Bushra Sajid

(501)590-5714 | bushrasajid79@gmail.com | U.S. Citizen | Woodbridge, Virginia (Willing to relocate)

Summary

PhD researcher with expertise in advanced machine learning, NLP, and data quality, applied to healthcare, cybersecurity, and government domains. Developed cutting-edge AI tools for entity resolution and error correction using pre-trained language models and custom ML pipelines. Conducted NSF-funded research and collaborated with the FDA on biomedical data analysis using BioBERT. Proficient in Python, SQL, Spark, and AI frameworks including LLMs, RAG, and diffusion models. Skilled at translating large-scale data into actionable insights and communicating complex findings to both technical and non-technical audiences. Experienced in Agile, stakeholder engagement, and mentoring graduate students in data governance and quality.

Experience

Data Scientist January 2025 - Present

Axicom world

- Developed automated data pipelines using Python (pandas, NumPy, and SQLAlchemy) to clean, transform, and integrate structured and semi-structured datasets across multiple sources
- Conducted exploratory data analysis (EDA) using pandas, seaborn, and matplotlib to uncover key business insights and inform architectural decisions
- Applied statistical models and machine learning algorithms (via scikit-learn) to forecast customer behavior and optimize business operations
- Implemented data validation and quality assurance protocols using Great Expectations and Python based test suites to maintain data integrity
- Contributed to client-facing documentation and presented technical findings to non-technical stakeholders, aiding in data-driven decision-making

Research Assistant - NSF DART August 2022 - December 2024

University Of Arkansas At Little Rock *Little Rock, AR*

- Conducted applied research in large-scale data quality improvement using Python, with a focus on enhancing accuracy, consistency, and duplicate detection in heterogeneous datasets through advanced Entity Resolution techniques
- Developed and integrated deep learning models (BERT, DistilRoBERTa) using PyTorch and Hugging Face Transformers into the Data Washing Machine (DWM) framework, replacing traditional rule-based matching with semantic similarity scoring for high-accuracy record linkage
- Designed a privacy-aware linkage method (”SSN Filtering with ML”) using pandas and scikit-learn, improving precision while maintaining balanced recall, aligning with ethical and privacy-compliant data practices
- Built scalable ML pipelines for unsupervised ER, including blocking, pair-wise comparison, and similarity classification to process millions of records while minimizing candidate pair generation efficiently
- Validated model performance across 18 real-world datasets using F1 score, precision, and recall, demonstrating effectiveness and generalizability of the ER models within the NSF DART initiative

Graduate Teaching Assistant - Department of Information Quality August 2023 - December 2024

University Of Arkansas At Little Rock *Little Rock, AR*

- Assisted seventy plus students in teaching graduate-level course on Information Quality, covering topics such as data governance, data quality metrics, and data profiling
- Taught hands-on sessions using tools such as OpenRefine, Talend, and Ataccama for data cleaning, transformation, and quality analysis
- Built custom Python scripts and labs for anomaly detection, duplicate identification, and schema validation
- Guided students on real-world data workflows and enterprise-level data quality solutions
- Supported the instructor in developing teaching materials and managing course logistics

Data Scientist intern July 2021 - August 2022

Sequretek *Little Rock, AR*

- Scripted python SQL, SQLAlchemy to extract data from Relational database (MySQL and Oracle)
- Implemented Talend to create basic ETL pipelines to extract data from multiple tables into a file to use in Python
- Customized an anomaly detection module to identify distinct user events, solving multiple use cases such as data ex-filtration and insider threats, and analyzing the Windows activity logs

- Developed and maintained Python and SQL scripts to streamline data processing and analysis tasks
- Emphasized using advanced deep learning techniques such as Variational Autoencoder, Hierarchical Temporal Memory, or Graph-based Approaches. Attention Models, GAN's to create and compare user and entity activity against their profiles and their peers' profiles

Researcher August 2020 - May 2021
U.S. Food and Drug Administration (FDA) *Jefferson, AR*

- Designed python-SQL library to connect and extract drug data from Oracle database
- Accessed the data by web scraping and used tokenization and text classification to manage the data
- Implemented Python Libraries and BERT Model for the named Entity recognition and visualized the data using tools like Matplotlib and Seaborn to create informative charts and graphs highlighting key patterns and insights
- Led NLP pipeline development for ADR extraction from drug labels using BioBERT and NER techniques, achieving 91 percent accuracy

Web Designer August 2019 - August 2020
University Of Arkansas At Little Rock *Little Rock, AR*

- Maintained and updated the UALR Computer Science department website, ensuring accurate and timely content updates
- Designed and implemented user-friendly web interfaces using HTML, CSS, and JavaScript to enhance the website's usability

Education

University Of Arkansas at Little Rock Little Rock, AR
Ph.D. in Computer and Information Sciences— 3.75 *August 2019 - December 2024*

Thesis Title: Advanced models for linking process in Data Washing Machine
PhD Supervisor: Dr. Ahmed Abu-Halimeh

University Of Arkansas at Little Rock Little Rock, AR
Bachelor of Science in Computer Science *May 2019*

Technical Skills

Languages: Python, R, SQL, C++, JavaScript, HTML/CSS
ML/AI Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost
MLOps and Infra: MLflow, Docker, Git, CI/CD **Data tools:**Spark, Hadoop, Talend, OpenRefine, MLFlow, Databricks
Databases: PostgreSQL, Oracle, MySQL, Vector DBs (Qdrant, Pinecone)
NLP/GenAI: Large Language Model (LLM), Retrieval Augment Generation (RAG), Diffusion model, BioBERT, NER, Prompt Engineering
Methodologies: Agile, Scrum

Publications

1. Sajid, B., Abu-Halimeh, A., Jakoet, N. (2024). "Pre-trained models for the linking process in Data Washing Machine." Computing and Artificial Intelligence. DOI: 10.59400/cai.v3i1.1450.
2. Sajid, B., Abu-Halimeh, A. Talburt, J. (March, 2025). "SSN filtering method with pre-trained models for entity matching in Data Washing Machine." *AI Insights*, Vol. 1 No. 1. DOI: 10.62617/aii1929.

Research Funding

- NSF EPSCoR-funded collaborative research initiative (Award No. OIA-1946391), focusing on introducing Machine Learning into the Data Washing Machine.