DSA 1080A FINAL PROJECT REPORT

Sentiment Analysis on Product Reviews

Student Name: Anas Abdiaziz

Registration Number: 674957

Unit Code: DSA 1080A

Semester: FS2025

Topic: Sentiment Analysis on Product Reviews

---

1. INTRODUCTION

Online customer reviews play a very important role in influencing purchasing decisions. Customers often rely on previous buyers' opinions before deciding whether to purchase a product. However, because businesses receive thousands of reviews daily, analyzing them manually becomes difficult.

This project focuses on building a machine learning system to automatically classify product reviews into **Positive, Negative or Neutral sentiment** using Natural Language Processing (NLP) techniques. The goal is to help businesses understand customer feedback more efficiently and make better decisions based on customer experience.

Sentiment analysis is used in:

- Monitoring customer satisfaction
- Measuring brand perception
- Improving product quality
- Automating feedback analysis

---

2. DATASET DESCRIPTION

The dataset used in this project is the **Amazon Reviews Dataset** obtained from Kaggle. The dataset contains customer reviews submitted on various products sold on Amazon.

Dataset attributes include:

- Review Text
- Sentiment Label (Positive / Negative / Neutral)

Each row in the dataset represents a customer's opinion about a product.

---

3. DATA PREPROCESSING AND CLEANING

The dataset was cleaned and prepared through the following steps:

Handling Missing Data:

- Rows with missing reviews or labels were removed.

Text Cleaning:

The following preprocessing steps were applied:

- Conversion of text to lowercase
- Removal of punctuation and special characters
- Removal of stopwords
- Tokenization

Encoding:

Sentiment labels were encoded into numeric form to allow machine learning processing.

---

4. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory analysis was conducted to identify patterns within the data.

Methods Used:

- Visualizing sentiment distribution
- Checking word frequency
- Analyzing review length

Observations:

- The dataset contained more positive reviews compared to negative and neutral reviews.
- Negative reviews often contained stronger emotional expressions.
- Positive reviews frequently contained words such as *good, great, excellent*.
- Negative reviews commonly used words like *bad, poor, terrible*.

---

5. FEATURE ENGINEERING

Text data was transformed into numerical data using:

TF-IDF Vectorization

This method converts words into numerical weights based on frequency and importance.

Irrelevant columns were removed and only features relevant for classification were used.

---

6. MODEL BUILDING

Four machine learning models were trained and evaluated:

- Logistic Regression
- Naive Bayes
- Linear Support Vector Machine (Linear SVM)
- Random Forest

The dataset was split into:

- 80% training data
- 20% testing data

Each model was trained and tested using the same dataset for fair comparison.

7. MODEL EVALUATION

The models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score

Model Performance Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7951	0.7356	0.7951	0.7380
Naive Bayes	0.7801	0.6534	0.7801	0.7111
Linear SVM	0.7848	0.7245	0.7848	0.7445
Random Forest	0.6452	0.5792	0.6452	0.5258

8. BEST PERFORMING MODEL

The **Linear SVM** model achieved the highest F1-score of **0.7445** and was therefore selected as the best performing model.

This indicates that Linear SVM provided the best balance between precision and recall, particularly for an imbalanced dataset where some sentiment classes had fewer samples.

9. CHALLENGES FACED

Some difficulties encountered included:

- Cleaning unstructured text
- Handling imbalanced data
- Feature extraction from text
- Model selection
- Training time for large datasets

---

10. CONCLUSION

The project successfully developed a sentiment classification model capable of identifying customer emotions from Amazon reviews.

The final model (Linear SVM) demonstrated reliable classification performance and is suitable for practical business analysis tasks such as product feedback evaluation and customer satisfaction monitoring.

---

11. RECOMMENDATIONS

Future improvements may include:

- Using deep learning models (LSTM, Transformers)
- Collecting more training data
- Improving preprocessing techniques
- Deploying the model with Streamlit
- Implementing real-time sentiment prediction

---

12. REFERENCES

- Kaggle Amazon Reviews Dataset
- Scikit-learn Documentation
- Pandas and NumPy Documentation
- NLP with Python Tutorials