Carlos Aguirre

scholar | pocaguirre | pocaguirre | pocaguirre.com | caguirre@cs.jhu.edu

Research Interests

My research focuses on measuring social biases and ensuring fairness of language models. In the past, I have primarily focused on application areas such as healthcare and have answered questions regarding ensuring fairness under missing demographic attributes and measuring fairness in decision making of LLMs. Other interests include: human-AI interactions, AI for social good, fairness and trustworthy-AI, and other applications of NLP, e.g. language generation and dialog systems.

Keywords: `fairness`, `NLP`, `healthcare`, `social media`, `Large Language Models`, `mental-health`, `Human-AI interaction`

Education

PhD Student Fall 2019 - present

Johns Hopkins University — CLSP affiliated — Advised by Mark Dredze

M.S.E. Computer Science Fall 2019 - Fall 2021

Johns Hopkins University — CLSP affiliated — Advised by Mark Dredze

B.S. Computer Science and minor in Mathematics Fall 2016 - Spring 2019

Kansas State University — GPA: 3.9 — Advised by William Hsu

Associates in Arts Fall 2014 - Spring 2016

Metropolitan Community College – Penn Valley — GPA: 4.0

Research Experience

Fairness & Biases in NLP 2022-present

Some of the current challenges for evaluation and ensuring fairness on LLMs are the lack of data availability (most datasets do not have demographic information available) and inability to finetune LLMs (new models are often hidden behind APIs and are immutable.) How we evaluate and ensure fairness under these conditions is still an open question.

Aguirre, Carlos and Mark Dredze. "Generalizing Fairness using Multi-Task Learning without Demographic Information". In: *arXiv preprint arXiv:2305.12671* (2023).

Aguirre, Carlos, Kuleen Sasse, et al. "Selecting Shots for Demographic Fairness in Few-Shot Learning with Large Language Models". In: *arXiv preprint arXiv:2311.08472* (2023).

Fairness & Biases of Depression Models using Social Media 2020-21

Trained language models using Twitter data to predict depression, measured the fairness of the models we trained along gender and racial/ethnic groups, and analyzed the models errors using topic models. Further, investigated the biases of depression models trained on Reddit data across gender groups.

Aguirre, Carlos, Keith Harrigian, and Mark Dredze. "Gender and Racial Fairness in Depression Research using Social Media". In: *Proceedings of the 16th EACL: Main Volume*. 2021, pp. 2932–2949.

Aguirre, Carlos and Mark Dredze. "Qualitative Analysis of Depression Models by Demographics". In: *Proceedings of the 7th CLPsych Workshop*. 2021, pp. 169–180.

Sherman, Eli et al. "Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models". In: *Proceedings of the 7th CLPsych Workshop*. 2021, pp. 217–223.

Relation of Stressors and Depression across Demographics 2021-22

Trained linear and finetuned neural language models using open-ended text responses about stressors to predict depressive symptoms, and investigated the differences in language of responses and performance of models across gender and racial/ethnic groups.

Aguirre, Carlos, Mark Dredze, and Philip Resnik. "Using Open-Ended Stressor Responses to Predict Depressive Symptoms across Demographics". In: *arXiv preprint arXiv:2211.07932* (2022).

Collecting Image Captions at Specific Levels of Detail 2021-22

Collected image captions at varied levels of detail by constraining the time that annotators were allowed to observe the image. Created a custom annotation tool and conducted experiments on MTurk. Designed manual evaluation protocol to assess fluency, correctness and amount of detail.

Aguirre, Carlos, Amama Mahmood, and Chien-Ming Huang. "Crowdsourcing Thumbnail Captions via Time-Constrained Methods". In: *27th IUI Conference*. 2022, pp. 36–48.

Aguirre, Carlos, Shiye Cao, et al. "Crowdsourcing Thumbnail Captions: Data Collection and Validation". In: *ACM Trans. Interact. Intell. Syst.* (Mar. 2023). ISSN: 2160-6455. DOI: 10.1145/3589346.

Social Media for Mental Health 2019-20

Performed a literature review of the state of research predicting various mental health disorders across social media platforms. Released a collection of publicly available datasets of mental health disorders using social media. Analyzed the effect of temporal and domain shifts across social media platforms on mental health models.

Harrigian, Keith, Carlos Aguirre, and Mark Dredze. "On the State of Social Media Data for Mental Health Research". In: *Proceedings of the 7th CLPsych Workshop*. Online: Association for Computational Linguistics, June 2021, pp. 15–24.

– "Do Models of Mental Health Based on Social Media Data Generalize?" In: *Proceedings of the 2020 EMNLP Conference: Findings*. 2020, pp. 3774–3788.

Campus and Community Involvement

Instructor — HEART: AI Ethics in Healthcare Applications 2022

Created and conducted a seminar class discussing ethical considerations of the use of machine learning systems in healthcare. Duties involved designing the class structure and creating all class materials.

TA — Introduction to Machine Learning 2020

TA for Prof. Mark Dredze. Duties included designing and writing class projects, homework and exams, as well as conducting and creating the content for recitation session every week.

Research Mentor 2021-present

Co-advised undergraduate and masters students on a variety of projects.

CS Social Committee 2022-present

Plan and assist in community and social events involving graduate students in the Computer Science department.

CLSP Graduate Admissions Committee 2021-present

Review graduate applications with faculty, helps make initial determination of candidates.

CLSP Diversity in Admissions Committee 2019-20

Work to increase the diversity of applicants to the university's PhD program.

Lecturer — Introduction to Programming with Python 2018

Taught an introductory course for programming using python for non-engineering students (and faculty) at Kansas State University. Duties included updating the course syllabus, lecturing, designing and grading homework.

Last updated: November 16, 2023