Mini Project 2: IMDB movie review sentiment classification

Dataset – The dataset under consideration contains a total of 50,000 movie reviews obtained from the Internet Movie Database. These reviews have been divided into two distinct sets: a training set comprising 25,000 reviews and a testing set consisting of the remaining 25,000 reviews. It is noteworthy that the reviews are not uniformly distributed across the entire spectrum of sentiment; rather, they are dichotomous in nature, with half of the reviews expressing positive sentiments and the other half expressing negative sentiments.

Link: Dataset

Tasks

Task 1 – In this exercise, you will construct **three distinct machine learning models** for sentiment classification using the provided dataset. As a machine learning expert, your objective is to identify suitable methods and evaluate their performance to achieve optimal results. You are permitted to utilize any model introduced during the course, in addition to approaches not covered that you deem applicable to your research. These may include traditional methods and deep neural networks (DNNs) with fine-tuning. Examples of such methods include CNNs, RNNs, Transformers architectures, Random Forests, Naïve Bayes, and SVM.

Task 2 – Compare and visualize the learning curves and performance of the three models.

Task 3 – Write a scientific report which includes

- Introduction (what is the problem you are solving?)
- Data processing (what are the choices you made in data processing and how you performed it?)
- Modelling (What are the modeling approaches? How have you performed them? Why do you think one model performed better than the other one? What can be done better? Can achieve over 80% accuracy?)
- Conclusions (what were the “scientific” bottlenecks? How did you overcome them? What is the result you obtained with your best model? Etc..)

You need to hand in your Python code (preferably Jupyter notebook or Google Colab notebook) alongside a written report.