

Beauty Product Reviews Dataset

A well structured revolutionary dataset for Sentiment Analysis & Recommendation System Research

A comprehensive AI-generated dataset of 50,000 reviews from 1,000 users across 200 beauty products, designed for machine learning and natural language processing research.



Research Team & Affiliations

Authors

- Mujahidul Islam
- Nafiz Khan Tasnul
- Korobi Sarkar
- Noor Shat Zahan

Institution

East West University, Dhaka,
Bangladesh

Contact: Mimujahid6@gmail.com



Dataset Overview

50K

Total Reviews

Comprehensive product
feedback

1K

Unique Users

Diverse customer base

200

Products

Beauty product variety

21

Attributes

Rich data dimensions

Synthetic dataset generated using AI, featuring user history, CTR, spend time, ratings, and sentiment classifications across three categories.

Key Dataset Features

User Data

ID, age range, gender, location, recommendation history, average ratings

Product Information

Name, ID, type, price, discount, average rating

Review Content

Title, description, date, packaging quality, sentiment classification

Engagement Metrics

CTR, spend time on product, user activity levels



Data Collection & Generation Process

01

Planning & Research

Analyzed beauty e-commerce websites and created surveys to identify key purchase factors.

02

Column Selection

Selected 21 essential attributes including user ID, product code, ratings, and packaging quality.

03

AI Generation

Used ChatGPT with specific prompts (pricing in USD, 5-10% discounts, multi-country locations) to generate synthetic data.

04

Finalization

Validated dataset for missing values, outliers, and quality before preprocessing and publication.

Data Preparation & Cleaning

Missing Values

- Ratings: filled with mean values
- Titles: generated from ratings
- Reviews: created using rating-based templates

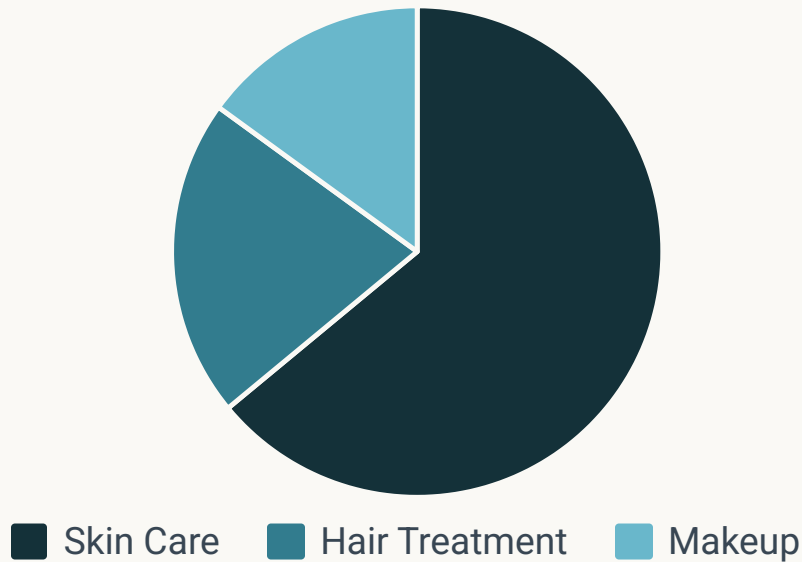
Data Cleaning

- Converted text to lowercase
- Removed special characters
- Trimmed whitespace

Sentiment Analysis

Applied Shaver et al's emotion model to classify reviews into three sentiment classes: positive (happy, love), neutral (no reaction), and negative (anger, fear, sadness).

Product Distribution & Sales Analysis



Market Insights

Skin Care dominates with 64% (31,905 units), indicating strong customer preference. Hair Treatment holds 21% (10,596 units), while Makeup represents 15% (7,499 units), presenting growth opportunities.

Geographic & Temporal Trends

Top 5 Cities by Sales

Among 25,333 users:

- **São Paulo:** 5,095 users
- **Berlin:** 5,093 users
- **Paris:** 5,063 users
- **Cape Town:** 5,059 users
- **Tokyo:** 5,023 users

Monthly Performance

October shows peak sales and revenue. February records lowest activity. May, July, and December show revenue spikes, indicating seasonal demand fluctuations and consistent pricing throughout the year.





Research Applications & Value

Sentiment Classification

Fine-grained analysis of customer emotions and product feature preferences.

Recommender Systems

Develop collaborative filtering and aspect-based hybrid systems using user history and CTR data.

NLP Algorithms

Train machine learning models for text classification and sentiment prediction.

Market Analysis

Understand consumer preferences, product design improvements, and targeted marketing strategies.



Dataset Impact & Future Research

Privacy-Protected Innovation: First comprehensive beauty product review dataset with user CTR and spend time metrics, enabling researchers to develop advanced recommender systems while maintaining user confidentiality through synthetic data generation.

→ Unique Contribution

Combines sentiment analysis with engagement metrics (CTR, spend time) rarely available together.

→ Research Potential

Supports aspect-oriented analysis, temporal sentiment tracking, and cross-platform consumer behavior studies.

→ Industry Application

Enables beauty brands to refine products, develop focused marketing campaigns, and improve customer satisfaction.

Thank you!!