SEG4300 - Assignment 5 Report - Anthony Nasr - 300170809

Dataset description:

For this assignment, I chose a news article dataset with roughly 120,000 entries in total. Here is a link to the dataset https://huggingface.co/datasets/fancyzhx/ag_news. The dataset is split into training and testing data. Each entry consists of a news article text, and its classified category. The 4 categories are mapped to integers. (World = 0, Sports = 1, Business = 2, Sci/Tech = 3). The data was tokenized with BERT's tokenizer. The text was also truncated and padded. Finally, the data was loaded into PyTorch using a DataLoader for batch processing.

Model architecture and justification:

The model is based on the pre-trained transformer model BERT. I chose BERT because of its superior contextual understanding in comparison to other traditional models. On top of the BERT embeddings, a layer was added to classify the text into the 4 mapped categories. Training of the model was done using Cross-Entropy Loss and an Adam optimizer at a learning rate of 0.0001. With 3 epochs, the model took roughly 33 minutes to train while using the CUDA compatible version of PyTorch.

Model performance in predicting Clusters:

The model achieved an accuracy of 97% on the test set. It also had high recall, precision, and F1-scores.

Category	Precision	Recall	F1-score
World (0)	0.99	0.95	0.97
Sports (1)	0.99	0.99	0.99
Business (2)	0.97	0.95	0.96
Sci/Tech (3)	0.93	0.98	0.96

Interpretation of the findings:

The best performing category was Sports (1) with the highest F1-Score. The worst performing one was Sci/Tech (3) with the lowest precision and tied lowest F1-Score. The confusion matrix provided some insights into some of the classifications that could cause errors in the model output. For example, World and Business categories showed some overlap most likely due to similar topics such as economy. Some improvements would be to fine-tune the model on more domain-specific news datasets and adjusting hyperparameters.