Assignment 3 Report - Sentiment Analysis Pipeline

Written Explanation

This machine learning pipeline performs sentiment analysis using Hugging Face's Transformers and Datasets libraries.
The IMDb dataset is loaded using the `datasets` library and preprocessed with the `bert-base-uncased` tokenizer.
The BERT model is fine-tuned on this binary classification task (positive/negative sentiment). Fine-tuning involves
training the model's classification head while retaining the pre-trained transformer layers.

After preprocessing, the dataset is split into training and test sets. Tokenized text inputs are wrapped in a DataLoader,
and training is performed using the Trainer API, which simplifies fine-tuning and evaluation. Accuracy and F1-score
are chosen as evaluation metrics to ensure both correctness and class balance sensitivity. Once trained, the model is
saved locally and can be reloaded for inference on new texts using a simple prediction function.

Challenges include the need for a GPU for efficient training, handling long texts during tokenization (with truncation),
and managing class imbalance. These are addressed by using batch processing, evaluation checkpoints, and leveraging
Hugging Face's robust utilities for tokenization and model training.

Overall, this design ensures modularity, reusability, and reproducibility of the sentiment analysis workflow.