IMDB Sentiment Analysis - Performance Analysis Report

Executive Summary

The Conv1D CNN achieved best performance at 86.97% accuracy, marginally outperforming Logistic Regression at 85.74%. Stopword removal consistently degraded performance across all models, with CNN showing the largest drop (0.76%). The Feed-Forward NN performed worst at 81.76%, actually underperforming simple Logistic Regression. None reached the 90% target.

Model Performance Rankings

1. Conv1D CNN: 86.97%
2. Logistic Regression (BOW): 85.74%
3. Logistic Regression (no stopwords): 85.64%
4. Conv1D CNN (no stopwords): 86.21%
5. Feed-forward NN: 81.76%
6. Feed-forward NN (no stopwords): 81.10%

Detailed Analysis

Logistic Regression (85.74%)

Strengths: Remarkably competitive with only 1.23% gap from best model. Fast training (128s), no GPU required, interpretable. Proves word frequency alone is powerful for sentiment.

Weaknesses: Ignores word order, cannot capture phrase-level sentiment or complex patterns.

Key Finding: Simple bag-of-words nearly matches deep learning, indicating strong lexical markers in movie reviews.

Feed-Forward Neural Network (81.76%)

Critical Issue: Worst performer, 4% below Logistic Regression despite greater complexity.

Root Cause: GlobalAveragePooling destroys sequential information. Averaging all word embeddings treats position-dependent patterns (e.g., "not good") identically to isolated words.

Implication: Architecture design matters more than model sophistication. Poorly structured neural networks underperform simpler alternatives.

Convolutional Neural Network (86.97%)

Strengths: Best accuracy. Three Conv1D layers capture n-gram patterns at multiple scales. MaxPooling extracts salient features. Preserves sequential context critical for phrases like "not bad" or "very good."

Architecture: 128/128/64 filters with kernel size 5, capturing 5-word phrase patterns. GlobalMaxPooling for position-invariant feature extraction.

Performance: 1.23% improvement over Logistic Regression validates sequential modeling for sentiment tasks.

Stopword Removal Impact

Model	With Stopwords	Without Stopwords	Change
Logistic Regression	85.74%	85.64%	-0.10%
Feed-forward NN	81.76%	81.10%	-0.66%
Conv1D CNN	86.97%	86.21%	-0.76%

Analysis

All models degraded with stopword removal. CNN suffered most because it learns sequential patterns that stopwords help define.

Why Stopwords Matter:

1. **Negations:** "not," "no," "never" reverse sentiment completely
2. **Intensifiers:** "very," "really," "so" modify sentiment strength
3. **Discourse markers:** "but," "however" signal sentiment transitions
4. **Context:** "if," "because," "while" provide conditional framing

Example: "The movie was not good but terrible" loses all meaning without stopwords, becoming "movie good terrible."

Contrast with Topic Modeling: Stopword removal helps topic classification (focused on content words) but harms sentiment analysis (dependent on function words).

Gap to 90% Accuracy

Likely Causes

1. Limited vocabulary (5,000 words)
2. Simple architecture (no recurrence or attention)
3. Short training (6-10 epochs)
4. No pre-trained embeddings
5. Fixed hyperparameters

Recommendations to Reach 90%

Architecture: Implement LSTM/GRU for long-range dependencies, add bidirectional processing, incorporate attention mechanisms, or fine-tune BERT.

Data: Expand vocabulary to 10,000-20,000 words, use GloVe/Word2Vec embeddings, train longer with early stopping.

Techniques: Ensemble multiple models, optimize hyperparameters (learning rate, dropout, batch size), apply data augmentation.

Key Findings

1. **Simplicity vs. Complexity:** Logistic Regression (85.74%) nearly matched CNN (86.97%), showing strong lexical signals. Poor architecture (FFNN) underperformed simple models by 4%.
2. **Sequential Structure Matters:** CNN's marginal improvement validates n-gram pattern learning. FFNN's averaging destroyed this advantage.
3. **Task-Specific Preprocessing:** Stopword removal, standard in NLP, proved harmful for sentiment analysis across all architectures.

Conclusions and Recommendations

For Production

- **Best Performance:** Conv1D CNN with stopwords (86.97%)
- **Best Efficiency:** Logistic Regression (85.74%, no GPU, 128s training)
- **Avoid:** Feed-Forward NN with averaging (81.76%, no advantages)

For Research

Explore LSTM/GRU architectures, attention mechanisms, or pre-trained transformers (BERT/RoBERTa) to exceed 90%.

Critical Lessons

1. Architecture design outweighs model complexity
2. Domain knowledge essential for preprocessing decisions
3. Sequential modeling crucial for text, but marginal gains suggest strong lexical features
4. Stopword retention non-negotiable for sentiment tasks

The 3% gap to 90% is addressable through architectural improvements (LSTM, attention) and better feature representations (pre-trained embeddings, larger vocabulary).