Assignment 3

1. **Creating a Machine Learning Pipeline with Hugging Face**

   Your task is to design and implement a machine learning pipeline using Hugging Face's libraries to perform sentiment analysis on a text dataset. Use the `transformers` library to fine-tune a pre-trained BERT model (`bert-base-uncas` on the IMDb dataset available through Hugging Face's `datasets` library. Your pipeline should include the following steps:

   i. Load the IMDb dataset using the `datasets` library.

   ii. Preprocess the dataset, including tokenization using the appropriate tokenizer for `bert-base-uncased`.

   iii. Fine-tune the BERT model for sentiment analysis (binary classification: positive or negative).

   iv. Evaluate the model's performance using accuracy and F1-score metrics.

   v. Save the fine-tuned model and demonstrate how to load it for inference on a sample text input.

   Provide the complete Python code for the pipeline, including necessary imports and comments explaining each step. Additionally, include a brief written explanation (150–200 words) describing the pipeline, its components, and the rationale behind your design choices. Discuss any challenges you anticipate (e.g., computational requirements, data preprocessing) and how you would address them. Submit your code as a Python script and the explanation as a separate section in your report.

   - Submit your Python code as a `.py` file named `sentiment_pipeline.py`.
   - Include the written explanation in your report, formatted clearly with proper headings.
   - Ensure your code is executable and includes error handling where appropriate.
   - Provide a link to the fine-tuned model if uploaded to the Hugging Face Model Hub.

Resources

- Hugging Face Datasets: https://huggingface.co/docs/datasets/
- Hugging Face Transformers: https://huggingface.co/docs/transformers/
- IMDb Dataset: https://huggingface.co/datasets/imdb