Report: Automated Customer Reviews Classification

Authors: Ginosca Alejandro Dávila & Natanael Santiago Morales
Ironhack Data Science and Machine Learning Bootcamp

---

1. Introduction

This project explores the classification of Amazon customer reviews into three sentiment categories: Positive, Neutral, or Negative. Our goal is to compare **Traditional Machine Learning (ML)** approaches against **Transformer-based models** to determine which method yields better performance in analyzing customer sentiment from textual reviews.

In this report, we present two approaches:

- **Traditional NLP & ML Approach**
- **Transformer Approach (HuggingFace)**

---

2. Problem Statement

Amazon receives thousands of customer reviews daily, making it difficult to manually categorize and analyze the sentiments expressed in the reviews. The company aims to build an automated system that classifies reviews as **Positive**, **Neutral**, or **Negative** to gain real-time insights into customer sentiment and feedback. This system is crucial for reviews without star ratings and for interpreting different user expectations.

---

3. Data Description

We used the **Consumer Reviews of Amazon Products** dataset, which contains thousands of reviews with corresponding star ratings (1-5) and product metadata. We transformed the star ratings as follows:

- **1, 2, 3 stars → Negative**
- **4 stars → Neutral**
- **5 stars → Positive**

After preprocessing, the dataset contained two key columns: the text (merged from the review title and review body) and the corresponding sentiment label.

---

4. Traditional NLP & ML Approach

4.1 Data Preprocessing

We performed the following preprocessing steps:

- **Text Cleaning**: Removed special characters, punctuation, and whitespace.
- **Tokenization and Lemmatization**: Broke down the text into tokens and applied lemmatization to convert words into their base forms.
- **TF-IDF Vectorization**: Transformed the cleaned text into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique.

4.2 Model Selection and Training

We explored the following ML algorithms:

- **Naive Bayes**: A simple baseline model.
- **Logistic Regression**: A strong linear model.
- **Support Vector Machines (SVM)**: Effective for high-dimensional text data.
- **Random Forest and XGBoost**: Ensemble methods that capture non-linear relationships.

Each model was evaluated using the following metrics: **Accuracy**, **Precision**, **Recall**, **F1-score**, and the **Confusion Matrix**.

4.3 Results

The **Logistic Regression** model outperformed the others with an **accuracy of 76.91%**. Below is a summary of the evaluation metrics for Logistic Regression:

Sentiment Class	Precision	Recall	F1-Score
Negative	86.40%	53.33%	65.95%
Neutral	55.61%	15.64%	24.41%
Positive	77.46%	96.93%	86.11%

Although the model performed well for **Positive** reviews, it struggled with the **Neutral** class.

---

5. Transformer Approach (HuggingFace)

5.1 Preprocessing

For the Transformer approach, we avoided heavy text cleaning to preserve the context and format of the input. Instead, we used **BERT's tokenizer** to split the text into tokens and encoded the input for model training.

5.2 Model Selection and Training

We used several Transformer models from HuggingFace, including **BERT base uncased** and **BERT base multilingual uncased (sentiment)**. Each model was fine-tuned on the training data using **transfer learning**.

The training parameters included:

- **Batch Size**: 16 for training, 32 for evaluation.
- **Learning Rate**: 2e-5.
- **Epochs**: 5.

We trained the model using a PyTorch-based Trainer and evaluated it using accuracy and F1-score metrics.

5.3 Results

The **BERT base multilingual uncased (sentiment) model** achieved **78.0% accuracy** after fine-tuning, with the following evaluation results:

Sentiment Class	Precision	Recall	F1-Score
Negative	81.00%	58.00%	68.00%
Neutral	57.00%	45.00%	50.00%
Positive	83.00%	91.00%	87.00%

This model performed better than the base (pre-trained) model and showed competitive performance compared to traditional ML models, particularly in classifying the **Neutral** class.

---

6. Conclusion

- **Logistic Regression** emerged as the best-performing traditional ML model.

- **BERT base multilingual uncased (sentiment)** outperformed the traditional models in classifying Neutral reviews and achieved an overall better performance.

The **Transformer-based approach** demonstrated the ability to understand sentiment nuances better, making it a superior choice for customer review classification.