UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Qualifying Exam

Thursday June 2, 2022 at 9:30am via Zoom

Hillary Dawkins

Detection and Mitigation of Gender Bias in Natural Language Processing

Chair: Dr. Joe Sawada
Advisor: Dr. Dan Gillis
Advisory: Dr. Graham Taylor [SoE]
Non-Advisory: Dr. Fei Song
Non-Advisory: Dr. Fattane Zarrinkalam [SoE]

Abstract:

The goal of this research proposal is to mitigate gender-biased outcomes produced by NLP systems by debiasing pretrained resources (both static word embeddings and language models) via simple post-processing methods. We focus on post-processing methods because they require minimal additional computation, and they are easy to concatenate with existing methods. Throughout, the performance of a debiasing method is quantified by its ability to eliminate or reduce unequal outcomes across binary genders (e.g. as differences in predictions across gender), without affecting task accuracy. As we will come to appreciate, mitigating bias in pretrained resources often requires an understanding of how intrinsic bias (some innate property of the pretrained resource) correlates with observable bias in downstream applications. Therefore, supporting contributions to this research are to propose and investigate intrinsic bias measures.