DESAUTELS FACULTY OF MANAGEMENT

INSY 669 - Text Analytics
Professor Changseung Yoo

Sephora - Save our Skin

Hon, Chelsea	261010089
Rao, Aishwarya	261008602
Samuel, Nancy	260948517
Jain, Nehal	260958896
Nadeem, Fatima	260973683

February 15, 2022

Table of Contents:

1. Introduction

The cosmetics beauty sector has been on a constant growth spurt for years as the business continues to defy gravity. Indeed, the industry revenue is predicted to exceed $680 billion worldwide by 2025 growing at an annual rate of 4.75% [11]. Even multi-brand stores, which in other consumer sectors have been struggling, are thriving within it. Forbes has outlined three major reasons behind this everlasting success. As the definition of 'beauty' is constantly changing, the consumers of the beauty sector are continuously exploring new products and brands. Moreover, with the power of social media, this consumer journey of exploration is becoming increasingly universal. It has led to a rise in creativity allowing the nurturing of young, independent brands who are supported by retail e-commerce beauty channels such as Sephora and Ulta resulting in a worldwide industry expansion.

However, due to the industry's dynamic nature, it is very important to constantly monitor the movement of trends within it in order to benefit from its prosperity. Forbes sheds light on the fastest growing beauty trends and concludes that consumers are currently infatuated with skin-care in relation to anti-aging instant fixes (e.g. wrinkles, eye bags) and natural, clean products along with heavy make-up, flawless 'doll' looks. Indeed, according to L'Oreal, despite the fact that the skin care segment contributes only 40% to the beauty market, it actually constitutes approximately 60% of the global cosmetics market growth [14]. Additionally, with reference to the US market, sales of skin-care grew by 13% last year while makeup product sales rose by 1% only [14]. Hence, it can be construed from the market analysis, that the future of the beauty industry is expected to advance predominantly towards skin-care.

2. Our Scope

Given the positive outlook of the beauty industry with the skin-care market representing its future course, we proceeded to narrow our project accordingly. However, further exploration in the skin-care sector was essential to understand the driving force behind it's growing popularity. Indeed, one of the biggest reasons behind this surge is change in user behavior as they have begun to develop a deeper relationship with the environment and their health and wellness. With increasing work stress and increasing pollution, harmful sun rays and animals, the modern day consumer is becoming increasingly worried about the inevitable consequences of these events on the human body and skin. To counteract such effects, they are leaning towards products with 'natural', 'cruelty-free', 'vegan' characteristics in an effort to be more environment-friendly and see better results. These are generally organic or plant-based products without harmful, anti-green ingredients such as parabens, synthetic colors and phthalates and which are not tested on animals. Certainly, the clean beauty hemisphere is expected to observe a rise in future demand as the global market value for natural cosmetics is projected to increase to $54 billion by 2027 [8]. Yet, despite the optimistic outlook, it is a relatively new concept and there is a lack of general purchase guides for organic, vegan products. Product ratings are somewhat indicative but do not capture the entire picture as they lose out lots of details regarding the effects of each product for particular skin types.

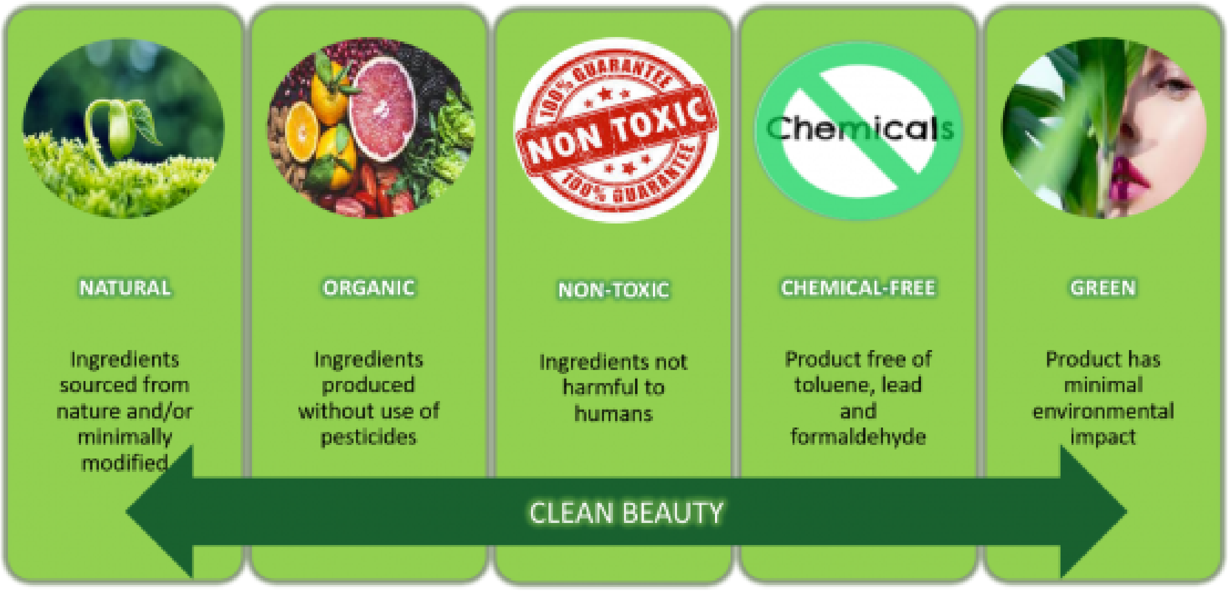

100% GUARANTEE
NON TOXIC
100% GUARANTEE
Chemicals
NATURAL
Ingredients sourced from nature and/or minimally modified
ORGANIC
Ingredients produced without use of pesticides
NON-TOXIC
Ingredients not harmful to humans
CHEMICAL-FREE
Product free of toluene, lead and formaldehyde
GREEN
Product has minimal environmental impact
CLEAN BEAUTY

Fig 1: Clean Beauty Characteristics

Therefore, observing the change in consumer thinking and behavior, we decided to pursue an analysis of the skin-care sector with a focus on clean, vegan products. The project's objective is to explore user perception of different vegan skin-care products in the Moisturizer and Eye-Cream categories through reviews and gain knowledge regarding their similarities with one another as well as their suitability for various skin types. With this information, a review-based recommendation system can be devised which would allow consumers to determine which product would be ideal for their skin. Sephora, a key player in the e-commerce cosmetics industry hosting more than 280 brands with a worldwide presence, was the ideal selection for our analysis. Presently Sephora's website does not possess a mechanism to search and filter by skin profile attributes. Moreover, it's recommendation system for users is based mainly on purchase, cart history and branding. Hence, this recommendation system would not only serve to educate users on vegan product performance but also allow companies such as Sephora to leverage this increased awareness to improve customer stickiness and boost sales.

The beauty you want, minus the ingredients you might not.

This seal means formulated without parabens, sulfates SLS and SLES, phthalates, mineral oils, formaldehyde, and more.

Feel-good formulas bursting with powerful ingredients for glowing skin.

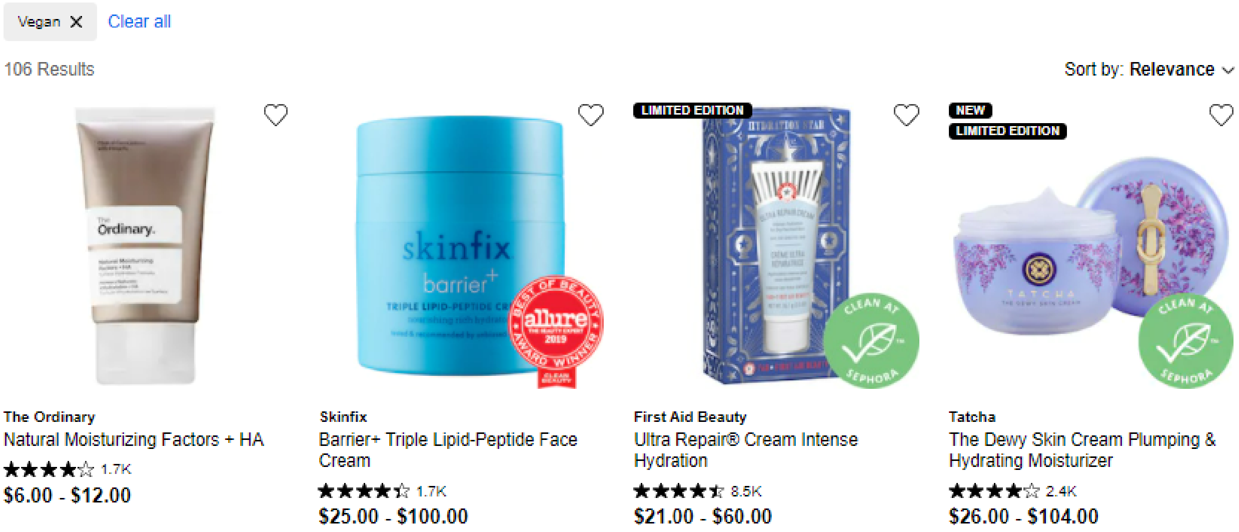
Fig 2: Clean Skin-Care at Sephora

3. Text Mining Methodology

Our objective is to mine the reviews related to the wide variety of vegan products from Sephora's online platform and look for top brands, best reviewed items, most popular attributes of the mentioned items and the sentiment related to it. In order to accomplish this, we have implemented a text-mining process that is oriented towards the steps involved in the creation of a Bag of Words model from reviews to analyze it along with other item information. In this section, we have provided the detailed framework and the need for these steps in the text mining process.

3.1. Web Extraction

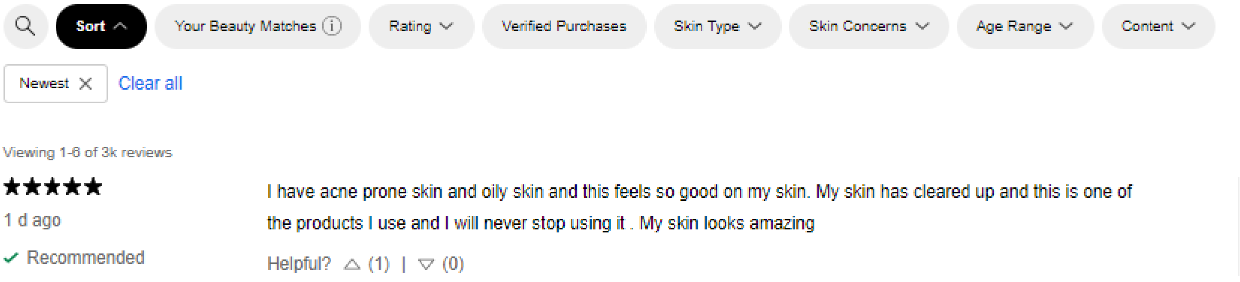
To begin our analysis on the various Vegan products, we started by extracting product information from Sephora's product page where we used the URLs for two product types- Vegan Moisturizers and Vegan Eye Creams. The item information like Item name, Brand Name, Price, link for the Item detail page etc. were extracted from the products page using Selenium webdriver. Selenium is a Python library and tool used for automating web browsers to do a number of tasks. One of such is web-scraping to extract useful data and information that may be otherwise unavailable. For extracting the reviews from the Items page, URL for each item detail page was opened and the reviews were sorted using “Newest” criteria. Then, the top 18 reviews (6 reviews per page) for each item were extracted as a list and added to the other item information as a pandas dataframe. One major issue faced while extracting the reviews was that the page becomes unresponsive after navigating to the next page of reviews intermittently. To deal with such issues, the scraping script kept the reviews extracted until then and ignored the error encountered. In total, we extracted information for a total of 90 items, snippet of which is shown in Fig 5.



106 Results

The Ordinary Natural Moisturizing Factors + HA ★★★★☆ 1.7K $6.00 - $12.00	Skinfix Barrier+ Triple Lipid-Peptide Face Cream ★★★★☆ 1.7K $25.00 - $100.00	First Aid Beauty Ultra Repair® Cream Intense Hydration ★★★★☆ 8.5K $21.00 - $60.00	Tatcha The Dewy Skin Cream Plumping & Hydrating Moisturizer ★★★★☆ 2.4K $26.00 - $104.00

Fig 3: Vegan Moisturizers - Product Page



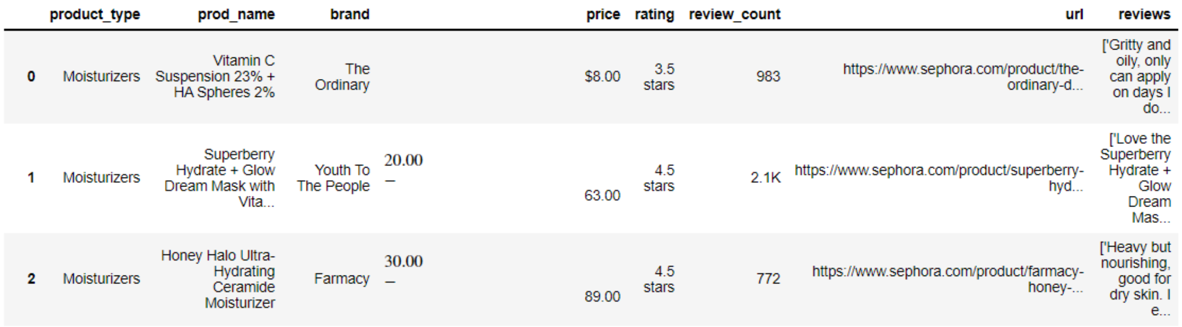
Viewing 1-6 of 3k reviews

★★★★★
1 d ago
✓ Recommended

I have acne prone skin and oily skin and this feels so good on my skin. My skin has cleared up and this is one of the products I use and I will never stop using it . My skin looks amazing

Helpful? △ (1) | ▽ (0)

Fig 4: Item Review Snippet



	product_type	prod_name	brand	price	rating	review_count	url	reviews
0	Moisturizers	Vitamin C Suspension 23% + HA Spheres 2%	The Ordinary	$8.00	3.5 stars	983	https://www.sephora.com/product/the-ordinary-d...	['Gritty and oily, only can apply on days I do...
1	Moisturizers	Superberry Hydrate + Glow Dream Mask with Vita...	Youth To The People	20.00 – 63.00	4.5 stars	2.1K	https://www.sephora.com/product/superberry-hyd...	['Love the Superberry Hydrate + Glow Dream Mas...
2	Moisturizers	Honey Halo Ultra-Hydrating Ceramide Moisturizer	Farmacy	30.00 – 89.00	4.5 stars	772	https://www.sephora.com/product/farmacy-honey-...	['Heavy but nourishing, good for dry skin. I e...

Fig 5: Scraped Item data

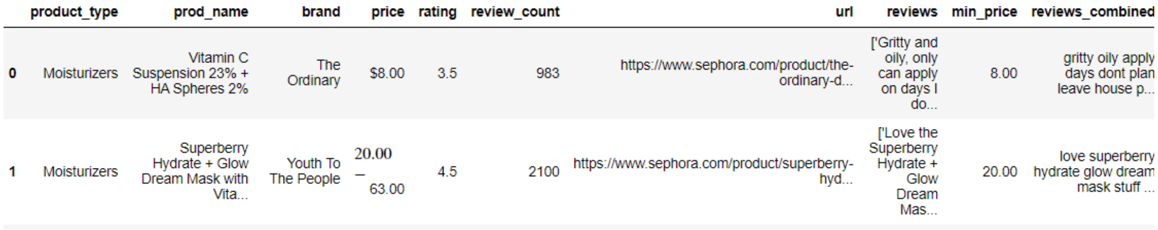
3.2. Data Preprocessing

Data preprocessing is the process of transforming raw data into a machine understandable format. It is also an important step in data mining as we cannot work with raw data and the goal is to produce text which machines can analyze without any errors. The data extracted using scraping was not consistent and clean. Hence, it required the below preprocessing steps to be performed for respective columns:

1. Price - The price contained a range for some items, so we extracted the min price for such items as a separate column ‘min_price’.
2. Rating - The rating was of string type containing word ‘stars’, which was removed and converted to a float type. Rating for items having ‘No Rating’ were considered as zero.
3. Review Count - For items having more than thousand reviews, the count was represented in ‘K’ for example 2.1K. Such counts were converted to actual numbers with data type as integer.
4. Reviews - The list of reviews for each item was converted to a single string in a separate column ‘review_combined’ and below transformations were applied to further.
    a. Normalizing text: We changed the capitalization of all review content to lowercase in order to maintain consistency throughout.
    b. Extract Alphanumeric Content: As the review data contained a lot of special characters, quoted, punctuations, unicode characters and emojis, we only extracted the alpha numeric data from the reviews.
    c. Removing Stop words: Stop words are basically a set of commonly used words in any language. Since they provide very little useful information they are usually removed from the text content. For getting the basics set of stop words we used the Gensim library’s stopwords function.

d. Lemmatization: Lemmatization is the process of converting a word to its base form or lemma. We do not use the stemming approach here as it may alter the meaning of the word and since we are looking at top attributes (words), we use lemmatization so that the meaning of the word stays intact. To implement lemmatization we use the WordNetLemmatizer module from NLTK library.

Fig 6 shows a snippet of the processed dataframe.



	product_type	prod_name	brand	price	rating	review_count	url	reviews	min_price	reviews_combined
0	Moisturizers	Vitamin C Suspension 23% + HA Spheres 2%	The Ordinary	$8.00	3.5	983	https://www.sephora.com/product/the-ordinary-d...	['Gritty and oily, only can apply on days I do...	8.00	gritty oily apply days dont plan leave house p...
1	Moisturizers	Superberry Hydrate + Glow Dream Mask with Vita...	Youth To The People	20.00 – 63.00	4.5	2100	https://www.sephora.com/product/superberry-hyd...	['Love the Superberry Hydrate + Glow Dream Mas...	20.00	love superberry hydrate glow dream mask stuff ...

Fig 6: Clean and Processed Item data

4. Approach & Analysis

Reviews are reflective of a customer's story in terms of experience. This provides social proof to other potential customers. In 2021, it was found out that 93% of customers read online reviews before buying a product [7]. And according to Yotpo, overall 94% of all purchases are made for products with an average rating of 4 stars and above [4]. Item reviews on Sephora were also very useful in determining important features or attributes of a particular item. For instance, reviews like the one below helped identify that the particular product addresses dry skin condition. It not only expressed a positive emotion but also what attribute about the product led to that positive review.

"I have dry and sensitive skin. This works to moisture the skin. I apply face oil at night, and I am happy with my soft skin when I awake! I have not tried it under makeup, but it does a good job in terms of its price!"

We analyzed such reviews from the clean dataset to explore the best products based on different criteria like popularity, ranking based on overall sentiment. Furthermore, various features associated with the items were discovered to create a targeted product guide for users.

4.1.Review-based Best Skincare Ranking

With reviews being one of the most resorted factors in deciding whether to purchase a specific product, given that they're more genuine, natural and that they're coming straight from the end-consumer, thereby adding to its authenticity, we found the need to first analyze the feedbacks given i.e., the reviews posted and understand the sentiment trend behind that product. Thus, we first performed sentiment analysis as part of our solution approach.

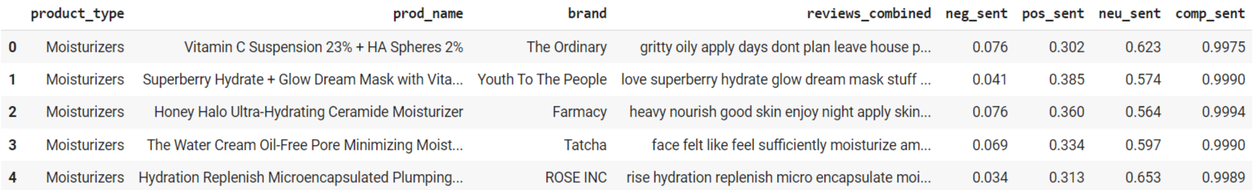
4.1.1 Sentiment Analysis

Sentiment analysis is a text analysis method that detects polarity (e.g. a positive or negative opinion) within the text of an entire document, paragraph, sentence, or clause. It aims to measure the attitude, sentiments and emotions of a speaker / writer based on the computational treatment of subjectivity in a text, that could help us transform text data into valuable customer insights: reveal trends and patterns over time, easily spot customer pain points, and make data-based decisions. [5]

Amongst the multiple sentiment analysis tools and packages available to us, we decided to proceed with the Vader (Valence Aware Dictionary for Sentiment Reasoning) model as Vader is optimized for social media data that contains a lot of informal writing - multiple punctuation marks, acronyms, emoticons and emphasis of capitalization. It focusses on a polarity-based approach, where pieces of texts are classified as either positive or negative, as well as a valence-based approach, where the intensity of the sentiment is taken into account. For example, the words ‘good’ and ‘excellent’ would be treated the same in a polarity-based approach, whereas ‘excellent’ would be treated as more positive than ‘good’ in a valence-based approach. [13]

As such, with the scraped data containing multiple metrics such as the type of the product, name, its corresponding brand, price, rating, minimum price of a product, reviews and so forth, we decided to narrow our scope to factors that aligned with our objective. Therefore, we focussed on columns such as the type of the product, its brand, the name of the product, and its reviews.

As part of our further preprocessing, we decided to drop rows that did not contain any reviews, as it would not allow us to perform meaningful sentiment analysis. Thereafter, using the polarity_scores() method for reviews of every product, we got four sentiment metrics from these reviews-split-into- lexicon ratings.



	product_type	prod_name	brand	reviews_combined	neg_sent	pos_sent	neu_sent	comp_sent
0	Moisturizers	Vitamin C Suspension 23% + HA Spheres 2%	The Ordinary	gritty oily apply days dont plan leave house p...	0.076	0.302	0.623	0.9975
1	Moisturizers	Superberry Hydrate + Glow Dream Mask with Vita...	Youth To The People	love superberry hydrate glow dream mask stuff ...	0.041	0.385	0.574	0.9990
2	Moisturizers	Honey Halo Ultra-Hydrating Ceramide Moisturizer	Farmacy	heavy nourish good skin enjoy night apply skin...	0.076	0.360	0.564	0.9994
3	Moisturizers	The Water Cream Oil-Free Pore Minimizing Moist...	Tatcha	face felt like feel sufficiently moisturize am...	0.069	0.334	0.597	0.9990
4	Moisturizers	Hydration Replenish Microencapsulated Plumping...	ROSE INC	rise hydration replenish micro encapsulate moi...	0.034	0.313	0.653	0.9989

Fig 7: Sentiment scores' distribution for each review for every product

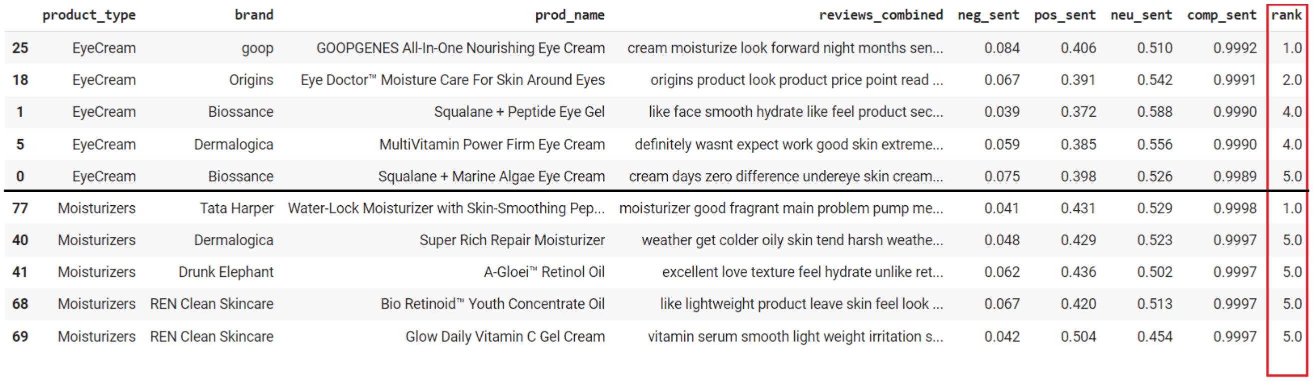
The first three scores, i.e, negative, positive and neutral represent the sentiment score of the review according to their semantic orientation. As you can see, our first review was rated as 62% positive, 30% neutral and 7% negative. The final metric, the compound score, is the sum of all of the lexicon ratings or sentiment scores (in this case 0.076, 0.302 and 0.623) which have been normalized to range between -1 and 1. In our case, the resultant score is a rating of 0.99, which is strongly positive.

Next, we filtered our reviews that have a compounded sentiment score of greater than 0.95 as that covered scenarios where the positive sentiment score was greater than that of the negative one. Moreover, since our end goal was to list the products that were most spoken in a good way, we wanted to focus on products who had reviews that were positively skewed. We saved the filtered output as a csv file (*sentiments-data.csv*) for future usage.

4.1.2 Top 5 Best Reviewed Moisturizer & Eye-cream (by-ranking)

With our review's data prepared, that contains reviews and their corresponding sentiment scores for each product, we segmented our products into two focus groups- moisturizers and eye creams.

For each of these two categories, we ranked our products based on reviews with the highest compounded sentiment score, in a descending manner, and thereby found the top 5 products that end-consumers loved. Please note that products having the same compounded sentiment score were ranked equally.



	product_type	brand	prod_name	reviews_combined	neg_sent	pos_sent	neu_sent	comp_sent	rank
25	EyeCream	goop	GOOPGENES All-In-One Nourishing Eye Cream	cream moisturize look forward night months sen...	0.084	0.406	0.510	0.9992	1.0
18	EyeCream	Origins	Eye Doctor™ Moisture Care For Skin Around Eyes	origins product look product price point read ...	0.067	0.391	0.542	0.9991	2.0
1	EyeCream	Biossance	Squalane + Peptide Eye Gel	like face smooth hydrate like feel product sec...	0.039	0.372	0.588	0.9990	4.0
5	EyeCream	Dermalogica	MultiVitamin Power Firm Eye Cream	definitely wasnt expect work good skin extreme...	0.059	0.385	0.556	0.9990	4.0
0	EyeCream	Biossance	Squalane + Marine Algae Eye Cream	cream days zero difference undereye skin cream...	0.075	0.398	0.526	0.9989	5.0
77	Moisturizers	Tata Harper	Water-Lock Moisturizer with Skin-Smoothing Pep...	moisturizer good fragrant main problem pump me...	0.041	0.431	0.529	0.9998	1.0
40	Moisturizers	Dermalogica	Super Rich Repair Moisturizer	weather get colder oily skin tend harsh weathe...	0.048	0.429	0.523	0.9997	5.0
41	Moisturizers	Drunk Elephant	A-Gloei™ Retinol Oil	excellent love texture feel hydrate unlike ret...	0.062	0.436	0.502	0.9997	5.0
68	Moisturizers	REN Clean Skincare	Bio Retinoid™ Youth Concentrate Oil	like lightweight product leave skin feel look ...	0.067	0.420	0.513	0.9997	5.0
69	Moisturizers	REN Clean Skincare	Glow Daily Vitamin C Gel Cream	vitamin serum smooth light weight irritation s...	0.042	0.504	0.454	0.9997	5.0

Fig 8: Top 5 eye creams and moisturizers ranked according to overall sentiment score

To help visualize the distribution of the individual sentiment scores of each of the top five moisturizers or eye creams better, we plotted the graphs as below.

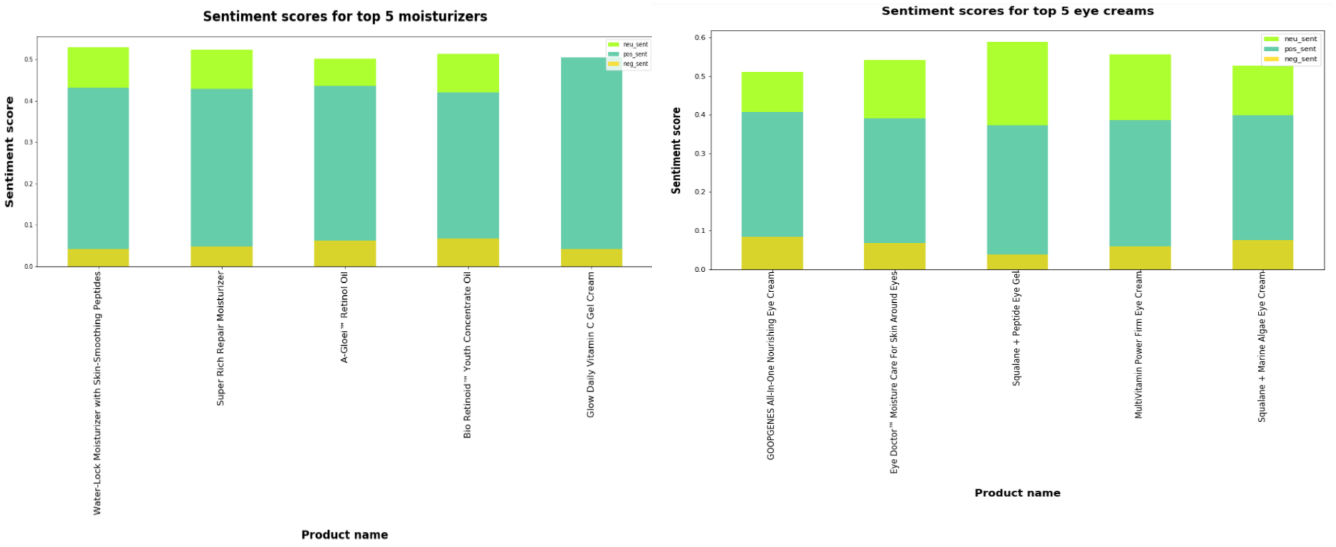


Fig 9: Sentiment score distribution for the top 5 products by category- moisturizers and eye creams

Similarly, for each of the product types - moisturizers and eye creams, we ranked brands according to the compounded sentiment score and sorted them in a descending order, resulting in our top 5 brands, the visualization is as follows.

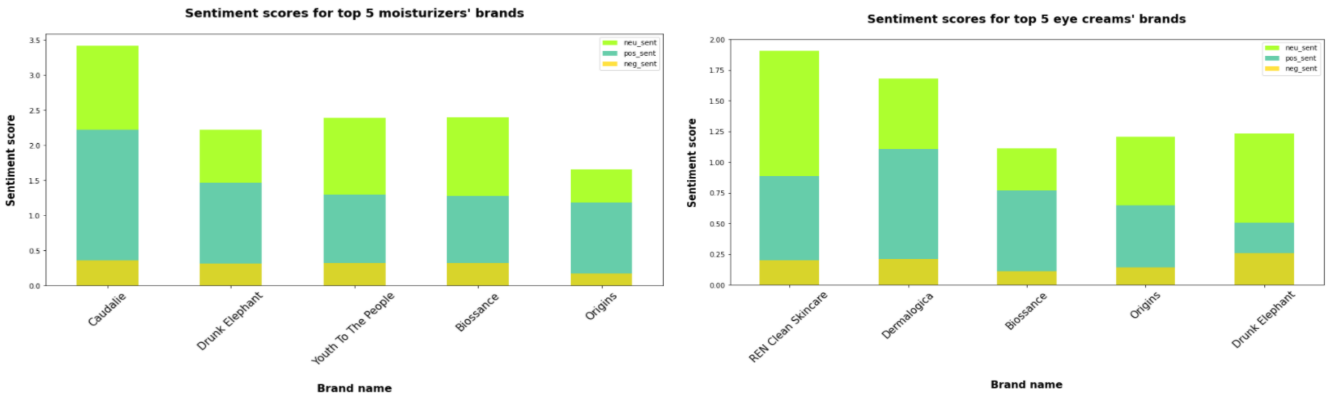


Fig 10: Sentiment score distribution for the top 5 brands by category- moisturizers and eye creams

4.2. Targeted Products Purchase Guide

The main objective of this initiative would be to help people in the individual skin care issues they are facing. Hence, a crowdsourced recommendation which would match products with their unique preferences would add value to the customers, and make Sephora a brand that cares about its customers.

This would allow for the consideration of more attributes than the traditional approach which is based on only purchase/cart history and review ratings.

4.2.1. Count Vectorization (by-most mentioned / popularity)

As our first step, once we perform sentiment analysis to filter only those products which have a combined sentiment score of greater than 0.95 to ensure that we exclude negatively reviewed products, we then go on to let our customers know what are the main topics that others are talking about. We need to tokenize and count each word in the entire extracted document of reviews. This has been implemented by the Bag of Words model which is a representation that transforms free text into fixed-length vectors by counting how many times each word appears in each document. This process is often referred to as vectorization.

For our use case, we cannot use the raw data directly since each document contains a variable length. In order to address this, we avail scikit-learn's tokenizing utility where it gives an integer id to each possible token along with occurrence counting of each token. Each individual token occurrence frequency is treated as a feature which results in a vector of such token frequencies for each document.

The final structure is a matrix with one row per document and one column per token. This entire process of turning a collection of text documents into numerical feature vectors is called vectorization and this strategy is called the Bag of n-grams representation. We should however note that this approach loses the relative position of the words in each document. The Count Vectorizer model from sklearn implements both tokenization and occurrence counting in a single class. The default parameters result in tokenizing words of at least 2 letters. It also has a parameter of ngram_range=(1, 2) by which we can extract 2-grams of words in addition to the individual words. However, we had to exclude this to account for computational resources.

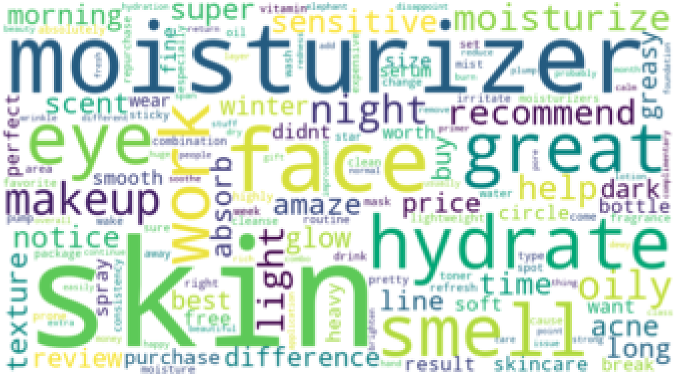


Fig 11: Word Cloud

As part of this count vectorization process, we find that the most frequent words that people talk about are skincare in terms of hydration, sensitivity; what are the features of the cream - are they dewy, are they misty, moisturizing;

regarding emotion - if they were amazed or just nice or if they loved the product and how functional it was in solving their issues like acne, clogged pores etc. So these were our assumptions of the main categories, however we wanted to use a more technically structured approach to find the topics that these words are associated with, and topic modeling helps in that.

4.2.1. Topic Modeling (by-trending topics)

Technology has developed some powerful methods which can be used to mine through the data and Topic Modeling is one such technique. It is a type of statistical model for discovering the abstract topics that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. [12] Topics can be defined as a repeating pattern of co-occurring terms in a corpus and are very useful for the purpose for document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection.

Latent Dirichlet Allocation (LDA) is the most popular topic modeling technique. LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. Given a dataset of documents, LDA backtracks and tries to figure out what topics would create those documents in the first place.

LDA is a matrix factorization technique. In vector space, any corpus (collection of documents) can be represented as a document-term matrix. LDA converts this Document-Term Matrix into two lower dimensional matrices which provides topic word and document topic distributions, However, these distributions need to be improved, which is the main aim of LDA. LDA makes use of sampling techniques in order to improve these matrices. It iterates through each word for each document and tries to adjust the current topic – word assignment with a new assignment. After a number of iterations, a steady state is achieved where the document topic and topic term distributions are fairly good.[10]

Before passing the reviews to the LDA model, we preprocessed the data to remove some more stopwords and performed tokenization. This was passed to the model to find 8 topics and 500 iterations. We decided on 8 topics to reach a number which would be able to bring attributes which were not majorly repeated across these topics. We chose the top 10 words from each of the 8 topics to understand what the categories of topics are. We also plotted an interactive plot with the objective to understand how closely related are these related to each other, but mainly to understand the distribution of the words inside each topic.

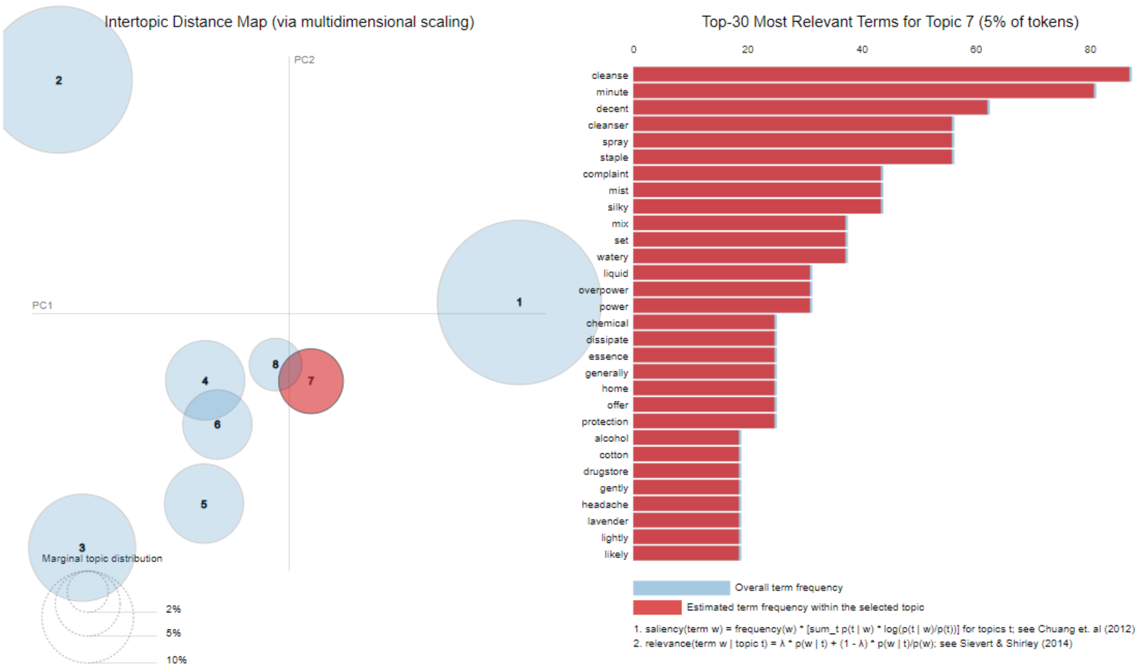


Fig 12: Plot of the associations between the various topics and their relevant terms inside

4.2.2. Insights from Association Analysis

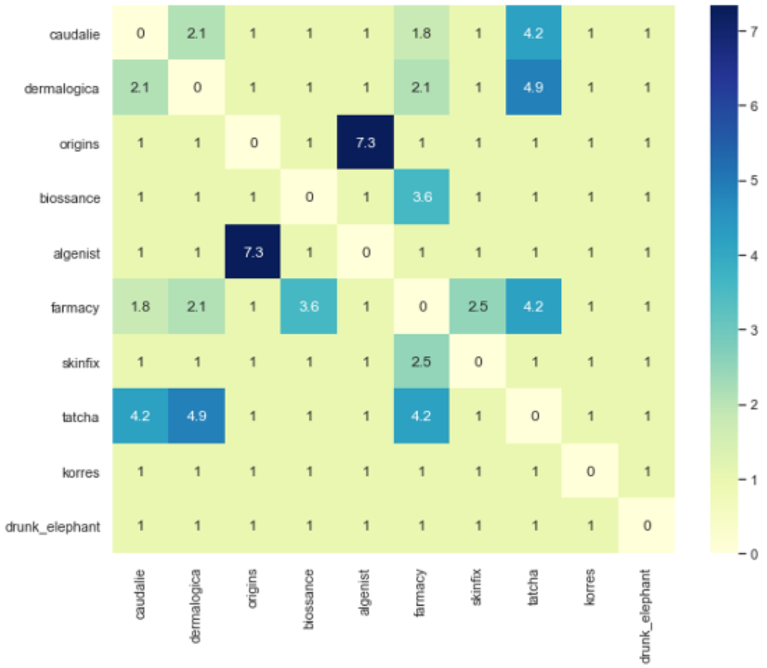
After understanding the words inside each topic and to move forward with our business strategy, we believe that Sephora can implement a concept similar to product lines. The direction would be to have topic lines, which would signify the main categories that people talk about. And since our product is healthy and vegan focussed, we associated these topic lines to the main spiritual traits that animals are known for. For example, If we talk of the product line 'fish' which is highlighted in the above figure, we sense an aquatic feeling and words like cleansing, misty, liquid, watery etc come in that category. Another topic line can be 'pandas', you would find that they are more associated with sleeping, so words like dark circles, puffiness, sleep come in that category. Additionally, we also have a product topic association which lets us know the main products which come under each category. Sure enough, our results tie in with our marketing strategy as the products under each topic are a perfect fit e.g. Blue Light Protect + Set Mist is a product under the topic line 'fish'.

Thus, these topic lines can be used as a marketing strategy to get Sephora's content in front of a healthy driven audience, cultivate a healthy living with skin instead of concerns, raise a positive impression about its products which can further be used to target a very specific group of people, boost its unique selling point, and use hot trends and topics to its advantage.

4.2.3. Lift Ratio associations

Unlike our previous assignment on Edmund's review which centralizes all discussion on entry-level luxurious sedans, Sephora's user reviews fall under each product and rarely mention a brand name, not to say a product name like the make and model on Edmund's. To explore if there is any potential brand association and better understand the ecosystem of vegan skin care products, we twisted the approach slightly to add corresponding brand and product name into the review before lift value ratios calculations. The heat map below shows the mode of lift ratios of 1, meaning there is no direct (positive or negative) association between the brands. This can also justify that further analyzing brands association and plotting MDS are not as essential for insights in our case.

Nonetheless, the highest lift ratio was observed in Algenist and Origins, followed by Tacha with three other brands - Dermalogica, Farmacy and Caudalie. When discussing continuity of brand collaborations, Sephora may take into consideration their potential competitions and synergies.



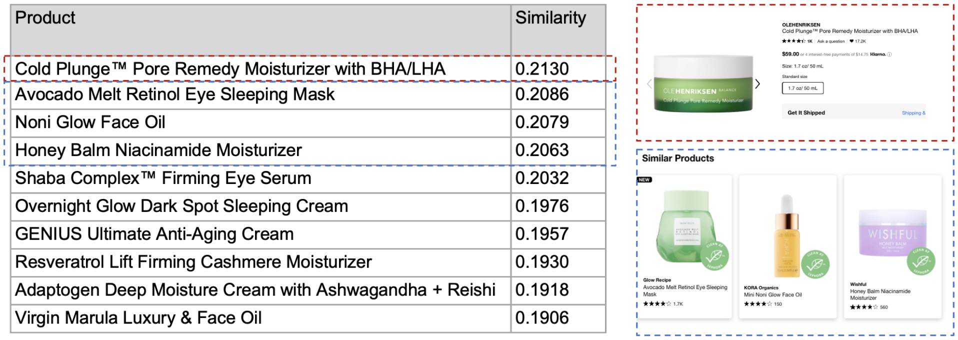
	caudalie	dermalogica	origins	biossance	algenist	farmacy	skinfix	tatcha	korres	drunk_elephant
caudalie	0	2.1	1	1	1	1.8	1	4.2	1	1
dermalogica	2.1	0	1	1	1	2.1	1	4.9	1	1
origins	1	1	0	1	7.3	1	1	1	1	1
biossance	1	1	1	0	1	3.6	1	1	1	1
algenist	1	1	7.3	1	0	1	1	1	1	1
farmacy	1.8	2.1	1	3.6	1	0	2.5	4.2	1	1
skinfix	1	1	1	1	1	2.5	0	1	1	1
tatcha	4.2	4.9	1	1	1	4.2	1	0	1	1
korres	1	1	1	1	1	1	1	1	0	1
drunk_elephant	1	1	1	1	1	1	1	1	1	0

Fig 13: Lift Ratios Matrix for Top 10 Brands Mentioned

4.2.4. Cosine Similarity

Cosine similarity measures how reviews are similar geometrically as angles between vectors in multidimensional space and mathematically as dot product of vectors irrespective of sizes. In this project, we conducted an overall cosine similarity analysis to identify the most similarly reviewed products and obtained the following top 10 products of highest cosine similarity score. From a scale from 0 no similarity to 1 exactly alike, we observed that Olehenriksen's Cold Plunge Pore Remedy Moisturizer is similar with Glow Recipe's Avocado Melt Retinol Eye Sleeping Mask, Kora Organics' Noni Glow Face Oil and Wishful's Honey Balm Niacinamide Moisturizer.

A potential use case here could be suggesting products 'you might be also interested in' with the similarity score that when users land on product detail pages. For example in Cold Plunge Pore Remedy Moisturizer product detail page, the three aforementioned products will be populated instead of the existing Sephora Collection's Nourishing Moisturizer with Prebiotics, Olehenriksen's C-Rush Vitamin C Gel Moisturizer and Herbivore Aquarius Pore Purifying BHA Cream. Still, we need to determine if we want to base such a recommendation around one product or products shortlisted, rather than overall similarity on reviews. With that being said, our team would consider this analysis as exploratory work only.



Product	Similarity
Cold Plunge™ Pore Remedy Moisturizer with BHA/LHA	0.2130
Avocado Melt Retinol Eye Sleeping Mask	0.2086
Noni Glow Face Oil	0.2079
Honey Balm Niacinamide Moisturizer	0.2063
Shaba Complex™ Firming Eye Serum	0.2032
Overnight Glow Dark Spot Sleeping Cream	0.1976
GENIUS Ultimate Anti-Aging Cream	0.1957
Resveratrol Lift Firming Cashmere Moisturizer	0.1930
Adaptogen Deep Moisture Cream with Ashwagandha + Reishi	0.1918
Virgin Marula Luxury & Face Oil	0.1906

OLEHENRIKSEN
Cold Plunge™ Pore Remedy Moisturizer with BHA/LHA
★★★★½ 1K | Ask a question | ♥ 17.2K
$59.00 or 4 interest-free payments of $14.75 Klarna. ⓘ
Size: 1.7 oz/ 50 mL
Standard size
1.7 oz/ 50 mL
Get It Shipped Shipping &

Similar Products

NEW
Glow Recipe
Avocado Melt Retinol Eye Sleeping Mask
★★★★☆ 1.7K

KORA Organics
Mini Noni Glow Face Oil
★★★★☆ 150

Wishful
Honey Balm Niacinamide Moisturizer
★★★★☆ 560

Fig 14: Top 10 Products of Highest Cosine Similarity Score

4.2.5. MDS (by-relevance)

<u>Brand Association</u>

To visualize the similarities between brands,a MDS plot was constructed. A dissimilarity matrix for the top 10 brands was formed by inverting their lift values. The MDS function was configured with default parameters and through it the matrix was transformed into a two dimensional array. After converting the transformed values into a dataframe and adding the subsequent brand labels, a K-means clustering was performed and the brands were categorized into five clusters as shown by the plot below. From the cluster visualization we can infer which brands are positively associated and similar with one another. Since the vegan beauty industry is relatively young, most brands are new and their association can be assumed to stem from working with similar ingredients or within the same product categories.

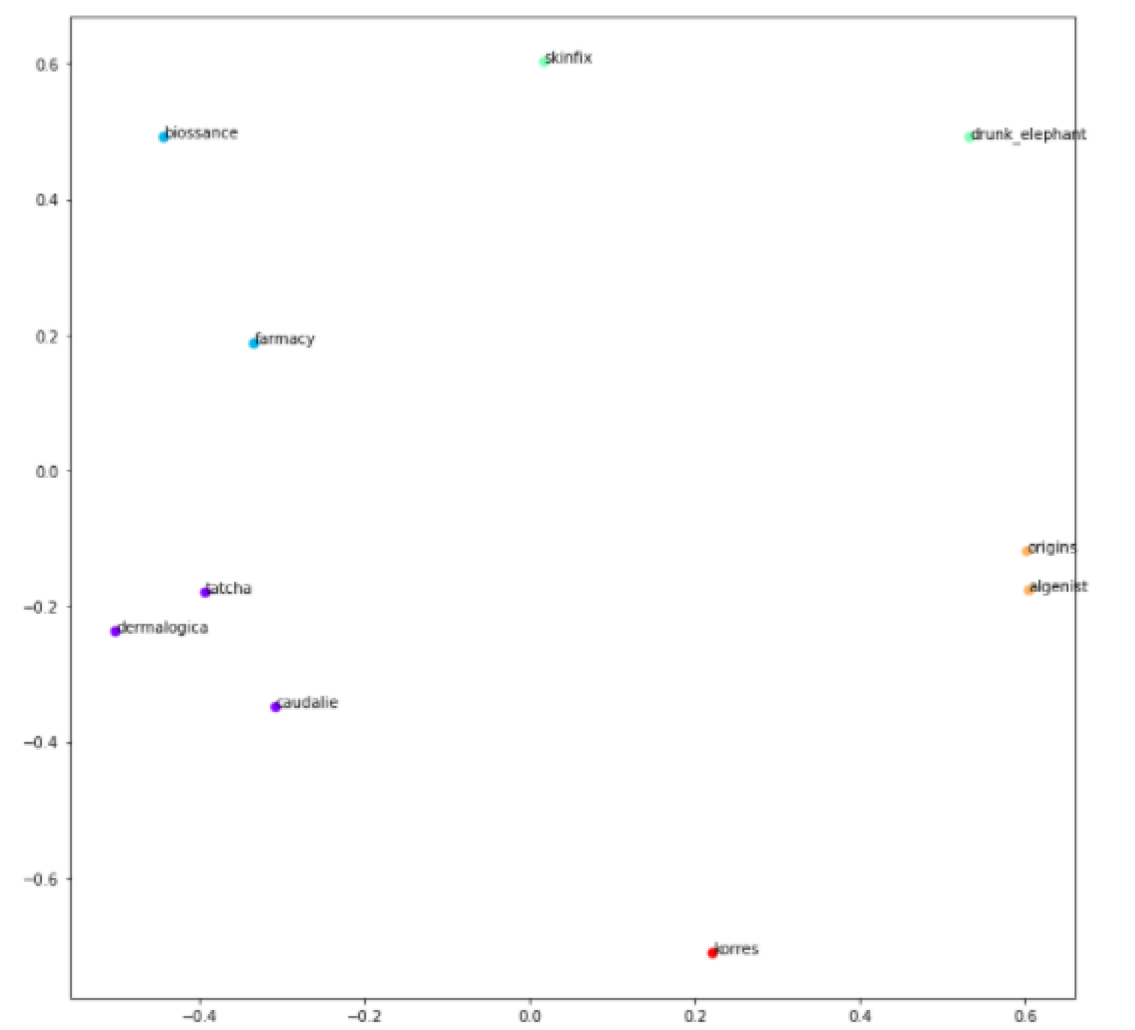


Fig 15: MDS plot of Association between Top 10 Brands

<u>Product-Topic Association</u>

Likewise, another MDS plot was formulated to visualize product topic similarity in a two dimensional space. First the distribution values of products for each topic were standardized using Standard Scaler to ensure symmetry. Following it, a dissimilarity matrix for all the products and topics was formed by inverting their standardized distribution values. As in the previous plot, it was configured with default parameters and the matrix was transformed into a two dimensional array. After converting the transformed values into a dataframe and adding the subsequent brand labels, a K-means clustering was performed and the brands were categorized into three clusters as shown by the plot below. However, since most of the products lie very close together, it inhibited us to derive further insights.

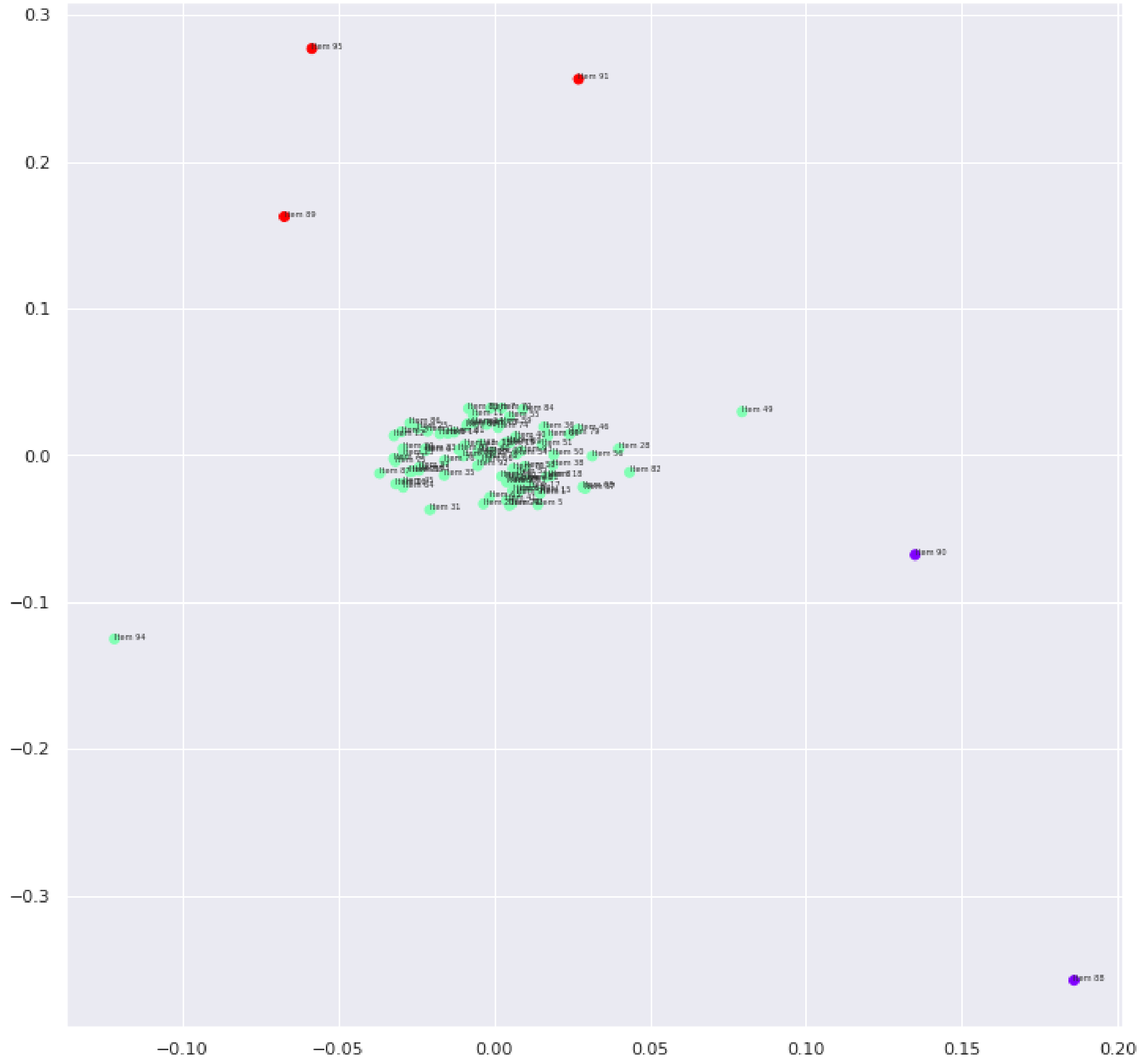


Fig 16: MDS plot of Association between Products and Topics

4.2.6. Recommendation System for Customers - Product Clustering

Bringing together all the analysis done from different perspectives, a recommendation system for customers can be developed. It will be established on crowdsourced recommendation matching products with customer preferences and thus allows the consideration of more attributes than the traditional approach which is based on only purchase/cart history and review ratings. Initially the products will be filtered on sentiment score so that only those who have a score of 0.95 or higher are incorporated. Next, each product's association with the skin profiles outlined in the topic modeling analysis will be calculated and clustered through the K-means and DBSCAN algorithms so products with similar associations for each skin profile are outlined and can be offered as recommendations. This could be done both as an extension for a particular product type in the same skin profile or as a bundle recommendation consisting of complementary products.

This concept can be illustrated through the following example of two different kinds of scenarios for two users. User 1 currently suffers from dry skin and is looking to explore clean moisturizers. In this case, products in the moisturizer category which lie in the dry skin profile (elephant) cluster will be ranked and recommended accordingly. On the other hand, User 2 is looking for a product in the eye-cream category to treat her dark circles (pandas) but wishes it to match her moisturizer's skin profile. For such scenarios, a bundle of products will be recommended based on overlapping skin profiles and may encompass various brands.

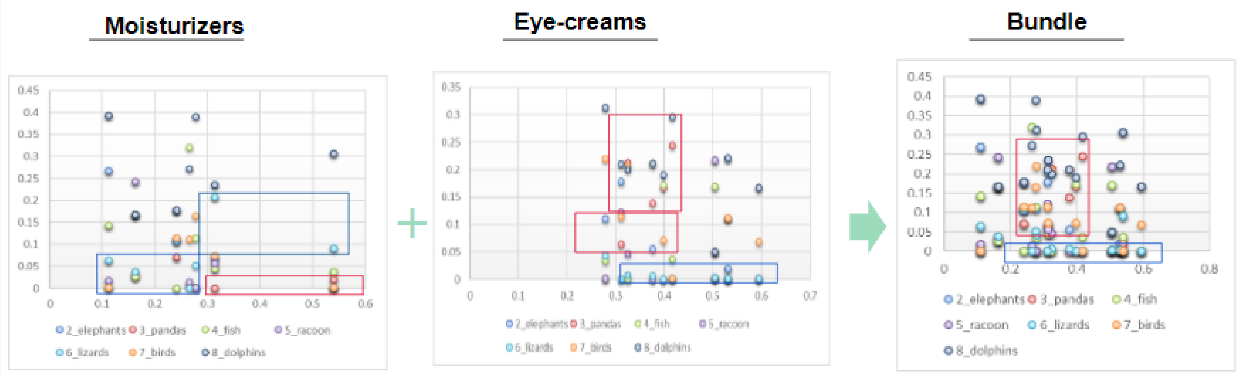


Fig 17: Illustration of product clustering for moisturizers (User 1) and eye creams (User 2)

Consequently, a recommendation system has been generated which is based wholly on reviews and can be immensely helpful for consumers who are not only able to link their skin types with products but can also obtain reassurance on the fact that the subsequent recommendations have been outlined and corroborated through evaluation of past user experiences.

5. Growth Strategy for Sephora

5.1 A Three-phase Customer-centric Marketing Strategy

With text analytics as an enable, we can formulate a customer centric marketing strategy for Sephora in three phases. In the short term, we want to gauge customers with review-based best skincare awards and be the most trusted beauty blog setting the trend by quarterly/half-yearly updated ranking so customers can explore many new high-quality beauty products. The success story of Cosme Cosmetics Ranking (https://www.cosme.net/bestcosme/) from Japan has inspired us to extend such influential online beauty events to the skincare sector, and showcased how real users' voices can be consolidated into rankings and awarded products surge further in popularity or even sell out in shops after the announcement. Leveraging its online community of 52M users and 270M+ posts[1], Sephora shall take vertical integration initiatives to create the buzz beyond the insider platform, apart from key opinion leaders (commonly referred as KOL) sharing their Sephora top 10 recommendations, and most importantly before other platforms monopolize in the North America market.

In the medium term, Sephora shall nurture their customers through targeted recommendations on the right alternatives/bundled products, not simply based on purchase history but also user reviews along their journey from exploration, acquisition, usage and different life stages. Sephora has taken steps to provide the Beauty Insider Community platform for sharing beauty advice, inspiration, news and recommendations with real users and encourages users to explore recommendations from people like us. Yet, a more catalytic and systematic way to streamline identification of customer needs and fulfillment of the needs with Sephora products is to be built from a sales perspective. In addition to ecommerce common practice recommending products purchased together with items saved / in cart and products purchased by users who shared similar beauty profiles they declared, Sephora can know the customers better than themselves based by analyzing their review contents, overlay another 'topic modeling' profiling dimension, and recommend the right bundled products that are cross-categorical while associating with that topic or the right alternative products that are either closer together when customers are comparing products or further apart when they are not satisfied and need something different.

Ultimately, customer centricity can be achieved by user-driven innovation. This is not bounded by Sephora platform but extendable to the industry trend. Beauty product manufacturers need to focus on research and development (R&D) and brand collaborations so as to fill the gaps between the dots in the

[1] Official statistics as of 15 Feb as published live at https://community.sephora.com/

product-topic scatter plot. We are aware of the uniqueness of skin and so as the unique need of skincare products. If customers still struggle to find the right products after trying out the targeted recommendations, it would be a corporate social responsibility for major players like Sephora to care for the individual stakeholders while generating revenue from the mainstream. Back in the days when veganism was not a trend, voices of vegan people had to be heard before more vegan products were offered and more people were willing to adopt and support. Through such user-centric innovations, customers can enjoy more product options which are tailored to their needs, better understand themselves, and co-create the beauty trend for the common good of all beings and our planet.

5.2 Way Forward (by-price, rating, sales and more)

The above strategy has led to an extended business discussion regarding the best recommendation system in the beauty industry. Throughout this report, we have introduced different dimensions and tools including by-ranking sentiment analysis, by-popularity frequently mentioned words cloud, by-association lift ratios, by-trending topics modeling, by-similarly cosine similarity and by-relevance product-topic clustering around the domain of text analytics.

In reality, companies also need to consider business operations like purchase history, prevailing products, sales target, partnership commitments, as well as customer experience and take-up like product satisfaction and digital behaviors. Some quick-wins for Sephora could be integrating text analytics insights as rule-based criteria in their existing product recommendation decisioning, as illustrated in our analysis applying filters on combined sentiment scores greater than 0.95 before any recommendation, and offering such filter dimensions on the e-commerce platform to let users decide on which factor they value more in product search.

The final answer to the most optimal recommendation algorithm might take further analytics and machine learning to evaluate how these factors play a significance in success metrics and shall be weighed varied across companies or industries in generating insights of best Sephora recommended.

6. References

1. (n.d.). Retrieved from Investigate AI: https://investigate.ai/investigating-sentiment-analysis/comparing-sentiment-analysis-tools

2. Ahiza Garcia. (2019, May). *Skincare industry* . Retrieved from CNN Business: https://www.cnn.com/2019/05/10/business/skincare-industry-trends-beauty-social-media/index.html

3. Afaf Athar. (2021, Jan). *Sentiment Analysis: VADER or TextBlob*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/01/sentiment-analysis-vader-or-textblob/

4. Aimee Millwood. (n.d.). *How Star Ratings Influence Customers' Behavior*. Retrieved from Yotpo: https://www.yotpo.com/blog/star-ratings-influence-customers/

5. Beri, A. (2020, May). *SENTIMENTAL ANALYSIS USING VADER*. Retrieved from https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

6. *Cosmetics Industry in the U.S. - statistics & facts*. (2022, Feb). Retrieved from Statista: https://www.statista.com/topics/1008/cosmetics-industry/#dossierKeyfigures

7. Diana Kaemingk . (2020, Oct). *Online reviews statistics to know in 2022*. Retrieved from Qualtrics XM: https://www.qualtrics.com/blog/online-review-stats

8. Jonas Sickler. (2021, July). *Beauty Industry: Cosmetic Market Share, Trends, and Statistics*. Retrieved from Terakeet: https://terakeet.com/blog/beauty-industry/

9. Richard Kestenbaum. (2019, Sep). *The Biggest Trends In The Beauty Industry*. Retrieved from Forbes: https://www.forbes.com/sites/richardkestenbaum/2018/09/09/beauty-industry-biggest-trends-skin-care-loreal-shiseido-lauder/?sh=111a3cc86982

10. Shivam Bansal . (2016, Aug). *Beginners Guide to Topic Modeling in Python*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

11. Statista. (2021, Dec). *Beauty & Personal Care*. Retrieved from Statista: https://www.statista.com/outlook/cmo/beauty-personal-care/worldwide#revenue

12. *Topic Modeling*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Topic_model

13. *Using VADER to handle sentiment analysis with social media text*. (2017, April). Retrieved from T-redactyl:

https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

14. *Why Is The Skincare Market Growing At Such Breakneck Speed*. (n.d.). Retrieved from Automat AI: https://automat.ai/resources/skincare-market-growth