Sentiment Analysis Pipeline Using Hugging Face

This sentiment analysis pipeline leverages Hugging Face's transformers and datasets libraries to fine-tune a pre-trained BERT model on the IMDb movie review dataset. The task is binary classification : predicting whether a review is positive or negative.

The pipeline begins by loading the IMDb dataset and applying the bert-base-uncased tokenizer to preprocess the text, including padding and truncation. A pre-trained BertForSequenceClassification model is fine-tuned using the Trainer API. The training configuration - number of epochs, batch size, evaluation strategy, logging interval and output settings - is defined through TrainingArguments. We use GPU acceleration where available, which significantly reduces training time.

To ensure quick iteration, a smaller subset of 2000 training and 500 test samples is used. The model is evaluated using accuracy and F1-score to assess performance across both classes. A progress bar and logging provide real-time feedback on training steps and estimated time remaining.

The trained model and tokenizer are saved locally and can be reloaded for inference. A simple predict() function demonstrates how the model can classify new input text. This modular pipeline is efficient, extensible, and well-suited for production-scale sentiment analysis tasks.