Measuring social biases in human annotators using counterfactual queries in Crowdsourcing

BHAVYA GHAI
PhD Candidate, Computer Science Department
Stony Brook University
Adviser: Prof. Klaus Mueller
Algorithmic Bias

When Algorithms exhibit preference for or prejudice against certain sections of society based on their identity. Such discriminatory behavior is termed as Algorithmic bias.

- Generally emanates from biased training data
- Minorities & underrepresented groups are worst hit.
- Which sub-domains of AI are affected? ALL

Algorithmic Bias is the imminent AI danger impacting millions daily.

Biased Algorithms Are Everywhere, and No One Seems to Notice

Forget Killer Robots—Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.

Algorithm mistakenly supplemented 'gorillas' and identified Teachers With a Bad Algorithm

The Value-Added Model has done more to confuse and oppress than...
Motivation

Tons of work has been done to prevent bias in these stages!!

Unlabeled Data

Human Annotator

Labeled Dataset

Model

Interpretation

Tackling Algorithmic bias in the crowdsourcing stage hasn’t been explored

Crowdsourcing for Machine Learning

We focus on Subjective labeling tasks because implicit bias may play a key role

E.g.- identify interesting tweet, best movie

E.g.- image labeling, transcribe audio

When Crowdsourcing got biased datasets

Crowdsourcing is not immune to social biases & may lead to Algorithmic bias

Sources of Bias

**Label Bias:** If the distribution of positive outcomes is skewed with respect to a demographic group

**Selection bias:** Samples chosen for labeling don’t represent the underlying population.

For e.g. Consider a graduate admissions scenario.

<table>
<thead>
<tr>
<th>CGPA</th>
<th>Gre_Verbal</th>
<th>TOEFL</th>
<th>Gender</th>
<th>International</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>168</td>
<td>117</td>
<td>Male</td>
<td>No</td>
<td>✓</td>
</tr>
<tr>
<td>3.7</td>
<td>165</td>
<td>119</td>
<td>Male</td>
<td>No</td>
<td>✓</td>
</tr>
<tr>
<td>3.4</td>
<td>167</td>
<td>118</td>
<td>Male</td>
<td>No</td>
<td>✓</td>
</tr>
<tr>
<td>3.8</td>
<td>155</td>
<td>106</td>
<td>Female</td>
<td>Yes</td>
<td>✓</td>
</tr>
<tr>
<td>3.9</td>
<td>160</td>
<td>108</td>
<td>Male</td>
<td>Yes</td>
<td>❌</td>
</tr>
<tr>
<td>3.7</td>
<td>157</td>
<td>110</td>
<td>Male</td>
<td>Yes</td>
<td>❌</td>
</tr>
</tbody>
</table>

In this study, we are just focused on Label bias
Types of Labelers

**Naive**

**Expert**

**Spammer**

**Adversarial**

**Biased** – A human annotator infested with serious social biases based on gender, race, etc. which are reflected in his/her labels. Their labels might reflect strong preference for or prejudice against a demographic group.

**In this study, we are trying to identify & control for biased labelers**
Our objective is to devise a new technique for measuring Individual Performance Label Quality control.

- Individual performance
  - Reputation score
  - Gold Questions
  - Self reported data

- Aggregation algorithms
  - Majority Voting
  - EM Algorithm
Reputation Score

Based on worker’s past performance. Eg.- percentage of previously approved HITs

Snippet from Amazon MTurk

. Drawbacks

- Requesters are approving HITs more than they should, thereby inflating workers’ reputation levels\(^1\)

- It is possible, that a biased user might achieve high reputation score by performing several objective tasks, so qualifies for a subjective task where his/her response(s) might be biased

Gold Questions

- Gold questions are the tasks for which ground truth is available. It’s one of the most common ways to evaluate noisy labelers like spammers, etc..

- If a worker correctly answers more than a threshold of gold questions, he/she is considered eligible for the study.

- Knowing how often someone is right is important. But in the context of social biases, it’s equally important to know when someone fails

High accuracy on Gold Questions doesn’t always mean low bias

Overall Accuracy: 75%
Male Accuracy: 100%
Female Accuracy: 33%
Self Reported data

Survey Questionnaire

1. No matter how accomplished he is, a man is not complete as a person unless he has the love of a woman.

2. Most women interpret innocent remarks or acts as being sexist.

3. Most women fail to appreciate what all men do for them.

4. When women lose to men in a fair competition, they typically complain about being discriminated against.

5. Women, as compared to men, tend to have a more refined sense of culture and good taste.

- One of the only measures designed to capture implicit social biases.
- The content of survey questions is quite different from the study. Hence, they make crowd workers conscious that they are being judged.
- Suffer from Social desirability & Social approval bias.
- Not very engaging.
- Inaccurate

It can serve as a good baseline for upcoming techniques to measure social bias

Our approach - Counterfactual Queries

Counterfactual tries to estimate the outcome in a hypothetical world where a different treatment was given.

In ML literature, an ML model is considered counterfactually fair if

\[ P(Y|X, A=1) = P(Y|X, A=0) \]

where A is the sensitive attribute like gender, race, etc.

We are trying to adopt this technique to identify biased workers in Crowdsourcing. Counterfactual query is created by flipping the sensitive attribute of the original query

Hypothesis: Unbiased worker will give consistent labels for counterfactuals

Use case- Toxic Comment classification

Rate the following statements on toxicity (1 to 10 scale) where 1 is non-toxic and 10 is highly toxic

**Q:** Homosexuality is a disease that must be cured

**CQ:** Heterosexuality is a disease that must be cured

\[
Worker\ Bias\ score = mean(|\ \text{Label}(Q) - \text{Label}(CQ)|)
\]

If Bias score > \(\lambda\) (threshold) => Worker is biased

Doesn't need Ground truth & blends with the task perfectly!
Conclusion & Future Work

- Datasets curated via crowdsourcing maybe polluted by social biases of crowd workers and may eventually lead to Algorithmic bias.

- Need for new label quality control techniques which incorporate fairness metrics apart from accuracy.

- Counterfactual queries can be one way to capture social biases without having any ground truth.

- Next, we intend to conduct a user study to test existing techniques and compare with our approach.
Thanks for your attention!

For any Questions, suggestions, feedback, criticism, please email me at:- bghai@cs.stonybrook.edu

Bhavya Ghai