KAMAL SHRESTHA

Machine Learning Engineer

kamalandshrestha@gmail.com | linkedin/shresthakamal | github/shresthakamal
Bangalore, India | (+91) 7893887563 | shresthakamal.com.np

SUMMARY

- **Proficient in End-to-End Machine Learning and Deep Learning Pipelines:** Extensive experience across the full ML lifecycle, from data generation and preprocessing to system design, model training, evaluation, deployment and maintenance, with a proven record in deploying robust, scalable solutions that drive tangible outcomes.
- **Specialist in Applied NLP with Expertise in Transformer Architectures:** Skilled in solving diverse NLP challenges, from synthetic data generation, fine-tuning novel architectures for classification and generation to deploying advanced LLMs based RAGs and agentic collaborations to address high impact business needs.
- **Proven Leader in Collaborative Projects and Technical Communication:** Recognized for strong leadership in team collaboration, technical documentation, and impactful presentations, supported by industry experience, academic research, and a track record of bridging technical expertise with clear, strategic communication.
- **Professional Career/Research Interests:** Intersection of applied NLP, DL, and ML Techniques for business applications

WORK EXPERIENCE

Research & Technology Center, BOSCH Global Software Technologies (BGSW) — Bengaluru, India
Machine Learning Engineer — August 2023 – Present

- Currently leading two high-impact projects utilizing custom **fine-tuned LLMs on 2TB+ of unstructured enterprise data** across various formats to enable agentic collaboration and retrieval augmented generation (RAG) for streamlined workflows and user engagement, **enhancing operational efficiency for 500+ users** (across multiple businesses) to optimize KPIs like query resolution rate, response accuracy
- **Received the Bravo Award X3 for excellent rigor and engineering skills** in successful completion of multiple Generative AI use cases.
- Developed and implemented **a novel deep learning model for classifying lengthy legal documents by relevance** within the ProCodex team, significantly improving compliance processes for various BOSCH products across departments and global jurisdictions and achieving **€10 million in annual savings by reducing manual human effort.**
- In parallel, **pursuing multiple research verticals** on advanced document processing, enhancement of core elements and architecture of LLM based RAG approaches like pre-retrieval, multimodality, agentic collaboration, Graph RAG alongside fine-tuning open-source models like Llama3 to make it compatible with understanding custom enterprise data like acronyms.

Fusemachines, Nepal — Kathmandu, Nepal
Machine Learning Engineer and Curriculum Engineer — July 2020 – December 2021

- Designed and developed Fuse Studio, **an automated video generation platform** that transforms presentations and scripts into lecture videos, reducing manual recording time and **boosting production efficiency by 75%.** Awarded top impact project in an in-house hackathon.
- Remodeled and optimized a **Question Answering and Difficulty Ranking Model** with enhanced representations, semantic ranking, and adaptive recommendations for quizzes, assignments, and exams, **dynamically adjusting quiz difficulty in real-time based** on student response accuracy and speed to assess intelligence levels.
- Worked as **a lead curriculum engineer to design, create, review, and refine materials** (including lesson plans, reading materials, slides, audio transcripts, graded assignments, hands-on implementations, and quizzes) for undergrad focused courses like DL and NLP.
- Represented the **company as industry expert in teaching CS concepts** like AI, Python programming, and data analysis to students at Q.I. Roberts Jr-Sr High School in Florida, USA, and undergraduates at Herald International College.

EDUCATION

M. Tech. in Computer Science and Engineering, CGPA: 9.06/10 — August 2021 – July 2023
Indian Institute of Technology, Hyderabad (IITH) — Hyderabad, India

Advisor: Dr. Maunendra Sankar Desarkar, NLIP Lab
Area of focus: Recommendation Systems and Hostility detection on online social media conversation threads
Relevant Courses: NLP, Information Retrieval, DL, Fundamentals of Machine Learning, Software Engineering

Bachelors in Computer Engineering, Percentage: 92.30% — August 2016 – November 2020
Kathmandu University (KU) — Dhulikhel, Kavre, Nepal

Relevant Courses: AI, ML, DSA, Algorithm and Complexity, Software Engineering, Probability and Statistics, Speech and Language Processing, C, C++, DBMS

TECHNICAL SKILLS

Programming Languages	Python, C, C++, SQL
Libraries	Pytorch/Lightning, Transformers, Microsoft Azure, Langchain/LanGraph/LlamaIndex, VectorDBs (FAISS/Chroma/Milvus), Chainlit Pandas, Numpy/SciPy, Matplotlib/Seaborn, Flask/FastAPI, MLOps (MLFlow, Docker, Hydra, DVC, DagsHub, CI/CD with Github)
Database	MySQL, MongoDB, Firebase, Elasticsearch
Management	Git, Github, JIRA, HRM Suite, Slack
Miscellaneous	Linux, Bash, Arduino, Latex

PUBLICATION

- Aditi Bagora*, **Kamal Shrestha***, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. 2022. Hostility Detection in Online Hindi-English Code-Mixed Conversations. Proceedings of 14th ACM Web Science Conference 2022 (WebSci '22). ACM, New York, NY, USA, 11 pages doi: 10.1145/3501247.3531579 *(* indicates equal contribution)*
- **Shrestha, K**. , Poudyal, P. , Karki, J. , Ranabhat, D. (2022). A Machine Learning Approach to Identify Fake News. Center for Project Management and Information Systems (PMIS) Review, 1–13. http://journal.pmis.du.ac.bd/journaldetails.php?pid=2203281648465920

PROJECTS

Inclusivity in Job Recommendation based on heuristic and learning approaches May 2022 – July 2023
M. Tech. Thesis, Patent Approved IIT, Hyderabad

- Developed a hybrid recommendation engine based on heuristics and transformer learning approaches for a personalized recommendation based on disability, skills, and preferences.
- Attained an impressive **F1 score of 0.9389** on the validation set and **65% accuracy on similar user analysis from human feedback** with minimal space usage and low latency in recommendations

Hostility Detection in Online Hindi-English Code-Mixed Conversations [Presentation], [Video] June 2022
14th ACM Web Science Conference 2022 (WebSci '22) IIT, Hyderabad

- Proposed a novel hierarchical neural network architecture to identify hostile posts/comments/replies in online Hindi-English Code-Mixed conversations as a part of HASOC 2021.
- Adapted multilingual pre-trained models like mBERT, XLMR, and MuRIL to generate contextual representations for natural abstraction and selection of the relevant context by exploiting the hierarchy of the conversations.

Federated Semi-Supervised Medical Image Classification via Inter-Client Relation Matching [Presentation 1], [Presentation 2] April, 2022
Dr C. Krishna Mohan, Visual Computing IIT Hyderabad

- Remodeled and evaluated **medical image classification with the addition of a self attention mechanism** in every convolutional block using CBAM to obtain better classification results.
- Outperformed the official implementation given a reduced dataset (only 2%) because of computational limitations
- Ranked with the **best Top 2%(A+) of the class** on the basis of two project presentations.

Secure chat communication with Openssl and Man-in-the-middle attacks [Application], [Interceptor] April, 2022
Dr. Bheemarjuna Reddy Tamma, Network Security IIT Hyderabad

- Implemented and demonstrated **a secure peer-to-peer chat application using openssl** along with how evil Trudy(user) cab intercept the chat messages to launch various attacks(Downgrade Attack by rejecting the request for TLS Encryption and MITM attack with two TLS connections at either end and Fake Certificates)

A Machine Learning Approach to Identify Fake News June, 2020
Dr. Prakash Poudyal Kathmandu University

- Focused on applying NLP sentence classification to generate contextual sentence representations passed over classical ma-chine learning classification heads to predict whether the provided sentence is fake or not with a certain degree of confidence.
- Evaluated using lexical/syntactical/grammatical/factual features based only on raw text and semantic features based on contextual representations with attentive weights.

CERTIFICATION

- Fundamentals of Deep Learning, Deep Learning Institute (DLI), *NVIDIA* April 26, 2022
- AWS Certified Machine Learning – Specialty, *Amazon AWS* August 31, 2021
- How to win Data Science Competition: Learn from Top Kagglers, *Coursera* October 1, 2020