# 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 large-scale 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 within the beauty industry. The project integrates Exploratory Data Analysis, Sentiment Analysis, and Association Rule Mining to examine trends across Sephora revies, specifically their face masks. The study demonstrates how text mining can reveal the emotional and behavioral drivers behind product perception. This provides practical implication for product development, consumer understanding, and continued decision making in the beauty industry.**

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

## 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. This comes from an oversaturation of products, trends, and reviews, but a field of rich content 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 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 takes a deep dive into these reviews and identifies 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 to review datasets. 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. Shoes-ACOSI*

Aspect-Based Sentiment Analysis (ABSA) traditionally focuses on explicit opinions using the ACOS schema (Pontiki et al., 2014), but struggles with implicit opinions, where sentiment must be inferred (Li et al., 2019). The Shoes-ACOSI dataset (Zhang et al., 2024) extends ABSA to a five-tuple schema (ACOSI) with explicitness labels, highlighting challenges such as long reviews, mixed sentiments, and implicit triggers. Inspired by this framework, our project applies ACOSI to skincare reviews, addressing the gap of implicit opinion extraction in this domain.

### *E. 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.

### B. Sentiment Analysis

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

### 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.

## 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.

## V. Narrowing the Data

Upon further examination of the data, it was found that there are two separate pieces of our data we have to take into consideration.

### A. Product Info

The product info data file has seven hundred and four observations and twenty-seven features. covering product identifiers, brand details, consumer engagement metrics, pricing, availability, and categorization. While rich in categorical and numerical attributes, some fields (ingredients, pricing variations, descriptions) have significant missing values, which may influence analysis. This will be a challenge that this project will need to address.

### B. Reviews

The second half of this project's data is the product reviews. The data in its original form was five files, each containing two hundred and fifty products names and their respective reviews. Each product has multiple reviews, so the dataset is very large, over one million reviews in total. This scale creates several challenges for a text-mining workflow, including data integration issues, quality control, computational constraints, modeling pitfalls, and governance concerns.

### C. Selecting our Sample Data

In order to narrow down the project and create a quality assessment, the product requirements had to be narrowed down. By examining the data, it was found that the majority of the reviews given products fell into the skincare quality. Even though the category was narrowed down to skincare, there were still over five hundred thousand reviews, which would not give us a quality assessment.

### D. Code Review

The updated Python workflow addressed several challenges brought upon by selecting the sample data, more specifically with the scale and inconsistency of the review data. The code was inputted with the scaled down product reference list, with only our sample data and therefore was relevant to the project scope. It also reads all of the review files provided in the dataset. It then outputs a .xml file with only the reviews, and their related column information, relevant to our sample data.

### E. Data Cleaning and Prepossessing

To prepare this accumulated data in the Python script, it also included a series of cleaning and normalization steps. To start off, the product names were standardized. This was achieved by all values converted to lowercase, stripped of punctuation, accents, and extra spaces. This ensured consistency between the reference product list and the review files and reduced false negatives during the matching stage. This also will be beneficial for further analysis stages.

The review was also altered to a form that is easier for the NLPs to understand. HTML tags, non-alphabetical characters, and excess spaces were removed. The text was also converted to lowercase lettering. The titles underwent a similar process to ensure consistency throughout the data.

Demographic variables like skin tone, skin type, hair color, and eye color were normalized into consistent lowercase categories. This was done to reduce fragmentation in the dataset. Numeric fields like ratings and helpfulness scores were cast into numeric types to support statistical aggregation.

Price data was altered to strip the data of any extraneous symbols to just produce a clean numeric column. This makes it easier to analyze trends related to price sensitivity or consumer preference correlations.

Data governance rules were enforced. This was done by removing the empty columns and unnamed metadata fields. Duplicate reviews were also reviewed and removed to avoid bias. With the cleaned text, the dataset was further prepared for deeper analysis.

## VI. Exploratory Data Analysis

To begin the project, the first step is to use Exploratory Data Analysis Techniques to gain in-depth knowledge of the complex relationships in the data.

### A. Values

The first piece was to evaluate if there are any missing values in the data after combining all of the reviews for the masks together. Using python, it was found that there were some missing values in the data. The most concerning insight was that out of 200,000 reviews the helpfulness score feature had 120,920 missing values. This is around 60 percent if all values for this feature are missing. This is problematic because if not addressed, these missing values can curve the results and prevent the authors from finding accurate results. To address this issue, the authors decided to drop the helpfulness feature and proceed with 19 features. There were also missing features in the 'review_text' and 'cleaned_review_text' columns, but because there were only 182 missing reviews out of the 200,000 observations, the authors decided to drop those with missing text, as the point of this project is to mine the textual part of the reviews.

The last set of missing values that need to be dealt with is in the review title features. The initial missing value script showed that the 'review_title' and 'cleaned_review_title' features also had some missing values with 58,778 titles missing out of the 200,000 reviews. This could be explained by the fact that the title of the review was optional. To solve this, the solution was to fill all empty titles with an empty string value of ''. This allows for all the rows to be kept while still preserving the accuracy of the data.

As part of the exploratory data analysis (EDA), value counts were used to summarize the frequency distribution of categorical variables, such as skin tone, skin type, hair color, and brand name. This helped identify which categories were most common and whether any levels were underrepresented or imbalanced. For continuous variables such as product price, EDA included examining the overall price range, mean, and spread of values through histograms and summary statistics. These analyses provided insight into the dataset's composition, potential skewness in pricing, and any outliers that might influence later analysis.

### B. Trends Over Time

Exploratory data analysis (EDA) was used to examine the distribution of reviews by month. After converting submission timestamps to datetime format, the dataset was grouped by month and year to identify seasonal trends in review activity.

The resulting table and visualization showed clear variation in review frequency across months, highlighting potential temporal patterns in consumer engagement.

### C. Product and Brand Analysis

By grouping the dataset by 'brand_name' and counting unique 'product_name' values, the analysis identified the brands with the widest product offerings and most reviews. Clinique, Shiseido, and Kiehl's Since 1851 appeared as leading brands in terms of product diversity, suggesting broad market representation and consumer reach.

Next, the relationship between product price and customer rating was explored by grouping reviews by rating and calculating the average 'price_usd' within each group. Results showed that higher-rated products did not necessarily correspond to higher prices, as average prices were relatively consistent across rating levels *(around $59–$62)*. This suggests that perceived product quality, as reflected by ratings, may be influenced more by product performance or brand reputation than by price alone.

### D. User Demographics

The User demographic information that is available in this dataset gives the unique opportunity to analyze the type of people creating the reviews. This will hopefully help to explain why certain products are getting their respective ratings.

The analysis showed that most reviewers reported *light*, *fair*, or *light-medium* skin tones, and over half identified as having *combination* skin. The most common hair colors were *brown* and *blonde*, and *brown eyes* were the predominant eye color. A notable share of entries (10–18%) listed demographics as *unknown*, which may reflect optional or incomplete user reporting.

To follow the analysis each feature previously evaluated, the 'skin_tone', 'skin_type', 'hair_color', and 'eye_color" were then grouped by rating to see which types of people gave higher and lower ratings to certain products. Average product ratings were generally consistent across demographic groups. Medium and light skin tones, combination skin, and lighter hair and eye colors showed slightly higher mean ratings, though differences were marginal (within 0.1 rating points). This indicates that reviewer satisfaction did not significantly vary by demographic attributes.

### E. Textual Analysis

As part of the textual exploratory data analysis, the average length of user reviews was calculated to assess the amount of written feedback. Additionally, all cleaned review texts were combined and tokenized to identify the most frequently used words across reviews. This analysis provided an initial understanding of reviewer language and key topics prior to any formal text mining or sentiment modeling.

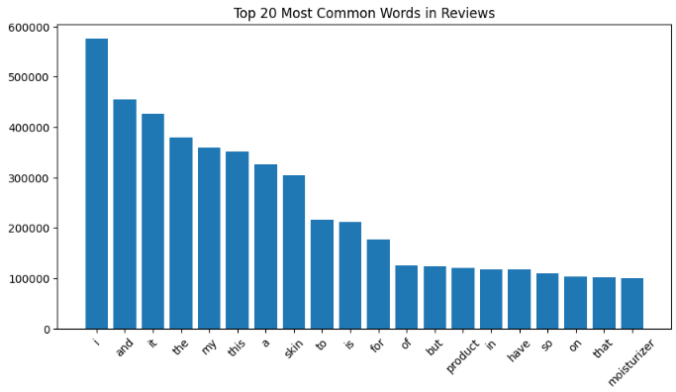



*Fig. 1 Top 20 Most Common Words in Reviews*

Figure 1 shows a word frequency distribution and bar plot of the top 20 words revealed that common terms such as *and*, *the*, *my*, and *skin* appeared most often, reflecting the conversational and product-focused nature of the dataset.

### F. Cleaning Biases

Regarding the influence and bias created by Sponsored Reviews. This was addressed through sorting through a set of suspicious keywords often associated with sponsored reviews. These words include "free", "sponsored", "promotion", "discounted", "in exchange", "gifted", "received", "complimentary", "influencer", "honest review", "testing purposes", "ambassador", "collab", "pr sample".

Other tests were created to compile this, including very short reviews that are often associated with sponsored reviews, so those were filtered out. Also, a 5-star review that most users don't find helpful could be a paid or biased review, and those were cleaned out as well.

## VII. Sentiment Analysis

The next big research step is to conduct a sentiment analysis of the data. A Sentiment Analysis is useful to uncover users' real sentiment toward the project by analyzing their text.

### A. NLP Specific Processing

In order to prepare the cleaned reviews for the sentiment analysis, there are a few steps to perform before the actual analysis.

Tokenization is the process of breaking the review of text into smaller units called tokens, typically words. This step converts each review from a long text string into a list of individual words that can be analyzed.

Stopwords are common words that do not add meaningful information for sentiment classification. Removing them helps reduce noise in the text and improves the model's focus on sentiment bearing words.

Lemmatization and stemming are used to reduce words to their dictionary definition, or their base. This is important because the models treat variations of a word as the same feature, which improves consistency in the dataset.

All of these were used to get the data ready for the VADER model and the Sentiment Analysis as a whole.

### *B. Weighted Classes*

VADER or Valence Aware Dictionary and Sentiment Reasoner was selected for this project because it is specifically designed to analyze short, informal, and emotionally expressive text such as product reviews. It accounts for sensitive text like capitalization, degree modifiers, and negations, allowing for more nuanced sentiment scores even in noisy text. It also outputs a single compound score which allows for easy comparison to support the sentiment-rating alignment.

The VADER sentiment analysis was performed to classify the reviews into Positive, Negative, and Neutral. There were 48,401 Positive reviews, 5,088 Negative reviews, 1,917 Neutral reviews. Then, for each sentiment class, it showed the most common words across the different reviews after tokenization.

Without weighing the sentiment of words that are often associated with positive, negative, and neutral words, there was frequent overlap between the categories. The most common words included “skin”, “mask”, “use”, and “product”. These words were the most frequent within all categories.

The script was then changed to incorporate an opinion lexicon NLTK that contained positive and negative sentiment words including “amazing”, “bad”, “love”, and “irritate”. Scanning through the text, it was able to count how often the opinion-bearing words appeared in each sentiment category.

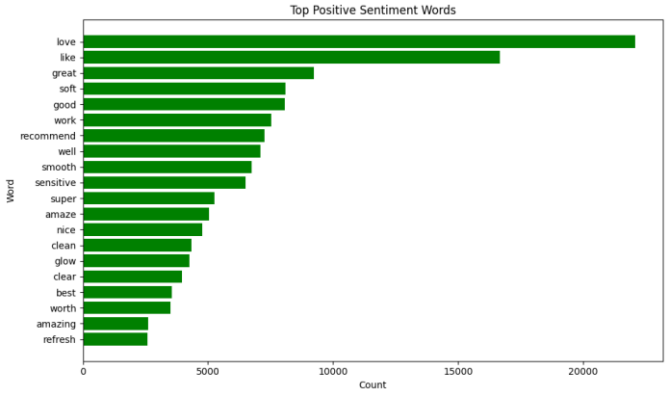



*Fig. 2 Top 20 Positive Review Sentiment Words*

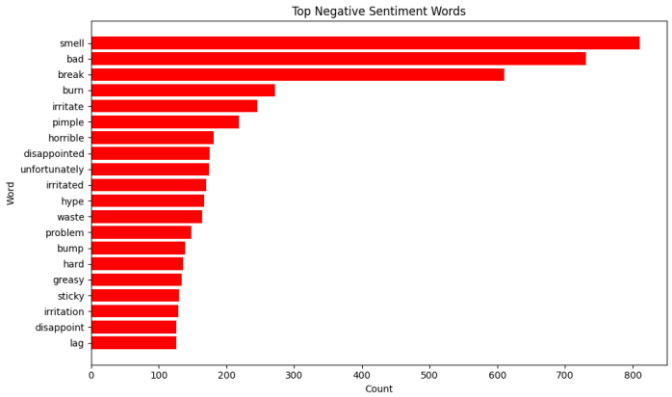



Fig 3. Top 20 Negative Sentiment Words

Analyzing Fig. 2 and Fig. 3 the positive and negative sentiment words can we see within these figures. The top 5 Positive Sentiment words were “love”, “like”, “great”, “soft”, and “good”. The word “love” is shown 22092 times, “like” 16685 times, “great” 9227 times, “soft” 8110 times, and “good” 8064 times. While the top 5 negative sentiment words include “smell”, “bad”, “break”, “burn”, and “irritate”. The word “smell” was shown 810 times, “bad” 731 times, “break” 610 times, “burn” 272 times, and “irritate” 245 times.

The word frequency analysis indicated that customer feedback on face masks is overwhelmingly positive. The frequency of strong positive words like “love” and “great” suggests an overall high level of satisfaction between all of the reviews. But within the negative reviews, the terms like “smell” and “irritate” show concerns within those products. These provide a solid foundation for further analysis of brand perception and product performance.

### *C. Sentiment Comparison*

To understand the impact and true validity of the sentiment analysis, the analysis can be compared to other features that are in the dataset. These include star ratings, prices, and brands.

Star Ratings represent the numerical measure of customer satisfaction provided by users when reviewing products on the Sephora website. Each rating typically ranges from 1 to 5 stars, with 1 star indicating strong dissatisfaction, and 5 reflects high satisfaction or product approval. These ratings serve as a key indicator of the user experience which complements the written reviews.

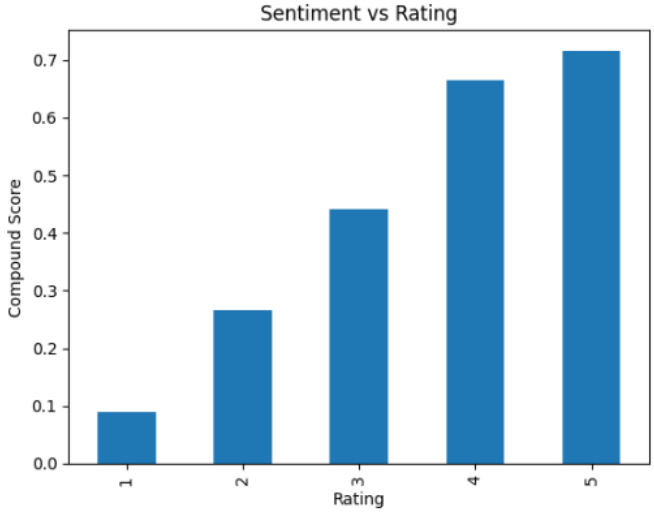



Fig 4. Sentiment vs Rating

The Figure above shows a clear positive correlation between sentiment and rating. Reviews with higher star ratings exhibit higher average compound sentiment scores, indicating that more favorable written opinions correspond to higher numerical ratings. Specifically, reviews rated 1 star have an average compound score near 0.1, reflecting negative

sentiment, while 5-star reviews reach an average of approximately 0.7, representing a strongly positive emotional tone.

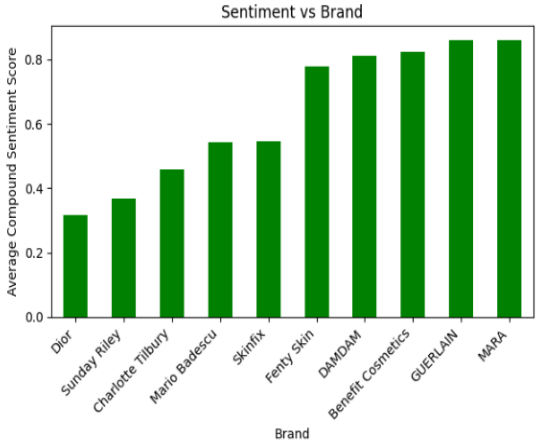
This pattern confirms that VADER successfully captures the emotional content of customer reviews and aligns well with the structured star rating data. The consistent upward trend across ratings demonstrates that the textual sentiment closely mirrors the reviewers' overall satisfaction levels, validating the reliability of sentiment analysis for assessing the customer feedback in the beauty product domain.

Prices play a key component in the analysis because they help to understand consumers' perception and satisfaction as it is a representation of product value and its market position.


Fig 5. Sentiment vs Price

Figure 5 shows the Sentiment Vs Price scatter plot. The plot shows a generally positive sentiment across all price ranges, with most compound scores clustering between 0.5 and 0.7, which indicates overall favorable perception of the products regardless of the cost. This graph shows no strong linear relationship between price and sentiment, which suggests that the customer is not solely dependent on product price. This pattern implies that price does not strongly influence the emotion tone in reviews, and customers evaluate products primarily on performance and individual experience rather than cost alone.

Brands also have a hand in a strong sentiment analysis, as brand loyalty and reputation can play an important part in how consumers perceive a product. In the context of this study, brand-level sentiment analysis highlights which beauty brands maintain the most positive consumer relationships and how emotional perception influences overall market success




. Fig 6. Sentiment vs Brand

The results of Figure 6, the Sentiment vs Brand analysis show that sentiment varies notably between brands. Luxury and skincare-focused brands such as GUERLAIN, MARA, and Benefit Cosmetics received the highest average compound scores, exceeding 0.8, indicating strong positive emotional responses from customers. In contrast, brands like Dior and Sunday Riley show comparatively lower sentiment averages, around 0.3–0.4, suggesting more mixed or moderate satisfaction levels.

The variations that have resulted from this analysis highlight how brand reputation, product quality, and customer expectations contribute to differences in perceived satisfaction from the customers. The results suggest that brand identity is a significant factor of consumer perception in the beauty market.

### *D. Naive Bayes Analysis*

The Naive Bayes Analysis is used within this paper to contribute to text classification tasks like sentiment analysis. Which was trained to automatically classify reviews as positive, negative, or neutral based on their text content. It learns patterns in word usage directly from the labeled data.

The assumption that all features are independent of each other is true. The text was first preprocessed with Term Frequency and Inverse Document Frequency (TF-IDF). This cleaned and tokenized the review text with sentiment labels obtained from the VADER analysis. With this trained data, the model can predict the sentiment of new reviews based on the language patterns.

The Niave Bayes model implemented a pipeline that ensured vectorization occurring only on the dataset. This prevented data leakage and domain specific stopword removal to reduce bias.

Label distribution was applied showing around 80% positive, 10% negative, and 4% neutral. This imbalance strongly influenced how the model performed. It predicted the majority class more often because that would minimize the overall errors.

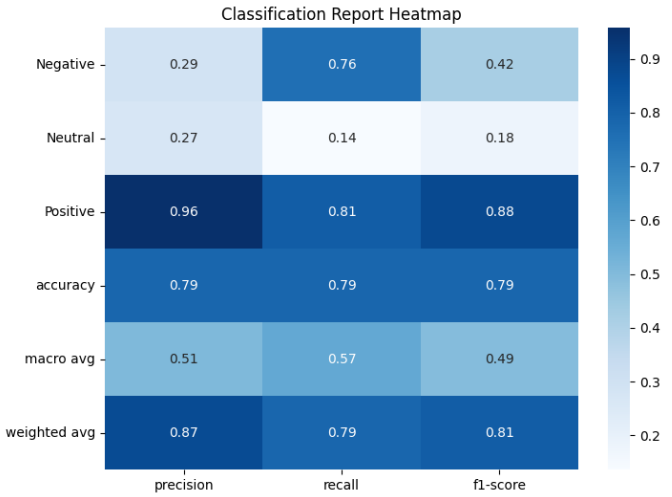



Fig 7 Classification Report Heatmap

The Classification Report was produced with these results; this model shows a 79% accuracy and is high because the dataset is dominated by positive reviews. The negative class precision 0.29 predicts negative is only correct for 29% of the time. The recall is 76% and is catching 76% of all actual negative reviews. While with an F1 score of 0.42 it shows that there is a moderate, with some imbalance. The model is overpredicting the negatives. Because the "neutral" boundary is fuzzy in language, and you have very few neutral samples to learn from. The model is very confident and accurate on positives.

The top features per class were calculated through this model as well. The words or phrases the model relies on to make predictions are displayed. The top 15 features for class Negative include: 'worst', 'smell awful', 'really disappointed', 'lip balm', 'bad mask', 'without use', 'kill', 'worse', 'dont care', 'red patch', 'unfortunately didnt', 'first two', 'no joke', 'old formula', 'negative review'.

The top 15 features for neutral were 'trs', 'produit', 'jai', 'ce', 'une', 'est', 'lip balm', 'un', 'facial treatment', 'several month', 'benzoyl peroxide', 'je', 'benzoyl', 'seed', 'get hormonal']. There were many non-English or factual terms within this, several French words and product mention which makes them harder to classify.

The top 15 features for positives included 'soft radiant', 'love help', 'awesome mask', 'amazing love', 'nice hydrate', 'night love'. Strongly positive adjectives and emotional expressions were produced with very clear praise patterns. The model easily learns these since positive language is common and distinctive.

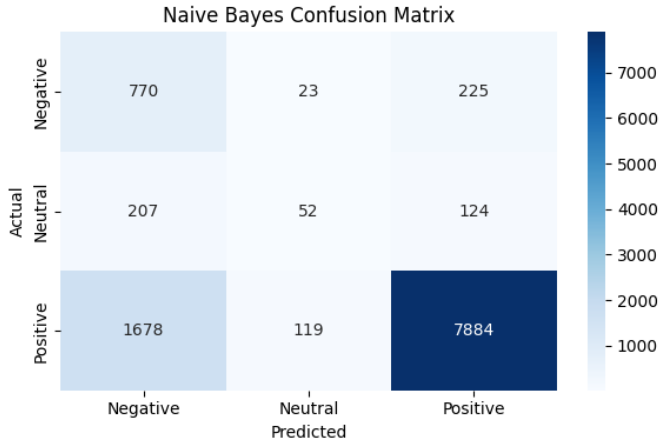



Fig. 8 Naive Bayes Confusion Matrix

The Naive Bayes confusion matrix shows strong performance on positive reviews but weaker accuracy for negative and neutral reviews. Most misclassifications involve neutral or mildly polarized text, suggesting that class imbalance and subtle tone differences impact the model's ability to distinguish sentiment of extremes. The model demonstrates overall reliable performance, especially in recognizing positive sentiment.

The Naive Bayes classifier performs well for positive sentiment detection. The model struggles with neutral and negative reviews highlight the importance of balanced training data and richer linguistic modeling. Enhancing preprocessing and testing alternative algorithms could lead to more reliable and equitable sentiment predictions.

## VIII. Association Rules

Association Rule Mining allows the authors to find recurring patterns between product descriptions, customer demographics, and satisfaction outcomes. It is an important tool for this analysis because it can reveal relationships that would otherwise go unnoticed but are still rich in information for analysis. Association Rules can reveal how customers actually talk about their products and can give key insights that can benefit both consumers and the brands/companies that create these products.

### *A. FP-Growth*

In order to find frequent patterns, FP-Growth was chosen because of its efficient runtime and ability to scale appropriately with our dataset. This algorithm has the ability to compress the data into an FP-tree and mines patterns without needing to generate every single candidate set, unlike other algorithms. After obtaining frequent itemset, Association Rule Mining was applied to compute support, confidence, and lift, allowing meaningful linguistic and brand-related relationships to be extracted from the review text.

### *B. Brand-Token Relationships*

To understand how customers talk about each brand, I applied FP-Growth to extract the frequent itemset and then used Association Rule Mining to generate the Brand to Token rules. These rules highlight which words are associated with certain

brands across the dataset. These patterns help reveal how customers perceive each brand and the over branching themes that consistently appear in their feedback.

| Brand | Key Tokens | Interpretation |
|---|---|---|
| **Fresh** | rise, fresh, scent, smell, price | Fresh is strongly tied to its own brand language, plus scent, samples, and price. People talk a lot about how it smells and whether it's worth the price. |
| **Youth To the People** | night, hydrate | Associated with nighttime skin care and hydration routines. |
| **Summer Fridays** | moisturizer, moisturize, hydrate, night | Seen as a hydrating, moisturizing brand, often used at night. |
| **Herbivore** | smell | For Herbivore, scent is a big talking point (good or bad). |
| **Dr. Jart+** | felt | Reviews emphasize how the product feels on the skin ("it felt soothing," "felt heavy," etc.). |
| **Origins** | clear, clean, pore | Origins is perceived as a pore-clearing / deep-cleaning / clarifying brand. |

Fig. 9 Brand to Token Insights

Figure 9 summarizes the strongest associations for some key brands, focusing on the tokens with high lift and confidence scores.

### C. *LDA: Key topics via Association Mining*

LDA or Latent Dirichlet Allocation is a model for unstructured data that can find abstract topics in a collection of documents, or in the case of this project review.

After generating brand-level associations, a FP-Growth and Association Rule Mining was applied to examine broader token co-occurrence patterns across all skincare reviews. By clustering rules with high lift and confidence, I identified five recurring linguistic themes that represent the main ways customers talk about skincare products. These themes summarize the biggest topics in the dataset.

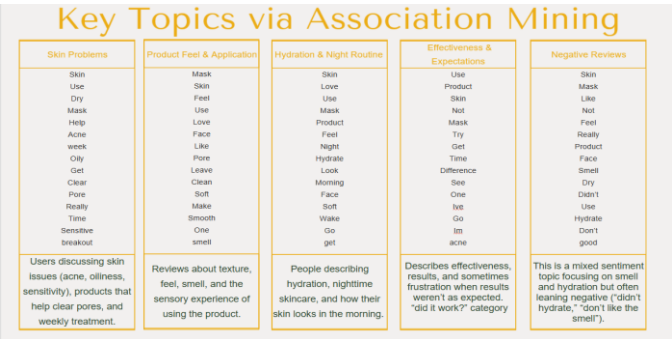



Fig. 10 Key Topics via Association Mining

The extracted topics in figure 10 reveal a consistent structure in how customers write about skincare products. The extracted topics show a consistent structure in how users describe their skincare experiences:

Skin Problems:

Discussions of acne, oiliness, pore care, sensitivity, and weekly treatment routines.

Product Feel & Application:

Emphasis on texture, scent, smoothness, and the sensory experience of using the product.

Hydration & Night Routine:

Comments about moisturizing effects, nighttime use, and how the skin looks in the morning.

Effectiveness & Expectations: User evaluations of whether the product "worked," including comparisons, results, and outcomes.

Negative Reviews:

Mix of complaints focused on dryness, scent issues, lack of hydration, or general dissatisfaction.

These five clusters align with common skincare concerns and help validate both the LDA topics and the sentiment patterns found elsewhere in the analysis.

## IX. Conclusion

Throughout this research, over 200,000 Sephora face mask reviews were analyzed to understand consumer experiences within the beauty industry. This was achieved exploratory data analysis, sentiment analysis, and association rule mining. By integrating the data with structured product information with the unstructured review, text patterns started to emerge that displayed consumer behavior, positive/negative sentiment, and other factors that shaped perceptions of product satisfaction.

### A. *Summary of Key Findings*

EDA revealed several insights about the characteristics, demographics, and rating trends the provided a great baseline for understanding and further analyzing the data. Despite the common assumption, the higher-priced items did not receive higher ratings, but rather price had no relationship with consumer satisfaction. Demographic characteristics like skin tone, skin type, and different colorizations showed little variations across different aspects such as rating levels. This shows that product quality did not falter or change through the demographic groups.

Sentiment Analysis showed strong alignment with the tone of the review with star ratings. Showing higher rated products with positive language and lower rated products with more negative language. The higher rated products were associated with ideas such as texture, hydration, and overall satisfaction. Negative reviews stemmed from words such as poor smell and irritation. Although, within the Sentiment Analysis we found some mistakes such as 5-star reviews containing negative

sentiment words and 1-star reviews containing positive language. This shows some potential positivity bias. The brand sentiment analysis showed variation among the companies with sometimes luxury brands performing worse than non-luxury brands.

Within the sentiment analysis, a Niave Bayes study was performed that achieved strong and significant accuracy on positive reviews but struggled with detecting negative and neutral statements due to a class imbalance.

Association Rule Mining provides key insights into how the customers interact and truly feel about these products. This further revealed recurring themes such as hydration, texture, application experience, skin problems, and effectiveness. These patterns allowed us to connect the features of the products with customer feedback to relay otherwise unseen relationships. This work was critical in creating a rich and meaningful analysis.

### *B. Strengths*

Throughout this study, strengths started to emerge starting at the dataset. One of the key strengths included the use and combination of structured and unstructured data from a large real-world dataset. The different approaches that this study utilizes build a multifaceted understanding of consumer perceptions. The focus on sentiment and emotional tone allowed for insights that analyzing star ratings couldn't show. The work shows demographic attributes, brand-level comparisons, and product characteristics that contributed to a full view of the landscape.

### *C. Limitations*

Several limitations should be addressed to fully understand the depth of the project. Within Niave Bayes, there was a large class imbalance, of roughly 80% of reviews classified as positive. This makes it challenging for the traditional models to properly capture minority classes. Some other data could have been provided and analyzed that would have provided a more in depth revies. This data could have included a helpfulness score, ingredients, and review titles. The narrowed subset ensured a manageable scope but reduced the generalizability of insights to other product categories. VADER, while good for short informal text, doesn't fully capture the contextual or smaller expressions. Transformer based models like BERT could analyze and pick up these expressions more successfully. Language variation like non-English reviews were common that introduced challenges for classification and normalization.

### *D. Implications and Applications*

The findings within the research offer practical implications for brands, retailers, and developers within the beauty industry. Sentiment-driven insights can help companies revise and refine the formulas and address common complaints. They also could utilize this information to improve product advertising for a more competitive edge. The presence of the positivity bias shows that consumers may not display their negative thoughts directly through the ratings so it is important to analyze the textual sentiment. Retailers can also use association rule patterns to improve personal recommendation to understand the product attributes most valued by certain audiences.

### *E. Future Work and Recommendations*

This research can be further developed in several different ways. Opening up the dataset to the full 1M+ reviews could provide a broader insight into the beaty industry. The research also could be focussed in other specific beauty product areas such as foundation, where colored products can be analyzed to see the change the amount of positive reviews. Doing this would also improve model robustness.

More advanced NLP models like BERT, DistilBERT, or other domain-specific transformers could be incorporated to improve sentiment classification, especially for more neutral and negative reviews that had a larger problem being analyzed with the used models.

Aspect-based sentiment classification could analyze specific attributes such as price or smell could provide further insight rather than just doing the sentiment alone. The temporal sentiment shifts could also show how brand reputation could change over time.

Expanding association rule mining throughout more product categories could allow cross-category comparisons. This would allow for a more in-depth comparison and understanding of consumer experiences and expectations across the beauty industry.

### *F. Concluding Remarks*

Overall, this study and research highlight the importance of combining the analysis of EDA, sentiment analysis, and association rule mining to uncover meaningful insights from a large-scale beauty product review. Although the research ran into some challenges including class imbalance and missing dimensions that could have strengthened the results, the results still demonstrated that consumer language provides valuable insights beyond traditional star ratings. The patterns of satisfaction, brand perception, and user experience emerged. As the beauty industry continues to expand and rely on the feedback that a text mining analysis offers, they could further identify product strengths, address pressing concerns, and guide future innovation.

## 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] Joseph J Peper, Wenzhao Qiu, Ryan Bruggeman, Yi Han, Estefania Ciliotta Chehade, and Lu Wang. 2024. Shoes-ACOSI: A Dataset for Aspect-Based Sentiment Analysis with Implicit Opinion Extraction. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15477–15490, Miami, Florida, USA. Association for Computational Linguistics.

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

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