Sentiment Analysis Pipeline: Written Explanation

This pipeline leverages Hugging Face's transformers and datasets libraries to perform sentiment analysis on the IMDb movie review dataset. The process begins by loading the IMDb dataset, which contains labeled movie reviews. Text data is preprocessed using the BERT tokenizer (bert-base-uncased), converting raw text into input IDs and attention masks with consistent length through padding and truncation. The core of the pipeline is a pre-trained BERT model, fine-tuned for binary classification to distinguish between positive and negative sentiments.

Model training is conducted using batch processing and the AdamW optimizer, with performance evaluated using accuracy and F1-score metrics. These metrics provide a balanced assessment of the model's predictive quality, especially important for binary classification tasks. After training, the fine-tuned model and tokenizer are saved for future use, and an inference function is provided to predict sentiment on new text samples.

BERT is chosen for its strong performance in text classification and transfer learning. The pipeline is designed for clarity, reproducibility, and ease of use. Anticipated challenges include high computational requirements and the need for careful data preprocessing, both addressed through batch processing, GPU support, and robust tokenization. This approach ensures reliable, scalable sentiment analysis suitable for real-world applications.