CLASSIFICATION OF SMS USING RNN AND LSTM

INTRODUCTION: The goal of the task is to classify the SMS whether or not they are spam. The models RNN and LSTM are built to do this task. Both of the models should have approximately similar number of parameters. The dataset contains mainly two columns. One is 'v1', containning the labels and the other is 'v2', containning the SMS messages, which are used to train and evaluate the models.

METHODOLOGY: For the data preprocessing, the unnecessary columns of the dataset are removed and the categorical column of label is encoded ('ham' as 0 and 'spam' as 1) for further process. A bar plot is drawn to visualize the class distribution of target variable. Then the text normalization is done using lemmatizer, which reduce words to their basic forms. Next, each message is tokenized with a vocabulary of up to 10,000 unique words and all tokenized sequences are padded to a fixed length of 100 tokens, ensuring uniformity for model input. Then the data is split into trainning and testing sets in a ratio 80:20.

Since the data is imbalanced, 'SMOTE' function is applied to balance the trainning data.

Now the RNN model is built starting with an Embedding layer to transform each word into a dense vector of size 32, then single SimpleRNN layer with 64 units, following to a fully connected layer with ReLU activation and sigmoid function for output layer for binary classification.

Similarly, LSTM model is built except the LSTM layer with 34 units to maintain similar number of parameters.

Both of the model are trained for 10 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss. Validation is performed on the testing data.

RESULT: Test accuracy for the RNN model is 95% and for LSTM, 93%.

Metric	**SimpleRNN**	**LSTM**
Training Accuracy	Reaches 100% quickly suggesting potential overfitting	Reaches 100%, though the increase is more gradual.
Validation Accuracy	Fluctuates around 94% initially, then stable around 95%	Varies significantly across epochs, indicating inconsistent generalization
Training Loss	Decreases consistently	Decreases consistently
Validation Loss	Fluctuates and increases over epochs, with an overall higher trend compared to trainning loss	Fluctuates initially and appears unstable with a large spike

Table 1: **Comparison Table for RNN and LSTM model**

Since the test accuracy for the RNN model is better than the LSTM model, the RNN model is slightly more suitable for this SMS classification task.