Explanation Document

Project Overview

This assignment developed a machine learning model for Named Entity Recognition (NER) and Parts of Speech (POS) tagging for the Bangla language.

Pre-processing Steps

- **Data Loading:** The dataset was loaded into a DataFrame, and made it three columns one word, pos, and ner for further process.
- **Tokenization:** Words were tokenized into integers using a tokenizer.
- **Padding:** Sequences were padded to ensure uniform input length.
- **Tag Mapping:** POS and NER tags were mapped to integers for model compatibility.
- **Data Splitting:** The data was split into training, validation, and test sets.

Model Architecture

- **Embedding Layer:** Transformed words into dense vectors for better semantic representation.
- **Bidirectional LSTM:** Captured context from both directions in sentences.
- **Model Compilation**: Used the Adam optimizer with sparse_categorical_crossentropy loss.

Model Evaluation

The model was evaluated based on accuracy, with a focus on handling class imbalance and data sparsity challenges.

API Integration and Deployment

- **FastAPI:** Created an inference API with endpoints for predictions and health checks.
- **Docker:** Containerized the application for consistent deployment across environments.

Challenges Faced

- **Data formatting:** After loading data, it has one column and two indexes that need to be reindexed, and the column name
- **Class Imbalance:** Addressed the skew in tag distribution.
- **Splitting:** Train, Test splitting time face shape problem, here two target features and it's unequal target. So fixed the shape because the model need to same shape.
- **Data Sparsity:** Managed unknown tokens during inference by handling out-of-vocabulary words.