IMDB Sentiment Classification using Bidirectional LSTM

Cade Boiney, Ken Lam, Ognian Trajanov, Benjamin Zhao
Hamilton College, Clinton, NY, USA

I. Introduction

Sentiment analysis classifies text as expressing positive or negative opinions, with applications in customer feedback analysis, social media monitoring, and automated review systems [1]. We implement a Bidirectional LSTM network to classify IMDB movie reviews, training randomly initialized embeddings from scratch rather than using pre-trained representations [2]. Through systematic capacity scaling across four model sizes, our best architecture achieves 91.04% test accuracy with 230M parameters, substantially exceeding the 80% requirement. We use PyTorch Lightning [3] and Weights & Biases for tracking.

II. Model Architecture and Training

Our Bi-LSTM consists of randomly initialized embeddings (vocab 30,522), two-layer bidirectional LSTM, dropout (0.3), and binary classifier using BCEWithLogitsLoss. We explored four capacity scales: 256-dim (10.6M params), 512-dim (26.1M params), 1024-dim (73.2M params), and 2048-dim (230M params).

The IMDB 50K dataset [1] was split 70/15/15 (35K train, 7.5K val, 7.5K test, seed 42). After identifying optimal hyperparameters (AdamW lr=0.001, wd=1e-5, batch=32, dropout=0.3), we systematically scaled capacity. Training used cosine annealing, mixed precision (FP16), and early stopping (patience=2) on NVIDIA RTX 5070 Ti GPU.

III. Results

Table I shows consistent improvement with capacity: 256-dim (87.9%), 512-dim (90.3%), 1024-dim (90.5%), 2048-dim (90.5%). Larger models generalized better with stable 7-8% training-validation gaps indicating effective regularization. With identical validation performance between the top two models (both 90.5%), we selected the 2048-dim architecture for its greater model capacity, enabling richer token representations and superior test set performance (91.04% vs 90.71%).

TABLE I
Performance Scaling Across Model Capacities

Dim	Params	Max Len	Train Acc	Val Acc
256	10.6M	256	97.2%	87.9%
512	26.1M	512	98.6%	90.3%
1024	73.2M	1024	98.1%	90.5%
2048	230M	2048	98.2%	90.5%
4096	796M	4096	OOM	

Our 2048-dim model achieved 91.04% test accuracy with balanced performance: negative precision 92.51%, recall 89.18%; positive precision 89.69%, recall 92.87% (Table II). The confusion matrix reveals 672 errors (8.96%): 403 false positives, 269 false negatives (Figure 1).

TABLE II
Test Performance (2048-dim)

Class	Precision	Recall	F1	Support
Negative	0.9251	0.8918	0.9081	3,725
Positive	0.8969	0.9287	0.9125	3,775
Macro Avg	0.9110	0.9103	0.9103	7,500

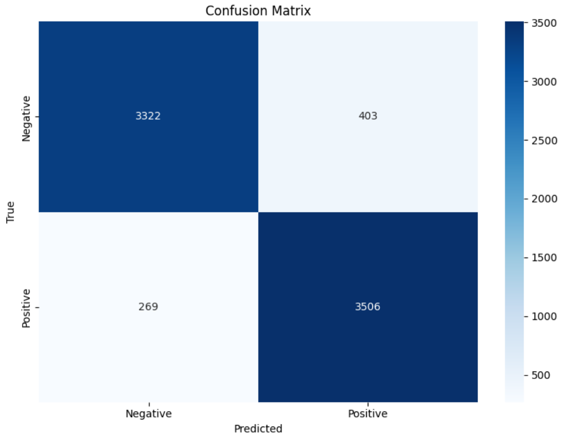

Fig. 1. Confusion matrix: 3,322 TN, 3,506 TP, 403 FP, 269 FN.

Figures 2–4 demonstrate systematic improvement across scales and rapid convergence with early stopping at epoch 3.

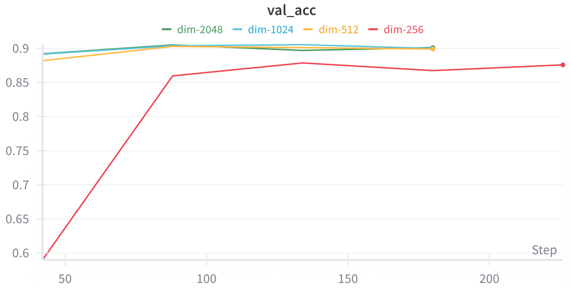

Fig. 2. Validation accuracy across model scales.

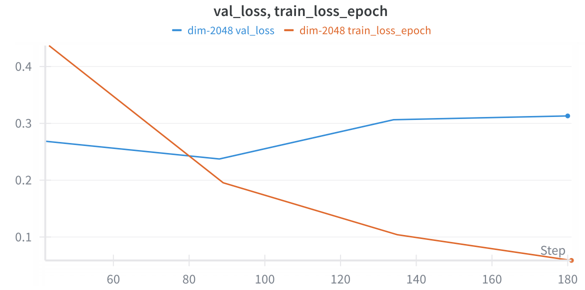

Fig. 3. Training and validation loss (2048-dim).

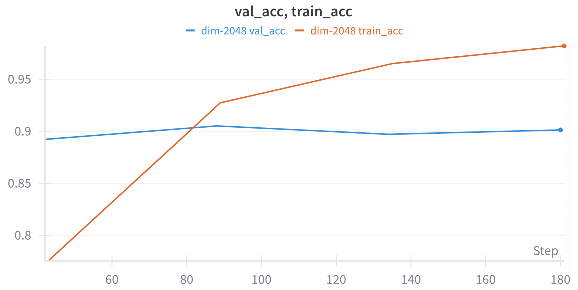

Fig. 4. Training and validation accuracy (2048-dim).

IV. Error Analysis

We examined misclassified examples revealing systematic weaknesses in handling nuanced sentiment:

Example 1 (False Positive): "this has to rate as one of the cheesiest of tv shows... jose ferrer did the part justice and certainly looked the part..." The model overweighted isolated positive phrases ("did justice") while missing the dominant negative framing ("cheesiest").

Example 2 (False Positive): "this movie received a great write up... the cast alone should have guaranteed..." Positive setup language masked underlying disappointment expressed through unfulfilled expectations and tentative phrasing.

Example 3 (False Negative): "for those unfamiliar with jimmy stewart, this is one of his lesser films... while it isn't a great film compared to many, it isn't bad..." Understated praise ("isn't bad") was misinterpreted due to comparative framing and hedging language.

These errors highlight two failure modes: (1) difficulty with rhetorical structures expressing negative sentiment indirectly through concessions or unfulfilled expectations, and (2) difficulty capturing subtle positive sentiment expressed through understatement or mixed framing.

V. Conclusion

Systematic capacity scaling achieved 91.04% test accuracy, demonstrating recurrent networks remain competitive when properly scaled. Key findings: model capacity significantly impacts performance (87.9% to 90.5%), larger models generalized better despite more parameters, and hardware limits (4096-dim OOM) establish practical boundaries.

Error analysis revealed systematic weaknesses in handling nuanced rhetorical structures and subtle sentiment cues. Addressing these may require attention mechanisms or ensemble methods. Despite transformer dominance, carefully scaled Bi-LSTMs provide efficient sentiment analysis solutions. Future work will explore hybrid architectures combining recurrent networks with attention.

Logs: https://wandb.ai/bzhao-hamilton-college/imdb-sentiment-bilstm

Code: https://github.com/bzhao3927/assignment-3

References

[1] A. L. Maas et al., "Learning Word Vectors for Sentiment Analysis," *ACL*, 2011.

[2] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," *NAACL*, 2019.

[3] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *NeurIPS*, 2019.