Sentiment Analysis of Amazon Reviews

Internship: EduLumos Data Science Intern

Submitted by: Syeda Faizah

Task Report – Week 3
Natural Language Processing – Sentiment Analysis

Tools & Technologies Used:
Python, Pandas, Scikit-learn, TF-IDF, Logistic Regression

1. Introduction

Sentiment analysis is a **Natural Language Processing (NLP)** technique used to determine the emotional tone expressed in textual data. In this project, sentiment analysis was performed on Amazon customer reviews to classify them as **positive** or **negative** based on their content.

The objective of this task was to apply **text preprocessing, feature extraction, and machine learning techniques** to build an effective sentiment classification model that can predict the sentiment of new customer reviews.

2. Dataset Description

The dataset consists of Amazon customer reviews with corresponding rating scores. The relevant columns used are:

- reviewText – the text of the review (input feature)
- overall – the rating score (used to derive sentiment labels)

The dataset contains **[insert number of rows] reviews**. All reviews were provided by the company for internship purposes.

3. Data Preprocessing

To prepare the dataset for modeling, the following steps were taken:

- **Missing and empty reviews** were removed to ensure clean input.
- **Ratings were converted into sentiment labels**:
  - overall >= 4 → **positive**
  - overall <= 2 → **negative**
- **Neutral reviews (overall = 3)** were excluded to simplify the problem into binary classification.
- **Duplicate reviews** were checked and removed to prevent bias in model training.

After preprocessing, the dataset was ready for feature extraction and model training.

4. Feature Extraction

Text data was converted into numerical features using **TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization**, which assigns importance to words based on their frequency in a review and across the entire dataset.

Parameters used:

- max_features = 5000 → limits the number of unique words/features
- stop_words = 'english' → removes common English stopwords

This transformed the text data into a **sparse numerical representation** suitable for machine learning.

5. Model Training

A **Logistic Regression** classifier was trained on the TF-IDF features to learn patterns associated with positive and negative reviews.

- Logistic Regression was chosen because it is **simple, fast, and effective** for binary classification tasks.
- The training set consisted of **80% of the dataset**, while **20% was used for testing**.

6. Model Evaluation

The trained model was evaluated on the test set using **accuracy** and other classification metrics:

- **Accuracy:** 94.45%
- **Precision, Recall, F1-score:** high for both positive and negative classes (detailed metrics included in the notebook)

The high accuracy indicates that the model performs well in distinguishing positive and negative reviews.

Sample Prediction:

sample_review = ["The product quality is excellent and delivery was fast"]
Predicted Sentiment: positive

7. Conclusion

The Logistic Regression model effectively classified Amazon customer reviews into positive and negative sentiments.

- Positive reviews dominate the dataset.
- The model achieved **94.45% accuracy** on the test set.
- Sample review predictions matched intuitive sentiment.

This project demonstrates the practical application of **NLP and machine learning techniques** for sentiment analysis in real-world data.

8. Future Scope

- Extend classification to include **neutral reviews** for multi-class sentiment analysis.
- Experiment with **advanced NLP models** such as BERT or LSTM for potentially higher accuracy.

- Analyze **sentiment trends over time** or across **product categories** to gain deeper insights.