NLP Sentiment Analysis Project

Introduction

The goal of this project is to build a sentiment analysis system for IMDB movie reviews, serving both as a practical learning experience in Natural Language Processing (NLP) and Machine Learning (ML) and as a portfolio project. The IMDB dataset contains 25,000 training and 25,000 test samples with balanced classes (positive and negative reviews). This project compares a classical machine learning baseline against a modern transformer-based approach, highlighting trade-offs in scalability and performance.

Abstract

This project applies sentiment analysis to the IMDB movie reviews dataset, comparing traditional machine learning with modern transformer-based approaches. A Logistic Regression baseline with TF–IDF features was contrasted with a fine-tuned DistilBERT model. Results show that the baseline slightly outperformed the transformer under constrained training, though transformers are expected to excel with full-scale training. This work highlights trade-offs between simplicity, performance and scalability in NLP.

Environment & Repository Structure

The repository is structured for clarity and reproducibility. Key components include:

- data/raw and processed datasets (ignored in Git)
- notebooks/exploratory data analysis, baseline modelling, transformer fine-tuning, and reporting
- src/scripts for inference and utilities
- reports/outputs including figures and final report
- requirements.txt: dependencies
- README.md: overview

Environment setup was managed using a virtual environment, with dependencies such as datasets, pandas, scikit-learn, matplotlib and transformers installed. The setup ensures reproducibility and ease of use.

Methods

Exploratory Data Analysis (EDA) included checking class balance, review length distributions and sampling example reviews. Visualisations confirmed balanced sentiment classes.

Baseline Model: A scikit-learn pipeline was used with TF–IDF vectorisation followed by

Logistic Regression. Evaluation metrics included accuracy, precision, recall and F1-score.

Transformer Model: DistilBERT was fine-tuned on a subset of 2,000 training and 1,000 validation samples for 2 epochs. Tokenisation and training were implemented with HuggingFace Transformers. Due to resource constraints, this setup was limited, but expected performance improvements with larger training budgets and hyperparameter tuning are noted.

Results

Two models were compared: Logistic Regression with TF–IDF features and DistilBERT with fine-tuning. The Logistic Regression baseline achieved slightly higher performance, though DistilBERT is expected to surpass it under full training. Evaluation metrics are shown below.

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.892	0.89	0.89	0.89
DistilBERT (subset, 2e)	0.870	~0.85	~0.87	~0.87

Discussion

Logistic Regression, despite its simplicity, performed strongly with TF–IDF features. DistilBERT, though initially weaker in this limited setup, shows strong potential and is expected to surpass the baseline when trained on the full dataset with more epochs and hyperparameter tuning.

Evaluation methods such as confusion matrices and ROC-AUC provide deeper insight into performance and can highlight areas for model improvement. These are recommended in future work.

This project highlights the value of both classical ML and modern NLP techniques: classical models offer efficiency and robustness, while transformers offer scalability and state-of-the-art accuracy.

Conclusion & Future Work

This project demonstrates the process of building an end-to-end sentiment analysis system, from data exploration and baseline modelling to advanced transformer fine-tuning. The Logistic Regression baseline provided a strong starting point, while DistilBERT showed clear potential for improvement with more resources.

Future work includes:

- Training DistilBERT on the full IMDB dataset (expecting >90% accuracy)
- Exploring larger models such as BERT-base or RoBERTa
- Adding advanced evaluation (confusion matrices, ROC-AUC)
- Implementing a robust inference script for real-time predictions