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 implementation and tracking.

II. Model Architecture and Training

Our Bi-LSTM consists of four components: randomly initialized embeddings (vocab 30,522), two-layer bidirectional LSTM processing sequences forward and backward, dropout (0.3) for regularization, and a binary classifier using BCEWithLogitsLoss. We explored four capacity scales proportionally varying embedding dimension, hidden dimension, and sequence length: 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 dimensions while holding other settings constant. 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 based on validation accuracy: 256-dim baseline (87.3%), 512-dim (90.0%), 1024-dim (90.0%), 2048-dim (90.1%). The 4096-dim model exceeded 16GB GPU memory. Larger models generalized better despite more parameters, with stable training-validation gaps (7-8%) indicating effective regularization. Based on validation performance, we selected the 2048-dim model for final test evaluation.

TABLE I
Performance Scaling Across Model Capacities

Dim	Params	Max Len	Train Acc	Val Acc
256	10.6M	256	91.2%	87.3%
512	26.1M	512	97.4%	90.0%
1024	73.2M	1024	98.1%	90.0%
2048	230M	2048	98.2%	90.1%
4096	796M	4096	OOM	

Our best model (2048-dim) achieved 91.04% test accuracy with 0.222 loss on the held-out test set. Table II shows balanced performance: negative precision 92.51%, recall 89.18%; positive precision 89.69%, recall 92.87%. The confusion matrix (Figure 1) reveals 672 errors (8.96%): 403 false positives, 269 false negatives.

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

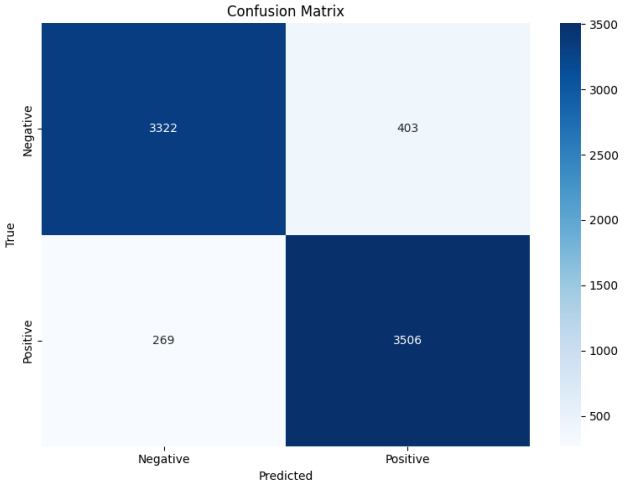

Fig. 1. Confusion matrix: 3,322 TN, 3,506 TP, 403 FP, 269 FN.

Figure 2 demonstrates consistent improvement in validation accuracy across model scales from 256-dim to 2048-dim. Figure 3 and Figure 4 show the 2048-dim model's training dynamics with rapid convergence at epoch 1 (90.1% val acc) and early stopping at epoch 3.

IV. Error Analysis

We examined three misclassified examples (all false positives). Example 1: "this has to rate as one of the cheesiest of tv shows... jose ferrer played the title character, nemo. he did the part justice and certainly looked the part..." Despite acknowledging one positive aspect, the overall rhetorical structure signals negative sentiment ("cheesiest"). The model overweighted isolated positive phrases. Example 2: "this movie received a great write up in blockbusters' coming attraction... the plot

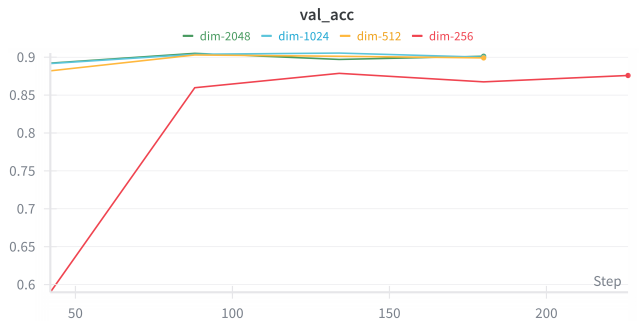

Fig. 2. Validation accuracy comparison showing systematic improvement across model scales.

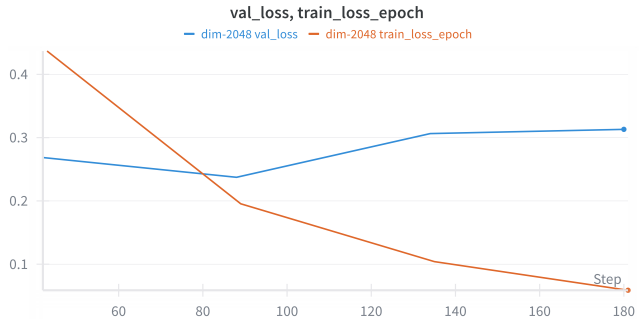

Fig. 3. Training and validation loss curves for 2048-dim model.

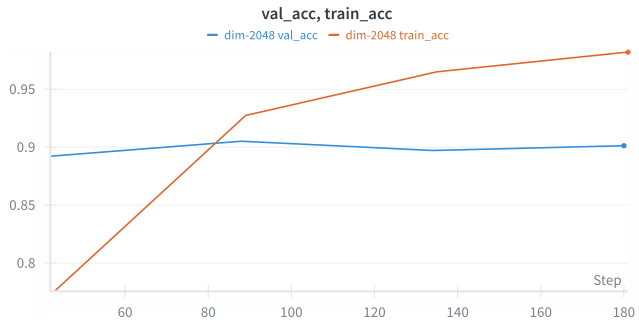

Fig. 4. Training and validation accuracy curves for 2048-dim model.

sounds reasonable, the cast alone should have guaranteed...” Positive setup masks underlying disappointment expressed through tentative language and unfulfilled expectations. Example 3: “barricade was viewed as a failure by the studio and shelved for a year before alice faye’s popularity reached such a high that the studio decided to release the film...” Factual recounting of failure confuses the model despite clear negative framing. Failure modes include difficulty recognizing rhetorical structures where negative sentiment is expressed indirectly through concessions, unfulfilled expectations, or factual descriptions of negative outcomes.

V. Conclusion

Systematic capacity scaling achieved 91.04% test accuracy, demonstrating recurrent networks remain competitive when properly scaled. Key insights: model capacity matters significantly (87.3% to 90.1% validation accuracy), larger models generalized better (challenging conventional overfitting assumptions), and hardware limits (4096-dim OOM) establish practical boundaries. The 70/15/15 split with 35K training examples effectively supported large models. Future work includes attention mechanisms, ensemble methods, and improved preprocessing. Despite transformer dominance, carefully scaled Bi-LSTMs provide efficient, interpretable sentiment analysis solutions.

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.