Toward Socially Aware Computing and Artificial Intelligence

Submission ID 3000315

Submission Type Poster

Topic Cognitive Science

Status Submitted

Submitter Pierre Karashchuk

Affiliation UC Berkeley

SUBMISSION DETAILS

Presentation Type Either Poster or Oral Presentation

Presentation Abstract Summary There is increasing concern that the proliferation of AI-driven automation may perpetuate and even amplify preexisting biases and social inequities facing certain groups of individuals. However relatively little attention has been paid to computational tools that can recognize and quantify social biases at the scale necessary to address these societal challenges. Here, we used models from social psychology relating stereotypes of specific groups to a few core dimensions in order to investigate the extent to which word embeddings reveal stereotypes about different social groups and their real-world consequences. We show that word embeddings trained on Google News can be used to make accurate predictions of stereotypes about social groups identified by occupation, geography, and race. We apply these models to quantify biases in historical data. Furthermore, we show that word embeddings can predict disparate treatment of different groups in lab and field settings. We conclude that word embeddings do not only capture "clusters" of social groups, but a general continuous representation of social biases. This could be used to help quantify biases for social research or to help AI systems be aware of social biases when making decisions.

Paper Upload (PDF) ccn_word2vec_stereotypes.pdf

Co-author Information

* Presenting Author

First Name	Last Name	Affiliation	E-mail
Pierre *	Karashchuk *	UC Berkeley	pierre@berkeley.edu
Ming	Hsu	UC Berkeley	mhsu@haas.berkeley.edu
Adrianna C.	Jenkins	UC Berkeley	adrianna.jenkins@gmail.com

Keywords

Keywords
social decision-making
social perception
word embeddings
natural language processing
bias