**Internship Task 4 Report — Sentiment Analysis of E-Commerce Reviews**

**1. Introduction**

This report documents **Internship Task 4 (Code Alpha)** based on the uploaded Jupyter notebook. The task focuses on **sentiment analysis of e-commerce reviews** using exploratory data analysis (EDA), lexicon-based sentiment scoring (VADER), and model-based approaches (ROBERTA transformer and LSTM neural networks).

**Objectives:**

* Explore the dataset and perform descriptive analysis.
* Apply **VADER** sentiment scoring for baseline sentiment classification.
* Build and evaluate **transformer-based (RoBERTa)** and **LSTM-based** models for sentiment classification.
* Compare outputs across approaches and visualize results.

**2. Methodology**

**Tools & libraries**

* **Data Handling:** pandas, numpy
* **Visualization:** matplotlib, seaborn
* **Text Processing:** nltk
* **Models:** transformers (RoBERTa), keras/tensorflow or torch (LSTM)
* **Evaluation:** scikit-learn (classification metrics)

**Workflow**

1. **Data Loading & Preprocessing**
   1. Load reviews.csv, check null values, shape, and preview.
   2. Tokenization and part-of-speech tagging (NLTK).
   3. Stopword handling and text cleaning.
2. **Lexicon-based Sentiment (VADER)**

* Compute compound, positive, neutral, and negative scores.
* Assign labels based on thresholds.
* Visualize sentiment distributions.

1. **Transformer Model (RoBERTa)**

* Tokenize using pretrained RoBERTa tokenizer.
* Fine-tune cardiffnlp/twitter-roberta-base-sentiment.
* Run inference on reviews.

1. **Evaluation & Visualization**

* Accuracy, Precision, Recall, F1-score.
* Confusion matrices.
* Sentiment distribution plots.
* Top positive/negative words.

**3. Key Findings**

* **VADER** provides quick baseline sentiment scoring but struggles with context-specific language.
* **RoBERTa** expected to yield the best accuracy (pretrained on large corpora).
* sentiment distributions and top word frequencies provide qualitative insights.

**4. Conclusions & Recommendations**

* **VADER**: good for baseline, lacks nuance.
* **RoBERTa**: strong performance expected; ensure fine-tuning is logged with validation metrics.