CAS2105 Homework 6: IMDB Sentiment Analysis Pipeline 🤗

Kanghan Lee (2024149005)

1 Introduction

This project explores a simple yet practical **AI pipeline** for binary sentiment analysis. The task is to classify movie reviews as either *positive* or *negative*. Sentiment analysis is a foundational NLP problem with real-world applications such as opinion mining, recommendation systems, and content moderation.

Rather than building a complex model from scratch, this project focuses on the full AI workflow: designing a naive baseline, applying a pre-trained transformer model, and comparing their performance through quantitative and qualitative evaluation.

2 Task Definition

- **Task description:** Classify IMDB movie reviews into positive or negative sentiment.
- **Motivation:** Movie reviews often contain nuanced sentiment that is difficult to capture with simple rules. This makes the task ideal for comparing heuristic methods against modern language models.
- **Input / Output:** Input is a raw text review; output is a binary label (positive or negative).
- **Success criteria:** Higher classification accuracy and F1-score on a held-out test set.

3 Methods

3.1 Naïve Baseline

The naive baseline is a **keyword-based classifier**. A small set of manually selected positive and negative keywords is used. For each review, the model counts occurrences of these keywords and predicts sentiment based on which count is higher.

- **Why naive:** The method ignores context, negation, and semantic meaning.
- **Failure modes:** Sarcasm, long reviews, implicit sentiment, and vocabulary mismatch.

3.2 AI Pipeline

The improved pipeline uses a pre-trained transformer model: **DistilBERT fine-tuned on SST-2**.

- **Model:** `distilbert-base-uncased-finetuned-sst-2-english`
- **Pipeline stages:**
  1. Text preprocessing and truncation
  2. Transformer-based sentiment inference

3. Label mapping to binary output

- **Justification:** DistilBERT offers strong performance while remaining lightweight and fast.

4 Experiments

4.1 Dataset

A subset of the **IMDB movie review dataset** from Hugging Face was used.

- **Source:** Hugging Face `datasets` library
- **Total examples:** 250
- **Train/Test split:** 200 training samples, 50 test samples
- **Preprocessing:** Lowercasing, truncation handled by the tokenizer

4.2 Metrics

Performance is evaluated using:

- **Accuracy**
- **F1-score**

4.3 Results

Method	**Accuracy**	**F1-score**
Naive Baseline	0.60	0.68
DistilBERT Pipeline	0.94	0.93

The transformer-based pipeline significantly outperforms the naive baseline. Qualitative inspection shows that DistilBERT correctly handles contextual sentiment and complex phrasing, while the baseline often misclassifies reviews with implicit or sarcastic sentiment.

5 Reflection and Limitations

This project demonstrated how powerful pre-trained models can be even without fine-tuning. The naive baseline was easy to implement but failed on most real-world language patterns. The evaluation metrics aligned well with the task, though accuracy alone may hide class imbalance effects.

A limitation of this study is the small dataset size, which may lead to variance in results. With more time, fine-tuning the model on IMDB-specific data or testing additional baselines such as TF-IDF would be valuable extensions.

References

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv preprint arXiv:1910.01108, 2019. https://arxiv.org/abs/1910.01108

[2] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. *Recursive deep models for semantic compositionality over a sentiment treebank.* In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013. https://aclanthology.org/D13-1170/