LAPTOP REVIEW CLASSIFICATION - NLP

Problem Statement

The goal of this project was to classify laptop product reviews into positive or negative sentiment using deep learning and natural language processing (NLP) techniques.
The dataset initially provided had many duplicate entries. Therefore, the task involved cleaning the dataset, generating sentiment labels, and building a binary classifier using LSTM neural networks.
The challenge was to work under strict conditions:

- No traditional machine learning models
- No use of external sentiment APIs (e.g., TextBlob, VADER)
- Dataset modifications allowed only through deduplication or synthetic review generation

Methods

1. **Data Cleaning**
    - Removed duplicate reviews using pandas.drop_duplicates() on the Review Text column.
2. **Synthetic Review Generation**
    - After deduplication, the dataset had only 42 unique entries.
    - We created 10 synthetic reviews (5 positive, 5 negative) using a template-based approach and manual labeling to maintain balance.
3. **Preprocessing**
    - Converted text to lowercase
    - Removed special characters and stop words
    - Applied lemmatization
    - Tokenized and padded text sequences using Keras tokenizer and pad_sequences

4. **Model Architecture**

    - Embedding Layer: Converts words into dense vectors
    - LSTM Layer: Captures sequential relationships in reviews
    - Dense Output Layer: Uses sigmoid activation for binary prediction

5. **Evaluation**

    - Split the dataset into 70% training and 30% testing
    - Used metrics such as accuracy, precision, recall, F1-score, and a confusion matrix

Insights

- Removing duplicates significantly reduced the dataset from ~5000 reviews to just 42.
- Adding synthetic reviews improved class balance and data diversity.
- The LSTM model performed better on positive reviews (Recall = 79%) than on negative ones (Recall = 25%).
- Overall model accuracy was around 53.8%, which is reasonable given the limited dataset size and constraint of not using external models or APIs.

Challenges

- High redundancy in the original dataset led to significant data loss after cleaning.
- No labeled data meant we had to manually label synthetic reviews or generate them using rule-based techniques.
- The dataset size was small, making it difficult for the LSTM to generalize well.
- Class imbalance influenced the model to favor positive reviews more than negative ones.

Conclusion

This project demonstrated how deep learning models like LSTM can be applied to sentiment classification of product reviews under constrained conditions. Despite the limited and synthetic nature of the dataset, we successfully built a functional pipeline for:

- Data preprocessing
- Text embedding
- Model training
- Performance evaluation

With access to a larger and more diverse labeled dataset, the model's performance could be further improved using advanced architectures like Bidirectional LSTM, attention mechanisms, or pretrained embeddings (e.g., GloVe or BERT).