Sentiment Analysis Pipeline Using Hugging Face Transformers

Overview

This project implements a complete sentiment analysis pipeline using the Hugging Face transformers and datasets libraries. The pipeline fine-tunes the pre-trained bert-base-uncased model on the IMDb movie review dataset to perform binary classification — predicting whether a review is **positive** or **negative**.

Pipeline Components

- **Dataset Loading**: The IMDb dataset is loaded using datasets.load_dataset("imdb").
- **Tokenization**: The BERT tokenizer (bert-base-uncased) tokenizes and pads/truncates text inputs for model compatibility.
- **Model Fine-Tuning**: We use Trainer from Hugging Face to fine-tune BERT using a training and validation split.
- **Evaluation**: Accuracy and F1-score are calculated to assess model performance.
- **Inference**: A function is provided to load the fine-tuned model and make predictions on custom text inputs.

Design Rationale

BERT was chosen due to its strong performance on NLP tasks and availability of robust tools within the Hugging Face ecosystem. The Trainer API simplifies the training loop, and the modular design allows easy extension.

Challenges & Solutions

- **Memory Constraints**: We use batching and truncation to avoid memory overload.
- **Tokenization Errors**: Exception handling ensures robust token processing during inference.