Sileo Nomen

Task Overview:

The project was aimed at building and evaluating a sentiment analysis model capable of distinguishing between positive and negative sentiments expressed in text reviews. The goal was to accurately classify text into these two categories based on the sentiment (+,-).

Dataset:

The dataset used consisted of reviews extracted from kaggle, labled online text reviews from anonymous users. The reviews were labeled as either "positive" or "negative" based on the sentiment expressed by the user. The dataset was pre-processed to remove punctuation and converted to lowercase to standardize the input.

Data Split:

The dataset was divided into three subsets:

Training set: 80% of the data, used to train the model.

Validation set: 10% of the data, used to tune the parameters and prevent overfitting.

Testing set:\* 10% of the data, used to evaluate the model's performance in predicting unseen data.

Model Scheme:

The primary architecture explored was a bidirectional LSTM (Long Short-Term Memory) model due to its effectiveness in handling sequence data and its ability to capture context from both past and future data points. The model included:

Embedding layer: To convert text data into dense vectors of fixed size.

Bidirectional LSTM layer: To process text sequences (I added padding to maintain consistency between hidden to hidden layers and most importantly hidden to output layers).

Dropout layer: To prevent over fitting by randomly setting a fraction of the input weights to 0 at each update during training time.

Fully connected layer: To direct the output layers of the LSTM into the final output labels.

Sigmoid activation function: To convert the output into a probability score between 0 and 1 (this would be found in def prediction).

Several hyper-parameters were varied during experimentation, including the number of LSTM units, dropout rate, and learning rate. The best results were obtained with a higher number of LSTM units and a moderate dropout rate.

Final Settings and Results:

The final network configuration used was:

Number of LSTM units: 512

Dropout rate:0.5

Learning rate: 0.001

Optimizer: Adam

Loss function: Binary Cross-Entropy

This configuration was chosen based on its performance on the validation set. The final model was trained using these settings and then evaluated on the test set.

Results:

The final model achieved the following results on the test set:

Accuracy: 50.00%

Precision: 50.00%

Recall: 100.00%

F1-Score: 66.67%

The results indicate that the model was 50% accurate in classifying the sentiment of the reviews. However, the close raw output scores (pre-sigmoid) for clearly positive and negative texts suggest that the model may still be struggling to distinguish effectively between the two classes. This could potentially be addressed by further tuning the model or experimenting with alternative text preprocessing techniques or model architectures.

Conclusion:

The bidirectional LSTM model demonstrated decent capability in classifying sentiments of text data, but there is room for improvement. Future work could explore more complex network architectures, different types of embeddings, or more advanced natural language processing techniques to enhance model performance, initially I was using a RNN with a stack LSTM layer but later switch to a bi-directional LSTM; these were my findings.