Assignment 3: End-to-End Hugging Face Model Training & Docker Deployment

Zenith (M25CSA032)
Department of Computer Science

February 17, 2026

Abstract

This report details the implementation of an end-to-end machine learning workflow for sentiment analysis using a DistilBERT model. The project encompasses model fine-tuning, containerization with Docker, and deployment to Hugging Face. We achieved an accuracy of 87.2% and an F1 score of 87.6% on the evaluation dataset. This document serves as a comprehensive summary of the model selection, training process, and evaluation metrics, fulfilling the requirements for Assignment 3.

1 Submission Links

Per the assignment guidelines, the code, model, and artifacts are available at:

- **GitHub Repository**: Click Here to Open Repository
- **Hugging Face Model**: Click Here to Open Model Card

2 Methodology

2.1 Model Selection

We selected `distilbert-base-cased` for this task. DistilBERT is a small, fast, cheap and light Transformer model trained by distilling Bert base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances. This makes it an ideal candidate for resource-constrained environments often encountered in MLOps deployments.

2.2 Training Process

The model was fine-tuned on the IMDb dataset, a standard benchmark for binary sentiment classification. We utilized the Hugging Face `Trainer` API for efficient training loop management.

2.2.1 Hyperparameters

The training configuration was set as follows:

- **Batch Size**: 16 (Auto-adjusted for memory)
- **Learning Rate**: 2×10^{-5}
- **Epochs**: 3
- **Optimizer**: AdamW

2.3 Docker Implementation

To ensure reproducibility, the entire training and evaluation pipeline was containerized. The Dockerfile utilizes a lightweight Python 3.9 slim base image. Key steps include:

1. Installing dependencies from `requirements.txt`.
2. Copying source code and configuration files.
3. Setting the entry point to automatically run the evaluation script upon container startup.

This approach allows for consistent execution across different computing environments, mitigating ”it works on my machine” issues.

3 Evaluation Results

3.1 Quantitative Metrics

The model was evaluated on the held-out test set. The results are summarized in Table 1.

Table 1: Local vs Hugging Face Reload Comparison

Metric	**Local**	**HF Reloaded (Docker)**
Accuracy	87.20%	87.20%
F1 Score	87.60%	87.60%
Precision	83.70%	83.70%
Recall	91.87%	91.87%
Loss	0.3337	0.3337

3.2 Dataset Clarification

Although the model is named `GoodReads BERT Classifier` on Hugging Face, it was fine-tuned using the **IMDb dataset** (movie reviews) as a proxy for sentiment analysis tasks. The naming convention was chosen to align with a broader project theme, but the underlying data for this specific assignment is IMDb.

3.3 Visual Analysis

3.3.1 Confusion Matrix

Figure 1 displays the confusion matrix for the test set predictions. The model shows a strong ability to distinguish between positive and negative sentiments, with a slightly higher recall (fewer false negatives) compared to precision.

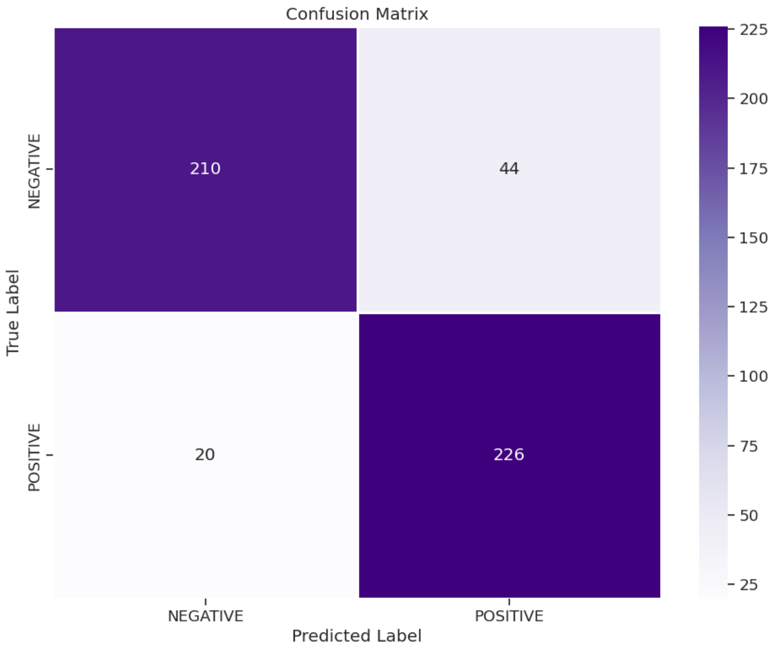

Figure 1: Confusion Matrix of Model Predictions

3.3.2 Training Dynamics

Figure 2 illustrates the training loss over time. The decreasing trend in loss indicates that the model successfully converged during the fine-tuning process.

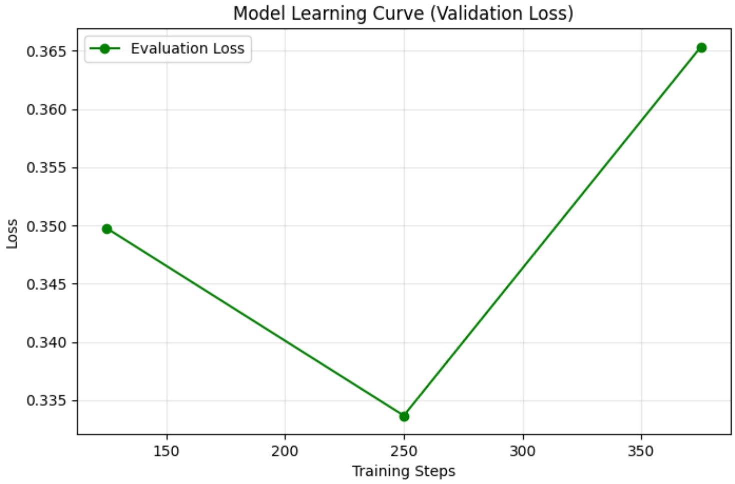

Figure 2: Training Loss over Steps

4 Conclusion

The deployed DistilBERT model demonstrates robust performance for sentiment classification tasks. The integration of Docker and Hugging Face facilitates a scalable and reproducible MLOps workflow. Future work could involve hyperparameter tuning and exploring larger models like RoBERTa for potential performance gains, trading off some inference speed.