Natasha Oberoi

DATA641

12 Nov 2025

Project Report (PDF)

1. **Dataset Summary:** Description of preprocessing and statistics (avg. review length, vocab size).
2. **Model Configuration:** Parameters (embedding dim, hidden size, number of layers, dropout, optimizer settings).
3. **Comparative Analysis:** Tables and charts comparing Accuracy, F1, and training time across experimental variations.
4. **Discussion:**
    - Which configuration performed best?
    - How did sequence length or optimizer affect performance?
    - How did gradient clipping impact stability?
5. **Conclusion:** Identify the **optimal configuration** under CPU constraints and justify your choice.

Dataset Summary:

The IMDb Movie Review dataset consists of 50,000 text samples divided into 25,000 training and 25,000 testing reviews. Each review is labeled as either positive or negative sentiment classes.

The reviews were lowercased, stripped of punctuation and special characters, and tokenized using the NLTK word_tokenize function. Then, a vocabulary of the top 10,000 most frequent words in the training set was built, including two special tokens: <PAD> for padding and <UNK> for unknown words. Each review was then mapped to the corresponding integer token IDs and padded or truncated to fixed sequence lengths of (25, 50, or 100 words).

The following are the summary statistics on the preprocessed reviews:

Average review length: 230.26 words
Minimum review length: 4
Maximum review length: 2469
Vocabulary size: 10000

Model Configuration:

All models were trained on a single A100 GPU in Google colab (on a limited plan). Each of the 3 types of network architectures began with an embedding layer that transformed input tokens into 100-dimensional embeddings. Two recurrent hidden layers were used, of size 64 units. A dropout rate between 0.5 was applied to reduce overfitting. A fully connected output layer with a single sigmoid-activated neuron produced the final binary probability for the positive class. The models were trained using binary cross-entropy loss, a batch size of 32, and a learning rate of 0.001.

During the experiments, three activation functions (ReLU, Sigmoid, and Tanh) and three optimizers (Adam, SGD, and RMSProp) were tested while keeping other parameters fixed. Sequence lengths of 25, 50, and 100 tokens were tested, and gradient clipping was tested with thresholds of 1.0 and 5.0 to evaluate its effect on training stability.

Comparative Analysis:

```
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| Model   | Activation  | Optimizer  |   Seq Length |   Grad Clipping |   Accuracy |     F1 |   Epoch Time (s) |
+=========+=============+============+==============+=================+============+========+==================+
| RNN     | Relu        | Adam       |           50 |             nan |     0.5934 | 0.5644 |             20.2 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| LSTM    | Relu        | Adam       |           50 |             nan |     0.7523 | 0.7555 |             20.3 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Relu        | Adam       |           50 |             nan |     0.7681 | 0.7705 |             21.5 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Sigmoid     | Adam       |           50 |             nan |     0.737  | 0.748  |             21.4 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Relu        | Adam       |           50 |             nan |     0.7681 | 0.7705 |             21.6 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |           50 |             nan |     0.7682 | 0.7748 |             21.5 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |           50 |             nan |     0.7682 | 0.7748 |             21.5 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Sgd        |           50 |             nan |     0.5126 | 0.3778 |             21.1 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Rmsprop    |           50 |             nan |     0.7543 | 0.7702 |             21.4 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |           25 |             nan |     0.7161 | 0.7113 |             21.8 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |           50 |             nan |     0.7682 | 0.7748 |             21.7 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |          100 |             nan |     0.7846 | 0.8058 |             21.9 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |          100 |             nan |     0.7846 | 0.8058 |             21.8 |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |          100 |               1 |     0.8082 | 0.814  |             22   |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
| BILSTM  | Tanh        | Adam       |          100 |               5 |     0.8051 | 0.8138 |             22   |
+---------+-------------+------------+--------------+-----------------+------------+--------+------------------+
```

Discussion:

The configuration that resulted in the best performance was the BiLSTM model, Tanh activation, Adam optimizer, sequence length 100, and gradient clipping threshold of 1.0. This model achieved an F1-score of 0.814 and an accuracy of 0.808, with an average epoch time of approximately 22 seconds.

This indicates that longer input sequences and gradient clipping incorporated at lower thresholds improved model performance and stability. Further, including more context (longer sequence length) allows the model to better capture sentiment information. Both Adam and Rmsprop were better optimizers than SGD, with Adam being marginally better than Rmsprop. Finally, since BiLSTM out performed RNN and LSTM, it was proven that bidirectional context in addition to persistent memory state improves sentiment analysis.

Conclusion:

Although all experiments were conducted using GPU acceleration, the optimal configuration for CPU-only environments takes into consideration the tradeoff between model performance and training time.

The Bidirectional LSTM (BiLSTM) with Tanh activation, the Adam optimizer, a sequence length of 100, and gradient clipping = 1.0 achieved the highest performance, however, the BiLSTM architecture is a heavier model due to the bi-directional processing. Since the LSTM model achieved comparable performance to the BiLSTM, I would use this slightly lighter model in a CPU environment to cut some of the training time at the very marginal cost of some accuracy.

Aside from this, Tanh activation, Adam optimizer, Sequence Length of 100, and Grad Clip at 1.0 would result in the optimal configuration under CPU constraints, balancing accuracy, stability, and efficiency.