MSML641

HW3 – Project Report

By: Bradley Scott

1. Dataset Summary

The IMDb Movie Review Dataset was split into a 50%/50% training/test split. The vocabulary size used was the 10,000 most frequent words. The text is cleaned in several ways to include replacing line breaks, HTML tags, URLS, hyperlinks, or any character that is not a letter, a digit, or a space with a space. We also lowercased all letters. Table 1 displays several statistics about the review datasets reviews word lengths.

Number of reviews:	50,000
Average review word length:	236
mean	236
std	174
min	6
25%	129
50%	177
75%	286
max	2,505

Table 1: Review word length statistics.

2. Model Configuration

Below is a table of all the parameters.

Architectures:	Simple RNN	LSTM	Bidirectional LSTM
Embedding Layer:	100-dimensional word embeddings		
Hidden Layers:	2 recurrent layers, 64 units each		
Dropout:	0.3-0.5 (recurrent and input)		
Output Layer:	Sigmoid for binary classification		
Optimizers:	Adam	SGD	RMSprop
Activations:	Tanh	ReLU	Sigmoid
Sequence Length:	25	50	100
Stability:	None or Gradient Clipping with norm 1		
Batch Size:	32		
Epochs:	6		
Loss Function:	Binary Cross-Entropy		

Table 2: List of all parameters to be tested.

In our search for the best model, we started with a baseline model of architecture = LSTM, activation = Tanh, optimizer = Adam, sequence length = 50 and stabilizer = None. From here we changed only our architecture while keeping all other parameters constant to find the best architecture. Then we varied our activation function while keeping all other parameters constant. We continued this process of changing one parameter to find the best model, while keeping all other parameters constant, for the optimizer, sequence length, and gradient stability. While an approach that checked all combinations would be ideal, it would also be computationally expensive. This method reduced our model checks from 162 models down to 15 models.

3. Comparative Analysis

Below is table of our results. Note that this is in testing order from top to bottom.

Model	Activation	Optimizer	Seq Length	Grad Clipping	Accuracy	F1	Epoch Time (s)
RNN	Tanh	adam	50	No	0.50	0.36	19.82
LSTM	Tanh	adam	50	No	0.76	0.76	54.38
Bidirectional LSTM	Tanh	adam	50	No	0.75	0.76	90.49
LSTM	Tanh	adam	50	No	0.76	0.77	42.68
LSTM	Relu	adam	50	No	0.71	0.73	40.61
LSTM	Sigmoid	adam	50	No	0.76	0.75	42.29
LSTM	Tanh	adam	50	No	0.75	0.74	42.23
LSTM	Tanh	sgd	50	No	0.58	0.45	40.77
LSTM	Tanh	rmsprop	50	No	0.78	0.78	42.12
LSTM	Tanh	rmsprop	25	No	0.73	0.73	40.51
LSTM	Tanh	rmsprop	50	No	0.77	0.78	57.88
LSTM	Tanh	rmsprop	100	No	0.83	0.83	89.12
LSTM	Tanh	rmsprop	100	No	0.83	0.83	107.83
LSTM	Tanh	rmsprop	100	Yes	0.83	0.83	83.26
LSTM	Tanh	rmsprop	100	No	0.83	0.83	85.42

Table 3: Model parameters and results in testing order from top to bottom.

It is easy to see from the chart above that our accuracy and F1 score increased as we continued to refine our model. Our epoch times had more variation but overall, it did appear to increase as our refinement continued. The RNN model was the least accurate, while LSTM was barely higher accuracy than the Bidirectional LSTM.

4. Discussion

The LSTM model with Tanh activation function, RMSprop optimizer, sequence length 100 and no gradient clipping performed the best. The model has an 83% accuracy with an F1-score of 83% with an epoch time of 85.42 seconds.

We can see from table 3 that accuracy and F1 score increased as sequence length increased. This is particularly apparent when we evaluate the sequence length with LSTM, tanh, rmsprop and sequence length 25, 50 and 100.

The rmsprop optimizer performed the best which we see when we evaluated LSTM, tanh, and sequence length 50 with adam, sgd and rmsprop.

Gradient clipping did not change our accuracy or our F1 score. It looks initially like it may have lowered our epoch but when we run the same model a second time, our epochs are much closer. Since there was no increase in accuracy or F1 score, our design chose no gradient clipping in our final result.

5. Conclusion

The LSTM model with Tanh activation function, RMSprop optimizer, sequence length 100 and no gradient clipping performed the best. The model has an 83% accuracy with an F1-score of 83% with an epoch time of 85.42 seconds.

6. References

1. Maas, Andrew L., et al. *"Large Movie Review Dataset." Stanford AI Lab*, 2011, https://ai.stanford.edu/~amaas/data/sentiment/.

2. *ChatGPT-5*. *OpenAI*, 2025. Code refinement and experimental design support.