UC SANTA BARBARA
THE Current

June 18, 2024
Andrew Masuda

Dirty data: an opportunity for cleaning up bias in AI

Haewon Jeong, an assistant professor in UC Santa Barbara's Electrical and Computer Engineering (ECE) Department, experienced a pivotal moment in her academic career when she was a postdoctoral fellow at Harvard University. She was investigating how machine learning (ML) models can discriminate against students in education-related applications. Discrimination, or bias, occurs when a model used to train algorithms makes incorrect predictions that systematically disadvantage a group of people. Bias in ML models can lead to inaccurate or unfair predictions, which can have serious consequences in fields such as healthcare, finance and criminal justice. For example, an unfair model that relies on historical data reflecting systematic social and economic inequities could result in mortgage applications being rejected more often for women than for men, or skin cancer being detected more for white patients than for Black patients, who might be denied treatment.

"I was working with education-related datasets collected by my collaborator, and I realized there was a lot of missing data," Jeong recalled.

Concerned about adding to the bias in the data, she searched for research papers on how to avoid adding more bias when substituting missing entries with new values, a process called *imputation*. That was when she made a shocking discovery.

"No one had studied the fairness aspect of imputation before, which was surprising because missing data is such a prevalent problem in the real world," she said.

“Nearly all of the research at the time centered around developing better training algorithms to eliminate bias, but not many people thought about addressing the bias that happened during data collection.”

That realization provided the framework for Jeong’s novel approach to identifying and mitigating the ever-evolving ethical challenges presented by AI-powered systems, launching her study of how various steps in the data-preparation pipeline can introduce bias or fairness.

“People in my field say, ‘Bad data in, bad algorithm out. Biased data in, biased algorithm out,’” said Jeong, “but I have proposed that if we focus on cleaning the bad data, we could reduce the bias from the start.”

As a testament to the potential impact of her proposed research, the National Science Foundation (NSF) has granted Jeong an Early CAREER Award, the federal agency’s most highly regarded honor for junior faculty. She said that the five-year, $558,000 grant provided a significant boost to her research group and to her, personally.

“I am honored and thrilled,” said Jeong. “This award has made me more confident that the direction of my research is meaningful and supported by the NSF.”

Her project, “From Dirty Data to Fair Prediction: Data Preparation Framework for End-to-End Equitable Machine Learning,” targets the data-preparation pipeline as a strategic opportunity for eliminating unwanted bias and bolstering desirable ethical objectives. Typically, Jeong said, data is “dirty” — missing values and entries, and including varying formats that require standardization. There are many steps required to prepare, or clean, the data, and underlying disparities can encode significant inaccuracies along the way. To mitigate the bias early in the process, Jeong has proposed a three-step process to insert fairness in, when addressing missing values, encoding data and balancing data.

“Right now, AI algorithms learn from examples, and algorithmic interventions can only do so much with the given data,” said Jeong, who earned her Ph.D. in ECE from Carnegie Mellon University. “I propose that supplying better examples and data to the algorithm will result in more fair and ethical learning.”

Datasets are often missing values. For example, when conducting a survey, some questions are not answered completely or are left empty. Before feeding any dataset into an ML algorithm, researchers have two main options for handling missing data: they can exclude the entries that contain missing data, or they can fill in the missing data with an estimate based on the other available information. Jeong's prior work showed that both methods significantly increased bias. She was the first researcher to publish a paper calling attention to that problem.

"In that paper, we proposed a simple algorithm to deal with bias created through imputation, but it was not very efficient," she said. "In this project, I want to dive deeper into the problem to investigate if there are more efficient ways to perform data imputation and consider fairness at the same time."

The second thread that she will address is data *encoding,* which is the process of changing raw data into a numerical format than an algorithm can read and interpret. Returning to the survey example, some answers may range from zero to five, while others include text fields. Data encoding involves converting the words into numbers. Encoding also enables computers to process and transmit information that is not numerically based, such as text, audio, and video.

"The process of encoding text is already known to cause gender bias and perpetuate social stereotypes, but it's unclear how these biases flow through the subsequent steps," explained Jeong, who will rely on her training in information theory to address data encoding. "By looking at it from an information-theory perspective, we hope to develop a fairer algorithm to preserve useful information and suppress information related to bias."

The third step involves increasing fairness when *balancing* data, which is the process of ensuring that a ML dataset represents the real-world population from which it is drawn. Having an uneven number of observations among different groups significantly impacts an ML models predictive performance and fairness. This particular thrust is driven by an experiment with education data that Jeong performed as a postdoctoral fellow. In the project, she grouped students into Black/Hispanic/Native American (BHN) and White/Asian (WA). The data was imbalanced, and a majority of the students were in the WA group. Seeking the best way to balance the data and mitigate bias, Jeong varied the proportion of the groups in the training set while keeping the size of the set constant. By varying the

percentage of the BHN student data in the training set to range from zero to one hundred percent, she made a surprising discovery.

“One might intuitively think that the mix of fifty-fifty or one that aligns with national demographics would yield the most equitable model, but it did not,” she explained. “We found that fairness increased most when we included more data points in the set from the majority group and fewer from the minority group.”

As part of her NSF project, Jeong wants to explore what causes the counterintuitive results and establish guidelines for data scientists on the optimal demographic mixture to use. She believes that the amount of noise in the data plays a role in how the data should be balanced. Noise here means inaccuracies in the data, such as people not answering surveys truthfully, giving incorrect answers, or problems created by a language barrier. Jeong hypothesizes that the fairest and least-biased mixture includes more data from the group having the lowest noise level.

Through her novel three-pronged approach to attacking real-world dataset issues, Jeong hopes to create guidelines and best practices in data preparation for equitable and fair ML. Given the skyrocketing use of ML and AI in nearly every sector of society, she believes that her work has significant real-world implications.

“Data and computer scientists want AI to embody and promote essential societal values, like fairness and diversity, not stereotypes,” said Jeong. “Removing unwanted bias and inserting ethical objectives into the data-preparation pipeline could make that possible.”

The end goal of Jeong’s project is to develop a software library that any data scientist or AI developer can use for fairness-aware data preparation. The library would include her group’s fair-imputation methods, bias-flow measurement toolkit and algorithms.

Jeong also proposed an educational agenda that prioritizes the attraction and retention of talented female students to the study of AI. Research shows that only 12% of AI researchers and a mere 6% of professional software developments in the AI field are women. Jeong plans to design and host the “Girls’ AI Bootcamp,” which will be specifically tailored to engaging female high school students and introducing them to the exciting possibilities within CS and AI.

“I have experienced firsthand the challenges of being in the minority in this field, and I am personally committed to closing the gender gap,” said Jeong. “I not only

want to pique the interest among female high school students, but also instill self-confidence in them that they can be leading innovators in the fields of AI and CS.”

Tags
Artificial Intelligence

Media Contact

Sonia Fernandez

Senior Science Writer

(805) 893-4765

sonia.fernandez@ucsb.edu

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.