Sunil Thapa

AI Engineer

+1 4379700802 | sunil43thapa@gmail.com | Toronto, Canada | linkedin.com/in/sunil-thapa99

Profile

Passionate about AI, Machine Learning, and NLP, with hands-on experience in LLMs, Retrieval-Augmented Generation (RAG), and AI-driven automation. Experienced in building scalable AI solutions for various industries, including finance and retail, using LangChain, Transformers, PySpark, and cloud platforms (AWS, Azure, GCP). Strong interest in applying AI to real-world challenges, particularly in document retrieval, predictive analytics, and automation. Continuously learning and exploring advancements in AI while collaborating with cross-functional teams to develop impactful solutions. Published research in NLP and deep learning, contributing to the AI research community.

Skills

- **Languages:** Python, SQL, JavaScript, Shell Scripting
- **NLP & AI:** Langchain, Transformers, RAG, Cohere, OpenAI, Agentic AI, Hugging Face, Spacy, Gensim, Stanza
- **Computer Vision & Time-Series Forecasting:** OpenCV, Pose Estimation, Gesture Recognition, LSTMs, ARIMA
- **Web Development & API Frameworks:** Flask, Django, FastAPI, HTML, CSS, JavaScript, jQuery
- **ML & Data Science:** TensorFlow, PyTorch, Keras, Scikit-Learn, XGBoost, Pandas, NumPy, SciPy
- **Big Data & Distributed Computing:** PySpark, Hadoop (HDFS, Hive), Spark SQL
- **Databases:** MySQL, MongoDB, VectorDB, Oracle SQL
- **Cloud & Deployment:** AWS (S3, Lambda, CloudWatch), Azure ML, GCP, Docker, Kubernetes, Terraform
- **MLOps & CI/CD:** Git, GitHub Actions, Jenkins, JIRA
- **Soft Skills:** Communication, Teamwork, Leadership, Consulting, Business Problem-Solving

Professional Experience

AI Engineer, Agilepitch *July 2024 - December 2024*

- Enhanced document retrieval accuracy by 28% by developing and deploying LLM-based Retrieval-Augmented Generation (RAG) pipelines, utilizing iterative testing and self-corrective algorithms.
- Optimized large-scale data processing by 35% through PySpark pipeline enhancements, reducing execution time and improving efficiency.
- Increased system reliability by 13% by integrating scalable cloud infrastructure with AWS S3, Lambda, and CloudWatch, ensuring real-time monitoring and automated failovers.
- Reduced deployment time by 40% by automating CI/CD pipelines using Terraform and GitHub Actions, enabling seamless integration and rapid iteration of updates.
- Enabled scalable deployments by containerizing ML services with Docker and Kubernetes, optimizing resource utilization and system efficiency.
- Collaborated with cross-functional teams to fine-tune LLMs for financial data processing, enhancing structured data insights for clients.

Software Engineer, MAGDOTNET *April 2022 - August 2022*

- Increased user engagement by 75% by developing a recommendation system using collaborative filtering, improving personalized product suggestions.
- Enhanced navigation services by integrating Google Maps API, enabling real-time location-based updates for an online platform.
- Reduced response time by 16% through automated Docker container deployment on Amazon ECS, improving scalability and service uptime.
- Boosted data-driven decision-making by conducting big data analytics with SQL and Hive, uncovering insights that led to a 20% increase in efficiency.

Software Engineer, InfoDevelopers Pvt. Ltd. *July 2019 - March 2022*

- Lowered inventory costs by 25% by designing and deploying an ML-powered inventory management system, accurately predicting stock requirements and optimizing supply chain efficiency.
- Achieved 95% face recognition accuracy by developing a face-attendance system using Support Vector Machines (SVM) and OpenCV, implementing robust image preprocessing techniques.
- Enhanced customer insights by leveraging NLP-based sentiment analysis, extracting valuable feedback data to improve product satisfaction.
- Cut manual testing time by 30% by automating end-to-end testing with Selenium, streamlining quality assurance processes.
- Boosted real-time object tracking efficiency by 35% by applying advanced computer vision techniques for video analysis, improving detection accuracy and speed.

PROJECTS

Sales Automation AI (Time-Series Demand Forecasting & Optimization)

- Developed an AI-driven inventory forecasting system that predicts demand for retail stores, reducing stock shortages by 30%.
- Automated order placement through warehouse integration, ensuring optimal stock levels.
- Implemented LSTMs and ARIMA models for accurate time-series forecasting, improving prediction accuracy by 20%.
- Deployed as a SaaS platform with Azure ML & FastAPI, allowing real-time AI performance monitoring.

Financial Report QA System (LLM-Powered Question Answering for Financial Data)

- Designed a LangChain-based financial QA system for NEPSE-listed companies, enabling structured query-based insights from financial reports.
- Integrated vector databases & NLP models to enhance structured financial data retrieval.
- Applied retrieval-augmented generation (RAG) techniques, achieving a 30% improvement in financial document understanding.
- Optimized deployment with Azure AI services, ensuring secure and scalable performance.

Rental Rasa Chatbot

- Built a conversational interface for real estate rental property search using the Rasa framework.
- Integrated natural language processing capabilities to interpret user requests based on location, price, and property type.
- Leveraged machine learning and rule-based approaches for accurate and context-aware responses, achieving seamless user experience.

Nepali Text Sentiment Analysis

- Developed a sentiment analysis model for Nepali text, enabling efficient classification of user sentiments in native language datasets.
- Created word embeddings for Nepali text using Word2Vec, GloVe, and BERT, addressing the lack of open-source solutions for the language.

Education

Post Graduate in Artificial Intelligence & Machine Learning, *Lambton College* *2024*

BSc in Computing (Software Engineering), *University of Northampton* *2019*

Research Paper

- **“NepaliBERT: Pre-training of Masked Language Model in Nepali Corpus” (I-SMAC, 2023):** Developed Word2Vec, Doc2Vec, and BERT embeddings for benchmarking NLP tasks in Nepali language.
- **“Adult Income Prediction Using ML Algorithms” (SSRN, 2023):** Conducted comparative analysis of ML models, achieving 86% accuracy for income prediction.
- **“Clothes Identification Using Inception ResNet V2 & MobileNet V2” (SSRN, 2021):** Developed a deep learning-based clothing classification model, achieving 84.1% accuracy.