# Assignment-4

# Text Data ( AML)

**Task:**

Our task is to apply RNNs to text and sequence data for the IMDB dataset. We use the following conditions, given in the assignment and applied on my code.

**About RNN's:**

Before I would like to provide the significance of RNN's in the deep neural network.

Recurrent Neural Networks (RNNs) are a type of Deep Neural Network (DNN) that have proven to be very useful in tasks that involve processing sequences of data, such as natural language processing, speech recognition, and time series prediction.

RNNs are also very flexible and can be adapte to a wide range of tasks by modifying their architecture and training regime. For example, variants of RNNs such as LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) has been develope to improve the network's ability to handle long-term dependencies and avoid the vanishing gradient problem.

**About Glove:**

The GloVe algorithm tracks the frequency with which words appear together in a huge corpus of text by creating a co-occurrence matrix. The statistical relationship among words can be stored as a phrase vectors by adjusting these matrices of data.. GloVe uses both the co-occurrence counts of words and their ratios to capture more complex relationships between words than other word embedding techniques like Word2Vec.

The modifications are applied on the code and results are below,

1.Cutoff reviews after 150 words

2) Restrict training samples to 100

3) Validate on 10,000 samples

4) Consider only the top 10,000 words

5) Before the layers. Bidirectional layer, consider a) an embedding layer, and b) a pretrained word embedding.

**Performance Results:**

| Train Sample Size | First Basic Sequence Model (Test Accuracy) | Embedded (Test Accuracy) | Embedded Masked (Test Accuracy) | Pre-Trained word embedding (Glove) (Test Accuracy) |
|---|---|---|---|---|
| 100 | 80.0 | 78.5 | 79.7 | 77.2 |
| 500 | 82.9 | 83.1 | 83.0 | 82.9 |
| 800 | 84.7 | 82.8 | 83.4 | 82.9 |
| 2500 | 84.4 | 82.8 | 84.0 | 83.2 |
| 5000 | 83.7 | 83.5 | 83.5 | 83.6 |

The results of the analysis showed that RNN's with Embedded layers performed significantly better than the other word embedding techniques., such as one-hot encoded sequence. The embedded layer models consistently outperformed the other techniques in terms of test accuracy.

It can also be observed that the performance of RNN models improved with increasing sample size. As the sample size increases the test accuracy will also increase.

Furthermore, with comparison of different types of embedded layer, including standard embedded and masked embedded. The standard embedded layer performs significantly slightly better than the masked embedded.

**Conclusion:**

Overall, this task with RNN's with different techniques shows varied performances on different sample sizes, one keynote is when we increase the sample size the performance of the model increases in the embedded and masked embedded. Based on the results obtained using the pre-trained glove embeddings showed better than basic sequence. Therefore, using the pre-trained glove embedding is efficient in terms of text analysis for the IMBD data.