## Part 3: Spam Classification using Recurrent Neural Networks

Objective: The objective of this assignment is to develop an RNN model for spam classification. You will use a dataset containing SMS messages labeled as spam or ham (not spam). You will use this data to train an RNN model to classify messages as spam or not spam.

Introduction

The objective of this report is to summarize the findings from training and evaluating recurrent neural network (RNN) models for SMS spam classification. We explored various hyperparameters and architectures to optimize model performance on this task.

Experiment Setup

We conducted experiments using a dataset of SMS messages labeled as spam or ham (not spam). The dataset was preprocessed by tokenizing messages, converting labels to binary (0 for ham, 1 for spam), and padding sequences to a fixed length of 100 tokens.

Model Architectures Explored

We experimented with two primary types of RNN architectures:

SimpleRNN Model: A basic RNN model with a single SimpleRNN layer.

LSTM Model: A more complex model using LSTM (Long Short-Term Memory) layers to capture long-range dependencies in text sequences.

Hyperparameters Explored

We varied the following hyperparameters during experimentation:

Embedding Dimension (embedding\_dim): Tested with values of 50 and 100.

Number of RNN Units (rnn\_units): Explored with 32 and 64 units.

Batch Size: Set to 32 for training the models.

Number of Epochs: Trained the models for 15 epochs.

Model Performance Analysis

After training and evaluating multiple models, the following key findings were observed:

SimpleRNN Model:

Best Configuration: embedding\_dim=50, rnn\_units=32

Test Accuracy: 87.80%

LSTM Model:

Best Configuration: embedding\_dim=100, rnn\_units=32

Test Accuracy: 94.26%

Factors Contributing to Success

Architecture Selection:

LSTM models consistently outperformed SimpleRNN models, indicating that the ability of LSTM to capture long-term dependencies was beneficial for SMS spam classification.

Embedding Dimension:

Higher embedding dimensions (e.g., embedding\_dim=100) led to improved performance, likely because of richer representations of words in the embedding space.

Number of RNN Units:

Models with fewer RNN units (e.g., rnn\_units=32) achieved competitive accuracy levels while being computationally efficient.

Conclusion

In conclusion, LSTM-based RNN models with higher embedding dimensions and moderate numbers of RNN units yielded the best performance for SMS spam classification. Continual exploration of advanced architectures and systematic hyperparameter tuning are essential for further improving model accuracy and robustness in real-world applications.