Mining Beauty Product Reviews: Sentiment, Demographics, and Patterns of Consumer Preference

1st Liu
Department of Data Science
Florida Polytechnic University,
Lakeland, Florida

2nd Majdoch
Department of Data Science
Florida Polytechnic University
Lakeland, Florida

3rd Thibodeau
Department of Data Science
Florida Polytechnic University
Lakeland, Florida

***Abstract*— The goal of this project is to integrate structured product data with unstructured customer reviews to uncover patterns in consumer preferences, analyze the alignment between ratings and sentiments, and build models that explain or predict product success and customer satisfaction.**

Keywords*— Exploratory Data Analysis, Sentiment Analysis, Association Mining, Text Mining, Customer Satisfaction, and Data Analysis (keywords***)**

I. Introduction

With the age of consumerism at an all-time high, but the economy in economic decline, it is more important now than ever to understand the product you are looking to buy. Product reviews can make or break a person's mind when it comes to purchasing a product, especially considering there are more than likely multiple similar products that have the same effect. The beauty world is no exception to this. According to McKinsey and Company, the global beauty market reached 450 billion dollars in 2024. Making it one of the biggest markets globally for consumerism. With this comes oversaturation of products, trends, and reviews, but a field of rich content for analysis.

The Beauty Industry provides a particularly rich context for studying consumer feedback. Products in this sector are highly subjective, with satisfaction influenced not only by measurable attributes such as price and availability but also by personal factors such as skin type, tone, and preferences. This gives for deeper reviews with more nuanced meanings, making for a melting pot of analysis and research questions. The dataset "Sephora Products and Skincare Reviews" gives around 8 thousand products and over 1 million user reviews from the Skincare category. This dataset will allow for a deep and meaningful analysis that works towards our goal of the paper.

This paper looks to take a deep dive into these reviews and identify meaningful patterns and relationships. The first method that will be used is Exploratory Data Analysis. It will be used to examine trends, find reviewer demographics, and find patterns in the reviews to find the truth about the product. The next approach that will be utilized is Sentiment Analysis. Sentiment analysis will be applied to customer reviews to assess whether written opinions align with numeric ratings and recommendation indicators. Finally, association rule mining will be used to identify frequent and co-occurring attributes to see the meaning between product descriptors, customer demographics, and satisfaction outcomes.

The goal of this project is to integrate structured product data with unstructured customer reviews to uncover patterns in consumer preferences, analyze the alignment between ratings and sentiments, and build models that explain or predict product success and customer satisfaction. By doing so, this research can contribute to understanding the user experience in the beauty sector and demonstrate how Data and Text mining concepts can help in that decision-making.

II. Related Work

A. Exploratory Data Analysis and Data Mining on Yelp Restaurant Review

Text mining and sentiment analysis of consumer reviews have been widely applied in different domains such as restaurants, e-commerce, and mobile applications. These studies demonstrate the importance of analyzing unstructured textual data to uncover patterns in consumer behavior and improve decision-making.

One relevant study is "Exploratory Data Analysis and Data Mining on Yelp Restaurant Review". The authors applied exploratory data analysis and data mining techniques to the Yelp restaurant review dataset. Their work focused on understanding the distribution of reviews, identifying frequent words across different sentiment categories, and analyzing the relationships between customer ratings and textual feedback. Specifically, they employed methods such as temporal and spatial analysis, Bag-of-Words representation, and topic identification to highlight key features that impact the customer experience.

This research provides a methodological foundation that can be extended to other review datasets. While the Yelp study centered on restaurant services, our project adapts similar techniques to the Sephora Product Reviews dataset, which contains customer opinions on beauty and skincare products. By

leveraging comparable EDA methods, sentiment classification, and keyword analysis, our project aims to extract consumer preferences and identify the factors that drive positive and negative perceptions in the beauty domain.

B. *Sentiment Analysis of Beauty Product Applications using the Naïve Bayes Method*

Sentiment analysis of product reviews has become an important area of research, as online reviews play a critical role in shaping consumer purchasing behavior. Previous studies have demonstrated the usefulness of machine learning techniques for classifying consumer opinions into positive, negative, or neutral categories.

A recent study by Rambe, Hasibuan, and Dar (2023), titled "Sentiment Analysis of Beauty Product Applications using the Naïve Bayes Method", applied the Naïve Bayes algorithm to analyze consumer reviews of beauty products. The authors divided reviews into three sentiment classes (positive, negative, and neutral) and combined text preprocessing steps such as TF-IDF feature extraction with classification modeling. Their experiments showed that the Naïve Bayes model achieved an accuracy of 90.08% using a training-test split of 90:10, demonstrating the efficiency of this relatively simple algorithm for sentiment classification tasks. The study also analyzed word frequency distributions, identifying "application," "product," and "price" as some of the most frequent terms, thereby highlighting key features influencing consumer sentiment.

This research provides valuable insights for our project, as it illustrates how supervised machine learning methods such as Naïve Bayes can be applied effectively in the beauty product domain. While their study focused on online shopping applications, we extend a similar methodology to the Sephora Product Reviews dataset, which includes both textual reviews and structured product features. By adopting comparable preprocessing techniques and baseline sentiment models, our project aims to validate and expand on their findings while also exploring additional methods such as logistic regression, SVM, or deep learning to improve predictive performance.

C. *Customer Perception on Online Cosmetic Product Purchases Using Association Rule Mining Based on Customer Feedback*

Understanding customer perception through online feedback has become a critical research direction, particularly in the cosmetics and beauty industry where consumer preferences strongly influence purchasing behavior. Mahesan et al. , in their study "Customer Perception on Online Cosmetic Product Purchases Using Association Rule Mining Based on Customer Feedback", investigated how consumer reviews can be analyzed to uncover hidden patterns that affect buying decisions. Their work applied association rule mining to identify relationships among customer feedback, product features, and purchasing behavior.

The study demonstrated that association rule mining is effective in highlighting the most influential factors that drive consumer interest and brand loyalty. For example, the authors showed that cosmetic products play a vital role in daily routines, and customer perceptions of attributes such as product quality, price, and brand reputation directly affect purchase plans. By analyzing patterns within consumer feedback, the research provided actionable insights for companies to improve product marketing strategies and customer retention.

This research is relevant to the project because it emphasizes the importance of analyzing customer feedback in the beauty domain. While Mahesan et al. focused on association rule mining for discovering purchase-related patterns, this project builds on this perspective by performing exploratory data analysis and sentiment analysis on the Sephora Product Reviews dataset. By combining text mining techniques with sentiment classification, our work aims to uncover not only patterns of consumer behavior but also the emotional drivers behind product evaluations, thereby providing a complementary view to the findings of Mahesan et al.

D. *Relevance to Our Project*

These studies collectively establish a strong foundation for research on consumer reviews. The Yelp study provides a methodological framework for EDA, the Naïve Bayes beauty product analysis demonstrates the effectiveness of sentiment classification models, and the association rule mining study emphasizes the role of feedback in shaping purchase decisions. Building upon these works, this project applies EDA, text mining, and sentiment analysis techniques to the Sephora Product Reviews dataset. By combining structured data analysis (e.g., product categories, ratings) with unstructured text mining (e.g., sentiment classification, keyword extraction), this project aims to uncover both the emotional and behavioral drivers of consumer preferences in the beauty industry.

III. Proposed Approaches

To achieve the goal of analyzing these product reviews, it is important that the right methods of evaluation and review are chosen in order to have successful results.

Three different methods were selected to do this. Exploratory Data Analysis, Sentiment Analysis, and Association Rule Mining.

A. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) is used to understand the dataset's characteristics. The characteristics of this dataset will be presented by summarizing the statistics and using different visual tools.

Using EDA in text analytics will focus on several different aspects, including counts and lengths of the reviews. This will count the words, sentences, and average the length of these. There is also a term frequency analysis, which will look at the most frequently occurring words and phrases.

The goal of using EDA is to uncover patterns, relationships, and outliers presented in the data. With this information, preparing the data for further analysis and developing a hypothesis will present itself in a more straightforward way.

The majority of the EDA will focus on exploring product categories, regular and discount prices, brand popularity, the impact of different characteristics on price, and ingredient trends. This will be using text mining analysis, comparative analysis, and correlation analysis.

The analysis will be divided into several different sections:

- Descriptive Statistics of the Structured Data: Taking the numerical and categorical information, such as prices, average ratings, and product type visualizations, will be created to compare.
- Exploration of Textual Data: Focus on word and sentence counts, length, and word frequency to figure out common phrases.
- Sentiment Rating Alignment: Compare the alignment between the star ratings (numerical) and the sentiment (textual)
- Product Feature Trends: The analysis of the data can be shown through uncovering insights and patterns, such as the development of new products and information. There is also an identification of data quality issues between the product and its ingredients. There is more analysis between brand popularity, price segmentation, and ingredient mentions.

B. *Sentiment Analysis*

Sentiment Analysis, also known as opinion mining or emotion AI, will be used to focus on the extraction and classification of the opinions in the customer reviews. This method focuses on analyzing and identifying the intent of the customer reviews.

The intent could consist of positive, negative, or neutral sentiments. By applying natural language processing (NLP) techniques, the dataset will be analyzed for emotional tone to compare the tone with the ratings.

The analysis will be divided into several different sections:

- Polarity Detection: Determining whether the review falls into a positive, negative, or neutral category
- Intensity Analysis: Measure how strong the emotions in the text are.
- Aspect-Based Sentiment: Identifying specific features of the product that are associated with the positive, negative, or neutral feedback.
- Compare Ratings: Analyze whether the sentiment found in the sentiment testing corresponds with the star rating.

C. *Association Rule Mining*

Association Rule Mining focuses on large data items and finds relationships between the items. These items could include product features, review terminology, and star ratings. It is often used with a multi-attribute dataset and Market-Based Analysis.

The goal of Association Rule Mining is to discover general patterns that predict success, customer satisfaction, and financial and market trends associated with the product.

The analysis will be divided into several different sections:

- Frequent Itemset Mining: Find frequent word combinations. These words could include ingredients or attributes that commonly appeared together in reviews.
- Rule Generation: Find and create association rules that explain the connections between the text features and structured data.
- Rule Reliability: Using the measures of support, confidence, and lift, determine the strength of the rules. This ensures the meaningfulness of the patterns.
- Cross-Category Analysis: Compare rules of different products, brands, and prices to showcase the most satisfied customers and their association.

IV. Planned Experiments

For the planned experiments, the authors will take the proposed methods and apply them to our data. This study will utilize the large dataset containing over 1.09 million customer reviews across more than 8,000 beauty products. In order to use this data, it must be preprocessed. TO prepare it, duplicate reviews will be removed, timestamps will be standardized, and missing information will be flagged and taken care of. Then normalization of product attributes along with review features. These steps, along with other preprocessing techniques, will produce both structured product-level data and unstructured text features for analysis.

A. *Exploratory Data Analysis*

EDA is a good first step after preprocessing, as it will establish baseline patterns in the dataset. Product analyses will examine how ratings, review counts, and prices can vary across categories in the beauty industry, such as fragrances, makeup, and skincare. At the review level, EDA will look into whether demographic groups provide systematically different ratings. Correlation Analysis will also be conducted to determine how strongly product attributes such as price and review count correlate with the overall ratings of the product. Together, these analyses will give a foundation to understand which types of products succeed in the marketplace and how certain factors shape success.

B. *Sentiment Analysis*

The next experiment will use **Sentiment Analysis** to evaluate the relationships between textual sentiments in the reviews and their corresponding star ratings and other indicators. Sentiment scores will be computed for each review and compared against numeric ratings to assess alignment and to identify cases where the written review text does not reflect the corresponding numeric score. Products that have unusually high ratings but negative textual sentiment, or vice versa, will also be examined to highlight mismatches. These experiments will test the hypothesis that textual sentiment generally aligns with numeric ratings but that meaningful discrepancies exist, particularly in subjective categories such as fragrance.

C. *Association Rule Mining*

The third set of proposed experiments will be using **Association Rule Mining** to find recurring patterns between product descriptions, customer demographics, and satisfaction outcomes. Reviews will be transformed into transactions by finding keywords and demographic attributes, along with outcomes such as high vs low ratings or recommendation status.

After that, association rules will be extracted to find patterns in positive and negative reviews. Other rules may also be researched and evaluated based on support, confidence, and lift, which then will be filtered for stability across categories and time.

Using all three planned experiments, the methods will provide a comprehensive analysis of consumer experience in the beauty industry, specifically Sephora customers and products. This integrated approach is expected to yield insights into both the measurable and emotional drivers of product success, while also identifying discrepancies between numerical ratings and the language of reviews.

REFERENCES

[1] E. S. Alamoudi and S. A. Azwari, "Exploratory Data Analysis and Data Mining on Yelp Restaurant Review," 2021 National Computing Colleges Conference (NCCC), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/NCCC49330.2021.9428850. keywords: {Visualization;Data analysis;Spatial databases;Data mining;EDA;Data Mining;BOW;Unigram;Bigram;Trigram;Yelp dataset;Restaurant data},

[2] Rambe, T. S., Hasibuan, M. N. S. ., & Dar, M. H. . (2023). Sentiment Analysis of Beauty Product Applications using the Naïve Bayes Method. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *7*(2), 980-989. https://doi.org/10.33395/sinkron.v8i2.12303

[3] S. S. Mahesan, U. R, S. Uma and R. Ganesan, "Customer Perception on Online Cosmetic Product Purchases Using Association Rule Mining Based on Customer Feedback," 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2023, pp. 714-720, doi: 10.1109/ICSSAS57918.2023.10331647. keywords: {Drugs;Footwear industry;Companies;Real-time systems;Hardware;Electronic commerce;Artificial intelligence;Association Rule;Customer Perception;Data Mining;Online Shopping},

[4] Inky, Nady. "Sephora Products and Skincare Reviews." Accessed Sept. 2025. Kaggle, Mar. 2023.

[5] Website: https://imajdoch.github.io/ProjectProposal_CAP5771/