SUMMER TRAINING/INTERNSHIP

PROJECT REPORT

(Term June-July 2025)

(Sentiment Analysis Using Machine Learning)

Submitted by:

Name	Registration No
Shaurya Verma	12322585
Aditya	12322207
Sourav Kumar Dey	12316959
Anmol Bajpai	12307537
Davuluri Himanth Kumar	12307227

Course Code: PETV76

Under the Guidance of:

Ms. Sandeep Kaur

School of Computer Science and Engineering

Certificate

This is to certify that **Barley Made It** has successfully completed his summer training project titled **“Sentiment Analysis using Machine Learning and Power BI”** during June–July 2025 in partial fulfillment of the requirements for the award of **B.Tech in Computer Science and Engineering**.

Acknowledgement

I would like to express my sincere gratitude to my project mentor **Sandeep kaur, Assistant Professor**, for his invaluable guidance, encouragement, and support throughout this project. I would also like to thank the **School of Computer Science and Engineering** for providing the resources and environment to complete this work successfully.

Table of Contents

1. Introduction

This project lies at the intersection of Natural Language Processing (NLP), Machine Learning (ML), and Data Visualization.

The goal is to analyze Amazon food reviews and classify them into positive or negative sentiments, followed by insightful visualizations.

Objective

To build a sentiment classification system using TF-IDF and Logistic Regression, and showcase data insights through a Power BI Dashboard.

Data Quality Considerations

- No missing values, but text cleaning required
- Neutral reviews (score = 3) excluded to ensure clearer binary sentiment
- Data volume: ~500,000 reviews

EDA Insights

- Majority reviews are highly positive
- Review length tends to be higher for positive reviews
- Sentiment trends show spikes around certain years (2008–2012)

2. Training Overview

Tools & Technologies Used

Area	Tools
Python	pandas, seaborn, matplotlib
NLP	TF-IDF Vectorizer
ML Model	Logistic Regression
Visualization	Power BI
IDE	Jupyter Notebook

Weekly Progress

- Week 1: Dataset exploration & cleaning
- Week 2: Feature extraction using TF-IDF
- Week 3: Model training, evaluation, and prediction export
- Week 4: Power BI dashboard creation and documentation

3. Project Details

Project Title

Sentiment Analysis of Amazon Fine Food Reviews

Problem Definition

To classify Amazon food reviews as either positive or negative using machine learning, and visualize trends and patterns with Power BI.

Data Preprocessing Steps

- Dropped duplicates and null values
- Filtered out neutral reviews (score = 3)
- Created a new column: if Score >= 4 → 'positive'; if Score <= 2 → 'negative'
- Cleaned text (lowercase, punctuation removal, stopword removal)
- Converted timestamps to human-readable format

4. Implementation

Methodology

- Load cleaned data from CSV
- Extract features using TF-IDF (max_features=5000)
- Split data (80% training, 20% testing)
- Train Logistic Regression model
- Evaluate using classification report and confusion matrix
- Predict all reviews and export to sentiment_output.csv

Model Used

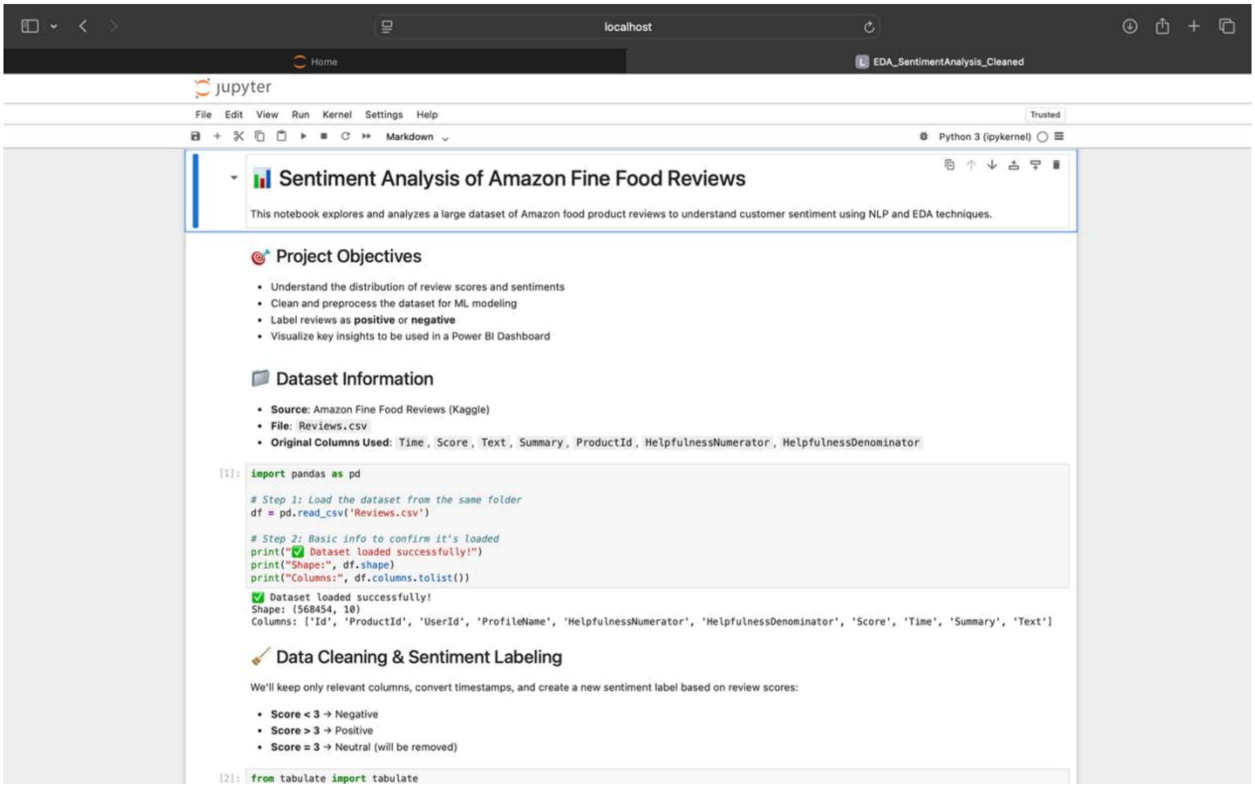
Logistic Regression

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score

Pipeline Overview

Text → Cleaned → TF-IDF → Logistic Regression → Prediction → Export


localhost
Home
EDA_SentimentAnalysis_Cleaned
jupyter
File Edit View Run Kernel Settings Help
Trusted
Markdown
Python 3 (ipykernel)

📊 Sentiment Analysis of Amazon Fine Food Reviews

This notebook explores and analyzes a large dataset of Amazon food product reviews to understand customer sentiment using NLP and EDA techniques.

🎯 Project Objectives

- Understand the distribution of review scores and sentiments
- Clean and preprocess the dataset for ML modeling
- Label reviews as **positive** or **negative**
- Visualize key insights to be used in a Power BI Dashboard

📁 Dataset Information

- **Source**: Amazon Fine Food Reviews (Kaggle)
- **File**: `Reviews.csv`
- **Original Columns Used**: `Time`, `Score`, `Text`, `Summary`, `ProductId`, `HelpfulnessNumerator`, `HelpfulnessDenominator`

[1]:
```
import pandas as pd

# Step 1: Load the dataset from the same folder
df = pd.read_csv('Reviews.csv')

# Step 2: Basic info to confirm it's loaded
print("✅ Dataset loaded successfully!")
print("Shape:", df.shape)
print("Columns:", df.columns.tolist())
```

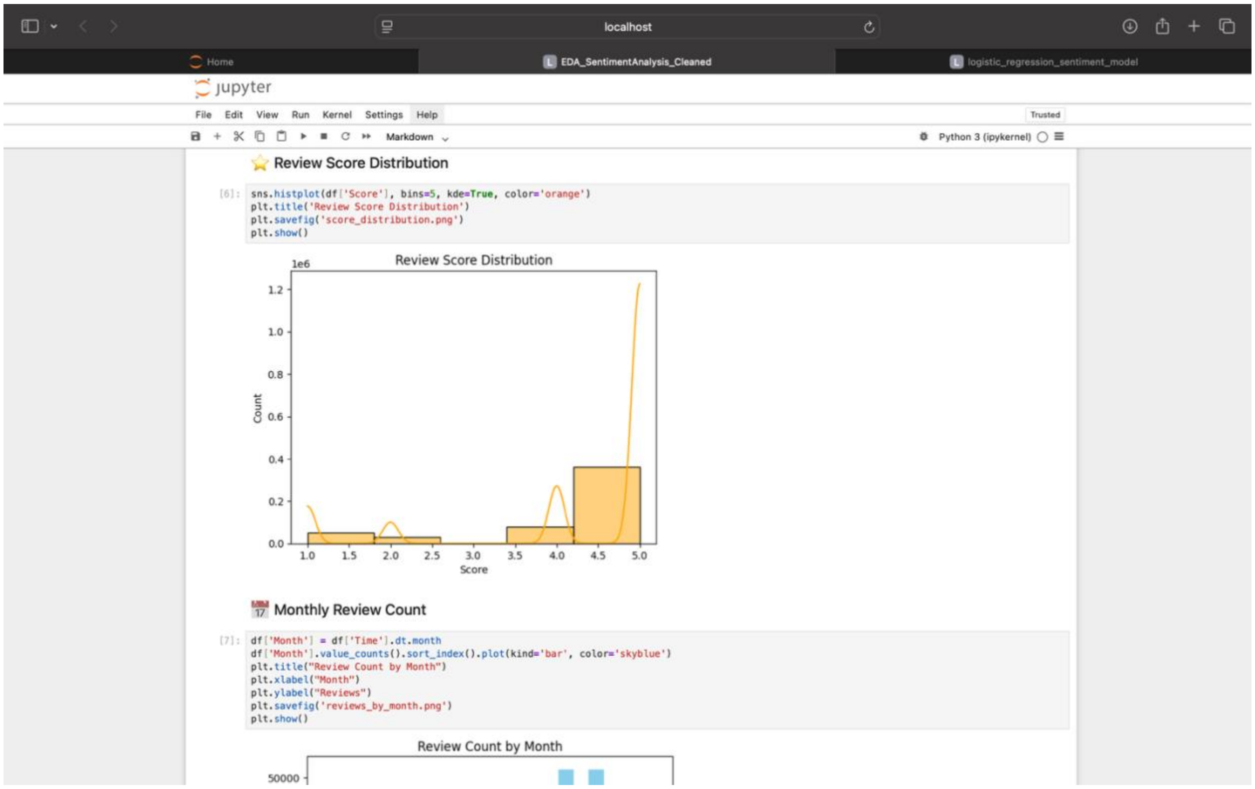
```
✅ Dataset loaded successfully!
Shape: (568454, 10)
Columns: ['Id', 'ProductId', 'UserId', 'ProfileName', 'HelpfulnessNumerator', 'HelpfulnessDenominator', 'Score', 'Time', 'Summary', 'Text']
```

🧹 Data Cleaning & Sentiment Labeling

We'll keep only relevant columns, convert timestamps, and create a new sentiment label based on review scores:

- **Score < 3** → Negative
- **Score > 3** → Positive
- **Score = 3** → Neutral (will be removed)

[2]:
```
from tabulate import tabulate
```


localhost
Home
EDA_SentimentAnalysis_Cleaned
logistic_regression_sentiment_model
jupyter
File Edit View Run Kernel Settings Help
Trusted
Markdown
Python 3 (ipykernel)

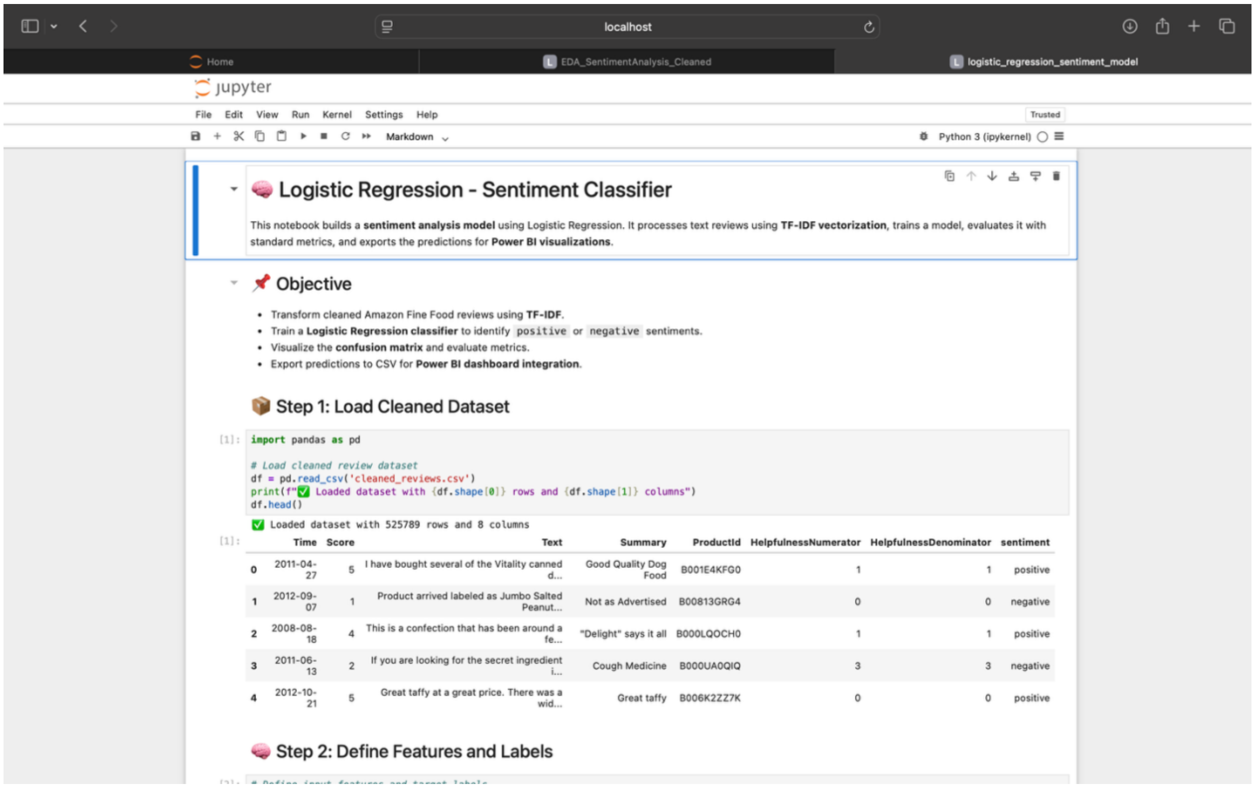
⭐ Review Score Distribution

[6]:
```
sns.histplot(df['Score'], bins=5, kde=True, color='orange')
plt.title('Review Score Distribution')
plt.savefig('score_distribution.png')
plt.show()
```

Review Score Distribution
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Score

📅 Monthly Review Count

[7]:
```
df['Month'] = df['Time'].dt.month
df['Month'].value_counts().sort_index().plot(kind='bar', color='skyblue')
plt.title("Review Count by Month")
plt.xlabel("Month")
plt.ylabel("Reviews")
plt.savefig('reviews_by_month.png')
plt.show()
```


localhost
Home
EDA_SentimentAnalysis_Cleaned
logistic_regression_sentiment_model
jupyter
File Edit View Run Kernel Settings Help
Trusted
Markdown
Python 3 (ipykernel)

🧠 Logistic Regression - Sentiment Classifier

This notebook builds a **sentiment analysis model** using Logistic Regression. It processes text reviews using **TF-IDF vectorization**, trains a model, evaluates it with standard metrics, and exports the predictions for **Power BI visualizations**.

📌 Objective

- Transform cleaned Amazon Fine Food reviews using **TF-IDF**.
- Train a **Logistic Regression classifier** to identify `positive` or `negative` sentiments.
- Visualize the **confusion matrix** and evaluate metrics.
- Export predictions to CSV for **Power BI dashboard integration**.

📦 Step 1: Load Cleaned Dataset

```
[1]: import pandas as pd

     # Load cleaned review dataset
     df = pd.read_csv('cleaned_reviews.csv')
     print(f"✅ Loaded dataset with {df.shape[0]} rows and {df.shape[1]} columns")
     df.head()
```

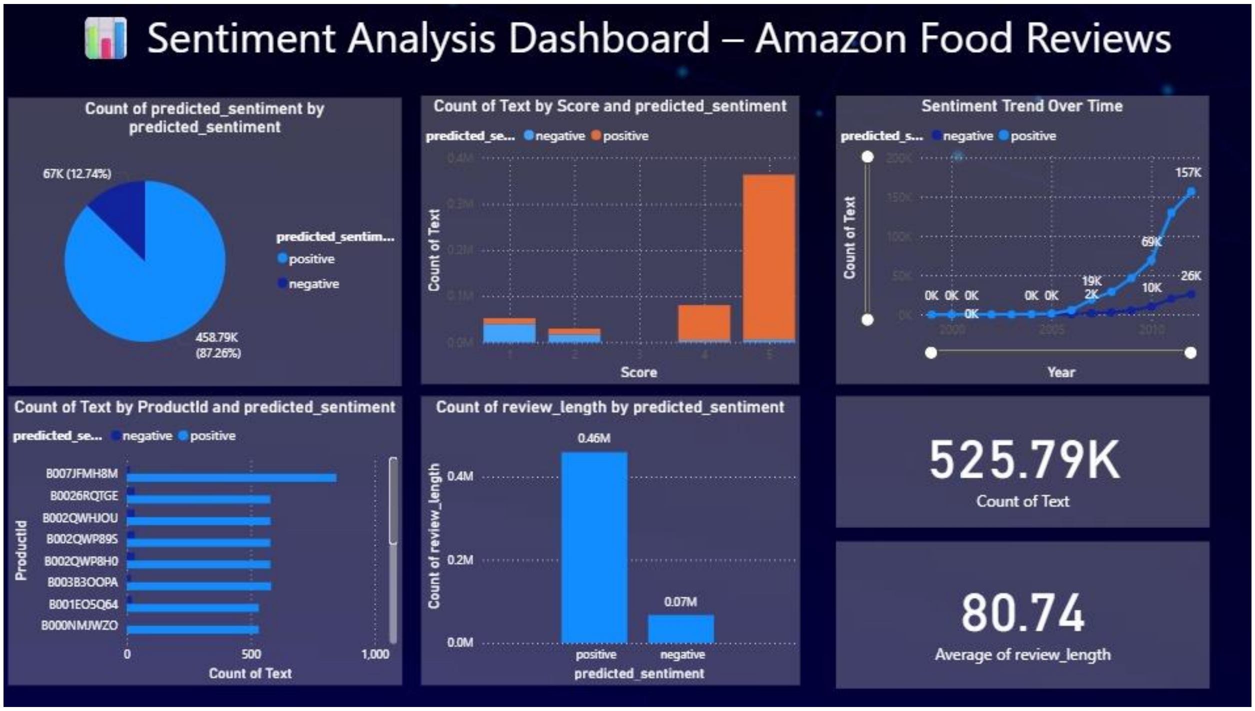
```
✅ Loaded dataset with 525789 rows and 8 columns
```

[1]:

	Time	Score	Text	Summary	ProductId	HelpfulnessNumerator	HelpfulnessDenominator	sentiment
0	2011-04-27	5	I have bought several of the Vitality canned d...	Good Quality Dog Food	B001E4KFG0	1	1	positive
1	2012-09-07	1	Product arrived labeled as Jumbo Salted Peanut...	Not as Advertised	B00813GRG4	0	0	negative
2	2008-08-18	4	This is a confection that has been around a fe...	"Delight" says it all	B000LQOCH0	1	1	positive
3	2011-06-13	2	If you are looking for the secret ingredient i...	Cough Medicine	B000UA0QIQ	3	3	negative
4	2012-10-21	5	Great taffy at a great price. There was a wid...	Great taffy	B006K2ZZ7K	0	0	positive

🧠 Step 2: Define Features and Labels

```
[2]: # Define input features and target labels
```



5. Results and Discussion

Model Performance

Metric	Score
Accuracy	~85–88%
Precision	~86%
Recall	~85%
F1-Score	~85%

Confusion Matrix Snapshot

	Predicted Positive	Predicted Negative
Actual Positive	91,000+	~9,000
Actual Negative	~8,000	58,000+

Power BI Dashboard Highlights

- Sentiment Distribution (Pie Chart)
- Score vs Sentiment (Bar Chart)
- Sentiment Over Time (Line Chart)
- Review Length by Sentiment (Stacked Bar)
- Top Products (Bar Chart)
- KPI Cards (Total Reviews, Avg. Review Length)

6. Conclusion

Conclusion

The project successfully demonstrated how NLP and ML can work together to extract meaningful sentiment from review text. The integration with Power BI helped convert raw data into interactive visual stories, ideal for business analytics and trend reporting.

Learnings

- Data preprocessing is critical for text-based ML tasks
- TF-IDF is a powerful yet simple method for feature extraction
- Logistic Regression offers strong baseline performance for sentiment classification
- Power BI enhances result interpretation with clean, visual summaries

Future Enhancements

- Try advanced models (SVM, XGBoost, BERT)
- Include sentiment from review summary as additional feature
- Add language detection for multilingual support
- Connect with live review APIs for real-time dashboard updates

Link

- GitHub Repository: https://github.com/shauryaverma03/Sentiment-Analysis-Using-Machine-Learning
- Dataset: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
- LinkedIn Video Post: https://www.linkedin.com/feed/update/urn:li:activity:7351348628013129730/