Carlos Aguirre

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

Overview

I create fairer and socially aware NLP & ML models. My current research focuses on measuring and then improving the fairness of such models, and adapting them across various domains. In addition to analyzing LLMs for fairness, I have developed a variety of methods to improve fairness, including a custom loss for use in finetuning. I have a strong background in developing models in data-limited settings, including mental healthcare. I have also pre-trained encoder-only transformers for Twitter, Bernice.

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

Publications

1. Aguirre, C. & Dredze, M. "Transferring Fairness using Multi-Task Learning with Limited Demographic Information". arXiv: `2305.12671 [cs.LG]` (2024).
2. Aguirre, C. *et al.* "Crowdsourcing Thumbnail Captions: Data Collection and Validation". *ACM Trans. Interact. Intell. Syst.* ISSN: 2160-6455 (Mar. 2023).
3. Aguirre, C. *et al.* "Selecting Shots for Demographic Fairness in Few-Shot Learning with Large Language Models". *arXiv preprint arXiv:2311.08472* (2023).
4. Aguirre, C., Dredze, M. & Resnik, P. "Using Open-Ended Stressor Responses to Predict Depressive Symptoms across Demographics". *arXiv preprint arXiv:2211.07932* (2022).
5. Aguirre, C., Mahmood, A. & Huang, C.-M. "Crowdsourcing Thumbnail Captions via Time-Constrained Methods". in *27th IUI Conference* (2022), 36–48.
6. DeLucia, A. *et al.* "Bernice: a multilingual pre-trained encoder for Twitter". in *Proceedings of the 2022 conference on empirical methods in natural language processing* (2022), 6191–6205.
7. Aguirre, C. & Dredze, M. "Qualitative Analysis of Depression Models by Demographics". in *Proceedings of the 7th CLPsych Workshop* (2021), 169–180.
8. Aguirre, C., Harrigian, K. & Dredze, M. "Gender and Racial Fairness in Depression Research using Social Media". in *Proceedings of the 16th EACL: Main Volume* (2021), 2932–2949.
9. Harrigian, K., Aguirre, C. & Dredze, M. "On the State of Social Media Data for Mental Health Research". in *Proceedings of the 7th CLPsych Workshop* (Association for Computational Linguistics, Online, June 2021), 15–24.
10. Sherman, E. *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), 217–223.

11. Harrigian, K., Aguirre, C. & Dredze, M. "Do Models of Mental Health Based on Social Media Data Generalize?" in *Proceedings of the 2020 EMNLP Conference: Findings* (2020), 3774–3788.
12. Bose, A. *et al.* "A novel approach for detection and ranking of trendy and emerging cyber threat events in twitter streams". in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2019), 871–878.
13. Yang, H. *et al.* "Pipelines for Procedural Information Extraction from Scientific Literature: Towards Recipes using Machine Learning and Data Science". in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* **2** (2019), 41–46.
14. Aguirre, C. A. *et al.* "Towards Faster Annotation Interfaces for Learning to Filter in Information Extraction and Search." in *IUI Workshops* (2018).
15. Behzadan, V. *et al.* "Corpus and Deep Learning Classifier for Collection of Cyber Threat Indicators in Twitter Stream". in *2018 IEEE International Conference on Big Data (Big Data)* (2018), 5002–5007.
16. Maria, F. *et al.* "MATESC: Metadata-Analytic Text Extractor and Section Classifier for Scientific Publications." in *KDIR* (2018), 259–265.
17. Aguirre, C. A. *et al.* "Learning to Filter Documents for Information Extraction using Rapid Annotation". in *2017 International Conference on Machine Learning and Data Science (MLDS)* (2017), 85–90.

Research Experience

Fairness & Social 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. I developed finetuning [1] and evaluation methods [3, 7] for fairness of language models

Language Analysis of Mental Health 2020-22

Trained language models from Twitter [6], Reddit [9] and open-ended survey responses [4] to predict depression, measuring the fairness of the models we trained along gender and racial/ethnic groups, and analyzed the models errors using topic models.

Human-AI interaction 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. [2, 5]

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. [8] Analyzed the effect of temporal and domain shifts across social media platforms on mental health models. [10]

Teaching and Mentoring

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 that have led to publications, e.g. [3]

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.

Service

Reviewer 2023-present

Serve as a reviewer for the ACL Roling Review.

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.