AWANTIKA SRIVASTAVA

Data Scientist | Software Engineer – AI | FastAPI | LLM

+91-8920482037 |sawantika81@gmail.com| LINKEDIN |Github

PROFILE SUMMARY

Junior Software Engineer (AI) with **2+ years of hands-on experience** in building and deploying **AI-powered backend services** using **Python, FastAPI, and modern ML/LLM workflows**. Strong exposure to **REST APIs, Docker, AWS deployment, and Git-based collaboration**. Experienced in developing **LLM and RAG-based systems**, writing production-ready code, and working across the **front-to-backend lifecycle**. Comfortable translating requirements into scalable implementations and iterating on solutions to improve end-user experience.

CORE TECHNICAL SKILLS

- **Programming Languages:** Python (Advanced), C++,SQL, Jupyter Notebook
- **Statistics & Mathematics:** EDA, Statistical Modeling, Hypothesis Testing, Confidence Intervals, A/B Testing, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Model Evaluation Metrics (Precision, Recall, F1-score, ROC-AUC).
- **Deep Learning :** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Transfer Learning, Model Fine-tuning, Computer Vision.
- **Deep Learning Frameworks:** TensorFlow, keras, Pytorch, TensorFlow Lite.
- **GenAI & LLMs:** LLMs, Prompt Engineering, RAG, LangChain.
- **Databases & Storage:** SQL Databases, Vector Databases (FAISS – exposure), Data Modeling.
- **DevOps & Cloud:** Docker, AWS (EC2, S3), Service Deployment, Containerized Applications.
- **Backend & APIs:** FastAPI, Flask, RESTful APIs, HTTP, Client–Server Architecture.
- **Tools & Collaboration:** Git, GitHub, Version Control, Code Reviews, Debugging.

EXPERIENCE

Data Scientist / Machine Learning Engineer | PPS International Pvt. Ltd. **January 2024-Present**

- Designed and developed **AI-powered backend services** using **Python and FastAPI**, exposing ML and LLM models through **RESTful APIs.**
- Built, containerized, and ran backend services using **Docker**, supporting consistent development and deployment workflows.
- Deployed and tested containerized services on **AWS (EC2, S3)** with guidance from senior engineers, ensuring reliability and scalability.
- Implemented **LLM-based workflows** including prompt engineering, document ingestion, and retrieval pipelines for production use cases.
- Applied foundational **HTTP and client–server concepts** to design clean request/response contracts and robust API interfaces.
- Collaborated with cross-functional teams using **GitHub**, contributing to code reviews, feature development, bug fixes, and testing.
- Wrote maintainable, well-documented code and participated in improving **service reliability and overall end-user experience**.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- **Designed** and deployed a real-time **computer vision** ML system for unsafe driver behavior detection using **CNN-based** models (**SSD MobileNet with TensorFlow**).
- Built **end-to-end ML pipelines** including data ingestion, preprocessing, feature extraction, model training, evaluation, and deployment.
- Optimized models using **TensorFlow Lite** to achieve **20–25 FPS real-time inference** with **<150 ms latency** on production/edge environments.
- Exposed model inference through **FastAPI-based REST services** for integration with downstream systems and monitoring tools.
- Performed **error analysis, bias checks, and performance monitoring** to improve model robustness and operational reliability.

Chatbot Using LLM & RAG | Applied ML Project

- Built a Lightweight **LLM-Powered chatbot** using **TinyLLaMA** to answer user contextual questions over structured resume data.
- Implemented a **Retrieval-Augmented Generation (RAG) pipeline**, including document indexing and similarity-based retrieval, to ground LLM responses.
- Applied **prompt engineering and response evaluation techniques** to improve answer relevance, consistency, and factual accuracy.
- Developed **FastAPI-based inference endpoints** to serve the LLM pipeline in a scalable manner.
- Deployed the application using **Streamlit**, focusing on low memory usage, fast inference, and user-friendly interaction.

YouTube Comments Sentiment Analyzer | link - https://youtube-ai-analyzer-ndzqo6r2mepjrtsdjmwaxl.streamlit.app/

- Built an end-to-end **NLP pipeline** for sentiment analysis using **transformer-based models (BERT/DistilBERT)**.
- Performed text preprocessing, feature extraction, model training, and evaluation on large volumes of user-generated data.
- Evaluated model performance using **precision, recall, and F1-score**, and iteratively improved results through error analysis.
- Served the trained model via **FastAPI APIs** for real-time inference and integration.

CERTIFICATION

- IBM Data Science & AI Certification
- AWS Generative AI with Large Language Models
- OpenCV Computer Vision Certification

EDUCATION

IMS Engineering College, Ghaziabad **September - 2020**

Bachelor of Technology (Electrical and electronics engineering)