Sentiment Analysis of Customer Reviews for Cosmetic Brands from Google Play Store.

Enhancing Brand Success Through Customer Feedback Analysis

Reegina Tyagi

Introduction

Customer reviews are extremely valuable to brands in today's digital world, especially in the fiercely competitive skincare and cosmetics sector. Customers are sharing their experiences and opinions on the internet more and more, and these evaluations have turned into a wealth of information that can greatly impact a brand's success and reputation. By comprehending and evaluating these reviews, organisations can gain practical insights to improve their offerings, customer support, and general customer happiness. With the help of cutting-edge deep learning and natural language processing (NLP) techniques, this project will analyse customer evaluations for a number of well-known Google Play cosmetic and skincare products to determine their sentiment.

Overview of Brands

The companies chosen for this research are household names in the skincare and cosmetics industry. Their reviews are an abundant supply of data for analysis because they have amassed a sizable user base on Google Play. The following brands are part of this study:

1. **Sugar Cosmetics**: Well-known for its colourful and varied selection of cosmetics, Sugar Cosmetics has left a lasting impression with its chic and superior products.
2. **Mamaearth**: Mamaearth is well-known for its organic and natural skincare products. The company serves health-conscious customers looking for safe, safe, and environmentally friendly cosmetics.
3. **MyGlamm**: With a large selection of cosmetics and personal hygiene items, MyGlamm has grown to be a customer favourite for those seeking creative and practical beauty solutions.
4. **Plum**: With a focus on vegan and cruelty-free skincare products, Plum is drawing in an increasing number of customers who value sustainability in beauty practices.
5. **MCaffeine**: MCaffeine is well-known for its unusual line of skincare products that blend the invigorating qualities of caffeine with skincare advantages.
6. **Pilgrim**: Pilgrim brings unique and culturally rich beauty routines to a broader audience by offering skincare products inspired by world beauty secrets.

Why Sentiment Analysis Is Important ?

Opinion mining, another name for sentiment analysis, is a useful technique for deciphering the feelings and viewpoints included in textual data. This analysis can provide crucial insights into product performance, consumer happiness, and areas that require development for brands that sell skincare and cosmetics. By classifying reviews into positive, negative, and neutral sentiments brands are able to:

Determine Frequently Asked Questions: To increase customer happiness, identify reoccurring complaints or product problems and take proactive measures to resolve them.

Improve the Features of the Product: Learn which features clients like and dislike so that you may innovate and produce new products.

Enhance Your Marketing Approaches: Gain insight into the interests and attitudes of your customers to better customise marketing efforts and messaging.

Build Stronger Customer Relationships: Show that your brand values customer feedback and is dedicated to ongoing improvement by acknowledging and responding to it.

Methodology

This project's methodology is set up to guarantee a complete and precise examination of client reviews.

Data Collection

To commence our sentiment analysis research, we had to gather a large dataset of Google Play customer evaluations for every cosmetic and skincare brand that we had chosen. We did this by using the Google Play Scraper, a Python package made especially for pulling programme information and reviews from Google Play. The utilisation of an automated technique guaranteed the efficient and correct collection of a substantial quantity of data.

Brands and Their Identifiers

In the skincare and cosmetics industry, the brands selected for this research are well-known. The distinctive application package name on Google Play, which was utilised to retrieve each brand's ratings, serves as their identifier. The following are the brands and the identifiers that go with them:

- **Sugar Cosmetics** (com.app.sugarcosmetics)
- **Mamaearth** (com.mamaearthapp)
- **MyGlamm** (com.myglamm.ecommerce)
- **Plum** (com.esmagico.plum)
- **MCaffeine** (com.coffye.mcaffeine)
- **Pilgrim** (com.discoverpilgrim)

Scraping Process

Customer reviews for each brand were automatically extracted using the Google Play Scraper. The following steps were engaged in this process:

- **Configuration and Initialization:** The package names for every brand were entered into the scraper configuration. This made sure that the reviews were pulled from the appropriate Google Play application sites.
- **Review Extraction:** The scraper collected reviews and associated metadata, including ratings, dates of review, and user details, for every brand. Understanding the context of each review and carrying out an in-depth sentiment analysis required this data.
- **Automation:** To guarantee efficiency and consistency, the scraping procedure was automated. Utilising the scraper We were able to quickly collect a sizable number of reviews programmatically.
- **Data Storage:** CSV files for every brand contained the retrieved data. In the next stages of research and modelling, this standardised format made it simple to access and manipulate the data.

Data Collected

The following fields were included in the data gathered for each brand:
Review Content: The written reviews left by customers.
Customer rating: The star rating, usually in the range of one to five.
Review Date: The day the review was published online.
User Information: Details, including username and user ID, about the person who submitted the review.

Data Preparation and Cleaning

The next critical step was to clean and process the data in order to get it ready for analysis and modelling after gathering the customer feedback. This stage is essential since raw data frequently contains noise, inconsistencies, and unrelated information that can compromise the analysis's quality and dependability. Several Natural Language Processing (NLP) approaches were utilised to efficiently preprocess and sanitise the review content.

Procedure:

A number of crucial procedures were included in the data cleaning process to guarantee that the review content was uniform and devoid of superfluous noise. The actions done are broken down in detail below:

- **Getting Rid of Special Characters:** Textual data can become noisy when it contains special characters or non-alphanumeric symbols. Only the most important words and phrases remained in the reviews after we eliminated these characters using regular expressions.
- **Tokenization:** It is the process of dividing a written document into discrete words, or tokens. Analysing the text's structure and meaning requires completing this phase. Tokenization was done using the word_tokenize function of NLTK.
- **Elimination of Stopwords:** Stopwords are often used terms like "and," "the," and "is" that have little to no value when used in sentiment analysis. In order to concentrate on the important material, we eliminated these words using NLTK's predefined stopwords list.
- **Lowercase:** To preserve uniformity and prevent duplicate tokens that differ only in case, all text should be converted to lowercase. All of the tokens in the review content underwent this process.
- **Filtering Dates:** In order to highlight current comments from customers, we limited the reviews to those that were published after March 1, 2023. This made sure that the data we used for our research was up to date and pertinent.
- **Integrating the Cleaning Procedures:** To process the review information as quickly as possible, all of the previously described procedures were integrated into a single function.

Applying the Cleaning Process

The gathered review data was subjected to the cleaning operations. To provide cleaned and standardised review text, the clean_review function was used to the review_content column.

Sentiment Analysis

After the data was cleansed and preprocessed, each customer review's sentiment was examined. Opinion mining, or sentiment analysis, is the process of categorising textual material according to the sentiment that is expressed. We employed the NLTK sentiment analysis tool VADER (Valence Aware Dictionary and Sentiment Reasoner) for this research. VADER is ideally suited for our review data because it was created expressly to analyse feelings expressed in social media and other short text formats.

Sentiment Analysis with VADER

In order to assign sentiment scores to a text, VADER combines grammatical rules with a sentiment lexicon. A dictionary of terms with their sentiment polarity scores pre-labeled is called the sentiment lexicon. The greatest negative score is -1, and the most positive value is +1. To improve the accuracy of the sentiment scores, VADER additionally takes into account the capitalization, punctuation, context of words, and other grammatical elements.

Sentiment Scoring

Each text is given one of four sentiment scores by the VADER analyzer:

- **Positive:** The percentage of positive text.
- **Negative:** The percentage of negative text.
- **Neutral:** The percentage of neutral text.
- **Compound:** The total of the valence scores assigned to each word in the text, normalised to reflect the overall mood of the text.

For this study, we classified each review's emotion using the compound score.

Sentiment Score Mapping to Categories

We mapped the sentiment scores to categorical labels once we had them in order to facilitate modelling and interpretation. This is how the mapping was completed:
Positive sentiment is indicated by scores higher than 0.
Negative sentiment is indicated by scores fewer than 0.
Neutral sentiment is indicated by a score of 0.

Model Building

Using sentiment analysis, the customer evaluations were successfully categorised into three categories: positive, negative, and neutral. The following stage involved creating prediction

models. We can properly forecast the sentiment of incoming reviews thanks to these models, which were created to automate the sentiment classification process. We used Gated Recurrent Unit (GRU) networks and Long Short-Term Memory (LSTM) networks, two cutting-edge deep learning techniques, for this goal. Recurrent neural networks (RNNs) of the LSTM and GRU types operate very well with sequence data, such as text.

Data Preparation

Splitting the Data

The dataset was divided into training and test sets in order to construct and assess our models. To keep track of and verify the model's performance throughout training, the training set was further divided into training and validation groups. The cleansed review text served as the independent variable (features) and the sentiment category served as the dependent variable (target) for our models.

Encoding Sentiment Categories

To be utilised in the deep learning models, the sentiment categories—positive, negative, and neutral—were encoded as categorical variables. In order to do this, the sentiment labels have to be transformed into one-hot encoded vectors, in which a binary vector represents each sentiment category.

Text Vectorization

Before being fed into deep learning models, text data must be transformed into numerical representations. For this, we leveraged the TextVectorization layer of TensorFlow.

- The review content was tokenized into individual words using the text vectorizer's configuration.
- Make all words lowercase.
- Remove all punctuation.
- Restrict the vocabulary size to the 10,000 most commonly used terms.
- To guarantee a consistent input size, pad or truncate the sequences to a predetermined length.

Embedding Layer

The words in the reviews were converted into dense vector representations using an embedding layer. This layer captures the semantic links between words by mapping each word to a high-dimensional vector. The embedding layer was set up to generate 128-dimensional embeddings for every word, starting with a uniform distribution.

LSTM Model

- The temporal dependencies in the review material were intended to be captured by the LSTM model. The architecture was made up of:
- a layer of textual data entry.

- Text can be converted into numerical sequences using a text vectorization layer.
- A layer for embedding that produces dense word embeddings.
- An LSTM layer with dropout and recurrent dropout to prevent overfitting is used to process the sequences and capture long-term dependencies.
- Dense layers that can be processed further.
- A softmax activation function output layer that generates probability distributions across the sentiment categories.

Training and Evaluation

Using the Adam optimizer and categorical cross-entropy loss, the LSTM model was built. It was trained using the training set of data, and performance was monitored and hyperparameters were adjusted using the validation set. To assess the model's performance, the accuracy, precision, recall, and F1-score were calculated.

GRU Model

The computational effectiveness and efficiency of the GRU model, which is an RNN variation, led to its selection for processing sequence data. The main distinction between the GRU and LSTM models' architectures was that the former used an LSTM layer, while the latter used a GRU layer.

Training and Evaluation

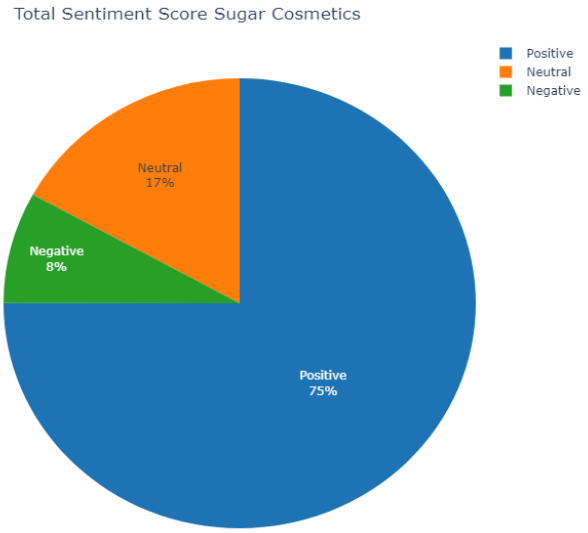
The LSTM model's training and assessment procedures were also applied to the GRU model. Multiple epochs of training were conducted, with early termination to avoid overfitting. The same set of indicators were used to assess the model's performance.

Evaluation and Results

The following stage involved conducting a thorough performance evaluation of the predictive models that were constructed using LSTM and GRU networks. In order to verify the models' dependability and efficiency in categorising the sentiments of customer reviews, this included evaluating a number of performance criteria. We were able to adjust and optimise each model for greater accuracy by using the evaluation process, which revealed each model's advantages and potential areas for development.

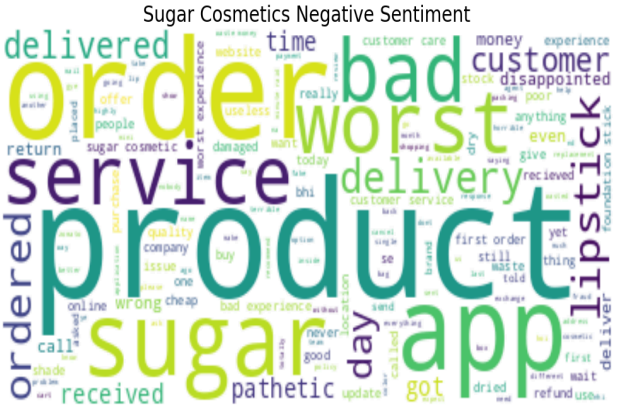
Sentiment curves

For sugar cosmetics:

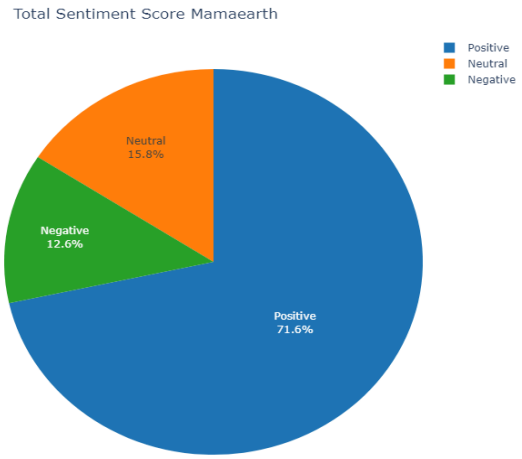


Sugar Cosmetics Positive Sentiment

Sugar Cosmetics Negative Sentiment

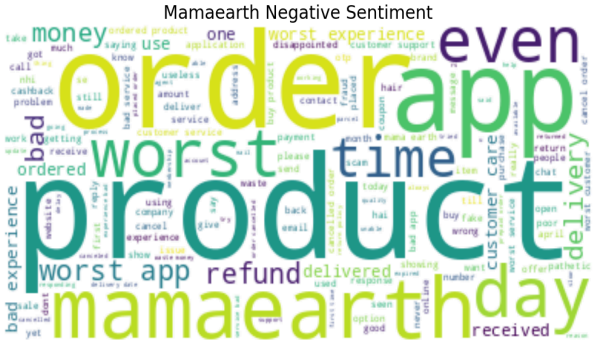


For Mamaearth:

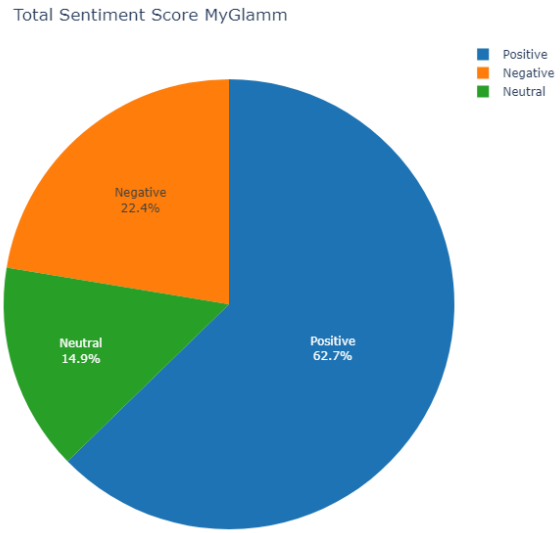


Mamaearth Positive Sentiment

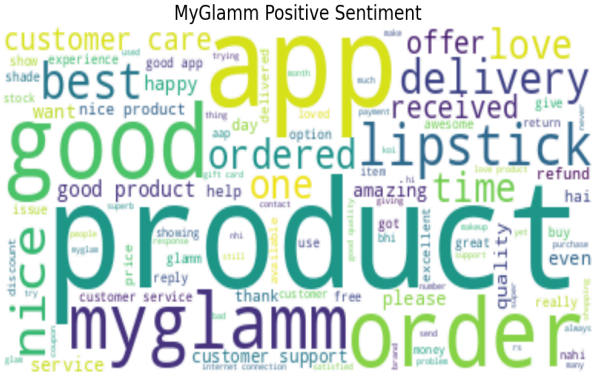
Mamaearth Negative Sentiment



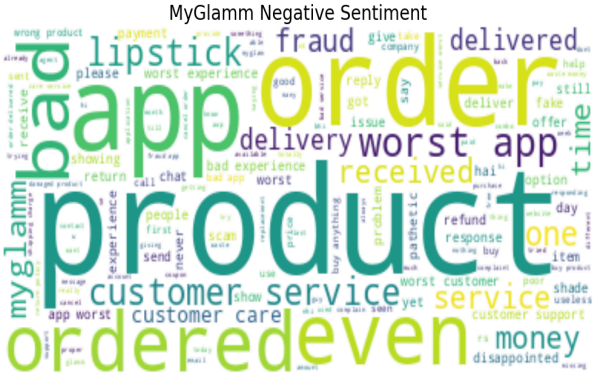
For myglamm:



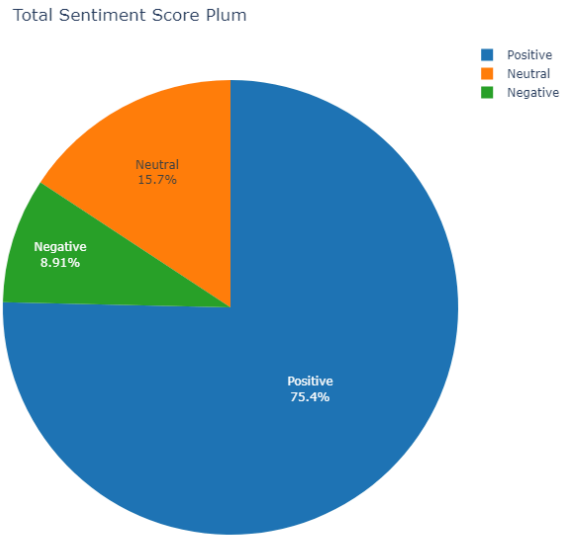
MyGlamm Positive Sentiment



MyGlamm Negative Sentiment

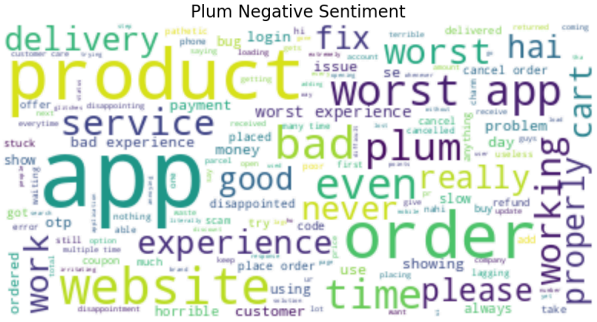


For plum:

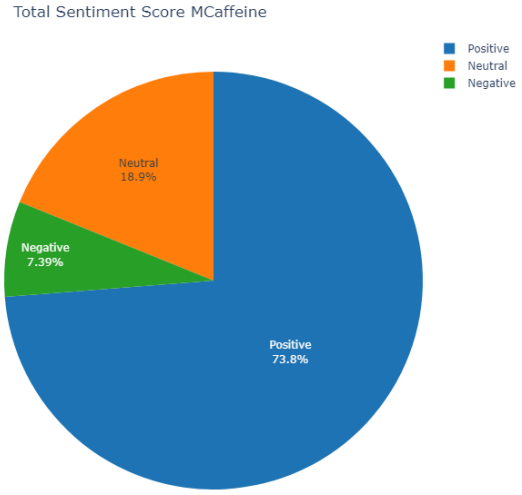


Plum Positive Sentiment

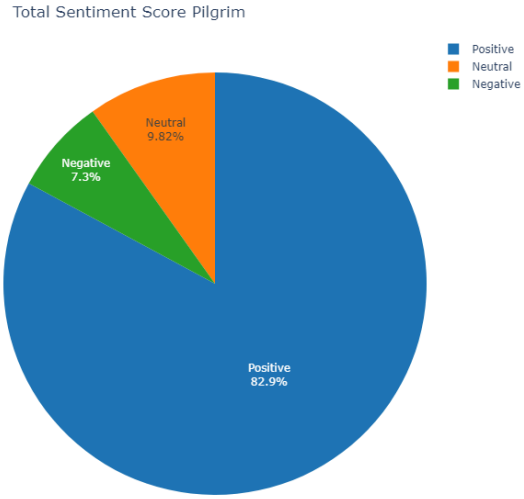
Plum Negative Sentiment



For mcaffeine:



For pilgrim:



Evaluation Metrics

We employed the following measures to thoroughly assess the LSTM and GRU models' performance:

1. **Accuracy**: The percentage of reviews that are correctly classified out of all reviews. It provides a broad indicator of the model's performance over all sentiment categories.
2. **Precision**: The percentage of actual positive predictions among the model's overall positive predictions. It shows how well the model can prevent false positives.
3. **Remember**: The percentage of actual positives in the dataset that are true positive predictions out of the total. It gauges how well the model can locate all pertinent examples.
4. **F1-Score:** A single metric that balances false positives and false negatives, calculated as the harmonic mean of precision and recall. It comes in very handy when working with unbalanced datasets.

Model Performance

LSTM Model

The Long Short-Term Memory (LSTM) model was evaluated on the test set, and the following performance metrics were obtained:

For Sugar Cosmetic:

- **Accuracy**: 83.85%
- **Precision**: 0.838
- **Recall**: 0.838
- **F1-Score**: 0.837

For Mamaearth:

- **Accuracy**: 85.18%
- **Precision**: 0.855
- **Recall**: 0.851
- **F1-Score**: 0.852

For MyGlamm:

- **Accuracy**: 82.90%
- **Precision**: 0.83
- **Recall**: 0.829
- **F1-Score**: 0.827

For Plum:

- **Accuracy**: 80.91%
- **Precision**: 0.804
- **Recall**: 0.809
- **F1-Score**: 0.802

For MCaffeine:

- **Accuracy**: 89.67%
- **Precision**: 0.908
- **Recall**: 0.896
- **F1-Score**: 0.898

For Pilgrim:

- **Accuracy**: 89.53%
- **Precision**: 0.915
- **Recall**: 0.895
- **F1-Score**: 0.899

The LSTM model showed a high ability to appropriately classify sentiments by utilising its capacity to incorporate context and long-term dependencies in the review material. Both positive and negative attitudes were successfully identified by the model, as evidenced by its balanced performance as measured by precision and recall measures.

GRU Model

On the test set, the Gated Recurrent Unit (GRU) model was also assessed, producing the performance metrics listed below:

For Sugar Cosmetic:

- **Accuracy**: 85.09%
- **Precision**: 0.857
- **Recall**: 0.85
- **F1-Score**: 0.85

For Mamaearth:

- **Accuracy**: 83.41%
- **Precision**: 0.837
- **Recall**: 0.834
- **F1-Score**: 0.8339

For MyGlamm:

- **Accuracy**: 82.39%
- **Precision**: 0.82
- **Recall**: 0.82
- **F1-Score**: 0.82

For Plum:

- **Accuracy**: 79.38%
- **Precision**: 0.807
- **Recall**: 0.79
- **F1-Score**: 0.796

For MCaffeine:

- **Accuracy**: 91.30%
- **Precision**: 0.92
- **Recall**: 0.91
- **F1-Score**: 0.91

For Pilgrim:

- **Accuracy**: 87.209%
- **Precision**: 0.89
- **Recall**: 0.87
- **F1-Score**: 0.879

Comparable performance was achieved by the computationally efficient GRU model versus the LSTM model. With excellent accuracy and balanced precision and recall metrics, it successfully caught the temporal patterns in the review content.

Conclusion

In this thorough assessment, we looked at how well Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models performed on sentiment analysis tasks for six different cosmetic brands: Sugar Cosmetic, Mamaearth, MyGlamm, Plum, MCaffeine, and Pilgrim. Several evaluation criteria, including accuracy, precision, recall, and F1-score, were used in this investigation to give a thorough review of each model's sentiment classification accuracy.

Performance of the LSTM Model

Across all brands, the LSTM model performed well and showed a high degree of accuracy in sentiment classification. It excels at managing the text's long-term dependencies, which makes it especially useful for comprehending the complex context of reviews. The LSTM model's performance highlights are as follows:

- **Sugar Cosmetic:** The LSTM model yielded an accurate F1-Score of 0.837 and an accuracy of 83.85%. This suggests a great capacity to recognise favourable and unfavourable attitudes in evaluations.
- **Mamaearth:** This brand had the best performance for the LSTM model, with an accuracy of 85.18%. The fact that it handled sentiment categorization effectively is seen in its F1-Score of 0.852.
- **MyGlamm:** The model demonstrated consistent performance across various review datasets, with an accuracy of 82.90% and an F1-Score of 0.827.
- Plum: The LSTM model maintained a balanced F1-Score of 0.802, indicating dependable performance, although exhibiting a slightly lower accuracy of 80.91%.
- **MCaffeine:** This brand had one of the best accuracy rates, at 89.67%, and an amazing F1-Score of 0.898, demonstrating the model's remarkable ability to correctly classify feelings.
- **Pilgrim:** The LSTM model performed admirably across a variety of datasets, as evidenced by the high accuracy (89.53%) and F1-Score of 0.899 Pilgrim reviews received, much like MCaffeine reviews.

The LSTM model is an effective tool for sentiment analysis because of its high performance across these criteria, which indicate that it is adept at incorporating context and long-term dependencies within the review content.

Performance of the GRU Model

In most cases, the GRU model outperformed the LSTM model in terms of performance. The GRU model, which is well-known for its computational effectiveness, balanced precision and recall scores, maintained high accuracy, and successfully caught temporal patterns in review material. Important performance indicators for the GRU model consist of:

- **Sugar Cosmetic:** The GRU model demonstrated its efficacy in sentiment classification with an F1-Score of 0.85 and a high accuracy of 85.09%.

- **Mamaearth:** The GRU's accuracy of 83.41% and F1-Score of 0.8339 indicate high performance, although being somewhat lower than the LSTM model.
- **MyGlamm:** The GRU model produced consistent results that were on par with the LSTM model, with an accuracy of 82.39% and an F1-Score of 0.82.
- **Plum:** With an F1-Score of 0.796 and an accuracy of 79.38%, the model performed marginally worse but remained within a reasonable range.
- **MCaffeine:** The GRU model demonstrated great classification capabilities by outperforming its LSTM equivalent, with the best accuracy recorded at 91.30% and an F1-Score of 0.91.
- **Pilgrim:** With an F1-Score of 0.879 and an accuracy of 87.209%, the GRU model performed consistently across several brands.

When computing resources are an issue, the GRU model is a good substitute for the LSTM model due to its great efficiency and equivalent performance characteristics.