

Investigating gender bias using artificial intelligence classification models on RAVDESS dataset

Sefara, Tshephisho J; Khosa, Marshal V; Kisten, Melvin

Abstract

Artificial intelligence (AI) classification models are increasingly being deployed across a wide array of sectors, becoming fundamental tools in decision-making processes that impact individuals and society. These models are utilised in critical applications such as healthcare diagnostics, financial risk assessment, criminal justice systems, and educational admissions, demonstrating their widespread influence. However, a significant challenge arises from the susceptibility of these models to biases, which can lead to outcomes that are unfair, discriminatory, and ultimately harmful to individuals and specific demographic groups. Artificial intelligence bias refers to systematic errors that occur within decision-making processes, ultimately leading to outcomes that are unfair or inequitable. This can manifest as skewed results stemming from human biases that have influenced the original data used to train the AI model, resulting in distorted outputs with potentially harmful results. In this paper, we mitigate the gender bias that occurred during data selection for a classification model. This research experiment was conducted on RAVDESS emotion recognition dataset. The experiments showed improvement in model accuracy by 6% after bias mitigation.