bias in Google Translate

In Turkish, *o* is a gender neutral pronoun (*he, she, or it*)

<table>
<thead>
<tr>
<th>Turkish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>o bir aşçı</td>
<td>she is a cook</td>
</tr>
<tr>
<td>o bir mühendis</td>
<td>he is an engineer</td>
</tr>
<tr>
<td>o bir doktor</td>
<td>she is a doctor</td>
</tr>
<tr>
<td>o bir hemşire</td>
<td>he is a cleaner</td>
</tr>
<tr>
<td>o bir temizlikçi</td>
<td>He-she is a police</td>
</tr>
<tr>
<td>o bir polis</td>
<td>he is a soldier</td>
</tr>
<tr>
<td>o bir asker</td>
<td>She's a teacher</td>
</tr>
<tr>
<td>o bir öğretmen</td>
<td></td>
</tr>
</tbody>
</table>


Lab Practicum Part 2: Word Embedding Association Tests

- Modeled after Implicit Association Test
- Target: object affected by bias
  Race, gender, religion, etc.
- Attribute: descriptors
  Pleasant, math, career, etc.
- Null hypothesis: attribute choice is independent of task performance
- WEAT: association between target words and attribute words is measured by vector similarity
Gender bias (from Caliskan et al paper)

- **Targets:**
  - Female names: Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna, ...
  - Male names: Adam, Chip, Harry, Josh, Roger, Alan, Frank, Ian, ...

- **Attributes:**
  - Career: executive, management, professional, salary, office, ...
  - Family: home, parents, children, family, marriage, relatives, ...

- Calculate pairwise similarities: female/career, female/family, male/career, male/family

- Null hypothesis: there is no difference based on gender
Wikipedia Gender Bias

Similarities Scores for Target/Attribute Pairs

Difference Scores For Each Target
Twitter Gender Bias

![Graph of Similarities Scores for Target/Attribute Pairs]

![Graph of Difference Scores For Each Target]
Twitter Racial Bias
Additional Exercises

Part 1: compare and contrast biases across different data sets
twitter, wikipedia, web, ...

Part 2: create a new bias to test; design experiment and run WEAT test

Part 3: case study
elaborate on ethical concerns that you may have if you were consulting
a hospital on using NLP for triage support
Strengths and Challenges

• No prior programming experience required
• Easily adapted for non-major, intro, or upper-level courses
• Assignment setup troubleshooting is low

• Ethics discussions are not straightforward
  • Students can get frustrated with the lack of clarity/easy answer
  • CS faculty are not naturally trained in these types of discussions
• This is a very specific tool for a very specific model for a very specific form of bias
Thank you!

Questions?
Student Feedback

• Received highest usefulness rating of all assignments in the course (mean: 4.0 out of 5)

• Ethics of AI, Bias in Algorithms were the two highest rated modules in the course (top choice for 66% of students)
I liked the lab practicum! Testing out and seeing the word biases made me realize how difficult it is to separate biases from machine learning.

[The] lab practicum helped me to think through the material by connecting the ethical frameworks that we learned to the technology that we were learning about.

Lab practicum showed how machine learning can be biased using real data. Very helpful to the more CS-oriented student.

Going through the data and correlations [sic] between groups as a class was more helpful to understanding bias than was writing a lab report.
Real-World Example: Gender Bias in Google Translate

In Turkish, *o* is a gender neutral pronoun (*he, she, or it*)

- o bir açıcı
- o bir mühendis
- o bir doktor
- o bir hemşire
- o bir temizlikçi
- o bir polis
- o bir asker
- o bir öğretmen