# Report - Assignment 3

I've used IMDB data set as it can be directly loaded into Keras. IMDB dataset is the Large Movie Review Dataset that contains 25,000 highly polar moving reviews (good or bad) for training and the same amount again for testing. The data was collected by Stanford researchers and was used in a 2011 paper (http://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf) where a split of 50/50 of the data was used for training and test. Reference - https://machinelearningmastery.com/predict-sentiment-movie-reviews-using-deep-learning/

## Task 1 - CNN

### Architecture of Model

1. A efficient embedding layer is used to maps the vocabulary indices into embedding_dims dimensions
2. Then a 1D Convolution Network is used which will learn the filters
3. Then we use MaxPooling
4. Add a Dense Layer
5. Project onto a single output layer and then use sigmoid activation function

### Training

Use Batch Size of 32 and train the model for 3 epochs

## Task 2 - RNN

### Architecture of Model

1. An efficient embedding layer is used to maps the vocabulary indices into embedding_dims dimensions
2. Then a LSTM Network with 128 units having dropout = 0.2
3. Project onto a single output layer and then use sigmoid activation function

### Training

Use Batch Size of 32 and train the model for 3 epochs

## Extending the approaches

1. We could clean the dataset by converting all letters to lowercase and remove stopwords and irrelevant whitespaces.
2. Tokenize the words in a sentence and use word embeddings like glove
3. We could use count vectors or TF-IDF as features
4. In RNN we could use Bi-directional LSTM