How can we develop more effective ways to mitigate bias in language models?

Abstract

Large Language models are a new technology which are predicted to have a significant impact on society. They are in use now and starting to transform society in a variety of ways. However, these systems still face many limitations, for example they learn historical biases from their training data and output discriminatory statements as a result. There is a growing body of research dedicated to bias in machine learning, but we are yet to develop effective ways to mitigate the harms that these systems pose, especially in the domain of natural language. In this paper I explore why this is by asking how language models threaten to harm certain demographics and why previous work on this topic has failed to result in conclusive answers. I find that many previous methods are disconnected from the way these systems are used in the real world. In addition, the concepts of bias and fairness are relatively new in machine learning which leads to them being treated very differently by different researchers. I argue that more research needs to be conducted on the effects these systems have when applied in social contexts, and doing this will require interdisciplinary collaboration with disciplines that have more experience examining bias in society. I use my findings to propose recommendations for future research to consider and advise ways to facilitate the necessary research. I also suggest how a framework from sociolinguistics can be adapted into a method to evaluate bias in language models.