Project Proposal for Machine Learning Semester Project

Group Members:

- Omer Faisal BSDSF22M007
- Umar Saleem BSDSF22M021

Project Title

Interactive QA and Summarization Bot for Paragraph Understanding

Problem Statement

As digital content continues to grow exponentially, understanding and extracting useful information from unstructured text becomes increasingly important. Users often require quick summaries or wish to ask specific questions about long paragraphs without reading the entire content. This project aims to develop a machine learning-based bot that can read a user-provided paragraph, summarize it, and answer questions interactively. Such a system can significantly enhance user productivity in educational, research, and content analysis scenarios by automating text comprehension tasks.

Objectives

- Accept paragraph input from users via an interface.
- Use pretrained summarization models to generate concise summaries.
- Engage users in an interactive question-answering session based on the input paragraph.
- Leverage state-of-the-art NLP models to ensure high-quality QA and summarization.
- Provide a unified, user-friendly interface for the entire process.

Proposed Methodology

- **Input Handling:**

Receive paragraph input from the user via a form or text box. Clean and validate the text before processing.

- **Text Summarization:**

Use facebook/bart-large-cnn to summarize the input paragraph. Summarization output will be presented as a concise version of the input for quick understanding.

• Question-Answering Session:

Employ distilbert-base-uncased-distilled-squad to handle user questions interactively. The model will take both the original paragraph and user queries to generate precise answers.

• User Interaction:

Display summary and provide a question input box for user-driven QA. Maintain session-based interaction to allow multiple questions per paragraph.

• Evaluation:

Manual testing of QA relevance and summarization accuracy. Feedback-based iteration to improve user experience.

Dataset Description

This project does not require training on custom datasets as it utilizes pretrained models. However, evaluation will be performed using benchmark datasets such as:

- SQuAD (Stanford Question Answering Dataset) for QA benchmarking
- CNN/DailyMail summarization dataset samples for summarization testing

Custom user-generated paragraphs will also be used for real-world testing.

Expected Outcomes

- A fully functional web-based or terminal-based tool that can summarize any paragraph and engage in a QA session based on it.
- Demonstration of how pretrained transformer models can be integrated into practical applications.
- A base for future improvements like multilingual support or domain-specific QA systems.