Beauty Intelligence Hub: Know Your Product Before You Glow

By
Pooja Dayanand Kabadi

Contents

Table of Figures

1 Introduction

The beauty and skincare industry is rapidly evolving, with consumers becoming increasingly aware of the ingredients in their products and the need for personalization. The challenge, however, lies in the overwhelming volume of data: customer reviews, ingredient lists, ratings, and conflicting opinions. Our objective was to transform this unstructured information into actionable insights through a unified digital interface.

This project, Beauty Review NLP Dashboard, harnesses natural language processing (NLP), sentiment analysis, and data visualization to empower users with transparent, data-driven insights into Sephora skincare products. It ensures users can confidently select products aligned with their skin concerns and ingredient preferences.

2 Data Description

The primary dataset used in this project is sourced from the publicly available Sephora Products and Skincare Reviews dataset on Kaggle. It contains over 1 million customer reviews, product ratings, ingredient lists, and brand metadata across a diverse range of skincare items available on Sephora. This dataset provides a rich foundation for performing natural language analysis, identifying ingredient-based health risks, and visualizing product trends over time. All personally identifiable information (PII) was excluded under data usage ethics.

1. This file aggregates customer reviews for Sephora skincare products and includes sentiment annotations:

Column Name	Description
review_id	Unique identifier for each review

product_name	Name of the skincare product reviewed
brand_name	Brand associated with the product
review_text	Full customer review in natural language
rating	Star rating (1 to 5) provided by the user
review_date	Timestamp of review submission
sentiment_score	Numeric sentiment polarity score from VADER (range: -1 to 1)
sentiment_label	Categorized label derived from score: Positive, Neutral, or Negative
product_id	Unique identifier for each product

2. This file contains metadata and ingredient lists for Sephora skincare products:

Column Name	**Description**
product_id	Unique identifier for each product
product_name	Name of the product
brand_name	Brand offering the product
ingredients	Raw ingredient list (text format)
category	Product category (e.g., moisturizer, cleanser)
skin_type	Suitable skin type(s) as listed by the product (e.g., oily, dry)
target_concern	Skincare concerns addressed (e.g., acne, wrinkles)

3 Problem Statement

Customers today face several challenges:

- Information Overload: Thousands of reviews per product across platforms.

- Lack of Ingredient Awareness: Potentially harmful components are hidden in complex scientific names.
- Emotional Misalignment: Star ratings don't always reflect how users feel.
- Product Saturation: With hundreds of similar products, making a decision is hard.
- No Unified View: No existing tool merges ratings, reviews, and ingredient safety into one dashboard.

4 Ingredient Risk Research & Classification Framework

4.1 Ideation

To ensure transparency and consumer safety, we conducted a thorough investigation into ingredient-level health risks, utilizing verified sources from global regulatory and advocacy bodies. Our goal was to systematically classify skincare ingredients based on their potential harm and to develop a scoring system that communicates risk in a user-friendly manner.

Regulatory Sources Consulted:

We based our ingredient assessment on a triangulation of information from the following internationally recognized institutions:

- **European Medicines Agency (EMA)** — for pharmacological safety and chemical exposure thresholds.
- **EU Cosmetic Ingredients Database (Cosing)** — for ingredient regulatory status, usage limits, and prohibited substances in the EU.
- **Environmental Working Group (EWG)** — for toxicity ratings, carcinogenicity reports, and allergen flags in consumer skincare products.
- **FDA Cosmetics Ingredient Hotlist (US)** — to identify banned or restricted ingredients within U.S. regulations.

4.2 Risk Categories Defined

Using the research corpus, we developed a risk taxonomy with three critical hazard categories:

1. **Carcinogens**

   Ingredients linked to increased cancer risk based on toxicological studies or regulatory warnings.

   Examples: Formaldehyde, Coal Tar, Benzene.

2. **Allergens**

   Substances known to trigger skin, eye, or respiratory allergies through repeated or prolonged exposure.

   Examples: Fragrance (Parfum), Methylisothiazolinone, Balsam of Peru.

3. **Endocrine Disruptors**

   Chemicals that interfere with hormonal systems and may lead to reproductive, developmental, or metabolic issues.

   Examples: Parabens, Phthalates, Triclosan.

4.3 Rating & Classification System

To enable clear interpretation for non-expert users, we translated ingredient-level findings into a **3-tier color-coded risk meter**, displayed within the app:

Risk Level	Icon & Label	Criteria	Color Code
Safe	Safe	No harmful ingredients detected	#4CAF50 (Green)
Moderate Risk	Moderate Risk	1–2 harmful ingredients detected	#FFC107 (Yellow)
High Risk	High Risk	3 or more harmful ingredients	#F44336 (Red)

The classification was implemented via Python dictionaries, cross-referencing the parsed ingredient lists against our risk taxonomy. Users receive immediate visual feedback through the dashboard's **Ingredient Risk Checker**.

5 Ideation & Solution Design

5.1 Ideation

We envisioned a solution that:

- Aggregates and cleans reviews across product lines.
- Classifies user sentiment using NLP.
- Flags harmful ingredients (carcinogens, allergens, disruptors).
- Visualizes product trends, reviews, and ratings interactively.
- Offers filtering options to personalize results.

5.2 Core Objectives

- **Transparency:** Make ingredient risks visible.
- **Clarity:** Visual summaries of ratings, sentiments, and reviews.
- **Interactivity:** Empower users to filter by brand, product, concern, and timeline.
- **Intelligence:** Use NLP and analytics to derive deeper insights.

6 How the Project Helps the Beauty Industry

- **Customers:** Make informed, personalized skincare decisions.
- **Brands:** Get feedback loops to enhance products.
- **Dermatologists:** Gain transparency into allergens and concerns.
- **Retailers:** Enhance product pages with dynamic visuals.

This dashboard bridges the trust gap between consumers and skincare companies by combining scientific and emotional insights.

7 Data Collection & Preprocessing

7.1 Source

Dataset: Sephora Skincare & Reviews on Kaggle

Files Used:

- merged_reviews_with_sentiment.csv (review text, rating, sentiment score)
- product.csv (ingredients, brand name, category, skin type)

7.2 Processing Steps

- **File Consolidation:** Merged multiple review files into one merged_reviews.csv, later enhanced with sentiment scores using VADER.
- **Date Parsing:** Converted submission timestamps to standard datetime format for filtering.
- **Text Cleaning:**
    - Removed special characters, symbols, digits.
    - Converted to lowercase.
    - Applied stop word removal for buzzword extraction.
    - Used regular expressions to clean ingredient lists.
- **Sentiment Analysis:**
    - Applied VADER (Valence Aware Dictionary for sEntiment Reasoning) to score polarity.
    - Classified reviews into Positive, Neutral, Negative based on thresholding.
- **Ingredient Preprocessing:**
    - Used Python's ast module to parse ingredient strings safely.

- Normalized ingredient text.
- Matched against dictionaries of carcinogens, allergens, and endocrine disruptors.

8 Project Architecture & Technology

8.1 Tools & Libraries

Tool/Library	Purpose
Streamlit	Front-end dashboard framework for Python
Pandas	Data manipulation and joining
Plotly	Rich interactive charts (bar, donut, line, scatter, etc.)
NLTK (VADER)	Sentiment classification based on lexical features
Scikit-learn	Buzzword extraction using CountVectorizer
Regex (re)	Cleaning and pattern matching of ingredient lists
AST	Safely parsing string representations of ingredient lists
WordCloud	Creating buzzword visual summaries

8.2 Architecture Flow

8.2.1 Data Ingestion:

- Load two core CSV files using Streamlit's @st.cache_data mechanism for speed.

8.2.2 Preprocessing:

- Clean and standardize review and ingredient columns.

8.2.3 Sentiment Scoring:

- VADER used for real-time score generation.

8.2.4 Buzzword Analysis:

- Extract top buzzwords per sentiment using CountVectorizer.

8.2.5 Ingredient Matching:

- Compare normalized ingredient names to pre-defined risk dictionaries.

8.2.6 Visualization:

- Display insights via interactive Streamlit tabs.

9 Dashboard Overview & Features

The Section has been linked with the Appendix for Visual references of dashboard screenshot.

9.1 Filters (Sidebar)

- **Date Range:** Enables users to track trends seasonally or historically.
- **Brand Name**: Compare performance and safety across major brands.
- **Product Name:** Focused insights per item.

9.2 Tab 1: Summary

Key Features

- Total Reviews, Unique Products, Brand Count.
- Bar Chart: Rating Distribution (1–5 stars)
- Line Chart: Review Submission Trends over Months.
- Horizontal Bar Chart: Most Reviewed Products.
- Horizontal Bar Chart: Average Rating of Top Products.

How Users Benefit

- **Users can:**

- Spot which products have long-term traction.
- Identify products with highest average ratings.
- Understand seasonal trends (e.g., winter skincare popularity).
- Quickly gauge the scale of user interaction across brands.

9.3 Tab 2: Sentiment Analysis

Features

- Donut Chart: Proportions of Positive, Neutral, and Negative sentiment.
- Histogram: Distribution of sentiment scores.
- Buzzword Chart: Top 5 words per sentiment.
- Sample Reviews: 3 hand-picked comments each for positive and negative.
- Brand Sentiment Score: Average sentiment by brand (top 10).

How Users Benefit

- Users gain emotional feedback beyond rating scores.
- Words like “burn”, “glow”, “gentle”, etc. provide qualitative feel.
- Negative reviews help avoid misaligned purchases.
- Brands can monitor perception over time.

9.4 Tab 3: Ratings Analysis

Features

- Star Distribution: Count of each rating (1–5).
- Top Rated Products by Category: Descending order.
- Line Chart: Rating Trends Over Time.
- Filters: Skin Type & Skin Concern from metadata.

- Scatter Plot: Rating vs. Sentiment.

How Users Benefit

- Match products with high ratings and specific skin needs.
- Identify categories consistently rated highly (e.g., moisturizers).
- Know if the emotional tone aligns with the numeric score.

9.5 Tab 4: Word Cloud

What It Shows

- Commonly used words in reviews filtered by brand/product.

Why It Matters

- Let's users sense trending descriptors instantly.
- Helps users relate the language with their own skin concerns.

9.6 Tab 5: Ingredient Risk Checker

Core Components

- Lists harmful ingredients in 3 categories:
    - ☣️ Carcinogens
    - 🌸 Allergens
    - 🧬 Endocrine Disruptors
- Risk Meter block based on total flags.
- Color Code: Red (High risk), Yellow (Moderate risk), Green (Safe)
- Definitions Added
    - Carcinogens: Cancer-causing agents.
    - Allergens: Substances causing reactions.

- Disruptors: Hormone-affecting chemicals.

How It Works

- Ingredients parsed and matched from product.csv
- Normalized and cross-referenced to master dictionaries

How Users Benefit

- Users can:
  - Avoid products that include ingredients known to irritate their skin.
  - Check for long-term health flags before purchase.
  - Choose brands that align with safe formulation standards.

10 Key Visual Elements

Color Themes:

- Blue for charts, Red-Yellow-Green for alerts.

Interactive Charts: Zoom, hover, select.

Responsive Layout: Filters affect all tabs dynamically.

11 Future Scope & Next Steps

11.1 Expansion

- Ingest larger datasets from Ulta, Amazon, etc.
- Extend to categories like haircare, supplements.
- Translate for multi-language support.

11.2 Technical Enhancements

- Migrate VADER to DistilBERT for contextual emotion detection.
- Host on AWS, add login, and personalization.

11.3 Productization

- Enable brand comparison report exports.
- Build subscription APIs for e-commerce stores.

12 Conclusion

The Beauty Review NLP Dashboard translates raw customer feedback and ingredient disclosures into meaningful, reliable insights. It balances user-centric design with deep analytics—paving the way for smarter, safer skincare choices.

From ingredient risks to emotional tone, this dashboard is a complete beauty intelligence system. It empowers users to shop smart, stay safe, and glow confidently.

13 References

Source	Purpose	Link
European Union Cosing Database	Official EU resource for cosmetic ingredients, including permitted uses, restrictions, and safety notes.	https://ec.europa.eu/growth/sectors/cosmetics/cosing_en
EMA (European Medicines Agency)	Evaluations and pharmacovigilance for substances used in topical and therapeutic products.	https://www.ema.europa.eu/en
U.S. FDA Cosmetics Ingredient Directory	Guidelines, warnings, and alerts for cosmetic ingredient safety.	https://www.fda.gov/cosmetics

EWG Skin Deep® Database	Non-profit rating database for consumer awareness on ingredient toxicity and risk.	https://www.ewg.org/skindeep/
PubChem (by NIH)	Chemical structure and toxicity profiles for cosmetic and pharmaceutical ingredients.	https://pubchem.ncbi.nlm.nih.gov/
ChemSec SIN List	Endocrine disruptors and other substances of high concern.	https://sinlist.chemsec.org/
Health Canada Cosmetic Ingredient Hotlist	Canada's restricted/prohibited substances in cosmetics.	https://www.canada.ca/en/health-canada/services/consumer-product-safety/cosmetics/cosmetic-ingredient-hotlist.html

Literary Sources:

- **Darbre, P. D., & Harvey, P. W. (2008).**

  "Paraben esters: review of recent studies of endocrine toxicity, absorption, esterase and human exposure, and discussion of potential human health risks."

  Journal of Applied Toxicology, 28(5), 561–578.

  https://doi.org/10.1002/jat.1358

- **Krause, M., et al. (2012).**

  "Sunscreens: are they beneficial for health? An overview of endocrine disrupting properties of UV-filters."

  International Journal of Andrology, 35(3), 424–436.

  https://doi.org/10.1111/j.1365-2605.2012.01280.x

- **Nohynek, G. J., et al. (2010).**

  "Safety assessment of personal care products/cosmetics and their ingredients."

Toxicology and Applied Pharmacology, 243(2), 239–259.

https://doi.org/10.1016/j.taap.2009.12.001

14 Appendix

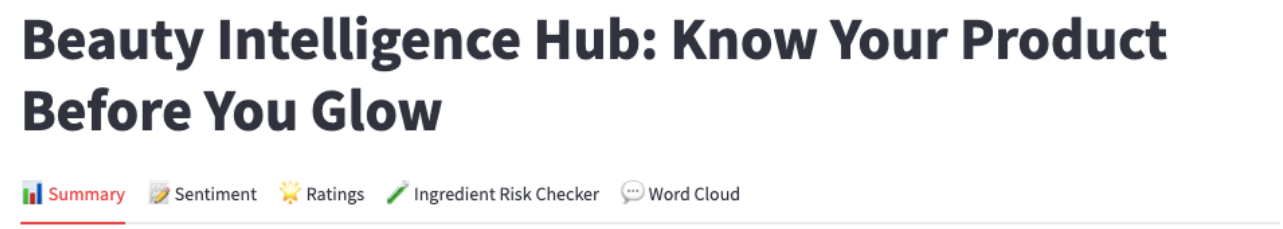

Figure 1Platform Title

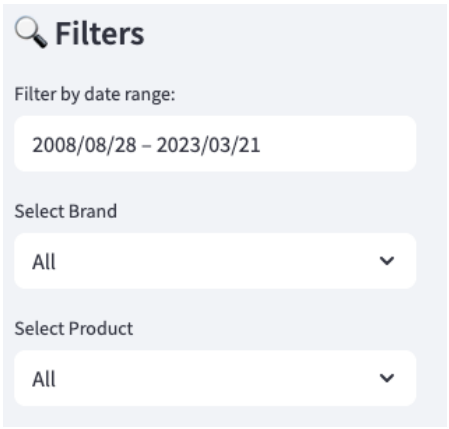

Figure 2 Global Filters controlling the entire dashboard

14.1 SUMMARY TAB

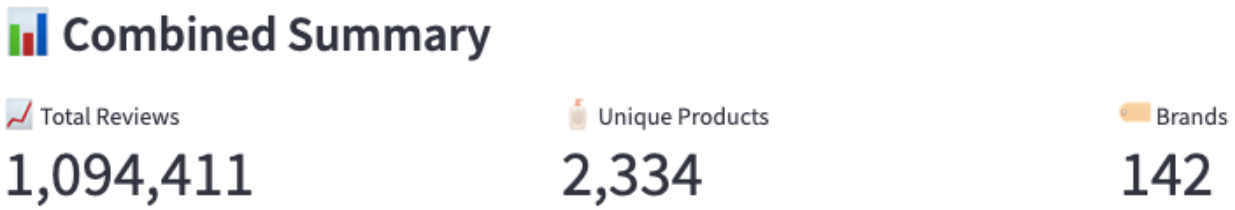

Figure 3 Summary KPIs with no brand or product selection

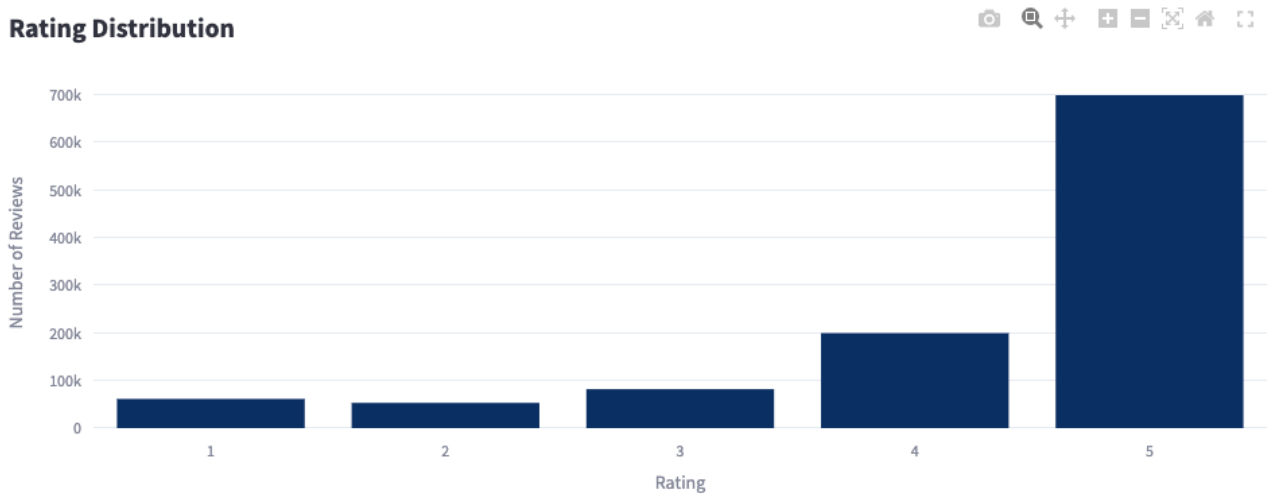

Figure 4 Summary Tab: Rating Distribution bar chart

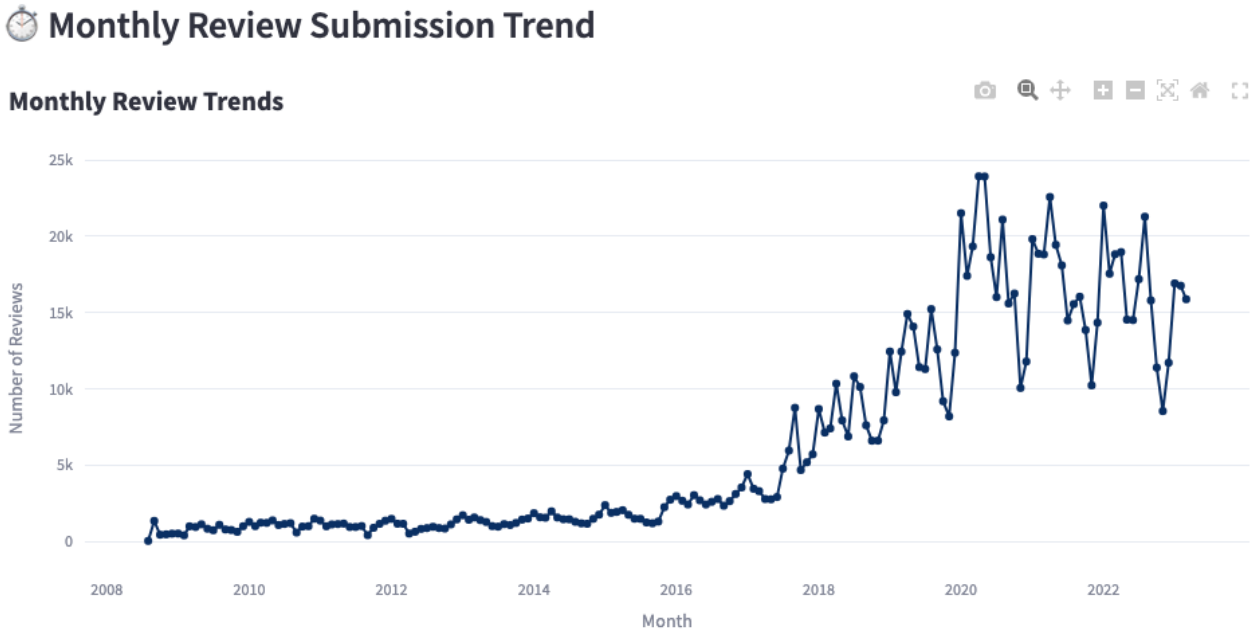

Figure 5 Summary Tab: Review Trend Time Series chart

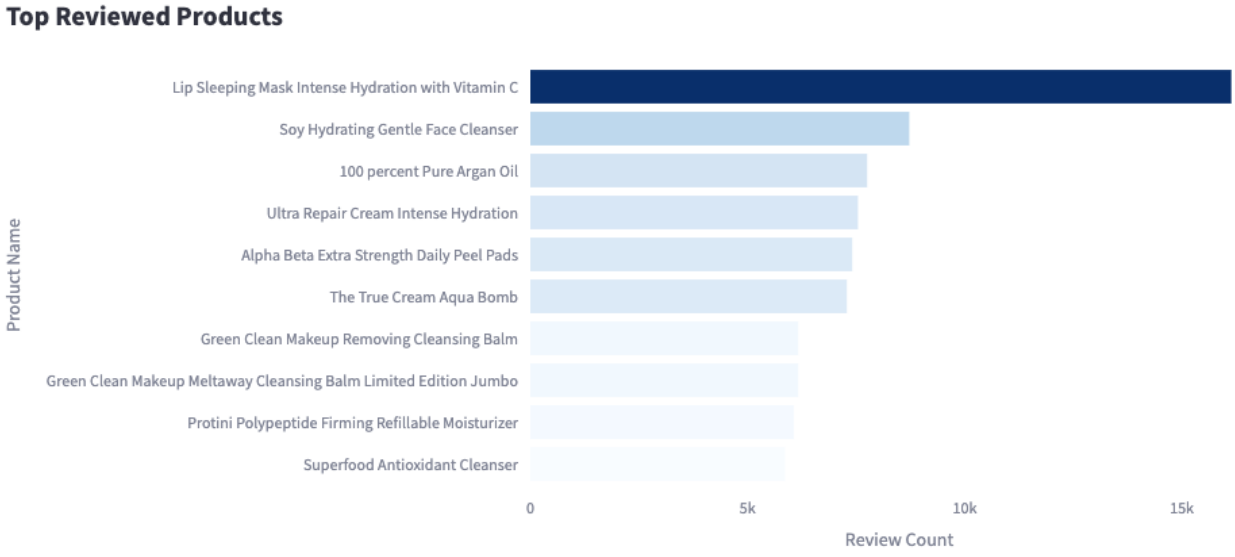

Figure 6 Summary Tab: Top Reviewed Products Bar chart

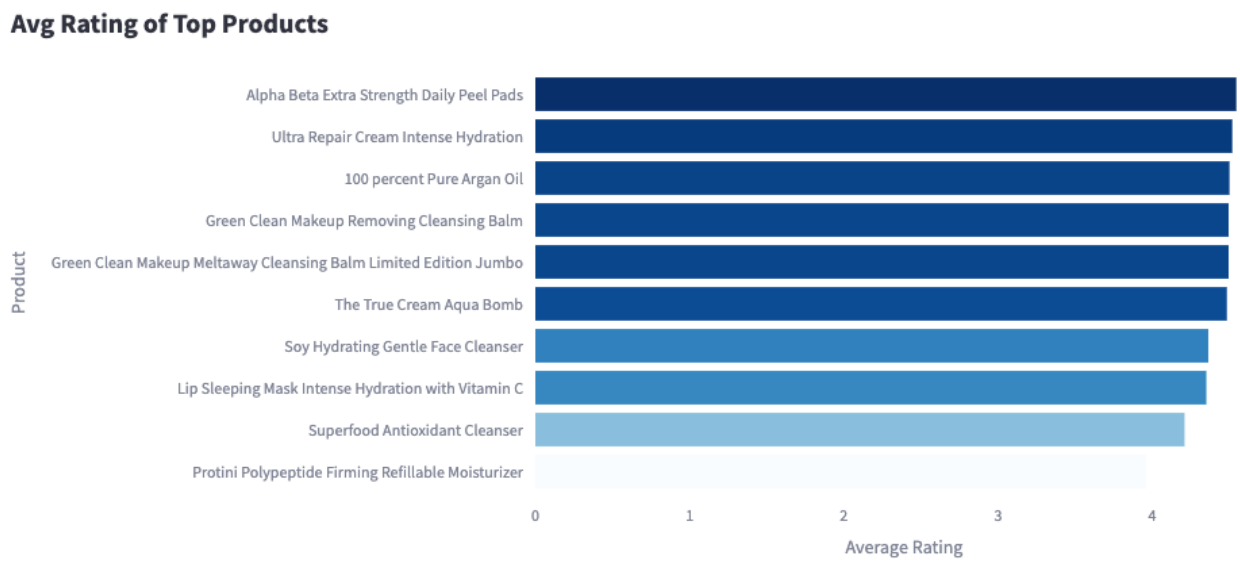

Figure 7 Summary Tab: Average Rating of Top Products Bar chart

14.2 SENTIMENT TAB

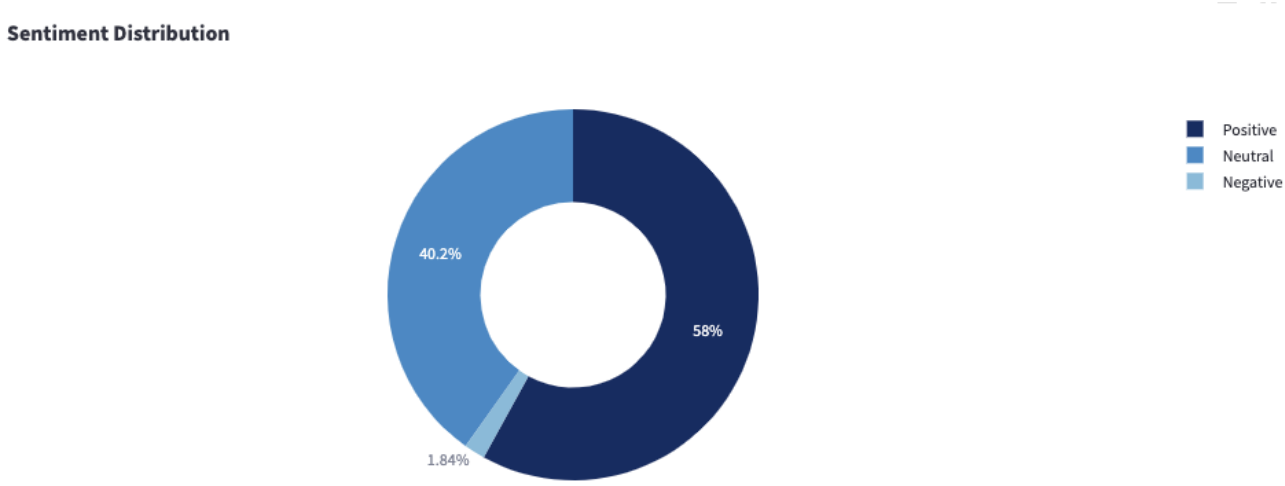

Figure 8 Sentiment Tab: Sentiment Distribution Donut Chart

Figure 9 Sentiment Tab: Sentiment Sample Review (Top 3 negative and positive reviews)

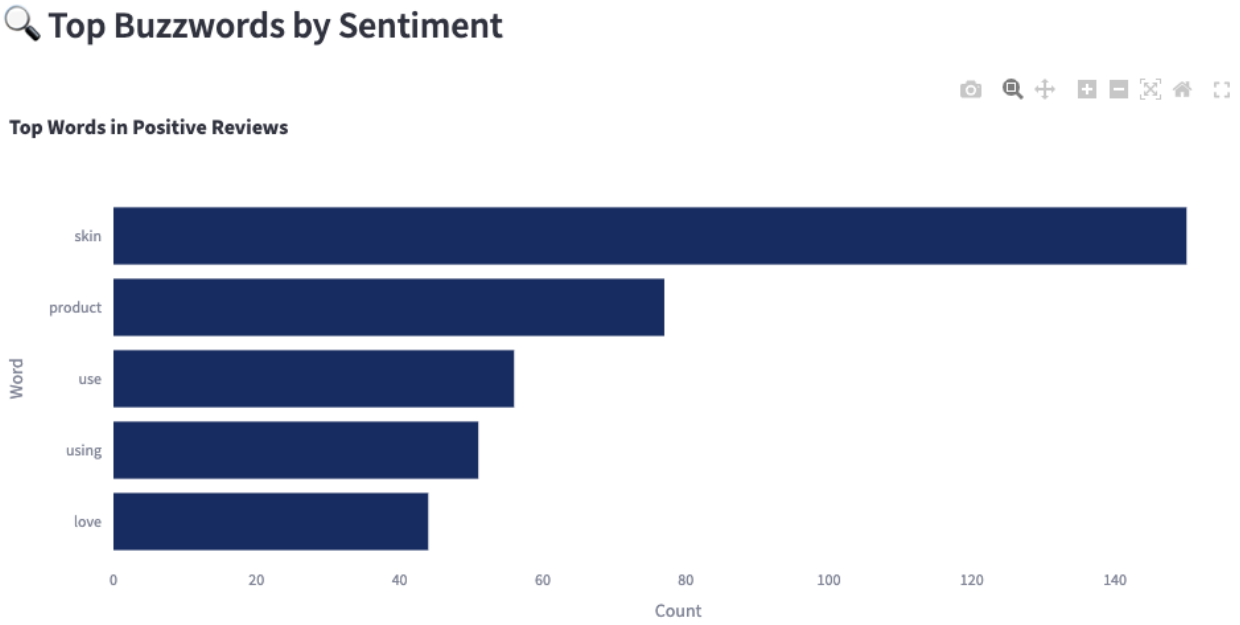

Figure 10 Sentiment Tab: Top buzzwords in Positive Sentiment (Similar buzzwords built for Neutral and Negative sentiment analysis)

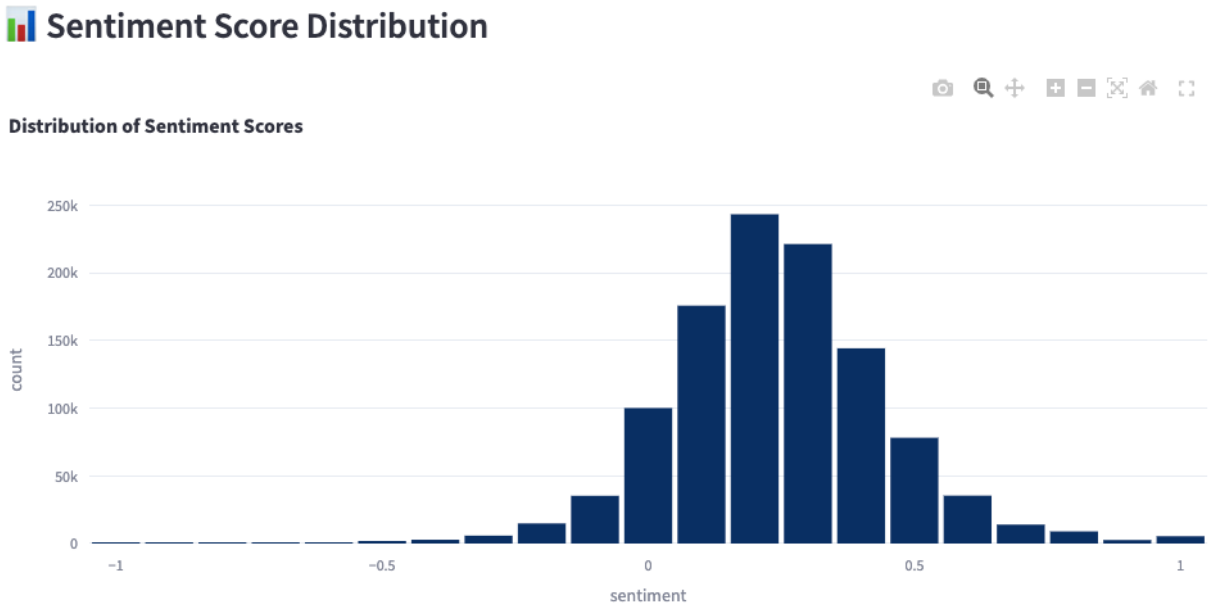

Figure 11 Sentiment Tab: Histogram to showcase Sentiment distribution

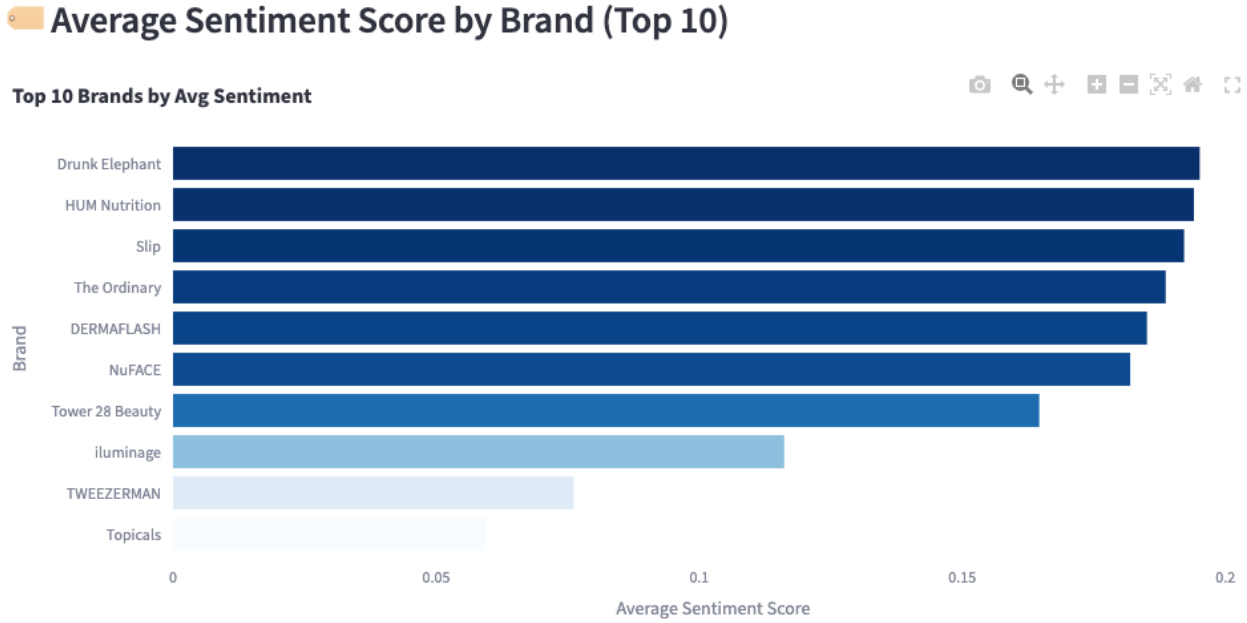

Figure 12 Sentiment Tab: Top 10 Brands by Average Sentiment Score

14.3 RATING TAB

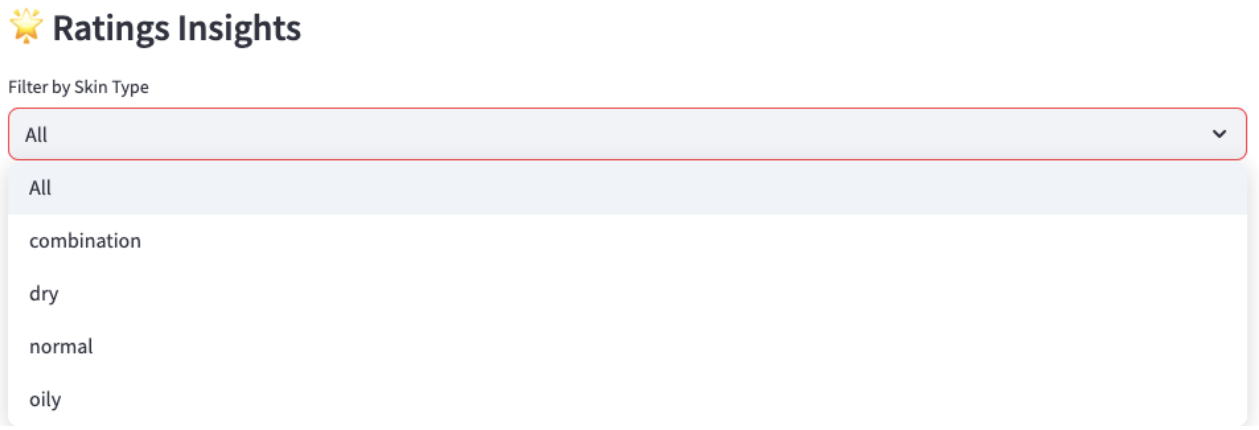

Figure 13 Rating Tab: Additional filter to filter through skin types for dermatology associated products

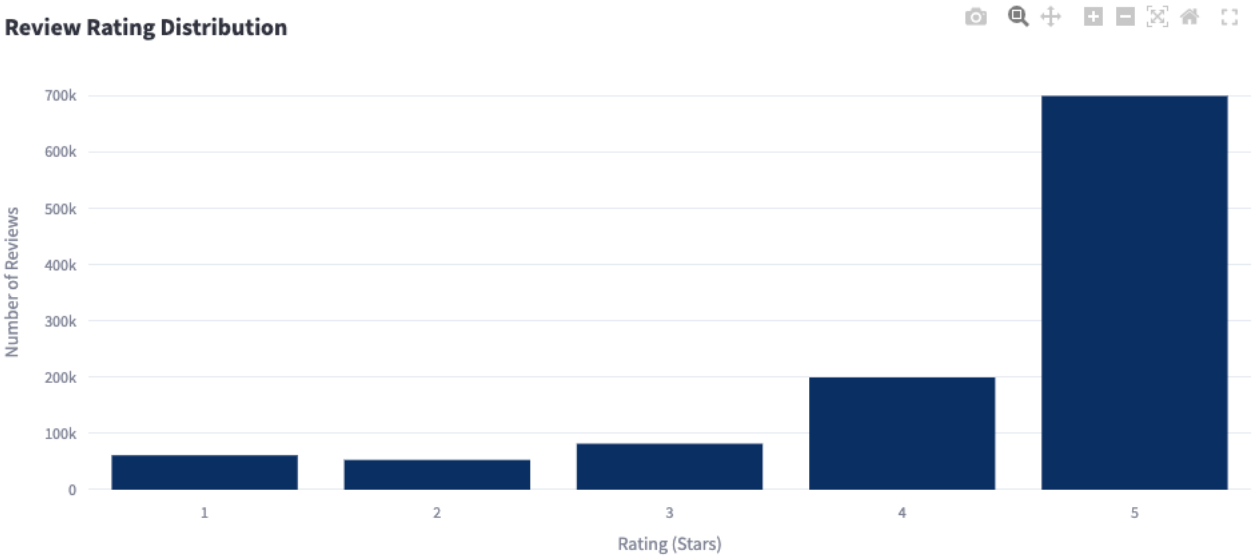

Figure 14 Rating Tab: Review Rating Distribution

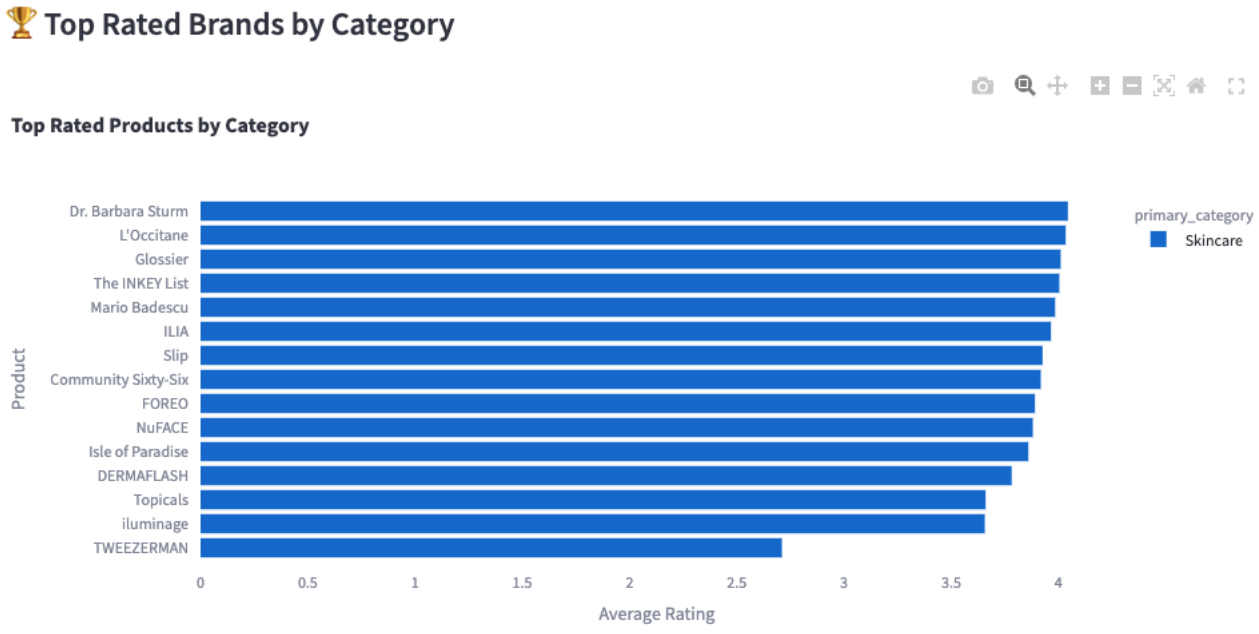

Figure 15 Rating Tab: Top-Rated Products by Category (Skincare)

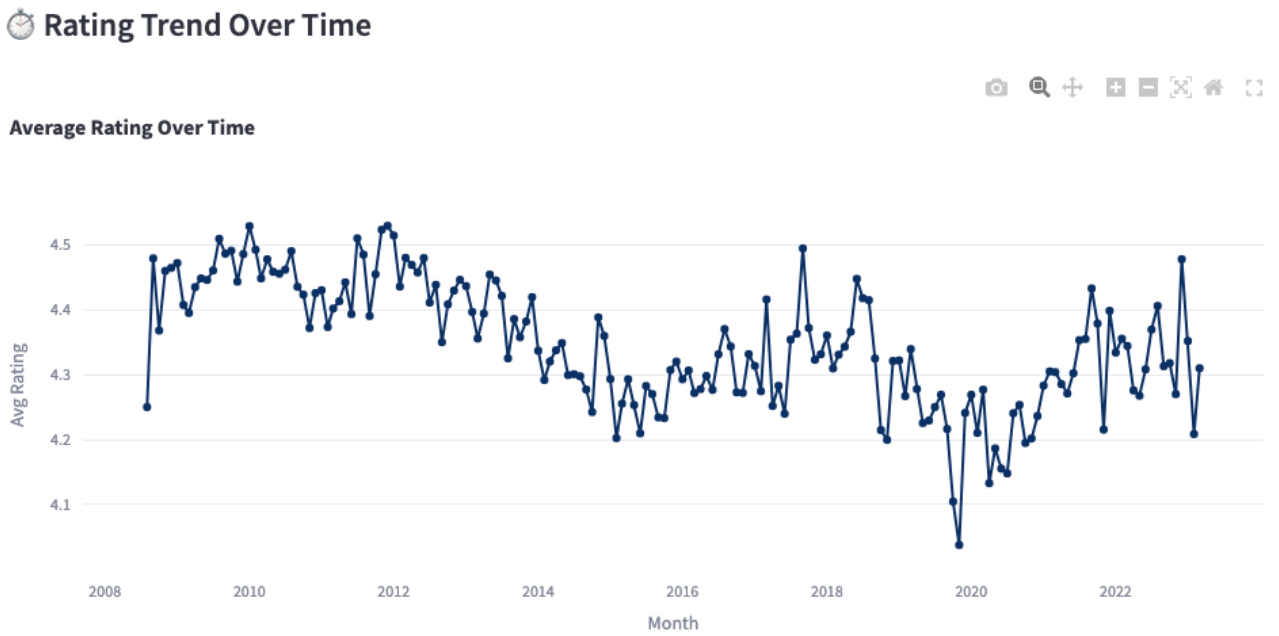

Figure 16 Rating Tab: Time Series chart for Rating trend analysis

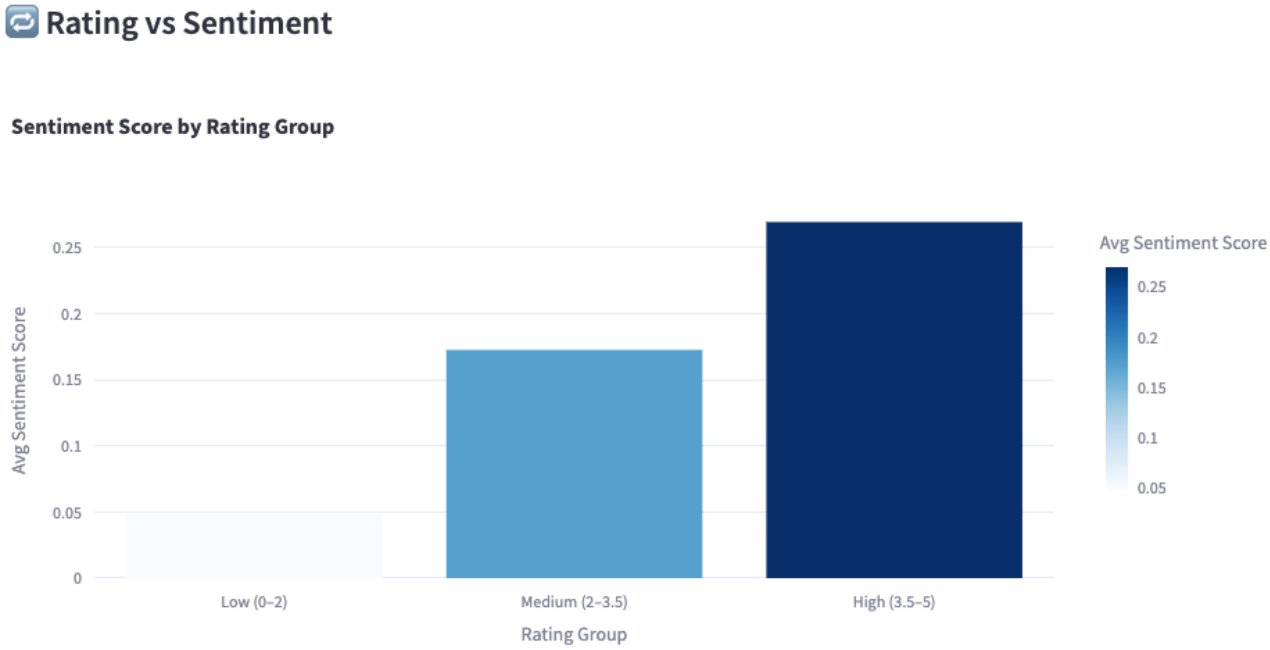

Figure 17 Rating Tab: Sentiment score vs Rating

14.4 INGREDIENTS TAB

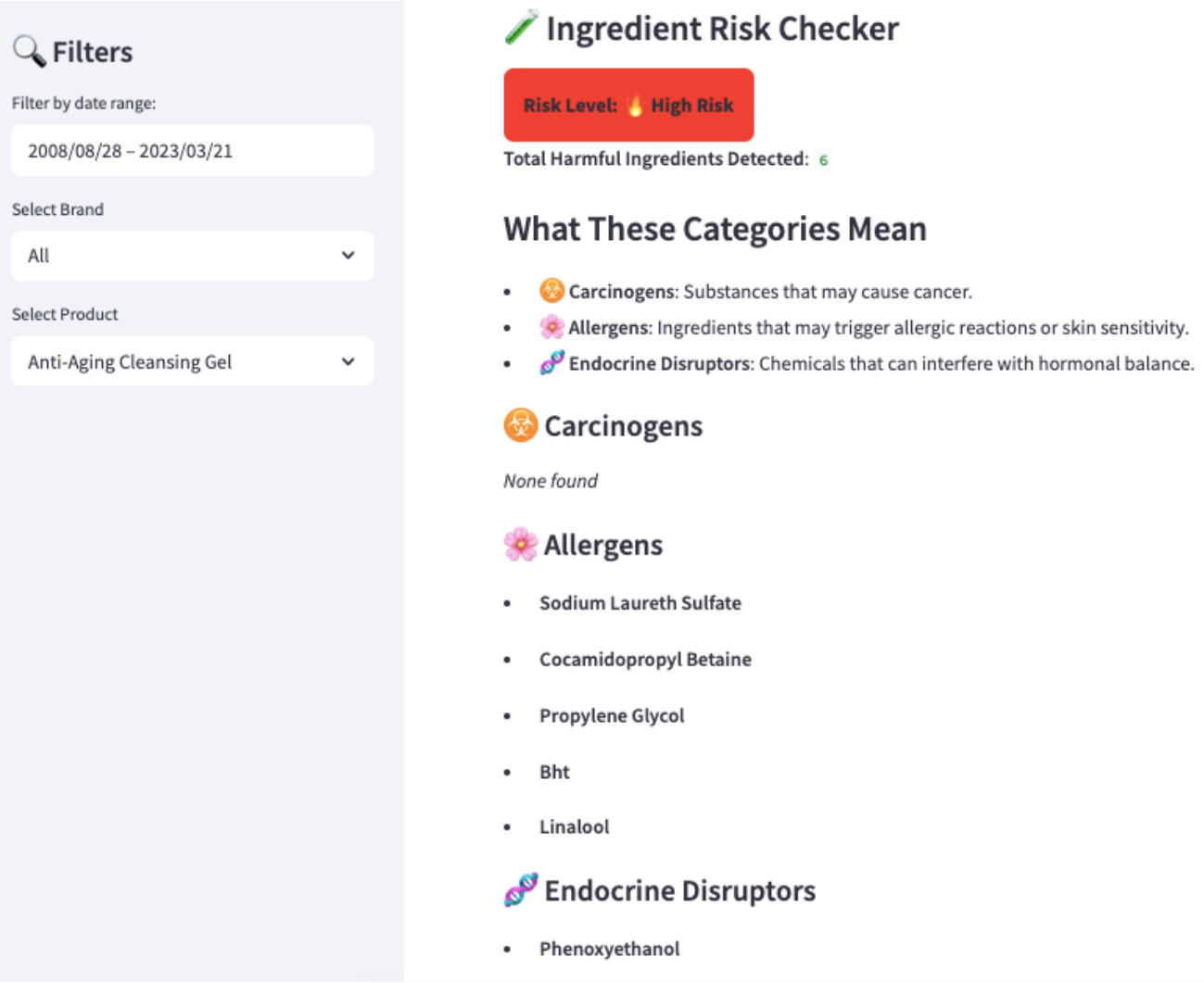

Figure 18 Ingredients Tab: A product being analyzed for harmful ingredients

14.5 WORDCLOUD TAB

Figure 19 Word cloud based on Customer Reviews