Sentiment Analysis Pipeline Report

Aman Nehra (23B1051)

June 29, 2025

1 Introduction

This assignment implements a machine learning pipeline for binary sentiment classification (positive or negative) using Hugging Face's `transformers` and `datasets` libraries. A pre-trained BERT model (`bert-base-uncased`) is fine-tuned on the IMDb dataset, which consists of movie reviews labeled as either positive or negative.

2 Pipeline Components

2.1 Data Loading

The IMDb dataset is loaded using the `datasets` library. It includes 50,000 labeled movie reviews, split evenly between training and testing subsets.

2.2 Tokenization

We use the `AutoTokenizer` for `bert-base-uncased` to tokenize the text data. Tokenization involves padding sequences to a fixed length and truncating longer texts to fit within BERT's input size limit.

2.3 Model Selection

The model used is `AutoModelForSequenceClassification` initialized with the pre-trained weights of `bert-base-uncased`. It is configured for binary classification by setting `num_labels=2`.

2.4 Training Setup

The training configuration is handled via `TrainingArguments`, specifying hyperparameters like learning rate, batch size, and number of epochs. GPU acceleration is enabled with mixed-precision training (`fp16=True`) to improve efficiency. Training and evaluation are managed using the `Trainer` API.

2.5 Evaluation

Post-training, the model is evaluated on the test dataset using accuracy and F1-score from `sklearn.metrics`. These metrics provide insights into the classifier's performance on unseen data.

2.6 Saving and Loading

The trained model and tokenizer are saved locally and reused with Hugging Face's `pipeline` API for inference. The model is also uploaded to the Hugging Face Model Hub using `huggingface_hub` for accessibility and deployment.

3 Rationale Behind Design Choices

- **BERT Model:** `bert-base-uncased` is a widely-used, powerful transformer model that captures rich contextual embeddings suitable for sentiment classification tasks.
- **Trainer API:** This simplifies model training and evaluation while ensuring integration with Hugging Face's ecosystem.
- **IMDb Dataset:** It is a balanced and clean dataset, making it ideal for binary sentiment classification benchmarking.
- **Tokenization Strategy:** Padding and truncation maintain consistent input shapes, necessary for batch processing in transformer models.

4 Anticipated Challenges and Mitigations

- **High Computational Requirements:** Fine-tuning `bert-base-uncased` requires significant GPU memory and processing time. We address this by enabling mixed-precision training (`fp16=True`) to reduce memory usage.
- **Long Sequence Truncation:** Movie reviews exceeding BERT's input length limit are truncated, potentially losing important information. Future improvements could involve hierarchical models .
- **Overfitting:** Regularization strategies like dropout and monitoring validation metrics can prevent overfitting on the training data.

5 Conclusion

This sentiment analysis pipeline demonstrates an effective and modular approach to fine-tuning BERT for binary classification tasks. Leveraging the Hugging Face ecosystem ensures efficient training, reproducibility, and easy deployment. While BERT offers strong performance, future improvements can explore scalable model variants and optimizations for deployment in resource-constrained environments.