ASSIGNMENT 4

NARESH CHIKKULA

1. Introduction

This project focuses on building a sentiment-classification system using the IMDB movie review dataset. The objective is to automatically determine whether a given review expresses a positive or negative opinion. The workflow includes data extraction, preprocessing, model construction, training, evaluation, and visualization.

2. Data Preparation

Dataset Overview

The IMDB dataset contains 50,000 movie reviews split evenly into:

- **25,000 training samples**
- **25,000 test samples**

Each subset is evenly divided into positive and negative examples. The dataset is widely used for benchmarking natural-language-processing models.

Downloading and Unpacking

The notebook downloads the compressed dataset, extracts it, and removes unnecessary folders (such as the “unsupervised” set) to focus exclusively on labeled sentiment data.

Directory Structure

After processing, the dataset is organized into:

```
train/
   pos/
   neg/
test/
   pos/
   neg/
```

3. Data Preprocessing

Text Cleaning and Label Extraction

The notebook uses a TensorFlow “text_dataset_from_directory” pipeline to:

- Load text files
- Infer labels based on directory names
- Shuffle and batch the data

Vectorization

Before feeding text into the model, reviews are transformed using a **TextVectorization** layer. This step:

- Applies standardization (lowercasing, punctuation removal)
- Tokenizes text into words
- Converts tokens into integer sequences
- Restricts vocabulary size to improve generalization

4. Model Development

Architecture

A simple but effective neural network is built using TensorFlow/Keras. The model includes:

- **Embedding layer** for representing words as dense vectors
- **Dropout layers** to reduce overfitting
- **GlobalAveragePooling** or similar mechanism to combine word embeddings
- **Dense output layer** with sigmoid activation for binary classification

Compilation

The model is compiled with:

- **Binary cross-entropy** loss
- **Adam** optimizer
- **Accuracy** as the primary metric

5. Training and Evaluation

Training

The model is trained using the prepared batches of text. A portion of the training data is set aside as a validation set to monitor performance during training.

Results

The notebook prints training progress, showing:

- Gradual reduction in loss
- Increasing training and validation accuracy
  Typical accuracy for this architecture on IMDB is around **85–90%**.

Final Evaluation

The trained model is evaluated on the test set to measure its generalization performance. Accuracy values close to validation accuracy confirm the model is not overfitting.

6. Model Deployment Preparation

Exporting the Model

The notebook demonstrates how to wrap the text vectorization layer and the trained network into a single exportable model capable of processing raw text directly.

Sample Predictions

Sample reviews are passed into the model to check whether their predicted sentiment aligns with expectations.

7. Conclusion

This assignment successfully walks through the entire pipeline of sentiment analysis using deep learning. Through proper data handling, preprocessing, and neural network modeling, the system achieves high accuracy in classifying movie reviews as positive or negative. The final exported model can be used for real-world sentiment-analysis tasks with minimal modifications.