Sentiment Analysis using Neural Networks

This notebook demonstrates a sentiment analysis task using the IMDb dataset. The process involves loading the data, preprocessing it, creating a neural network model, and evaluating its performance.

Data Loading and Preprocessing

Data Loading: The IMDb reviews and labels are loaded into pandas dataframes. Labels are converted to a binary format indicating positive or negative sentiment.
Data Splitting: The dataset is split into training, validation, and test sets using train_test_split from sklearn.
Bag-of-Words Representation: A Bag-of-Words (BoW) representation is created using CountVectorizer from sklearn, limited to the 10,000 most frequent words.

Exploratory Data Analysis

Vocabulary Analysis: The vocabulary size and sample words are displayed to understand the BoW representation.
BoW Representation: The BoW representation of the first review is shown to illustrate how text data is transformed into numerical format.

Model Building and Training

Neural Network Architecture: A Convolutional Neural Network (CNN) with one hidden layer is built using TensorFlow Keras. The model includes Conv1D, MaxPooling1D, Flatten, Dense, and Dropout layers.
Model Compilation: The model is compiled with the Adam optimizer and binary cross-entropy loss function.
Training: The model is trained on the training set with early stopping based on validation accuracy to prevent overfitting.

Model Evaluation

Performance Metrics: The model's performance is evaluated on the test set, achieving an accuracy of approximately 88.34%.
Custom Sentence Classification
Sentiment Prediction: The trained model is used to classify the sentiment of custom sentences. The sentences are transformed into BoW representation and reshaped to fit the CNN input requirements. The model predicts whether each sentence has a positive or negative sentiment.
This notebook provides a comprehensive workflow for performing sentiment analysis using neural networks, from data preprocessing to model evaluation and application on new data.