Minhwa Lee

minhwalee@umass.edu | Amherst, MA | github.com/mimn97 | linkedin.com/in/minhwalee

EDUCATION

University of Massachusetts Amherst — Amherst, MA
M.S. in Computer Science (Data Science Concentration); GPA: 3.8 *(*in transition to PhD program)* — *2021-* Current

The College of Wooster — Wooster, OH
B.A. in Computer Science; B.A in Mathematics (*Graduated with Magna Cum Laude*) — 2017-2021

Coursework: Advanced ML, Advanced NLP, Neural Networks, Probabilistic Graph Models, Reinforcement Learning, Responsible AI, Probability & Statistics, Mathematical Modeling, Data Visualizations, Master's Research Project

WORK EXPERIENCE

Microsoft — Cambridge, MA
Data Scientist Intern – NLP — Jan 2023 – Feb 2023

- Currently developing a named entity recognition (NER) model to predict and extract product-related entities from customer reports using pre-trained BERT and SpaCy's NER tagging models.
- Creating a Power BI dashboard to address major problems and challenges in current NER models and present the performance of our own NER models.

Bloomberg LP — Remote, MA
Graduate Student Researcher — *Jan 2022 – May 2022*

- Designed and implemented a Longformer-based language model architecture for document-level sequential sentence classification tasks, using PyTorch and Huggingface Transformers frameworks.
- Pre-processed several documents (e.g., abstracts, full papers, clinical notes) and applied domain adaptation techniques to these datasets for further developing a long-range cross-domain language model.

RESEARCH EXPERIENCE

Biomedical Informatics NLP Laboratory, UMass Amherst — Amherst, MA
Graduate Student Researcher (+ 2022 Summer Research Assistant) — *Aug 2021-Current*

- Public health applications of **NLP on Social Media [Abstract submitted to 2022 AMIA Symposium]**
  - Developed and tested zero-shot performance of two clinical language models on ~100K tweets: (1) a BioClinicalBERT and (2) a RoBERTa fine-tuned on) electronic health records.
  - Examined statistical association between the NLP models' performances and each of the following user characteristics: (1) the frequency of grammatical errors in the tweets and (2) degrees of neighborhoods disadvantage that the tweets were created in.
  - Constructing a large-scale Twitter dataset of users who tweeted health-related posts during COVID-19.
  - Developing a named entity recognition model using BERTweet and BioClinicalBERT to extract evidence of health-related mentions on general tweets.
- Public health applications of **Data Science [In preparation for Nature Food]**
  - Conducted linear regression analysis and statistical tests to identify the relationship between the travel time to a food pantry in the U.S and a neighborhood's characteristics within the food pantry's service area.
  - Hosted a website to post several interactive visuals of our research results using Tableau.
- Visual Word Sense Disambiguation Task **(SemEval-2023 Task 1) [Submitted to ACL 2023]**
  - Proposed a novel approach of using Bayesian inference to incorporate sense definitions of each polysemous word from the SemEval-2023 dataset into image-text matching models (e.g., CLIP, FLAVA).
  - Developed a context-aware definition generator of polysemous words using GPT-3 and employed it into our CLIP-architecture models, thus significantly increasing the original CLIP's performance by 10%.

SELECTED PROJECTS

[1] Honors Thesis: ML for Depressive Disorders among US adults: Developed supervised machine learning models (CART, Logistic Regression, SVM) that detect and predict the U.S. adults' depressive disorders from their socio-demography and health records, thereby achieved the precision of 84% in prediction *(Finalists, ACM Student Research Competition at GHC 2021).*

[2] Answering COVID-19 Questions with Medical Chatbot Applications: Trained a medical question-answering model by finetuning DialoGPT with BioBERT sentence embeddings of COVID-19 question-answer (QA) pairs. Improved precision scores of generated answers by 30%.

TECHNICAL SKILLS

- Languages/Software: Python, R, PostgreSQL, C, Bash, Git, Tableau, Power BI
- Frameworks: PyTorch, Huggingface Transformers, sklearn, Pandas/Numpy/Matplotlib, Seaborn, NLTK, SpaCy

LEADERSHIP EXPERIENCE

- **Organization Committee of Voices of Data Science 2022** (Oct 2021 - Mar 2022): Hosted a college event for 157 participants of underrepresented groups in CS at UMass Amherst, thus promoting diversity and inclusion in CS fields.