**Project-5**

**Amazon Sentiment Sleuth Analyzing Product Reviews with IIMS**

**Overview:**

This report outlines the process of conducting sentiment analysis on Amazon reviews for personal care appliances. The analysis leverages natural language processing (NLP) techniques using the Distil BERT model from Hugging Face's transformers library. The goal was to classify customer reviews into three sentiment categories: positive, neutral, and negative.

**Stepwise Process:**

**Step 1: Data Acquisition**

* **Dataset:** amazon\_reviews\_us\_Personal\_Care\_Appliances\_v1\_00.tsv
* The dataset was obtained from Amazon's open data repository, containing reviews related to personal care appliances.

**Step 2: Data Loading**

* Utilized pandas to load the TSV file.
* Handled potential bad lines in the dataset to ensure smooth data importation.

**Step 3: Data Preprocessing**

* Selected essential columns: review\_body (review text) and star\_rating (review rating).
* Removed missing values to maintain data integrity.
* **Sentiment Labeling:**
  + **Negative (0):** Ratings ≤ 2
  + **Neutral (1):** Rating = 3
  + **Positive (2):** Ratings ≥ 4

**Step 4: Data Splitting**

* Split the dataset into **training (80%)** and **validation (20%)** sets using train\_test\_split from sklearn.

**Step 5: Tokenization**

* Employed DistilBertTokenizer to convert review texts into token IDs.
* Applied truncation and padding with a maximum length of 64 tokens for uniformity.

**Step 6: Dataset Creation**

* Defined a custom SentimentDataset class compatible with PyTorch.
* Prepared data loaders for batch processing during training and evaluation.

**Step 7: Model Setup**

* Implemented the DistilBertForSequenceClassification model for sentiment classification.
* Defined:
  + **Loss Function:** CrossEntropyLoss
  + **Optimizer:** Adam optimizer

**Step 8: Model Training**

* Trained the model on the training set over multiple epochs.
* Monitored loss and accuracy to track learning progress.

**Step 9: Model Evaluation**

* Evaluated the model's performance on the validation set using:
  + **Accuracy:** Overall correctness of predictions
  + **Precision, Recall, F1-Score:** Metrics to assess performance for each sentiment category

**Step 10: Results & Insights**

* The model effectively classified sentiments with reasonable accuracy.
* Identified areas for improvement, including potential hyperparameter tuning and advanced model architectures for better performance.

**Conclusion:**

The sentiment analysis pipeline effectively processed Amazon review data, trained a robust NLP model, and delivered meaningful insights into customer sentiments. This workflow can be adapted and scaled for similar sentiment analysis projects across different domains.