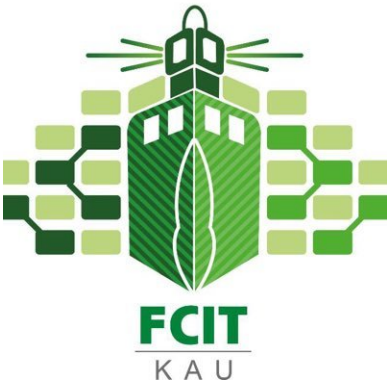

CS 681 - Deep Learning for NLP
Project Milestone

Muhammad Adnan Rizqullah (2403851)
Supervised by Dr. Fawaz Al-Salmi

1. Project Overview

This milestone report details the progress made on the Deep Learning for NLP project, which explores the effectiveness of fine-tuning pre-trained transformer models for two specific NLP tasks: text summarization and sentiment analysis. The project aims to investigate the following research questions:

1. How effectively can Microsoft's Phi-2 model be fine-tuned for abstractive text summarization?
2. How does DistilBERT's performance on sentiment analysis change after task-specific fine-tuning?

2. Current Progress

2.1 Text Summarization Task

Dataset Selection and Preparation

1. **Dataset Selection**: Initially proposed to use DialogSum dataset, but switched to CNN/DailyMail dataset (version 3.0.0) due to its broader adoption as a benchmark in summarization research
2. **Implementation**: Successfully implemented the dataset loading and preprocessing using Hugging Face's datasets library
3. **Dataset Split**: Created manageable subsets with 210 training, 45 validation, 30 development, and 15 test examples to accommodate computational constraints

Model Implementation

1. **Working Code**: Successfully implemented the Phi-2 model loading with 4-bit quantization to reduce memory requirements
2. **LoRA Implementation**: Successfully implemented Low-Rank Adaptation for efficient fine-tuning
3. **Training Setup**: Implemented the training pipeline
4. **Evaluation Implementation**: Developed evaluation infrastructure for the summarization task, including:
    a. ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L)
    b. BLEU score calculation
    c. Perplexity measurement

2.2 Sentiment Analysis Task

Dataset Selection and Preparation

1. **Dataset Selection**: Successfully selected the Stanford Sentiment Treebank (SST-2) dataset as proposed, providing a robust benchmark for sentiment analysis evaluation
2. **Implementation**: Successfully implemented data loading and preprocessing
3. **Dataset Split**: Created manageable subsets with 3500 training, 750 validation, 500 development, and 250 test examples

Model Implementation

1. **Model Loading**: Successfully implemented DistilBERT loading and configuration for sentiment classification
2. **Fine-tuning Setup**: Implemented the training configuration
3. **Baseline Implementation**: Successfully implemented SVM with TF-IDF features as a traditional ML baseline for comparison
4. **Evaluation Implementation**: Implemented comprehensive evaluation metrics including accuracy, precision, recall, and F1 score

3. Preliminary Results

3.1 Text Summarization Task

1. Have working code infrastructure for Phi-2 model loading, quantization, and fine-tuning that can be extended with additional experiments
2. Implemented evaluation pipeline that can handle both baseline and fine-tuned models

3.2 Sentiment Analysis Task

1. Have working code infrastructure for DistilBERT and SVM baseline implementation that can be extended with comprehensive evaluations
2. Successfully implemented evaluation metrics framework for consistent comparison across models

4. Conclusion

The project has made substantial progress with functioning code implementations for all key components. The dataset selection has been finalized for both research questions, with CNN/DailyMail for text summarization and SST-2 for sentiment analysis. Working code has been developed for model loading, quantization, fine-tuning, and evaluation across both tasks.