IJoCED

Indonesian Journal of Computing, Engineering, and Design

Journal homepage: http://ojs.sampoernauniversity.ac.id/index.php/IJOCED

# Sentiment Analysis of Multi-Brand Skincare Product Reviews Using IndoBERT Fine-Tuning

*Irsalina Yumna*[1], *Alam Rahmatulloh*[2*]

[1,2] Department of Informatics, Faculty of Engineering, Siliwangi University, Tasikmalaya, 46115, Indonesia

Corresponding email: alam@unsil.ac.id

## ABSTRACT

The Female Daily platform predominantly features consumer reviews characterized by informal Indonesian and English code-mixing. This specific linguistic complexity presents a significant impediment to traditional classification methods, such as classical machine learning algorithms (e.g., Naive Bayes and SVM) and lexicon-based approaches, which often fail to accurately capture semantic nuances and contextual dependencies in unstructured text. To fill this research gap, this study uses a deep learning method with fine-tuned IndoBERT for multi-brand sentiment analysis. Using a dataset of 12,418 reviews across five popular skincare brands, the model achieved an accuracy of 85%, an F1-score of 0.84, a precision of 0.85, and a recall of 0.84. A key contribution of this research is the multi-brand analysis, which reveals distinct consumer perception patterns: Wardah and Emina achieved the highest proportions of positive sentiment, while Skintific and Garnier demonstrated a more balanced distribution between positive and negative reviews. In contrast, MS Glow exhibited more varied and diverse consumer opinions. These findings confirm that IndoBERT's self-attention mechanism is highly effective and adaptive in processing the informal, code-mixed vocabulary of the beauty community, outperforming traditional methodologies in both robustness and contextual understanding.

## ARTICLE INFO

***Article History:***
*Received 01 Dec 2025*
*Revised 09 Mar 2026*
*Accepted 25 Mar 2026*
*Available online 24 Apr 2026*

***Keywords:***
*Code-mixed,*
*Consumer Opinion,*
*Deep Learning,*
*Digital Platform,*
*Female Daily,*
*Transfer Learning.*

## 1. INTRODUCTION

The beauty industry in Indonesia is currently experiencing rapid growth (Wildana and Kautsar, 2025). This phenomenon is driven not only by increasing public awareness of the importance of skin care but also by the massive penetration of digital technology and social media, which facilitates access to product information (Fadilah, Muflikhah, and Perdana, 2025). Changes in consumer behavior, who are increasingly critical and actively seeking references

before purchasing products, have also triggered the development of various online review platforms (Habibi and Kusumaningtyas, 2023). One of the largest review platforms shaping consumer preferences is Female Daily (Yutika, Adiwijaya, and Faraby, 2021), which provides thousands of reviews on various popular skincare brands in Indonesia. This platform reflects consumers' honest and real experiences in assessing the quality, effectiveness, and level of satisfaction with the products they use. The existence of these reviews makes Female Daily not just a product catalog but a data ecosystem rich in consumer behavior insights (Hidayat and Handayani, 2023).

The large volume of consumer reviews makes platforms such as Female Daily a valuable source of opinion data, both for understanding public perception, evaluating brand image, and formulating data-driven marketing strategies (Dheanis et al., 2021). However, with high user activity, the volume of reviews generated is enormous, and the data is unstructured. The diverse nature of text data makes manual analysis inefficient, time-consuming, and prone to human subjectivity bias (Mutinda, Mwangi, and Okeyo, 2023). Therefore, a computational approach is needed that can automatically extract information and emotions contained in review texts to produce fast and accurate analysis (Mughal et al., 2024).

Various approaches have been used in sentiment analysis to understand user opinion trends, ranging from lexicon-based methods to classic machine learning algorithms such as Support Vector Machine (Jasmarizal et al., 2024) and Naïve Bayes (Puspita et al., 2024). Although these methods provide fairly good results on formal texts, their performance tends to decline on Indonesian review texts, which often use a mixture of Indonesian and English, or code-mixed text, due to the unique linguistic features and informal expressions commonly found in such reviews. The emergence of modern deep learning architectures such as the Transformer model has brought about major changes in the field of Natural Language Processing, particularly through the Bidirectional Encoder Representations from Transformers (BERT) model, which is capable of understanding two-way semantic context in sentences. Research (Jayadianti et al., 2022) shows the effectiveness of IndoBERT in improving the accuracy of sentiment analysis in various domains, such as e-commerce product reviews (Huang, Zavareh, and Mustafa, 2023) and hotel reviews (Ounacer et al., 2023).

However, existing research generally focuses on a single brand or a specific type of product, and not much research has examined multi-brand sentiment analysis in the skincare domain. In fact, variations in opinion between brands can provide a comprehensive picture of consumer preferences in the Indonesian beauty market (Mei, Tiun, and Sastria, 2024). Based on research (Patwardhan, Marrone, and Sansone, 2023), emphasize that BERT is more versatile and effective than traditional recurrent or convolutional models because its pre-training approach allows it to learn rich contextual representations of words while utilizing parallel processing for greater efficiency.

This study proposes the implementation of IndoBERT fine-tuning method to classify sentiment in multi-brand skincare reviews from the Female Daily platform, following the established effectiveness of encoder-only transformer architectures for discriminative tasks. This method takes advantage of BERT's ability to learn rich contextual representations of words, which is important for better capturing the subtleties of mixed Indonesian and English text than

traditional models. By doing so, this research aims to produce a highly accurate and adaptive sentiment analysis model that offers in-depth insights into consumer perceptions across the Indonesian skincare industry.

## 2. RESEARCH METHODOLOGY

The research method was structured to develop a sentiment analysis model capable of classifying multi-brand skincare reviews through the IndoBERT fine-tuning approach (Mughal et al. 2024). The workflow consisted of sequential stages ranging from data collection to model evaluation. A Transformer-based NLP technique was employed to categorize consumer opinions into positive and negative sentiments (Prottasha et al. 2022). The overall research flow is presented in **Figure 1**.

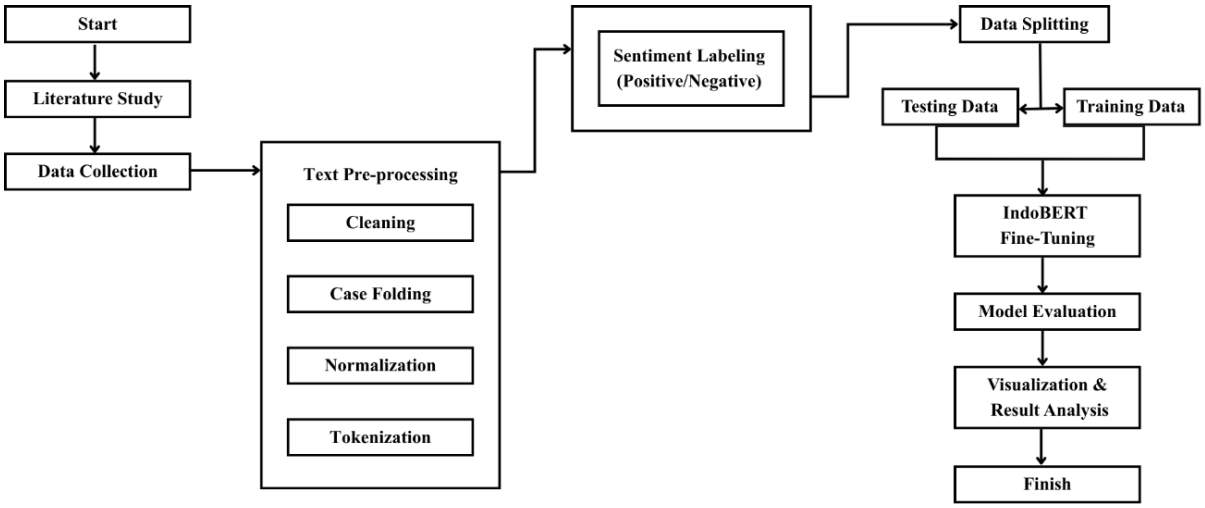


**Figure 1. Research Flow Diagram**

### 2.1. Literature Review and Data Collection

This research began with a literature study of various references related to sentiment analysis, natural language processing, and the implementation of the IndoBERT-based Transformer model (Jayadianti et al., 2022). The literature study was conducted to understand the methods, techniques, and current trends in the application of language models to Indonesian-language data and Indonesian-English code-mixed (Astuti et al., 2023).

Next, data collection was carried out using web scraping techniques from the Female Daily platform (Al Aufar and Romadhony, 2025). This platform was chosen because it has a large volume of reviews, an informal writing style, and a dominance of code-mixed Indonesian and English texts that are relevant for testing the capabilities of IndoBERT(Widya Astuti and Sari, 2023). The collected data covered five basic product categories, namely facial wash, toner, serum, moisturizer, and sunscreen, each from five popular brands. Each review contained attributes such as product name, brand, rating, review content, and upload date (Anandia, Purnomo, and Setiawan, 2025).

### 2.2. Text Pre-processing

Text preprocessing was performed to clean and standardize the review text before further analysis (Yutika, Adiwijaya, and Faraby, 2021). This stage included several main steps, which are outlined as follows.

*a. Cleaning*

The cleaning stage is carried out to remove various irrelevant elements that

have the potential to cause noise in the analysis process (Wildana and Kautsar, 2025). This process includes removing URLs, mentions, hashtags, numbers, emojis, non-alphabetic characters, and excessive punctuation marks; performing reduction of repeated characters in words, for example, *'bangettt*' becomes '*banget*'; and removing meaningless words.

*b. Case Folding*

All text is converted to lowercase to standardize the writing format and avoid unnecessary variations in meaning due to differences in capitalization. This step is taken because user reviews are often written informally and inconsistently in terms of capitalization (Hidayat and Handayani, 2023).

*c. Normalization*

Normalization is performed using a slang dictionary that has been compiled and loaded into the program (Fadilah, Muflikhah, and Perdana, 2025). Non-standard words, abbreviations, and slang terms such as "*bgt*", "*gimana*", or "*mukaku oily*" are replaced with standard or more common forms. In addition, normalization also includes the removal of words that are too short, with 1–3 characters, and do not contribute to the sentiment context.

*d. Tokenization*

The tokenization stage is carried out to break down the text into tokens using a word tokenizer that produces a list of tokens for the normalized reviews. This tokenization ensures that the text is in a structure suitable for sentiment labeling and IndoBERT model training (Singh et al., 2025).

**2.3. Sentiment Labeling**

Automatic labeling is performed using the pre-trained model mdhugol/indonesia-bert-sentimentclassification from Hugging Face. This model automatically classifies reviews into two polarity classes, namely positive (1) and negative (0) (Prottasha et al. 2022). Each classification result is accompanied by a confidence score to indicate the model's level of confidence. Reviews with confidence scores below the threshold of 0.75 are eliminated to maintain label quality. The labeled dataset is then saved for use in the model training stage.

**2.4. Splitting Data**

The labeled dataset is divided into three parts: 70% training data, 20% validation data, and 10% test data. This division aims to ensure that the model can be tested on data different from its training data so that the results obtained are objective and do not experience overfitting (Abdullah, Hassan, and Mustafa, 2024).

**2.5. Fine-Tuning the IndoBERT Model**

The indobenchmark/indobert-base-p1 model is used as the base model. The training process is carried out using a batch size of 16, a learning rate of 2e-5, and three epochs, with the AdamW optimizer and cross-entropy loss function. These parameters were selected based on stable values for the BERT model so that the training process was consistent (Prottasha et al., 2022). This training used PyTorch and the Transformers library, with monitoring of loss and accuracy values at each epoch.

**2.6. Model Evaluation**

Model performance evaluation was carried out using four main metrics, namely Accuracy, Precision, Recall, and F1-score (Manoppo et al., 2025). In addition, a Confusion Matrix was used to analyze the distribution of model predictions for positive and negative classes. All of these metrics were used to provide a comprehensive overview of the model's performance after the training process was complete. This evaluation approach ensures that the model's performance is

not only assessed based on the overall prediction accuracy but also on the balance between positive and negative errors produced (Imaduddin, Yusfida A'la, and Nugroho, 2023).

### 2.7. Visualization and Analysis of Results

Data visualization is used to strengthen understanding of the characteristics of multi-brand skincare reviews, ranging from reviewer activity patterns and the dominance of popular products to the trends of words that frequently appear in the review corpus. These various exploratory graphs form the basis for reading consumer opinion dynamics before entering the modeling stage. In addition, the Confusion Matrix is used to show the ability of the IndoBERT model to clearly recognize sentiment polarity so that prediction quality can be evaluated not only numerically but also through the distribution patterns of errors and classification successes (Jasmarizal et al., 2024).

## 3. RESULTS AND DISCUSSION

The analysis results begin with examining the basic patterns in the dataset through the Exploratory Data Analysis (EDA) stage to understand the review patterns and language characteristics in the dataset. These exploration results form the basis for understanding the data context before transformation through the text pre-processing stage, which aims to improve text quality so that it is ready for use in the labeling and model training processes (Habibi and Kusumaningtyas, 2023). Once the data is clean, the next step involves automatic sentiment labeling, IndoBERT model training, and performance evaluation to assess the model's ability to recognize sentiment polarity in skincare reviews.

### 3.1. Reviewer Activity Analysis

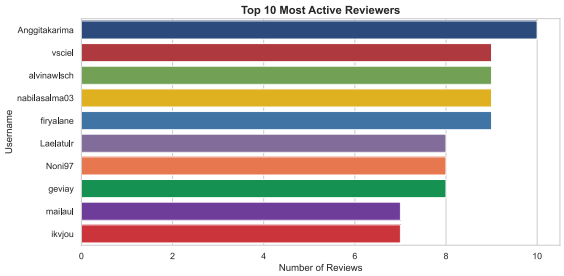


**Figure 2. Most Active Reviews**

Based on **Figure 2**, we can see the distribution of user activity, where the account named Anggitakarima is the most active contributor, followed by vsciel and alvinawlsch. The top ten contributors write an average of between 7 and 10 reviews. The existence of this group of top reviewers adds value to the quality of the research dataset. Active users tend to write reviews that are more descriptive and detailed, and use specific beauty terminology compared to casual users. This group enriches the variety of the data corpus that will be studied by the IndoBERT model, so that the model can better recognize the language patterns of the beauty community.

### 3.2. Reviewer Activity Analysis

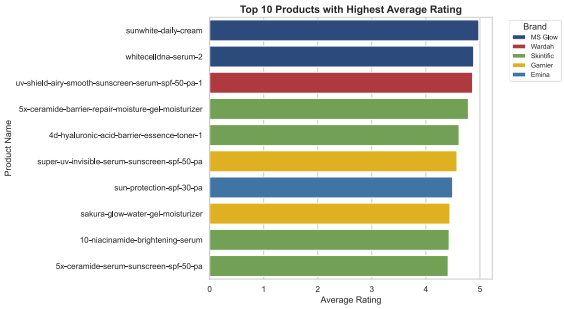


**Figure 3. Products with the Highest Average Rating**

**Figure 3** shows the ten products with the highest average ratings, with MS Glow's Sunwhite Daily Cream and Wardah's UV Shield occupying the top positions. The dominance of high ratings is held by basic skincare products that are protective and repair the skin barrier, such as sunscreen, moisturizer, and serum. This phenomenon indicates that Indonesian consumers currently prioritize basic

functionality and skin protection in their routines. The high level of satisfaction with local brands such as MS Glow and Wardah also reflects strong consumer confidence in the effectiveness of domestic product formulations that are able to compete with global products.

### 3.3. Analysis of Products with the Most Reviews

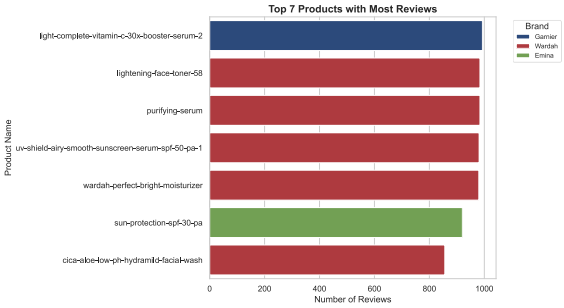


**Figure 4. Products with the Most Reviews**

As shown in **Figure 4**, Garnier's Light Complete Vitamin C Serum and Wardah's product line are the most reviewed items. The high volume of reviews on these products indicates a high level of market penetration and popularity among Female Daily users. In sentiment analysis, the large number of reviews on these popular products is very beneficial because it provides a wide range of opinions. The review data covers a complete spectrum of sentiments, from complaints about side effects or breakouts to praise for product texture, which is very useful for training the model to be adaptive to various sentence contexts.

### 3.4. Word Frequency Analysis in Reviews

The word frequency analysis in **Figure 5** shows that the words "this," "I," and "and" are the three most frequently occurring words in user reviews. In addition, words such as "skin," "really," "use," and "make" are also frequently used. This pattern indicates that the majority of users write reviews in a personal and subjective manner, often recounting their experience of using the product based on their own skin condition. This finding also reinforces the indication that the review text has an informal style and is a mixture of Indonesian and English terms.

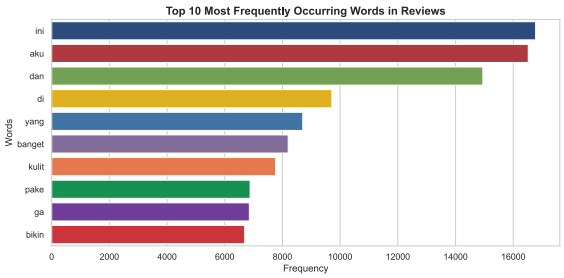


**Figure 5. Frequently Occurring Words**

### 3.5. Text Pre-processing Results

The pre-processing stage is carried out to ensure data quality before entering the labeling and model training processes. This section shows the changes that occur at each step, from cleaning, case folding, and normalization to tokenization, to demonstrate how text quality improves after going through all stages of preprocessing.

a. Cleaning

The cleaning stage removes irrelevant elements such as URLs, emojis, numbers, excessive punctuation, mentions, and meaningless words. This process produces cleaner text that is ready for processing in the next stage. The results of the cleaning process can be seen in **Table 1**, which illustrates the transformation of raw review text into a cleaner and more structured form.

**Table 1. Demographic Profile of Surveyed Respondents**

| no | Review_text | clean_review |
|---|---|---|
| 0 | Kulitku sensitif & suka merah, tapi sejak pakai ini cuci muka jadi nyaman. Gak kering, ga ketarik. Teksturnya lembut, low pH, dan calming banget! | Kulitku sensitif suka merah tapi sejak pakai ini cuci muka jadi nyaman Gak kering ga ketarik Teksturnya lembut low pH dan calming banget |
| 1 | Texturenya sedikit thick tapi pas di apply bs cpt menyerap ke kulit, lightweight & ga lengket sm sekali. | Texturenya sedikit thick tapi pas di apply bs cpt menyerap ke kulit lightweight ga lengket sm sekali |
| ... | ... | ... |
| 12416 | Sunscreen ini teksturnya milky berwarna putih. | Sunscreen ini teksturnya milky berwarna putih |
| 12417 | Aku emg punya flek hitam di wajah dan kulit agak kusam, tapi setelah nyoba ini dalam waktu 7hari udah sedikit memudar fleknya, lebih cerah & ga kusam lagi | Aku emg punya flek hitam di wajah dan kulit agak kusam tapi setelah nyoba ini dalam waktu hari udah sedikit memudar fleknya lebih cerah ga kusam lagi |

**Table 2. Case Folding Results Data**

| No | clean_review | casefold_review |
|---|---|---|
| 0 | Kulitku sensitif suka merah tapi sejak pakai ini cuci muka jadi nyaman Gak kering ga ketarik Teksturnya lembut low pH dan calming banget | kulitku sensitif suka merah tapi sejak pakai ini cuci muka jadi nyaman gak kering ga ketarik teksturnya lembut low ph dan calming banget |
| 1 | Texturenya sedikit thick tapi pas di apply bs cpt menyerap ke kulit lightweight ga lengket sm sekali | texturenya sedikit thick tapi pas di apply bs cpt menyerap ke kulit lightweight ga lengket sm sekali |
| ... | ... | ... |
| 12416 | Sunscreen ini teksturnya milky berwarna putih | sunscreen ini teksturnya milky berwarna putih |
| 12417 | Aku emg punya flek hitam di wajah dan kulit agak kusam tapi setelah nyoba ini dalam waktu hari udah sedikit memudar fleknya lebih cerah ga kusam lagi | aku emg punya flek hitam di wajah dan kulit agak kusam tapi setelah nyoba ini dalam waktu hari udah sedikit memudar fleknya lebih cerah ga kusam lagi |

b. Case Folding

Case folding converts all characters to lowercase to standardize text representation. This step ensures that capitalization differences are not considered different word entities by the model. The results of the case folding process are presented in **Table 2**, which shows the conversion of all text into lowercase to ensure consistency in text representation.

c. Normalization

During the normalization stage, non-standard words, slang, abbreviations, and other writing variations are converted using a compiled slang dictionary. This step helps to standardize the informal language structure that often appears in Female Daily reviews. **Table 3** illustrates the normalization stage, where informal expressions, abbreviations, and slang words are converted into their standard equivalents to improve text consistency.

**Table 3. Normalization Result Data**

| No | Casefold_review | Normalization |
|---|---|---|
| 0 | kulitku sensitif suka merah tapi sejak pakai ini cuci muka jadi nyaman gak kering ga ketarik teksturnya lembut low ph dan calming banget | kulitku sensitif suka merah tapi sejak pakai ini cuci muka jadi nyaman tidak kering tidak ketarik teksturnya lembut low ph dan calming banget |
| 1 | texturenya sedikit thick tapi pas di apply bs cpt menyerap ke kulit lightweight ga lengket sm sekali | texturenya sedikit thick tapi pas di apply bisa cepat menyerap ke kulit lightweight tidak lengket sama sekali |
| ... | ... | ... |
| 12416 | sunscreen ini teksturnya milky berwarna putih | sunscreen ini teksturnya milky berwarna putih |
| 12417 | aku emg punya flek hitam di wajah dan kulit agak kusam tapi setelah nyoba ini dalam waktu hari udah sedikit memudar fleknya lebih cerah ga kusam lagi | aku memang punya flek hitam di wajah dan kulit agak kusam tapi setelah mencoba ini dalam waktu hari sudah sedikit memudar fleknya lebih cerah tidak kusam lagi |

**Table 4. Tokenization Results Data**

| No | Normalization | Tokenization |
|---|---|---|
| 0 | My skin is sensitive and prone to redness, but since using this, washing my face feels comfortable—it doesn't dry out or feel tight, and the texture is soft, low pH, and very calming. | ['my skin', 'sensitive', 'tends to', 'redden', 'but', 'since', 'using', 'cleansing', 'face', 'has', 'become', 'comfortable', 'not', 'dry', 'not', 'tight', 'texture', 'soft', 'calming', 'very'] |
| 1 | The texture is slightly thick, but when applied, it absorbs quickly into the skin. It's lightweight and not sticky at all. | ['texture', 'slightly', 'thick', 'but', 'apply', 'can', 'quickly', 'absorb', 'skin', 'lightweight', 'not', 'sticky', 'at', 'all'] |
| ... | ... | ... |
| 12416 | This sunscreen has a milky texture and is white in color. | ['sunscreen', 'texture', 'milky', 'white'] |
| 12417 | I do have dark spots on my face, and my skin is a bit dull, but after trying this for a few days, the spots have faded a bit, and my skin is brighter and no longer dull. | ['indeed', 'have', 'dark spots', 'face', 'skin', 'a bit', 'dull', 'but', 'after', 'trying', 'within', 'a few', 'days', 'they', 'have', 'faded', 'a bit', 'and', 'look', 'brighter', 'and', 'less', 'dull', 'again'] |

d. Tokenization

Tokenization uses the IndoBERT tokenizer to break the text into subword tokens as needed by the Transformer model. This step handles slang and code-mixed expressions by breaking them into valid token fragments. **Table 4** illustrates the tokenization stage, where each normalized sentence is segmented into smaller units (tokens) to facilitate further processing by the IndoBERT model.

### 3.6. Sentiment Labeling Results

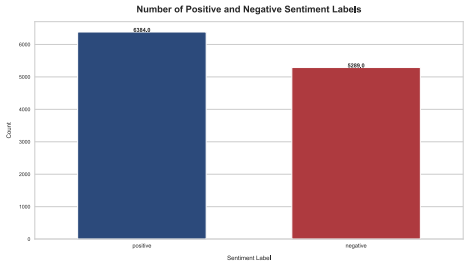


**Figure 6. Number of Positive and Negative Labels in the Dataset**

The sentiment labeling process was carried out automatically using the pre-trained mdhugol/indonesia-bert-sentiment-classification model from Hugging Face (J. Huang, 2023). The model automatically classified reviews into two

categories, namely positive (1) and negative (0), including a confidence score as the model's level of confidence in the predictions generated. This stage aims to produce a dataset that has been labeled with sentiment so that it can be used in the IndoBERT model training process. The labeling results show that out of the total reviews obtained, 6,384 were classified as positive and 5,289 as negative. This comparison shows that the majority of users tend to provide positive reviews of the skincare products they use.

**Figure 6** illustrates the sentiment class distribution in the dataset, consisting of 6,384 positive reviews and 5,289 negative reviews. The difference in numbers between classes is relatively small, so the dataset can be categorized as having a satisfactory balance or *near-balanced class distribution*.

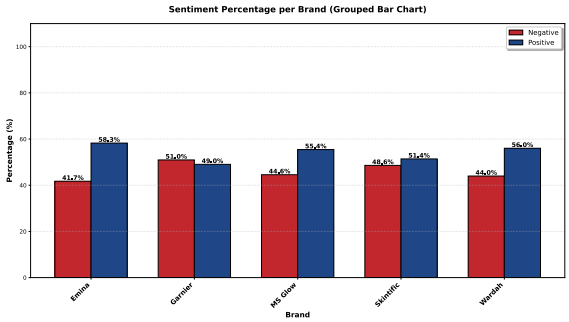


**Figure 7. Sentiment Percentage per Brand**

To gain a better understanding of consumer preferences, a multi-brand analysis was conducted, as illustrated in **Figure 7**. The visualization details the percentage of positive and negative sentiments across the five selected skincare brands. As shown in the chart, Wardah and Emina achieved the highest proportion of positive sentiments, indicating strong consumer satisfaction and confidence in these products. Conversely, Skintific and Garnier demonstrated a more balanced distribution between positive and negative reviews, suggesting diverse user experiences. MS Glow, on the other hand, had a wider range of feelings than the other brands. This brand-specific breakdown highlights the diverse consumer perception patterns in the Indonesian beauty market and enriches the contextual variety of the dataset for the IndoBERT model.

### 3.7. Model Training Results

The model training process was carried out using the indobenchmark/indobert-base-p1 model with a labeled dataset. The model was retrained to recognize sentiment patterns in Indonesian text and code-mixed Indonesian and English, which are characteristic of Female Daily users. Training was conducted in the Google Colab environment with an active GPU using the PyTorch framework and the Transformers library from Hugging Face. The training parameters used are shown in **Table 5**.

The implementation of these hyperparameter configurations became the main foundation in the fine-tuning process to ensure that the model could adapt optimally to the distribution characteristics and linguistic patterns in the review data. During the training process, the optimization process showed a stable convergence pattern, marked by a continuous decrease in loss and gradual improvement in all evaluation metrics. A numerical summary of the training dynamics is presented in **Table 6** to provide a quantitative overview of the model's performance development at each epoch.

**Table 6** represents the progressive dynamics of the IndoBERT model's performance over three epochs, showing consistent improvement in all evaluation parameters. The simultaneous decrease in Training Loss and Validation Loss from the initial to the final epoch indicates that the model successfully minimized prediction errors effectively without experiencing

significant overfitting. The linear increase in accuracy and F1-score with training iterations confirms the model's ability to absorb complex linguistic patterns in review data, proving that the fine-tuning mechanism works optimally in adapting IndoBERT's semantic representation to the specific domain of skincare.

**Table 5. IndoBERT Model Training Parameters**

| Parameter | Value |
|---|---|
| Base Model | IndoBERT-basep1 |
| Batch Size | 16 |
| Learning Rate | 2e-5 |
| Epoch | 3 |
| Optimizer | AdamW |
| Loss Function | Cross-Entropy |
| Train/Valid/Test Data Ratio | 70%/20%/10% |

**Table 6. IndoBERT Model Training Parameters**

| Epoch | Train Loss | Valid Loss | Accuracy | F1-Score | Precision | Recall |
|---|---|---|---|---|---|---|
| 1 | 0.52 | 0.42 | 0.80 | 0.79 | 0.79 | 0.79 |
| 2 | 0.35 | 0.36 | 0.83 | 0.83 | 0.84 | 0.83 |
| 3 | 0.28 | 0.33 | 0.85 | 0.84 | 0.85 | 0.84 |

The narrowing gap between training and validation metrics across epochs also shows that the model generalizes effectively to unfamiliar data. This pattern indicates that the chosen learning rate and optimization configuration were well-aligned with the characteristics of the dataset.

The model training process during fine-tuning shows a consistent pattern of training loss and validation loss reduction in each epoch, as visualized in **Figure 8**. Training loss decreases sharply from the first to the third epoch, while validation loss also decreases but at a slower rate.

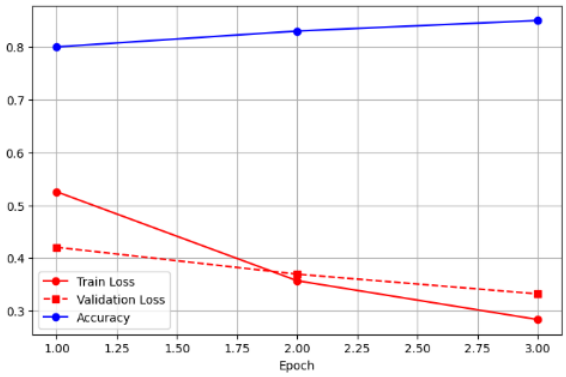


**Figure 8. Accuracy and Loss Curve of the IndoBERT Model**

The fact that the two curves remain close together without significant divergence indicates that the model does not overfit and is able to maintain its generalization ability on the validation data. On the same dataset, accuracy increased consistently at each epoch, indicating that IndoBERT successfully internalized linguistic patterns in the dataset progressively. The combination of the downward trend in loss and the upward trend in accuracy is a strong indication that the fine-tuning process was effective and resulted in stable convergence.

### 3.8. Model Evaluation

Model evaluation was conducted to measure the ability of fine-tuned IndoBERT to classify reviews correctly. Testing was performed using 10% of the total dataset as test data, producing the following evaluation values.

Based on the results in **Table 7**, the model shows solid performance, with balanced scores across accuracy (0.85), precision (0.85), recall (0.84), and F1-score (0.84), indicating that it is effective in classifying sentiment in skincare reviews. In addition to using numerical metrics, a more in-depth analysis was conducted using a Confusion Matrix to comprehensively map the details of the prediction class distribution. This visual approach is essential for understanding how accurately the model distinguishes

sentiment classes and for identifying specific classification error patterns, as illustrated in **Figure 8**.

**Table 7. Model Evaluation Results Test Data**

| Metric | Value |
|---|---|
| Accuracy | 0.85 |
| Precision | 0.85 |
| Recall | 0.84 |
| F1-Score | 0.84 |

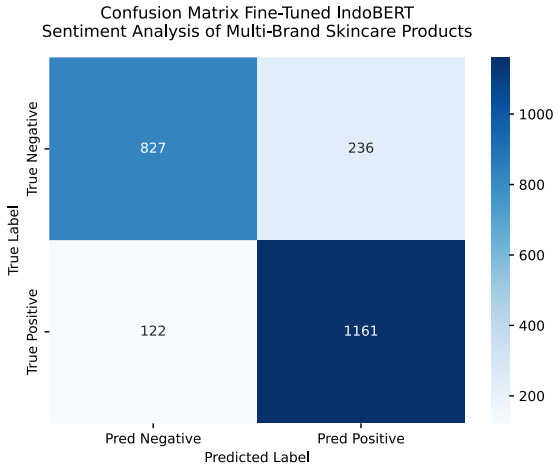


**Figure 9. Confusion Matrix of Fine-Tuned IndoBERT**

The Confusion Matrix visualization in **Figure 9** highlights the details of the model's classification performance. The matrix reveals that the model accurately predicted 1,161 positive reviews (True Positive) and 827 negative reviews (True Negative). The high numbers on the main diagonal of the matrix indicate that the model has excellent generalization capabilities. But there are still mistakes in the predictions, with 236 false positives and 122 false negatives. Reviews containing sarcastic language or implicit sentences that are difficult to detect based on word patterns alone likely cause these classification errors. But the low error rate shows that the fine-tuned IndoBERT model is generally very robust when it comes to processing different types of language structures, including the features of mixed languages. This stability shows that the model is able to maintain consistent performance even when faced with irregular syntactic patterns and vocabulary.

### 3.9. Discussion

The test results show that fine-tuning IndoBERT is capable of providing superior performance in classifying code-mixed text reviews with an accuracy rate of 85%. This proves that Transformer-based models have an advantage in understanding sentence context compared to classical methods such as TF-IDF and Naive Bayes, which only utilize static word representations (Yutika, Adiwijaya, and Faraby, 2021). The self-attention capability of IndoBERT allows the model to capture semantic relationships between words, including specific terms in the world of beauty such as "light texture," "fast absorption," or "oily T-zone," which often appear in user reviews. This is a major factor in why the model is able to produce more accurate predictions.

Overall, the results of this study are in line with previous studies that state that Transformer models have better generalization capabilities on product reviews written in natural language (Prottasha et al., 2022). The performance of the IndoBERT model in this study shows its potential application for opinion analysis in other domains such as e-commerce and the hotel sector (H. Huang, Zavareh, and Mustafa 2023; Ounacer et al., 2023).

## 4. CONCLUSION

This study successfully applied the IndoBERT Transformer fine-tuning method to analyze sentiment in multi-brand skincare product reviews obtained from the Female Daily platform. Based on the results of the experiments conducted, the IndoBERT model showed excellent performance compared to the traditional

method in classifying user reviews into two polarity categories, namely positive and negative. The fine-tuning process, which was carried out over three epochs, resulted in an accuracy of 85% with an F1-score of 0.84, precision of 0.85, and recall of 0.84, indicating stable performance in recognizing positive and negative opinion polarities. The results of automatic labeling using the pre-trained IndoBERT model produced a balanced dataset between the two classes, thereby supporting optimal model training. Multi-brand analysis shows that Wardah and Emina received the highest positive sentiment, while Skintific and Garnier had balanced sentiment distribution, and MS Glow displayed more varied opinions.

Overall, this study demonstrates that IndoBERT outperforms traditional approaches in understanding Indonesian-language contexts, particularly in texts that contain code-mixed expressions and informal slang commonly found in online skincare reviews. Unlike conventional methods such as TF-IDF combined with SVM, which rely on surface-level word frequency and often fail to capture contextual relationships between words, IndoBERT leverages a Transformer-based architecture with a self-attention mechanism. This mechanism allows the model to capture complex semantic dependencies and interpret words based on their surrounding context. As a result, IndoBERT is better equipped to understand informal linguistic patterns, mixed-language expressions, and community-specific terminology frequently used in the Indonesian skincare discourse, leading to more accurate contextual representation and classification performance.

Despite its strong performance, this study has specific limitations that direct future research. First, to address the model's difficulty with sarcasm and implicit complaints, future studies should integrate context-aware or sarcasm-detection techniques. Second, developing aspect-based sentiment analysis is crucial to move beyond simple binary classification and evaluate specific product features, such as texture or ingredients. Finally, validating the model using cross-platform datasets is necessary to ensure robust generalization beyond the unique linguistic style of the Female Daily community, which may include diverse user demographics and varying expressions of sentiment across different platforms

## REFERENCES

Abdullah, A. A., Hassan, M. M., & Mustafa, Y. T. (2024). Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach. *Heliyon*, 10(2). https://doi.org/10.1016/j.heliyon.2024.e24188

Anandia, N., Purnomo, W., & Setiawan, N. Y. (2025). Product monitoring terhadap Emina Sun Battle SPF 30 PA+++ pada Sociolla berdasarkan hasil analisis sentimen. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 9(4), 2548–2964.

Al Aufar, A. P., & Romadhony, A. (2025). Aspect-based sentiment analysis on beauty product reviews using BERT and long short-term memory. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, 14(2), 364–373. https://doi.org/10.23887/janapati.v14i2.94392

Ayu Puspita, N., Anggraeny, F. T., & Rizki, A. M. (2024). Analisis sentimen dengan algoritma naïve Bayes classifier terhadap ulasan aplikasi My F&B ID. *Jurnal Mahasiswa Teknik Informatika*.

Dheanis, K., Salsabila, A., & Trianasari, N. (2021). Analisis persepsi produk kosmetik menggunakan metode sentiment analysis dan topic modeling (Studi kasus: Laneige Water Sleeping Mask). *Jurnal Teknologi dan Manajemen Informatika*, 7(1).

Fadilah, G. T., Muflikhah, L., & Perdana, R. S. (2025). Analisis sentimen produk hijab pada e-commerce Tokopedia menggunakan algoritma support vector machine dan IndoBERT embedding. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 9(2), 2548–2964.

Habibi, M., & Kusumaningtyas, K. (2023). Customer experience analysis of skincare products through social media data using topic modeling and sentiment analysis. *Journal of Science and Applied Engineering*, 6(1), 1. https://doi.org/10.31328/jsae.v6i1.4169

Hidayat, E. Y., & Handayani, D. (2023). Penerapan 1D-CNN untuk analisis sentimen ulasan produk kosmetik berdasarkan Female Daily review. *Jurnal Nasional Teknologi dan Sistem Informasi*, 8(3), 153–163. https://doi.org/10.25077/teknosi.v8i3.2022.153-163

Huang, H., Zavareh, A. A., & Mustafa, M. B. (2023). Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. *IEEE Access*, 11, 90367–90382. https://doi.org/10.1109/ACCESS.2023.3307308

Huang, J. (2023). Practical application of platform comment sentiment binary classification based on deep learning. *Academic Journal of Science and Technology*.

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment analysis in Indonesian healthcare applications using IndoBERT approach. *International Journal of Advanced Computer Science and Applications (IJACSA)*.

Jasmarizal, J., Rahmaddeni, R., Junadhi, J., & Anam, M. K. (2024). Penerapan metode support vector machine untuk analisis sentimen. *Indonesian Journal of Computer Science*, 13(1).

Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Dwiyanto, F. A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah*, 14(3), 348–354. https://doi.org/10.33096/ilkom.v14i3.1505.348-354

Manoppo, M. R., Kolang, I. C., Fiat, D. N. N., Mawara, R. M. C., Sumarno, A. D. P., Yusupa, A., & Tarigan, V. (2025). Analisis sentimen publik di media sosial terhadap kenaikan PPN 12% di Indonesia menggunakan IndoBERT. *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, 4(2), 152–163. https://doi.org/10.69916/jkbti.v4i2.322

Mei, N. C., Tiun, S., & Sastria, G. (2024). Multi-label aspect-sentiment classification on Indonesian cosmetic product reviews with IndoBERT model. *International Journal of Advanced Computer Science and Applications (IJACSA)*.

Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., & Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12, 60943–60959. https://doi.org/10.1109/ACCESS.2024.3386969

Mutinda, J., Mwangi, W., & Okeyo, G. (2023). Sentiment analysis of text reviews using lexicon-enhanced BERT embedding (LeBERT) model with convolutional neural network. *Applied Sciences*, 13(3). https://doi.org/10.3390/app13031445

Ounacer, S., Mhamdi, D., Ardchir, S., Daif, A., & Azzouazi, M. (2023). Customer sentiment analysis in hotel reviews through natural language processing techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*.

Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the real world: A survey on NLP applications. *Information*, 14(4). https://doi.org/10.3390/info14040242

Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., & Baz, M. (2022). Transfer learning for sentiment analysis using BERT-based supervised fine-tuning. *Sensors*, 22(11). https://doi.org/10.3390/s22114157

Singh, D., Barve, S. S., & Dwivedi, A. K. (2025). OptiASAR: Optimized aspect-based sentiment analysis of reviews with BiLSTM-GRU and NER-BERT in healthcare decision-making. *IEEE Access*, 13, 47459–47473. https://doi.org/10.1109/ACCESS.2025.3549303

Widya Astuti, L., & Sari, Y. (2023). Code-mixed sentiment analysis using transformer for Twitter social media data. *International Journal of Advanced Computer Science and Applications (IJACSA)*.

Wildana, A., & Kautsar, A. (2025). Sistem rekomendasi pemilihan produk skincare. *Jurnal Global Ilmiah*, 2(5).

Yutika, C. H., Adiwijaya, A., & Faraby, S. A. (2021). Analisis sentimen berbasis aspek pada review Female Daily menggunakan TF-IDF dan naïve Bayes. *Jurnal Media Informatika Budidarma*, 5(2), 422. https://doi.org/10.30865/mib.v5i2.2845