FAITH WAVINYA MUTINDA

wavinya.faith@gmail.com | linkedin.com/in/faithwmutinda | github.com/faith-wm | Philadelphia, PA

Summary

Innovative Data Scientist with extensive expertise in NLP, specializing in fine-tuning language models for diverse tasks such as information extraction, summarization, and question answering. Skilled in end-to-end model development, including data preparation, distributed training, and benchmarking. Looking to leverage my technical skills in a data-driven role within a forward-thinking company.

Skills

Natural Language Processing	Machine Learning	Communication
Retrieval-Augmented Generation	Large Language Models	Team Collaboration
Distribute Training	Prompt Engineering	Problem Solving

Experience

NLP Postdoctoral Research Fellow *July 2024 - Present*
Children's Hospital of Philadelphia, U.S.A.

- Developed a **RAG-based LLM pipeline** using Llama and DeepSeek-Distill-Llama models to **extract and normalize** dysmorphic findings from unstructured genetics notes, improving model performance by **13%**.
- Fine-tuned **clinical QA LLMs** using **HPC GPU clusters** and **Cloud TPUs**; optimized performance, built benchmarking frameworks, and applied **RLHF** to improve alignment and accuracy.
- Mentored undergraduate students, providing **technical guidance** and overseeing **research projects**.

Python Data Scientist/Analyst *May 2024 - July 2024*
Turing, Remote

- Developed and curated robust datasets in data science and Python coding domains for LLM training and evaluation, ensuring quality and consistency.

Data Scientist *Apr 2023 - Nov 2023*
Biofourmis Singapore PTE. Limited, Singapore

- Developed **predictive machine learning models** using electronic health records, achieving high accuracy in **identifying patients at risk of chronic conditions**, bolstering precision medicine initiatives.
- Enhanced **hospital readmission prediction models**, **increasing accuracy by 10%** through extensive exploratory data analysis, feature engineering, and hyperparameter optimization.
- Collaborated with **cross-functional teams** to transform complex business requirements into effective and scalable data-science solutions.

Graduate Researcher (Data Scientist) *Apr 2019 - Mar 2023*
Nara Institute of Science and Technology, Japan

- Developed a **Longformer-based transformer model** that streamlined the meta-analysis process, achieving an **F1 score above 80%** by efficiently **extracting core concepts** from biomedical research articles.
- Trained a **transformer-based deep learning model** to compute the degree of **semantic similarity** in clinical texts, achieving a **correlation of 90%** with human scores.
- Fine-tuned **BERT-based** model for **medication and context extraction** from unstructured clinical notes, attaining a **94% F1 score** and earning a **top 10 ranking** in a shared task.

Education

Ph.D. in Information Science & Engineering **2023**
Nara Institute of Science and Technology, Japan

Master's in Information Science & Engineering **2020**
Nara Institute of Science and Technology, Japan

Projects

RAG-HPO: A Retrieval-Augmented LLM for Normalizing Physical Examinations in Genetic Clinical Notes

- Developed an AI model to identify and standardize descriptions of phenotypic abnormalities in unstructured genetic clinical notes, mapping them to Human Phenotype Ontology (HPO) terms.
- Utilized Retrieval Augmented Generation (RAG) with Llama 3.3-70B and DeepSeek-Distill-Llama models, achieving an F1 score of 0.82.

AUTOMETA: Automatic Meta-Analysis System Employing Natural Language Processing

- Fine-tuned a LongFormer-based model to extract key clinical elements - Participants, Intervention, Control, Outcomes (PICO) - from breast cancer clinical trial articles, achieving an F1-score over 0.80 for most entities.
- Parsed numeric patient outcomes, converting them to a structured format for statistical analysis.

Contextualized Medication Event Extraction

- Fine-tuned a BERT-based model to classify medication mentions by context into Disposition (medication change discussed), NoDisposition (no change discussed), and Undetermined, achieving an F1-score of 0.94.

Semantic Textual Similarity in Clinical Domain Texts

- Fine-tuned a BERT-based approach to measure semantic similarity in clinical texts pairs, achieving high Pearson correlation with human scores (0.904 on Japanese case reports, 0.875 on English clinical notes data).

30-Day Hospital Readmission Prediction for Heart Failure Patients

- Developed a LightGBM model to predict 30-day hospital readmission risk for heart failure patients using over 1 million patient records and features including comorbidities, demographics, labs, vitals, procedures, medications, admissions, emergency visits, outpatient visits, achieving an AUROC of 0.766.

Selected Publications

- **Mutinda, F.W.,** Liew K., Yada, S., Wakamiya, S., & Aramaki, E. (2022). Automatic Data Extraction to Support Meta-Analysis Statistical Analysis: A Case Study on Breast Cancer. BMC Medical Informatics and Decision Making.
- **Mutinda, F. W.**, Liew K., Yada, S., Wakamiya, S., & Aramaki, E. (2022). PICO Corpus: A Publicly Available Corpus to Support Automatic Data Extraction from Biomedical Literature. In Proceedings of the First Workshop on Information Extraction from Scientific Publications. Asia-Pacific Chapter of the Association for Computational Linguistics.
- **Mutinda, F.W.,** Yada, S., Wakamiya, S., & Aramaki, E. (2022). AUTOMETA: Automatic Meta-Analysis System Employing Natural Language Processing. MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation.
- **Mutinda, F.W.,** Yada, S., Wakamiya, S., & Aramaki, E. (2021). Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT. Methods of Information in Medicine.
- **Mutinda, F.W.,** Nigo, S., Shibata, D., Yada, S., Wakamiya, S., & Aramaki, E. (2020). Detecting Redundancy in Electronic Medical Records Using Clinical BERT. The Association for Natural Language Processing.
- **Mutinda, F.W.,** Nakashima, A., Takeuchi, K., Sasaki, Y., & Onizuka, M. (2019). Time Series Link Prediction Using NMF. IEEE International Conference on Big Data and Smart Computing (BigComp).