Dataset Summary:

- The dataset contains 50,000 data with 2 columns. The first column is named “review”, which holds the comments from users. The second column is called “sentiment”, which is the true sentiment label of each comment. It has two unique values: “positive”, refers to positive sentiment, and “negative”, refers to negative sentiment. The average review length is about 231 words, with the longest comment of 2470 words and the shortest comment of 4 words. The vocabulary size of this dataset “review” column is 438729. The preprocessing of the dataset include lowercasing all text, removing punctuation and special characters such as single character, then keep the top 10000 most frequent words and tokenization using nltk.word_tokenization

Model Configuration:

- We have 3 different neural networks models: RNN, LSTM, and Bidirectional LSTM. Each model has an embedding dimension of 100, with 2 fully connected hidden layers and hidden size of 64. The dropout rate is set to 0.5 for all 3 models. The batch size of all models is 32 and all models use binary cross-entropy loss. For the optimizer setting, we have tested 3 different optimizers for all 3 models: Adam, Stochastic Gradient Descent (SGD), RMSProp.

Comparative Analysis and Discussion:

- Here according to the result, we have the top 5 model configurations on accuracy with following table:

	Architecture	Activation	Optimizer	Sequence_Length	Gradient_Clip	Accuracy	F1_Score	Avg_Epoch_Time(s)	Total_Train_Time(s)
143	Bidirectional LSTM	relu	rmsprop	100	Yes	0.8196	0.8206	2.67	36.92
161	Bidirectional LSTM	tanh	rmsprop	100	Yes	0.8174	0.8208	2.23	34.12
128	Bidirectional LSTM	relu	adam	100	No	0.8163	0.8115	2.19	33.62
140	Bidirectional LSTM	relu	rmsprop	100	No	0.8162	0.8222	2.12	33.71
131	Bidirectional LSTM	relu	adam	100	Yes	0.8157	0.8258	2.41	34.58

For the top 5 model configurations on F1 score, we have this table:

	Architecture	Activation	Optimizer	Sequence_Length	Gradient_Clip	Accuracy	F1_Score	Avg_Epoch_Time(s)	Total_Train_Time(s)
131	Bidirectional LSTM	relu	adam	100	Yes	0.8157	0.8258	2.41	34.58
122	Bidirectional LSTM	sigmoid	rmsprop	100	No	0.8070	0.8229	2.24	33.84
125	Bidirectional LSTM	sigmoid	rmsprop	100	Yes	0.8135	0.8223	2.38	34.44
140	Bidirectional LSTM	relu	rmsprop	100	No	0.8162	0.8222	2.12	33.71
161	Bidirectional LSTM	tanh	rmsprop	100	Yes	0.8174	0.8208	2.23	34.12

From these tables, both top 5 in Accuracy and F1 are Bidirectional LSTM, which indicates that this model architecture has obvious advantages on the IMDB comment

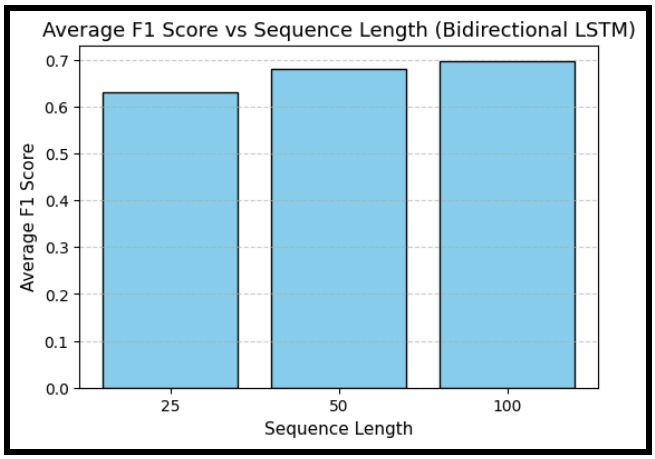
sentimental classification task. Moving to the sequence length, it is also obvious that the sequence length of 100 generates the best accuracy and F1 result, with all these 10 data having the sequence length of 100. Moving to the optimizer parameter, for both top 5 accuracy and F1 score tables, the RMSProp optimizer contributes most of the data, with 3 data at top 5 accuracy table and 4 data at top 5 F1. For activation function and gradient clip, they do not generate obvious advantages for a single parameter but we can say that the activation function of ReLU and the presence of gradient clips tends to produce better results. Therefore, if we consider the accuracy and F1 score, the best configurations according to my result is:

Architecture	Activation	Optimizer	Seq_Length	Gradient_Clip
Bidirectional LSTM	ReLU	RMSProp	100	Yes

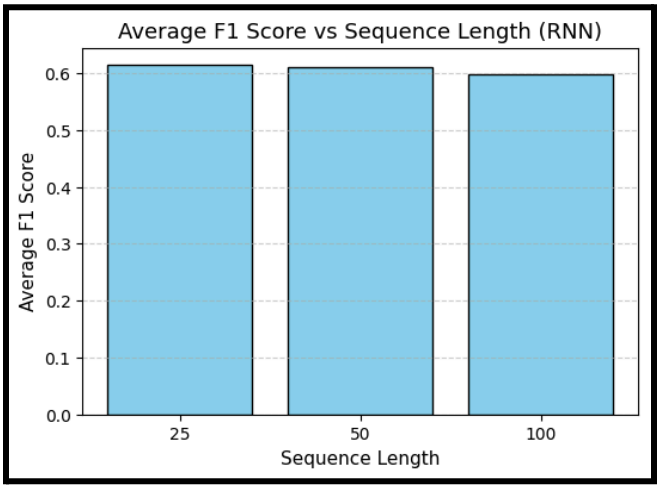
- On the other hand, if we consider about the total training time, we have a different result with the following table:

	Architecture	Activation	Optimizer	Sequence_Length	Gradient_Clip	Accuracy	F1_Score	Avg_Epoch_Time(s)	Total_Train_Time(s)
43	RNN	tanh	sgd	50	No	0.5006	0.5056	1.22	27.93
7	RNN	sigmoid	sgd	50	No	0.5038	0.4768	1.28	28.13
24	RNN	relu	sgd	25	No	0.5005	0.4845	1.29	28.16
6	RNN	sigmoid	sgd	25	No	0.4982	0.5028	1.29	28.22
25	RNN	relu	sgd	50	No	0.5052	0.4852	1.30	28.39

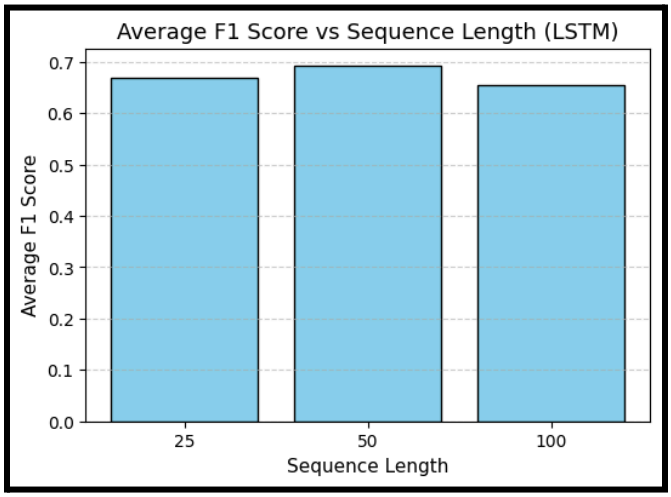
  In this table, we can see that RNN generally takes less time than other 2 models, with the optimizer all SGD, sequence length of mostly 50, and no gradient clip presence for the top 5 least training time table. However, all 5 of them perform significantly lower on both accuracy and F1 score, with their training time not much improvement (less than 10 second improvement with 25000 training data). Therefore, in this 50000 IMDB comment sentiment task, we can still consider that the above chart would be the best configuration.
- The influences of sequence length toward the performance depends on different models. For example, in Bidirectional LSTM, a longer sequence tends to generate better performance, which we can see in the bar graph below:



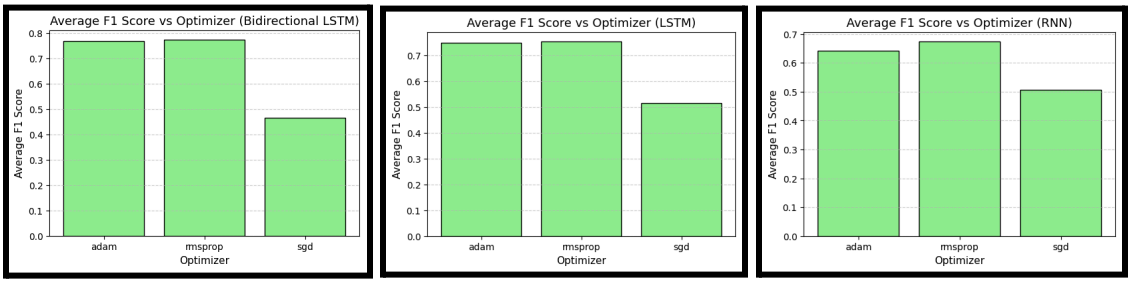
However, in the RNN model, longer sequence length tends to have a little bit negative influence on the performance, which can be seen in the graph below:



In LSTM model, on the other hand, the sequence length of 50 actually tends to generate the best performance:



- For optimizer influence, we can see the bar graph below:



  According to these results, we can see that the influences of optimizers are pretty much the same across all 3 models, with RMSprop has the best performance; Adam has relatively smaller F1 score on all 3 models; and SGD has significantly lower performance in all models compared with the previous 2 optimizers. However, even though the SGD has the worst performance, it tends to take the shortest training time.
- For the gradient clip, all 3 models are set to renormalize the gradient that has norm greater than 1 to prevent gradient explosion. However, for gradients to vanish, the RNN model does not have any mechanism to prevent it, while LSTM and Bidirectional LSTM has its built-in mechanism to prevent gradient vanish. Therefore, this indicates why LSTM and Bidirectional LSTM tend to have better performance than RNN.