CSE508_Winter2024_A3_2021311

Project Report – Ankit Raj

Amazon Review Electronics Product Analysis

This report offers a comprehensive analysis of Amazon review data for electronics products, focusing specifically on headphones. It delves into data exploration, preprocessing, sentiment analysis, machine learning modeling, and collaborative filtering techniques. The report aims to provide valuable insights into customer preferences and product recommendations.

1. Introduction

This analysis investigates customer experiences with headphones on Amazon. By exploring and modeling review data, we aim to:

- Understand user sentiment and product perception through descriptive statistics and exploratory data analysis (EDA).
- Leverage text processing and machine learning models to predict rating classes.
- Develop recommender systems using collaborative filtering techniques to personalize product suggestions.

2. Dataset Description

The analysis utilizes the Amazon Reviews Dataset, specifically the 5-core dataset for the Electronics category. This dataset features information about product reviews, including reviewer ID, product ID (ASIN), review text, rating, and additional metadata.

3. Data Preprocessing

- **Loading:** Pandas DataFrames are used to load the dataset from provided file paths.
- **Cleaning:** Missing values and duplicate rows are addressed appropriately.
- **Filtering:** The dataset is filtered to include only headphone products based on a relevant identifier (e.g., title column).

4. Descriptive Statistics

- **Total Reviews:** The total number of reviews for headphones after preprocessing is calculated.
- **Average Rating:** The average customer rating for headphones is determined.
- **Product Variety:** The number of unique headphone products within the dataset is identified.
- **Rating Distribution:** Ratings are categorized as 'Good' (>= 3) and 'Bad' (< 3). The counts for each category and the distribution of ratings across all reviews are analyzed.

5. Text Preprocessing

- **HTML Removal:** HTML tags are removed from the review text for cleaner analysis.
- **Normalization:** Accented characters are transformed into their standard ASCII counterparts.
- **Acronym Expansion:** Common acronyms are converted to their full forms for improved understanding.
- **Special Character Removal:** Special characters not essential for sentiment analysis are removed.
- **Lemmatization:** Words are reduced to their base forms (e.g., "running" becomes "run") to capture root meaning.

6. Exploratory Data Analysis (EDA)

- **Brand Analysis:** The top 20 most and least reviewed headphone brands are identified to understand market presence.
- **Top-Rated Headphones:** The headphone product with the highest average rating is determined, indicating customer preference.
- **Rating Trends:** Counts of ratings across 5 consecutive years are analyzed to identify trends in popularity or sentiment.
- **Sentiment Visualization:** Word clouds are generated for 'Good' and 'Bad' ratings, highlighting frequently used words in positive and negative reviews.
- **Rating Distribution:** A pie chart is created to depict the distribution of

ratings compared to the number of reviews for each rating value.

7. Feature Engineering and Modelling

- **Feature Engineering:** Text data is transformed into numerical features using techniques like Bag-of-Words (BoW) or TF-IDF vectorization.
- **Rating Class Encoding:** Ratings are converted into categorical variables representing 'Good', 'Average', and 'Bad'.
- **Model Selection and Evaluation:** Five machine learning models (e.g., Logistic Regression, Random Forest) are trained and evaluated on the processed data. Performance metrics like precision, recall, F1-score, and support are used to assess the effectiveness of each model in predicting rating classes.

8. Collaborative Filtering

- **Recommendation Systems:** Two collaborative filtering approaches are implemented: User-User and Item-Item.
- **Evaluation:** Mean Absolute Error (MAE) is calculated for different values of N (number of nearest neighbours considered) for both user-user and item-item recommender systems. Lower MAE indicates better prediction accuracy.

9. Top 10 Products by User Sum Ratings

The top 10 headphone products are identified based on the total number of ratings received from users. This provides insights into the most popular headphones within the dataset.

10. Conclusion

This analysis delves into the world of customer reviews for headphones on Amazon. Through data exploration, preprocessing, text processing, machine learning, and collaborative filtering, we gain valuable insights into user sentiment, product perception, and recommendation opportunities. This knowledge can be leveraged to enhance customer experience by providing personalized product suggestions and improving overall satisfaction.