Fine-Tuning DistilBERT for Movie Review Sentiment Analysis: A LoRA-Enhanced Approach

Author: Jinit Patel
Date: June 2025

Keywords: Sentiment Analysis, DistilBERT, LoRA, Fine-tuning, Movie Reviews, Natural Language Processing

---

Abstract

This project presents the development of a high-performance sentiment analysis model for movie reviews using fine-tuning techniques on the IMDB dataset. The model employs DistilBERT as the base architecture, enhanced with Low-Rank Adaptation (LoRA) for efficient parameter tuning. Through systematic preprocessing, training, and evaluation, the model achieved over 93% accuracy on validation data, demonstrating superior performance in binary sentiment classification. The project utilized cloud-based GPU resources (Kaggle P100) for efficient training while maintaining reproducibility through structured preprocessing pipelines.

1. Introduction

Sentiment analysis represents a critical application in natural language processing, enabling automated understanding of human emotions and opinions expressed in text. Movie review sentiment analysis, in particular, serves as a valuable benchmark for evaluating text classification models due to the rich emotional content and varied linguistic expressions found in user-generated reviews. This project addresses the challenge of accurately classifying movie reviews as positive or negative sentiment using state-of-the-art transformer architectures.

The primary objective was to develop a robust sentiment classification system capable of handling diverse review styles and lengths while maintaining high accuracy and computational efficiency. The approach leverages transfer learning principles, building upon pre-trained language models and adapting them to the specific domain of movie review sentiment analysis.

2. Tools Used

The project utilized a comprehensive technology stack optimized for deep learning and natural language processing tasks.

Core Framework and Libraries: - **PyTorch**: Primary deep learning framework for model implementation and training - **Transformers (Hugging Face)**: For pre-trained model access and fine-tuning utilities - **PEFT (Parameter Efficient Fine-Tuning)**: Implementation of LoRA adaptation techniques - **Datasets**: Efficient data loading and processing for large-scale datasets

Development Environment: - **Kaggle Notebook**: Cloud-based training environment with P100 GPU acceleration - **JupyterLab**: Local development and experimentation platform - **Hugging Face Hub**: Dataset hosting and model sharing platform

Data Processing and Evaluation: - **Pandas**: Data manipulation and preprocessing operations - **Scikit-learn**: Additional machine learning utilities and metrics - **Evaluate**: Comprehensive model evaluation metrics and benchmarking

Supporting Tools: - **IPython**: Interactive development and progress visualization - **tqdm**: Training progress monitoring and visualization - **NumPy**: Numerical computations and array operations

3. Implementation Steps

3.1 Data Acquisition and Preprocessing The project began with acquiring the IMDB movie reviews dataset containing approximately 50,000 reviews (~60MB). A comprehensive preprocessing pipeline was implemented in `preprocess.py` to ensure data quality:

- **HTML Tag Removal**: Systematic cleaning of HTML markup using regex patterns
- **Text Normalization**: Unicode character decoding and whitespace standardization
- **Quote Character Removal**: Elimination of quotation marks that could interfere with tokenization
- **Label Encoding**: Binary mapping of sentiment labels (negative: 0, positive: 1)
- **Data Splitting**: Strategic 80-20 split for training and validation sets with fixed random state for reproducibility

3.2 Model Architecture Selection The project utilized **DistilBERT-base-uncased** as the foundational model, chosen for its optimal balance between performance and computational efficiency. DistilBERT maintains 97% of BERT's performance while being 60% smaller and significantly faster, making it ideal for sentiment classification tasks.

3.3 Parameter-Efficient Fine-Tuning Implementation To optimize training efficiency and prevent overfitting, **Low-Rank Adaptation (LoRA)** was implemented with the following configuration: - **Rank (r)**: 4 - controlling the bottleneck dimension - **Alpha**: 32 - scaling factor for LoRA updates - **Dropout**: 0.01 - regularization to prevent overfitting - **Target Modules**: Query linear layers ('q_lin') for focused attention adaptation

3.4 Tokenization and Data Pipeline A robust tokenization strategy was implemented to handle variable-length reviews: - **Maximum Sequence Length**: 512 tokens with left-side truncation - **Dynamic Padding**: Batch-level padding for computational efficiency - **Special Token Handling**: Proper pad token configuration for DistilBERT compatibility

3.5 Training Configuration and Optimization The training process was carefully configured for optimal performance: - **Learning Rate**: 1e-3 with weight decay of 0.01 - **Batch Sizes**: 16 for training, 32 for evaluation - **Training Epochs**: 5 epochs with early stopping based on validation accuracy - **Mixed Precision**: FP16 training for memory efficiency and speed - **Evaluation Strategy**: Epoch-based evaluation with best model retention

4. Results and Discussion

4.1 Model Training and Validation The experimental results demonstrate the effectiveness of the proposed approach. Training was conducted on Kaggle's P100 GPU infrastructure, achieving the following performance metrics: - **Final Training Accuracy**: 93.17% - **Final Validation Accuracy**: 93.17% - **Training Loss Convergence**: From 0.293 to 0.110 over 10 epochs - **Validation Loss Stability**: Consistent performance around 0.24 - **F1 Score**: 0.932, indicating balanced precision and recall

5. Conclusion

This project successfully demonstrates the effectiveness of combining modern transformer architectures with parameter-efficient fine-tuning techniques for sentiment analysis tasks. The achieved accuracy of over 93% represents strong performance on the challenging IMDB movie review dataset, while the LoRA adaptation approach significantly reduced computational requirements compared to full fine-tuning.

Key Achievements: - Developed a production-ready sentiment analysis model with 93.17% accuracy - Implemented efficient training pipeline using LoRA for reduced computational overhead - Established reproducible preprocessing and training workflows - Successfully leveraged cloud computing resources for scalable model development