Sentiment Analysis Pipeline Using Hugging Face

In this assignment, I built a sentiment analysis pipeline using Hugging Face's transformers and datasets libraries. I used the IMDb dataset, which contains movie reviews labeled as positive or negative. The model I used was a pre-trained BERT (bert-base-uncased), which I fine-tuned to classify the reviews.

The pipeline includes loading the dataset, tokenizing the text using BERT's tokenizer, training the model, and then evaluating it using accuracy and F1 score. I trained the model for 4 epochs on the full training dataset and evaluated it on a smaller part of the test data. The accuracy was high (above 93%), but the F1 score was around 0.49, likely because the model predicted one class more often than the other.

One challenge was the training time– it took over 2 hours even with a GPU on Google Colab. Another issue was class imbalance in predictions, which I tried to address by using a better scoring method (macro F1). Finally, I added a function to test the model on new input text, which works well for quick predictions.