Joanna Bryson – Abstract

Discovering Individual and Collective Bias via Automated Language Model Analysis

We study bias in individuals and groups through automated text analysis by incorporating machine learning and natural language processing techniques. Such an automated method makes it possible to analyze bias at the large scale for different cultures, time periods, and languages. Our approach is the first step towards a principled method for quantifying bias and its effect in digital communications.

Machine learning models are criticized for incorporating bias from their training data. Eliminating bias in models have been limited to controlling algorithm's parameters to avoid overfitting, which doesn't prevent bias from revealing itself at the contextual level. Based on this knowledge, we train language models on writings' of subjects of interest to generate a semantic space of word embeddings. Embeddings are numeric vectors whose dimensions correspond to combinations of contexts. We focus on concepts that have been used in bias studies, such as gender, professions, racism, and religion. We then calculate distances between concepts and potentially biased terms to observe implicit bias through spatial associations.

We investigate bias in famous individuals, the Enron Corporation, Wikipedia, Twitter, and GoogleNews. We discuss the implications of bias present in widely used language models in digital communications for text generation, automated speech, and translations.