AWANTIKA SRIVASTAVA

AI/ML Engineer | GenAI | NLP | LLMs

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

PROFILE SUMMARY

AI / Machine Learning Engineer with **2+ years of hands-on experience** in designing, building, and deploying **production-grade machine learning and Generative AI systems**. Strong expertise in **Python, Deep Learning, NLP, LLM-based models, and Retrieval-Augmented Generation (RAG)**. Proven experience across the **end-to-end ML lifecycle**, including data ingestion, model training, deployment, monitoring, and optimization, delivering scalable AI solutions aligned with real-world business needs.

CORE TECHNICAL SKILLS

- **Programming & Data Science:** Python, C++, SQL., Pandas, Numpy, Scikit-learn, Jupyter Notebook.
- **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 :** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Transfer Learning, Model Fine-tuning, Computer Vision.
- **Deep Learning Frameworks:** TensorFlow, keras, Pytorch, TensorFlow Lite.
- **Generative AI:** Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Prompt Engineering, Semantic Search.
- **NLP:** Text preprocessing, Tokenization, Sentiment Analysis, Transformer-based models.
- **Model Development & Deployment:** End-to-End ML Pipelines, Model Training & Tuning, Model Serving, REST API Development, FastAPI.
- **MLOps & Production:** MLflow, Model Monitoring, Model Performance Tracking, CI/CD for ML, Experiment Tracking, Version Control (Git).
- **Cloud & DevOps:** AWS (EC2, S3, SageMaker – basic), Docker.
- **Databases & Tools:** SQL, Vector Databases (FAISS / similar), Git, Linux.

EXPERIENCE

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

- Designed and implemented **end-to-end pipelines** covering data ingestion, preprocessing, model training, **evaluation** and deployment.
- Developed and deployed **machine learning and deep learning models** using **Python**, **TensorFlow**, and **PyTorch** for real-world applications.
- Built **computer vision–based ML systems**, applying image preprocessing and CNN models to extract meaningful insights from visual data.
- Integrated ML pipelines with **MongoDB** to manage structured and unstructured data reliably.
- Deployed and tested ML models on **AWS environments**, ensuring scalability, stability, and performance.
- Analyzed and optimized deployed models to improve **efficiency, latency, and reliability**.
- Collaborated with cross-functional teams to understand project requirements and deliver impactful ML solutions.
- Documented ML workflows, experiments, and deployment processes to ensure clear knowledge transfer.

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 **production-ready Generative AI application** using **Large Language Models (TinyLLaMA)** and **Retrieval-Augmented Generation(RAG).**
- Designed pipelines for **document ingestion, text chunking, embeddings, vector retrieval, and LLM-based inference orchestration**.
- Applied **prompt engineering and response evaluation strategies** to improve accuracy, relevance, and reliability.
- Implemented **context-aware semantic search** for grounded and factual responses.
- Deployed the chatbot as a **Python-based inference service** with observable workflows aligned to modern LLM systems.

YouTube Comments 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 (Electrical and electronics engineering)