Sentiment Analysis on IMDb Movie Reviews

**Objective:**

To design and implement a deep learning model that can predict the sentiment (positive or negative) of movie reviews from the IMDb dataset.

**Approach:**

1. Data Preprocessing:

- Loaded the IMDb movie reviews dataset.

- Used BeautifulSoup to remove any HTML tags from the reviews.

- Removed URLs and texts within square brackets.

- Tokenized and padded the reviews to create uniform input sequences.

2. Initial Model - LSTM:

* Initially, an LSTM-based model was chosen for this task. LSTMs (Long Short-Term Memory networks) are a type of RNN that are well-suited for sequence-based tasks like sentiment analysis.
* However, during the training process, the LSTM model exhibited significant overfitting, where it performed very well on the training data but poorly on the validation data.
* Attempts to mitigate overfitting using dropout and regularization did not yield satisfactory results.

3. Transition to 1D CNN:

* Given the challenges with the LSTM model, the approach was pivoted to using a simpler 1-dimensional Convolutional Neural Network (1D CNN).
* CNNs, while typically associated with image processing, can also be effective for sequence data. By using convolutional layers, the model can identify local patterns or sequences of words that are indicative of sentiment.
* Dropout layers were added to the CNN model to prevent overfitting. The dropout rate was set relatively high to ensure that the model generalizes well.

5. Evaluation:

- To ensure a robust evaluation of the model, 5-fold cross-validation was implemented. This entails splitting the dataset into five parts and training the model five times, each time using a different part as the validation set.

- This method provides a more comprehensive assessment of the model's performance on different data splits and reduces the risk of overfitting on a single validation set.