Sentiment Analysis on Sephora Product Reviews

An in-depth exploration using machine learning and deep learning models

By: Shahnaz Binti Husain Sukri | 1211101888 Yusriena Aqielah Binti Muhammad Nafiz Hans | 1221302876 Nor Aliah Syuhaidah Binti Sharuddin | 1211102031

Project Overview

- Analyzed Sephora product reviews using sentiment analysis techniques
- Goal: Classify reviews as positive, neutral, or negative
- Used traditional ML models (Logistic Regression, SVM, Naive Bayes)
- Applied transformer-based models (BERT, RoBERTa, DistilBERT) for deeper understanding
- Included Aspect-Based Sentiment Analysis (ABSA) to capture opinions on specific features like texture, price, and smell
- Performed opinion mining to identify common phrases (e.g., “dry skin”, “great product”)

Problem Statement

- Sephora receives a high volume of unstructured customer reviews.
- Manually analyzing these reviews is time-consuming and inefficient.
- Valuable insights about product satisfaction and customer preferences may be overlooked.
- Lack of detailed analysis on specific product aspects (e.g., packaging, texture, price).
- A scalable, automated sentiment analysis system is needed to extract meaningful feedback and support data-driven decisions.

Literature Review

Traditional ML

- Models: Naïve Bayes, SVM, Logistic Regression
- Pros: Simple and fast for small datasets.
- Cons: Struggles with complex sentiments like sarcasm or mixed feelings.

Deep Learning

- Models: CNN, RNN (LSTM, BiLSTM)
- Advantage: Captures hierarchical relationships and long-range dependencies in reviews.
- Result: Achieved better accuracy compared to traditional ML methods, particularly on large datasets.

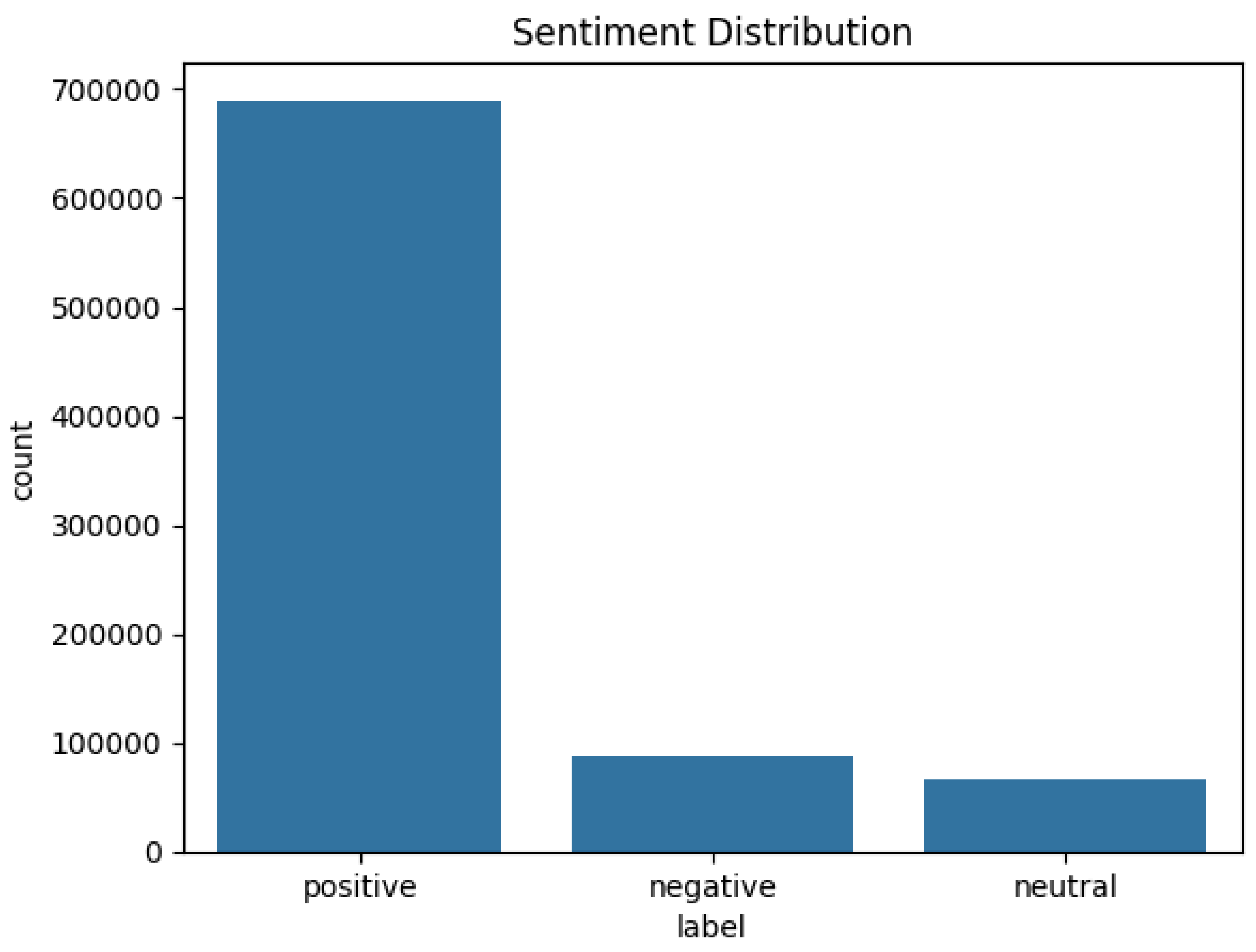
Transformer Models

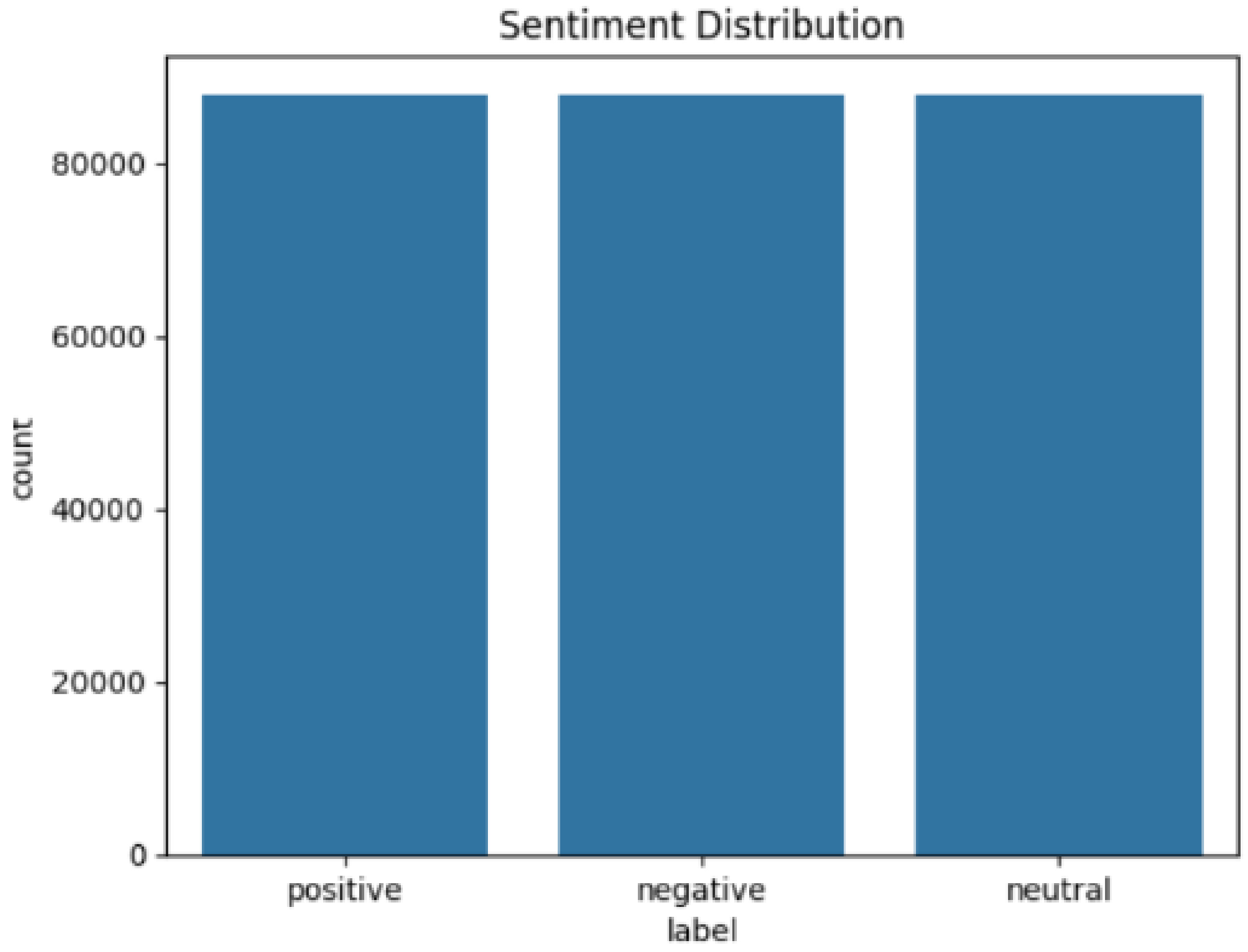
- Models: BERT, RoBERTa, DistilBERT
- Strength: These models leverage self-attention mechanisms to capture context over long distances in text, resulting in superior performance for sentiment analysis tasks.

Methodology

Data Collection & Cleaning:

- Dataset: Sephora product reviews sourced from Kaggle.
- Preprocessing Steps:
  - Removal of irrelevant columns (e.g., review title, author ID).
  - Text normalization: converting to lowercase, removing punctuation.
  - Lemmatization and stopword removal using NLTK for better feature extraction.
- Sentiment Labeling: Reviews were labeled as positive, neutral, or negative based on their rating .
- Resampling Data: To address class imbalance in the dataset, we performed resampling to achieve a balanced distribution of 88,000 samples per class.





Methodology

Feature Engineering:

- Used TF-IDF (Term Frequency-Inverse Document Frequency) to convert text into numerical features, allowing models to understand the importance of words across reviews.
- Enhanced model performance by adding features like review length and product information.

M**odel Selection and Training:**

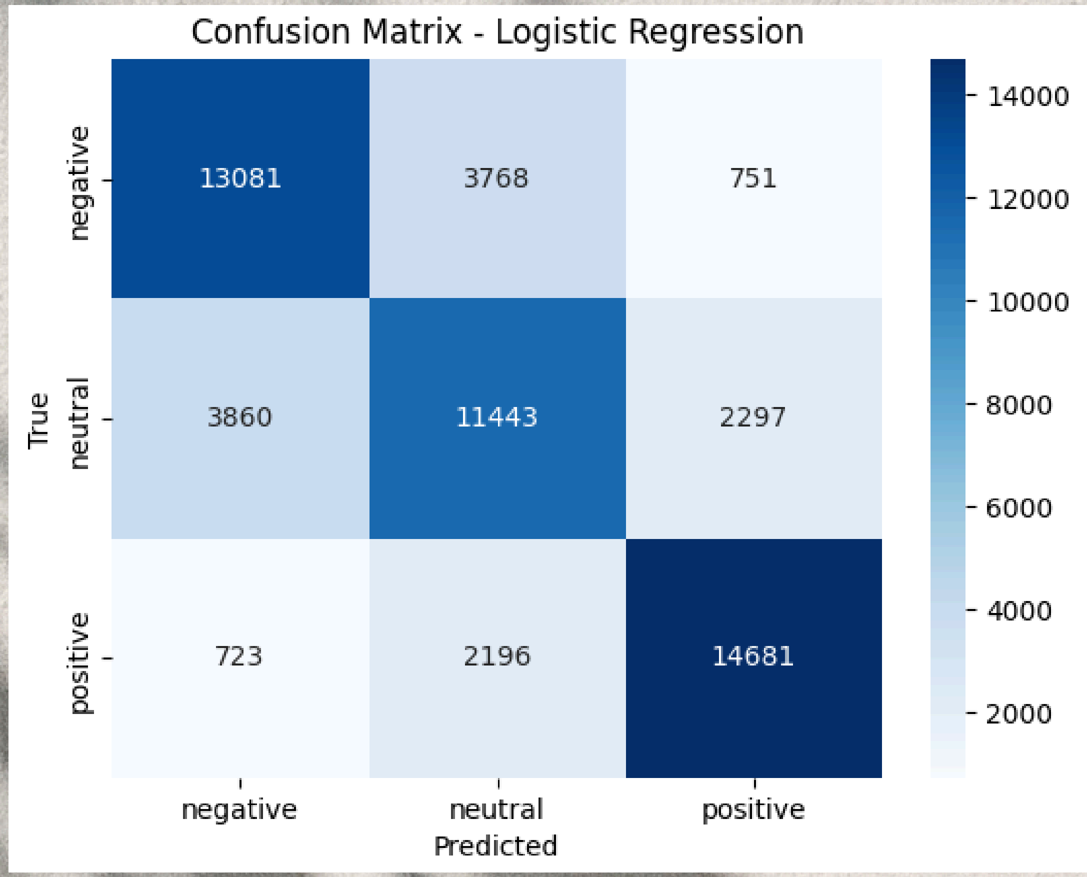
- Trained models included traditional machine learning algorithms like Logistic Regression, Naive Bayes, and SVM, as well as transformer models like BERT and RoBERTa.

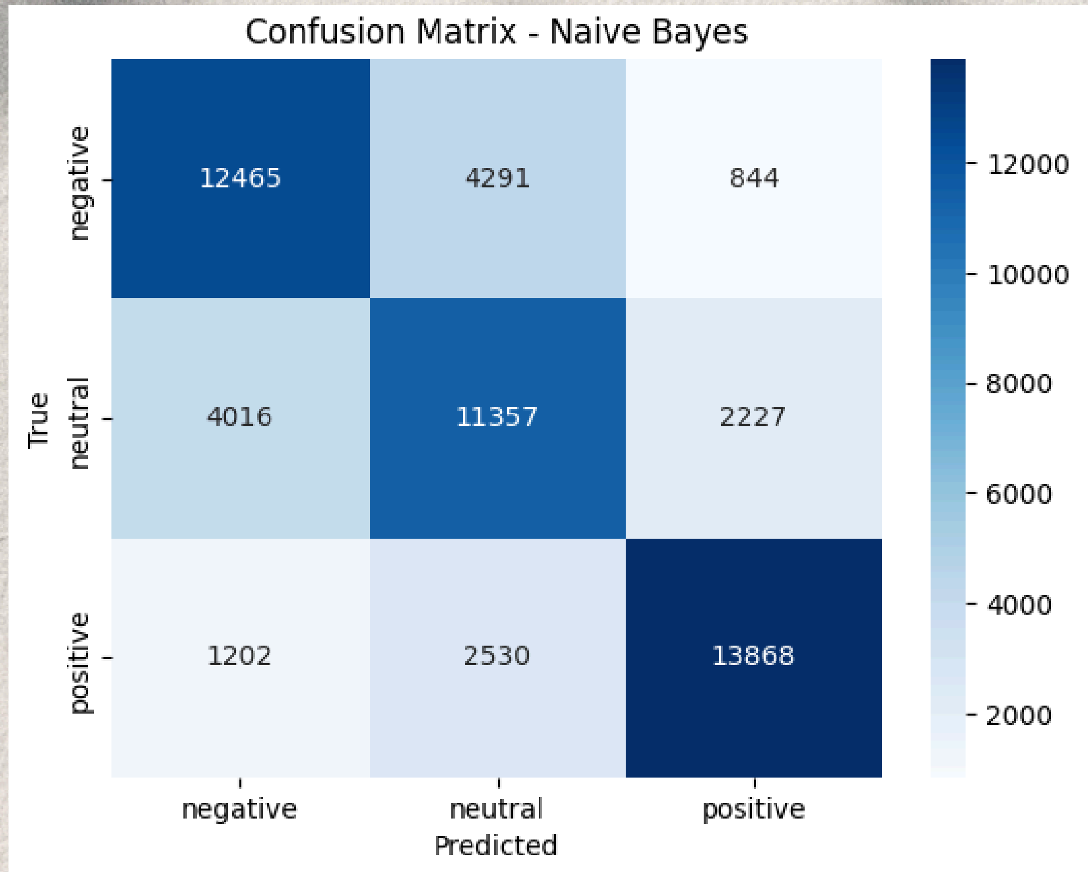
M**odel Evaluation:**

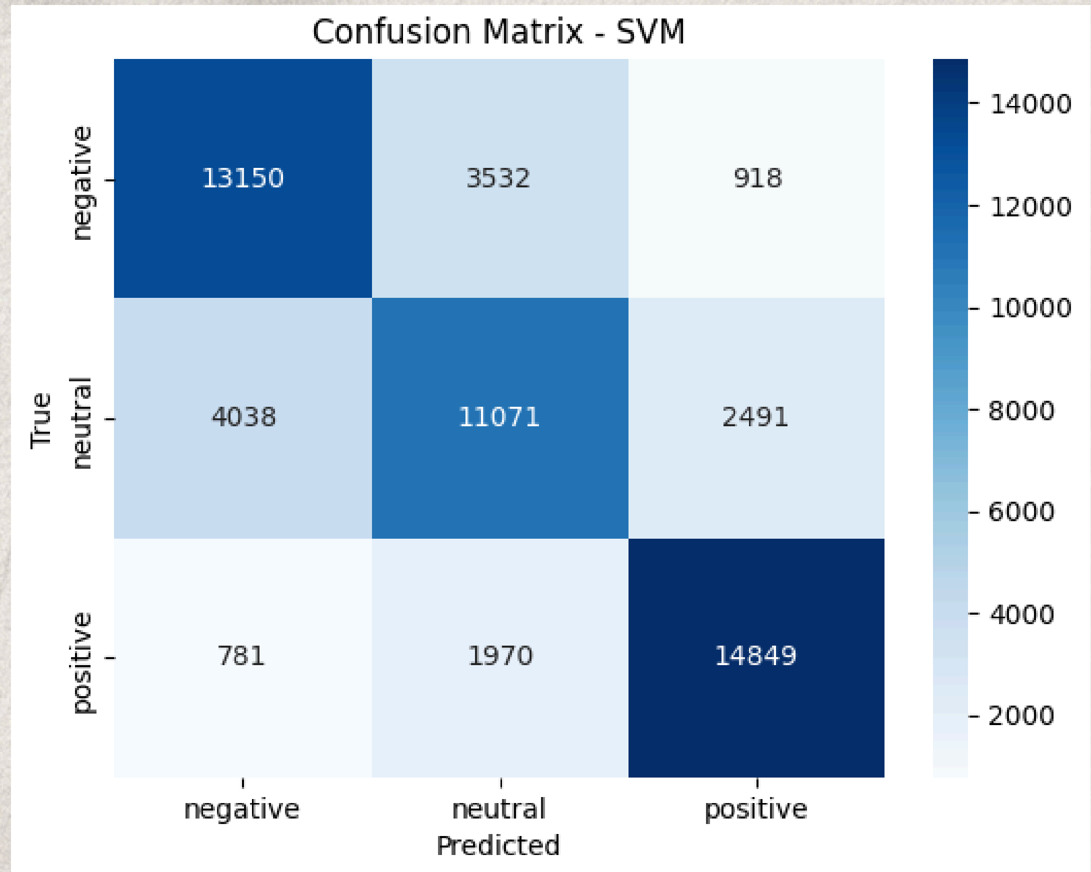
- Model was assessed using metrics such as accuracy, precision, recall, F1 score, and confusion matrices.
- K-fold cross-validation was used to ensure the robustness and stability of the results, providing a comprehensive evaluation of each model's strengths and weaknesses.

Model Evaluation

These models were trained using TF-IDF features extracted from the cleaned product reviews. Their performance was evaluated on metrics like accuracy, precision, recall, F1-score, and confusion matrix.







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

Model Evaluation

These transformer models were trained using a subset of 20,000 reviews from the dataset and showed better performance than traditional models, with BERT achieving the highest accuracy of 79.58%.

[4000/4000 11:52, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.608000	0.598760	0.757000	0.777076	0.757000	0.755111
2	0.392400	0.585076	0.795750	0.799680	0.795750	0.797120

[4000/4000 06:20, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.635100	0.551639	0.766250	0.776920	0.766250	0.769271
2	0.410600	0.592725	0.784000	0.787495	0.784000	0.785402

[4000/4000 13:06, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.738100	0.635902	0.733750	0.768446	0.733750	0.727540
2	0.493300	0.536426	0.788250	0.788801	0.788250	0.788481

Transformer	Accuracy	Precision	Recall	F1-Score	Loss
BERT	0.79575	0.799680	0.79575	0.797120	0.585076
DistilBERT	0.78400	0.787495	0.78400	0.785402	0.592725
RoBERTa	0.78825	0.788801	0.78825	0.788481	0.536426

Opinion Mining

Objective:

- Opinion Mining aims to extract recurring themes or sentiment pairs from product reviews, capturing key customer opinions and insights.

Method:

- Process:
  - Identify common opinion pairs from the reviews (e.g., "good coverage," "too oily").
  - Aggregate these pairs to highlight frequently discussed features and sentiments associated with those features.
  - Analyze these pairs in relation to customer satisfaction to understand product performance or issues

Key Insights:

- Skin Type Concerns: Terms related to skin type like "dry skin", "sensitive skin", and "oily skin" appear across all sentiment categories, indicating they are major factors in the customer experience with Sephora products.
- Positive vs. Negative Sentiment:
  - "Great product" appears frequently in positive reviews, reflecting satisfaction.
  - "Left skin" appears in negative reviews, suggesting dissatisfaction with product results.

Aspect-Based Sentiment Analysis (ABSA)

ABSA offers deeper insights by:

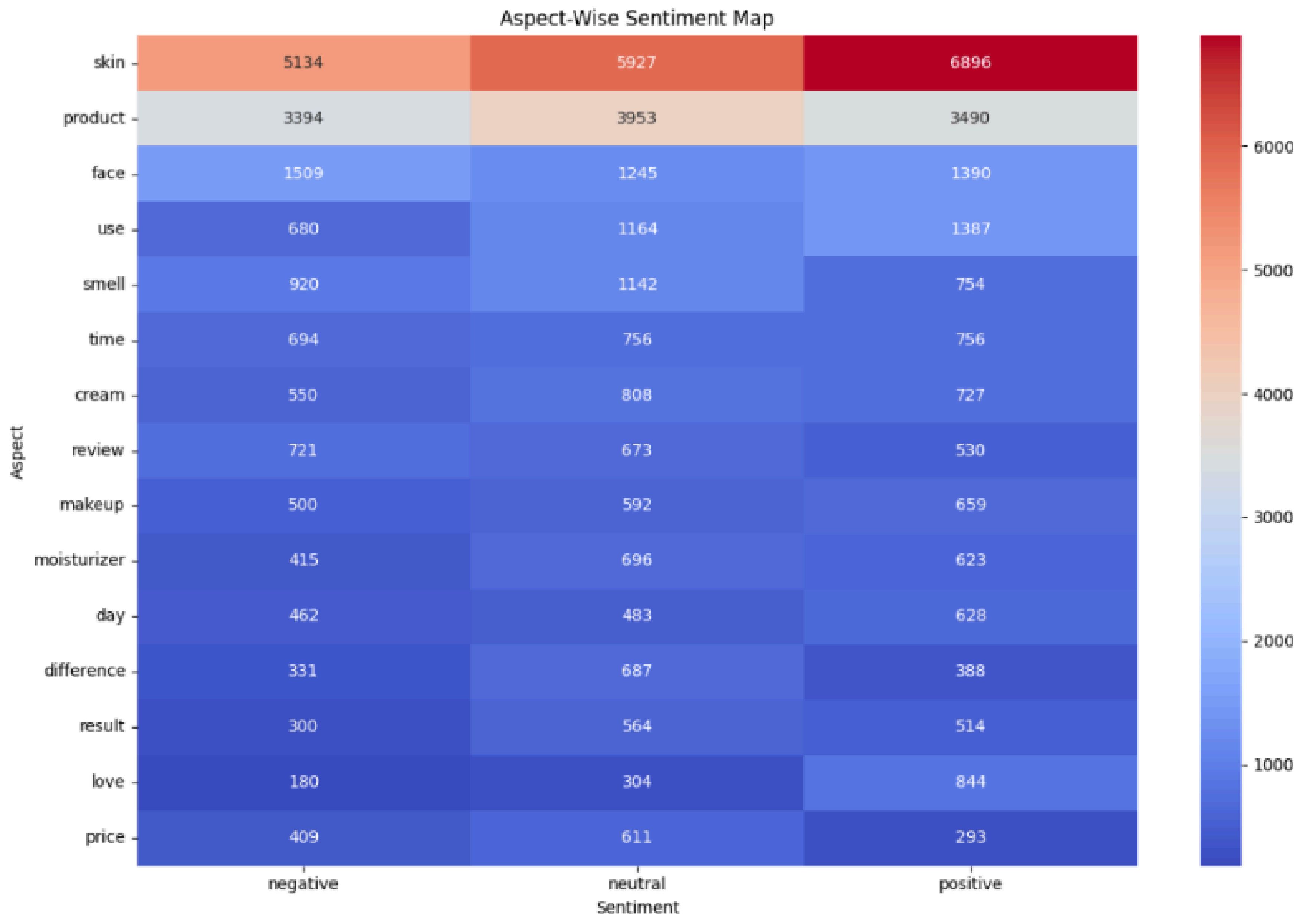
- Identifying specific product features (aspects) mentioned in reviews
- Determining the sentiment (positive, neutral, negative) toward each aspect

How It Was Done:

- Extracted key aspects using the spaCy library
- Analyzed sentiment tied to each aspect

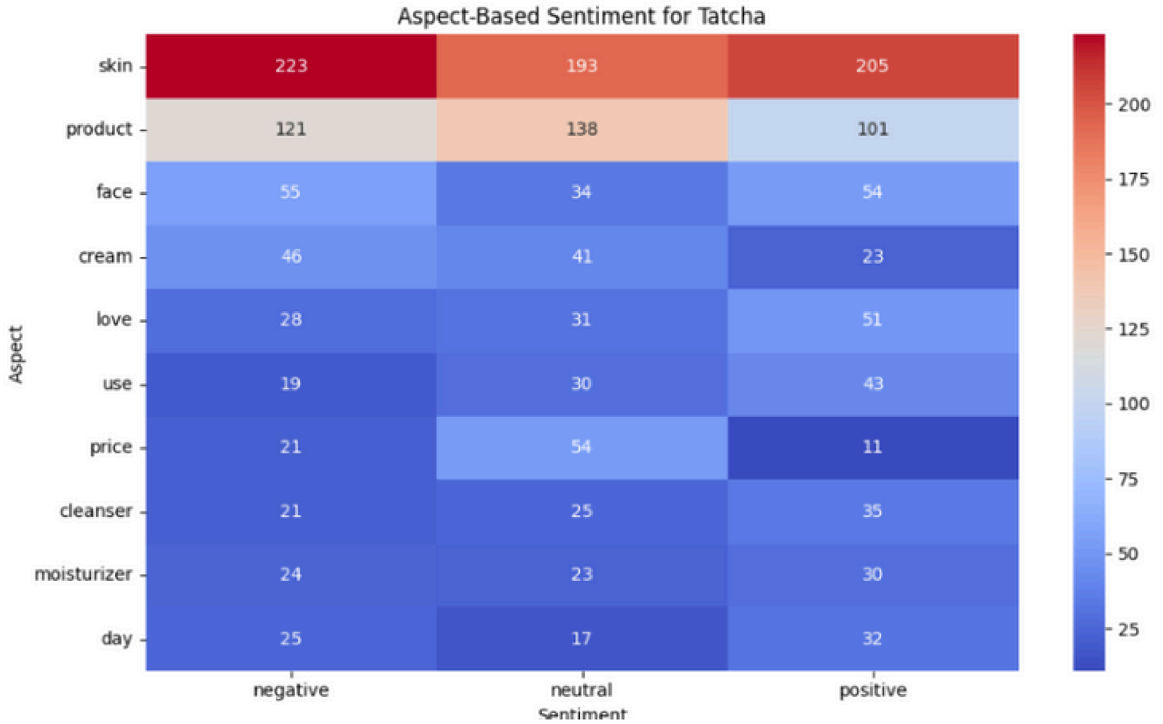
Why It Matter:

- Reveals exactly what customers liked or disliked
- Helps brands enhance products, pricing, and marketing strategies



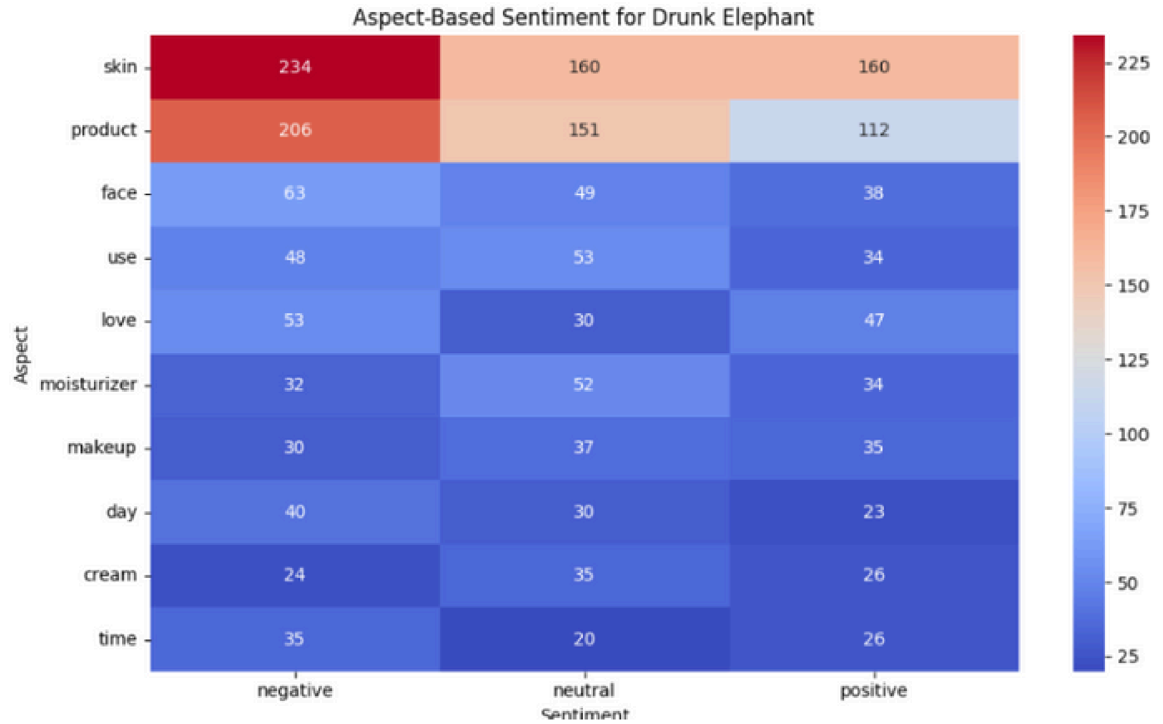
ABSA - Brand Comparison

Tatcha



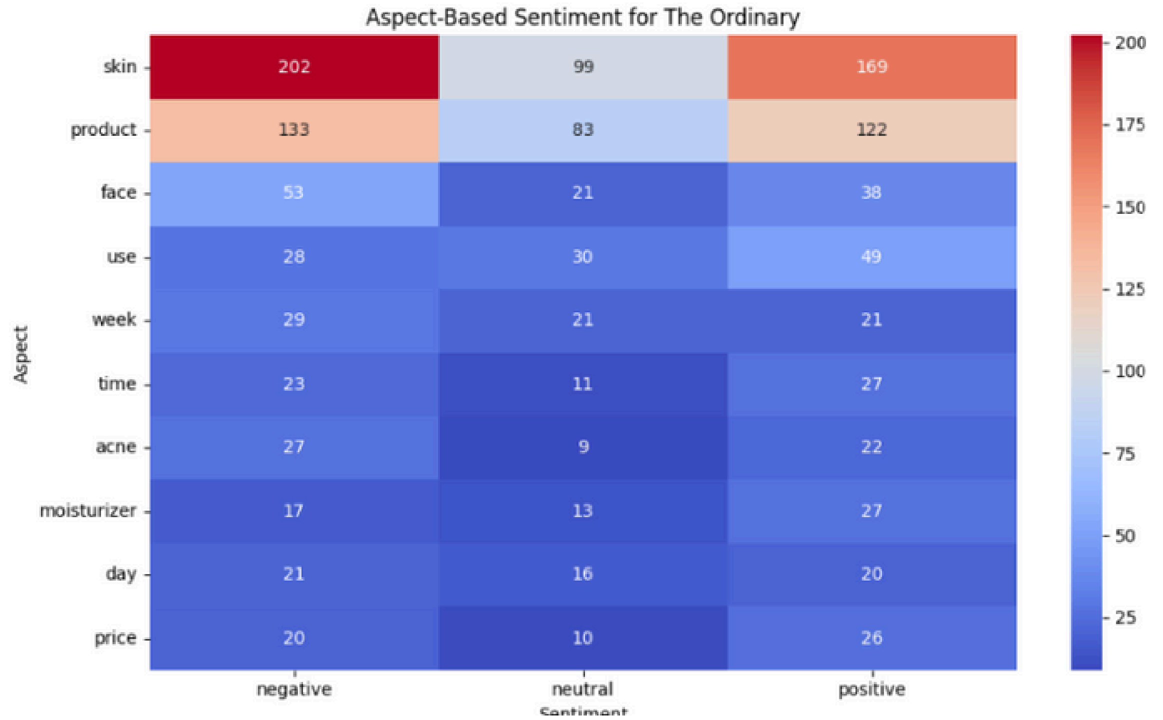
- Most Discussed Aspect: Skin (Balanced sentiment: 205 positive, 193 neutral, 223 negative)
- Product and Face (Mentioned often with a slight tilt toward negative/neutral sentiment)
- Love (Mostly positive sentiment (51 positive vs. 28 negative))
- Price (Frequently mentioned in neutral or negative context & Indicates concerns about high-end pricing)

Drunk Elephant



- Most Mentioned Aspects: Skin and Product (Both have a higher number of negative mentions (234) compared to positives (160)
- Face, Use, and Love (More evenly distributed sentiments)
- Moisturizer and Makeup (Moderately mentioned, with balanced or slightly positive sentiment)
- Day, Cream, Time (Less frequent mentions with mixed sentiment)

The Ordinary



- Key Aspects: Skin and Product (High mentions with polarized sentiment (many positives and negatives)
- Face, Use, Week (Balanced distribution, showing diverse experiences)
- Acne and Moisturizer (Less frequent but mostly positive sentiment)
- Price (Mentioned least, but when mentioned, it's viewed positively & Reflects affordable brand perception)

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
- Hardware limitations restricted the number of epochs and model size that could be tested, particularly for transformer-based models.

Conclusion & Future Work

- Analyzed Sephora reviews using classical ML & transformer models
- Used k-fold validation for consistent results
- BERT achieved highest accuracy (79.58%)
- Identified key aspects: texture, scent, price
- Future improvements:
  - Train on full dataset
  - Add multilingual support
  - Create real-time sentiment dashboard

Thank You