Task 4: Sentiment Classification of a Movie Review Dataset

Hyper-parameters

BERT Model	Validation Split	Epochs	Batch Size
BERT-Tiny (L-2_H-128_A-2)	0.2	6	12
BERT-Mini (L-4_H-256_A-4)	0.2	5	12
BERT-Small (L-4_H-512_A-8)	0.2	5	12
BERT-Medium (L-8_H-512_A-8)	0.2	10	12

These hyper parameters were chosen for standardisation and efficiency. The epoch values were chosen because it was at that epoch that the model returned diminishing increases in accuracy.

Results

BERT Model	Train Accuracy (%)	Test Accuracy (%)	Precision	Recall	F1-Score
BERT-Tiny (L-2_H-128_A-2)	80.55	79.64	0.8100	0.7977	0.6663
BERT-Mini (L-4_H-256_A-4)	82.27	81.84	0.8241	0.8203	0.7038
BERT-Small (L-4_H-512_A-8)	86.08	84.44	0.8621	0.8587	0.7593
BERT-Medium (L-8_H-512_A-8)	93.95	93.54	0.9413	0.9373	0.8390

Summary

BERT-Medium (L-8_H-512_A-8) gave the best accuracy in both train and test scores along with the best precision, recall and F1-Score. This BERT model is bigger than that of the other models. It is well known that larger BERT models tend to be more accurate, but lack the speed of the smaller BERT models. This is because the number of parameters, or weights, increase and the number of attention heads increase. This leads to more accurate results on average.

Pre-processing

A vocab file was created from the Bert_layer variable. Tokenizing was done in lower case. A tokenizer was using the vocab file and lower case variables. In the pre-processing function *bert_encode* each review was tokenized and the input ids, input masks and input segments calculated. All this data is compiled into a singular numpy arrays for training and testing input.