Assignment 5: Fit a Deep Learning Model to a HuggingFace Dataset

Keith Tran - 300240028

The dataset chosen for this assignment is the “imdb” dataset from HuggingFace. This dataset contains 50,000 movie reviews labeled as either positive or negative. It was selected for its simplicity and easily lends itself to binary text classification, a fitting task for deep learning models. Since it is a large dataset, 5,000 reviews were sampled for training and 1,000 for testing.

Preprocessing involved tokenizing the text (with a maximum vocabulary of 10,000 words), converting each review into a sequence, and then padding/truncating sequences to 200 tokens. The training data subset was split into 80% training and 20% validation splits.

A Sequential model – a linear stack of layers where data flows sequentially from one layer to the next – was used and comprised of:

- An embedding layer: maps 10,000 words to 128-dimensional vectors; captures semantic relationships
- A Long Short-Term Memory (LSTM) layer: processes sequential data and captures long-term text dependencies
- Dropout layers: two dropout layers that set half of the outputs of the previous layer to zero to prevent overfitting and force the model to generalize better
- Dense layers: a fully connected layer that applies the rectified linear unit activation function to introduce nonlinearity, prevent overfitting, and avoid the vanishing gradient problem

Below is the classification report on the model’s performance on the test set of 1000 samples. The F1-score is fairly balanced for both positive and negative, indicating a lack of bias for one class over the other. The accuracy achieved is 0.80.

	Precision	Recall	F1-score	Support
Negative	0.80	0.81	0.81	512
Positive	0.80	0.78	0.79	488
Accuracy			0.80	1000
Macro average	0.80	0.80	0.80	1000
Weight Average	0.80	0.80	0.80	1000

The training loss curve decreased steadily over time until hitting a plateau at epoch 4. However, the validation loss curve initially decreased, but started to increase after epoch 1, indicating overfitting; the model increasingly relied on memorization of training data rather than generalization. This could be improved by increased regularization or making changes to the data. The accuracy for training hovered around 1.0, while the accuracy for validation remained much lower, at 0.80.