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. We implement a Bidirectional Long Short-Term Memory (Bi-LSTM) network to classify IMDB movie reviews as positive or negative. Unlike transfer learning approaches using pre-trained BERT, our model trains randomly initialized embeddings from scratch, learning task-specific representations directly from sentiment data.

Our architecture achieves 85.35% test accuracy with 6.3M trainable parameters, exceeding the 80% validation accuracy requirement while demonstrating efficient convergence in 7 epochs. We use PyTorch Lightning for modular implementation and Weights & Biases for experiment tracking, providing a reproducible framework for sentiment classification research.

II. Model Architecture and Training

Our Bi-LSTM model consists of four components. The embedding layer uses randomly initialized 128-dimensional vectors with vocabulary size 30,522 (BERT tokenizer). Unlike pre-trained embeddings, our approach learns sentiment-specific representations during training. The core architecture is a two-layer Bidirectional LSTM with 256 hidden units per direction, processing sequences both forward and backward to capture contextual information from both directions. This bidirectional design is critical for sentiment analysis, where negations and sentiment reversals depend on surrounding context.

Dropout regularization (rate 0.5) is applied between LSTM layers and before the final classifier to prevent overfitting. The classification head maps concatenated forward and backward hidden states (512 dimensions) to a single output logit using BCEWithLogitsLoss for numerical stability. Total model parameters: 6.3M (embedding: 3.9M, LSTM: 2.4M, classifier: 513).

Following assignment requirements, we split the IMDB 50K dataset into training (21,250 reviews, 42.5%), validation (3,750 reviews, 7.5%), and test (25,000 reviews, 50%) sets with random seed 42 for reproducibility. Preprocessing used BERT's uncased tokenizer for consistent WordPiece subword tokenization. All sequences were padded or truncated to 256 tokens with special tokens ([CLS], [SEP], [PAD]).

We optimized using AdamW (learning rate 0.001, weight decay 1e-5, batch size 32) with cosine annealing learning rate schedule (T_max=20 epochs). Mixed precision training (FP16) reduced memory usage and accelerated computation on NVIDIA RTX 5070 Ti GPU. Early stopping monitored validation loss with patience 3. Training terminated at epoch 6 when validation loss increased, restoring the best checkpoint from epoch 3 (val loss=0.351, val acc=86.03%). Total training time: approximately 3 minutes.

III. Results

A. Quantitative Performance

Our model achieved 85.35% test accuracy with 0.359 test loss, exceeding the assignment requirement. Final training accuracy reached 98.05%, indicating sufficient model capacity, though the gap to validation accuracy (86.03%) suggests moderate overfitting that dropout and early stopping controlled effectively.

TABLE I: Test Set Performance by Class

Class	Precision	Recall	F1	Support
Negative	0.8898	0.8070	0.8464	12,500
Positive	0.8235	0.9000	0.8600	12,500
Macro Avg	0.8567	0.8535	0.8532	25,000

Table I shows balanced performance across sentiment classes, with higher recall on positive reviews (90.0%) than negative (80.7%), while precision is higher for negative (88.98%) than positive (82.35%). This asymmetry suggests the model is more conservative predicting negative sentiment.

The confusion matrix (Figure 1) reveals 3,662 total errors (14.65% error rate): 2,412 false positives (negative classified as positive) and 1,250 false negatives (positive classified as negative).

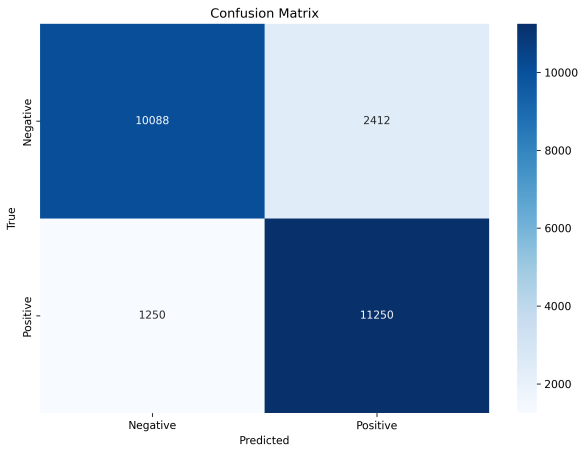
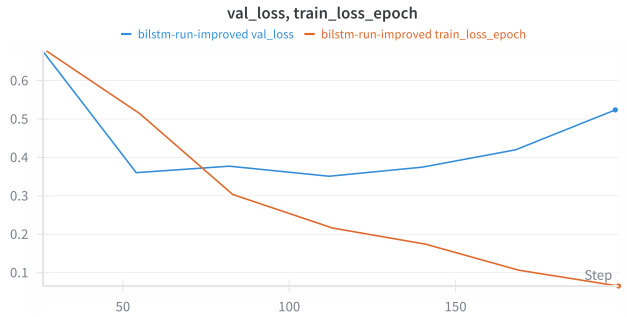


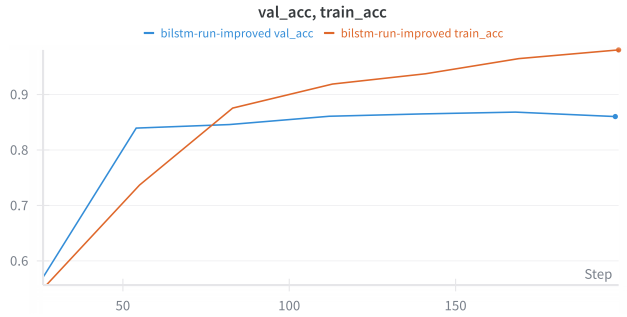
Fig. 1: Confusion matrix showing 10,088 true negatives, 11,250 true positives, 2,412 false positives, and 1,250 false negatives.

B. Training Dynamics

Figure 2 shows training and validation curves from Weights & Biases. Validation loss improved from 0.676 (epoch 0) to 0.351 (epoch 3), then increased to 0.524 (epoch 6), triggering early stopping. Validation accuracy reached 86% at epoch 3 and remained stable. Training accuracy climbed steadily to 98%, demonstrating successful learning without erratic fluctuations. Smooth convergence validates our hyperparameter choices and learning rate schedule.



(a) Training and validation loss curves



(b) Training and validation accuracy curves

Fig. 2: Training dynamics showing loss and accuracy evolution. Best performance at epoch 3 (val loss=0.351, val acc=86.03%).

W&B Project: https://wandb.ai/bzhao-hamilton-college/imdb-sentiment-bilstm

IV. Error Analysis

We examined misclassified examples to identify failure patterns. Three representative cases illustrate common errors:

Example 1 (False Positive): True=Negative, Predicted=Positive. Text: "naach a more detailed review can be obtained anywhere else in the web. this one is a good portrayal, although i do not agree with it entirely..." Analysis: Mixed sentiment with explicit positive phrase "good portrayal" overwhelmed subtle negative framing.

Example 2 (False Negative): True=Positive, Predicted=Negative. Text: "critics claim that this film is one of the worst films ever. watchers also claim the same. but there is a flip side to that coin. they are wrong, very wrong. it's the most clever film i've seen..." Analysis: Strong initial negative sentiment before reversal caused misclassification. The model weighted early negative phrases too heavily, failing to capture the sentiment shift.

Example 3 (False Negative): True=Positive, Predicted=Negative. Text: "1969 was the year. new york city was the place. putney swoope was the second robert downey film to achieve some recognition..." Analysis: Factual, descriptive language without explicit sentiment markers made classification difficult.

Three primary failure modes emerged: (1) sarcasm and irony where surface lexical features contradict pragmatic meaning, (2) genuinely mixed sentiment with balanced positive and negative aspects, and (3) implicit sentiment expressed through factual descriptions rather than explicit evaluative language.

V. Discussion and Conclusion

Our Bi-LSTM achieved 85.35% test accuracy on IMDB sentiment classification, demonstrating that recurrent architectures remain competitive when properly trained. Several design choices contributed to performance: randomly initialized embeddings learned task-specific representations, bidirectional processing captured contextual dependencies including negations, two-layer depth provided sufficient capacity without excessive parameters, and dropout with early stopping prevented severe overfitting.

The 14.65% error rate stems primarily from sarcasm, mixed sentiment, and implicit sentiment expression. The training-validation gap (98.05% vs 86.03%) indicates moderate overfitting that regularization controlled. Fixed 256-token length may truncate longer reviews, potentially discarding late-appearing sentiment information.

Future directions include attention mechanisms to focus on sentiment-bearing phrases, ensemble methods for improved robustness, incorporation of linguistic features (negation handling, sentiment lexicons), and semi-supervised learning using unlabeled data. Despite transformer dominance in NLP, our results show recurrent architectures provide efficient, interpretable, competitive solutions for sentiment analysis.

Code: https://github.com/bzhao3927/Assignment-3.
Logs: https://wandb.ai/bzhao-hamilton-college/imdb-sentiment-bilstm

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.