Sentiment Analysis and Insights from E-Commerce Reviews

Data Science Project

Siddharth Acharya[HCE080BCT037]

Sneha Rawal[HCE080BCT040]

🗓️ **19th February 2025**

Introduction and Objective

Introduction

Customer reviews play a crucial role in shaping purchasing decisions in e-commerce. Analyzing these reviews can provide valuable insights into customer sentiment, product performance, and areas for improvement. This project focuses on sentiment analysis of women's clothing e-commerce reviews to extract meaningful patterns and trends.

Objective

The goal of this project is to analyze customer reviews using Natural Language Processing (NLP) and machine learning techniques to:

- **Perform sentiment classification** (positive, neutral, negative).
- **Identify key themes and topics** discussed in reviews.
- **Detect trends in customer feedback** over time.
- **Provide data-driven insights** to improve product offerings and customer satisfaction.

Motivation and Background

Motivation

Customer reviews are a vital source of information for both businesses and consumers. Understanding customer sentiment helps businesses:

- Improve product quality and customer satisfaction.
- Enhance marketing strategies based on customer feedback.
- Identify key pain points and areas for improvement.

With the rise of e-commerce, analyzing vast amounts of textual data manually is inefficient. Automating this process through **Natural Language Processing (NLP)** and **Machine Learning (ML)** allows businesses to extract valuable insights efficiently.

Background

E-commerce platforms generate large volumes of customer reviews daily. Traditional methods of analyzing feedback, such as surveys and focus groups, are time-consuming and limited in scope. Advances in **text mining**, **sentiment analysis**, and **topic modeling** enable more efficient and scalable ways to process and interpret customer feedback.

This project focuses on applying **EDA (Exploratory Data Analysis), NLP techniques, and machine learning models** to analyze women's clothing reviews and derive actionable insights.

Data Description

Data Source

The dataset used in this project is the **Women's Clothing E-Commerce Reviews** dataset. It contains customer reviews and related metadata collected from an online retail platform.

Type of Data

- **Structured Data:** Includes numerical and categorical features such as age, clothing category, and review ratings.
- **Unstructured Data:** The primary textual data consists of customer-written reviews, which require **Natural Language Processing (NLP)** techniques for analysis.

Key Features

1. **Review Text** – Customer-provided textual feedback.
2. **Rating** – A numerical score (1-5) indicating customer satisfaction.
3. **Recommended IND** – A binary indicator (1 = recommended, 0 = not recommended).
4. **Age** – Customer age.
5. **Division Name** – High-level category of clothing (e.g., General, Petite).
6. **Department Name** – Department classification (e.g., Tops, Bottoms).
7. **Class Name** – Specific product type (e.g., Dresses, Sweaters).

This dataset enables sentiment analysis, trend identification, and customer behavior insights to improve product recommendations and marketing strategies.

Methodology

1. Data Wrangling and ETL Processes

- **Data Cleaning:** Handled missing values, removed duplicates, and corrected inconsistencies.
- **Data Transformation:** Converted categorical variables into numerical format.
- **Text Preprocessing:** Tokenization, lowercasing, stopword removal, stemming, and lemmatization were applied to process review text.
- **ETL (Extract, Transform, Load):** Extracted data from CSV, transformed it for analysis, and loaded it into a structured format.

2. Feature Extraction and Selection Techniques

- **TF-IDF (Term Frequency-Inverse Document Frequency):** Extracted important words for NLP analysis.
- **Word Embeddings (Word2Vec, GloVe):** Captured contextual meanings of words.
- **Sentiment Scores:** Computed sentiment polarity of reviews.
- **Numerical & Categorical Features:** Selected relevant structured data attributes (e.g., rating, age, recommendation indicator).

3. Algorithms/Models Used

- **Exploratory Data Analysis (EDA):** Used visualization and statistical analysis to understand patterns.
- **Machine Learning Models:**
  - **Logistic Regression & Random Forest:** For sentiment classification.
  - **Naïve Bayes:** Applied for text-based classification.
  - **Support Vector Machine (SVM):** Used for sentiment analysis.
- **Deep Learning Models:**
  - **LSTM (Long Short-Term Memory):** Used for sentiment prediction based on sequential text.
  - **BERT (Bidirectional Encoder Representations from Transformers):** Applied for advanced NLP tasks.

This methodology ensures effective data preprocessing, feature engineering, and model selection for optimal analysis and predictions.

Exploratory Data Analysis (EDA)

1. Key Insights from EDA

- **Review Length Distribution:** Most reviews have a moderate length, with some outliers having very long text.
- **Sentiment Distribution:** A significant portion of the reviews are positive, with fewer negative reviews.
- **Rating Trends:** Higher ratings (4 & 5) are more common, indicating overall customer satisfaction.
- **Age Group Analysis:** Certain age groups, particularly 30-40, contribute the most reviews.
- **Recommendation Impact:** Customers who recommend a product tend to give higher ratings.
- **Frequent Words in Reviews:** Common words include "love," "fit," and "comfortable," indicating common themes.

2. Visualizations to Support Insights

- **Histogram of Review Lengths:** Shows the distribution of text lengths.
- **Word Cloud:** Highlights frequently used words in positive and negative reviews.
- **Sentiment Distribution Bar Chart:** Displays the proportion of positive, neutral, and negative sentiments.
- **Boxplot of Ratings by Age Group:** Analyzes how ratings vary across different age groups.
- **Heatmap of Feature Correlations:** Identifies relationships between rating, recommendation, and sentiment scores.

These insights help in understanding customer feedback trends, identifying key factors influencing reviews, and guiding model development.

Results and Evaluation

1. Model Performance Metrics

- **Classification Metrics:**
  - **Accuracy:** Measures overall correctness of predictions.
  - **Precision:** Proportion of true positive predictions out of all positive predictions.
  - **Recall:** Measures how well the model captures actual positives.
  - **F1-Score:** Harmonic mean of precision and recall for balanced evaluation.
  - **ROC Curve & AUC:** Evaluates the model's ability to distinguish between classes.
- **Clustering Metrics (if applicable):**
  - **Silhouette Score:** Measures how well clusters are defined.
  - **Inertia:** Evaluates clustering compactness.
  - **Davies-Bouldin Index:** Assesses cluster separation.

2. Comparison of Different Models

- **Baseline Model:** Performance of a simple model (e.g., Logistic Regression, Naïve Bayes).
- **Advanced Models:** Performance of more complex models (e.g., Random Forest, XGBoost, LSTM for NLP).
- **Comparison Table:**
  - Displays accuracy, precision, recall, and F1-score for each model.
  - Highlights trade-offs between interpretability and performance.

3. Key Findings

- Best-performing model based on selected evaluation metrics.
- Strengths and weaknesses of different approaches.
- Insights on misclassified cases and potential improvements.

These evaluations guide model selection, interpretability, and improvements for better decision-making.

Conclusion and Future Work

1. Summary of Findings

- Key insights derived from **Exploratory Data Analysis (EDA)** and modeling.
- Best-performing model(s) based on **evaluation metrics** (e.g., accuracy, precision, recall).
- Notable trends and patterns in the dataset that influenced predictions.

2. Limitations of the Study

- **Data Limitations:** Quality, bias, or missing values affecting model performance.
- **Model Limitations:** Trade-offs between complexity and interpretability.
- **Computational Constraints:** Challenges in processing large datasets efficiently.
- **Generalization Issues:** Model performance on unseen data and real-world applications.

3. Future Work

- **Data Enhancements:** Collecting more diverse or higher-quality data.
- **Feature Engineering:** Exploring additional features for better prediction accuracy.
- **Advanced Models:** Implementing deep learning or ensemble methods for improved results.
- **Real-World Deployment:** Developing an API or dashboard for practical use.
- **Hyperparameter Tuning:** Further optimizing models for better generalization.

By addressing these limitations and expanding the scope, the project can evolve into a more robust and scalable solution.