AWANTIKA SRIVASTAVA

AI Engineer | Machine Learning Engineer | LLMs

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

PROFILE SUMMARY

AI/ML Engineer with **2+ years of hands-on experience** in building, evaluating, and operationalizing **machine learning and data science solutions**. Strong foundation in **supervised and unsupervised learning**, feature engineering, and model evaluation, with growing exposure to **deep learning, LLM workflows, and AWS-based ML systems**. Experienced in collaborating with engineers and data scientists to deliver **production-ready AI components** using Python, scikit-learn, and cloud-native tools.

CORE TECHNICAL SKILLS

- **Programming & Data Science:** Python, C++, SQL., Pandas, Numpy, Scikit-learn, PySpark
- **Statistics & Mathematics:** Statistical Modeling, Descriptive Statistics, Hypothesis Testing, Probability, Sampling, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Model Evaluation Metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC).
- **Deep Learning & AI:** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Transfer Learning, Model Fine-tuning, Time-series Forecasting, ARIMA, SRIMA.
- **Deep Learning Frameworks:** TensorFlow, keras, Pytorch, TensorFlow Lite.
- **NLP:** Text preprocessing, Tokenization, Chunking, Sentiment Analysis, Topic Modeling (LSA, LDA), Transformer-based models, Computer Vision.
- **Generative AI:** LLMs (LLaMA, GPT), RAG, Prompt Engineering & Prompt Strategies, LLM Fine-Tuning (PEFT/LoRA), HuggingFace, Semantic Search, Query Understanding, Reranker Design (Cross-Encoder), Multi-Agent Systems, Routing & State Management
- **Vector & Search Systems:** Vector Databases, Embeddings, OpenSearch / Elasticsearch, Hybrid Search & Semantic Retrieval
- **Cloud & MLOps:** AWS Sagemaker, S3, EC, DataBricks, MLflow, Streamlit, Langchain
- **Tools & Deployment:** Microservices Architecture, REST APIs, Model Deployment & Monitoring, CI/CD, Git, Docker, Kubernete.

EXPERIENCE

AI/ML Engineer | PPS International Pvt. Ltd. **January 2024-Present**

- Implemented **end-to-end GenAI powered RAG and multi-agent systems** using LangChain and HuggingFace.
- Designed **LLM agents with routing, state management, validation rules, and prompt strategies** for contextual enterprise solutions.
- Built embedding-based **vector search pipelines integrating OpenSearch and reranker models** for improved semantic retrieval.
- Fine-tuned LLM models using **PEFT/LoRA techniques** for domain-specific adaptation.
- Worked extensively on **Databricks (PySpark, Spark SQL)** for distributed data processing, feature engineering, and ML workflows.
- Translated business requirements into technical ML solutions including data preparation, modeling, evaluation, and deployment.
- Developed scalable ML APIs using **Python (OOP principles) and microservices architecture**.
- Deployed ML and GenAI systems on **AWS**, ensuring low-latency real-time inference.
- Supported end-to-end project delivery including business understanding, data analysis, modeling, evaluation, and stakeholder communication.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- **Designed and deployed a** real-time computer vision–based ML system **for unsafe** driver behavior detection using **CNN-based SSD MobileNet** model.
- Trained and optimized models on large-scale video datasets, achieving **20–25 FPS** real-time processing with **<150 ms inference latency**.
- Implemented **end-to-end ML pipelines** for data ingestion, preprocessing, model training, evaluation, and production inference.
- Deployed optimized models using **TensorFlow Lite** on edge/production environments for continuous monitoring.
- Built a **Flask-based web dashboard** to visualize detections and automatically record **30-second event clips**, reducing manual review effort.

Chatbot Using LLM & RAG | Applied ML Project

- Built a Lightweight **LLM-Powered chatbot** using **TinyLLaMA** to answer user queries over content.
- Implemented a **Retrieval-Augmented Generation (RAG) pipeline** to retrieve relevant resume sections for contextual question answering.
- Selected **TinyLLaMA** to ensure **low memory footprint and fast inference**, making the solution suitable for resource-constrained environments.
- Applied **prompt engineering techniques** to improve response relevance and consistency.
- Deployed the chatbot as an interactive **Streamlit web application** for real-time user interaction.

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

- Deployed transformer-based **NLP models** as **production-ready** services with **REST APIs**.
- **Fine-tuned** and served a **DistilBERT**-based **sentiment** classification model for large-scale text inference.
- **Built** and deployed an interactive **streamlit web application** to perform real-time **sentiment analysis** on YouTube comments.
- **Processed** high-volume text **data** with sub-second inference **latency** for real-time sentiment analysis.

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