From Simplicity to Focus: A Comparative Study of RNN Variants with Attention in Sentiment Analysis

Sara Alsanajleh
Department of Artificial Intelligence
Jordan University of Science and Technology
sjalsanajleh22@cit.just.edu.jo

Ghada Abu Shaqra
Department of Artificial Intelligence
Jordan University of Science and Technology
gmabushaqra22@cit.just.edu.jo

***Abstract*—This Report investigates binary sentiment classification using deep learning models, the corpus was collected from three public sources: Amazon product reviews, IMDb movie reviews, and Yelp restaurant reviews. We used LSTM, BiLSTM, BiRNN and BiLSTM with Attention, leveraging pre-trained GloVe word embeddings. We present PyTorch implementations, hyperparameter tuning, evaluation metrics, and a comparative analysis of the models' performances.**

I. Task Overview

The task is binary sentiment classification on a dataset of short English sentences. The goal is to build and compare different deep learning models using PyTorch: LSTM, BiLSTM, BiRNN and (BiLSTM+Attention), with GloVe pre-trained word embeddings.

II. Dataset Description

The dataset used in this project is a balanced sentiment analysis corpus designed for binary classification. Each entry consists of a `sentence` and a `label`, where the label indicates whether the sentiment is positive (1) or negative (0). The data is divided into separate subsets for training, development, and testing.

III. Model Implementations

Four models were implemented in PyTorch:

- **LSTM:** A unidirectional LSTM using GloVe embeddings.
- **BiLSTM:** A bidirectional LSTM for richer context representation.
- **BiRNN:** A bidirectional Recurrent Neural Network that processes the input sequence in both forward and backward directions to capture contextual information from both past and future tokens.
- **BiLSTM+Attention:** A bidirectional LSTM model enhanced with an attention mechanism to assign different importance weights to tokens, allowing the model to focus on more informative parts of the input during classification.

IV. LSTM Vs BiLSTM

A. Hyperparameter Tuning Summary

The following hyperparameters were tuned for both LSTM and BiLSTM models:

- **Embedding Dimension:** {100d, 300d}
- **Hidden Layer Size:** LSTM: {64, 128, 256}, BiLSTM: {32, 64, 128}
- **Dropout Rate:** {0, 0.2, 0.3, 0.5}
- **Optimizer:** Adam, SGD
- **Learning Rate:** LSTM: {5e-4, 1e-2, 1e-1}, BiLSTM: {5e-4, 1e-4, 1e-3}
- **Batch Size:** LSTM: {16, 32, 64}, BiLSTM: {32, 64, 128}
- **Number of Epochs:** {10, 15, 20}

B. The best combination of hyperparameters

TABLE I: Best Hyperparameter Settings for LSTM and BiLSTM

Model	Embed	Hidden	Dropout	Batch	LR	Epochs	Optimizer
LSTM	100	128	0.3	16	1e-2	15	Adam
BiLSTM	300	64	0.0	32	1e-3	20	Adam

C. Evaluation Results on both the development and test sets

TABLE II: Full Evaluation Metrics for LSTM and BiLSTM Models

Model	Set	Precision (0 / 1)	Recall (0 / 1)	F1-score (0 / 1)
LSTM	Dev Set	84.42% / 70.87%	63.73% / 88.24%	72.63% / 78.60%
	Test Set	83.39% / 72.02%	66.28% / 86.80%	73.86% / 78.72%
BiLSTM	Dev Set	84.53% / 77.53%	75.00% / 86.27%	79.48% / 81.67%
	Test Set	82.15% / 74.81%	71.55% / 84.46%	76.49% / 79.34%

Model	Set	Accuracy	Macro Precision	Macro Recall	Macro F1-score
LSTM	Dev Set	75.98%	77.64%	75.98%	75.61%
	Test Set	76.54%	77.71%	76.54%	76.29%
BiLSTM	Dev Set	80.64%	81.03%	80.64%	80.58%
	Test Set	78.01%	78.48%	78.01%	77.91%

D. Comparative Analysis

Which Model Performed Better and Why: Based on the final evaluation on the test set, the BiLSTM model outperformed the standard LSTM. It achieved a higher accuracy of 78.01% compared to 76.54% for the LSTM. In addition, BiLSTM showed better macro-averaged F1-score and more balanced precision-recall across both classes. This improvement can be attributed to the bidirectional nature of BiLSTM, which allows it to capture both past and future context in a sentence—especially useful in sentiment analysis where important cues can appear at either end of the sentence.

Trade-offs Between Performance, Complexity, and Training Time: While BiLSTM delivered better performance, it also comes with increased complexity. Since it processes the input sequence in both directions, it has twice the number of parameters compared to LSTM with the same hidden size. This resulted in longer training time and more memory usage. On the other hand, the LSTM was simpler and trained faster, which could be more practical in resource-constrained settings. So, the trade-off here is clear: BiLSTM offers improved accuracy and generalization, but at the cost of higher computational overhead.

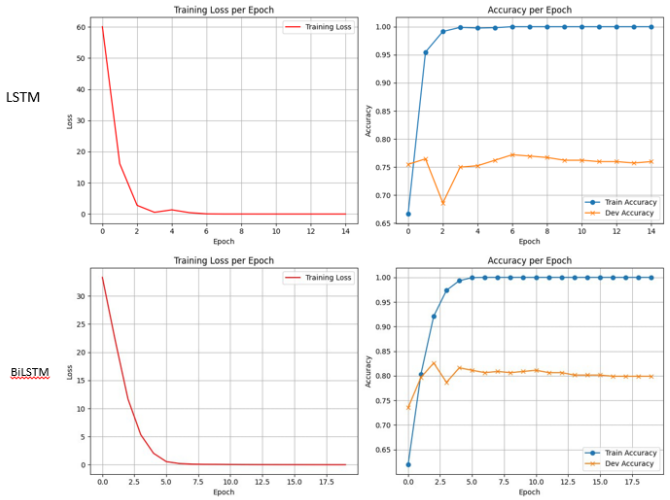

Fig. 1: Training loss and accuracy curves for LSTM and BiLSTM models on the training and development sets.

Insights from the Confusion Matrix: Looking into the confusion matrices, both models showed stronger performance in identifying positive sentiment (class 1). However, LSTM had more false negatives—meaning it often misclassified positive reviews as negative. BiLSTM reduced this issue noticeably. For instance, BiLSTM misclassified 53 positive examples versus LSTM's 45, but it significantly reduced misclassification of negative examples from 115 in LSTM to only 97. This suggests that BiLSTM achieved better class balance and was more consistent across both sentiment classes.

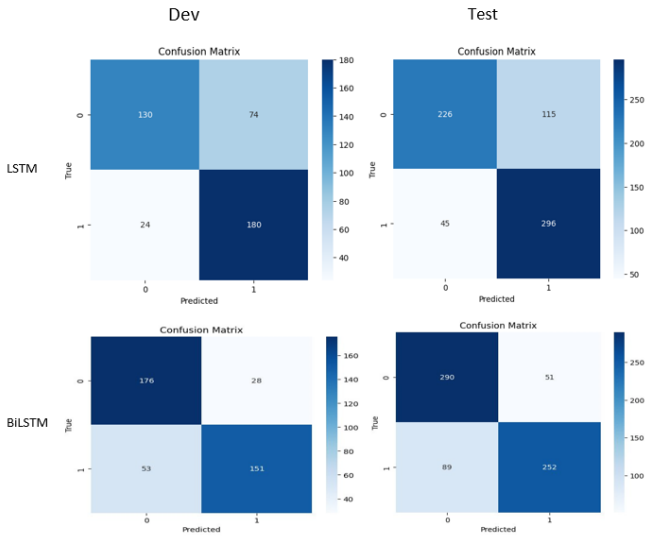

Fig. 2: Confusion matrices of LSTM and BiLSTM models on Dev and Test sets.

Notable Patterns in Per-Class Performance: A closer look at per-class metrics reveals that:

- LSTM had higher precision for class 0 (negative), but lower recall, meaning it was more conservative in predicting negatives.
- BiLSTM, on the other hand, showed more balanced precision and recall for both classes, especially improving the recall for class 0 without sacrificing much precision.

This indicates that BiLSTM is better at catching subtle positive and negative cues, leading to more stable performance.

V. BiLSTM Vs BiRNN

A. Hyperparameter Tuning Summary

The following hyperparameters were tuned for the BiRNN model:

- **Embedding Dimension:** {100d, 300d}
- **Hidden Layer Size:** {64}
- **Dropout Rate:** {0.2, 0.3, 0.5}
- **Optimizer:** Adam
- **Learning Rate:** {1e-3, 5e-4, 5e-3}
- **Batch Size:** {32}
- **Number of Epochs:** {10, 20, 30}

B. The best combination of hyperparameters

TABLE III: Best Hyperparameter Settings for BiLSTM and BiRNN

Model	Embed	Hidden	Dropout	Batch	LR	Epochs	Optimizer
BiLSTM	300	64	0.0	32	1e-3	20	Adam
BiRNN	100	64	0.5	32	5e-4	20	Adam

C. Evaluation Results on both the development and test sets

TABLE IV: Full Evaluation Metrics for BiLSTM and BiRNN Models

Model	Set	Precision (0 / 1)	Recall (0 / 1)	F1-score (0 / 1)
BiLSTM	Dev Set	84.53% / 77.53%	75.00% / 86.27%	79.48% / 81.67%
	Test Set	82.15% / 74.81%	71.55% / 84.46%	76.49% / 79.34%
BiRNN	Dev Set	84.38% / 65.71%	52.94% / 90.20%	65.06% / 76.03%
	Test Set	84.39% / 64.78%	50.73% / 90.62%	63.37% / 75.55%

Model	Set	Accuracy	Macro Precision	Macro Recall	Macro F1-score
BiLSTM	Dev Set	80.64%	81.03%	80.64%	80.58%
	Test Set	78.01%	78.48%	78.01%	77.91%
BiRNN	Dev Set	71.57%	75.04%	71.57%	70.55%
	Test Set	70.67%	74.59%	70.67%	69.46%

D. Comparative Analysis

The performance difference between BiLSTM and BiRNN: The BiLSTM model consistently outperformed the BiRNN model on both the development and test sets. On the test set, BiLSTM achieved an accuracy of 78.01% compared to 70.67% for BiRNN. Additionally, BiLSTM had higher macro-averaged precision, recall, and F1-score. For example, the macro F1-score on the test set was 77.91% for BiLSTM, while it was only 69.46% for BiRNN. This indicates that BiLSTM is better at capturing the contextual dependencies in both directions, leading to stronger overall performance and more balanced classification between the two sentiment classes.

Trade-offs in terms of accuracy, model simplicity, and training time: Although BiLSTM achieved better performance, this came at the cost of increased model complexity and training time. BiLSTM models use more parameters due to their bidirectional structure and tend to be slower during both training and inference. On the other hand, BiRNNs are relatively simpler and train faster, which makes them more suitable in scenarios where computational resources or time are limited. However, the noticeable drop in accuracy and class-level metrics makes BiRNN a less favorable option when high performance is critical.

Common limitations of vanilla RNNs: Vanilla RNNs, including BiRNN, suffer from well-known limitations. One major issue is the vanishing gradient problem, which makes it difficult for the model to learn long-term dependencies. This often results in poorer performance on tasks like sentiment analysis, where important context can appear in distant parts of the sentence. Additionally, vanilla RNNs are generally less expressive than LSTMs, as they lack the internal gating mechanisms that help LSTMs retain and control information flow through time. These limitations likely contributed to the weaker performance of the BiRNN model in this task.

Key insights from the confusion matrix and class-level metrics: The confusion matrix for BiRNN shows a significant imbalance in class-specific performance. On the test set, BiRNN misclassified a large number of negative samples as positive — 168 out of 341 — which severely affected the precision for class 1. In contrast, BiLSTM was more balanced, showing fewer misclassifications for both classes. Also, BiRNN achieved a higher recall for class 1 (positive sentiment), but at the cost of lower precision for the same class. This suggests that BiRNN tends to overpredict positives, which could be problematic in real-world scenarios where both false positives and false negatives carry consequences.

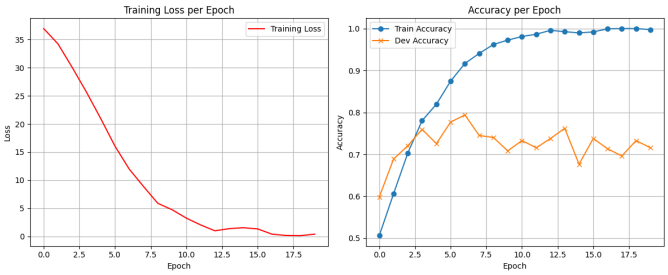

Fig. 3: Training loss and accuracy curves for BiLSTM and BiRNN models.

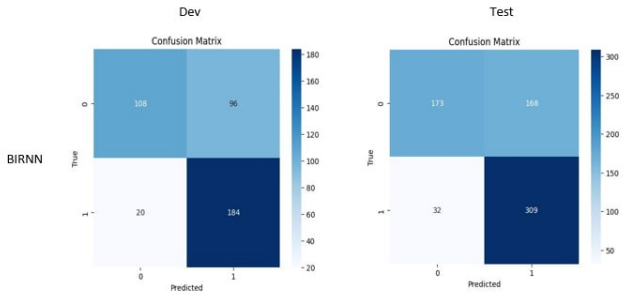

Fig. 4: Confusion matrices of BiLSTM and BiRNN models on Dev and Test sets.

VI. BiLSTM With Attention

A. Hyperparameter Tuning Summary

The following hyperparameters were tuned specifically for the BiLSTM with Attention model:

- **Learning Rate:** {5e-4, 3e-4, 5e-5, 1e-3,1e-2, 1e-4}
- **Number of Epochs:** {10, 15, 20, 32, 25, 30}

B. Evaluation Results on both the development and test sets

TABLE V: Evaluation Metrics for BiLSTM with Attention

Set	Accuracy	Precision (0 / 1)	Recall (0 / 1)	F1-score (0 / 1)	Macro Precision	Macro F1-score
Dev Set	81.37%	78.32% / 85.16%	86.76% / 75.98%	82.33% / 80.31%	81.74%	81.32%
Test Set	79.91%	75.76% / 85.66%	87.98% / 71.85%	81.41% / 78.15%	80.71%	79.78%

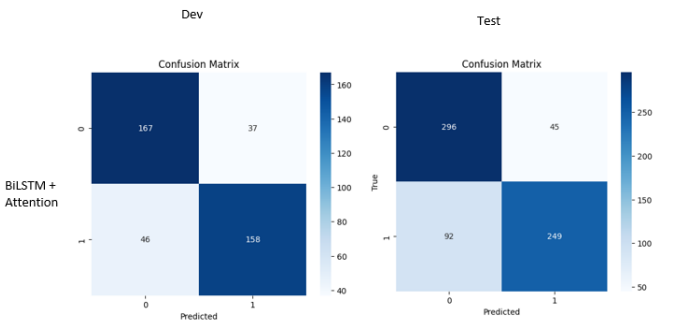

Fig. 5: Confusion matrices of BiLSTM with Attention on Dev and Test sets.

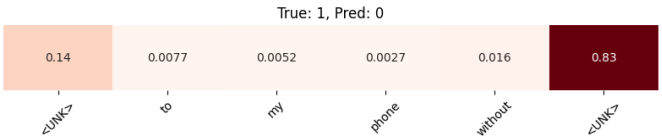

Fig. 6: Attention distribution for a misclassified example (True: 1, Pred: 0).

Fig. 7: Attention distribution for a correctly classified example (True: 1, Pred: 1).

C. Comparative Analysis

Did attention improve performance and why?: Compared to the BiLSTM baseline, the BiLSTM model with attention demonstrated modest but meaningful performance gains. On the development set, accuracy improved from 80.64% to 81.37%, and macro F1-score increased from 80.58% to 81.32%. The improvements were also evident on the test set, where accuracy increased from 78.01% to 79.91%, and macro F1-score rose from 77.91% to 79.78%. These results indicate that the attention mechanism effectively enhanced the model's ability to identify sentiment-relevant tokens, leading to more accurate and balanced predictions across sentiment classes.

Insights from the attention distribution: The attention visualizations reinforce the quantitative metrics. In the correctly classified sentence (Figure 7), the model directs attention toward the word *best*, highlighting its role in conveying strong sentiment. Meanwhile, in the misclassified example (Figure 6), attention was overly concentrated on a `<UNK>` token, causing the model to overlook more meaningful words like *without*. This showcases the potential of attention mechanisms to improve interpretability and decision-making—when attention is properly guided toward informative parts of the input, the model performs better.

VII. Final Summary and Comparative Analysis

A. Summary Table of Results

TABLE VI: Performance Summary of All Models on Development and Test Sets

Model	Set	Accuracy	F1 (Class 0)	F1 (Class 1)	Macro Precision	Macro Recall	Macro F1
LSTM	Dev	75.98%	72.63%	78.60%	77.64%	75.98%	75.61%
	Test	76.54%	73.86%	78.72%	77.71%	76.54%	76.29%
BiLSTM	Dev	80.64%	79.48%	81.67%	81.03%	80.64%	80.58%
	Test	78.01%	76.49%	79.34%	78.48%	78.01%	77.91%
BiRNN	Dev	71.57%	65.06%	76.03%	75.04%	71.57%	70.55%
	Test	70.67%	63.37%	75.55%	74.59%	70.67%	69.46%
BiLSTM + Attn	Dev	81.37%	82.33%	80.31%	81.74%	81.37%	81.32%
	Test	79.91%	81.41%	78.15%	80.71%	79.91%	79.78%

TABLE VII: Average Training Time Per Epoch for Each Model

Model	Avg. Time per Epoch
LSTM	1m 53s
BiRNN	1m 50s
BiLSTM	3m 21s
BiLSTM + Attention	5m 35s

TABLE VIII: Total Trainable Parameters for Each Model

Model	Total Trainable Parameters
LSTM	651,318
BiRNN	554,806
BiLSTM	1,787,550
BiLSTM + Attention	1,795,870

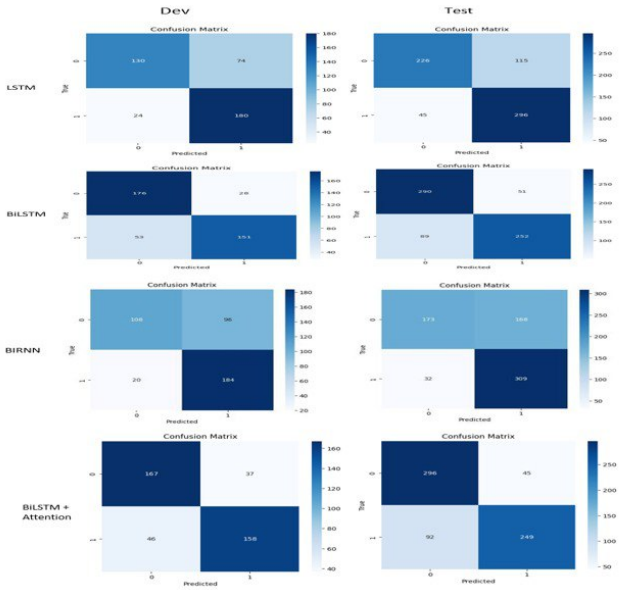

Fig. 8: Confusion Matrices of All Four Models on the Test Set (LSTM, BiLSTM, BiRNN, BiLSTM+Attention)

B. Comparative Analysis

This final comparison addresses the five required analytical points using results from all previous parts.

How did different architectures affect classification performance?: Model architecture had a clear and measurable effect on sentiment classification performance. While LSTM was fast and moderately effective, it was outperformed by more complex models in almost every metric. BiRNN, although bidirectional, did not surpass LSTM—its performance on both the dev and test sets was weaker across all macro metrics and F1 scores. In contrast, BiLSTM improved substantially upon both by leveraging gated memory, and BiLSTM with attention achieved the highest results overall.

Did bidirectionality or attention mechanisms lead to improvements?: Bidirectionality helped only when paired with a gating mechanism. BiLSTM significantly outperformed BiRNN, proving that bidirectionality alone (as in BiRNN) is not sufficient without controlled memory flow. Attention added to BiLSTM offered further gains, proving especially effective in emphasizing sentiment-rich words.

Which models were strongest or weakest overall and why?: BiLSTM with attention was the strongest model, dominating across all test metrics. LSTM, while limited, still outperformed BiRNN, which showed the weakest results. This reinforces the importance of memory mechanisms—BiRNN struggled due to the vanishing gradient problem and lack of internal control.

Was the added complexity of attention justified by the performance gain?: Yes. The increase in macro F1 from 77.91% (BiLSTM) to 79.78% (BiLSTM+Attn), along with improved interpretability, makes the additional training time (5m 35s vs 3m 21s) worthwhile. In sentiment tasks requiring high confidence, the trade-off favors complexity.

Any patterns or surprises across models and metrics?: Contrary to initial expectations, BiRNN was the weakest model, even compared to the simpler LSTM. It suffered in macro F1, precision, and recall. Meanwhile, BiLSTM with attention stood out with balanced class-wise performance and reliable accuracy, making it the most effective and explainable architecture.