Equitable Artificial Intelligence: A Guide for Meta-Analysis of Techniques for De-Biasing Machine Learning Models

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Sophie Meyer

Spring, 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Computer Science

ABSTRACT

Biased outputs based on machine learning can have real and devastating effects on people; for instance, the denial of loans and Green Cards based on biased machine learning models. Proposed methods for debiasing focus on everything from refining data collection to improving deployment of the machine learning model. I propose a systematic evaluation and meta-analysis of the data to determine the best existing method or combination of methods. The anticipated result of such an evaluation would be a strategy to improve the use of machine learning models and reduce inequality perpetuated by biased artificial intelligence. Future research includes the execution of a de-biasing meta-analysis and additional research into the most promising methods.

1. INTRODUCTION

Machine learning is often perceived as an objective method of evaluation, but in actuality machine learning models can share the same biases as the humans who create them. According to mathematician, data scientist, and former hedge fund trader O’Neil: "Models are opinions embedded in mathematics" (O’Neil, 2016, p. 8). This sentiment becomes extremely apparent when examining the machine learning models involved in loan approval, where racial biases affect the models irrespective of other factors. O’Neil refers to these biases as “weapons of math destruction” (WMDs).

Housing loans fall prey to these WMDs, as described by O’Neil, who cites a study that determined a disturbing fact about a certain unnamed bank’s loan approval machine learning model. It was less likely to approve loan from zip codes that had higher percentages of Black people. This was controlling for income, credit score, and all other outside factors. Race alone had determined loan approval. Since this study was published IN DATE, there have been numerous attempts at counteracting biases embedded in machine learning models. Some attempts have been successful, at least in certain contexts, while others leave something to be desired. Determining which de-biasing methods are best is essential to avoiding creating WMDs. In order to do so, inclusion criteria must be established.

2. RELATED WORK

O'Neil (2016) introduces different ways that bias is expressed in machine learning, and what types of algorithms are particularly dangerous. Specifically, those that cause the most harm are large scale, opaque, and have the propensity to cause harm. O'Neil describes car insurance algorithms that give people much higher rates because of credit score, but barely factor in DUIs. Teachers are assessed (and fired) using opaque algorithms that do not relate to student or parent satisfaction. She also describes the aforementioned racially biased mortgage approval system.

In a frequently cited and influential paper, Fu, et. al. (2021) discussed machine learning as an alternative to crowd lending. Crowd lending is a form of raising funds for a business without using a bank. Instead, crowds of investors choose which businesses to loan money to. When a machine learning algorithm was implemented and tested, it outperformed crowds in regard to predicting which businesses would default on their loans. Additionally, using the model led to a higher rate of return for lenders and more "funding opportunities for borrowers with few alternative funding options" (p. 89). However, the researchers found that the model was biased in race and gender, even when these factors were not explicitly used as parameters. The authors attributed this to redundant encodings, features that the model used as proxies for gender and race. They then implemented a debiasing algorithm to remove these redundant encodings so that race and gender cannot be inferred. Fu, et al. also explain why their method is best for this application, contrasting it with other methods that are only useful for certain types of machine learning algorithms or only certain tasks. Overall, their debiased model had only a small reduction in accuracy for the test set, around 2%.

A 2021 study on natural language processing (NLP) by B, et. al. states that "Data in general encodes human biases by default" (p. 1). This is a common sentiment in the machine learning community. If humans are biased, then those biases are reflected and "encoded" in algorithms created based on humans. NLP is the branch of artificial intelligence dealing with the processing and recognition of human speech. NLP is commonly used in voice assistants like Siri and Alexa. The B, et al. paper is, in itself, similar to a small meta-analysis, with methods taken from other papers and combined.

3. PROPOSED DESIGN

There are several steps to performing a meta-analysis of debiasing techniques.

First, one must find reputable studies, then find the techniques outlined in these studies, and determine whether they seem like a good fit for the meta-analysis.

Next, the techniques found in the studies must be sorted. Not all debiasing methods work for all applications, so they must be sorted by application in order to compare apples to apples. Some methods are limited to only certain types of machine learning algorithms such as tree-based models. Other methods are limited to only certain types of tasks, like binary classification (placing something into one of two categories, like labelling a picture as a dog or a cat). The techniques that have been chosen for analysis should be categorized by both the type of algorithm that they can be used on and the type of task that they can be used for.

Next, tests must be developed for each category that the techniques have been sorted into. These tests must have criteria for what makes a “good” or “bad” algorithm and may look at correlations between outcomes and demographics. Finally, tests must be performed and the results gathered and analyzed.

4. ANTICIPATED RESULTS

There will likely be multiple different best techniques depending on the category that is being analyzed. Deciding on how to test will likely be difficult and may determine the “winners” of each category Overall, if combinations of techniques are tested, that may lead to the best results. Such combinations might, for example, start by removing biased data before it trains the model, then adding in additional randomness to reduce likelihood of redundant encodings, and finally changing the algorithm to detect possible signs of bias. The results would be helpful to future researchers and developers.

5. CONCLUSION

Bias is very common in machine learning models, and it can have very real effects on peoples’ lives. A meta-analysis of de-biasing techniques is essential as artificial intelligence becomes more and more a part of our lives. Such a meta-analysis could be performed by finding reputable studies, sorting the de-biasing techniques, and testing the techniques to determine the effectiveness of each category.

6. FUTURE WORK

Future work includes performing the de-biasing meta-analysis. The results from this research can be used to guide further exploration and development of the techniques that are found to be most promising. Additionally, as the field of machine learning is constantly evolving, new examples of bias in machine learning will appear, and those must be examined and mitigated with similar analyses.

REFERENCES

O’Neil, C. (2016). *Weapons of Math Destruction*. Penguin Books.

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, Lending, Machine, and Bias. *Information Systems Research*, *32*(1), 72–92. https://doi.org/10.1287/isre.2020.0990

B, S. K., Chandrabose, A., & Chakravarthi, B. R. (2021). An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline. Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 34–45. https://aclanthology.org/2021.ltedi-1.5