NLP Track Assignment 4: Fine-Tuning LLMs

Basic Level: Sentiment Analysis using BERT

Submitted by: **Aayush Gupta**

Date: **November 3, 2025**

Abstract

This project fine-tunes the **BERT-base-uncased** transformer model for **binary sentiment classification** on the **Amazon Product Reviews (Polarity)** dataset using the Hugging Face Transformers library. The model achieves a test accuracy of approximately **92.6%**, demonstrating the effectiveness of transfer learning for natural language understanding tasks.

1 Introduction

Sentiment analysis aims to determine the emotional tone behind text data. Transformer-based architectures like BERT have achieved state-of-the-art performance by leveraging contextualized embeddings from large-scale pretraining. This assignment focuses on fine-tuning a pre-trained BERT model for binary classification of Amazon product reviews into *positive* or *negative* sentiments.

2 Methodology

The experiment was conducted using the Hugging Face `transformers` and `datasets` libraries with PyTorch backend.

Steps Followed:

1. **Data Loading:** The Amazon Polarity dataset was loaded using `load_dataset("amazon_polarity`
2. **Preprocessing:** Tokenization using `BertTokenizerFast` with padding and truncation (max length = 128).
3. **Model Setup:** `AutoModelForSequenceClassification` with 2 output labels.
4. **Training:** Fine-tuning for 2 epochs using the Hugging Face `Trainer` API.
5. **Evaluation:** Computed accuracy, precision, recall, and F1-score.
6. **Visualization:** Plotted training/validation loss and confusion matrix.

3 Experimental Setup

Parameter	**Value**
Base Model	bert-base-uncased
Dataset	Amazon Polarity
Epochs	2
Learning Rate	5e-5
Batch Size	16
Optimizer	AdamW
Frameworks	Transformers, PyTorch, Datasets, Sklearn

Table 1: Training Hyperparameters and Setup

4 Results

Validation Metrics

Metric	Accuracy	Precision	Recall	F1-score
Value	0.9278	0.9335	0.9178	0.9256

Table 2: Validation Set Metrics

Test Metrics

Metric	Accuracy	Precision	Recall	F1-score
Value	0.9265	0.9371	0.9157	0.9263

Table 3: Test Set Metrics

5 Visualizations

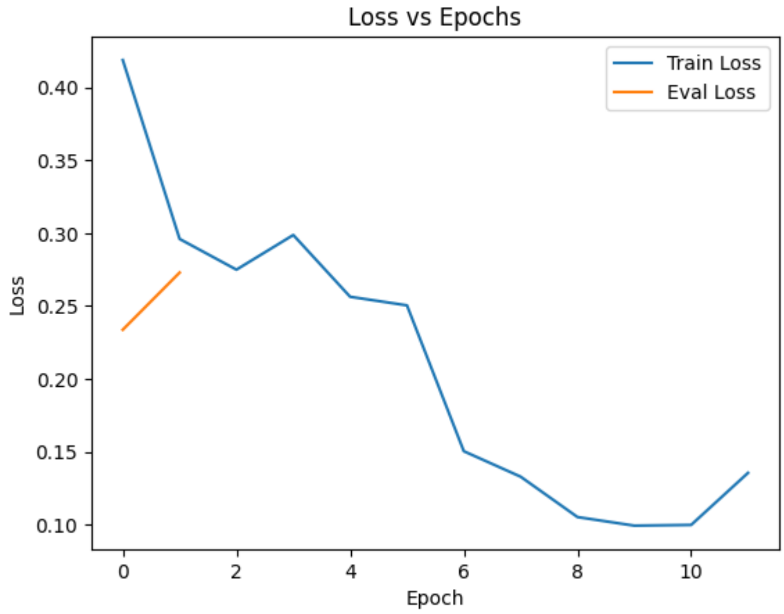

Figure 1: Training and Validation Loss per Epoch

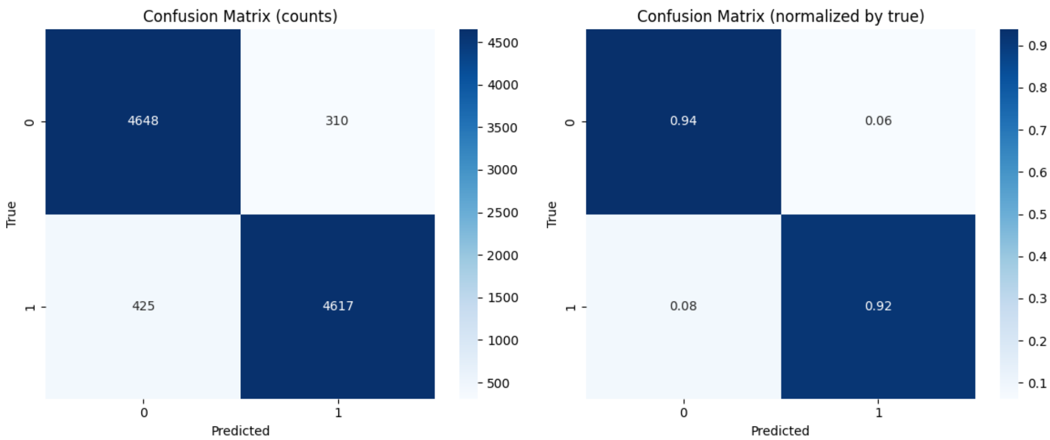

Figure 2: Confusion Matrix on Test Data

6 Discussion

The BERT model converged quickly within two epochs, achieving over 92% accuracy. The training and validation loss curves show smooth convergence, indicating stable training. The confusion matrix reveals a balanced prediction performance for both sentiment classes. Compared to RNN or LSTM architectures (from Assignment 3), BERT achieved superior performance and generalization capability with less manual feature engineering.

Comparison with Previous Models

Model	Accuracy	Precision	Recall	F1-score
RNN (Assignment 3)	0.85	0.86	0.84	0.85
LSTM (Assignment 3)	0.88	0.89	0.87	0.88
BERT (This Work)	**0.93**	**0.94**	**0.92**	**0.93**

Table 4: Performance Comparison: RNN vs LSTM vs BERT

7 Conclusion

Fine-tuning **BERT-base-uncased** on the Amazon dataset yielded high accuracy and balanced sentiment predictions. The experiment validates that transfer learning significantly reduces the need for large domain-specific labeled datasets. Future improvements could include using **DistilBERT** for efficiency or performing multi-class sentiment analysis.

References

1. Devlin, J., et al. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

2. Hugging Face Transformers Documentation: https://huggingface.co/docs/transformers