Sentiment Analysis Project on IMDb dataset Using Python

Overview:

This project focuses on sentiment analysis of movie reviews from the IMDb dataset. The primary objective is to classify the reviews as positive or negative using a **Recurrent Neural Network (RNN),** specifically a Long Short-Term Memory (LSTM) network. The project involves several key steps: data collection, data preprocessing, model building, model training, model evaluation, and making predictions on new data.

Data Collection:

The dataset used in this project is the IMDb movie reviews dataset, which is readily available in the Keras library. The dataset consists of 25,000 highly polar movie reviews for training and 25,000 for testing. Reviews are pre-processed into sequences of integers representing the words in the reviews.

Data Preprocessing:

To prepare the data for model training, the sequences of words (represented by integers) are padded to ensure uniform length across all reviews. Padding is essential for batch processing in neural networks, especially for models like RNNs that require input sequences to be of the same length.

Key actions in data preprocessing include:

- Limiting the vocabulary size to 10,000 most common words.

- Padding sequences to a maximum length of 200 words.

Model Building:

An LSTM network is constructed using the Keras Sequential API. The architecture of the model includes:

- An Embedding layer to convert integer-encoded words to dense vectors of fixed size.

- An LSTM layer with 128 units, which captures the temporal dependencies in the sequence data.

- A Dense output layer with a single neuron and a sigmoid activation function for binary classification (positive or negative sentiment).

The model is compiled with binary cross-entropy as the loss function and Adam optimizer. Accuracy is used as the performance metric.

Model Training:

The model is trained using the training dataset. Key training parameters include:

- Batch size: 32

- Number of epochs: 1

- Validation split: 20% of the training data is used for validation during training to monitor the model's performance on unseen data.

Model Evaluation:

The trained model is evaluated on the test dataset to assess its performance. The evaluation metrics include:

- Test Loss: A measure of the model's prediction error on the test data.

- Test Accuracy: The proportion of correctly classified reviews out of the total reviews in the test dataset.

Predictions:

The model makes predictions on new data by outputting probabilities that the reviews are positive. These probabilities are converted to binary labels (positive or negative) based on a threshold of 0.5. The results are then compared with the actual labels to verify the model's performance.

Conclusion:

The project successfully demonstrates the process of building an LSTM-based sentiment analysis model using the IMDb movie reviews dataset. The steps include data collection, preprocessing, model construction, training, evaluation, and prediction, highlighting the importance of each phase in developing an effective sentiment analysis.