Addressing Bias Against the Poor in Artificial Intelligence by Observing Individual Bias in Twitter

Imani Mathenge

Abstract

With an increase in Artificial Intelligence (AI) usage, there is an increasing demand to address biases and implement fairness in AI. This white paper focuses on the importance of first recognizing and addressing societal biases against the poor to address biases against the poor in AI. The results of the study are from the first collection of data collected from Twitter. The data is used to provide an index of how people in societies perceive poor people.

Why AI skeptics should care

- Bias, particularly bias against poor people, is not just present in AI. Humans also display biases.
- Addressing human biases can help address and minimize AI biases because algorithms are programmed by and learn from human behavior.
- AI has already been adopted in many industries like healthcare and finance. So, the focus should be less of eliminating AI and more of improving it.

Introduction

There is an increasing reliance on AI in decision making, yet algorithms are not neutral (Curto, Jojoa Acosta, Comim, Garcia-Zapirain). Whether biases are positive or negative, they are unfair. Some people attribute biases in AI algorithms as technological problems. Whereas others relate those biases to societal or user biases (Kirsten). Societal biases are impactful especially when AI algorithms learn from previous inputs or data. Common biases discussed are gender and racial biases; however, socio-economic biases are also a cause for concern.

For the purpose of this white paper, bias is defined as the tendency for algorithms to evaluate certain individuals differently (Kirsten). In this paper, I use Adela Cortina's idea of prejudice towards the poor and eventual rejection of poor people to identify societal bias against the poor.

Cortina, a Spanish philosopher and professor of Ethics and Political Philosophy, coined the term "aporophobia." because "we can't recognize something, we can't name" (Cortina). Aporophobia is more than not showing compassion towards others. It is an acquired feeling of hatred and fear of the poor. That baseline is one which will be used when categorizing tweets as showing prejudice or not. Cortina's reason for coining the term is relevant to the discussion

of biases in AI because in recognizing biases, people have more control and get to choose whether to act on their biases.

Although human biases are a cause for concern, human involvement is sometimes necessary. For example, when casework was automated in Indiana, "removing human discretion from frontline social servants and moving it instead to engineers and private contractors, the Indiana experiment supercharged discrimination" (Eubanks). Eubanks suggests that people in different occupations have different information and biases. Eubanks does not blame the engineers and private contractors. She rather suggests that the decision-making process in casework should rely heavily on how caseworkers make decisions without automation.

This paper contributes to the discussion of societal biases by providing an approach to seeing biases against poor people and the results of the approach.

Approach

To find out how societies view the poor, I observed Twitter tweets from several countries. Svetlana Kiritchenko, Ph.D. pulled over 1,420,000 tweets in the month of November. Those tweets were related to target words such as poor, poor people, poor families, homeless, welfare recipients, low-income, and disadvantaged. The tweets were also organized in 142 topics with some topics including the target words. Because the data process and topic modeling were done automatically and the word poor being polysemic some of the tweets were not about poor people and were less coherent.

Georgina Curto Rex, Ph.D manually selected eight of the most related topics for preliminary evaluation. 99 tweets were in the 8 topics selected. The goal was to create a baseline to see whether I and future labelers can come to an agreement on how to categorize the tweets.

We had six initial categories that mainly intersected with other forms of biases like xenophobia or racism. After going through the tweets, we modified the categories because there was a need for specificity and focus on just prejudice against the poor. The final categories used for labeling were divided into two large categories: direct and reporting. The direct tweets are tweets in which the author describes their personal experience. The "reporting" category is for the tweets when the person who is talking is reporting the action.

For each of the two categories there were more specific categories: belief (ex. the poor are lazy), negative belief (ex. the poor are not lazy), avoidance (ex. I do not want to be next to the poor), antilocution (ex. we do not care what the poor think), discrimination (ex. the poor cannot go to the concert), physical attack (ex. I want to kill the poor). The expansion of the categories allowed us to focus on biases people may have on the poor.

Because not everyone uses Twitter, tweets do not represent everyone in the population; however, the tweets do represent some people. This limits our ability to accurately understand the whole population but still gives us an index of how poor people are viewed.

Findings

From the topic modeling, Kathleen Fraser, Ph.D. created a word cloud using the initial 1,420,000 tweets. The word cloud illustrates that common associations with the poor are kill, criminals, white, black. The problem is that it did not account for possible interpretation mistakes. For example, a tweet that read "Yess kill the poor!" would be interpreted literally but the reality could be that the person was expressing sarcasm or something else. Therefore, manually going through tweets was needed to leave out unclear tweets.

I expected labeling and categorizing tweets to be easier and produce more definite results. However, that was not the case. Many tweets needed additional context to them. For the first topic, I categorized 14% of the tweets. These tweets were ones that discussed addiction. The second set of tweets were associated with race, and I categorized 6% of the tweets. The third set of tweets were related to immigrants, and I labeled 5% of the tweets. The fourth set of tweets were related to crime. I labeled 23% of them. The fifth set of tweets were related to hatred. I labeled 34% of the tweets. The sixth set of tweets were related to a specific crime, stealing. I labeled 3% of them. The seventh set of tweets were related to taxes, and I did not label any of the tweets. The last set of tweets were related to blame. I labeled 2% of the tweets.

These findings suggest that there is evidence of some prejudice towards the poor, especially when the tweets mentioned hate and crime. The low labeling rate reflects how hard it is to spot bias.

Practical Application

Generalizing is important and allows humans to think broadly about topics. It is important though to be careful when generalizing based on just personal experience. Many tweets pulled from the study described homeless people as drug abusers. While making a general statement like "many homeless people are drug abusers'" would not be put into any category or marked as prejudice, this generalization can easily become prejudice when it is applied to every person that is homeless. A way in minimizing person bias is by increasing the information you have before making conclusions about people. Because it is easier to point out your own biases than it is to point out others', encouraging people to introspect about their attitudes towards people who are poor.

Next Steps

I conclude that there is evidence of negative attitudes towards the poor on Twitter. For a proper index and less biased index, there needs to be more data from beyond social media. The data used is from a pool of people who use Twitter. So, the results reflect Twitter users' experiences and viewpoints.

To move out of convenience sampling and provide a better measure of societies' views on poor people, information would be to include data from polls or even Google searches. Additionally, having multiple people from different backgrounds agree on the interpretation of tweets will

make the categorizations of the tweets more reliable. In general, it is harder to point out others' biases. So, using Twitter was a good place to look for frank viewpoints.

References

Cortina, Adela (2022). *Aporophobia: Why We Reject the Poor Instead of Helping Them.* Princeton University Press.

Curto, G., Jojoa Acosta, M.FComim, F. et al. *Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. AI & Soc (2022).* https://doi.org/10.1007/s00146-022-01494-z

Eubanks, Virginia. "Automating Eligibility in the Heartland." *Automating Inequality: How Hight-Tech Tools Profile, Police, and Punish the Poo*r, New York, 2018, p. 81.

Martin, Kirsten. *Ethics of Data and Analytics: Concepts and Cases. First edition.*Boca Raton, FL: CRC Press, 2022.

Parikh RB, Teeple S, Navathe AS. *Addressing Bias in Artificial Intelligence in Health Care.* JAMA.2019;322(24):2377–2378. doi:10.1001/jama.2019.18058