IMDb Sentiment Analysis

Transformer Architecture vs. TF-IDF + Logistic Regression

Introduction

Sentiment analysis is a fundamental NLP task with applications across domains from marketing to social media monitoring. This study compares transformer-based and traditional machine learning approaches for sentiment classification on movie reviews. Specifically, we investigate the following research question:
How do transformer-based models (DistilBERT) compare to traditional ML pipelines (TF-IDF + Logistic Regression) for IMDb sentiment analysis in terms of performance and computational efficiency?

Methods

Data Processing

We used the IMDb movie reviews dataset with binary sentiment labels (positive/negative) from HuggingFace's datasets library.. The data processing pipeline included:

1. Loading the dataset directly using the `load_dataset()` function
2. Standardizing column names
3. Creating a balanced dataset of 10,000 reviews through stratified sampling (5,000 positive, 5,000 negative)
4. Implementing a 90/10 train/validation split while preserving class balance
5. Using a separate test set of 2,000 reviews with equal class distribution

Model Implementation

Baseline: TF-IDF + Logistic Regression

Our baseline model combined TF-IDF vectorization with logistic regression:

1. **Vectorization**: n-gram range of 1-3 (unigrams, bigrams, trigrams)
2. **Hyperparameter Optimization**: Bayesian search with 20 iterations using 3-fold cross-validation
3. **Search Space**:
    a. max_features: 5,000-20,000 features
    b. min_df: 2-15 minimum document frequency

    c. max_df: 0.6-0.95 maximum document frequency
    d. sublinear_tf: True/False for log scaling
    e. C: 0.01-100 regularization parameter (log-uniform prior)
4. **Training**: Multi-threaded training with early stopping based on validation performance

Transformer: DistilBERT

The transformer approach used a fine-tuned DistilBERT model:

1. **Model Initialization**: Pre-trained "distilbert-base-uncased" with added classification head
2. **Text Processing**:
    a. Tokenization with padding and truncation to 512 tokens
    b. Conversion to PyTorch tensors with attention masks
3. **Training Configuration**:
    a. Batch size of 16
    b. Learning rate of 1e-5 with AdamW optimizer
    c. Weight decay of 0.01
    d. Linear learning rate scheduler with 10% warmup
    e. 3 epochs with early stopping (patience=3)
    f. Gradient clipping with max_grad_norm=0.8

Parameter Choice Rationale

For the transformer model, we employed reasonable parameters rather than extensive tuning due to computational constraints. Full hyperparameter optimization for transformers requires prohibitive resources, so we limited optimization to early stopping. We selected a batch size of 16 and maximum of 3 epochs as a balance between training stability and computational efficiency. Learning rate (1e-5) is within the recommended range for DistilBERT fine-tuning. The dynamic warmup schedule (10% of total steps) was implemented to stabilize early training and prevent gradient issues, while weight decay (0.01) helps control overfitting.

Evaluation and Visualization

We evaluated both models using:

1. **Performance Metrics**: Accuracy, precision, recall, and F1 score on the test set
2. **Confusion Matrices**: Normalized matrices showing class-specific performance
3. **Length Analysis**: Performance stratified by review length categories (0-100, 101-200, 201-300, 301-500, 501-1000, 1000+ words)

Results

Figure 1: Visualisations of results

Figure 1a: Performance metrics comparison between Logistic Regression and DistilBERT models.

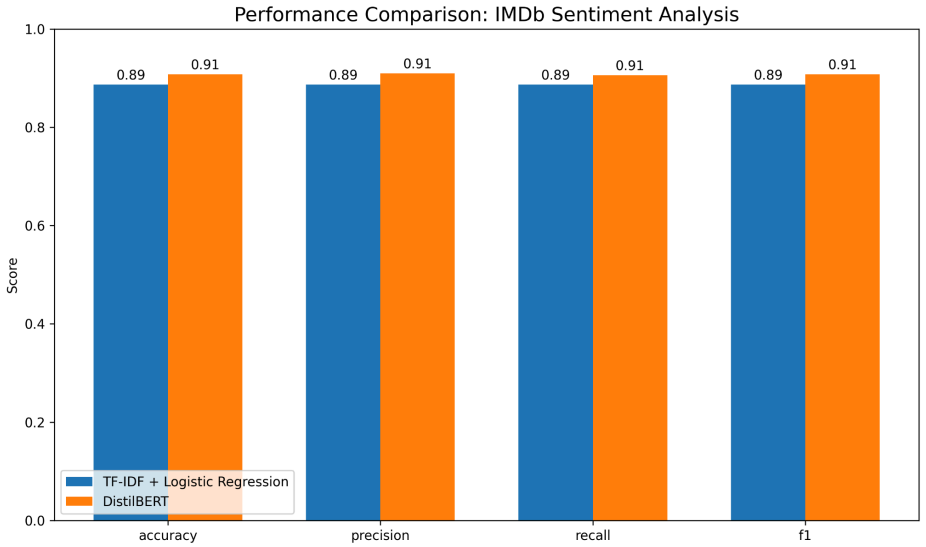


Figure 1b: Model accuracy across different review length categories

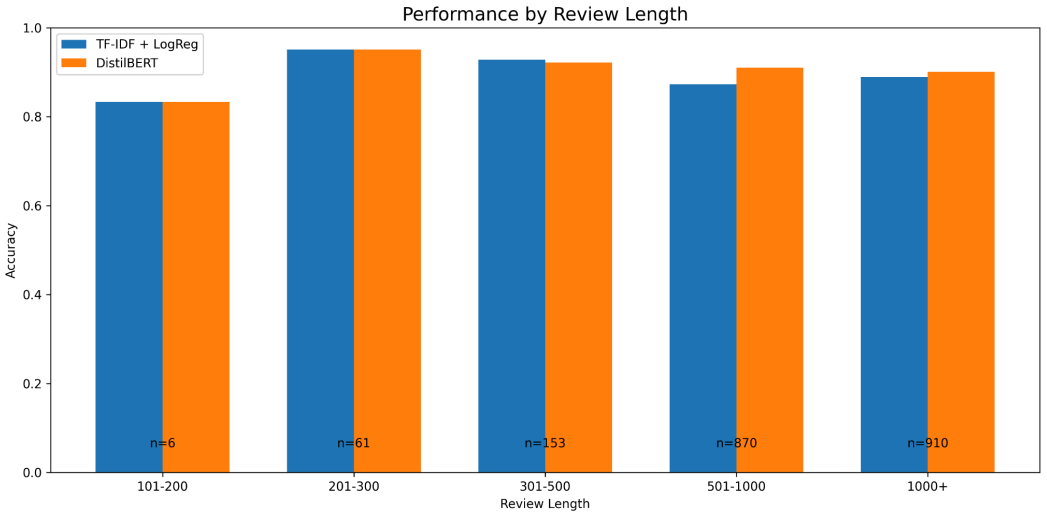
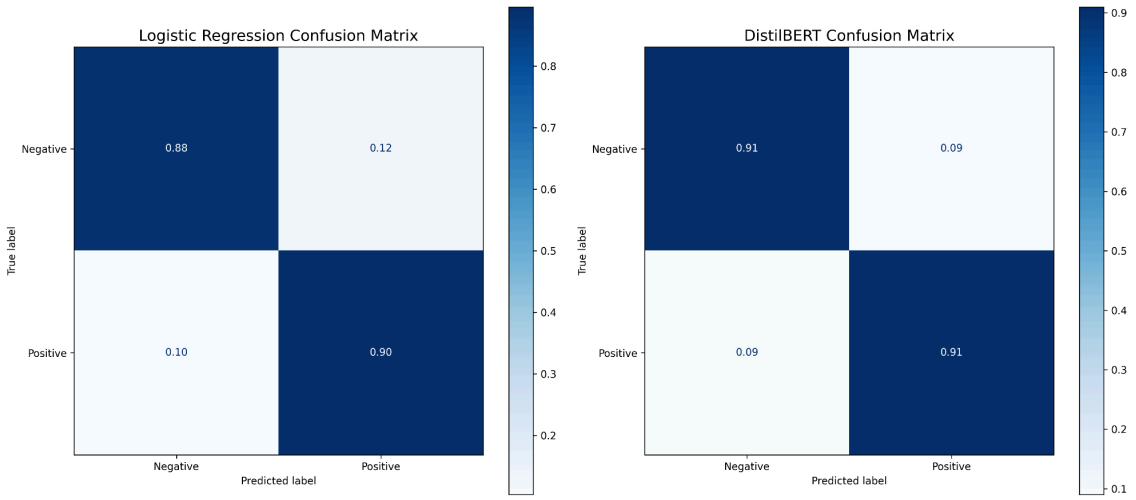


Figure 1c: Confusion matrices for Logistic Regression (left) and DistilBERT (right) models



The transformer model achieved 90.8% accuracy (precision: 0.910, recall: 0.906, F1: 0.908) compared to the logistic regression model's 88.7% accuracy (precision: 0.887, recall: 0.887, F1: 0.887), representing a 2.1 percentage point improvement. However, this performance gain came with significantly higher computational costs: 35 minutes training time for DistilBERT versus approximately 2 minutes for the logistic regression pipeline, including hyperparameter optimization. See Appendix A for full details.

Performance analysis by review length revealed both models achieved peak accuracy (>95%) on medium-length reviews (201-300 words). DistilBERT demonstrated stronger performance on longer reviews (501+ words), while the traditional model performed comparably on shorter texts. The confusion matrices showed slightly better classification across both positive and negative classes for the transformer model.

Discussion

The results suggest that transformer models provide measurable but modest performance improvements over traditional approaches.

Several observations are notable:

1. The optimized logistic regression model performs remarkably well, achieving nearly 89% accuracy with minimal computational resources
2. The transformer's 2.1% accuracy improvement represents an 18.6% reduction in error rate, which may be significant for certain applications
3. We observed training instability in the transformer, with gradient norm spikes reaching 23.5, suggesting potential benefit from more aggressive gradient clipping

The performance advantage of transformers varies by review length, with stronger benefits for longer documents where attention mechanisms likely help capture long-range dependencies better than bag-of-words approaches. This suggests application-specific considerations when choosing between these approaches.

For practical implementation, decision factors should include:

1. Available computational resources for training and inference
2. Performance requirements (is the modest accuracy gain worth the resource investment?)
3. Typical document length in the target application
4. Frequency of model retraining

Limitations

Our study used a simplified subset of the IMDb dataset and focused only on binary classification. The transformer implementation used a smaller model with fixed parameters, while the logistic regression pipeline received full hyperparameter optimization.

Future work could explore fine-grained sentiment analysis, performance on other domains, hybrid approaches, and systematic transformer hyperparameter optimization.

References

Hugging Face. (2024, March 11). *distilbert/distilbert-base-uncased*. https://huggingface.co/distilbert/distilbert-base-uncased

Hugging Face. (n.d.). *scikit-learn/imdb – IMDb dataset*. Retrieved May 4, 2025, from https://huggingface.co/datasets/scikit-learn/imdb

Appendix

Appendix A: Table of Model Results

Model	Accuracy	Precision	Recall	F1	Training Time
Logistic Regression	0.887	0.887	0.887	0.887	~2.15 minutes
DistilBERT	0.908	0.910	0.906	0.908	~35 minutes

Table 1: Model performance metrics and training time comparison