Kathmandu University

Department of Computer Science and Engineering

Dhulikhel, Kavre

Mini-Project Proposal

on

“Building a Nepali Large Language Model (LLM)”

[Code No: COMP 488]

(For partial fulfillment of IV Year/ I Semester in Computer Engineering)

Submitted by:

Ishan Panta [34]
Bibhushan Saakha [41]
Samir Wagle [60]
Manish Shivabhakti [63]

Submitted to:

Dr. Bal Krishna Bal

Department of Computer Science and Engineering

Submission Date: 01/10/2024

Building a Nepali Large Language Model (LLM)

1. Data Set

We will use a diverse dataset comprising:

- **Nepali news articles**, sourced from publicly available news websites.
- **Books and literature** written in Nepali.
- **Wikipedia entries** in Nepali language.
- **Social media posts and comments** in Nepali.
- **Government documents and publications** available online.
- **Legal Documents** available online

A proper preprocessing will be done such as text normalization, tokenization, stop word removal, emojis removal and **Stemming/Lemmatization. If any further preprocessing is needed, we will discuss and apply those accordingly.**

2. Project Idea

This project aims to develop a Nepali Large Language Model (LLM) from scratch, using the Devanagari script, to enhance natural language understanding and generation tasks in the Nepali language. Given the increasing demand for AI tools tailored to specific languages, we aim to create a language model capable of performing tasks such as text classification, question-answering, sentiment analysis, and text summarization, all within the context of Nepali. The model will be trained on a large corpus of Nepali text, covering a range of domains, such as news, literature, social media, and government publications.

This project is being undertaken as an **educational initiative**. While limited in hardware, we plan to optimize the model architecture for efficiency using techniques such as **mixed precision training** and **gradient checkpointing**. Our goal is to make the LLM scalable and accessible for educational purposes and local language support in addition to gaining a basic understanding of the LLM training process.

3. Software and Tools

- **Programming Languages**: Python
- **Deep Learning Frameworks**: Transformers, PyTorch
- **Preprocessing Tools and Dataset**: AI4Bharat, NLTK
- **Model Training**: Custom transformer architecture (GPT-2/3-style), mixed precision training (FP16)
- **Storage and Data Management**: Pandas, Numpy, and custom scripts for dataset processing.
- **Version Control**: GitHub for collaboration and project tracking.

4. Teammates and Work Division

Role: Data Collection and Preprocessing

Assigned to: Ishan Panta, Samir Wagle

Responsibilities: Responsible for gathering, cleaning, and preprocessing the Nepali corpus. They will ensure that the data is properly tokenized and suitable for training. They will also manage data augmentation and ensure that the dataset represents various domains like news, literature, and social media.

Role: Model Training, Optimization and Evaluation

Assigned to: Bibhushan Saakha, Manish Shivabhakti

Responsibilities: In charge of designing the transformer architecture and managing the training process and evaluating the model's performance. They will handle model optimization using techniques like gradient accumulation, mixed precision training. They will also benchmark the model on various NLP tasks such as sentiment analysis, text classification, and question-answering, ensuring its effectiveness for Nepali text.

5. Conclusion

This proposal outlines the roadmap for building a Nepali LLM from scratch using local resources. The project will focus on developing an efficient model optimized for our hardware constraints, and its primary goal is to serve the educational community and beyond with localized NLP capabilities.