Manish Lama

NLP Engineer

Driven by a profound interest in Machine Learning and Natural Language Processing, I am a dedicated and self-motivated learner. I bring a combination of strong programming skills and innovative thinking, enabling me to swiftly develop solutions that elevate organizational performance and support my professional growth.

manishlama5053@gmail.com

+977 9808015715

New Baneshwor, Kathmandu 44600, Nepal

Skills

- Technologies: LangChain, Guidance, DSPy, Elasticsearch, Chroma DB, ColbertV2, RASA, Ollama
- Programming Languages: Python, JavaScript
- Frameworks and Tools: Docker, LangChain, Ollama Python Client
- Languages: Python, Java, C, C++, CSS, HTML5
- Frameworks: Rasa, Huggingface, scikit-learn, Keras, PyTorch
- Web Technologies & Methodologies: gRPC, Protobuf
- Web Services: FastAPI
- Version Control: GitHub

Languages

- Nepali - Fluent
- English - Fluent
- Hindi - Intermediate
- Spanish - Beginner
- French - Beginner
- German - Beginner
- Chinese - Beginner
- Italian - Beginner

Work Experience

Software Engineer, AI Team Lead, Associate Principal Engineer • eSewa

• Chatbot Development and Integration

• Designed and Developed Advanced Chatbots: Led the design and development of sophisticated chatbots by integrating cutting-edge technologies such as LangChain, Guidance, and DSPy for LLM management, alongside Elasticsearch, Chroma DB, and ColbertV2 for implementing a robust Retrieval-Augmented Generation (RAG) system.

• Enhanced Capabilities with RASA and Proprietary LLMs: Directed the integration with RASA and proprietary LLM engines, significantly enhancing the chatbot's capability to handle both basic and advanced queries. This included developing a strategic AI roadmap to guide future advancements.

• Fine-Tuning and Experimentation with LLMs

• Fine-Tuning LLMs: Fine-tuned several large language models, including LLaMA2, Mistral v0.1, Mistral v0.2, and Gemini. This involved optimizing model performance and ensuring high relevance and accuracy in responses.

• Conversion to Ollama: Experimented with converting these models to run on Ollama, a platform designed for running open-source large language models locally. Ollama provides a streamlined process for managing model weights, configurations, and data, significantly reducing latency and enhancing privacy by processing data locally

Advanced RAG Implementation:

• Hybrid Approach to RAG: Implemented an advanced RAG system using a hybrid approach that combines Chroma DB and ColbertV2, ensuring efficient and accurate information retrieval.

• Reducing Model Hallucination: Reduced model hallucination through optimized instruction fine-tuning provided by the DSPy framework. This included employing a teacher-forcing mechanism where one model generates an answer, and another model validates its relevance and accuracy.

November 2022 - Present

Software Engineer, NLP Team Lead • Treeleaf Technologies Pvt Ltd

• Directing and overseeing the in-house Personal Assistant project, which entails receiving user input and automating a spectrum of tasks encompassing appointment scheduling, ticket generation, team communication via calls and messages, reminder establishment, and to-do list management.

• Leading the development and oversight of a Meeting Assistant project, which demonstrates the capability to intelligently transcribe spoken content into text form, enabling the generation of comprehensive meeting transcripts, summaries, and minutes. Furthermore, this advanced AI system actively analyzes the conversation to provide automated suggestions for tasks such as ticket creation, reminder setting, to-do list management, and other pertinent actions.

• Collaborating closely with the Project Manager to ascertain project requirements, skillfully delegating tasks to team members, fostering seamless coordination with fellow team leads, and producing a comprehensive monthly progress report.

• Designing instructive materials and structured training programs aimed at facilitating the professional growth and seamless integration of new employees and trainees into ongoing NLP projects.

November 2021 - November 2022

Software Engineer, NLP • Palm Mind Technology Pvt. Ltd

• Framework Transition: Led the migration of the chatbot system from Snape to the Industry Standard Rasa Framework.

• Architectural Expertise: Designed and implemented a comprehensive architecture for data handling, pipeline evaluation, result analysis, and model deployment within the chatbot system.

• Language Specialization: Developed and managed chatbots specializing in the Nepali (Devnagari) language, with a focus on Romanized Nepali.

• Training and Skill Enhancement: Developed instructional resources and structured training programs aimed at facilitating the growth and smooth integration of new team members and trainees into active NLP projects

November 2022 - April 2022

Software Engineer, NLP • Treeleaf Technologies Pvt Ltd

- Developing multilingual chat bots for languages such as English, Spanish, Hindi and Nepali using Rasa Framework
- Work on the optimization of chat bots, specifically reducing loading time and improving accuracy
- Work on development and optimization of the myriad of NLP tasks such as question generation, paraphrasing, summarization

July 2021 - November 2021

Solution Consultant • The Real PBX

- Responsible to handle Voice over Internet sales inquiries received over phone, chat or email
- Understand the requirements of the clients and draft an appropriate hosted solution for them
- Pitch drafted solution to the clients and explain to them the merits of the solution
- Responsible to do a continuous follow-up with leads/clients until the client is converted
- Responsible to handle client retention process under the guidance of TL/Manager

March 2018 - March 2021

Network Engineer • Worldlink Communications Ltd.

- IP Networking, Routing and Switching and its protocols, security, VPN and Internet
- MPLS technology and its different services (L2VPN, L3VPN)
- Layer 3 - IP and related technologies (ICMP, TCP, GRE, QoS)
- Responsible for installing, configuring, troubleshooting and documentation of enterprise solutions

December 2016 - March 2018

Communications Engineer Intern • Nepal Telecom

- Learn about the working mechanism of 3G and 4G technologies
- To get acquainted with the technologies used

November 2016 - November 2016

Education

2012 - 2016

Kathmandu Engineering College

Electronics and Communication Engineering

2008 - 2010

Trinity International College

High School

Key Projects

• eVA(eSewa Virtual Assistant)

1. eVA is Nepal's first payment-enabled AI assistant, currently serving over 3 million registered users. It efficiently handles a wide range of user inquiries, from simple sign-up questions to complex transaction-related issues. As Nepal's pioneering AI assistant, eVA facilitates transactions such as mobile top-ups, wallet-to-wallet money transfers, and the purchase of airline tickets, setting a new standard for digital payment solutions in the region.

eVA

• Anydone

1. Anydone is an AI-powered system that enables us to create a chatbot for your service as a service provider and also allows the customers to communicate with the chatbot, along with a task management system, and a professional communication platform.

anydone

• Chat Bot created based on the paper "End-to-End Memory Networks" by Rob Fergus,Jason

Weston,Arthur Szlam,Sainbayar Sukhbaatar

1. Chat bot project was executed using BaBi / baby dataset released by Facebook research using concepts of

Keras Tokenizer and pad sequences,encoders alongside the concept of above-mentioned papers where

LSTM,Input, Activation, Dense, Permute, Dropout,add, dot, concatenate layers were used alongside

numpy and set (for creating vocabulary)

2. Link: https://github.com/manishl7/chatbot_endtoendnw.git

• Text Classification

1. Ham/spam classification completed using a Pipeline concept (Tfidf and classifier combined) where

linearSVC(support vector classifier) classifier was used from SVM (Support Vector Method)

2. Tweet Classification Project completed using Scikit Learn (Pipeline using Tfid Vectorizer and Support

Vector Module's SVM, Naive Bayes's Multinomial NB model, Logistic Regression model)

3. Link: https://github.com/manishl7/TweetClassification.git

- **Semantics and Sentiment Analysis**

1. Cosine similarities calculated using Semantics and word vectors with spacy

2. Sentiment Analysis project completed using NLTK's (Natural Language Tool Kit), VADER (Valence Aware Dictionary for sEntiment reasoning)

3. Link: https://github.com/manishl7/sentimentanalysis_Vader.git

- **Topic Modelling**

1. -Topic Modelling Project executed using LDA (Latent Dirichlet's Allocation) with Scikit learn and Count Vectorizer on corpus data.

2. -Topic Modelling Project executed using NMF (Non-Negative Matrix Factorization) with Scikit learn and TF-IDF (Term Frequency - Inverse Document Frequency Vectorizer) on corpus data.

3. Link: https://github.com/manishl7/Topic-Modelling.git

- **Classification using Neural Network**

1. Iris classification using concepts of SciKit Learn, ReLu (Rectified Linear Unit), concept of scaling using MinMaxScaler and simultaneously using concepts of Hot encoding along with Keras library's Sequential model, Dense layer and categorical crossentropy

2. Link: https://github.com/manishl7/Iris_classification_RNN.git

- **Text Generation using LSTM (Long short-term memory) based model**

1. Text generation project executed using concepts of LSTM, Keras Tokenizer, numpy array, Sequential model, Embedding layer, LSTM layer and Dense Layer.

2. Link: https://github.com/manishl7/text_generation_LSTM.git

- **Binary & Multi Image Classification:**

1. Using CNN, Keras, Convolution, Pooling, Numpy, checkpoint and check models to create a model that carries out binary & multi-image classification with high accuracy

- **Object Detection using SSD and YOLO**
- **Image creation using GAN**

Awards and Certification:

- NLP-Natural Language Processing with Python certification from Udemy.com – Jose Portilla, Head of Data Science, Pierian Data Inc.
- Natural Language Processing with Classification and Vector Spaces from DeepLearning.AI
- Fine-Tuning Large Language Models course from DeepLearning.AI
- Machine Learning with Python Certification from Broadway Infosys
- Recipient of a merit-based full scholarship for high school education from NeBA(Nepal Basketball Association).

Personal Traits:

- Adept self-learning skills
- An amicable team worker and with a hard-working ability.
- Comfortable working in Windows, Linux or similar environment
- Good written and verbal communication skills both in English and Nepali

Links:

- GitHub: https://github.com/manishl7
- LinkedIn: https://www.linkedin.com/in/manish-lama-b63776185/