Project Report

Overview

This project aimed to develop a comprehensive recommendation system for electronic products based on Amazon reviews. We focused on predicting user ratings, evaluating review usefulness, and identifying top products and brands. The project involved data preprocessing, analysis, collaborative filtering, and machine learning (ML) models to achieve these objectives.

Methodologies

Data Preprocessing

- Data Parsing: Extracted relevant information from JSON files containing reviews and metadata.
- Text Preprocessing: Cleaned review texts, removing HTML tags, special characters, and applying lemmatization to standardize text data.

Collaborative Filtering

- User-Item Rating Matrix: Created matrices for both user-user and item-item recommendation systems, normalizing ratings using Min-Max scaling.
- Similarity Computation: Calculated cosine similarities among users and items to identify relationships based on review patterns.
- K-Folds Validation: Applied 5-fold cross-validation to both systems, predicting missing ratings and calculating Mean Absolute Error (MAE) for model evaluation.

Machine Learning Models

- Model Selection: Evaluated five ML models: Naive Bayes, Linear SVM, Random Forest, Gradient Boosting, and an undisclosed model.

- Evaluation Metrics: Focused on precision, recall, F1-score, and support for three rating categories: Good, Average, and Bad.

Assumptions

- Reviews containing the keyword "Headphones" were specifically analyzed to tailor recommendations within this category.
- Ratings >= 3 were considered good, =3 average, and <= 2 bad, to categorize review sentiments.
- A review was deemed useful if it contributed positively to the accuracy of the recommendation system.

Results

Collaborative Filtering

- Item-item and user-user collaborative filtering approaches were developed, with MAE used to assess accuracy. Detailed MAE values for varying numbers of neighbors (N=10, 20, 30, 40, 50) are omitted for brevity but were instrumental in tuning the models.

Machine Learning Models

- The Logistic Regression model demonstrated the highest accuracy (0.827) and balanced performance across rating categories.
- Linear SVM and Random Forest showed promising results, particularly in correctly predicting good ratings.
- Naive Bayes struggled with average ratings, showing a preference for predicting reviews as good.
- Gradient Boosting had a lower overall accuracy (0.798) but showed a relatively balanced performance across different metrics.

Top Products and Brands

- Analysis identified the top 10 most and least reviewed brands, providing insights into brand popularity and market presence.
- The most positively reviewed headphone (based on ASIN) was highlighted, showcasing a product that stood out in the headphones category.

Conclusion

The project successfully leveraged collaborative filtering and machine learning techniques to analyze Amazon reviews for electronic products, specifically headphones. The combination of user-user and item-item recommendation systems, alongside sophisticated ML models, provided a robust framework for predicting ratings and evaluating review usefulness. Identifying top products and brands further enriched the analysis, offering valuable insights for both consumers and manufacturers.

Future work could explore deeper text analysis, integrate more sophisticated NLP techniques for feature extraction, and expand the recommendation system to encompass a wider range of products and review characteristics.