Zih-Dwo Yeh (Doyle)

(+1) 814-862-8109 doyleyeh.job@gmail.com linkedin.com/in/doyle980216 github.com/doyleyeh

Machine Learning Engineer | Software Developer | Cybersecurity Researcher

Education

The Pennsylvania State University University Park, PA, USA
Master of Science in Informatics August 2023 – May 2025

- Coursework: Computer Security, Software Security, Cybersecurity Analytics, Network Management, Cloud Security, Data Mining, Machine Learning

National Taipei University of Technology Taipei, Taiwan
Master of Science in Computer Science September 2021 – July 2023

- Coursework: Pattern-oriented Software Design, Design and Analysis of Algorithms, Cloud Computing, Data Science, Operating Systems, Network Security and Penetration Testing

National Taipei University of Technology Taipei, Taiwan
Bachelor of Science in Electrical Engineering September 2016 – July 2020

- Coursework: Object-Oriented Programming, Computer Networks, Image Processing, Python Programming

Skills

Programming Languages: C++, Python, MATLAB, SQL, JavaScript (or TypeScript)
Development: Node.js, Next.js, Express.js, Tailwind CSS
Machine Learning/AI: PyTorch, TensorFlow, Scikit-Learn, Hugging Face Transformers
Packages and Tools: Git/GitHub, Docker, Conda, VSCode, Shell/Bash, Linux, Wireshark

Experience

SYSTEX Corporation Taipei, Taiwan
IT Intern September 2018 – June 2019

- Maintained over 100 company workstations and laptops, troubleshooting hardware and software issues to ensure system stability and smooth operation.
- Configured and managed network settings, servers, and edge devices, ensuring compliance with security policies and company standards.
- Performed virus updates using FortiClient, safeguarding company systems from malware threats and optimizing network security.
- Conducted packet analysis with Wireshark to diagnose and resolve network issues related to Ethernet switching and IP routing.

Projects

Spam Classification API 2025

- **Developed** REST API for spam detection with **DistilBERT** and **FastAPI**, achieving **95% accuracy**.
- **Deployed and scaled** the system on **AWS Lambda** with **Docker**, enabling both server and serverless solutions, achieving approximately **50 ms inference latency**.
- **Created** an end-to-end ML training pipeline for preprocessing, fine-tuning, evaluation, and deployment, achieving a **prediction accuracy over 95%**.

Active Retrieval-Augmented Generation (RAG) by Small Language Models 2024 – 2025

- **Engineered** an active retrieval system dynamically integrating external knowledge for multi-hop QA tasks, enhancing model accuracy by **15–20%** on 2WikiMultihopQA.
- **Reduced** retrieval overhead significantly, achieving near-commercial performance using smaller models (**Llama, Gemma, Qwen**), saving infrastructure costs by approximately **30%** compared to full-scale LLMs.

Large Language Models (LLMs) Attribution Classifier 2023 – 2024

- **Built and labeled** a dataset of **25,000 prompt-completion pairs** using an automated pipeline for text extraction, truncation and generation.
- **Fine-tuned ALBERT**, **DistilBERT**, and **RoBERTa**, achieving **67.6% accuracy** (**0.68 F1**).
- **Performed** extensive **ROC-AUC evaluations** and detailed error analysis to identify confusion patterns, guiding future enhancements and boosting classifier precision.