Vikrant

New Delhi,
9411063933
vikranty622@gmail.com

www.linkedin.com/in/vikrantyadav1234
https://github.com/vikrantyadav11234

AI/ML Engineer with expertise in developing scalable, multilingual AI solutions for text, document, and speech-based applications. Proficient in fine-tuning and optimizing Large Language Models (LLMs) for low-resource languages, document intelligence systems, and speech-to-speech models. Successfully built end-to-end products like **DocSense**, an AI-powered document extraction and classification platform, and a **PDF translation system** preserving original layouts in Indian languages. Experienced in audio data processing pipelines, real-time conversation analytics, LLM quantization (GPTQ), and deploying lightweight, optimized models on cloud infrastructure. Passionate about AI for multilingual and enterprise applications with a strong foundation in Python, PyTorch, NLP, Computer Vision, and backend API development.

EDUCATION

Master of Technology, *National Institute of Technology, Warangal, Telangana*
Bachelor of Technology, *Institute of Engineering and Technology, Lucknow*

SKILLS

Data Science & AI Skills	Statistics, Machine Learning, Deep Learning, CNN, NLP/NLU, Transformers, Attention Mechanism, MLFLOW, LLMs
Frameworks/Libraries:	PyTorch, TensorFlow, Keras, Scikit-learn, FastAPI, Flask, SentencePiece, Hugging Face, ONNX
Languages Known	SQL, Python
Optimization & Deployment	LoRA, GPTQ, vLLM, Nvidia dynamo, GCP, Azure
Data Visualization	Matplotlib, Seaborn, Pandas, NumPy

EXPERIENCE

AI/ML Engineer* | *Aiverbalyze Technologies Pvt. Ltd **06/2024-Present**

- **Built DocSense:** A document AI product for automatic data extraction, classification (Invoices, Bank Statements, Aadhar, PAN, Legal Documents, etc.), and JSON conversion from diverse document types including digital PDFs, scanned images, and handwritten prescriptions. Different Types of Documents one Solution.
- **Performed LLM quantization (GPTQ)** on models like **Qwen2.5-VL-7B-Instruct** at **8-bit and 4-bit precision**, followed by deployment using **vLLM** for high-speed, efficient inference in production environments.
- Developed an AI-based **PDF translation tool** converting English PDFs into Indian regional languages while retaining layout and design.
- Automated **audio data extraction and chunking** workflows for STS model training and multilingual speech applications.
- Enhanced client onboarding APIs using **FastAPI**, adding a **Customer-Agent Conversation Analysis** feature to detect unanswered customer queries in real-time, store them in a database, and integrate management-reviewed responses into future conversations.
- Optimized large language and translation models including **facebook/seamlessM4T**, **Meta-LLaMA 3.1-8B**, and **Qwen2.5-VL** for production environments.
- Currently working to make an **STS AI Agent** for end-to-end speech-to-speech response generation.
- Built backend APIs for client onboarding systems using **FastAPI.**
- Fine-tuned google/mt5-large for NMT purpose for 22 languages.
- Developed a Neural Machine Translation (NMT) model for low-resource Indian languages, converting it and other models like facebook/seamlessM4T and Meta-LLaMA 3.1-8B from PyTorch to ONNX for optimized inference and deployment using NVIDIA Triton Inference Server.
- Fine-tuned Meta-LLaMA 3.1-8B with LoRA and the Unsloath AI library for faster, more efficient text generation on low-resource setups, and supported model deployment on cloud platforms like GCP VM and OVH Cloud.

Assistant Professor **02/2021— 08/2023**

Institute of Engineering and Technology Lucknow

- Taught Digital Electronics, Machine Learning, Artificial Neural Network to Instrumentation Engineering students at IET Lucknow. Led labs for core instrumentation subjects.
- Completed a certification in Data Science and Advanced AI, gaining expertise in Python, SQL, Statistics, Machine Learning, Deep Learning, CNN, NLP, and Generative AI.

CERTIFICATIONS

- Deep Learning and NLP Professional Certificate Course- Learnbay Pvt. Ltd

OTHER PROJECTS

CUSTOMER SEGMENTATION | *Python / Unsupervised Learning(clustering) / Seaborn / Matplotlib / Data preprocessing*

- segment(group/cluster) customer on the basis of buying pattern RFM (Recency Frequency Pattern)
- Identify sales trends for days, months and season time by invoice number, Identify highest sales trend item wise

HOMESTAYS DATA ANALYSIS AND PRICE PREDICTION | *ML / NLP / Hyperparameter tuning / Data Preprocessing*

- Build a robust predictive model to estimate the `log_price` of homestay listings based on comprehensive analysis of their characteristics, amenities, property_type and host information.
- Geospatial Analysis: Map listings using latitude and longitude data to visualize price distribution. Investigate whether specific neighborhoods or proximity to city centers impact pricing, offering a spatial view to the pricing strategy.
- Sentiment Analysis on Textual Data: Use advanced NLP techniques to analyze sentiment in description texts. Assess if positive or negative descriptions affect listing prices. Integrate these insights into the predictive model during training.

ML MODEL FOR CAB CANCELLATION DATA | *Data Preprocessing/ ML / Seaborn / Matplotlib*

- The business problem tackled in this project is trying to improve customer service for YourCabs.com, a cab company in Bangalore. The problem of interest is booking cancellations by the company due to unavailability of a car. The challenge is that cancellations can occur very close to the trip start time, thereby causing passengers inconvenience.
- Create and compare classification ML models for this problem.

EDA AND DATA PREPROCESSING | *Pandas/ NumPy / Seaborn / Matplotlib*