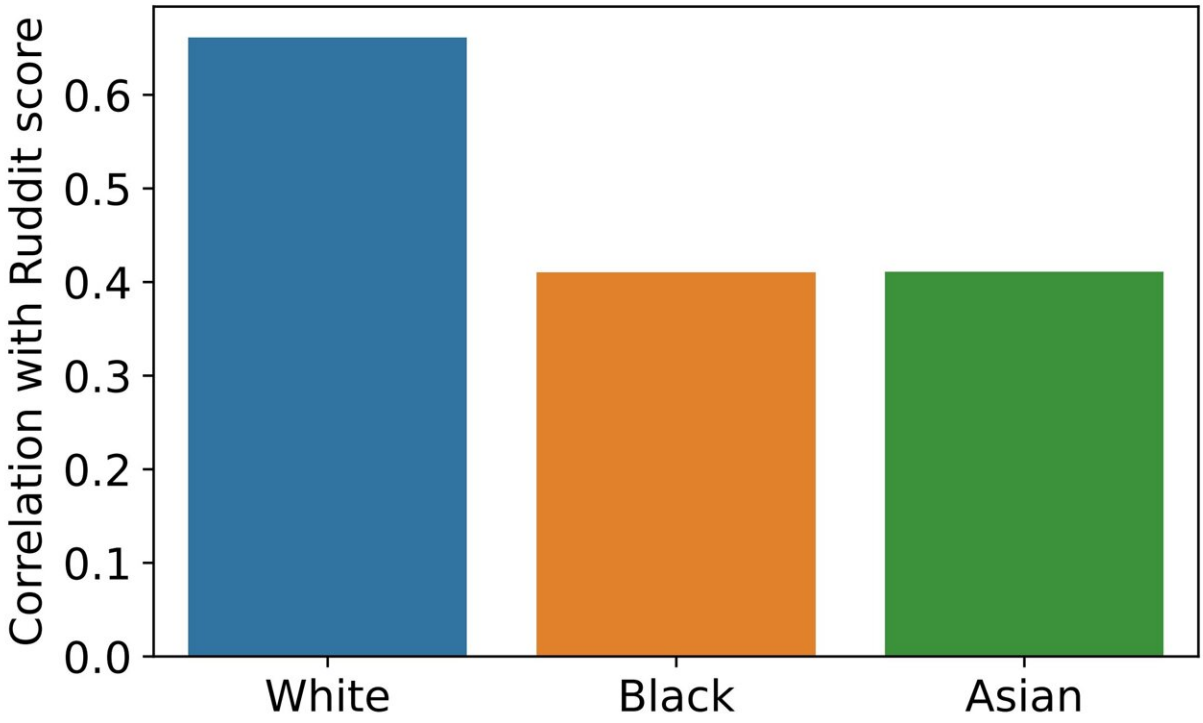
Building reliable AI models requires understanding the people behind the datasets

August 8 2023, by Jared Wadley



Correlation with the original Ruddit offensiveness score by race. Annotations by White participants have the highest correlation with the Ruddit score, while annotations by Asian and Black participants are significantly less correlated. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2306.06826

Social media companies are increasingly using complex algorithms and artificial intelligence to detect offensive behavior online.

These algorithms and AI systems all rely on data to learn what is offensive. But who's behind the data, and how do their backgrounds influence their decisions?

In a new study, University of Michigan School of Information assistant professor David Jurgens and doctoral candidate Jiaxin Pei found that the backgrounds of data annotators—the people labeling texts, videos and online media—matter a lot.

"Annotators are not fungible," Jurgens said. "Their demographics, life experiences and backgrounds all contribute to how they label data. Our study suggests that understanding the background of annotators and collecting labels from a demographically balanced pool of crowdworkers is important to reduce the bias of datasets."

Through an analysis of 6,000 Reddit comments, the study shows annotator beliefs and decisions around politeness and offensiveness impact the learning models used to flag the online content we see each day. What is rated as polite by one part of the population can be rated much less polite by another.

"AI systems all use this kind of data and our study helps highlight the importance of knowing who is labeling the data," Pei said. "When people from only one part of the population label the data, the resulting AI system may not represent the average viewpoint."

Through their research, Jurgens and Pei set out to better understand the differences between annotator identities and how their experiences impact their decisions. Previous studies have only looked at one aspect of identity, like gender. Their hope is to help AI models better model the beliefs and opinions of all people.

The results demonstrate:

- While some existing studies suggest that men and women may have different ratings of toxic language, their research found no statistically significant difference between men and women. However, participants with nonbinary gender identities tended to rate messages as less offensive than those identifying as men and women.
- People older than 60 tend to perceive higher offensiveness scores than middle-aged participants.
- The study found significant racial differences in offensiveness ratings. Black participants tended to rate the same comments with significantly more offensiveness than all the other racial groups. In this sense, classifiers trained on data annotated by white people may systematically underestimate the offensiveness of a comment for Black and Asian people.
- No significant differences were found with respect to annotator education.

Using these results, Jurgens and Pei created POPQUORN, the Potato-Prolific dataset for Question Answering, Offensiveness, text Rewriting and politeness rating with demographic Nuance. The dataset offers social media and AI companies an opportunity to explore a model that accounts for intersectional perspectives and beliefs.

"Systems like ChatGPT are increasingly used by people for everyday tasks," Jurgens said. "But whose values are we instilling in the trained model? If we keep taking a representative sample without accounting for differences, we continue marginalizing certain groups of people."

Pei said that POPQUORN helps ensure everyone has equitable systems that match their beliefs and backgrounds.

The study is published on the *arXiv* preprint server.

More information: Jiaxin Pei et al, When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset, *arXiv* (2023). DOI: 10.48550/arxiv.2306.06826

Provided by University of Michigan