Casey K. Cooper

caseykcooper14@gmail.com • 206-409-8851 • https://www.linkedin.com/in/ckcooper/ • https://caseykcooper.me/

EDUCATION

University of Colorado Boulder **Boulder, CO**
Master of Science in Data Science Aug 2022 – Dec 2023
Cumulative GPA: 4.00

University of Washington, Department of Biochemistry **Seattle, WA**
Bachelor of Science in Biochemistry; Minor in Chemistry, Informatics Sep 2017 - June 2021
Cumulative GPA: 3.35
Dean's List, 5 quarters

SKILLS

Programming Languages: Python, R, SQL
Technical Skills: Scikit-Learn, Pandas, NumPy, Caret, e1071, TensorFlow, PyTorch, Machine Learning, Docker, GitHub, Jupyter Notebook, Spyder, Rstudio, Data Wrangling, Data Visualization (Matplotlib, seaborn, ggplot2), NLP, Prompt Engineering, LangChain, RAG, Microsoft Office
Soft Skills: Quick Learner, Work Ethic, Attention to Detail, Teamwork, Leadership

WORK EXPERIENCE

InfinitIntel **Boulder, CO**
Data Science Intern Jun 2023 – Present

- Assisted with development of a model to classify patients as at MMI or not at MMI based on their electronic health record
- Worked on three main projects: treatment extraction, synthetic data generation, feature engineering
- Treatment Extraction: Extract desired treatments that a patient has received from their electronic health record
    - Designed prompts for LLMs and dealt with long documents using the LangChain Framework
    - Created a method to extract desired treatments from an electronic health record with 98% accuracy
- Synthetic Data Generation: Generate synthetic text data to supplement small existing dataset
    - Used LLMs to generate synthetic electronic health records
    - Used text augmentation to generate synthetic electronic health records
    - Currently an ongoing project
- Feature Engineering: Find/generate features to improve model accuracy
    - Designed prompts for LLMs and dealt with long documents using the LangChain Framework
    - Found four general topics that could be added to the model as features: Pre-existing conditions, MRI report, X-Ray report, Range of motion details
    - Used LLMs to successfully extract these features from each electronic health record
    - Extracted features provided an improved model accuracy

InBios International **Seattle, WA**
Quality Control Associate I Aug 2021 – July 2022

- Ran PCR, ELISA, and Rapid assays to qualify each new lot of infectious disease diagnostic kits
- Interacted with various teams to troubleshoot kit lots that did not pass QC testing
- Performed PCR, ELISA, and Rapid test data analysis to ensure quality of products was maintained during manufacturing and throughout product shelf life
- Ensured that GMP regulations were followed through all aspects of product development
- Helped qualify ten million COVID rapid tests during the pandemic
- Trained new team members on all aspects of working in the quality control team

Chu Lab UW Medicine **Seattle, WA**
Research Assistant Nov 2020 – Aug 2021

- Administered COVID-19 test to study participants
- Ensured COVID-19 samples were stored and transported securely
- Greeted and checked in study participants for COVID-19 tests
- Certified in GCP/HIPAA, Biosafety, and Bloodborne Pathogens

DATA SCIENCE PROJECTS: More in depth details can be found on https://caseykcooper.me

Predicting the Total Result for NFL Games

- Goal: Create a model to predict whether an NFL game goes over or under the total set by sports books with an accuracy of at least 52.4% (this is the percentage that sports bettors need to win at to be profitable)
- Models Used: Decision Trees, Naïve Bayes, Support Vector Machines, Logistic Regression, XGBoost, Random Forest, ANN
- Results: The best model was SVM and ANN with accuracies of approximately 54%
- Conclusion: Successfully created a model that would be profitable for sports bettors, but not by much.

Tweet Sentiment Analysis

- This project explores the use of common NLP techniques (text pre-processing, TF-IDF Vectorizer with basic classification models, neural networks with learned embeddings)
- Goal: Create a model to accurately classify a tweet as Positive, Negative, Neutral, Extremely Positive, or Extremely Negative
- Models Used With TF-IDF: Logistic Regression, Naïve Bayes, XGBoost, Random Forest
- Neural Network Architectures Used: ANN, LSTM
- Results: Best TF-IDF Baseline (Logistic Regression): 0.5824, Best ANN: 0.6717. Best LSTM: 0.7120
- Conclusion: Successfully established baseline TF-IDF models. Successfully improved upon those models using more complex neural networks with learned embeddings.

NLI for Clinical Trials

- This project explores the use of LLMs for natural language inference
- Goal: Create a model to accurately classify whether a statement about a clinical trial entails or contradicts the information in the eligibility criteria, intervention, results, or adverse events sections
- Models Used: BioBert, Prompt Engineering
- Results: BioBert gave a ~0.6 F1 Score, Prompt Engineering gave a ~0.85 F1 Score
- Conclusion: Successfully used prompt engineering to create a classifier that accurately predicts whether a statement about a clinical trial entails or contradicts information from specific sections of the clinical trial

LEADERSHIP ACTIVITIES/VOLUNTEER EXPERIENCE

Special Olympics, *Student Volunteer* April 2018 – July 2018

- Coached and played for the University of Washington's Special Olympics soccer team
- Served food during the opening ceremony to the contestants/families

JDRF International, *Student Volunteer* April 2019

- Set up a course for a 5k race and helped administer snacks for the participants

Delta Tau Delta Fraternity, *Philanthropy Chair* Dec 2018 – Oct 2019

- Facilitated fundraisers for Cancer for College
- Coordinated with other chapters to increase participation in each other's philanthropy events

ADDITIONAL INFORMATION

Interests: Soccer, Baseball, Football, Snowboarding, Reading, Strategy Board Games