Measuring social biases in human annotators using counterfactual queries in Crowdsourcing

Algorithmic bias has been termed as the imminent AI danger faced by our society. Recent studies have shown that machine learning(ML) algorithms are capable of exhibiting social biases like gender, race, etc. A major source of bias in the ML pipeline arises from the training dataset. Crowdsourcing is a popular way to gather labeled data for different ML tasks. As crowdsourcing tasks might involve a subjective component, it's important to gauge implicit social biases of human annotators and prevent them from spreading into the curated dataset.

We propose a novel way to measure social biases in crowdworkers using counterfactuals queries. Here, we are considering a supervised learning scenario with numeric or categorical input features. A counterfactual to a given query is the most similar query in an alternate world where its sensitive attributes like race, gender, etc. is flipped. Counterfactual queries can be generated using causal inference by measuring the impact of flipping the sensitive attribute on other input features. During the training phase of the user study, we can ask human annotators to label a set of counterfactual queries and measure the deviation in the responses. Zero deviation characterizes perfect unbiased behavior and higher values symbolize more biased behaviour. If the deviation is beyond a specific threshold, we can consider such annotators to be unfit for the study and terminate the study for those labelers. This methodology doesn't need unbiased labels and biased labelers are disqualified in the training phase itself. Hence, this can serve as a cost-effective way to tackle social biases in crowdsourcing.