Arshitha Basavaraj

linkedin.com/in/arshitha/

Email : arshithab@proton.me
GitHub : github.com/Arshitha

Skills

- **Languages:** Python (PyTorch, HuggingFace, Pandas, Polars, NumPy, Scikit-learn), Shell Scripting, SQL
- **Tools, Technologies & GenAI Techniques:** HPC, AWS, Docker, Metabase, Git/GitHub, DataLad, Prompt engineering, Retrieval-Augmented Generation (RAG), Context Engineering

Projects

- **Summarization of Biomedical articles & Radiology Reports**
  - Built a unified prompting, retrieval, and multimodal LLM framework across 6 biomedical and radiology datasets, eliminating dataset-specific models and improving generalization.
  - Achieved **Top-3 leaderboard performance** (2nd in Radiology Translation, 3rd in Lay Summarization with external knowledge) using Llama-3.3-70B and GPT-4.1.
  - Boosted radiology translation and summarization quality by **28–35% over baselines** using similarity-based few-shot retrieval, structured role prompting, and UMLS-powered RAG.
  - Benchmarked small vs. large LLMs and showed that optimized prompting & retrieval strategies outperformed fine-tuned models without any training compute, reducing cost while improving accuracy.
  - **Published** at Association for Computational Linguistics (ACL) 2025 conference proceedings. 10.18653/v1/2025.bionlp-share.27

- **Language Usage Checker**
  - Built a real-time multilingual language usage checker that analyzed input text against a large crawled corpus.
  - Increased data collection throughput by **8×** using a multi-threaded BFS web crawler.
  - Designed a graph-based NoSQL database achieving **99.9% compression efficiency** for text storage.
  - Enabled multilingual processing using Stanford CoreNLP with custom tokenizers
  - **GitHub Repository:** https://github.com/Arshitha/EC504-Language-Correction

- **Toxic Comments Classification**
  - Built a toxic speech detection pipeline for a highly imbalanced 1:10 dataset, benchmarking TF-IDF, count vectorizers, Word2Vec, and deep models.
  - Improved AUC-ROC from **0.892 → 0.906**, outperforming the **0.84 human baseline** through optimized preprocessing, feature engineering, and combined word + character n-gram representations.
  - Demonstrated that tuned classical models (SVM, Logistic Regression) can match Bi-LSTM with attention on imbalanced text data, achieving **0.903** ROC-AUC with lower compute and complexity.
  - **GitHub Repository:** https://github.com/abhaysarda/jigsaw-toxic-comment-classification

Experience

- **International Institute of Information Technology** — Bangalore
  Data Engineer — *Jan 2025 – Present*
  - **Data Warehousing**: Improved data harmonization efficiency across **8 sites for 9,000 participants** time by automating data ingestion and transformation processes, saving an average of 20 hours per week.
  - **Data Pipeline**: Built automated scoring, de-identification (DPDP/HIPAA-compliant), and QC reporting pipelines that **improved data quality by 10%** through faster error detection, better record completeness, and scalable participant reporting.
- **National Institutes of Health** — Bethesda
  Data Engineer — *Sep 2019 - Sep 2024*
  - **Data Pipelines**: Built scalable, policy-compliant data pipelines to prepare and share research data across NIH, improving reproducibility of analyses.
  - **Data Curation**: Standardized multi-modal datasets for **10,000+ participants** across **40 NIMH labs**, enabling reproducible analyses for over **12 peer-reviewed studies**
  - **Anatomical Scans Defacer**: Developed open-source MRI de-identification and QC tools adopted within NIMH, and **reducing error rates by 14%** and cutting curation timelines by several weeks per release.
  - **Open-source contributions**: Contributed to the BIDS (Brain Imaging Data Structure) standard, authoring tabular data curation guidelines and automating PDF generation of the specification.

- **dataxu (acquired by Roku)** Boston
  Engineering Intern *May 2018 – Aug 2018*
  - **Event-driven data transfer**: Improved data transfer systems' efficiency by **95%** by migrating from CRON-based pipeline to an event-triggered, scalable pipeline.
  - **Technical Documentation**: Delivered a proof-of-concept automation tool integrating Sphinx for documentation generation across Python repositories.
- **Indian Statistical Institute** Bangalore
  Research Assistant *May 2015 – Aug 2015*
  - **Neural Networks**: Designed a two-layer feedforward neural network to classify six facial expressions from frontal face images that improved accuracy by **2%** over existing methods.

Education

- **Boston University** Boston
  Master of Science in Electrical and Computer Engineering *Sep 2017 – May 2019*
- **National Institute of Technology, Karnataka** Surathkal
  Bachelor of Technology in Electrical and Electronics *Jul 2012 – May 2016*

Publications

- **Prompts, Retrieval, and Multimodal Fusion** – ACL Anthology 2025 (LLM summarization)
- **Demonstrating QC procedures in fMRI**– Frontiers in Neuroscience 2023 (MRI data quality)
- **NIMH Healthy Volunteer Dataset** – Scientific Data 2022 (Large-scale multi-modal data curation)
- **Facial Expression Recognition and Classification** – IEEE 2015 (Neural Networks)
- More on **Google Scholar**