Guiding Radiance: Making Informed Skin Care Decisions by Navigating Ingredients and User Experiences in SPF Face Creams

Executive Summary

This project's paper shows the alignment between marketing claims and user experiences in SPF[1] face creams, focusing on the top ten products of 2024. Leveraging sentiment analysis and topic modelling techniques, user reviews from both online forums and reputable sources are analyzed to uncover insights into consumer perceptions and preferences. The study aims to bridge the gap between marketing promises and real-world user experiences, empowering consumers to make educated skincare choices. The data collection process involves web scraping data from The Independent online newspaper to identify the top ten SPF face creams, and retrieving data from the Reddit API for user experiences with those creams. Initial findings suggest a notable variance between marketing claims and user experiences, with sentiments ranging from positive to neutral, and negative. While some products receive praise for affordability and effectiveness, others face criticism for issues like white cast residue and greasiness. These insights highlight the importance of considering authentic user feedback in skincare decision-making.

Keywords

Sunscreen face products, web scaping, Reddit API, sentiment analysis, and topic modelling

1. Introduction and Motivation

The main goal of this project report is to identify the top ten best sunscreen protective factors for the face, analyze reviews from people who have used these SPF face products, and learn about their experiences. With the user experience, I will investigate text analysis approaches and acquire insights from the analysis that will be performed.

As someone who has always been skeptical of beauty trends, I'm exploring the world of SPF face creams for my final project. These magic' creams may be found all over social media, as well as in popular newspapers and publications. But do they truly keep their promises? As I explore the realm of SPF face creams, my attention shifts to real-life experiences using sentiment analysis and topic modeling of user evaluations. I hope to clarify not just what's in the products, but also how people truly feel about them, offering a nuanced viewpoint for making better skincare decisions.

[1] *Sun protector factor*

Before we begin with the project's analysis, we must first understand what corpus linguistics is. Corpus linguistics is a rapidly expanding field of linguistics that studies massive text collections (corpora) to detect language patterns and use trends. Corpus linguistics investigates distributional differences and interactions between linguistic characteristics, which gives information on language structure and variation (Gries, S.T., 2009).

The paper from (Osterwalder & Herzog, 2009) highlights how sunscreen effectiveness, often misunderstood due to factors like SPF numbers, needs a closer look. It suggests moving away from just focusing on high SPF values and instead considering the overall quality of UV protection (Osterwalder & Herzog, 2009).

To what extent do the advertised benefits of popular SPF face creams align with their actual formulations and real-world user experiences, and how can this information guide consumers in making informed choices about SPFs?

In order to answer this question, different natural language processing techniques were used, such as sentiment analysis and topic modelling. Additionally, text mining techniques were used such as web scraping from webpages and also retrieving data directly from the Reddit API.

The research question that this project aims to answer is: ***"Will the sentiment analysis and topic modelling of user reviews for the best SPF face creams in 2024 expose a divergence between marketing assurances and genuine user experiences, suggesting that the proclaimed effectiveness may not consistently match the diverse realities encountered by users?"***

2. Sample, Data, Corpus

2.1. Web scraping from the Independent online newspaper

The data was web scraped from the online British newspaper, called The Independent, for the ten best SPF face creams for 2024. Web scraping the data was made using Python programming language, with the specific libraries for web scraping such as BeautifulSoup.

The ten best face sunscreen creams based on Independent for the year 2024 include La Roche-Posay anthelios UVmune, Garnier Ambre Solair ultra-light sensitive SPF50+, Boots SPF+ niacinamide moisturising lotion, Medik8 advanced day ultimate protect, Supergoop mattescreen SPF30, Hello Sunday the take out one invisible sun stick, Chanel UV essential, Glossier invisible shield, Omorovicza mineral UV shield SPF30, and Bioré UV aqua water essence sunscreen.

The final dataset from this website contains ten SPFs products names in a column, the newspaper's review, what it is best known for, the price, and the new created column for ingredients, which was taken from the SPF products' official websites[2].

2.2. Retrieving user's reviews from the Reddit API

Then, based on this dataset and the newspaper, I will be retrieving data directly from the Reddit API. For this first, I had to create a developer profile and create a project under which I was given the project's name and the secret key that would be needed in order to retrieve the data from the Reddit API.

From the Reddit API, I have chosen two specific subreddits that are more about beauty trends. These subreddits are SkincareAddiction and AsianBeauty, and then I will be searching directly from the query for the specific name of the product itself. Many people, when they have doubts or are not sure which kind of SPF face products to try, choose social media to get reviews from, and from these two subreddits, I will be retrieving the name of the subreddit, the post title, the score, the specific ID that is unique to each post, the author, the date when it was created, comments, comment's scores, and so on.

The data obtained from the Reddit API consists of the top thousand posts that are most relevant to the query being requested. The data is not for a certain period of time, but rather from all eras. There are 10 datasets, each with the same number of variables; however, the number of observations changes depending on the posts and comments made for these specific products. On this project, I will concentrate on analyzing the comments' content, which I will first pre-process before performing sentiment analysis, clustering, or topic modeling. The pre-processing steps here include turning the comment text lowercase, eliminating stopwords, and removing symbols from the text, as many individuals use symbols or emojis to communicate their opinions in social media data. I will next tokenize the text and finally lemmatize it. Lemmatizing is the process by which the bag of words 'run, runs, and running' is reduced to a single word 'run' with the same meaning or context in a particular phrase.

3. Descriptive Statistics

3.1. Overview of the datasets

Following the pre-processing of the text, which I will use to extract information from, further tools were used to determine the size of the corpus in each dataset and to identify the top ten most frequently used terms in the corpus using TF. From the provided appendix containing

[2] *All of the websites where the ingredients were taken are given as references, at the references section*

code, it is evident that the prevailing terms across all datasets include "sunscreen," "skin," "SPF," "product," "oil," "acid," "face," and "good." This observation offers insight into the predominant topics discussed within the datasets sourced from the Reddit API, suggesting a focus on SPF face products. This preliminary analysis lays the groundwork for further investigation into the overall discourse surrounding SPF face products within the Reddit beauty forums, specifically 'SkincareAddiction' and 'AsianBeauty'.

3.2. Insights into Textual Structure and Sentiment Analysis

3.2.1. Text Pre-processing and Sentiment Analysis

Text pre-processing is like preparing ingredients before cooking a dish. In this case, we're cleaning up the text data by converting it to lowercase, removing unnecessary punctuation marks and common words (stopwords), and reducing words to their root form (lemmatization). Lemmatization is the process of determining a word's basic or dictionary form, known as the lemma, after eliminating inflections and variations. To acquire a word's normalized form, it is necessary to identify and modify its suffixes (Plisson, 2004).

Sentiment analysis is where things get interesting. We're trying to understand the emotions conveyed in the text. The VADER sentiment analyzer helps us assign a sentiment score to each comment, indicating whether it's positive, negative, or neutral. Sorting the dataset based on these scores allows us to prioritize comments based on their emotional intensity. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a rule-based sentiment analysis approach developed exclusively for social media material. It is proven scientifically that it exceeds traditional benchmarks and it even individual human critics in sentiment classification with excellent accuracy (F1 Classification Accuracy is 96 percent). VADER combines a gold-standard sentiment vocabulary with five generic principles that represent grammatical and syntactical standards for expressing feeling. Its performance across several areas, including as social media, NY Times editorials, and product evaluations, proves its efficacy and adaptability (Hutto & Gilbert, 2014).

3.2.2. TF-IDF Vectorization and KMeans Clustering

TF-IDF[3] vectorization is a powerful technique used to represent text data numerically. It converts words into numerical vectors, considering both their frequency in a document and their rarity across all documents in the dataset (Jayaswal, 2020). KMeans clustering is like grouping similar items together on a shelf. In this case, we're clustering comments based on their TF-IDF representations to identify common themes or topics. Each cluster represents a distinct group of comments that share similar characteristics (Kharkar, 2023).

[3] *Term Frequency-Inverse Document Frequency*

3.2.3. Word Cloud Generation for Each Cluster

Word clouds offer a visually appealing way to explore the most prominent words within each cluster of comments. They help us identify recurring themes or keywords that characterize each cluster's content. By generating word clouds for each cluster, we can quickly grasp the main topics or sentiments expressed within different segments of the datasets.

4. Main Analysis and Methodology

4.1. Analytical Approaches and Insight

In the first dataset, which I received from the Independent newspaper, I want to investigate how each brand's product price differs. In addition, I would try to arrange the ingredients into two separate clusters and plot them in a two-dimensional space with two principal components, to see whether there are any products that share the same ingredients and how similar or dissimilar they are. Furthermore, I will pre-process the ingredient variable into a bigram because most of the components have two words, and I will be looking at the ten most utilized ingredients in these items to let the buyer determine whether specific products include the elements that are important for an SPF.

The following data collected were gathered from Reddit, primarily from two subreddit groups: 'SkincareAddiction' and 'AsianBeauty'. I explored these two subreddits for each of the ten SPFs, and then from the dataframe that was generated, I had to do various text analysis and natural language processing algorithms to obtain information on these datasets. In these datasets, sentiment analysis was carried out utilizing the VADER sentiment.

The datasets are not labelled; therefore, the data needs to be analyzed using unsupervised approaches. The reason why the VADER sentiment analysis was done, was to show the overall sentiment of the people who commented for those specific products.

After this, cluster analysis was performed, where I wanted to see which words occur more often together in a sentence. Additionally, topic modelling was applied to the 'Comment' variable in all the datasets while using the LDA model, with a selection of five topics and a bag of words of ten. LDA[4] is a widely used method in topic modelling, aiding in data mining and latent data discovery across various fields. It generates topics as lists of words occurring together statistically, helping to understand large collections of text documents. While topic models don't grasp word meanings, they assign words to topics iteratively, aiming for the most probable distribution (Jelodar, et al., 2018).

[4] *Latent Dirichlet Allocation*

Finally, after the topic modelling was done on each one of the datasets, another analysis was taken to see that, based on some negative experiences that people might have from SPF face products, which may be such as the SPF might leave the skin oily or greasy or make it even dry, the products might leave a white cast, which may not look good on the face, it can cause breakouts, and it can lead to acne. Based on these negative words, I checked the occurrence of each of the words, and based on how big each corpus of the datasets is, the percentage of each bad experience compared to the corpus size of each dataset is illustrated through barplots per dataset. These visualizations were done with the intention of showing the best ten SPF products based on the Independence newspaper and whether they are really causing any bad experiences for the users.

4.2. Analysing Price Disparities and Ingredient Similarities among Top SPFs

Based on figure 1 (please refer to the appendix), we can see that the price of the best spf ranges from less than 20 British pounds to 80 British pounds. Another point in this project has been to look more into the ingredients of each of the best spf face creams and do different products have the same ingredients or not. In this analysis, the clusters were set to be two[5], in we plot them in a two-dimensional space, with principal components. First principal component shows the most of the variance of the data, and from this plot we can see that some of the products do share the same ingredients, for instance the brands La Roche-Posay and Garnier are very close to each other, which indirectly means that these two products have almost the same ingredients, but if we look back at the price, Garnier is cheaper than La Roche-Posay, so for a future purchase it would be better to choose Garnier, since the price is cheaper. Also, when comparing Bioré and Medik8, we can see that they almost have the same ingredients, but when comparing prices between these two, Bioré is much cheaper than Medik8.

5. Results

5.1. User perceptions through sentiment analysis and topic modelling

Figures four to twenty-three in the appendix provide visual representations of the sentiment scores and topic modelling clusters for each sunscreen brand. These figures offer further insight into the distribution of sentiment and the key topics discussed within user reviews. They complement the textual findings by presenting the data in a graphical format, allowing for a deeper understanding of the sentiments and themes associated with each brand.

From the analysis, it is evident that sentiment scores across various sunscreen brands tend to be neutral to positive. La Roche-Posay SPF receives positive feedback, particularly for its

[5] *The clusters were set to be two using the elbow criterion for selecting how many clusters there should be (even though some of them are not well defined)*

affordability and effectiveness. However, some users mention a white cast as a drawback. Garnier SPF also garners positive reviews, but is cautioned for its potential strength and tendency to leave a white cast. Boots' product is generally well-received, with users praising its ingredients and effectiveness. The Medik8 product elicits mixed sentiments, with mentions of its effectiveness against hormonal acne but also reports of skin stripping. Supergoop spf receives mostly positive feedback, though some users note a greasy finish. The Hello Sunday brand is positively anticipated, with users seeking advice and expressing excitement to try it. Chanel's product generally receives positive comments but is associated with acne for some users due to its oil content. Glossier product receives neutral to positive sentiment, with cautionary notes about its potential to cause breakouts for acne-prone skin. Overall, sentiment tends to be positive for most brands, with users expressing gratitude for effective products like the Omorovicza product.

5.2. Negative experiences in the SPF creams

The analysis from Figure 24 to 31 on the appendix focuses on identifying the prevalence of negative experiences associated with the top ten SPF face products, as reported by the Independent. This examination aims to ascertain which products are more likely to induce adverse effects among users. Upon scrutinizing the corpus for the term 'oily,' it becomes apparent that Supergoop exhibits the highest percentage, followed by Medik8, whereas Hello Sunday and Omorovicza demonstrate the lowest occurrence. Furthermore, concerning acne-related issues, Chanel and Garnier emerge as the most possible ones, with Supergoop being the least implicated. In terms of breakouts, Garnier is identified as highly likely, given its notable association with acne, while Hello Sunday, La Roche-Posay, Boots, and Biore are less prone to causing breakouts. Notably, Medik8 is notably linked with dry skin, which may stem from its oily composition, suggesting its affinity with individuals experiencing dry skin conditions. Conversely, individuals with oily skin are advised against using Medik8. Regarding SPF face creams leaving a white cast, La Roche-Posay, Garnier, and Supergoop are implicated. Additionally, Supergoop is notably associated with greasiness, while Glossier and Biore are linked with allergies, and Medik8 is associated with causing rashes. This analysis offers valuable insights into the potential adverse effects of SPF face products, enabling users to make informed decisions based on their specific skin concerns and preferences.

6. Conclusion

The primary aim of this project was to investigate whether the sentiment analysis and topic modelling of user reviews for the top SPFs in 2024 would reveal discrepancies between marketing claims and authentic user experiences, thereby indicating potential inconsistencies in the effectiveness of these products. Through comprehensive text analysis techniques and natural language processing methodologies, this study delved into user-generated content from

both online forums and reputable sources, aiming to provide insights into the experiences surrounding SPFs.

The findings of this analysis suggest that there exists a notable variance between the marketing assurances propagated by SPFs brands and the actual experiences reported by users. While marketing campaigns often emphasize the efficacy, affordability, and overall benefits of these products, user reviews reveal a more nuanced reality characterized by diverse sentiments and experiences.

The sentiment analysis of user reviews indicated a spectrum of emotions ranging from positive to neutral and negative. While some users expressed satisfaction with the affordability, effectiveness, and overall performance of certain SPFs, others raised concerns regarding common issues such as white cast texture, greasiness, and propensity to cause breakouts or acne. Through topic modelling, distinct themes and topics emerged within user reviews, shedding further light on the specific attributes and characteristics of each product.

By considering real-world user feedback, consumers can make more informed decisions regarding SPFs, taking into account factors such as individual skin type, preferences, and specific concerns. In conclusion, these findings highlight the importance of prioritizing authentic user feedback when making informed decisions about skincare products, ultimately empowering consumers to navigate the complex landscape of SPFs with confidence.

7. Limitations and Outlook

7.1. Limitations

Some of the limitations that this project has faced could be:

- The datasets from the Reddit API are not labelled, which makes it challenging to analyse the data accurately.

- Web scraping data from official websites could provide additional valuable information, such as ratings, but restrictions on web scraping prevent access to this data.

- The lack of context in Reddit comments may pose challenges in accurately interpreting user sentiments, as comments may be brief or lack detailed information about users' experiences with specific SPF products.

- The use of VADER sentiment analysis alone may oversimplify the complexity of user sentiments, as it may not capture nuances or sarcasm effectively.

7.2. Outlook

Some of the outlooks for this project are:

- To enhance sentiment analysis, exploring additional techniques beyond VADER, such as HuggingFace or Roberta, could offer more insights and comparative analysis.

- Incorporating user engagement metrics from social media platforms beyond Reddit, such as Instagram, X, or YouTube, could provide a more comprehensive understanding of user experiences with SPF products.

8. References

Biore. (n.d.). *bioré uv aqua essence sunscreen*. Retrieved April 15, 2024, from Biore: https://www.biore.com/en-gb/products/sunscreen-spf-50/

Boots. (n.d.). *Boots SPF+ Niacinamide Moisturising Lotion*. Retrieved April 15, 2024, from Boots: https://www.boots.com/boots-spf-niacinamide-moisturising-lotion-spf50-50ml-10331095

Chanel. (n.d.). *Chanel UV Essential*. Retrieved April 15, 2024, from Chanel: https://www.chanel.com/de/hautpflege/p/141897/uv-essentiel-umfassender-schutz-uv-umweltschadstoffe-antioxidativ-spf50/?param=value&gad_source=1&gclid=CjwKCAjwwr6wBhBcEiwAfMEQs17QJPwFV7v4wzdNYriPaT5zEYyxuL6Flvv0tQJJJ2Su8W_yjMYLaBoCDBsQAvD_BwE

Garnier. (n.d.). *Garnier Ambre*. Retrieved April 15, 2024, from Garnier DE: https://www.garnier.de/sonnenschutz/sonnenschutz-marken/ambre-solaire/super-uv/gesicht-uv-schutz-fluid

Glossier. (n.d.). *Glossier Invisible Shield*. Retrieved April 15, 2024, from Glossier: https://www.glossier.com/en-de/products/invisible-shield-spf30

Gries, S.T. (2009), What is Corpus Linguistics?. Language and Linguistics Compass, 3: 1225-1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

Independent. (2024, February 12). *16 best sunscreens for your face 2024: Daily SPF protection, from sensitive to non-greasy formulas*. (L. Cunningham, Editor) Retrieved April 15, 2024, from Independent: https://www.independent.co.uk/extras/indybest/fashion-beauty/skincare/best-sunscreen-face-mineral-facial-spf-tint-non-greasy-sensitive-skin-a9569096.html

Jayaswal, V. (2020, October 4). *Text Vectorization: Term Frequency — Inverse Document Frequency (TFIDF)*. Medium. https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d

Jelodar, H., Wang, Y., Juan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018, November). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 15169–15211. doi: https://doi.org/10.1007/s11042-018-6894-4

Kharkar, D. (2023, June 11). K-Means Clustering Algorithm. *Medium*. https://medium.com/@dishantkharkar9/k-means-clustering-algorithm-ce4fbcac8fb0

Medik8. (n.d.). *Medik8 Advanced Day Ultimate*. Retrieved April 15, 2024, from Medik8: https://eu.medik8.com/products/advanced-day-ultimate-protect

Omorovicza. (n.d.). *Omorovicza Mineral UV Shield SPF30*. Retrieved April 15, 2024, from Omorovicza: https://www.feelunique.com/p/Omorovicza-Mineral-UV-Shield-SPF-30-100ml#:~:text=Aqua%20(Hungarian%20Thermal%20Water)%2C,Cetearyl%20Alcohol%2C%20Saccharomyces%20(Hungarian%20Thermal

Plisson, J. L. (2004). A rule based Approach to Word Lemmatization. Retrieved April 19, 2024, from https://www.semanticscholar.org/paper/A-Rule-based-Approach-to-Word-Lemmatization-Plisson-Lavra%C4%8D/5319539616e81b02637b1bf90fb667ca2066cf14

Osterwalder, U., & Herzog, B. (2009). Sun protection factors: world wide confusion. *British Journal of Dermatology*, *161*(s3), 13-24. doi:https://doi.org/10.1111/j.1365-2133.2009.09506.x

Roche-Posay, L. (n.d.). *La Roche-Posay Anthelios*. Retrieved April 15, 2024, from La Roche-Posay DE: https://www.larocheposay.de/anthelios/anthelios-uvmune-400-oil-control-fluid-lsf50plus?gad_source=1&gclid=Cj0KCQjw8pKxBhD_ARIsAPrG45mNSarMeiOrUxsUVNgkDbbf_M1A3tFHHFFbsXjaSSDJv21wQOAW_98aAv01EALw_wcB&gclsrc=aw.ds

Sunday, H. (n.d.). *Hello Sunday Invisible Sun Stick SPF*. Retrieved April 15, 2024, from Hello Sunday: https://www.hellosundayspf.com/products/invisible-sun-stick-spf-50

Supergoop. (n.d.). *Supergoop mineral mattescreen spf*. Retrieved April 15, 2024, from Supergoop: https://supergoop.com/products/smooth-and-poreless-mattescreen?variant=39816821407842

9. Code file and the datasets

The Jupyter code file and also the datasets can be found in this link[6] as a zip file, because the maximun size of that Github allows is 100 MB.

[6] *https://github.com/leonahasanii/TANLP_finalproject*