ISSN 2089-8673 (Print) | 2548-4265 (Online)
Volume 14, Issue 2, 2025, pp. 364-373

ASPECT-BASED SENTIMENT ANALYSIS ON BEAUTY PRODUCT REVIEWS USING BERT AND LONG SHORT-TERM MEMORY

Arya Prima Al Aufar*, Ade Romadhony

School of Computing, Master of Computing, Telkom University Bandung, Jawa Barat, Indonesia

Abstract

In e-commerce, product reviews play a crucial role in influencing potential buyers by sharing user experiences and assessing product quality. This is especially important for beauty products, where poor quality can lead to physical harm. Reviews also help increase consumer interest in purchasing. Previous research has shown that product reviews differ in various aspects and content, making it challenging for consumers to quickly analyze them from multiple perspectives. This study applies aspect-based sentiment analysis to beauty product reviews on the Female Daily Network using a combination of BERT and LSTM. The goal is to provide more precise sentiment classification across different aspects, aiding consumers in selecting the best products. Several evaluation scenarios were conducted to assess different aspects of product reviews, including price, packaging, staying power, moisture, and aroma. The F-1 score revealed that the price aspect achieved the highest performance, reaching 100% in a 90%:10% test data scenario. However, the aroma aspect proved the most challenging to analyze, indicating that the model struggles to capture features related to scent effectively under the given evaluation setup.

Keywords: Aspect, Beauty Product, BERT, LSTM, Sentiment Analysis

Received: 14-04-2025 | **Accepted:** 07-07-2025 | **Available Online:** 16-07-2025
DOI: https://doi.org/10.23887/janapati.v14i2.94392

I. INTRODUCTION

Technological advancements have transformed various aspects of life, including online shopping and selling [1]. The shift from traditional face-to-face transactions to digital platforms has encouraged companies to expand their marketing channels through the internet [2]. The convenience offered by e-commerce attracts users [3], particularly in purchasing beauty products, which is the focus of this study [4].

Beauty products come in various brands and functions [5]. In e-commerce, product reviews are crucial for prospective buyers as they reflect other users' experiences [6]. Reviews help determine product suitability, ensuring safety and compliance with BPOM standards to prevent potential harm [7]. Research by Jain et al. (2021) confirms that reviews significantly influence purchasing decisions [8]. However, reviews often cover different aspects and content [6], [9] making it challenging for consumers to quickly analyze them [10]. Thus, a system is needed to process and summarize review data based on sentiments and aspects [11]. Sentiment analysis categorizes opinions into positive, negative, or neutral [12] and operates at the document, word, and aspect levels [6], making aspect-based sentiment analysis essential for obtaining more specific sentiment insights.

Zing Fang and Jie Tao [13] explored aspect-based sentiment analysis using BERT for text representation, focusing on restaurant reviews across aspects such as food, price, service, and ambiance, achieving 61.65% accuracy. Murthy et al. [14] employed LSTM for sentiment analysis on movie and Amazon product reviews, attaining 85% accuracy. Singh et al. [15] analyzed the impact of COVID-19 on social life using BERT with Twitter data, achieving 94% validation accuracy. Rai et al. [16] integrated BERT with LSTM for fake news classification, improving accuracy to 88.75% on PolitiFact and 84.10% on GossipCop, outperforming the standard BERT model by 2.50% and 1.10%, respectively.

Based on the description above, this research conducted sentiment analysis of beauty product reviews using aspect-based sentiment analysis with the aim of obtaining more specific sentiments based on their aspects [6]. The aspects used include packaging, moisture, price, staying power, and aroma. The aspects used are based on the research of Yutika, et al [17] is price, packaging, aroma. Aspect-based sentiment

*Corresponding author: aryaprima@student.telkomuniversity.com (A.P.A. Aufar)

analysis aims to identify the sentiment of customer reviews regarding different aspects of the product, such as Packaging, Moisture, Price, Durability, and Scent. Using a tokenization technique, each review is analyzed to extract the sentiment, which is then classified as positive (pos), negative (neg), or unknown.

The dataset has several aspects consisting of price, packaging, staying power, moisture and aroma. In addition, the labels used are positive, negative and unknown. Datasets that get the unknown label because the data does not have enough information to be classified into positive or negative labels, or because the data is ambiguous, so it is difficult to categorize.

Based on the collected user reviews, several key insights were identified regarding consumer perceptions of the product. In terms of application effectiveness, one review noted that reapplying the product several times did not result in any significant difference, suggesting that the product's performance may be limited for certain users.

Regarding the moisturizing effect, there were contrasting opinions. One user positively highlighted that the product did not cause lip dryness, indicating a beneficial moisturizing property. However, another user stated that the product only kept their lips moisturized for the first two hours, after which it caused dryness. This mixed feedback suggests that the moisturizing capability may vary depending on individual usage or skin type.

From the perspective of aroma, the response was generally positive. One review mentioned that the product had no disturbing smell or taste, which is considered a favorable attribute, especially for users sensitive to fragrances.

In terms of packaging, the product received positive remarks. A user complimented its slim and visually appealing design, indicating that packaging aesthetics and practicality play a role in consumer satisfaction.

However, there was no explicit information found in the reviews related to price or staying power of the product, hence both attributes remain undetermined based on the current data.

Meanwhile, aspects of moisture and staying power are based on research by Tran, et al [18]. In conducting sentiment analysis, researchers use a combination of BERT embedding and LSTM as a classification method to produce good performance. The use of LSTM as a classification method because it has the advantage of being able to model and learn long-term relationships between data and has strong potential in dealing with complex high dimensions on very large data [17], so that LSTM can recognize patterns in data to make predictions about sentiment [18]. The use of LSTM as a classification method is also due to the fact that it is a deep leaning model with a type of RNN architecture that is dedicated to sequential modeling such as text classification, and the sequential modeling capability is not possessed by other deep learning models [19]. While the use of BERT in this study is because the model has provided output in the form of pre-trained models that can be adopted for various NLP tasks, such as text classification [20]. From this description, the researcher conducted a combination of BERT Embedding and LSTM, this combination has the advantage of increasing the ability of the model [21] because BERT has a contextual sentence-level representation that can help LSTM in understanding sentence semantics better [22].

There are many studies related to sentiment analysis that have been done before. Some of these studies were conducted by Syiti Liviani Mahfiz and Ade Romadhony [23] which discusses sentiment analysis on beauty products. The reason behind this is that among Indonesian women, beauty products are the most discussed topic, so there are more and more reviews. Given the numbers, extracting aspect-based information from unstructured review text is a challenging task for consumers. Therefore, providing automated aspect-based opinions is a very valuable service for consumers. The research carried out aspect-based opinion extraction and polarity classification using Naive Bayes, with overall aspect F1-score is 50.55%. In addition, the study carried out 10 different preprocessing settings that combined filtering and stemming for Indonesian and English with an f-1 score of 53.04%.

Furthermore, research conducted by Chiorrini, et al [24] which discusses emotional and sentiment analysis on Twitter. The background of this research is that there has never been an investigation to determine people's opinions and emotions on Twitter. To investigate sentiment analysis using BERT. The results of this study show that for two datasets using BERT resulted in an accuracy of 92% for sentiment, and 90% for emotion analysis.

Further related research was conducted by Alsharef, et al [25] which discusses the classification of toxicity in social media using LSTM and word embedding. The background of this research is that automatic identification of toxicity is important in social media, so a classification of toxicity on social media using LSTM and word embedding is carried out. The results obtained for the combination of LSTM and Glove produced 93% accuracy, while LSTM and BERT produced 94% accuracy.

Further related research was conducted by research conducted by Dewi Ayu Khusnul Khotimah and Riyanarto Sarno [26] discussed the analysis of hotel sentiment aspects using probabilistic latent semantics, as well as word embeddings and LSTM. This research addresses the challenge of analyzing large volumes of unstructured product review data, which requires appropriate analytical techniques. These reviews can influence the hotel's image based on several aspects including location, food, service, comfort and cleanliness. In this research, the PLSA method was used to generate hidden topics and categorize five hotel aspects using semantic similarity, as well as TF-ICF to obtain important terms. Based on the results of this research, it shows that the combination of PLSA+TF ICF 100% + semantic similarity method gets superior results of 0.840 in five categorizations of hotel aspects. Meanwhile, the LSTM method with word embedding obtained precision of 0.932, recall of 0.960, and f-1 Measure of 0.946. The results of this research also show that the service aspect received the highest positive sentiment at 45,545 compared to other aspects. Meanwhile, the comfort aspect received the highest negative sentiment at 12,871 compared to other aspects

In addition to earlier works on general sentiment classification and aspect-based analysis, recent studies have addressed sentiment analysis in the specific context of e-commerce, particularly in beauty product reviews. Ng et al. [27] proposed a multi-label aspect-sentiment classification approach using the IndoBERT model on Indonesian cosmetic product reviews from Female Daily. Their work emphasized the effectiveness of contextual embedding in handling overlapping and nuanced sentiment expressions. Similarly, Cahyani [28] analyzed user-generated reviews on the Shopee marketplace using aspect-based sentiment analysis with Doc2Vec and TF-IDF, highlighting the importance of semantic representation in understanding product attributes. These studies provide recent evidence of growing interest and technical development in the field, reinforcing the relevance of this study's approach using BERT embeddings and LSTM architecture for beauty product sentiment classification.

Based on several related studies, this research aims to analyze sentiment in beauty product reviews by applying a combination of BERT embedding and LSTM. A dataset consisting of beauty product reviews was used, categorized into three sentiment classes: positive, negative, and unknown, across five aspects including packaging, moisture, price, staying power, and aroma. By leveraging BERT and LSTM methods, this study is expected to assist consumers in selecting the best beauty products based on sentiment analysis results.

The main contribution of this paper is the application of a hybrid BERT-LSTM model specifically trained on Indonesian-language beauty product reviews with fine-grained aspect-based sentiment labels. Unlike prior works that mostly relied on general sentiment classification or traditional machine learning approaches (e.g., TF-IDF + SVM), or that employed only BERT or only LSTM in isolation, this study demonstrates how combining contextual embeddings from BERT with sequential modeling from LSTM improves classification performance on subjective product aspects such as aroma and moisture. Furthermore, this research introduces a manually annotated corpus of 1503 Indonesian reviews labeled per aspect, which can serve as a benchmark for future studies.

II. METHOD

A. Data Collection

The dataset was collected using web scraping from the Female Daily Network, specifically targeting reviews related to lip-based beauty products, including lipstick, lip balm, and lip tint. These product types were chosen due to their high frequency of user reviews and rich descriptions across multiple aspects such as moisture, staying power, packaging, aroma, and price. The dataset was inspired by research by Syiti Liviani Mahfiz and Ade Romadhony [23] totaling 1503 data with aspects used, namely packaging, moisture, price, staying power and aroma.

B. Labelling

The labeling process was used to classify the dataset into positive, negative, and unknown classes across five aspects: packaging, moisture, price, staying power, and aroma. The five aspects are labeled using a manually created corpus, which serves as a reference to ensure consistency and accuracy in the annotation process.

C. Preprocessing

Preprocessing is the process of cleaning data by removing noise and handling missing values. This study applies case folding to convert text to lowercase and tokenization using NLTK to separate text into words. These steps prepare the data for further processing, such as stemming, and filtering. Additionally, BERT tokenization is used at the word embedding stage to ensure compatibility with the BERT model.

D. Split Data

In this process, a split data process was carried out with the aim of avoiding overfitting, which is a condition where the model learns too much training data so that it is unable to predict data that has never been seen before properly. In this splitting process, a method of separating datasets based on their labels is used, then recombining them into train data and test data. Following a prior study on the use of dataset splitting carried out by Kaur et al [29], is a comparison of 60% training data and 40% testing data, 70% training data and 30% testing data, 80% training data and 20% testing data and 90% training data and 10% testing data.

E. BERT Embedding

The word embedding process is performed using BERT embedding after the data splitting process. BERT Embedding was used to produce embedding tokens, which were subsequently used in the classification process.

F. Classification

In this process, the beauty product review classification process is carried out according to its aspects using BERT and LSTM. The classification results were then evaluated to test the performance of the BERT Embedding and LSTM models, so that the model performance of the BERT method and the LSTM method with BERT embedding can be known.

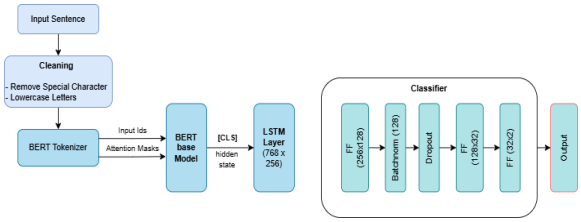

Figure 1. BERT-LSTM Architecture.

The combination of BERT embedding and LSTM model in performing aspect-based sentiment analysis on beauty products, utilizing the advantages of BERT in understanding better context as well as the advantages of LSTM in capturing sequential information. BERT embedding is used on the dataset so that it can generate embedding tokens. Furthermore, the data was trained and then predicted on the testing data in classifying beauty product reviews using the LSTM method [16].

G. Evaluation

The evaluation in this study was carried out using the following metrics: accuracy, precision, recall, and F-1 scores. The experiments carried out consisted of using test data, train data, and train data without unknown labels.

III.RESULT AND DISCUSSION

A. Presentation of Dataset

At this stage, the results of the research experiment are presented, starting from a dataset that already has a label and goes through a preprocessing stage, then the data is split into train data and test data, then an experiment is carried out on the BERT-LSTM model. The BERT-LSTM model experiment was carried out on test data, train data and test data without unknown labels.

Preprocessed Dataset

The final dataset includes 1029 preprocessed reviews distributed across five aspects: Price (123), Packaging (127), Staying Power (355), Moisture (316), and Aroma (108). Each review was manually annotated with one or more relevant labels and corresponding sentiment labels (positive, negative, or unknown). These counts reflect the usable review data after preprocessing, including case folding, tokenization, and filtering, which reduced the dataset from the original 1503 raw entries to a clean and consistent subset suitable for aspect-based sentiment classification. One of the sample datasets used in the preprocessing stage can be seen in Table 1 and Table 2.

Thus, the dataset that used in conducting sentiment analysis of beauty product reviews based on aspects can be seen in Table 3.

Table 1. Casefolding Result.

Before	After
saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel :d / when used there is no particular smell or taste that can make you feel disgusted :d	*saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel d* / When used there is no particular smell or taste that can make you feel uncomfortable d

Table 2. Tokenizing Result.

Before	After
saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel d / When used there is no particular smell or taste that can make you feel uncomfortable d	*['saat', 'dipakai', 'juga', 'tidak', 'ada', 'bau', 'atau', 'rasa', 'tertentu', 'yang', 'bisa', 'bikin', 'ilfeel']* / ['when', 'used', 'there', 'is', 'no', 'particular', 'smell', 'or', 'taste', 'that', 'can', 'make', 'you', 'feel', 'uncomfortable']

Table 3. Preprocessing Result.

review_sent	Casefolding	Tokenizing
produk ini tidak bikin bibir kering, sih, di saya. / This product doesn't make my lips dry, though.	*produk ini tidak bikin bibir kering sih di saya* / This product doesn't make my lips dry	*produk ini tidak bikin bibir kering sih saya* / This product doesn't make my lips dry
saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel :d / when used there is no particular smell or taste that can make you feel disgusted :d	*saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel d* / When used there is no particular smell or taste that can make you feel uncomfortable d	*saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel* / When used there is no particular smell or taste that can make you feel uncomfortable
…	…	…
worth the price lah. / worth the price.	worth the price lah worth the price	worth the price lah / worth the price
tapi lama-lama bosen juga, terus nyoba rebel ini, eh sekarang malah jadi seneng warna yang bold kayak gini teksturnya cair dan ringan, pas di bibir nggak kerasa tebal, nggak ada sensasi kayak ketarik gitu, untuk lipstik matte begini buat ku masih tergolong nyaman dan nggak terlalu bikin bibir terasa kering juga (selama masih pake lip balm dulu sebelumnya). / but after a while I got bored, then I tried this rebel, eh now I actually like bold colors like this, the texture is liquid and light, it doesn't feel thick on the lips, there's no pulling sensation, for a matte lipstick like this for me it's still quite comfortable and doesn't make my lips feel too dry (as long as I use lip balm first).	*tapi lamalama bosen juga terus nyoba rebel ini eh sekarang malah jadi seneng warna yang bold kayak gini teksturnya cair dan ringan pas di bibir nggak kerasa tebal nggak ada sensasi kayak ketarik gitu untuk lipstik matte begini buat ku masih tergolong nyaman dan nggak terlalu bikin bibir terasa kering juga selama masih pake lip balm dulu sebelumnya* / but after a while I got bored of trying this rebel but now I actually like bold colors like this the texture is liquid and light on the lips it doesnt feel thick theres no pulling sensation for a matte lipstick like this its still comfortable for me and doesnt make my lips feel too dry as long as I use lip balm beforehand	*tapi lamalama bosen juga terus nyoba rebel ini sekarang malah jadi seneng warna yang bold kayak gini teksturnya cair dan ringan pas bibir nggak kerasa tebal nggak ada sensasi kayak ketarik gitu untuk lipstik matte begini buat masih tergolong nyaman dan nggak terlalu bikin bibir terasa kering juga selama masih pake lip balm dulu sebelumnya* / but after a while I got bored of trying this rebel and now I actually like bold colors like this, the texture is liquid and light on the lips, it doesn't feel thick, there's no pulling sensation, for a matte lipstick like this, it's still relatively comfortable and doesn't make the lips feel too dry as long as you use lip balm beforehand

From the preprocessing results in Table 4, there are five kinds of labels used as the final result in this study, namely price_label for product reviews related to price, packaging_label for product reviews related to product packaging, staying_power_label for product reviews related to product durability, moist_label for product reviews related to the moisture provided by the product after use, and aroma_label for product reviews related to product aroma. One of them in the review “saat dipakai juga tidak ada bau atau rasa tertentu yang bisa bikin ilfeel :d” is included in the positive reviews on the aroma_label.

Training

The training process for aspect-based sentiment analysis leverages BERT embedding, and LSTM. The following diagram illustrates the stages of the training process.

The Results section is a critical component of the research article, presenting the main findings derived from the study, including outcomes of hypothesis testing. The use of tables and figures (e.g., charts, graphs) is strongly encouraged to visually convey the data in a clear and effective manner.

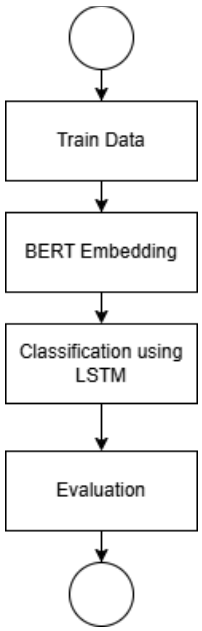

Figure 2. Train Flowchart.

1. Packaging Aspect
   The Packaging aspect achieved an F1-score of 99.86% in the 80%:20% scenario (Table 4). This high performance is due to the structured nature of packaging reviews, which consistently describe attributes like design, material, and color. However, its slightly lower performance compared to the Price aspect may result from greater variability in descriptive terms.

Table 4. BERT-LSTM Model Experiment on Training Data for Packaging Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	99.62	99.62	99.62	99.60
70%:30%	96.07	93.76	96.07	94.84
80%:20%	99.86	99.86	99.86	**99.86**
90%:10%	95.45	92.90	95.45	93.99

2. Moisture Aspect
   The Moisture aspect achieved an F1-score of 99.34% in the 70%:30% scenario (Table 5). Its subjective nature poses challenges, as terms like "moisturizing" or "oily" vary in interpretation. Despite this variability, the model demonstrated strong performance in handling complex sentiment data.

Table 5. BERT-LSTM Model Experiment on Training Data for Moisture Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	88.17	87.38	88.17	87.59
70%:30%	99.35	99.35	99.35	**99.34**
80%:20%	88.70	91.24	88.70	85.52
90%:10%	93.00	93.02	93.00	92.99

3. Price Aspect
   The Price aspect achieved the highest F1-score of 99.75% in the 90%:10% scenario (Table 6) due to its clear sentiment indicators, such as "cheap" or "expensive." Its structured nature allowed the model to generalize effectively, ensuring high precision and recall.

Table 6. BERT-LSTM Model Experiment on Training Data for Price Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	95.99	97.02	95.99	94.46
70%:30%	97.55	98.27	97.55	97.65
80%:20%	98.86	98.90	98.86	98.87
90%:10%	99.75	99.75	99.75	**99.75**

4. Staying Power Aspect
   The Staying Power aspect achieved its highest F1-score of 99.86% in the 80%:20% scenario (Table 7). Its complexity arises from nuanced descriptions of durability, making generalization challenging despite the model's strong pattern recognition.

Table 7. BERT-LSTM Model Experiment on Training Data for Staying Power Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	99.81	99.81	99.81	99.81
70%:30%	78.71	78.25	78.71	78.07
80%:20%	99.86	99.86	99.86	**99.86**
90%:10%	75.06	65.29	75.06	69.62

5. Aroma Aspect
   The Aroma aspect achieved its highest F1-score of 99.84% in the 70%:30% scenario (Table 8). Its complexity stems from subjective fragrance descriptions and varied terminology, making sentiment patterns harder to generalize. The absence of some test data terms in the training set further impacted on model performance.

Table 8. BERT-LSTM Model Experiment on Training Data for Aroma Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	99.81	99.81	99.81	99.81
70%:30%	99.84	99.84	99.84	**99.84**
80%:20%	94.28	91.17	94.28	92.34
90%:10%	99.75	99.75	99.75	99.74

Based on Table 4 to Table 8, the BERT-LSTM model exhibited varying performance across sentiment aspects, influenced by their distinct characteristics. The Price aspect achieved the highest F1-score of 99.75% due to clear and consistent terms like "cheap" or "expensive," which the model easily recognized. The Packaging aspect followed with 99.86%, benefiting from structured descriptions of design, color, and material. In contrast, the Aroma aspect (99.84%) faced challenges due to its subjective nature and variability in fragrance-related terms, some of which were absent in the training data. Similarly, the Moisture aspect (99.34%) was affected by diverse interpretations of terms like "moisturizing" or "oily." The Staying Power aspect (99.86%) performed well but struggled with indirect expressions of durability. These findings highlight that aspects with clearer terminology enhance model performance, while

subjective or variable language hinders generalization [29].

The result is the critical part of the research article containing research findings and hypothesis testing results. A table and graphics are highly recommended to visualize the result.

Testing

The BERT-LSTM model experiment for test data was carried out using a scenario of different proportions of dataset sharing on each sentiment aspect used. The dataset proportion scenarios used are 60%: 40%, 70%:30%, 80%:20%, 90%:10%. The results of the model experiment on test data in each aspect can be seen in Table 9 to Table 13.

1. Packaging Aspect
   The Packaging aspect achieved its highest F1-score of 92.31% in the 80%:20% scenario, benefiting from structured and well-defined terminology that enhanced model generalization and classification accuracy.

Table 9. BERT-LSTM Model Experiment on Testing Data for Packaging Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	76.47	76.47	100.0	86.67
70%:30%	77.78	77.78	100.0	87.50
80%:20%	88.24	92.31	92.31	**92.31**
90%:10%	81.82	81.82	100.01	90.00

2. Moisture Aspect
   The Moisture aspect achieved a peak F1-score of 83.02%, lower than Packaging due to the subjective interpretation of terms like "hydrating" or "oily," leading to classification challenges.

Table 10. BERT-LSTM Model Experiment on Testing Data for Moisture Aspect.

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	73.79	76.00	86.36	80.85
70%:30%	75.34	77.19	89.9	**83.02**
80%:20%	66.67	66.67	100.0	80.00
90%:10%	64.52	70.00	73.68	71.79

3. Price Aspect
   The Price aspect exhibited the best performance overall, with a perfect F1-score of 100% in the 90%:10% scenario. This is expected, as terms related to price, such as "*murah*/cheap," "*mahal*/expensive," or "*terjangkau*/affordable," are usually very clear and straightforward, which makes it easier for the model to accurately identify and classify the sentiment.

Table 11. BERT-LSTM Model Experiment on Testing Data for Price Aspect

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	75.50	72.50	100.0	84.06
70%:30%	75.0%	90.91	66.67	76.92
80%:20%	75.0%	82.35	82.35	82.35
90%:10%	100.0%	100.0	100.0	**100.0**

4. Staying Power Aspect
   The Staying Power aspect achieved the second-highest F1-score of 95.24% in the 90%:10% scenario. While the model performed well, the subjective nature of staying power-related terms led to slightly lower precision than the Price aspect.

Table 12. BERT-LSTM Model Experiment on Testing Data for Staying Power Aspect

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	74.17	77.67	90.91	83.77
70%:30%	83.13	83.10	96.72	89.39
80%:20%	76.09	81.82	84.38	83.08
90%:10%	90.91	90.91	100.0	**95.24**

5. Aroma Aspect
   The Aroma aspect showed high variability, with an F1-score peaking at 80.0% in the 70%:30% scenario but dropping to 0% and 66.67% in other cases. This inconsistency stems from the subjective nature of fragrance-related terms and discrepancies between training and test data, affecting model precision and recall.

Table 13. BERT-LSTM Model Experiment on Testing Data for Aroma Aspect

Evaluation Scenario	Acc (%)	Pre-cision (%)	Recall (%)	F1 (%)
60%:40%	75.00	80.00	66.67	72.73
70%:30%	82.35	85.71	75.00	**80.00**
80%:20%	60.00	0.0	0.0	0.0
90%:10%	80.00	100.0	50.00	66.67

The testing results highlight distinct performance characteristics across the five sentiment aspects. The Price aspect exhibited the best results, achieving an F1-score of 100% in the 90%:10% scenario,

primarily due to clear and consistent sentiment indicators such as "cheap" and "expensive." The Staying Power aspect followed with an F1-score of 95.24%, though it presented slight ambiguity due to varying expressions of durability. The Packaging aspect yielded a high F1-score of 92.31%, supported by consistent descriptions of visual and structural features. In contrast, Moisture produced a moderate F1-score of 83.02%, as user perception of hydration tends to be more subjective and varied. The most challenging aspect was Aroma, where the model struggled with inconsistent terminology and ambiguous phrasing, resulting in fluctuating F1-scores between 0% and 80% depending on the data split. These differences affirm that model performance correlates with the clarity and consistency of aspect-specific expressions found in the dataset.

B. Data Analysis

The BERT method was applied for word embedding, which produces token embeddings that are then utilized in the classification process. The classification was carried out using the LSTM model, which utilized an architecture consisting of 64 and 128 units with the tanh activation function, a dropout rate of 0.1, and a dense layer with 128 units and the ReLU activation function. During the training process, a batch size of 32 was used, and the model was trained over 100 epochs. An evaluation of the model's performance was carried out using a confusion matrix to obtain values for accuracy, precision, recall, and F1-score. Experiments were conducted across various aspects of sentiment analysis for beauty product reviews, specifically focusing on Packaging, Price, Moisture, Staying Power, and Aroma. These aspects were evaluated to understand how well the model could perform across different types of sentiments and conditions, which was applied to balance the dataset between positive and negative classes.

From the experiments that have been carried out for sentiment analysis of beauty product reviews using the BERT-LSTM method, the following conclusions are obtained:

1. The evaluation results indicate that the model performed well on training data for Packaging, Moisture, and Aroma aspects but showed variability on testing data, particularly for Price and Staying Power. Since training performance does not always reflect generalization, error analysis on test data is crucial to accurately assessing the model's predictive capability. Thus, the BERT-LSTM model's performance is evaluated based on its results with testing data.
2. From the results of the research conducted, an error analysis was performed on the most difficult prediction conditions, which can be observed from the test data. When the evaluation results show the lowest values, specifically in the Aroma aspect with the 80%:20% evaluation scenario, the accuracy obtained is 60.0%, and the F1-score is 0.0%. The misidentification in the Aroma aspect occurred because the model struggled to distinguish between positive and negative sentiments, especially in words or phrases that are ambiguous in product reviews. Some examples of reviews that fall into the error category are:
   a. "*Baunya emang agak ganggu waktu dipake*" / "The smell is a bit annoying when I use it" = should be neg (negative) but predicted as unknown.
   b. "They only cont, the smell realy don't like" = should be neg (negative) but predicted as pos (positive).
   c. "*Baunya khas, tapi nggak terlalu menyenangkan buat saya*" / "The smell is distinctive, but not too pleasant for me" = should be neg (negative) but predicted as unknown.

These errors may arise due to the model's difficulty in understanding sentiment nuances in aroma-related reviews and data imbalance, which biases predictions toward the dominant class. To address this, improvements in data preprocessing and model training, such as data augmentation or parameter adjustments, are needed

IV. CONCLUSION

This study successfully conducted sentiment analysis on beauty product reviews from the Female Daily Network using aspect-based sentiment analysis with a BERT-LSTM model. The model's performance is influenced by data distribution across labels and the quality of data obtained during the scraping process. Compared to previous research by Syiti Liviani Mahfiz and Ade Romadhony, which reported an F1-score of 50.55%, the BERT-LSTM model in this study achieved an F1-score exceeding 60%, with the highest reaching 100.0%, due to the enhanced word embedding process provided by BERT. The results demonstrate that the model effectively analyzes sentiment across five aspects, as indicated by accuracy, precision, recall, and F1-score metrics. However, error analysis highlights challenges in classification due to the absence of an unknown label, dataset quality issues, and

imbalanced data distribution. For future work, applying stratified k-fold cross-validation and exploring more robust architecture such as RoBERTa or attention-based LSTM are recommended to enhance generalization and performance, particularly for subjective aspects like aroma and moisture.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Telkom University for providing the research grant and continuous support throughout this study. This support has been instrumental in facilitating data collection, analysis, and the overall development of this research.

REFERENCES

[1] "[01] Statistik E-Commerce 2019".
[2] V. Sima, I. G. Gheorghe, J. Subić, and D. Nancu, "Influences of the industry 4.0 revolution on the human capital development and consumer behavior: A systematic review," *Sustainability (Switzerland)*, vol. 12, no. 10, May 2020, doi: 10.3390/SU12104035.
[3] M. Yunus, J. Saputra, and Z. Muhammad, "Digital marketing, online trust and online purchase intention of e-commerce customers: Mediating the role of customer relationship management," *International Journal of Data and Network Science*, vol. 6, no. 3, pp. 935–944, Jun. 2022, doi: 10.5267/j.ijdns.2022.2.003.
[4] "[04] BPS, Statistik E-Commerce 2020".
[5] P. M. Coelho, B. Corona, R. ten Klooster, and E. Worrell, "Sustainability of reusable packaging–Current situation and trends," May 01, 2020, *Elsevier B.V.* doi: 10.1016/j.rcrx.2020.100037.
[6] N. Putri Arthamevia and M. Dwifebri Purbolaksono, "Aspect-Based Sentiment Analysis in Beauty Product Reviews Using TF-IDF and SVM Algorithm." [Online]. Doi: 10.1109/icoict52021.2021.9527489.
[7] I. Fadillah, M. A. Firmansyah, S. Hadi, M. A. Danurwindo,) Universitas, and M. Surabaya, "PROSIDING SEMINAR NASIONAL EKONOMI DAN BISNIS 1 Fakultas Ekonomi dan Bisnis Universitas Muhammadiyah Surabaya PENGARUH BRAND AWARENESS, KUALITAS PRODUK, DAN WORD OF MOUTH (WOM) TERHADAP MINAT BELI SKIN CARE LOKAL DI SOCIOLLA STORE SURABAYA."
[8] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," Aug. 01, 2021, *Elsevier Ireland Ltd*. doi: 10.1016/j.cosrev.2021.100413.
[9] M. A. A. T. Utami, P. Silvianti, and M. Masjkur, "Algoritme Support Vector Machine untuk Analisis Sentimen Berbasis Aspek Ulasan Game Online Mobile Legends: Bang-Bang," *Xplore: Journal of Statistics*, vol. 12, no. 1, pp. 63–77, Jan. 2023, doi: 10.29244/xplore.v12i1.1064.
[10] "[10] Aspect Based Sentiment Analysis With Combination Feature Extraction LDA and Word2vec".
[11] Muhammad Rio Pratama, Faza Abdillah Gunawan Soerawinata, Rafdi Reyhan Zhafari, Rendy, and Helena Nurramdhani Imanda, "Sentiment Analysis of Beauty Product E-Commerce Using Support Vector Machine Method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 2, pp. 269–274, Apr. 2022, doi: 10.29207/resti.v6i2.3876.
[12] S. Pandya, P. Mehta, and D. Pandya, "A Review On Sentiment Analysis Methodologies, Practices And Applications," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, p. 2, 2020, [Online]. Available: www.ijstr.org
[13] Mohammad. Alsmirat and Yaser. Jaraweh, *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) : Granada, Spain, October 22-25, 2019*. IEEE, 2019.
[14] G. S. N Murthy, S. Rao Allu, B. Andhavarapu, M. Bagadi, and M. Belusonti, "Text based Sentiment Analysis using LSTM; Text based Sentiment Analysis using LSTM.", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 05 (May 2020), doi: 10.17577/IJERTV9IS050290.
[15] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Soc Netw Anal Min*, vol. 11, no. 1, Dec. 2021, doi: 10.1007/s13278-021-00737-z.
[16] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, "Fake News Classification using transformer based enhanced LSTM and BERT," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 98–105, Jun. 2022, doi: 10.1016/j.ijcce.2022.03.003.
[17] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF

dan Naïve Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 422, Apr. 2021, doi: 10.30865/mib.v5i2.2845.

[18] Q.-L. Tran, P. Thanh, and D. Le, "Aspect-based Sentiment Analysis for Vietnamese Reviews about Beauty Product on E-commerce Websites." [Online]. Available: https://shopee.vn/

[19] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep Learning for Air Quality Forecasts: a Review," *Curr Pollut Rep*, vol. 6, no. 4, pp. 399–409, Dec. 2020, doi: 10.1007/s40726-020-00159-z.

[20] A. Rolangon *et al.*, "Perbandingan Algoritma LSTM Untuk Analisis Sentimen Pengguna Twitter Terhadap Layanan Rumah Sakit Saat Pandemi Covid-19 The Comparison of LSTM Algorithms for Twitter User Sentiment Analysis on Hospital Services During the Covid-19 Pandemic."

[21] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/j.neucom.2019.01.078.

[22] S. Mohammadi and M. Chapon, "Investigating the Performance of Fine-Tuned Text Classification Models Based-on Bert," in *Proceedings - 2020 IEEE 22nd International Conference on High Performance Computing and Communications, IEEE 18th International Conference on Smart City and IEEE 6th International Conference on Data Science and Systems, HPCC-SmartCity-DSS 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 1252–1257. doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00162.

[23] S. L. Mahfiz and A. Romadhony, "Aspect-based Opinion Mining on Beauty Product Reviews," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 488–493. doi: 10.1109/ISRITI51436.2020.9315350.

[24] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using BERT," 2021. [Online]. Available: https://www.researchgate.net/publication/350591267

[25] C. I. and Neuroscience, "Retracted: An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding," *Comput Intell Neurosci*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9850820.

[26] D. A. K. Khotimah and R. Sarno, "Sentiment analysis of hotel aspect using probabilistic latent semantic analysis, word embedding and LSTM," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 4, pp. 275–290, 2019, doi: 10.22266/ijies2019.0831.26.

[27] N. C. Mei, S. Tiun, and G. Sastria, "Multi-Label Aspect-Sentiment Classification on Indonesian Cosmetic Product Reviews with IndoBERT Model," 2024. [Online]. doi: 10.14569/ijacsa.2024.0151168.

[28] A. D. Cahyani, "Aspect-Based Sentiment Analysis from User-Generated Content in Shopee Marketplace Platform," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 2, pp. 444–454, Jun. 2023, doi: 10.26555/jiteki.v9i2.26367.

[29] R. Kaur, R. Kumar, and M. Gupta, "Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence," *Endocrine*, vol. 78, no. 3, pp. 458–469, Dec. 2022, doi: 10.1007/s12020-022-03215-4.