UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Seminar 2

Wednesday December 7, 2022 at 9am via Zoom

Hillary Dawkins

Detection and Mitigation of Gender Bias in Large Pre-trained Language Models

Advisor: Dr. Judi McCuaig
Co-Advisor: Dr. Dan Gillis
Advisory: Dr. Graham Taylor (SoE)
Advisory: Dr. Stefan Kremer

Abstract:

Mitigation of gender bias in NLP has a long history tied to debiasing static word embeddings. More recently, attention has shifted to debiasing pre-trained language models. We study to what extent the simplest projective debiasing methods, developed for word embeddings, can help when applied to BERT's internal representations. Projective methods are fast to implement, use a small number of saved parameters, and make no updates to the existing model parameters. We evaluate the efficacy of the methods in reducing both intrinsic bias, as measured by BERT's next sentence prediction task, and in mitigating observed bias in a downstream setting when fine-tuned.