Comparative Analysis of Machine Learning vs Deep Learning for Hotel Review Sentiment Classification

Benchmarking Classical ML and Neural Network Approaches

Problem Statement

While traditional machine learning algorithms excel at text classification, deep learning approaches promise improved performance through automatic feature learning. This project conducts a rigorous empirical comparison between classical ML methods (scikit-learn) and deep neural networks (TensorFlow/Keras) for sentiment analysis, evaluating accuracy, training efficiency, and interpretability tradeoffs on the same TripAdvisor hotel review dataset.

Research Objective

Determine which modeling paradigm delivers superior performance for multi-class sentiment classification while considering practical deployment factors: training time, computational resources, model interpretability, and accuracy. Results inform architecture selection for production sentiment analysis systems.

Methodology

Dataset: 20,491 TripAdvisor hotel reviews with 5-point rating scale, identical to standard sentiment analysis benchmarks. Stratified split maintains class distribution across train/validation/test sets.

Text Preprocessing Pipeline:

- Lowercasing and tokenization
- Stop word removal using NLTK corpus
- Punctuation and special character cleaning
- Lemmatization for morphological normalization
- Sequence padding for neural network inputs (max length standardization)

Approach 1: Classical Machine Learning (scikit-learn)

- Feature extraction: TF-IDF vectorization with n-gram range (1,2)
- Models evaluated: Logistic Regression, Random Forest Classifier, Support Vector Classifier (SVC), Decision Tree
- Hyperparameter optimization via GridSearchCV with 5-fold cross-validation
- Ensemble voting classifier combining top performers

Approach 2: Deep Learning (TensorFlow/Keras)

- Word embedding layer (dimension 128) for dense text representation
- LSTM (Long Short-Term Memory) architecture with 128 hidden units

- Dropout regularization (rate 0.3) to prevent overfitting
- Dense output layer with softmax activation for multi-class classification
- Adam optimizer with learning rate scheduling
- Early stopping callback monitoring validation loss (patience 5 epochs)

Evaluation Metrics: Accuracy, precision, recall, F1-score (macro-averaged for class balance), confusion matrix analysis, training time, and inference latency.

Results

Performance Comparison:

- Classical ML (Best: Random Forest): 82% accuracy with fast training (¡ 3 minutes)
- Deep Learning (LSTM): 82% accuracy with longer training (15-20 minutes, 20 epochs)
- Both approaches achieved similar final accuracy, indicating task difficulty rather than model limitation
- Classical ML showed more stable performance across folds (lower variance)

Detailed Analysis:

- SVC achieved competitive 80% accuracy but with quadratic training time complexity
- Logistic Regression provided fast baseline at 78% accuracy
- LSTM captured sequential dependencies but required extensive hyperparameter tuning
- Confusion matrices revealed both approaches struggled with adjacent rating categories

Key Insights:

- For this dataset size (20K samples), classical ML matched deep learning performance
- TF-IDF with Random Forest offers best accuracy-efficiency trade-off
- Deep learning advantages may emerge with larger datasets (100K+ samples)
- Model interpretability favors classical ML (feature importance analysis readily available)

Production Recommendations: For sentiment analysis tasks with moderate dataset sizes and requirement for fast deployment, classical machine learning (Random Forest or Logistic Regression with TF-IDF) provides optimal balance of accuracy, training efficiency, and interpretability. Deep learning should be considered for datasets exceeding 50K samples or when transfer learning from pre-trained embeddings is viable.

This comparative study demonstrates that algorithm selection should be guided by empirical evaluation on target data, rather than assumptions about model sophistication.