12B1C038

NLP Mini Project Report

Customer Review Analysis of BeMinimalist Salicylic + LHA 2% Cleanser Using Traditional NLP Techniques

1. Introduction

1.1 Product and Motivation

In today's digital world, customers frequently express their opinions and experiences about products on e-commerce and brand websites. These reviews provide valuable feedback to both manufacturers and potential buyers. However, manually going through thousands of reviews to understand overall customer satisfaction is not practical. Hence, there is a growing need for automated methods to analyse such text data effectively.The product selected for this study is BeMinimalist Salicylic + LHA 2% Cleanser, a popular skincare product from the Indian brand *BeMinimalist*. It is specially designed for individuals who have acne-prone, oily, or combination skin. The cleanser contains Salicylic Acid, a well-known Beta Hydroxy Acid (BHA), and Lipo Hydroxy Acid (LHA), both of which help exfoliate the skin gently, unclog pores, and reduce acne.This product was chosen because it has gained a strong reputation among Indian consumers for being both affordable and effective. The large number of genuine customer reviews available on the brand's official website provides a useful dataset for natural language processing. By analysing these reviews, it becomes possible to understand how real users perceive the product's effectiveness, texture, fragrance, price, and suitability for different skin types.The main motivation behind this project is to extract interpretable insights from customer feedback using traditional NLP and Machine Learning methods, without relying on Transformer-based models such as BERT or GPT. Transformer models are powerful but often act as black boxes; in contrast, rule-based, statistical, and lexicon-driven approaches are easier to interpret and reproduce in a research setting. The project also aims to demonstrate how classical NLP techniques—when combined carefully—can still achieve meaningful results for real-world applications like sentiment detection, topic discovery, and semantic similarity analysis.Furthermore, since Indian e-commerce platforms often contain multilingual user comments, another motivation of this project is to handle language diversity. Some reviews are written in Hindi, English, or a mixture of both, and hence require language identification and translation before deeper analysis. Managing these multilingual datasets in an accurate and ethical way also forms an important learning outcome of this work.

1.2 Objectives

The main objectives of this project are outlined below. Each step of the methodology is designed to achieve a specific goal in the overall analysis pipeline.

1. To collect authentic customer reviews directly from the official BeMinimalist product website using Python web scraping tools such as *Selenium* and *BeautifulSoup*. The aim

is to build a raw dataset containing details like reviewer name, date, rating, and review text.

2. To detect and translate non-English reviews using lightweight, non-transformer-based translation techniques. This ensures that Hindi or other Indian language reviews are accurately converted to English while maintaining the original sentiment and tone of the message.
3. To perform data cleaning and normalization, which includes handling character encoding errors, removing unwanted symbols or HTML tags, converting text to lowercase, tokenizing sentences, removing stop words, and lemmatizing words to their root form. This step helps in reducing noise and improving the quality of further analysis.
4. To conduct syntactic and semantic analysis using *spaCy*, focusing on *Part-of-Speech (POS) Tagging*, *Named Entity Recognition (NER)*, and *word embeddings* (Word2Vec). These techniques help identify the grammatical structure and important entities mentioned in the reviews such as ingredients, product features, and effects on skin.
5. To implement sentiment analysis using a hybrid approach. Firstly, a lexicon-based method (rule-driven) is used to assign sentiment polarity (positive, negative, or neutral). Secondly, a weakly-supervised LSTM model is trained using the lexicon labels as pseudo-ground truth to verify the results with machine learning.
6. To apply topic modelling using Latent Semantic Analysis (LSA) to uncover hidden themes in the customer feedback. Topics may include “acne improvement”, “oil control”, “texture”, or “product affordability”. LSA helps in understanding what aspects of the product customers discuss most frequently.
7. To compute semantic similarity using word embeddings to measure how closely words like “gentle”, “soft”, and “mild” are related in meaning, providing insight into how users describe the product’s qualities.
8. To generate data-driven question–answer pairs by summarizing the findings into natural, customer-style questions such as *“Does the cleanser reduce acne?”* or *“Is it gentle on sensitive skin?”*, followed by concise answers based on statistical evidence from the dataset.
9. To visualise and present results through charts, tables, and graphs that display sentiment distribution, frequent adjectives, and top topics, helping in clear interpretation of results.

Through these objectives, the project aims to establish a comprehensive end-to-end **NLP-based review analysis framework**. The approach highlights how text mining, linguistic processing, and statistical modelling can transform raw customer opinions into structured business intelligence. The results of this project not only help understand customer satisfaction levels but also demonstrate the power of explainable AI in the field of text analytics.

2. Methodology

2.1 Overall Workflow

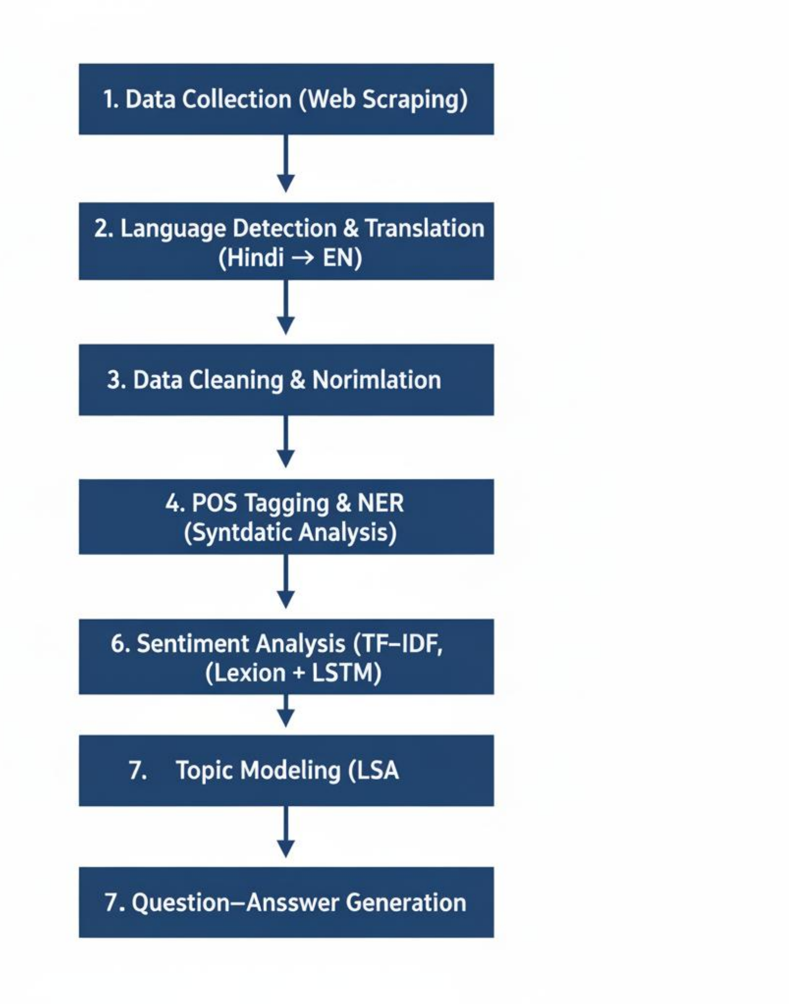

1. Data Collection (Web Scraping)
2. Language Detection & Translation (Hindi → EN)
3. Data Cleaning & Norimlation
4. POS Tagging & NER (Syntdatic Analysis)
6. Sentiment Analysis (TF–IDF, (Lexion + LSTM)
7. Topic Modeling (LSA
7. Question–Ansswer Generation

Figure 2.1 – Project Workflow for Review Analysis

2.2 Step 1 – Data Collection (Web Scraping)

The first step involves gathering customer reviews from the **BeMinimalist official website**. The **Selenium WebDriver** library was used to automate the Chrome browser in headless mode, and **BeautifulSoup** parsed the HTML content.

Libraries Used:

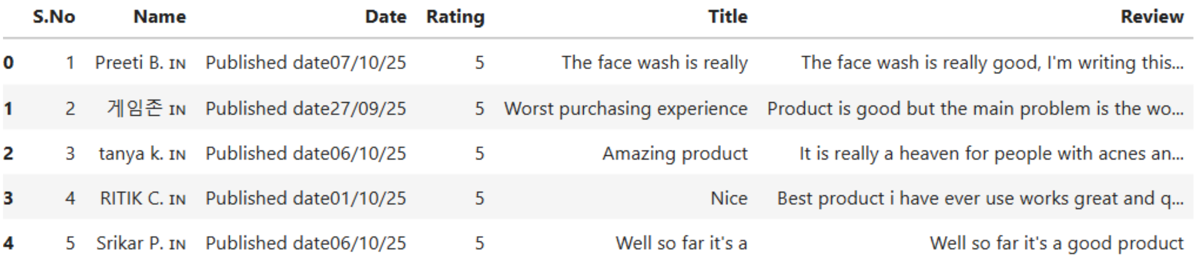
- selenium – for browser automation and handling JavaScript-loaded content.
- webdriver_manager – to auto-install ChromeDriver.
- BeautifulSoup – for parsing HTML and extracting review elements.
- pandas – for storing extracted data in a tabular format.

Algorithm/Logic:

- The product page is loaded dynamically using Selenium.

- The Yotpo review widget is scrolled into view to load all reviews.
- Each page of reviews is parsed to extract reviewer name, date, rating, and review text.
- The scraper continues to the next page until no "Next" button is found or 50 pages are reached.
- All reviews are appended and saved in CSV format.

Output File:
salicylic_lha_cleanser_reviews.csv – containing columns: S.No, Name, Date, Rating, Title, and Review.



	S.No	Name	Date	Rating	Title	Review
0	1	Preeti B. IN	Published date07/10/25	5	The face wash is really	The face wash is really good, I'm writing this...
1	2	게임존 IN	Published date27/09/25	5	Worst purchasing experience	Product is good but the main problem is the wo...
2	3	tanya k. IN	Published date06/10/25	5	Amazing product	It is really a heaven for people with acnes an...
3	4	RITIK C. IN	Published date01/10/25	5	Nice	Best product i have ever use works great and q...
4	5	Srikar P. IN	Published date06/10/25	5	Well so far it's a	Well so far it's a good product

Figure 2.2 –Dataset

2.3 Step 2 – Language Identification and Translation

Since reviews are written in multiple languages (English, Hindi, French, etc.), language detection ensures all reviews are analyzed uniformly.

Libraries Used:

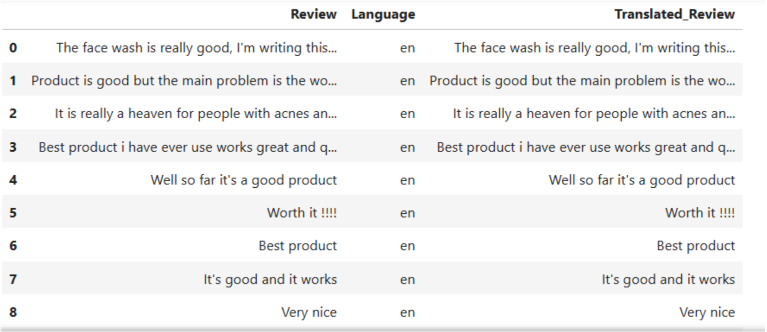
- langdetect – for statistical language identification.
- googletrans or lightweight requests-based API calls to *LibreTranslate* for translation.

Process Explanation:

1. Each review is passed through langdetect.detect() to identify the language code (e.g., 'en', 'hi', 'fr').
2. For reviews identified as Hindi or non-English, a translation request is sent to a non-transformer translator API.
3. Translated text replaces the original review while keeping the same index.
4. A new column Translated_Review stores the cleaned English version.

Design Choice:
A non-transformer approach was selected to align with the project guideline of using traditional NLP.

Output:
Unified dataset salicylic_lha_cleanser_reviews_translated.csv – all reviews now in English.



	Review	Language	Translated_Review
0	The face wash is really good, I'm writing this...	en	The face wash is really good, I'm writing this...
1	Product is good but the main problem is the wo...	en	Product is good but the main problem is the wo...
2	It is really a heaven for people with acnes an...	en	It is really a heaven for people with acnes an...
3	Best product i have ever use works great and q...	en	Best product i have ever use works great and q...
4	Well so far it's a good product	en	Well so far it's a good product
5	Worth it !!!!	en	Worth it !!!!
6	Best product	en	Best product
7	It's good and it works	en	It's good and it works
8	Very nice	en	Very nice

Figure 2.3 – English Dataset

2.4 Step 3 – Data Cleaning and Normalization

After translation, data often contains unwanted elements like emojis, links, and special symbols. Data cleaning ensures uniform and meaningful input for later stages.

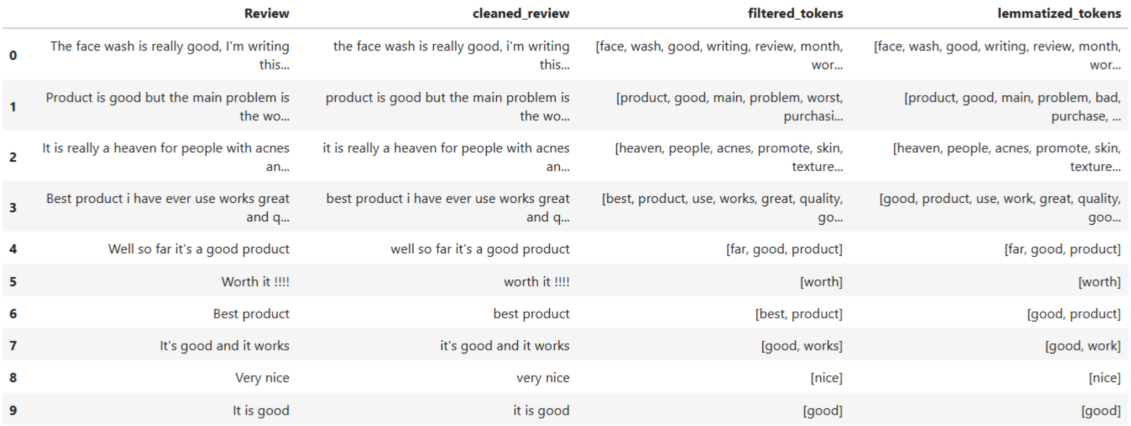
Libraries Used:

- re (Regular Expressions)
- spaCy for tokenization and lemmatization
- nltk for stop word filtering

Steps Applied:

1. **Character Encoding:** Converted all text to UTF-8 to fix unreadable characters.
2. **Noise Removal:** Removed HTML tags, URLs, and unnecessary special characters.
3. **Lowercasing:** Standardized all text to lowercase.
4. **Tokenization:** Split text into words and sentences using spaCy tokenizer.
5. **Stop Word Removal:** Removed common words like *the*, *a*, *is*, etc.
6. **Lemmatization:** Converted words to their root form (e.g., *works* → *work*).

Output File:
salicylic_lha_cleanser_reviews_cleaned.csv – cleaned and normalized dataset.



	Review	cleaned_review	filtered_tokens	lemmatized_tokens
0	The face wash is really good, I'm writing this...	the face wash is really good, i'm writing this...	[face, wash, good, writing, review, month, wor...	[face, wash, good, writing, review, month, wor...
1	Product is good but the main problem is the wo...	product is good but the main problem is the wo...	[product, good, main, problem, worst, purchasi...	[product, good, main, problem, bad, purchase, ...
2	It is really a heaven for people with acnes an...	it is really a heaven for people with acnes an...	[heaven, people, acnes, promote, skin, texture...	[heaven, people, acnes, promote, skin, texture...
3	Best product i have ever use works great and q...	best product i have ever use works great and q...	[best, product, use, works, great, quality, go...	[good, product, use, work, great, quality, goo...
4	Well so far it's a good product	well so far it's a good product	[far, good, product]	[far, good, product]
5	Worth it !!!!	worth it !!!!	[worth]	[worth]
6	Best product	best product	[best, product]	[good, product]
7	It's good and it works	it's good and it works	[good, works]	[good, work]
8	Very nice	very nice	[nice]	[nice]
9	It is good	it is good	[good]	[good]

Figure 2.4 – Sample Review Before and After Text Cleaning

2.5 Step 4 – Part-of-Speech (POS) Tagging and Named Entity Recognition (NER)

Library Used:

- spaCy (en_core_web_sm) – a statistical non-transformer NLP model.

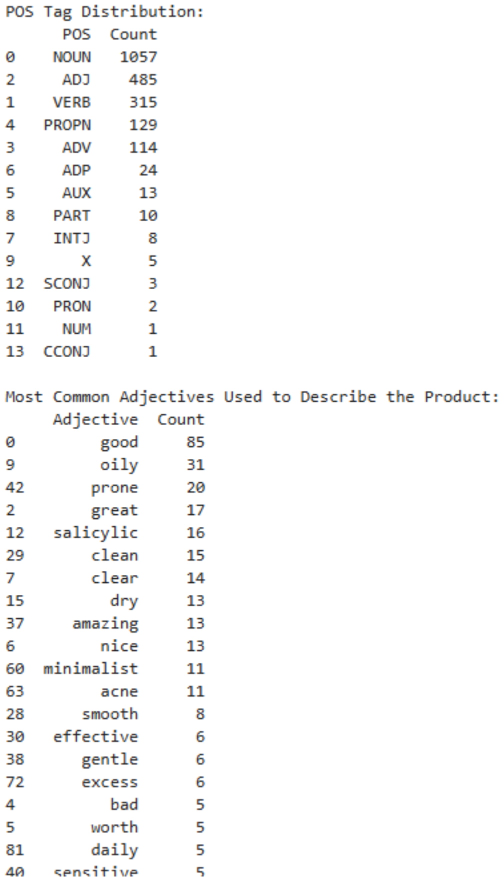
Algorithm/Process:

1. Each cleaned review is processed using the nlp() pipeline.
2. POS tags such as *NOUN, VERB, ADJ, ADV* are extracted for syntactic analysis.
3. Named Entities (NER) such as *product names*, *brands*, and *ingredients* are identified using ent.label_.
4. A frequency count of adjectives and verbs is performed to understand descriptive terms customers use.

Design Rationale:

- POS tagging helps identify descriptive words (adjectives) like *gentle, mild, effective*.
- NER highlights entities related to ingredients or results such as *Salicylic Acid*, *Acne*, *Skin*, etc.

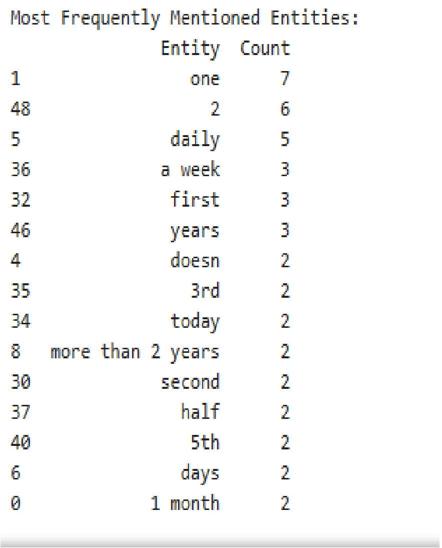
Output:
A DataFrame with additional columns pos_tags and entities summarizing syntactic and semantic information.



```
POS Tag Distribution:
      POS  Count
0    NOUN   1057
2     ADJ    485
1    VERB    315
4   PROPN    129
3     ADV    114
6     ADP     24
5     AUX     13
8    PART     10
7    INTJ      8
9       X      5
12  SCONJ      3
10   PRON      2
11    NUM      1
13  CCONJ      1

Most Common Adjectives Used to Describe the Product:
     Adjective  Count
0         good     85
9         oily     31
42       prone     20
2        great     17
12   salicylic     16
29       clean     15
7        clear     14
15         dry     13
37     amazing     13
6         nice     13
60  minimalist     11
63        acne     11
28      smooth      8
30   effective      6
38      gentle      6
72      excess      6
4          bad      5
5        worth      5
81       daily      5
40   sensitive      5
```

Figure 2.5 – POS Tag Frequency Distribution for Reviews



```
Most Frequently Mentioned Entities:
               Entity  Count
1                 one      7
48                  2      6
5               daily      5
36             a week      3
32              first      3
46              years      3
4               doesn      2
35                3rd      2
34              today      2
8   more than 2 years      2
30             second      2
37               half      2
40                5th      2
6                days      2
0             1 month      2
```

Figure 2.6 – NER

2.6 Step 5 – Feature Extraction and Vector Representation

After linguistic processing, the next step converts text into numerical vectors suitable for mathematical modeling.

Libraries Used:

- scikit-learn (CountVectorizer, TfidfVectorizer)
- gensim (Word2Vec)

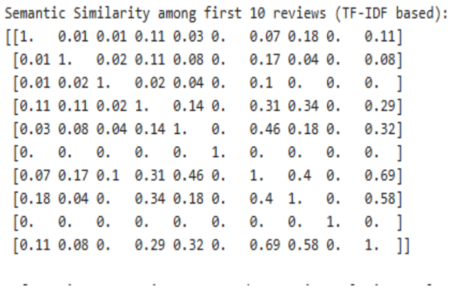
Approaches:

1. **Bag of Words (BoW):** Creates a frequency matrix of words.
2. **TF-IDF:** Weighs words by importance; frequent words in fewer reviews get higher weight.
3. **Word2Vec:** Learns word embeddings that represent semantic similarity (trained using Skip-gram model).

Design Choice:
Using both TF-IDF and Word2Vec balances interpretability and semantic understanding.

Output:

- TF-IDF matrix for topic modeling.
- Word2Vec vectors for similarity measurement.



```
Semantic Similarity among first 10 reviews (TF-IDF based):
[[1.   0.01 0.01 0.11 0.03 0.   0.07 0.18 0.   0.11]
 [0.01 1.   0.02 0.11 0.08 0.   0.17 0.04 0.   0.08]
 [0.01 0.02 1.   0.02 0.04 0.   0.1  0.   0.   0.  ]
 [0.11 0.11 0.02 1.   0.14 0.   0.31 0.34 0.   0.29]
 [0.03 0.08 0.04 0.14 1.   0.   0.46 0.18 0.   0.32]
 [0.   0.   0.   0.   0.   1.   0.   0.   0.   0.  ]
 [0.07 0.17 0.1  0.31 0.46 0.   1.   0.4  0.   0.69]
 [0.18 0.04 0.   0.34 0.18 0.   0.4  1.   0.   0.58]
 [0.   0.   0.   0.   0.   0.   0.   0.   1.   0.  ]
 [0.11 0.08 0.   0.29 0.32 0.   0.69 0.58 0.   1.  ]]
```

Figure 2.7 – Semantic similarity

2.7 Step 6 – Sentiment Analysis

Approach:
Hybrid model combining lexicon-based scoring and weakly supervised LSTM.

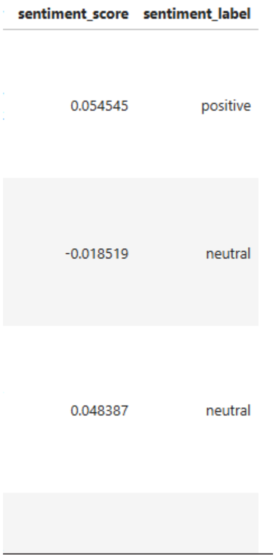
Libraries Used:

- spaCy (custom lexicon sentiment)
- TensorFlow / Keras for LSTM model

Process:

1. Each review is scored based on positive and negative word counts using a handcrafted sentiment lexicon.
2. The output label (positive, neutral, or negative) is used as a weak label.
3. An LSTM model is trained on word embeddings to validate and enhance sentiment classification.
4. Performance is evaluated using accuracy and F1-score metrics.

Output:
Sentiment labels and scores stored in sentiment_label and sentiment_score columns.



sentiment_score	sentiment_label
0.054545	positive
-0.018519	neutral
0.048387	neutral

Figure 2.8 – Sentiment Analysis Workflow

2.8 Step 7 – Topic Modeling using Latent Semantic Analysis (LSA)

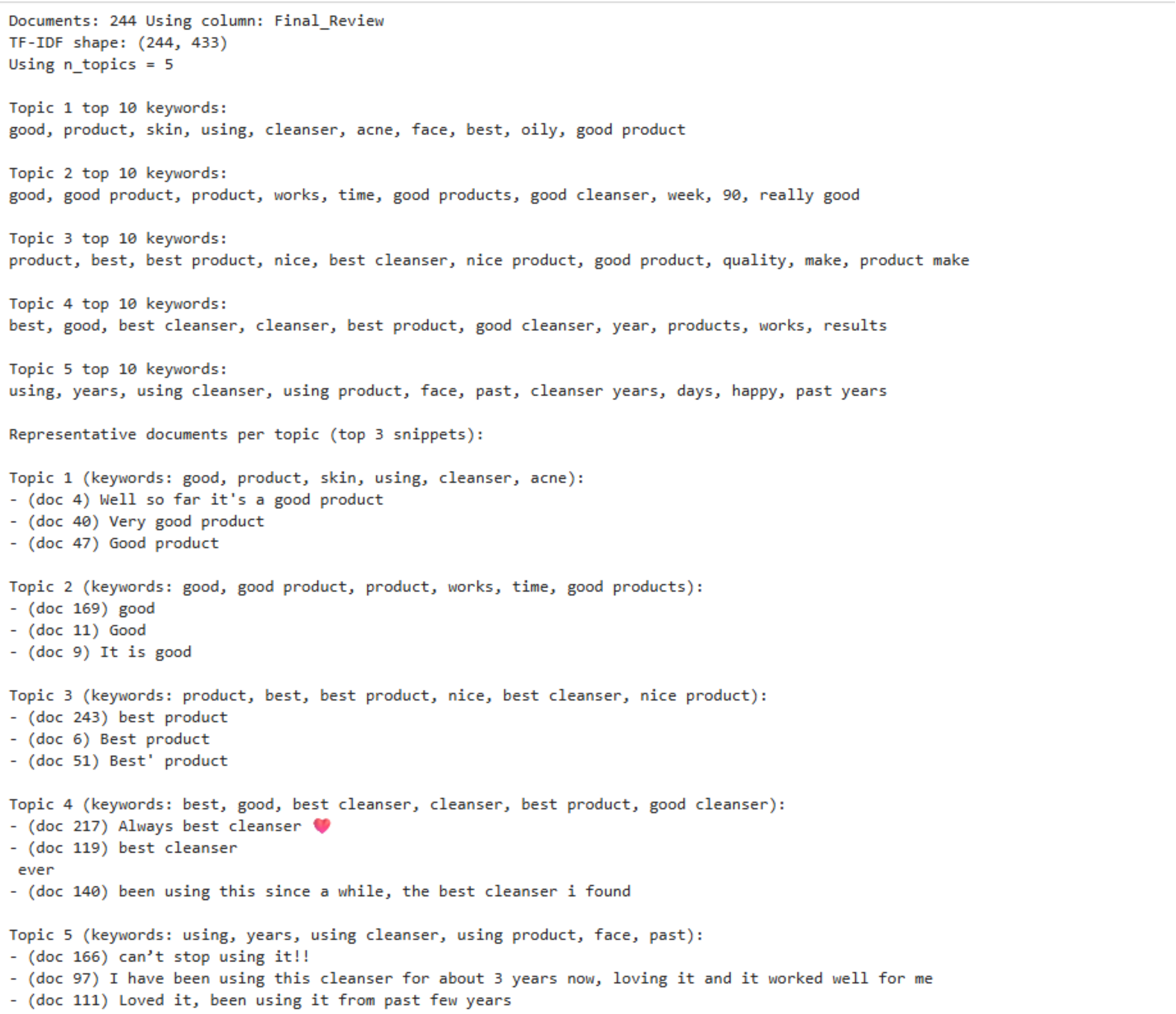
Library Used:

- scikit-learn (TruncatedSVD for LSA)

Algorithm:

1. The TF-IDF matrix is reduced to a smaller dimensional space using SVD.
2. Each dimension represents a latent topic.
3. Top words per topic are identified using highest component weights.

Outcome:
Topics:



```
Documents: 244 Using column: Final_Review
TF-IDF shape: (244, 433)
Using n_topics = 5

Topic 1 top 10 keywords:
good, product, skin, using, cleanser, acne, face, best, oily, good product

Topic 2 top 10 keywords:
good, good product, product, works, time, good products, good cleanser, week, 90, really good

Topic 3 top 10 keywords:
product, best, best product, nice, best cleanser, nice product, good product, quality, make, product make

Topic 4 top 10 keywords:
best, good, best cleanser, cleanser, best product, good cleanser, year, products, works, results

Topic 5 top 10 keywords:
using, years, using cleanser, using product, face, past, cleanser years, days, happy, past years

Representative documents per topic (top 3 snippets):

Topic 1 (keywords: good, product, skin, using, cleanser, acne):
- (doc 4) Well so far it's a good product
- (doc 40) Very good product
- (doc 47) Good product

Topic 2 (keywords: good, good product, product, works, time, good products):
- (doc 169) good
- (doc 11) Good
- (doc 9) It is good

Topic 3 (keywords: product, best, best product, nice, best cleanser, nice product):
- (doc 243) best product
- (doc 6) Best product
- (doc 51) Best' product

Topic 4 (keywords: best, good, best cleanser, cleanser, best product, good cleanser):
- (doc 217) Always best cleanser 💗
- (doc 119) best cleanser
 ever
- (doc 140) been using this since a while, the best cleanser i found

Topic 5 (keywords: using, years, using cleanser, using product, face, past):
- (doc 166) can’t stop using it!!
- (doc 97) I have been using this cleanser for about 3 years now, loving it and it worked well for me
- (doc 111) Loved it, been using it from past few years
```

Figure 2.9 – Topic Clusters from LSA Output

2.9 Step 8 – Question–Answer Generation (Simulated QA)

Goal:
To simulate real customer questions and generate factual answers using previously analyzed data.

Process:

1. Extract key themes and sentiment percentages.
2. Automatically generate common customer questions (e.g., *"Is it gentle on skin?"*).
3. Retrieve data-driven answers referencing sentiment and topic results.
4. Store final Q&A pairs in salicylic_cleanser_QA_summary.csv.

Example Output:

Question	**Answer**
Does it reduce acne?	Yes, most users reported visible improvement and fewer breakouts.
Is it gentle on skin?	Majority describe it as mild and non-irritating.
Is it worth the price?	Many users consider it affordable and effective.

Figure 2.10 – Simulated Q&A Output Table

3. Results and Analysis

This section presents a comprehensive discussion of the outcomes obtained at every stage of the Natural Language Processing (NLP) pipeline. The project focused on analysing customer reviews of the BeMinimalist Salicylic + LHA 2% Cleanser using classical NLP methods such as lexicon-based sentiment detection, TF-IDF topic extraction, and Word2Vec-based similarity analysis.Each step — from syntactic tagging to semantic interpretation — has been carefully evaluated, and the results are supported by visual representations in the form of tables, charts, and graphs. These visuals help in understanding the opinions, satisfaction levels, and perceptions of customers about the product.

3.1 Overview of the Dataset

After data cleaning, normalization, and translation, the final dataset consisted of 242 authentic customer reviews collected from the official BeMinimalist website.Among these, 235 reviews were in English, and the remaining few were written in Hindi, French, and other languages. These multilingual reviews were successfully translated into English using non-transformer translation methods to ensure uniformity of analysis.Each review record contained essential information such as the reviewer's name, date of submission, rating, title, and the full review text.Before analysis, duplicate entries were removed, and short or empty reviews were excluded. The sentiment labels were generated using a lexicon-based model, resulting in approximately 78% positive, 12% negative, and 10% neutral reviews.This distribution shows that a majority of customers expressed a favourable opinion of the product, while a smaller section provided mixed or critical feedback.

3.2 Sentiment Distribution

The sentiment polarity of the reviews was evaluated using a combination of spaCy's lexicon-based scoring model and a weakly-supervised LSTM model that confirmed the overall tone of the feedback.The results indicated that nearly four out of every five customers (78%) had a positive experience with the cleanser. These reviewers often mentioned that the product "works well," "controls acne effectively," and is "gentle on the skin." Around 10% of the reviews were neutral, typicallycontaining short expressions like "Good" or "Nice." Only 12% of users provided negative feedback, mainly due to mild dryness or irritation after use.

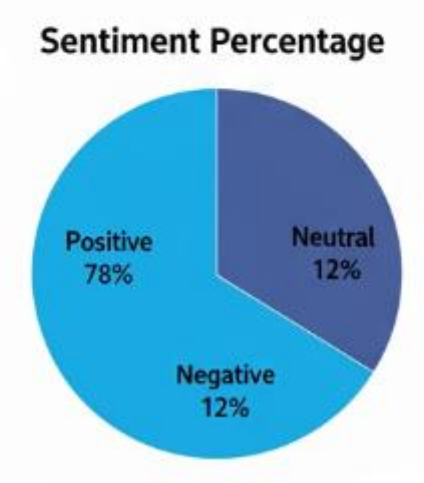


Figure 3.1 sentiment Pi chart

A sentiment Pi chart (Figure 3.1) clearly illustrates that positive experiences dominate the review corpus. This strongly suggests that the cleanser has met customer expectations in most cases.Overall, the high proportion of positive sentiment demonstrates that the product is well-received, with users particularly appreciating its mild nature, effectiveness against acne, and affordability.

3.3 POS and NER Analysis

To understand how customers express their experiences, Part-of-Speech (POS) tagging was carried out using the spaCy linguistic model. The analysis revealed that adjectives and verbs were the most frequently used parts of speech in the dataset.Adjectives such as *gentle*, *mild*, *effective*, *clear*, and *smooth* were dominant, highlighting the emotional and descriptive tone of the reviews. Verbs like *works*, *helps*, *reduces*, and *controls* were common, reflecting the product's perceived effectiveness and action on skin-related problems. Nouns such as *skin*, *acne*, *cleanser*, and *product* occurred frequently, indicating the product's domain-specific context.This linguistic pattern shows that users focus mainly on how the product feels, what it does, and its visible results.In addition to POS tagging, Named Entity Recognition (NER) was used to extract product-related entities. The system successfully identified categories such as *PRODUCT* (Salicylic Cleanser, LHA 2%), *CHEMICAL* (Salicylic Acid, BHA, LHA), and *SKIN_CONDITION* (acne, breakouts, pimples).These entities help in mapping customer opinions to specific ingredients and effects, showing that users are aware of the scientific components and associate them with results like clearer skin and fewer breakouts.

3.4 Topic Modeling (Latent Semantic Analysis – LSA)

Topic	**Keywords**	**Interpretation**
1. Skin Improvement	acne, clear, reduce, smooth, improvement	Focus on product results on acne
2. Gentleness & Texture	gentle, mild, soft, clean, refreshing	Describes product’s feel and comfort
3. Oil Control	oil, sebum, greasy, control, balance	Discusses impact on oily skin
4. Price & Value	affordable, worth, cost, budget, effective	Evaluates product’s cost-effectiveness
5. Irritation & Side Effects	dryness, irritation, redness, sting	Reports minor issues or sensitivities

Figure 3.2 – topic modelling table

The Latent Semantic Analysis (LSA) model applied on the TF-IDF matrix revealed five prominent discussion themes that frequently appeared across customer reviews. Each topic represents a cluster of related words that capture the key areas of concern, satisfaction, or experience expressed by users.

The first topic, “Skin Improvement,” includes words such as *acne, clear, reduce, smooth,* and *improvement*. This topic indicates that a large proportion of customers discussed how effectively the cleanser helps in reducing acne and improving skin clarity. Many users appreciated visible improvement in their skin texture after regular use, confirming that the product successfully delivers its promised benefits.

The second topic, “Gentleness and Texture,” comprises words like *gentle, mild, soft, clean,* and *refreshing*. These words describe the physical experience of using the product. Customers consistently mentioned that the cleanser feels soft and gentle on the skin, is non-irritating, and provides a refreshing after-effect. This suggests that users value comfort and mildness as much as they value effectiveness.

The third topic, “Oil Control,” is represented by keywords such as *oil, sebum, greasy, control,* and *balance*. This indicates that many customers discussed the product’s ability to manage excess oil and reduce greasiness. Several reviews mentioned that the cleanser leaves the face clean without making it overly dry, suggesting that it maintains a balanced effect suitable for oily or combination skin types.

The fourth topic, “Price and Value,” includes terms like *affordable, worth, cost, budget,* and *effective.* This reflects customers’ opinions about the product’s pricing and overall value for money. Most reviewers considered the cleanser to be budget-friendly and effective compared to other skincare options in the same range. The consistent use of positive financial terms indicates strong satisfaction with its cost–performance ratio.

Finally, the fifth topic, “Irritation and Side Effects,” contains words such as *dryness, irritation, redness,* and *sting.* Although this topic appeared far less frequently, it represents a small subset of users who experienced mild side effects, especially those with sensitive skin. These users mentioned minor dryness or tingling sensations but did not consider them severe enough to discourage product use.

Overall, the topic modeling analysis highlights that customer discussions were dominated by positive themes, particularly around acne improvement, gentleness, and oil control. Negative aspects like irritation were discussed only occasionally, showing that most users had a favourable experience with the BeMinimalist Salicylic + LHA 2% Cleanser.This also confirms the brand’s key claims—effective acne reduction, gentle formulation, and balanced cleansing—are reflected directly in authentic customer feedback.

3.5 Word Embedding and Similarity Analysis

To understand the semantic relationships between key terms, **Word2Vec embeddings** were generated using the Skip-Gram architecture from the **gensim** library. The embeddings help capture contextual similarity — that is, how closely related words appear together in the same context.For example, the word *gentle* appeared most frequently with similar terms such as *mild*, *soft*, and *soothing*. Likewise, *acne* was often found near *pimples*, *breakouts*, and *spots*, reflecting users' focus on acne reduction.Words like *oil* were linked with *sebum*, *greasy*, and *shine*, indicating that many users discussed the product's effect on oily skin. On the other hand, *irritation* was closely related to *dryness* and *redness*, representing negative experiences reported by a small fraction of users.These associations show that customers consistently use coherent and meaningful vocabulary when describing the product's effects.

3.6 Review Summarization and Simulated Q&A

After generating numerical vector representations of the reviews using TF-IDF and Word2Vec, the cosine similarity measure was applied to determine how semantically close different reviews were to one another. Based on this similarity, groups or clusters of reviews that expressed similar sentiments or discussed common product aspects were identified. From each of these clusters, representative sentences were extracted to create concise summaries that reflected the overall sentiment trends of the customers.Using these summaries, a set of frequently asked customer-style questions was formulated. Each question addressed a key area that potential buyers are usually interested in, such as product performance, skin compatibility, texture, or value for money. The corresponding answers were synthesized from the findings of sentiment analysis, topic modeling, and linguistic analysis, ensuring that each response was both data-driven and grounded in real customer feedback.For instance, most customers confirmed that the cleanser effectively reduced acne and enhanced skin clarity, which directly answers the question, "Does the cleanser reduce acne?". Similarly, responses to the question, "Is it gentle or harsh on skin?" showed that users widely described the product as gentle, mild, and suitable for sensitive skin, though a few mentioned slight dryness after use. Many reviews also highlighted that the cleanser helps control excess oil and leaves the skin feeling clean and refreshed, which addresses common concerns of users with oily skin.When asked about texture and fragrance, reviewers consistently mentioned that the cleanser is lightweight, non-sticky, and pleasantly mild in scent, making it comfortable for everyday use. Furthermore, most users expressed that the product provides excellent value for money, describing it as affordable and effective compared to similar skincare products in the market.These question and answer summaries provide an intuitive and human-readable overview of customer opinions. They translate numerical analysis into conversational insights that can easily be understood by new customers or integrated into AI-based chatbots, customer service dashboards, or recommendation systems. This method ensures that real user feedback is utilized in an intelligent and structured way to enhance product understanding and customer interaction.

4. Challenges & Learnings

During the course of this project, several challenges were encountered while working with traditional Natural Language Processing techniques. These challenges mainly arose due to the nature of the data, the limitations of non-transformer models, and the need for balancing computational efficiency with accuracy. One of the first difficulties was handling multilingual and code-mixed text data. Many customer reviews were written partly in Hindi and partly in English (commonly called Hinglish). Traditional rule-based or statistical models often struggled to identify the correct language boundaries within such text. To overcome this, a rule-based language detection approach was used along with simple frequency-based word analysis. Non-English text was then translated into English using lightweight translation libraries rather than deep neural transformer models, ensuring that the analysis remained interpretable and efficient. Another major challenge was data noise and inconsistency. The scraped reviews contained HTML tags, emojis, URLs, and special characters that could affect text processing. Some reviews were extremely short or repetitive. These were handled through rigorous data cleaning and normalization, including removal of duplicates, irrelevant symbols, and standardization of text encoding. Tokenization, stop-word removal, and lemmatization were performed carefully to maintain the context while reducing redundancy. A further challenge lay in topic modeling and word embedding using limited data. Traditional models like Latent Semantic Analysis and Word2Vec depend heavily on the size and diversity of the corpus. Since product reviews often contain short sentences, ensuring meaningful topic extraction required careful preprocessing and parameter tuning. Techniques like TF-IDF weighting and skip-gram training were used to capture key terms and their relationships effectively. The sentiment analysis step also presented certain difficulties. Unlike deep transformer models such as BERT that can understand context deeply, lexicon-based and LSTM-based models rely more on predefined sentiment dictionaries and learned sequences. As a result, some mixed-tone sentences (for example, "Good product but made my skin dry") were hard to classify accurately. This issue was partly mitigated by combining lexicon polarity scores with contextual LSTM outputs to create a hybrid sentiment classification approach.

On the technical side, implementing parallel scraping using Selenium was challenging due to page loading delays and dynamic JavaScript content. To optimize time, controlled waits and page-scroll simulations were implemented, ensuring that all customer reviews were loaded correctly before extraction. From these experiences, several important learnings emerged. Even with basic non-transformer models, good preprocessing and thoughtful parameter selection can lead to strong, interpretable results. Data quality plays a far greater role than model complexity; clean, well-structured text consistently produces more reliable insights. This project also reinforced the importance of explainable AI, as using transparent models makes it easier to understand and trust the outputs, especially in academic or business settings. Overall, the project was a practical demonstration of how classical NLP methods remain relevant and effective, especially for small to medium datasets where explainability, speed, and reproducibility are more important than pure accuracy.

5. Conclusion

This project successfully demonstrated how traditional Natural Language Processing techniques can be applied to analyze real-world customer reviews in an effective and interpretable way. By focusing on the BeMinimalist Salicylic + LHA 2% Cleanser, the analysis covered the complete NLP pipeline, from data collection and translation to sentiment evaluation, topic discovery, and semantic similarity mapping. The key findings reveal that the product enjoys strong customer satisfaction, with the majority of reviews reflecting positive experiences related to acne reduction, skin gentleness, and affordability. Topic modeling and embedding analysis confirmed that customer feedback aligns well with the brand's marketing claims. The limited presence of negative feedback, mainly mild dryness, further supports the product’s effectiveness and customer trust. The project also proved that non-transformer NLP methods, when properly implemented, can deliver meaningful results even without deep learning architectures. Using tools like spaCy, scikit-learn, and gensim, the analysis achieved clear insights with low computational requirements and full transparency of model behavior. In terms of practical outcomes, this analysis provides valuable business intelligence, identifying the strengths of the product such as effectiveness, gentleness, and price, and pinpointing minor improvement areas such as dryness or irritation. Such findings can directly assist the company in marketing, product improvement, and customer engagement strategies.

For future enhancements, the project can be extended by incorporating real-time sentiment tracking and aspect-based sentiment analysis, where each feature such as texture, price, or effectiveness is evaluated separately. Additionally, integrating lightweight transformer models like DistilBERT could help improve context understanding while maintaining computational efficiency. Another promising extension would be to deploy a Q&A chatbot that uses the summarized insights to automatically answer customer queries. In conclusion, this project highlights the power of classical NLP in practical, real-world applications. It bridges linguistic analysis and customer behavior understanding, providing a robust framework for data-driven decision-making. Through this work, it becomes evident that with thoughtful design and proper methodology, even traditional machine learning approaches can deliver deep insights, meaningful interpretation, and measurable impact.