Pilot Study: Computer Scientists' Understanding of Modifying Datasets to Achieve a Fairer Dataset/Model

Hossein Abedi
7930589
abedikhh@myumanitoba.ca

Soheil Changizi
7905520
changizs@myumanitoba.ca

March 10, 2023

Abstract

Data is one of the most crucial components of machine learning. Despite this, we believe very little attention is given to the quality of data in academia. To investigate our idea, we designed a pilot study and conducted semi-structured interviews with 11 participants who previously studied machine learning in university. By performing a qualitative analysis of our data, we extracted several themes on participants' understanding of data, how they learned ML, and what they think about dataset modification practices.

1 Introduction

Even though data is an essential component of machine learning, it is rarely discussed in teaching materials or university courses. The main focus of many institutions and online courses is the theoretical elements of machine learning. Looking through the top 10 online courses, only a handful of them discuss data work that is limited to simple data preprocessing and small datasets [1]. As for the courses taught in universities, we can see that there is no mention of curating and cleaning datasets in Stanford's machine learning course (CS229) syllabus [2].

Another critical topic rarely discussed in academic courses is fairness in machine learning. CS students either do not get introduced to fairness and responsible AI or are rushed through condensed and simplified overviews of broad ethical thinking. This can cause students to be unprepared for these issues and make wrong decisions when implementing machine learning models [3].

Since these courses emphasize more on theory aspects of ML, the datasets utilized for instruction are either relatively simple or have had the relevant operations already performed on them [2, 1]. Also, courses that try to allow students to collect and process their datasets use simple examples, such as collecting handwriting datasets or cats/dogs images, aimed at confronting students with model-related challenges rather than data work [1]. On the other hand, academia follows the same path by using standardized datasets such as UCI [4] as benchmarks. While this strategy allows for creating unique architectures and machine learning models such as ResNet [5], GoogleLeNet [6] and AlexNex [7], it also minimizes the value of data gathering, preparation, and processing. Papers in ML are often focused on abstract evaluation metrics such as RMSE and accuracy, and there is a lack of connection between active research and relevant real-world problems [8].

As AI becomes more integrated into everyday decision-making, the sanctity, and quality of the data used to run these models become increasingly important. Ignoring datasets throughout a machine learning application development will result in a cascade of damaging effects leading to technical debt such as fairness issues over time [9].The data gathered from the real world from communities can have various biases [10] and paying close attention to data and data-related procedures can help control the fairness issue [10, 9].

In this paper, we conducted a pilot survey of graduate students studying or working with machine learning to see how well they understand the relevance of data and how much academia has contributed to their understanding. Furthermore, we asked them how much they knew about algorithmic unfairness present in machine learning and their views on dataset debiasing methods.

The literature on existing data debiasing toolkits and related studies on fairness issues is presented in the next section. In section 3 we discuss our methodology for conducting semi-structured interviews for this study and provide our early hypotheses. We provide our results and analysis in sections 4 and 5. After noting our our limitations in section 6, we conclude our study in section 7.

2 Related Work

2.1 Debiasing tools

Open-source auditing toolkits like Aequitas [11] and LIME [12] can be used to evaluate a trained model for any discriminatory impact from unfair datasets, while toolkits such as AI Fairness 360 [13] also eliminate some of these biases but in limited problem settings.

Aequitas provides a user interface for developers and policymakers to evaluate a prediction model based on multiple fairness metrics and a decision tree that directs the selection of the fairness notion. The tool then creates a bias report based on a per-subgroup basis, indicating which ideas were violated and to what extent. Aequitas can check AI systems for biases in actions or results based on inaccurate assumptions about different demographic groups [11].

LIME toolbox evaluates feature importance and explains local behavior using a highly interpretable linear or tree-based model. It will use these interpretable models to fit the model's (the one we try to unbias) predictions. We can learn how the network works by training a linear model to mimic its behavior. After that, human decision-makers can use LIME to analyze the model's decisions and make a final decision [12].

AI Fairness 360 is a more comprehensive toolbox that includes an interface for detecting and mitigating the unfairness of an algorithmic system. It includes over 70 different metrics for individual fairness, group fairness, and general classification measurements, allowing for the development of fairness notions. The AI Fairness 360 package (AIF360) offers pre-processing algorithms that balance the dataset, in-processing algorithms that penalize undesirable bias as the model is being built, and post-processing algorithms that balance out the results following a prediction [13].

The bias detection and mitigation methods in AIF360 can only work for binary classification problems and do not cover multiclass and regression models. Aequitas and LIME, on the other hand, have good metrics for some more complex models, but they only detect bias and can not automatically fix these issues. However, knowing that a model is biassed before it goes into production can be helpful. These three tools are accessible on GitHub and can be used as a Python library.

2.2 Related studies

Using these toolkits to mitigate and assess fairness may come with potential risks. ML practitioners may misuse these tools in which they might change or remove critical features. They can also misinterpret the reported metrics or employ the incorrect metric, leading to not finding the core source of the bias. Lastly, they also argue about the limited consideration of real-life circumstances, which indicates a gap between the showcased test benches in these toolkits and real-world use cases. Lee and Singh looked at six popular open-source fairness toolkits (including Aequitas and AI Fairness 360), outlined their benefits and drawbacks, and showed what ML practitioners must know about their functionality and usability [14].

It is also necessary to figure out why people believe particular features in algorithms are fair or unfair. It is argued that people's opinions on the fairness of using features can be used to learn how to make fair algorithmic decisions [15]. Also, by asking, "Is this feature fair to use?" researchers can see how different fairness factors influence their decisions. Grgic-Hlaca et al. also investigated how people feel about the role of several characteristics in predicting criminal recidivism risk. They discovered that people's concerns extend beyond prejudice and include qualities like relevance and dependability [16]. In our study, we want to see the computer scientists' opinions on modifying dataset attributes and samples to achieve a degree of fairness.

Paullada et al., in their survey, examined the various stages of dataset analysis. Starting with the design and development of datasets, they underline negative social consequences and poor system performance. Next, they address data filtering and augmenting strategies and modeling tools for reducing the influence of bias in datasets. Finally, they cover studies of data practices, cultures, and disciplinary norms and their implications for the field's legal, ethical, and functional difficulties. Based on their findings, they urge that qualitative and quantitative methodology be used during the development and utilization of the dataset for better documentation [17].

The correctness of the model is a typical criterion for ML systems in the perspective of most ML practitioners, whereas the data collection is presumed to be of sufficient quality. Even though machine learning models are increasingly being utilized in high-stakes domains where data quality is crucial, data quality remains the most undervalued aspect of machine learning. Sambasivan et al. studied the features and consequences of negative data cascades that occur in industrial sectors due to popular machine learning approaches [9]. While their research focuses on organizational challenges, they mention a lack of proper data education as a contributing cause. We believe it is critical to delve deeper into this topic and determine how well ML students understand dataset properties.

3 Methodology

We conducted semi-structured interviews with 13 students who have had previous academic experience in Machine Learning. Interviews focused on how each participant had learned ML, what they learned about data practices and their view on dataset modification and fairness. After designing an initial interview guideline, we decided to conduct a pilot-pilot interview with two of the participants to refine our guideline. Given the poor quality of the two initial interviews, we have excluded them from our analysis. The primary interviews were conducted with the rest of the participants. All interviews were conducted online in Persian (participants preferred language) and lasted about 20-35 minutes.

Interview questions were asked of the participants in three parts. In the first part, they were asked about their field of work and experiences in machine learning and data science. Questions such as "What were the ML projects you have worked on so far?", "What were the resources/courses you took to learn about ML/DS?". In the second part, we focused on their knowledge of the data work. We probed the participants on their experiences in gathering, curating, or processing datasets. We then continued the inquiry on the challenges they faced and how they resolved them. Lastly, we asked how much time they spent working on the model vs. on the data during their projects. In the last part, we asked the participants if they had heard about algorithmic fairness. If not, we would give a brief definition with an example and then move on to questions regarding ML fairness issues. We asked for participants' opinions on data modification to achieve fairness, noting that it was not necessary to provide any technical reasoning. Finally, we concluded by asking if they had any further suggestions on the issue.

Participant recruitment Given the limitation of the pilot study, our sample is from close friends. All participants have successfully completed a bachelor of science degree (2 Electrical Engineering, 9 Computer Engineering) and had at least one machine learning course during their undergraduate studies. They are currently persuading a graduate degree (6), employed full-time (2), or planning to continue their studies (3). A summary of participants' demographic is given in table 1.

Analysis Following [18], we conducted an open coding and refined the result through several iterations. We identified a total of 8 themes of participants' experiences in machine learning and datasets and their views on data modification. We present these themes in table 2.

Research ethics . During interviews, participants were informed of the purpose of the study, the focus of the study on machine learning and their experience. At the beginning of each interview, the moderator additionally obtained verbal informed consent. We stored all data in a private Google Drive folder and removed all personally identifiable information. All audio files and notes taken from the interviews will be deleted after April 30 and only this report will remain.

Type	**Count**
Bachelor degree	Computer Engineering (9), Electrical Engineering (2)
Gender	Man (6), Woman (5)
Current field of work/study	Machine Learning (5), Software Engineering (1), Computer Engineering (1), Data Science (1), Visual Computing (1), Human Technology Interaction (1), Control engineering (1)

Table 1: Summary of participant demographics.

4 Findings

In this section, we present our findings regarding our research questions. We identify a total of eight themes from our data. One regarding participants' understanding of data, two on how they learned ML and five for their stance on data modification for fairness. List of themes can be seen in table 2.

4.1 What do students know about the importance of data

Acquired understanding of data

Whether self studying machine learning or perusing a degree, most of our participants (8 of 11) showed an understanding of how vital data is to their tasks. When accounting of their projects, they began by describing their data, its properties and shortcoming before moving on to their model. Furthermore, five of the participants explicitly mentioned how they are currently focused

Research Question	Theme	Count
Understanding Data	Acquired understanding of data	8 of 11
Learning ML	Unhelpful academia	11 of 11
	Learning the model, working the data	6 of 11
Views on data modification for a fairer system	Accepting	8 of 11
	Rejecting synthetic data	3 of 11
	Reservations for removing data	3 of 11
	Concern for critical domains	2 of 11
	Against use of ML	3 of 11

Table 2: List of extracted themes

on data for their current project. "[My focus is] on the data. It has lots of empty fields. Like, it's 130 thousand observations, so there's gonna be some issues" (P6).

While most participants clearly understood the importance of the data, only one student had explicitly studied working with data: "As far as I've read, much time has to be spent on data cleaning..." (P1). whereas other students mostly addressed working with data based on their experience in past or current projects, emphasising the challenges and difficulties they faced. "I mainly focused on data, because I think working on network models is the most straightforward part of this [work]... this is from experience; I can't really explain why..." (P3).

4.2 How did the students learn about machine learning

Unhelpful academia

All of our participants mentioned seeking external courses and materials for study machine learning. One of the most commonly noted resources was the Deep Learning course by Andrew Ng. Furthermore, two of the participants outright discredited their studies at university, while three others simply noted their lack of readiness for the industry or real-world projects. Students commonly criticized courses' lack of practicality, usefulness and comprehensibility. "We had a machine learning course, but it was rudimentary ... and [it] wasn't very practical... So I enrolled in a deep learning course on Coursera. It was an excellent experience" (P2).

Learning the model, working the data

Another interesting trend, closely coupled with the last theme, is how students mainly work with data, while they learned machine learning from a model perspective. Half the participants (6 of 11) note they began studying machine learning from studying models, while their current work focuses more on data. "In machine learning, there is a big gap between academia and the industry. In the industry, models' hyperparameters may not be so important. ...we need to focus on both data and model, but data is more important" (P8).

4.3 What do students think of data modification for algorithmic fairness

We note that of our participants, only two (P3 and P8) were familiar with the concept of algorithmic fairness. For other students we explained the fairness problem with an imaginary loan approval system being biased based on gender. After ensuring the participants understood the concept, we moved on to the last part of the interview. Surprisingly we received a mixed response to data modification, with most participants accepting data modification to a certain degree.

Overall, a majority of the participants (8 of 11) were accepting of modifying data to achieve a fairer system: "I think the only way is to kinda force fairness into your system. Like with data, or models..." (P11) Nonetheless, the same participants expressed concerns about certain aspects of these methods; in particular they were concerned about using synthetic data (4 of 11), removing real data (3 of 11), and applicability in critical domains such as medicine (3 of 11). Lastly, several of the students (3 of 11) were firmly against the use of machine learning in human affected tasks. In the following sections, we expand on the various responses and their reasoning.

Rejecting Synthetic Data

Three participants expressed their concern about the use of synthetic data for training models. Overall, a key concern for synthetic data is its validity. P4, questioned the methods for creating synthetic fair data: "But still, there are thousands of ways [to create synthetic data], ... I don't think it'll be a good method" (P4). P5, who has been working with generative networks in a professional capacity, strongly emphasized the difference between real and artificial data: "We must bear in mind that artificial data is never like real data. Even the best model in the strongest laboratory is not able to produce artificial data close to reality" (P5).

Reservations for removing data

P4 and P5, who raised concerns about synthetic data, had similar views on data modification, particularly data deletion. Whereas P5 was unsure of the effectiveness of this method, P4 scrutinizes how we chose what data is to be deleted: " See, having someone to remove data just induces a bias from the other side. Then you have to handle that too" (P4). Additionally, P2 questioned the soundness of models being trained on incomplete datasets: "...[the model] won't correctly process the data that it's not trained on" (P2).

Concern For Critical Domain

Of our participants, only two (P2 and P6) raised concerns about using modifying data in critical domains, both emphasizing the extreme costs of irreversible mistakes.

> In my opinion, a crucial factor is the situation in which the model will be used. Like, in the safety field, in self-driving cars, or in medicine, where we are dealing with a person's life, I don't accept the risk of relying too much on fake data. I'd rather use real data, no matter how unfair. That's my preference. Because in my opinion, it's more realistic (P6).

Against use of ML

Interestingly, three participants expressed their opposition for the use of ML in making decisions for the future of people. Responses indicated a general distrust of a computer system, especially in the case of P7, whose application was rejected before being reviewed by a human. “Really, I think it’s not fair at all. When they use ML, they also need a human supervisor to check the person’s file; because these things have errors. I personally had this problem; my application got rejected because of an AI system” (P7). P3, with seven years of experience in ML, was the most determined student against the use of ML in human oriented tasks. “...to solve this [fairness] problem, we should not involve machine learning in human issues at all. Machine learning should be used only as a tool to solve problems, not to make decisions about people’s lives. Machine learning is being used in the wrong fields” (P3).

5 Discussion

As we expected, our pilot study suggests that most students studying machine learning have some understanding of importance of data, however this understanding comes from experience rather than academia. As several participant noted, they were "not ready" for the industry and faced many challenges that they were not taught about. Our participants reported learning (either self-taught or in university) machine learning by focusing on the models, its mathematical details and method of operation. However during their projects they faced issues on data management, augmentation and cleaning. We believe this is a consequence of how machine learning is commonly taught. Majority of machine learning courses focus on *how* the models work, their mathematical details, optimizations and architectures. In contrast, little time is devoted to *what* models need to work. This shortcoming of academia results is countless setbacks for most students transitioning to industry after their undergraduate studies.

Surprisingly, in contrast to our expectations we met with considerable support for data modification to achieve a fairer system. Most participants explicitly emphasized identifying the bias factor and eliminating it in the database. However, one participant (P5) pointed out that modifying the dataset may not cover all bias cases and mentioned the power of mathematical models as a better alternative for bias detection.

Despite the overall support, participants had a mixed response on how the data modification should be done. While some students were wary of artificial data performing as well as read data, other students questioned the validity of removing real data from a dataset. Lastly, several students stated their opposition for any data manipulation in critical domains. Even tough some participants such as P6 were in favour of data manipulation at first, they views changed completely when the discussion led to critical domains. We believe that even though students are in favour of a fairer system, they intuitively distrust altered datasets, which they expressed in domains they closely associate. However, determining whether this distrust arises from discrediting modified data as a principle or a lack of familiarity with data modification practices requires more extensive research.

An unexpected discovery for us was a minority of our participants (P3, P7, P10) were opposed to giving machine too much control over human lives. Perhaps one of the reasons for this view was the lack of trust in existing methods for fairness in machine learning. One of the participants (P5) pointed out that manipulating the dataset itself could create new biases. On the other hand, some participants (P3, P10, P11) considered the lack of a comprehensive metric to measure fairness as the main factor of intractability. Due to these reasons, they argue that it is best that we do not use ML models in critical scenarios at all.

6 Limitations

Our work has several limitations. First, given the nature of the pilot study, our participants were extremely limited and lacked diversity. Second, most of our participants were initially unfamiliar with the concept of algorithmic fairness, and they learned about fairness and provided their opinions in one interview setting. Finally, Given the lack of research in this area, we relied on our research on two pilot-pilot interviews to design our interview guidelines. We believe an iterative study with a broader pool of participants would be able to gather more comprehensive responses.

7 Conclusion And Future work

Even though data is one of the most important parts of machine learning, it is rarely discussed or taught in conjunction with it. In this research, we provide the results of a pilot study of 11 university students who studied machine learning and how they learned about data, and their opinions on data manipulation. According to our findings, our participants learned machine learning primarily through popular online courses and were unsatisfied with their local education. Also, we found out that students learned to work with data only after encountering challenges at work, not due to coursework during their education.

Furthermore, while the majority of participants supported data alteration as a means of attaining a more fair system, many questioned the methods adopted. We believe this stems from a natural apprehension about data modification. This was particularly apparent when participants discussed potential hazards in critical fields such as medicine and safety.

We note that our pilot study is limited by the diversity of our participants and our understanding of the investigated subject. However, this unexplored avenue of study is crucial in training a more knowledgeable generation of machine learning engineers for both academia and industry. We hope our pilot study encourages wider and more comprehensive studies to be conducted.

References

[1] "10 Best Machine Learning Courses to Take in 2022," Jan. 2022. https://www.freecodecamp.org/news/best-machine-learning-courses/.

[2] "CS229: Machine Learning." http://cs229.stanford.edu/syllabus.html.

[3] I. D. Raji, M. K. Scheuerman, and R. Amironesei, "You can't sit with us: Exclusionary pedagogy in ai ethics education," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 515–525, Association for Computing Machinery, 2021.

[4] D. Dua and C. Graff, "UCI machine learning repository," 2017. http://archive.ics.uci.edu/ml.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," pp. 770–778, 2016.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper With Convolutions," pp. 1–9, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

[8] K. Wagstaff, "Machine Learning that Matters," *arXiv:1206.4656 [cs, stat]*, June 2012. arXiv: 1206.4656.

[9] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

[10] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.. California Law Review*, vol. 104, no. IR, p. 671.

[11] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016.

[13] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[14] M. S. A. Lee and J. Singh, "The landscape and gaps in open source fairness toolkits," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[15] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[16] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proceedings of the 2018 World Wide Web Conference*, pp. 903–912, 2018.

[17] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *CoRR*, vol. abs/2012.05345, 2020.

[18] V. Braun and V. Clarke, "Thematic analysis.," in *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* (H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher, eds.), pp. 57–71, Washington: American Psychological Association, 2012.