# Ali Naeim abadi

Database Lab, Computing Science Dept., Edmonton, AB | (438) 924-0342 | naeimaba@ualberta.ca | LinkedIn | GitHub | Personal Website

## Highlights of Skills

- Over 5 years of experience in developing and maintaining complex ML/Deep Learning projects on a large scale, such as ML-based client risk management models for a major Canadian financial institution
- Developed client Risk management Model/fraud management model using multi-modal data (entity relationships, transaction history, news, tweets, financial reports) for AML/ATF Department of a major Canadian financial institution
- Experienced in LLMs for developing risk management, entity matching, and tweet classification models
- Experienced in quantitative methods and time-series forecasting using multi-variate ensembles of deep learning models
- Experienced in large-scale DataBase Management Systems (DBMS) using distributed computing platforms such as Hadoop, Distributed File System (HDFS), Hive, Spark, PySpark, Azure Data Factory, Azure Databricks, etc.
- Fluent in SQL/NoSQL databases using Oracle DB, IBM DB2, MySQL, PostgreSQL, SQLite, MongoDB
- Experienced in ETL development for large datasets: financial data, client data, e-commerce data
- Experienced in analytical and Machine Learning (ML) models through different projects (e.g., client risk management, entity matching and tweet classification models) and by leveraging scikit-learn, pandas, NumPy, matplotlib, seaborn libraries, XGB, etc.
- Experienced in cloud computing using AWS, GCP, Azure, etc.
- Demonstrated strong teamwork skills by being part of more than 4 teams of different sizes and collaborating with cross-functional teams over various large-scale projects

## Related Experience

### Data Scientist Intern | Scotiabank

September 2021 - April 2022

Responsibilities:

- Developed a client risk management model using multi-modal data (entity connections, transaction history, news, and financial reports) for Wealth department (GRM) - AML/ATF Department
- Implemented a database integration system using a 2-stage string matching methodology: blocking stage, machine learning-based filtering
- Constructed a Knowledge Graph using internal client data, Wikidata, and third-party databases
- Built an internal search engine using Elasticsearch for internal unstructured data
- Implemented the deep learning and machine learning Proof Of Concept (POC) models into a scalable pipeline
- Applied database extraction and database cleaning

### Research Assistant | University of Alberta

August 2022 - August 2023

Responsibilities:

- Designed SOTA few-shot LLM-based e-commerce product entity matching models using LLMs (RoBERTa, GPT3) and tabular data
- Designed unsupervised attribute ranking modules for tabular data using LLMs (RoBERTa, SBERT)
- Developed ETL (Extract, Transform, and Load) solution for our datasets
- Created two e-commerce product entity matching datasets including tabular and textual information for each product, by collecting and scraping online data: Walmart*-Amazon* and Amazon*-Google
- Applied Database extraction and Dataset sanitization

### Machine Learning Developer | Faroob Zaman Inc.

September 2016 - June 2020

Responsibilities:

- Developed time series forecasting models for stock price prediction and sale price prediction using ensembles and deep learning models
- Designed deep learning for sentiment analysis on website reviews to evaluate customer satisfaction
- Designed AI-assisted microscope image processing software for the thin-layer lab using deep learning (CNN)

## Technical skills

- **Programming Languages:** Fluent in Python | Fluent in SQL, C++ | R, Scala, Java, MATLAB, C#
- **Databases (SQL/NoSQL)**: Oracle DB, IBM DB2, MySQL, PostgreSQL, SQLite, MongoDB, Redis, Casandra, Hive, Neo4j
- **Machine Learning:** Data Science, Data analytics, Data Mining, Analytical models, Time series data forecasting, Machine Learning algorithms
- **Deep Learning**: PyTorch, Hugging Face, TensorFlow, Keras
- **Parallel Processing**: Apache, Spark, Hadoop, CUDA, PyTorch Lighting, PyTorch DistributedDataParallel
- **Cloud Computing**: Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, IBM Cloud, Oracle Cloud
- **Other**: Docker, GitHub, JIRA, Software Development Processes (Agile, Waterfall, DevOps, TDD: Test-Driven Development, Spiral), Data Structures and Algorithms, Applied statistics, qualitative model, Clean Code, Code Refactor and Review, Multithreading, Big Data, data

analytics, statistics, SAS, SPSS, STATA, Tableau, Microsoft Power BI, Google Data Studio, BI, Business Intelligence, business analysis, PowerPoint, MS Excel, MS Access, Visual Studio, scikit-learn, SciPy, NumPy, Pandas, NTLK, XGB, XGBoost, GENSIM, Seaborn, Matplotlib, Linux

# Education

### MSc. in Computing Science | University of Alberta

- Thesis: Product entity matching by leveraging tabular data and Large Langauge Models.
- Supervisor: Prof: Davood Rafiei.
- GPA: 3.84

### BSc. in Computing Engineering | Amirkabir University

- Thesis: Microscope images processing to predict fracture strength of nanoclay/polyamide 12 nanocomposites using CNN.
- Supervisor: Dr. Farkhindeh Hemmati.
- GPA: 3.96

# Selected Projects

### Implementing a risk management model using multi-modal data

A robust deep learning model flags high-risk clients given multi-modal sources of data including transaction history, connections with other entities, news feeds on Twitter, and financial journals (e.g., Harvard Business Review, Wall Street Journal, etc.). To discover the connection between clients, a knowledge graph was constructed based on enriched client databases. | Scotiabank
Skills: risk management, GRM, financial risk management, data modeling, multi-modal data analysis, data fusion, data analytics, data extraction, ETL, NLP, information retrieval, IR, cloud computing, parallel computing, knowledge graph construction, deep learning, LLM, DBSM, database management systems, HDFS, Hadoop, Apache, Spark, PyTorch, Scikit-learn, SQL, PostgreSQL, Python, data cleansing, data pre-processing.

### Time series forecasting models using deep multi-variate LSTM and ensemble models

A robust, high-performance time series forecasting model is implemented to predict the stock price. I designed ensembles by combining LSTM and decision-tress-based models. The ensemble model of XGBoost+multi-variate-LSTM outperformed other baselines for the AAPL stock price prediction task [GitHub Code]. | Faroob Zaman
Skills: time-series forecasting, data modelign, time series data modeling, deep time series forecasting, quant, quantitative model development, multi-variate LSTM, Prophet model, Neural Prophet model, decision tree, XGBoost, Light GBM, ensemble models, ensemble deep models, tensorflow, sci-kit learn, APPLE stock price prediction, TESLA stock price prediction, marketing analytics, data cleansing, data pre-processing.

### E-commerce product matching model using Large Language Models and tabular data: TATEM

A fast, highly informed, and effective record-matching model is designed to find the matching pairs among online e-commerce platforms: Google-Amazon and Walmart-Amazon. For the first time, the model benefits from product titles and other complementary textual and tabular information. | University of Alberta
Skills: LLM, table serialization, tabular data encoding, high-dimensional similarity search, information extraction, ETL development, Extract, Transform, and Load development, machine learning, marketing analytics, data analytics, user behavior modeling, GPT3, BERT, RoBERTa, SBERT, Universal Sentence Embedding, SBERT, transformers, deep learning, PyTorch, SQL, Hadoop, Spark, business intelligence, BI, AWS, GCP, cloud computing, Large Language Models Python, data cleansing, data augmentation, data generation.

### LLM-based Disaster Tweet Classification

A robust disaster news classification model was implemented using LLMs such as BERT, RoBERTa, and GPT3. Given the text of a tweet, the model checks if the tweet is about a piece of real disaster news or a fake one [GitHub Code]. | University of Alberta
Skills: data modeling, NLP, LLM, BERT, GTP3, text summarization, data manipulation, data augmentation, neural networks, deep learning, transformers, text classification, Python, PyTorch, cloud computing, AWS, Natural Language Processing, Large Language Models, data cleansing, data pre-processing.

### Introduced two product entity matching datasets: Amazon*-Google, Walmart*-Amazon*

We created new product entity matching datasets in which, for the first time, product-specific tabular data is given as complementary source of domain knowledge. This additional knowledge helps us disambiguate hard negative examples. | University of Alberta
The Tabular data was extracted from Amazon.com through a scraper tool that I developed: Amazon Textual and TAbulaR Information extractIon (ATTARII) [GitHub Code].
Skills: data collection, data quality, data governance, Web Scraping, information extraction, IE, Information retrieval, IR, tabular data, online product matching, user behavior modeling, marketing analysis, business intelligence, BI, ETL development, Extract, Transform, and Load, Beautiful Soup, Selenium, Requests, SQL, Python.

### Database integration between two large-scale databases

Database integration was applied through a 2-stage string matching method: LSH and a Snorkel rule-based classifier | Scotiabank.
Skills: databases, database integration, PySpark, Hadoop, SQL, PostgreSQL ETL, data governance, machine learning, data analytics, data cleaning, locality sensitive hashing (LSH), cloud computing, snorkel.

# Selected Publications

- **A. Naeim abadi**, T. Nayeem, D. Rafiei, TATTOO: Product Entity Matching as a Topology Construction, WSDM 2024 [submitted].
- **A. Naeim abadi**, T. Nayeem, D. Rafiei, Product Entity Matching via Tabular Data, CIKM 2023, Oct 21-23, 2023, Birmingham, United Kingdom.
- T. Firoozi, O. Bulut, C. Demmanse Epp, **A. Naeim abadi**, D. Barbosa, The Effect of Fine-tuned Word Embedding Techniques on the Accuracy of Automated Essay Scoring Systems Using Neural Networks, NCME 2022, April 21-24, 2022, San Diego, USA