**Sentiment Analysis of Product Reviews**

Author: Syed Adil Shah

Project Duration: May 31 - May 02, 2025

Date of Submission: May, 02, 2025

Project Duration: 3 Days

**1. Introduction**

This project focuses on developing a machine learning model to classify the sentiment of customer reviews

for a fictional product. The sentiment categories include positive, negative, and neutral. The primary goal is to

understand the entire pipeline of sentiment analysis - from data preprocessing to model training and

performance evaluation.

**2. Data**

The dataset consists of product reviews in a CSV file named product\_reviews.csv with columns:

- review\_id - Unique identifier for each review

- review\_text - Text content of the review

- rating - Numerical rating (1 to 5 stars)

Sentiments are derived from the rating as follows:

- 1-2 stars -> Negative

- 3 stars -> Neutral

- 4-5 stars -> Positive

**3. Methodology**

3.1 Data Preprocessing

- Text Cleaning: Removed punctuation, converted text to lowercase, and eliminated non-alphanumeric

characters.

- Tokenization: Split the review text into individual tokens (words) using NLTK's tokenizer.

- Stop-word Removal: Common English stop-words were removed to reduce noise.

- Lemmatization: Words were reduced to their root form to normalize variations.

3.2 Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer was applied to convert preprocessed text

into numerical features, capturing word importance while minimizing common terms' impact.

3.3 Model Training

Three classification algorithms were trained and evaluated:

- Logistic Regression

- Multinomial Naive Bayes

- Support Vector Machine (Linear SVM)

The dataset was split into training and test sets, with models trained on the training data.

3.4 Evaluation Metrics

Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrix.

**4. Results**

- The dataset was balanced with a similar number of positive and negative reviews and fewer neutral reviews.

- Visualization: Sentiment distribution was plotted as a bar chart, and a word cloud was generated to visualize

common terms in reviews.

**Sentiment Analysis of Product Reviews**

- Performance:

\* Logistic Regression achieved an accuracy of ~50% on the test set.

\* Naive Bayes and Linear SVM models showed slightly better accuracy (~67%).

\* Evaluation metrics indicated challenges in correctly classifying neutral sentiment due to data imbalance

and limited sample size.

- The confusion matrices highlighted some misclassifications, particularly for neutral reviews.

**5. Challenges**

- Data Size: The dataset was small, limiting the model's ability to generalize well.

- Class Imbalance: Neutral reviews were fewer, making it harder for the models to learn their characteristics.

- Metric Warnings: Some metrics were ill-defined due to no predicted samples for certain classes, reflecting

data and model limitations.

**6. Potential Improvements**

- Larger Dataset: Collecting or generating more data would improve model robustness.

- Advanced Feature Extraction: Incorporating word embeddings such as Word2Vec or GloVe could capture

semantic relationships better than TF-IDF.

- Hyperparameter Tuning: Fine-tuning model parameters might improve classification performance.

- Additional Models: Testing neural networks or ensemble methods may boost accuracy.

- Handling Imbalance: Techniques like oversampling or class-weighting could address class imbalance.

**7. Conclusion**

This project demonstrated the complete sentiment analysis workflow: preprocessing raw text data, extracting

meaningful features, training multiple classification models, and evaluating their performance. Despite

dataset and class imbalance challenges, models showed promising results. Further improvements could

enhance accuracy and generalization.