TOLGA ŞAKAR

Data Scientist, Msc in Applied Data Science

Contact Information

Email: tolgasa2@gmail.com Github: github.com/dfavenfre
Phone: +90 539 772 13 42 Kaggle: kaggle.com/dfavenfre/code
Medium: medium.com/@bauglir LinkedIn: linkedin.com/in/tolga-şakar

Programming Languages & Skills

- Python (Pandas, Numpy, Statsmodel, Scikit-learn, Tensorflow, Keras, PyTorch, LangChain, LangGraph), Julia (Flux, Zygote),
- SQL (SQLite, MSSQL)
- Development Frameworks (Docker, FastAPI, Streamlit, HTML, CSS)
- LangChain (Conditional Tool Calling, Customized Tool Architecture, Few-shot Prompt Engineering, Chatbot Security)
- LangGraph (Customized Multi-Tool Agent Architecture, Flow Engineering with multi-tools, Chatbots with Human-Feedback)
- Transfer Learning (HuggingFace, Kaggle)
- Predictive Modelling (Time-Series Forecasting, CARTs)
- Optimization (HyOPT, Keras Pruner, Optuna, Randomized Search, Grid Search Bayesian)
- ML models, LLMs and Chatbot Development and Monitoring (Weights & Biases, LangSmith, Predibase, OpenAI Fine-tuning)
- Web Scraping (Selenium, PlayWright)

PUBLISHED RESEARCH

- Cambridge University Journal of Natural Language Processing (Q1) : **Maximizing RAG efficiency: A comparative analysis of RAG methods**

OPEN-SOURCE PROJECTS

- **Multi-Modal RAG**

Developed a multi-modal **Retrieval-Augmented Generation (RAG)** system that integrates both text and image data for improved information retrieval and question-answering. The project uses a dataset from a Macroeconomics 101 course, including a PDF textbook and embedded images (e.g., graphs and charts). Preprocessing involved text extraction and summarization of images and tables. Implemented a **multi-modal ChromaDB** vectorstore for storing and retrieving content, allowing for multi-index search across text, images, and tables. Evaluated system performance based on retrieval accuracy, image description generation, and latency.

- **TalkYou**

TalkYou is an open-source project that enables users to chat with YouTube videos using a chatbot interface powered by **LangChain** and **LangGraph**. The system supports both **Retrieval-**

Augmented Generation (RAG) for information queries and image retrieval from videos. The backend, built with **FastAPI** and containerized with **Docker**, integrates **Nvidia CUDA** support for efficient **Speech-to-Text (STT)** processing via the **Whisper** model. The frontend uses **Streamlit** for a lightweight, interactive interface, making TalkYou a dynamic tool for engaging with YouTube content in real time.

- **RAG Optimization**

Developed an optimization framework for **Retrieval-Augmented Generation (RAG)** systems using publicly available insurance documents. Key contributions include document preprocessing through PDF chunking, vectorstore creation using **FAISS**, and **grid-search** optimization across various **RAG** methods and embedding models (**OpenAI, Cohere, BGE**). Evaluated model performance using multiple metrics such as coherence, context accuracy, relevance, helpfulness and conciseness evaluated by **GPT-4**. Achieved significant improvements in contextual accuracy and relevance in RAG-generated responses.

- **LLMRoboFund**

LLMRoboFund is an innovative investment chatbot designed to provide real-time information on funds and ETFs by utilizing **LLM** with **Retrieval-Augmented Generation (RAG)**. This system taps into financial platforms like Turkey's **TEFT** and **PDP** to access up-to-date data, enabling users to query investment strategies, financial risks, management fees, and other critical fund details. By integrating **vectorstore (Text-to-Text)** and **SQL databases (Text-to-SQL)**, LLMRoboFund ensures comprehensive, current insights, enhancing the investment research process without the need for time-consuming manual efforts.

- **MobileNet Julia Implementation**

This project implements MobileNet v1 from scratch in **Julia** using the **Flux** framework, optimized for mobile and embedded devices through **depthwise separable convolutions**. The model was trained on the **CIFAR100** dataset, achieving efficient performance with reduced parameters. The architecture is adjustable with **width (α)** and **resolution multipliers (ρ)**, allowing it to scale for different environments. Training was conducted on an **NVIDIA RTX 3050 Ti**, leveraging CUDA for faster computation. The repository provides a streamlined approach for deploying MobileNet in resource-constrained settings.

- **Face Recognition**

This project implements a custom **Convolutional Neural Network (CNN)** in **PyTorch** for image classification using the **Olivetti-Faces** dataset, which includes 400 grayscale images of 40 individuals. The CNN architecture features three convolutional layers with ReLU activations, max-pooling, and fully connected layers. The model was trained on a **Google Colab A100 GPU** with CUDA, completing training in 3.5 hours. The training achieved a validation accuracy of **93.33%** after 256 epochs, and results were monitored using **Weights and Biases**. This model provides a robust solution for facial recognition tasks.

- **Electricity Price Forecasting**

This project focuses on forecasting hourly electricity prices using various modeling approaches based on data from the **EXIST Market Transparency Platform**. Initially, a baseline XGBoost (XGBM) model was developed using lag-1 of the exogenous variable for predictions, yielding performance metrics for comparison. The model was then fine-tuned with hyperparameters such as max_depth, n_estimators, and subsamples, resulting in improved performance indicated by lower root mean squared error (RMSE). Subsequently, a **Long Short-Term Memory (LSTM)** model was implemented to capture long-term dependencies in time-series data, outperforming the fine-tuned XGBM model with an RMSE of **57.348** compared to **59.972** for the XGBM. The project demonstrates effective strategies for electricity price forecasting by leveraging both traditional and deep learning models.

WORK EXPERIENCE

Data Scientist

Naviga AI (Ankara,TR)

Dates: 01/2024 – Present

I am responsible for developing chatbots and AI agents utilizing the LangChain and LangGraph frameworks to create customized and multi-tasking and self-reinforced AI tools. I take direct responsibility for agent, tool and chatbot development processes, such as RAG optimization, multi-modal RAG, developing customized tools and agents, and prompt engineering to both instruct large language models and to prevent prompt injection attacks.

I have hands-on experience working with LLMs, Speech-to-Text, Text-to-Speech models, Multi-Modal RAG, and fine-tuning LLMs with few-shot examples. Additionally, I engage in containerized, with Docker, back-end development utilizing the FastAPI framework to provide endpoints for chatbots.

- **ChatTEDU** – I've worked in chatbot development for TED University. The chatbot is equipped with Text-To-SQL, Text-To-Text, Speech-To-Text and Text-to-Speech capabilities, along with security measures on conditional tool calling for preventing prompt injection attacks.

- **Talk With Data** – I led the development of an innovative talk-with-data chatbot that empowers users to engage with complex data tasks through conversational interactions. The chatbot simplifies advanced data exploration, enabling users to ask nuanced questions without coding. It facilitates data preprocessing and feature engineering through intuitive prompts, and guides users in creating predictive models and conducting predictive analytics. Additionally, the chatbot incorporates automated hyperparameter tuning, optimizing model performance in real-time for enhanced outcomes. This project revolutionizes how professionals extract insights and visualize data, making sophisticated analytics accessible to all.

Data Scientist

Bosphorus AI (Ankara,TR)

Dates: 10/2023 – 01/2024

Achievements/Tasks

I was involved in a variety of projects, including the development of a fraud detection model for the EXIT Platform. This model aims to identify market participants engaged in price manipulation within order books. In addition, I was tasked with creating chatbots that guide customers toward recommended insurance products for an insurance company. Furthermore, I was also responsible for developing recommendation models based on customer segmentation tasks.

Equity Research Analyst

Valens Research (Istanbul, TR)

Dates: 01/2022 – 03/2023

Achievements/Tasks

My responsibilities encompassed the preparing of investment recommendation reports achieved through a sequence of rigorous data-driven analyses. This included intricate financial and economic data modeling, as well as thorough peer comparisons. I frequently used SPSS IBM for statistical inference and Python when conducting time-series analysis.

Quantitative Research Intern

E2T (Remote, UK)

Dates: 07/2021-10/2021

Achievements/Tasks

I worked in quantitative research projects, acquiring methodologies to derive insights using statistical inference techniques, revealing trends in finance and economics.

EDUCATION

TED University

Master of Science in Applied Data Science, Turkey (GPA: 3.62 / 4)

Lodz University of Technology, Poland

Bachelors of Science in Econometrics (Erasmus +) (GPA : 3.1 / 4)

TED University, Turkey

Bachelors of Science in Economics & Business Administration (GPA : 3.3 / 4)