NISHAN JAIN

7410576180 | gr8nishan@gmail.com
LinkedIn | Github

SUMMARY

Senior Machine Learning Engineering Leader with 13+ years of experience architecting enterprise-scale production AI systems across NLP, Generative AI, Computer Vision, and RAG applications. Proven expertise in end-to-end ML pipeline automation, model operationalization, and leading cross-functional teams to deliver scalable AI solutions. Skilled in mentoring technical talent, establishing ML best practices, and transforming business requirements into production-ready AI systems.

WORK EXPERIENCE

Principle AI Engineer, Daxa AI **April 2024- Present**

- Architected and deployed an end-to-end MLOps pipeline for RAG-based QnA system, implementing automated model versioning, monitoring, and serving capabilities to support enterprise-scale question answering.
- Established a comprehensive model validation framework using statistical analysis for topic classification dataset quality verification, implementing automated data drift detection and model performance monitoring pipelines.
- Operationalized BERT-based topic classification system with CI/CD deployment pipeline, including automated model retraining triggers and performance monitoring dashboards for production environments.
- Implemented scalable LLM serving infrastructure by deploying Gemma 9B model using VLLM on GPU.
- Built an automated model evaluation pipeline using DeepEval for RAG systems, implementing continuous monitoring (CM) and automated model quality assessment with configurable performance thresholds.
- Built Entity and Topic Classification systems using LLMs with an automated MLOps pipeline, including Weights & Biases experiment tracking, dataset versioning, and CI/CD deployment.
- Built intelligent chatbot with LangGraph that routes user queries to specific URLs and APIs

Machine Learning Lead, Canwill Technologies **Oct 2021 – March 2024**

- Built a large-scale data processing pipeline that handles terabytes of data with automated data ingestion, transformation, and quality checks.
- Operationalized GPT-3.5 integration for resume parsing system, building production API with automated prompt engineering, response validation, and structured data transformation pipelines. This helped in reducing resume parsing cost from 7 cents to 3 cents.
- Built an end-to-end image classification pipeline using YOLO with automated model deployment, monitoring, and serving infrastructure—integrated object detection models with business logic to generate automated property descriptions through a microservices architecture. Used a semi-supervised approach to create training data.
- Built a zero-shot medical text classification, achieving 90% false positive reduction for ICD-10 detections

---

Senior Associate, MSCI **Mar 20214- Sep 2024**

- Led a project on corporate event extraction from news feeds, enhancing financial event tracking and decision-making.

Machine Learning Lead, Canwill Technologies **Oct 2020 – May 2021**

- Built production-ready NER pipeline using CRF and LSTM models for email text processing, implementing automated model retraining workflows and performance monitoring with configurable drift detection
- Fine-tuned and deployed a YOLO model for document categorization, establishing an end-to-end pipeline from data ingestion to model serving with automated model versioning and deployment strategies
- Implemented text classification microservices using transfer learning on BERT models, building containerized deployment with Docker, and establishing CI/CD pipelines for model updates and monitoring
- Architected feature engineering pipeline extracting 40+ text attributes with automated data transformation, validation, and serving capabilities supporting downstream ML applications at scale
- Built an Entity Linking production system using BERT and various other NLP techniques like POS, coreference resolution, etc
- Developed an Active Learning pipeline for the PII deduplication system, establishing automated data labeling workflows

---

Product Technical Lead | Senior System Engineer , Infosys Limited **Sep 2011 – Mar 2020**

- Production Video Analytics System: Built a real-time video analytics pipeline for weapon detection with automated model deployment, monitoring, and alert generation capabilities. For the same system, we developed a BERT-based harmful text detection system

- Built a speech analytics pipeline using DeepSpeech for speech transcription and DBSCAN clustering on TF-IDF features to identify frequently asked questions from call center audio transcriptions

- Architected a scalable AI-powered content mining platform with an intuitive template builder, integrating sophisticated NLP techniques, including part-of-speech tagging, named entity recognition, and text classification, to enable non-technical users to extract structured data from unstructured documents

- Led a cross-functional team of 10 professionals managing enterprise gamification platform implementations across multiple client engagements, driving 30% improvement in user engagement and delivering tailored solutions for employee recognition and performance management

---

TECHNICAL SKILLS

Programming Languages - Python, .NET

Databases - SQL, MongoDB, ElasticSearch

Vector DB - Qdrant**,** Vespa**,** Pinecone

Model Deployment: Docker, Kubernetes, VLLM

API Framework - Flask, FastAPI, Django

LLM Frameworks - AWS Bedrock, OpenAI, Litellm, Langgraph, Langchain, CrewAI, Langsmith, Ragas, DeepEval, Together AI

AI ML Frameworks - Numpy, Pandas, Scikit Learn, Keras, Pytorch, Hugging Face, Transformers, Spacy, NLTK

Model Architectures - Naive Bayes, BERT, Roberta, Random Forest, YOLO, FasterRCNN, Deepspeech, Llama, GPT, Claude, Deepseek, Mistral, Resnet, Transformer, Encoder-decoder, Layout Analysis

OCR - Tesseract OCR, PaddleOCR, EasyOCR, Unstructured IO, Docling, Pypdf,

MLOPS - MLFlow, Weights and Biases

EDUCATION

Bachelor of Engineering, Information Technology **Aug 2007 - June 2011**

MBM Engineering College