(507)338-8898
Minneapolis, MN
lixx3013@umn.edu

Changye (Erin) Li

Data Scientist | Natural Language Processing

changyeli.github.io
github.com/changyeli
linkedin.com/in/changye

Data Scientist and Research Engineer with expertise in ML, NLP, and ASR. Specializing in developing explainable NLP methods to assist healthcare professionals and cross-disciplinary researchers.

Education

Doctors of Philosophy in Health Informatics, ABD, *University of Minnesota* 09/2019 — 05/2024 (exp.)
Master of Science in Data Science, *University of Minnesota* 09/2016 — 10/2018
Bachelor of Science in Economics, *University of Minnesota* 01/2013 — 05/2016
Bachelor of Arts in Statistics, *University of Minnesota* 01/2013 — 05/2016

Skills

Languages Python, R, SQL
Frameworks & Tools PyTorch, scikit-learn, NLTK, transformers, HuggingFace, Git, Docker

Experience

Research Intern **01/2023 — 04/2023**
Truveta *Bellevue, WA*

- Created and deployed in-house Transformer-based models for the automatic evaluation and improvement of millions of manual annotations, achieving an impressive classification accuracy and recall rate of 85%.
- Leveraged reinforcement learning and human feedback to enhance the efficiency and effectiveness of in-house models.

Graduate Research Assistant **03/2020 — Present**
University of Minnesota, College of Pharmacy *Minneapolis, MN*

- Developed *GPT-D*, a Transformer-based model achieving SOTA performance on detecting and generating dementia-related linguistic anomalies, *without* requiring corresponding large dataset produced by dementia patients.
- Developed *TRESTLE* (**T**oolkit for **R**eproducible **E**xecution of **S**peech **T**ext and **L**anguage **E**xperiments) that unifies the pre-processing audio and text data from Dementia Bank to address the reproducibility challenge from previous works.
- Designed and built a knowledge base system to support *HealthMarshall*, a voice-based personal assistant for cognitively impaired patients that answers their questions through every step of a clinical interaction.

Natural Language Processing Engineer **12/2018 — 08/2019**
University of Minnesota, Institute of Health Informatics *Minneapolis, MN*

- Created APIs to facilitate seamless concept integration between UMLS, BioPortal, and *iDISK*.
- Developed Python-based ETL pipelines to build and maintain the knowledge base for *iDISK*.

Digital Data Analyst Intern **07/2017 - 09/2017**
Padilla *Minneapolis, MN*

- Generated consumer profiles by evaluating personality scores with Machine Learning models for various clients, including consumer goods, automobile and oil companies.
- Collaborated with digital marketing and creative departments, transferred their business interests into data science problems to perform data analysis, including time series and regression analysis.
- Reduced 25 hours per month working time to 2 mins by building R Shiny web dashboards for product and user analysis.

Selected Publications

Li, C., Xu, W., Cohen, T., Michalowski, M., & Pakhomov, S. (2023, Mar.). TRESTLE: Toolkit for Reproducible Execution of Speech, Text and Language Experiments. In *AMIA Informatics Summit*

Li, C., Knopman, D., Xu, W., Cohen, T. and Pakhomov, S., 2022. GPT-D: Inducing Dementia-related Linguistic Anomalies by Deliberate Degradation of Artificial Neural Language Models. In *ACL*.

Guo, Y., **Li, C.**, Roan, C., Pakhomov, S. and Cohen, T., 2021. Crossing the "Cookie Theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Frontiers in Computer Science, 3, p.642517*.

Activities

Reviewer, **AMIA** Informatics Summit, **AMIA** Annual Summit, **ML4H** 2023
Programming Club leader, Institute of Health Informatics Student Group, University of Minnesota 2021 — 2022
Data Hackallenge Organizer, 6th International Workshop on Health Intelligence, **AAAI** 2022 2022