Sunil Ghimire

NLP Engineer

+977 9841070311 | info@sunilghimire.com.np | sunilghimire.com.np | linkedin/ghimiresunil | github/ghimiresunil

NLP Engineer renowned for reshaping the NLP landscape with a project that achieved remarkable success in Large Language Models (LLMs). Distinguished for spearheading a transformative project where the implementation of Mistral model with DPO trainer led to an unprecedented 20% increase in accuracy compared with LlaMa, GPT-J, GPT-Neo and Llama2, setting a new industry standard for NLP applications. Specialized in LLMs, including Named Entity Recognition (NER), relation extraction, document classification, and pretraining/fine-tuning. With over 3 years of experience, I excel in collaborative team efforts, optimizing NLP pipelines, and seamlessly integrating models into chatbot systems. Proficient in TensorFlow, PyTorch, NLTK, and SpaCy, my expertise extends to end-to-end model development and deployment, complemented by a robust background in Data Mining, Machine Learning, Statistical Modelling, Business Intelligence, SVM, Logistic Regression, XGboost, and Neural Networks.

WORK EXPERIENCE

Machine Learning Engineer, Fusemachines Aug 2021 - Present

- Developed and fine-tuned a conversational generative AI model using LlaMa, GPT-J, GPT-Neo, Mistral, and LlaMa2 with the Hugging Face Trainer API, leveraging medical domain knowledge to simulate a realistic doctor-patient interaction. Conducted a comprehensive comparative analysis, revealing that LlaMa emerged as the superior performer in this context. Additionally, implemented Mistral by training it with one epoch using the Trainer API and further fine-tuned the Mistral model with the DPO trainer for three epochs, achieving a remarkable 20% improvement in accuracy compared to other models.
- Built and implemented a document classification and document segmentation model where sentence transformers were used to find the embedding of textual content and the XGboost model was implemented for document classification. Similarly, Zero-Shot Classifier was used for document segmentation.
- Built and fine-tuned using HuggingFace transformers with spaCy 3 to perform Name Entity Recognition on text data with an accuracy of 70%.
- Trained a Joint Entities and Relation Extraction Classifier using BERT Transformer with spaCy 3 and compared relation classifier using transformers and tok2vec algorithms.
- Developed end-to-end parsing and scoring pipeline that focuses on the development and deployment of NLP based services for job portals.
- Built an Email Automation system, where for email classification SVM and Xgboost models were implemented and RTE models were implemented for macros suggestion.
- Worked on a propensity modeling project for identifying potential customers for all-flash storage solutions for the next 3 to 6-month timeframe using telemetric data. Employed data science skills to clean and

simplify data, reducing over 400 features to a concise 250 features, through hypothesis testing and feature selection techniques. Conducted feature engineering to derive new features correlated with the target feature. Trained tree-based models like Random Forest, Decision Trees, and XGboost and achieved an impressive 90% accuracy rate on the test data set. Presented stakeholders with feature importance scores and identified the key features influencing customer propensity.

- Designed and implemented an advanced ETL pipeline for Client controllers, seamlessly integrating asynchronous programming for efficiency and optimizing MongoDB pipelines. Achieved an exceptional 5-6 minute workflow, surpassing the 30-minute target, emphasizing our commitment to project responsibility and performance optimization.
- Faced with the challenge of retrieving data from 1M records, each requiring a 1-second API call, and an additional challenge of hourly token updates. A data engineering pipeline was designed, boosting efficiency by 15 times and reducing the retrieval time for 1 million records from 11-12 days to just 16 hours. The utilization of parallel processing played a key role, showcasing expertise in overcoming complex data retrieval challenges.
- Worked as an active member in a group of 25 people to create relevant content, project, and speaker for 3 different workshops on AI Education & Training 3 three different educational institutions both physically and virtually.
- Worked on content creation for a Monthly Newsletter about "The Art of Programming", "Tech Trivia", "Tech Quiz", "This Month in Tech", "Most-known Tools", and "Machine Learning Series" with the motive to AI Support Community.

Graduate Teaching Assistant, Herald College Kathmandu Jan 2021 - Aug 2021

- Taught an AI course toBSc (Hons) Computer Science students and supervised their final-year projects.
  Topics Covered: Python, Data Science, Machine Learning, Deep Learning, Computer Vision, Natural Language Processing

Founder & Content Writer- Graspcoding and Tech Tutor Jul 2021 –Present

- Manage a blog named **graspcoding.com** where I provide the latest code of Python and Artificial Intelligence along with quizzes and have led to over 30,000 traffic in a month
- Revamped a business page on Instagram named **@_tech_tutor** that has led to over 5000+ followers (up by 20% in 2 months) and where I get a number of queries regarding jobs, professional education, technology, internet, and projects regarding artificial intelligence on my email and other social networking platforms, where I with utmost interest shares my perspective.

PROJECT EXPERIENCE

Nepali News Classification and Summarization

- Trained and tested on a large corpus of Nepali language data, including news articles from various sources.
- Able to identify and classify different types of Nepali news, such as political, social, economic, and entertainment news.
- Capable of providing summaries of news articles in the Nepali language, accurately conveying the main points and key details of the article.

Accident Severity Prediction

- Worked on a project to predict the severity of accidents using data analysis and machine learning techniques.
- Conducted exploratory data analysis to gain insights into the accident data.
- Built a predictive model using scikit-learn that accurately predicted the severity of accidents based on various features.

Retrieval Based Chatbot

- Implemented machine learning and deep learning techniques to assist businesses in improving customer interactions and experiences.
- Prioritized accuracy, customization, reliability, and scalability to ensure the best possible solutions for each business need.
- Collaborated with business teams to understand their unique needs and developed customized solutions tailored to their specific requirements.
- Built and optimized models using various machine learning and deep learning techniques, such as neural networks and decision trees, to improve business outcomes.

SELF LEARNING

- Broadened NLP capabilities through independent study and hands-on exploration, including contributing to my own LLM-PowerHouse repository on GitHub. Honed skills in key NLP areas like custom training and inferencing on single and multiple GPU, comprehensive analysis and optimization of both individual and ensemble encoder architectures, in-depth understanding of encoder-decoder dynamics, and evaluating LLM applications across diverse domains. Quantified this expertise by successfully training X LLMs for diverse tasks and achieving 20% - 25% improvement in model accuracy. [**Link**]
- Expanded machine learning knowledge through independent practice, including building my own repository. Developed practical skills in implementing core algorithms from scratch, gaining a deeper understanding of their mechanisms and potential applications. [**Link**]

PUBLICATION

- **Ultimate Guide to Python Basics** [Published on 23rd May 2020, Amazon Kindle Edition]
- **Implementation of Machine Learning Algorithm From Scratch** [**Under Review**]
- SMS Spam Detection using Relevance Vector Machine [**Paper Accepted**]
- Application of Nepali Large Language Models to Improve Sentiment Analysis [**Paper Accepted**]
- A Comparative Study of Transfer Learning Approaches for Strengthening Face Antispoofing Security [**In Progress**]

SKILLS

Languages	Python
Operating Systems	Windows, Linux, Slurm
ML/ DL Tools	Keras (TensorFlow), PyTorch, Scikit-learn, Numpy, Pandas, PySpark
CV Tools	Yolo, OpenCV, MediaPipe, OpenVino
NLP Tools	NLTK, Spacy, Gensim, OpenNLP, CoreNLP, LSTM, BERT, FastText, Haystack, FAISS, LangChain, Transformers, HuggingFace, GPT, Milvus, ChromaDB
Visualization	Matplotlib, Seaborn, Plotly, Tensorflow Embedding Project
Development Tools	Sublime, PyCharm, VSCode
MLOps	Docker, Kubernetes, AWS Lambda, EMR, EC2, ECR, S3, SQS, Jenkins
Web Tools	Django, Flask, FastAPI, Streamlit, Gradio
Databases	SQL/NoSQL
Version Control	Git (GitHub, GitLab, Bitbucket)
Document Skills	Word, Powerpoint, Xcel, Latex
Soft Skills	Team Player, Presentation Skills, Communication Skills, Leadership

TRAININGS

- Machine Learning, Andrew Ng - YouTube Series
- Applied Data Science with Python - Coursera
- Linear Algebra by Gilbert Strang - MIT Lecture Series, Ongoing
- Machine Learning & Data Science A-Z – Udemy
- Master Python with NumPy For Data Science & Machine Learning – Udemy
- AI Fellowship - FuseMachines

Awards and Recognitions

- Awarded Best Team at Fusemachines Hackathon [April, 2023] (2 days) out of 12 teams for fine tuning a medical generative model on a custom dataset. Recognized for advanced NLP skills, innovation in healthcare technology, effective collaboration, and earning a Rs. 10,000 cash prize.

EDUCATION

Herald College Kathmandu	B.Sc. (Hons.) Computer Science	Dec 2017 – July 2020