Hatred is in the Eye of the Annotator: Hate Speech Classifiers Learn Human-Like Social Stereotypes

Aida Mostafazadeh Davani
University of Southern California, Los Angeles, California, United States

Mohammad Atari
University of Southern California, Los Angeles, California, United States

Brendan Kennedy
University of Southern California, Los Angeles, California, United States

Shreya Havaldar
University of Southern California, Los Angeles, California, United States

Morteza Dehghani
University of Southern California, Los Angeles, California, United States

Abstract

Social stereotypes impact individuals' judgement about different social groups. One area where such stereotyping has a critical impact is in hate speech detection, in which human annotations of text are used to train machine learning models. Such models are likely to be biased in the same ways that humans are biased in their judgments of social groups. In this research, we investigate the effect of stereotypes of social groups on the performance of expert annotators in a large corpus of annotated hate speech. We also examine the effect of these stereotypes on unintended bias of hate speech classifiers. To this end, we show how language-encoded stereotypes, associated with social groups, lead to disagreements in identifying hate speech. Lastly, we analyze how inconsistencies in annotations propagate to a supervised classifier when human-generated labels are used to train a hate speech detection model.