MULTIMEDIA UNIVERSITY

FACULTY OF COMPUTING AND INFORMATICS

BCS.

SOCIAL MEDIA COMPUTING – CDS6344

TRIMESTER, Session 2024/2025

Sentiment Analysis on Sephora Product Reviews

By: Shahnaz Binti Husain Sukri | 1211101888

Yusriena Aqielah Binti Muhammad Nafiz Hans | 1221302876

Nor Aliah Syuhaidah Binti Sharuddin | 1211102031

1. Introduction

In this digital age, the amount of user generated content such as reviews, comments, and feedback has become a valuable source of insights for both consumers and businesses. Sephora, a leading beauty retailer, receives thousands of unstructured product reviews from its customers. These reviews contain valuable insights into customer satisfaction, preferences, and experiences with various products. However, the sheer volume and unstructured nature of this data make it challenging to extract actionable information manually.

Traditionally, sentiment analysis focuses on determining the overall sentiment of a document, sentence, or phrase as positive, negative, or neutral. However, this general-level sentiment is often insufficient in real-world scenarios. For example, a product review might praise the packaging but criticize the delivery time. To capture such detailed information, Aspect-Based Sentiment Analysis (ABSA) is used to break down text into specific aspects and analyze the sentiment associated with each.

This project aims to use sentiment analysis techniques to analyze Sephora product reviews. By applying natural language processing (NLP) and both classical and deep learning techniques, we seek to identify patterns in customer opinions, classify sentiments, and provide actionable insights that can help business decision making.

1.1 Project Overview

This project focuses on extracting and analyzing customer sentiments from a large dataset of Sephora product reviews. The primary objectives of this project are:

- To preprocess and clean the raw review data for effective analysis.
- To apply sentiment labeling, categorizing reviews as positive, neutral, or negative based on their content.
- To explore the distribution of sentiments across different brands and products.
- To implement and compare various machine learning and deep learning models for sentiment classification.
- To visualize the results and provide insights.

The dataset used in this project, "Sephora Products and Skincare Reviews," was sourced from Kaggle and contains detailed information about product reviews, ratings, and customer information. Through this project, we aim to demonstrate the value of automated sentiment analysis in understanding customer feedback.

2. Problem Statement

With the increasing reliance on online reviews, businesses like Sephora are challenged with understanding vast amounts of customer feedback. While these reviews offer valuable insights, they are mainly unstructured, making it difficult to analyze them efficiently or consistently using manual methods.

This lack of structure affects Sephora's ability to fully understand what customers like or dislike about specific products. Without a system to systematically analyze sentiments across product features, important feedback may be overlooked, and customer dissatisfaction may go unaddressed.

Therefore, there is a need for an automated and scalable solution to process these reviews, extract relevant opinions, and categorize sentiments in a meaningful way. Addressing this problem can help Sephora enhance product development, marketing strategies, and overall customer experience.

3. Literature Review

The literature on sentiment analysis of product reviews has evolved rapidly over the past decade, moving from classical machine learning techniques to deep learning and, most recently, Transformer-based architectures. In parallel, Aspect-Based Sentiment Analysis (ABSA) has emerged as a critical subfield for extracting fine-grained opinions about specific product features. Below, we review key contributions in each of these areas.

3.1 Traditional Machine Learning Approaches

Early work in review sentiment classification relied on manually engineered features (e.g., bag-of-words, TF-IDF) [1] combined with classifiers such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests. These methods achieved reasonable accuracy on balanced datasets but often struggled with nuanced language (e.g., sarcasm, mixed sentiments) and required extensive feature engineering.

Key studies:

- Naïve Bayes and SVM models applied to product review sentiment yielded baseline accuracies of 70–80% using unigrams and bigrams with bag-of- features [1]

3.2 Deep Learning Techniques

The advent of deep learning enabled models to learn hierarchical representations of text, reducing reliance on manual feature extraction. Convolutional Neural Networks (CNNs) captured local ngram features, while Recurrent Neural Networks (RNNs), especially LSTM and BiLSTM variants, modeled longer-range dependencies. Comparative studies indicate that deep architectures often outperform classical methods, particularly when large labeled datasets are available.

Key findings:

- CNN and RNN models achieved 5–10% higher accuracy over SVM baselines, thanks to their ability to learn semantic patterns directly from embeddings [2]
- Hybrid models combining BERT embeddings with BiLSTM layers further improved F1-scores on e-commerce review datasets [3]

3.3 Transformer-Based Models

Recent advances in NLP have been driven by transformer architectures, such as BERT [4], RoBERTa [5], and DistilBERT [6]. These models leverage self-attention mechanisms to capture long-range dependencies and contextual information in text. Pre-trained on massive corpora, transformers can be fine-tuned for specific tasks, achieving state-of-the-art results in sentiment classification and other NLP applications.

3.4 Aspect-Based Sentiment Analysis (ABSA)

While document-level sentiment classification provides a coarse view, ABSA breaks down text into aspects (e.g., “packaging,” “ingredients,” “delivery”) and classifies the sentiment associated with each. Recent surveys highlight key subtasks—aspect for term extraction, aspect polarity classification, and end-to-end architectures—and review trends in solution paradigms (rule-based, supervised, and deep learning).

Survey highlights:

- Systematic reviews of ABSA identify attention-based and graph-convolutional methods as leading approaches for capturing aspect–context interactions [7]
- Challenges include handling implicit aspects and domain adaptation; public datasets span domains from restaurants to electronics, with relatively few in cosmetics [8]

3.5 Summary

This review shows a clear trajectory from traditional machine learning through deep learning to Transformer-based and aspect-aware models. While domain-agnostic architectures (e.g., BERT) deliver strong baseline performance, cosmetic review analysis benefits from domain-specific fine-tuning and ABSA techniques to extract actionable insights on distinct product features. The present project builds on these foundations by comparing classical and deep learning models on a Sephora reviews dataset, and by incorporating aspect-level sentiment classification to uncover granular customer perceptions.

4. Methodology

To analyze Sephora product reviews, our project followed a complete NLP pipeline that included data preprocessing, sentiment classification, and aspect-based sentiment analysis (ABSA). Both classical machine learning models and transformer-based models were used to compare performance and accuracy.

4.1 Data Collection & Cleaning

The dataset used, “Sephora Products and Skincare Reviews,” was sourced from Kaggle. It contains customer reviews, ratings, product and brand information, and various user attributes. The dataset was then preprocessed by doing the following:

- Data Cleaning: Irrelevant columns (such as author ID, review title, product ID) were removed. Reviews with missing values in critical fields were dropped to ensure data quality.
- Text Normalization: All review texts were converted to lowercase, and punctuation was removed.
- Stopword Removal & Lemmatization: Using NLTK, common stopwords were removed and words were lemmatized to their base forms, resulting in a cleaned review text column.
- Sentiment Labeling: Reviews were labeled as ‘positive’, ‘neutral’, or ‘negative’ based on their rating (e.g., ratings ≥4 as positive, 3 as neutral, ≤2 as negative).
- Resampling Data: To address class imbalance in the dataset, we performed resampling by downsampling the majority classes (positive and negative) and upsampling the minority class (neutral) to achieve a balanced distribution of 88,000 samples per class.

4.2 Feature Engineering

For the machine learning models to process textual data, the cleaned review texts were transformed into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique. This approach captures the importance of words in each review relative to the entire corpus. In some models, additional features such as review length and product or brand information were also considered to enhance predictive performance.

4.3 Model Selection and Training

A range of models was implemented to classify the sentiment of reviews. Traditional machine learning algorithms, including Logistic Regression, Naive Bayes, and Support Vector Machine (SVM), were trained using the engineered features. In addition, transformer-based models BERT, DistilBERT, and RoBERTa were fine-tuned on the dataset to leverage their ability to capture complex linguistic patterns and context.

Each model was trained on the processed dataset, with hyperparameters tuned to optimize performance. The training process involved splitting the data into training and testing sets to evaluate the models' generalization capabilities.

4.4 Model Evaluation

The performance of each model was assessed using metrics such as accuracy, precision, recall, F1 score, and confusion matrices. K-fold cross-validation was employed to ensure the robustness and stability of the results, providing a comprehensive evaluation of each model's strengths and weaknesses.

5. Sentiment Analysis

Sentiment analysis is a core component of this project, enabling the extraction of emotional tone and subjective opinions from Sephora product reviews. The goal is to classify not only the overall sentiment of a review but also to dig deeper into what exactly customers like or dislike.

We approached sentiment analysis in three layers: general sentiment classification, aspect-based sentiment analysis (ABSA), and opinion mining

5.1 General Sentiment Classification

Each review was labeled as: positive, neutral, and negative. The sentiment was determined based on customer ratings. Specifically, ratings of 4 or 5 were considered positive, a rating of 3 was neutral, and ratings of 1 or 2 were labeled negative. This method provided a simple and consistent way to assign sentiment labels without manual tagging.

Before running any analysis, the review text was cleaned and prepared. This included steps like converting text to lowercase, removing punctuation and common stopwords, and simplifying words to their root forms through lemmatization. The sentiment labels were also converted into numbers so they could be used in the models. After that, the data was split into training and testing sets, with 80% used for training and 20% set aside for testing.

To create a starting point for sentiment classification, 3 traditional machine learning models were tested. The review text was transformed into numerical features using a TF-IDF vectorizer, which captured the most important words and short phrases. Then, three models were trained: Logistic Regression, Support Vector Machine (SVM), and Naive Bayes. These models gave a starting point for measuring how well sentiment could be predicted before using more advanced methods. The models were then evaluated using standard classification metrics and 5-fold cross-validation was applied to ensure robustness.

5.2 Aspect-Based Sentiment Analysis (ABSA)

While general sentiment analysis gives an overview of customer satisfaction, Aspect-Based Sentiment Analysis (ABSA) allows for more granular insights. ABSA identifies specific aspects of a product (e.g., “texture”, “packaging”, “price”) mentioned in the reviews and determines the sentiment associated with each aspect. Aspects key nouns representing product features were extracted from the cleaned review texts using the spaCy library.

The ABSA results were visualized using aspect-sentiment heatmaps, frequency bar charts, and brand-aspect sentiment comparison charts, which revealed trends like common strengths and weaknesses across products and brands.

5.3 Opinion Mining

Beyond aspect-level analysis, opinion mining was conducted to extract commonly expressed opinions across all reviews. Opinion pairs (e.g., “good coverage”, “too oily”) were aggregated to understand how users describe various product features.

The top opinion pairs were identified for both the entire dataset and per-brand basis. This allowed us to discover not only which features were most talked about but also the sentiment behind those opinions.

6. Transformers

Recent advances in NLP have been driven by transformer-based models, which have set new benchmarks for a variety of text classification tasks, including sentiment analysis. In this project, several transformer models were fine-tuned and evaluated on the Sephora product reviews dataset to compare their performance with traditional machine learning approaches.

6.1 Overview of Transformer Models

Transformers are deep learning models that utilize self-attention mechanisms to capture contextual relationships in text. Unlike earlier models, transformers can process entire sentences in parallel and understand long-range dependencies, making them highly effective for sentiment analysis.

The following pre-trained transformer models were used in this project:

- BERT (Bidirectional Encoder Representations from Transformers): A widely used model that learns deep bidirectional representations by jointly conditioning on both left and right context in all layers.
- DistilBERT: A lighter and faster version of BERT that retains most of its performance while being more efficient.
- RoBERTa (Robustly Optimized BERT Approach): An improved variant of BERT, trained with more data and optimized training strategies for better performance.

6.2 Fine Tuning and Training

Each model was loaded from Hugging Face's transformers library using its respective tokenizer and classification head. The review text was tokenized with truncation and padding, and mapped into a custom dataset format compatible with PyTorch's Dataset class. A GPU was used when available to accelerate training.

Due to computational constraints and long training times, only a 20,000 sample subset of the dataset was used for fine-tuning each model. Stratified train-test splits ensured balanced sentiment classes across training and evaluation sets.

Model Settings:

- Max Sequence Length: 128 tokens
- Batch Size: 8
- Epochs: 2
- Loss Function: CrossEntropyLoss
- Metrics: Accuracy, Precision, Recall, F1-Score

7. Result & Visualization

This section presents the outcomes of the sentiment analysis performed on Sephora product reviews, comparing the performance of traditional machine learning models and transformer-based deep learning models. Visualizations are used throughout to illustrate key findings and support the interpretation of results.

7.1 Sentiment Distribution

The distribution of sentiment classes (positive, neutral, negative) before and after balancing the dataset was visualized using a bar plot. This helped confirm that the resampling process resulted in relatively equal representation for all classes, reducing bias in model training.

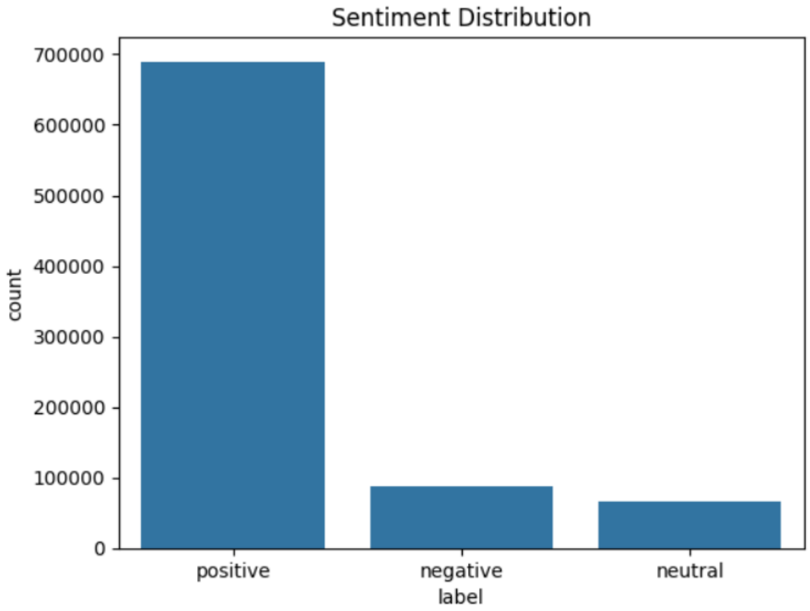

Figure 1: Sentiment Distribution Before Data Resampling

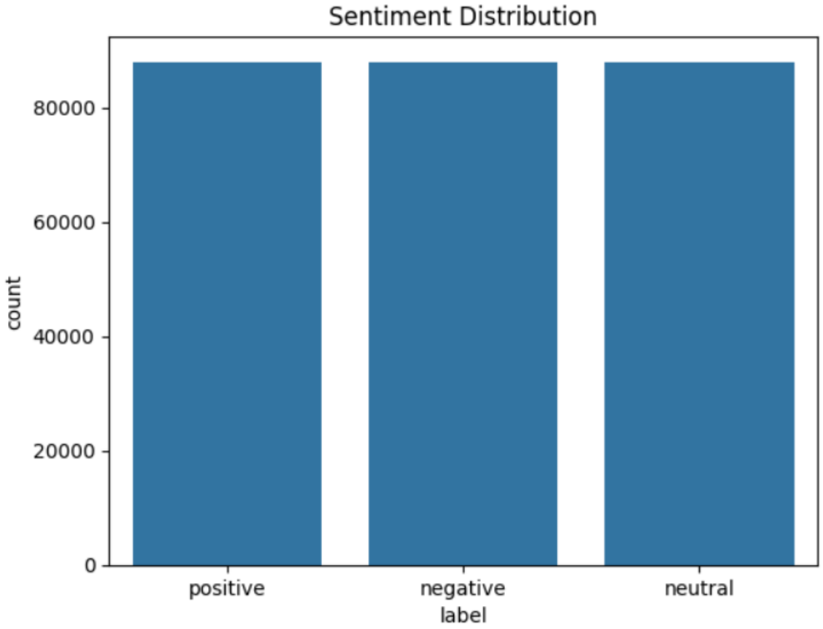

Figure 2: Sentiment Distribution After Data Resampling

7.2 Word Clouds

Word clouds were generated for positive and negative reviews to highlight the most frequently used terms in each sentiment category. These visualizations provide qualitative insights into the language and topics associated with different sentiments.

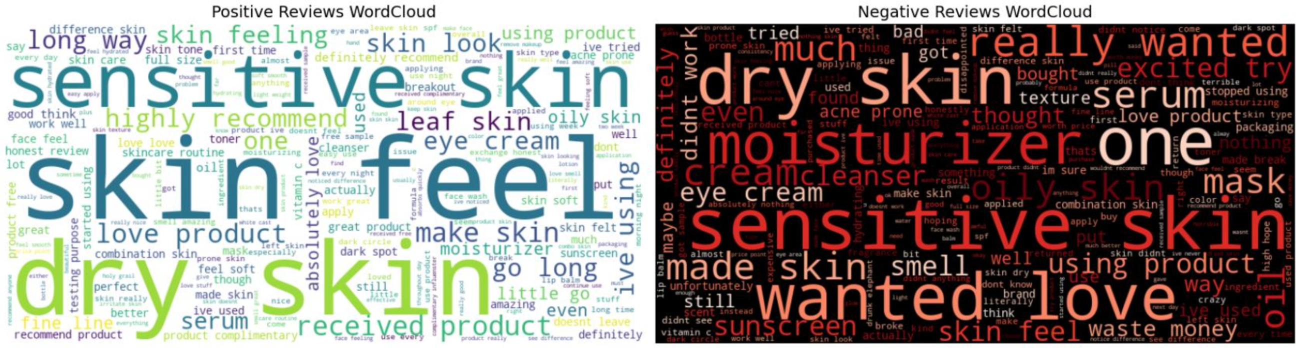

Figure 3: Word Clouds for Positive and Negative Class

In positive reviews, terms such as “love,” “feel,”, and “great” frequently appear, reflecting satisfaction with the product's feel, effect on sensitive skin, and overall performance.

On the other hand, negative reviews prominently feature words like “wanted,”, “waste,” “bad”, and “didn’t,” indicating unmet expectations, poor hydration, or ineffectiveness.

7.3 Traditional Machine Learning Results

The following classifiers were trained using TF-IDF features on the cleaned reviews:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Naïve Bayes (NB)

Each model was evaluated on the test set with metrics including accuracy, precision, recall, and F1-score. Below are sample results:

Table 1: Traditional Machine Learning Results

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Logistic Regression	0.7425	0.7420	0.7425	0.7422
Naive Bayes	0.7138	0.7161	0.7138	0.7148
SVM	0.7400	0.7377	0.7400	0.7385

Observations:

- Logistic Regression performed slightly better than SVM in terms of all metrics, making it the best-performing classical model overall.
- Naive Bayes, while faster and simpler, performed slightly worse than both Logistic Regression and SVM, which is expected due to its assumption of feature independence.
- The metrics are fairly balanced, indicating the models are not overfitting to any single class.

Confusion matrices were plotted for each classifier, providing a visual representation of misclassifications among the three sentiment classes.

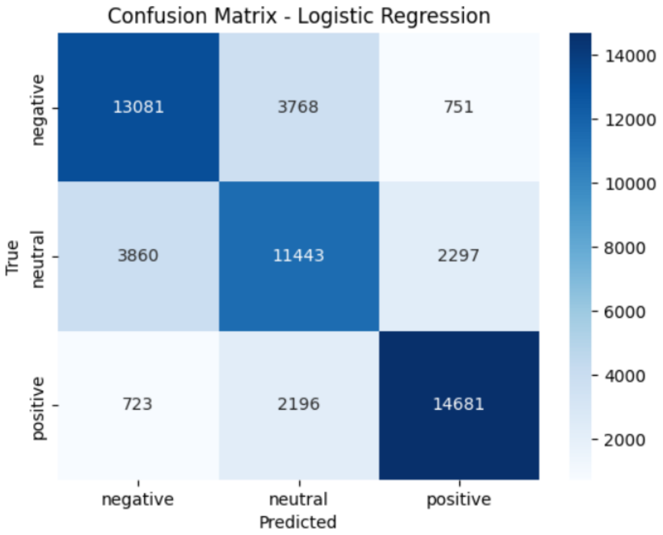

Figure 4: LR Confusion Matrix

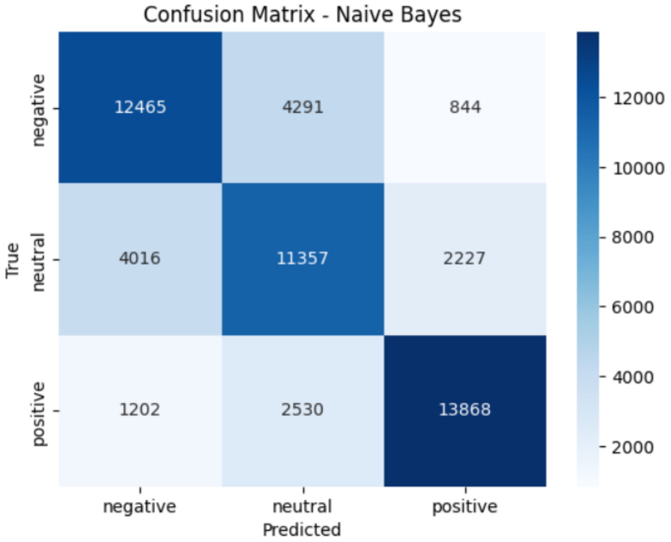

Figure 5: NB Confusion Matrix

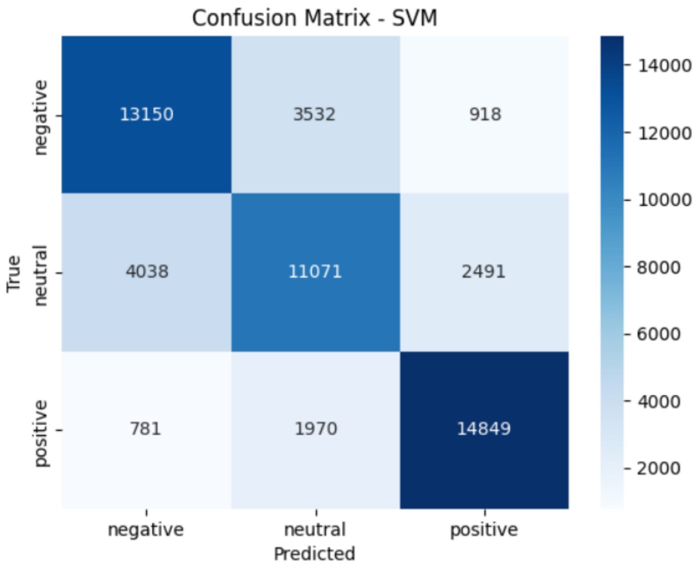

Figure 6: SVM Confusion Matrix

7.4 K-Fold Cross Validation

To further validate classical models, 5-fold Stratified Cross-Validation was applied. Below are the accuracies for each fold:

Table 2: K-Fold Cross-Validation Accuracy Across Models

Model	*Fold 1*	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean Accuracy*	*Std. Dev*
Logistic Regression	0.7474	0.7468	0.7487	0.7456	0.7436	0.7464	0.0017
Naive Bayes	0.7152	0.7173	0.7172	0.7151	0.7136	0.7157	0.0014
SVM	0.7437	0.7419	0.7460	0.7429	0.7403	0.7430	0.0019

7.5 Transformer-Based Model Results

To evaluate the performance of modern deep learning models in sentiment classification, we implemented three transformer-based models: BERT, RoBERTa, and DistilBERT. Due to computational constraints, we trained each model on a subsample of 20,000 reviews to ensure efficient experimentation without compromising representativeness.

Each model was fine-tuned for multi-class sentiment classification (positive, negative, neutral). We used standard evaluation metrics: accuracy, precision, recall, F1-score, and loss to assess model effectiveness.

Table 3: Transformer-Based Model Results

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Loss*
BERT	0.7958	0.7997	0.7958	0.7971	0.5851
RoBERTa	0.7883	0.7888	0.7883	0.7885	0.5364
DistilBERT	0.7840	0.7875	0.7840	0.7854	0.5927

All three models performed well, with BERT achieving the highest accuracy and F1-score. These results show the potential of transformer models in sentiment classification, especially for handling longer and more complex reviews.

7.6 ABSA Results

To gain deeper insight into which specific aspects of products customers react to, ABSA was conducted on the review dataset. Each review was parsed to extract aspect terms (e.g., packaging, price, texture) along with the associated sentiment.

1. **Review 1**: its the best toner the skin is fully hydrated and the pores are less visible i love the texture and it absorbs quickly

   Aspects: ['toner', 'skin', 'pore', 'love', 'texture']
   Sentiment: positive

2. **Review 2:** good product but size is too small when i opened the jar is was not even fun

   Aspects: ['product', 'size', 'jar', 'fun']
   Sentiment: neutral

3. **Review 3:** meh if you dont have a prescription for a good retinol you might try this as a starter after seven days minimal peel prescription grade retinol or tretinoin cream works better faster stronger almost too strong using the packets there was nothing to return unsatisfied

   Aspects: ['meh', 'retinol', 'starter', 'day', 'peel', 'prescription', 'grade', 'retinol', 'tretinoin', 'cream', 'work', 'packet', 'return']
   Sentiment: negative

These example demonstrates how ABSA allows us to break down a review into specific components, helping us understand exactly which parts of the product or experience the user is dissatisfied with.

To visualize which aspects were most frequently discussed and the sentiments attached to them, an aspect-wise sentiment heatmap was generated.

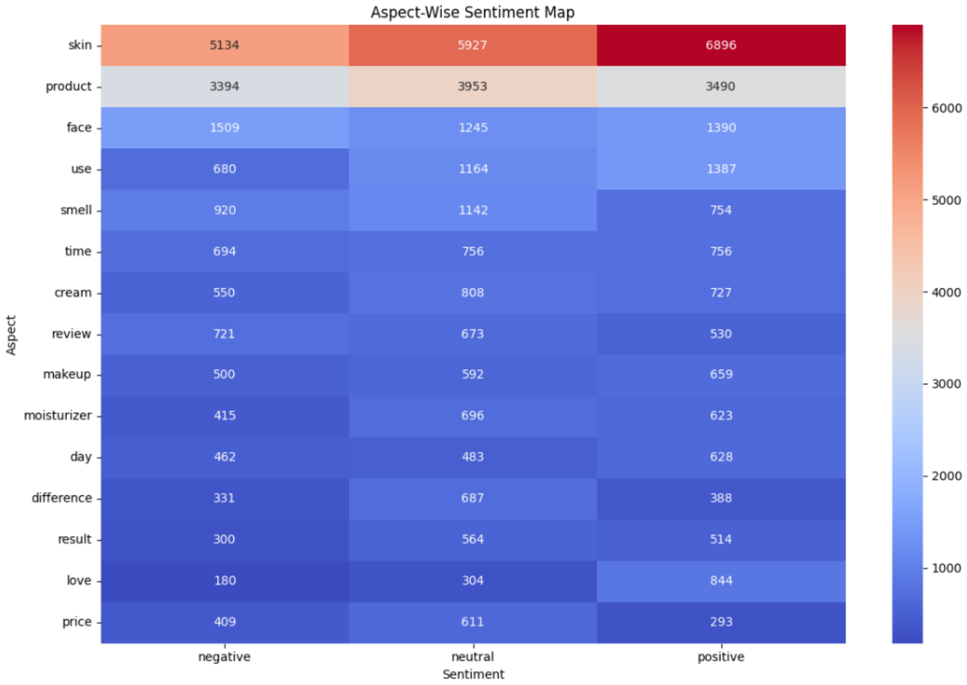

Figure 7: Aspect Wise Sentiment Map

The heatmap above visualizes the sentiment distribution for the top aspects mentioned in the reviews. Each row represents a key aspect (such as “skin,” “product,” “face,” “smell,” or “price”), while the columns correspond to negative, neutral, and positive sentiments. The color intensity and the numbers indicate how often each aspect was mentioned in a particular sentiment context.

Key observations from the aspect-wise sentiment map:

- Skin is the most discussed aspect, with a high number of both positive and negative mentions, indicating its central importance to customers and a diversity of experiences.
- Product and face are also frequently mentioned, with a relatively balanced sentiment distribution.
- Aspects like smell, cream, and moisturizer show a predominance of positive sentiment, suggesting customer satisfaction with these features.
- Price and difference have a higher proportion of negative or neutral mentions, which may indicate concerns or unmet expectations regarding value or noticeable results.
- The aspect love is overwhelmingly associated with positive sentiment, reflecting strong customer approval when this term is used.

Lastly, we visualize the results of ABSA for three leading skincare brands: Tatcha, Drunk Elephant, and The Ordinary using heatmaps.

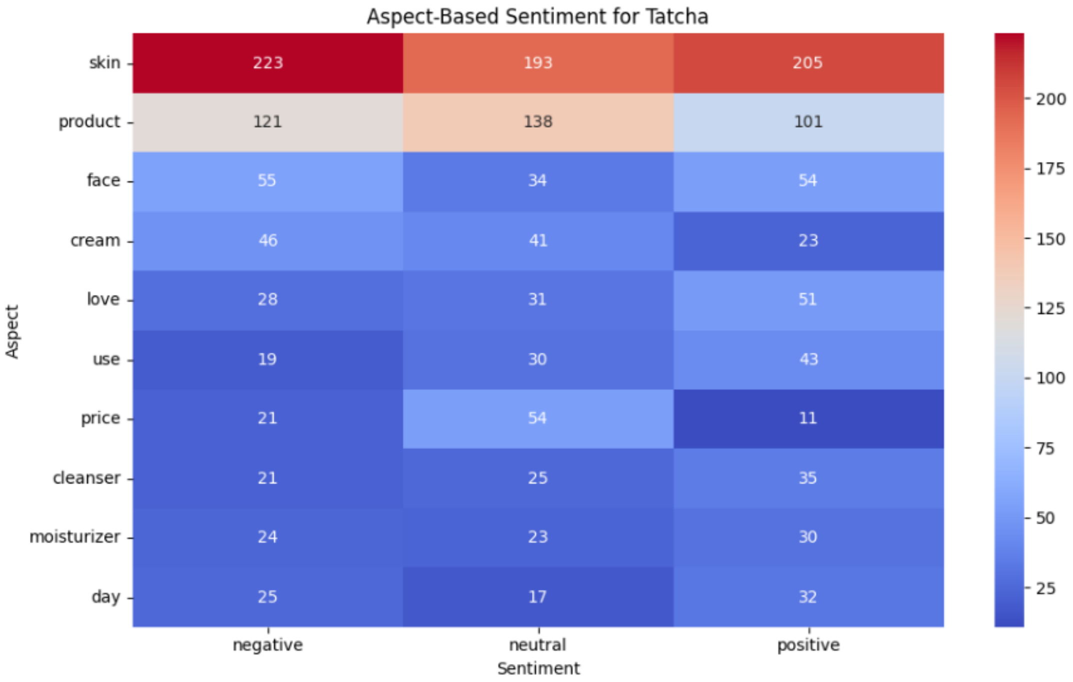

Figure 8: ABSA for Tatcha

- Skin is the most discussed aspect, with a relatively balanced distribution across negative (223), neutral (193), and positive (205) sentiments. This suggests that while many customers are satisfied with how Tatcha products affect their skin, a significant number also report neutral or negative experiences.
- Product and face are also frequently mentioned, with a slight tilt towards negative and neutral sentiments, indicating mixed customer perceptions.
- Aspects like cream, moisturizer, and cleanser show a more even sentiment distribution, but with fewer overall mentions.
- Love stands out as an aspect with more positive (51) than negative (28) or neutral (31) mentions, reflecting strong approval when customers use this term.
- Price is mentioned more often in a neutral or negative context, suggesting that cost may be a concern for some customers considering the Tatcha's high end prices.

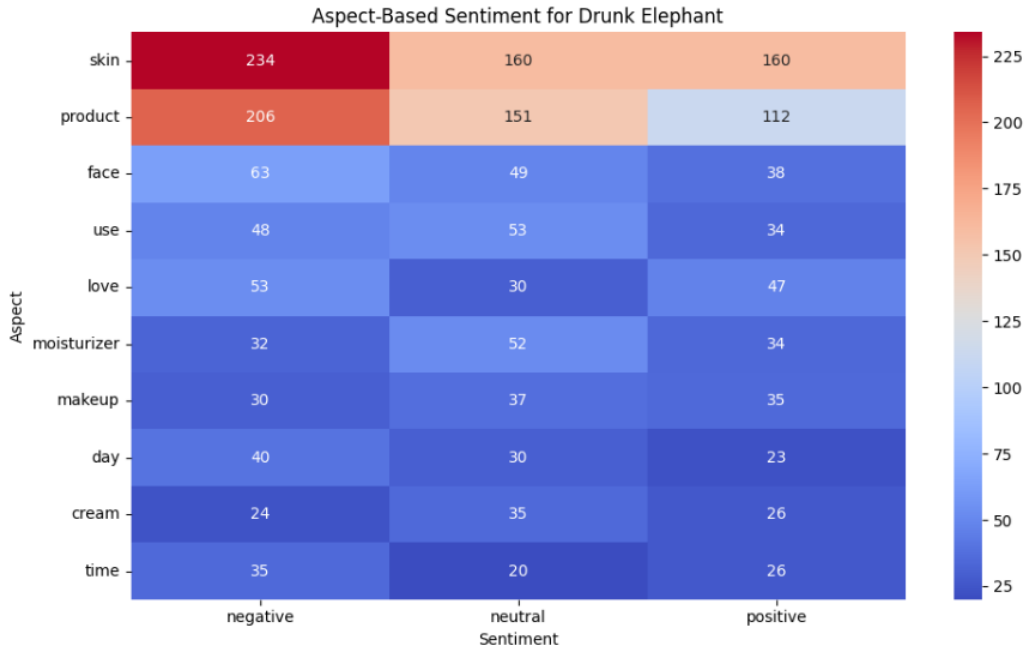

Figure 9: ABSA for Drunk Elephant

- Skin and product are the most frequently discussed aspects, with a higher number of negative mentions (234) compared to positive (160 and 160). This may indicate that some customers have concerns or unmet expectations regarding these aspects.
- Face, use, and love are also common, with sentiment more evenly distributed, though still with a slight negative skew.
- Moisturizer and makeup aspects have a relatively balanced sentiment, but with a modest number of positive mentions.
- Day, cream, and time are less frequently mentioned, with sentiment distributed fairly evenly across categories.

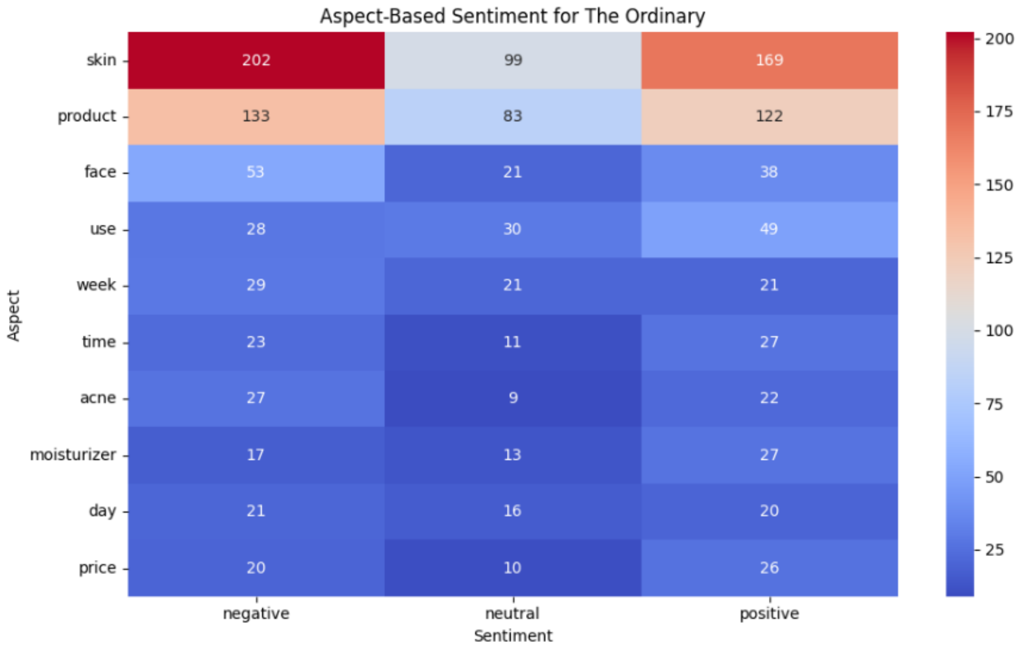

Figure 10: ABSA for The Ordinary

- Skin and product are again the most discussed aspects. For skin, negative (202) and positive (169) mentions are both high, but there are fewer neutral (99) mentions, indicating polarized experiences among customers.
- Product shows a similar pattern, with a notable number of both negative and positive mentions.
- Aspects such as face, use, and week have a more balanced sentiment distribution, though with fewer total mentions.
- Acne and moisturizer are mentioned less frequently, but positive sentiment is slightly higher for these aspects.
- Price is mentioned least, but when it is, it tends to be in a positive context, which is in line with the brand's affordability.

Across all three brands, certain patterns emerge from the aspect-based sentiment analysis. The aspects "skin" and "product" consistently dominate customer discussions, underscoring their central importance in the skincare experience. For each brand, these aspects receive a mix of positive and negative sentiments, indicating that while many customers are satisfied, a significant portion also report neutral or negative experiences, reflecting the diversity of individual skin responses and expectations. Notably, the aspect "love" is associated with a higher proportion of positive sentiment across brands, highlighting features or experiences that particularly resonate with and delight customers. Interestingly, the aspect of "price" received varying sentiments across different brands where it was often mentioned negatively for premium or high-end brands, while being viewed more positively or neutrally for affordable brands, highlighting how perceived value is brand dependent. Overall, while each brand has its unique sentiment profile, the distribution of opinions across key aspects provides valuable guidance. This comparative analysis enables brands to better understand their market position and prioritize improvements that align with customer feedback.

7.7 Opinion Mining Results

Opinion mining was conducted to identify the most common descriptive phrases or opinion pairs expressed in the reviews. These pairs capture recurring themes and user sentiments related to product performance, experience, or personal preferences.

Below are the top 10 opinion pairs extracted from each sentiment category:

Table 4: Top Opinions in Positive Reviews

Opinion Pair	*Frequency*
('dry', 'skin')	485
('sensitive', 'skin')	437
('long', 'way')	292
('using', 'product')	209
('oily', 'skin')	193
('great', 'product')	176
('using', 'skin')	171
('honest', 'review')	159
('fine', 'line')	152
('great', 'skin')	135

Table 5: Top Opinions in Negative Reviews

Opinion Pair	*Frequency*
('sensitive', 'skin')	445
('dry', 'skin')	417
('oily', 'skin')	197
('using', 'product')	174
('using', 'skin')	119
('prone', 'skin')	108
('used', 'product')	97
('first', 'time')	95
('left', 'skin')	91
('fine', 'line')	80

Table 6: Top Opinions in Neutral Reviews

Opinion Pair	*Frequency*
('dry', 'skin')	506
('sensitive', 'skin')	395
('oily', 'skin')	308
('using', 'product')	182
('fine', 'line')	142
('honest', 'review')	134
('long', 'way')	112
('using', 'skin')	108
('good', 'product')	107
('using', 'week')	107

From the analysis of top opinion pairs, it's evident that skin type-related terms such as “dry skin”, “sensitive skin”, and “oily skin” are frequently mentioned across all sentiment categories, indicating that skin compatibility is a major concern for Sephora customers. Interestingly, while “great product” and “long way” appear in positive reviews, phrases like “left skin”, and “used product” are prominent in negative reviews, reflecting dissatisfaction with perceived value and product performance. This distinction helps identify which aspects drive positive or negative user experiences.

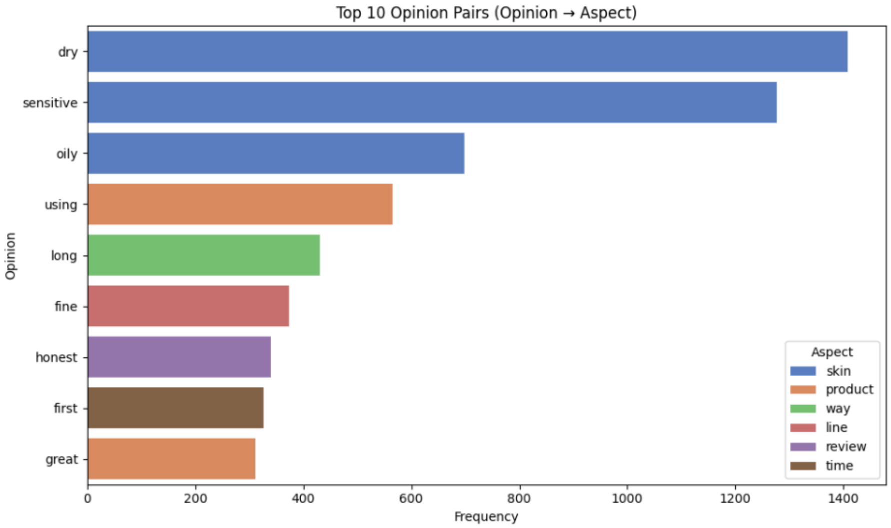

Figure 11: Top 10 Opinion Pairs

8. Discussion

This section reflects on the performance of the different models used and interprets the sentiment analysis results from both classical and transformer-based approaches.

8.1 Traditional Machine Learning Models

Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) were evaluated using TF-IDF features. Among them, Logistic Regression performed the best with an accuracy of 74.25%, followed closely by SVM (74.00%). Naive Bayes, while simpler and faster, yielded a slightly lower accuracy of 71.38%.

K-fold cross-validation (k=5) was also applied to ensure reliability. The average cross-validation accuracy for Logistic Regression was approximately 74.64%, with low standard deviation, indicating stable performance across different folds. Despite their simplicity, these classical models demonstrated competitive performance and served as effective baselines for comparison with transformer-based models.

8.2 Transformer Based Models

Three transformer models, BERT, RoBERTa, and DistilBERT were fine-tuned on a smaller 20,000 sample subset due to computational constraints. All three outperformed the classical models in terms of accuracy and F1-score, with BERT achieving the highest accuracy (79.58%).

Interestingly, RoBERTa, despite having a lower loss than BERT, slightly underperformed in overall metrics. This may be due to sensitivity to training hyperparameters or class distribution in the sampled data. DistilBERT, being a lighter model, performed slightly lower but still showed strong results compared to traditional approaches.

8.3 Limitations & Considerations

- The transformer models were trained on a reduced dataset, which may have limited their full potential.
- The neutral class required oversampling, which might have introduced some noise or class imbalance artifacts.
- Hardware limitations restricted the number of epochs and model size that could be tested, particularly for transformer-based models.

9. Conclusion & Future Work

9.1 Conclusion

This project explored sentiment analysis and aspect-based sentiment analysis (ABSA) on Sephora product reviews using both classical machine learning and modern transformer-based models.

Classical models such as Logistic Regression, Naive Bayes, and SVM served as strong baselines, with Logistic Regression achieving the highest accuracy of 74.25%. Transformer models, particularly BERT, demonstrated superior performance, achieving an accuracy of 79.58% on a subset of the data.

Beyond overall sentiment classification, ABSA and opinion mining provided insights into customer opinions tied to specific aspects like product texture, scent, and price. These findings are highly relevant for product teams and marketing strategists aiming to improve customer satisfaction and brand perception.

The application of k-fold cross-validation further ensured the robustness and consistency of the models, especially in the presence of imbalanced sentiment classes.

9.2 Future Work

While the results are promising, there are several directions for improvement and expansion:

- Full Dataset Training: Fine-tuning transformer models on the entire dataset (instead of a 20k subset) may yield better generalization and performance.
- Hyperparameter Tuning: Experimenting with learning rates, batch sizes, and training epochs can further optimize transformer performance.
- Multilingual Support: Incorporating multilingual models could help analyze reviews in languages other than English.
- Real-Time Sentiment Dashboard: Building an interactive dashboard would allow businesses to monitor aspect-level sentiment trends over time.

In summary, this project demonstrated how combining NLP with deep learning can extract meaningful insights from unstructured customer reviews and empower data-driven decision-making in the beauty industry.

10. References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," in Proc. ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86, 2002. [Online]. Available: https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

[2] L. G. Atlas, D. Arockiam, A. Muthusamy, B. Balusamy, S. Selvarajan, T. Al-Shehari, and N. A. Alsadhan, "A modernized approach to sentiment analysis of product reviews using BiGRU and RNN based LSTM deep learning models," *Scientific Reports*, vol. 15, no. 1, pp. 1–24, Jun. 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-01104-0

[3] S. K. Suriya and P. Ponsenthil, "Sentiment analysis of product reviews using deep learning techniques," *Int. J. Adv. Eng. Manag.*, vol. 4, no. 4, pp. 345–350, 2022. [Online]. Available: https://ijaem.net/issue_dcp/Sentiment%20analysis%20of%20product%20reviews%20using%20Deep%20Learning%20techniques.pdf

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/pdf/1810.04805

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/pdf/1907.11692

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, Oct. 2019. [Online]. Available: https://arxiv.org/pdf/1910.01108

[7] Y. C. Hua, P. Denny, J. Wicker, and K. Taskova, "A systematic review of aspect-based sentiment analysis: domains, methods, and trends," *Artificial Intelligence Review*, vol. 57, pp. 2239–2294, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10906-z

[8] M. Wankhade, C. Kulkarni, and A. C. S. Rao, "A survey on aspect-based sentiment analysis methods and challenges," *Applied Soft Computing*, vol. 150, 2024, Art. no. 110238. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494624010238