Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Integrated aspect extraction and sentiment classification for aspect-based sentiment analysis using fine-tuned indoBERT on indonesian e-commerce reviews

Feliks Victor
Parningotan Samosir[1], Gavrila Louise Tumanggor[2]
[1,2]Department of Informatics, Universitas Pelita Harapan, Indonesia

Article Info

Article history:

Received Mar 19, 2026
Revised April 9, 2026
Accepted April 12, 2026

Keywords:

Aspect based sentiment analysis
IndoBERT
E-commerce review
Transfer learning
Fine tuning

ABSTRACT

The rapid growth of Indonesian e-commerce has generated vast volumes of consumer reviews, yet extracting actionable aspect-level sentiment from informal Indonesian-language texts remains challenging due to the limited availability of domain-specific Aspect-Based Sentiment Analysis (ABSA) models. This study aimed to develop and evaluate an integrated IndoBERT-based ABSA model that combines aspect extraction and aspect sentiment classification within a single framework, applied to Indonesian beauty product reviews. A corpus of 500 beauty product reviews was processed through aspect extraction, yielding approximately 10,000 aspect-level data points labeled as positive or negative. The IndoBERT model was fine-tuned with optimized hyperparameters. The model achieved 86% accuracy, 85.71% F1-score, and 88% balanced accuracy. Aspect-level evaluation revealed F1-scores of 100% for seller, 98% for product, and 86% for shipping. Inference throughput of 33,173 samples per second confirmed real-world deployment feasibility. These results demonstrate the effectiveness of integrated IndoBERT fine-tuning for ABSA on Indonesian e-commerce reviews and provide a foundation for enhancing data-driven marketing strategies in the beauty product sector.

Corresponding Author:

Parningotan Samosir,
Department of Informatics,
Universitas Pelita Harapan, Indonesia
Email: feliks.parningotan@uph.edu
https://doi.org/10.52465/joscex.v7i1.26

1. INTRODUCTION

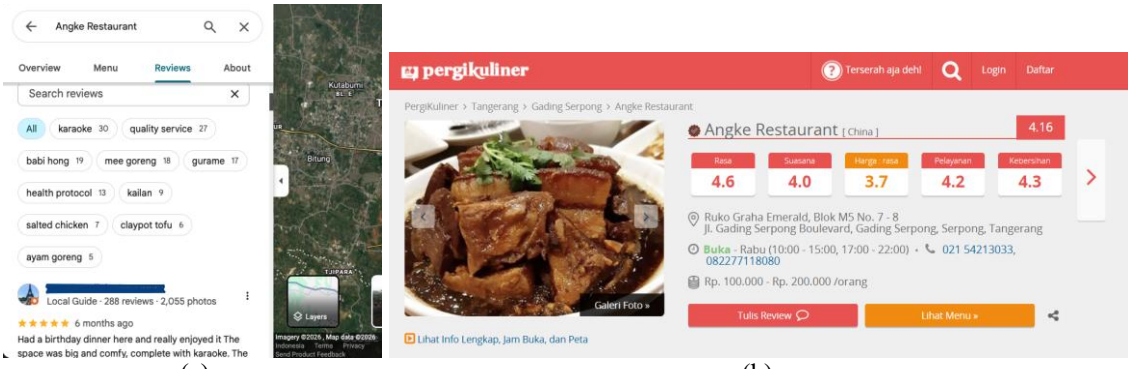
In the current digital era, the growth of e-commerce platforms has transformed the landscape of consumer behavior and interaction. Fernandes et al. [1] noted that consumers increasingly rely on online product reviews to guide their purchasing decisions. These reviews encompass a wide range of opinions, from positive praise to detailed criticism, offering valuable insights into product quality and functionality. The sheer volume of reviews on e-commerce platforms underscores the importance of leveraging review data for comprehensive analysis and decision-making. Among the various product categories in Indonesian e-commerce, beauty products represent a particularly active domain characterized by rich and diverse consumer feedback [2].

Research conducted by Thamrin et al. [2] provided an in-depth analysis of beauty product reviews, highlighting the significant role of beauty products in contemporary consumer life. Their study drew on sales data from leading Indonesian e-commerce platforms such as Shopee, Tokopedia, and Bukalapak, indicating strong demand for beauty products in the online marketplace. Furthermore, Google search trend data revealed substantial consumer interest in beauty product brands. These findings confirm that beauty products have become an integral part of consumers' daily lives, reflecting the growing emphasis on personal appearance and self-care. The presence of e-commerce as a primary platform for purchasing beauty products offers convenience in searching for and buying a wide range of items, from skincare to cosmetics. This abundance of user-generated content presents both an opportunity and a challenge for automated sentiment analysis [3].

Given the diverse range of available products, consumers often face the challenge of determining which products best suit their needs. Understanding the specific aspects that influence positive and negative consumer sentiment toward beauty products is essential for designing more targeted marketing strategies and product development. Research by Salsabila and Sibaroni [3] demonstrated that consumers explicitly express opinions based on aspects such as price, packaging, and fragrance, where price and packaging tended to receive positive sentiment, while fragrance was more likely to elicit negative sentiment. Similar findings were reported by Wardani et al. [4], who employed the Random Forest method to analyze beauty product reviews on e-commerce platforms and successfully identified that consumer perceptions were strongly influenced by technical and aesthetic aspects such as texture, fragrance, and packaging design. Both studies confirmed that aspect-based sentiment mapping not only helps companies understand consumer perceptions in greater detail but can also serve as a basis for store owners and researchers to adjust marketing strategies and product and service development more effectively in accordance with market demands. However, these approaches relied on traditional machine learning methods, which may not fully capture the contextual nuances of informal Indonesian-language text.

Aspect-Based Sentiment Analysis (ABSA) is a specialized form of sentiment analysis that provides deeper granularity [5]. ABSA focuses on identifying and extracting specific aspects from review text, then determining the sentiment associated with each aspect. The foundational theory underlying this method encompasses aspect identification, aspect extraction, and sentiment classification. ABSA consists of two main components: Aspect Extraction (AE) and Aspect Sentiment Classification (ASC) [6]. The underlying architecture that has revolutionized ABSA is the Transformer model, introduced by Vaswani et al. [7], which employs self-attention mechanisms to capture long-range dependencies in text, forming the basis for pre-trained language models such as BERT [8] and its derivatives. Sun et al. [9] further demonstrated that constructing auxiliary sentences from aspect terms significantly improves BERT's performance on ABSA tasks, establishing a methodological foundation for subsequent transformer-based ABSA research. Recent surveys have systematically examined the evolution of ABSA methodologies from traditional machine learning pipelines toward end-to-end transformer-based approaches, highlighting the advantages of integrated models that jointly handle aspect extraction and sentiment classification [10], [11].

Figure 1 (a) presents Google Review as an example of a platform that implements AE, where user reviews are analyzed to extract frequently occurring keywords such as "food taste" or "service." However, Google Review only extracts words without determining the sentiment associated with each aspect.



(a) (b)

Figure 1. Aspect extraction on google review (a) and aspect sentiment classification on PergiKuliner (b)

Furthermore, Karimi et al. [5] explained that ASC is a task within ABSA that aims to classify the sentiment expressed toward aspects extracted from a review. In ASC, the model analyzes the sentiment contained in consumer reviews and classifies it into positive, negative, or neutral categories. Figure 1 (b) shows the PergiKuliner platform as an example where ASC is applied, with predefined aspects such as "Taste" or "Cleanliness," so that each review is classified based on sentiment for each respective aspect

Research on ABSA in Indonesia has been extensively conducted using various deep learning and machine learning approaches. To contextualize the present study, the following review examines key prior works, their methodological approaches, and their limitations. A study by Imron et al. [12] examined product reviews on the Bukalapak platform using a combination of BERT as word embedding, LSTM for aspect classification, and CNN for sentiment classification. Through manual labeling of aspects and sentiments, the study achieved an accuracy of 93.91%, precision of 93.44%, recall of 91.84%, and an F1-score of 92.50%. These findings indicated that the combined LSTM-CNN approach with BERT embedding delivered superior performance compared to single-model approaches, offering an effective solution for understanding consumer opinions on an aspect basis.

Febrianto et al. [13] analyzed visitor reviews of the Baturraden tourism site using the IndoBERT model to classify sentiments based on four main aspects: Attraction, Accessibility, Amenities, and Ancillary Services. The results showed that IndoBERT achieved an accuracy of 94.61%, with precision of 83.22%, recall of 96%, and an F1-score of 88.11%, despite facing challenges in handling language variation and imbalanced data distribution across aspects.

Additionally, Audyna et al. [14] applied IndoBERT to perform aspect-based sentiment analysis on the topic of Electric Vehicles (EV) in Indonesia through the social media platform X. This study compared IndoBERT's performance with various other machine learning and deep learning methods, showing that IndoBERT achieved superior performance with a sentiment accuracy of 82% and an aspect accuracy of 85%. The analysis also found that negative sentiment dominated public discussion regarding electric vehicles, particularly concerning infrastructure and cost aspects.

Another ABSA study was conducted by Bahri and Suadaa [15] on tourism destination reviews in Indonesia, specifically at Bromo Tengger Semeru National Park, comparing conventional machine learning models (SVM, Complement Naïve Bayes, and Logistic Regression) with transfer learning models such as BERT, IndoBERT, and mBERT. The results demonstrated that transfer learning-based models, particularly IndoBERT, achieved the best performance with an accuracy of 91.48% and an F1-score of 71.56%. This achievement was attributed to IndoBERT's ability to comprehend the diverse range of Indonesian language used in social media and public reviews, as well as its advantage in leveraging large-scale Indonesian-language corpora through pre-training and fine-tuning processes.

More recently, Yulianti and Nissa [16] investigated three IndoBERT-based approaches for ABSA on Indonesian customer reviews: feature-based extraction paired with CNN, fine-tuned single-sentence classification, and fine-tuned sentence-pair classification. Their results demonstrated that the sentence-pair classification approach achieved the highest effectiveness, with a significant improvement of 23.6% in F1-score over deep learning baselines and 2.2% over multilingual BERT baselines, further confirming IndoBERT's superiority for Indonesian-language ABSA tasks.

Beyond the Indonesian-language context, transformer-based ABSA has also been applied to other domains and languages. Samosir and Ferdynand [17] applied DistilBERT for ABSA on Amazon book reviews in English, achieving high accuracy in classifying sentiments related to specific book aspects such as plot, character development, and writing style, demonstrating the versatility of transformer models for aspect-level sentiment classification across different product domains. In a separate study, Samosir and Riyaldi [18] utilized IndoBERT for sentiment analysis of TikTok comments on Indonesian presidential elections, further demonstrating the model's adaptability to diverse Indonesian-language text genres beyond product reviews. Additional studies have explored IndoBERT for various Indonesian NLP tasks including text classification with CNN and Bi-LSTM architectures [19], hate speech detection on Twitter [20], depression detection in health-related texts [21], multilabel news classification [22], and sarcasm detection using domain-adapted IndoBERTweet [23], collectively confirming the model's versatility across diverse classification tasks in Indonesian.

The aforementioned prior studies demonstrate that ABSA with IndoBERT is effective in classifying aspect-based sentiment across various domains and is capable of handling the diverse characteristics of Indonesian-language text data. IndoBERT, as a monolingual model developed from the BERT architecture with adaptation for the Indonesian language [24], relies on bidirectional attention mechanisms and pre-training techniques such as Masked Language Model (MLM) and Next Sentence Prediction (NSP) [8]. These capabilities enable IndoBERT to understand the context of words both preceding and following them in a sentence, making it more effective in capturing semantic meaning and inter-word relationships in Indonesian.

Despite the growing body of research on IndoBERT-based ABSA, several research gaps remain. First, existing studies have primarily focused on tourism destinations [13], [15], electric vehicles [14], general

marketplace products [12], and healthcare [25], with the beauty product review domain in Indonesian e-commerce remaining unexplored. Second, most prior approaches employ separate models for aspect extraction and sentiment classification [12], lacking an integrated single-model framework that offers computational efficiency. Third, computational efficiency metrics essential for real-world deployment, such as inference throughput and training time, are rarely reported in comparable studies. To address these gaps, this study proposes the application of IndoBERT for the ABSA task on Indonesian-language beauty product reviews. This domain presents unique NLP challenges due to its characteristic informal language, brand-specific terminology, colloquial expressions about skin conditions, and nuanced subjective evaluations that differ substantially from the more structured language found in tourism or automotive reviews. Furthermore, whereas most prior studies employ separate models for aspect extraction and sentiment classification [12], this study integrates both Aspect Extraction (AE) and Aspect Sentiment Classification (ASC) within a single IndoBERT model, providing computational efficiency and representational consistency. Additionally, this research contributes granular per-aspect F1-score evaluation and computational efficiency metrics (inference throughput and training time per iteration), which are rarely reported in comparable studies but are essential for assessing real-world deployment feasibility in e-commerce applications. The objectives of this research are: (1) to fine-tune IndoBERT for integrated ABSA on Indonesian beauty product reviews, (2) to evaluate the model's per-aspect classification performance, and (3) to assess computational efficiency for potential deployment in e-commerce environments.

2. METHOD

The overall methodology of this study is illustrated in Figure 2, which presents the experimental workflow from data collection through model evaluation.

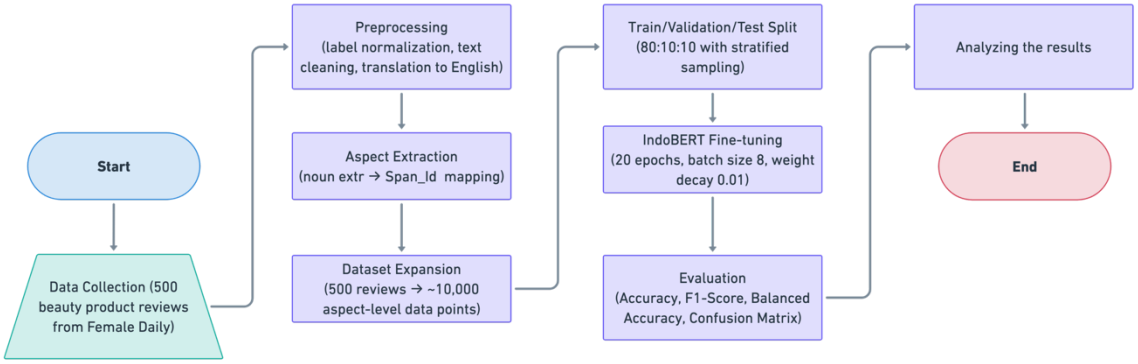


Figure 2. Methodology diagram

Data Collection and Preparation

In this study, the data collection and preparation process began by obtaining a beauty product review dataset from GitHub, previously provided by Claudiah [26]. This dataset comprises reviews of facewash products from the Female Daily application. Table 1 presents the structure of the dataset.

Table 1. Dataset structure and completeness overview

index	ID Data	Product Name	Skin Type	Product Review	Rating	Label
0	A001	Cetaphil: Gentle Skin Cleanser	Oily	Aku suka banget sama ini. Cu…	5	Positive
1	A002	Cetaphil: Gentle Skin Cleanser	Oily	Walaupun produknya diklaim…	5	Positive
2	A003	Cetaphil: Gentle Skin Cleanser	Dry	Thanks God for this product…	5	Positive
3	A004	Cetaphil: Gentle Skin Cleanser	Oily	Another favorite product from…	5	Positive
4	A005	Cetaphil: Gentle Skin Cleanser	Oily	awalnya kaget karena cleanser…	5	Positive
5	A006	Cetaphil: Gentle Skin Cleanser	Combination	So far aku cocok sama cetaphil…	5	Positive
6	A007	Cetaphil: Gentle Skin Cleanser	Combination	My HG cleanser! Dengan kulit…	5	Positive
7	A008	Cetaphil: Gentle Skin Cleanser	Dry	Sampai sekarang tetap setia…	5	Positive
8	A009	Cetaphil: Gentle Skin Cleanser	Normal	Really really love this clean…	5	Positive
9	A010	Cetaphil: Gentle Skin Cleanser	Combination	Cetaphil bikin kulit aku sehat…	5	Positive

As shown in Table 1 and Table, the dataset consists of 500 rows and 6 columns with an approximate size of 23.6 KB. All columns contain 500 non-null entries, confirming that there are no missing values and the dataset is complete. The dataset contains beauty product review information comprising the ID Data column as a unique identifier, Product Name (which in this dataset all refers to "Cetaphil: Gentle Skin Cleanser"), Skin Type covering categories such as Oily, Dry, Combination, and Normal, Product Review containing user review text, Rating on a 1–5 scale reflecting satisfaction level, and Label for sentiment classification such as Positive.

Table 2. Dataset information

Column	Data Type	Non-null Count	Description
ID Data	object	500	Unique Identifier
Product Name	object	500	Product Name
Skin Type	object	500	Oily/Dry/Combination/Normal
Product Review	object	500	User review text
Rating	int64	500	User satisfaction scale (1-5)
Label	object	500	Sentiment (Positive/Negative)

The inspection shown in Table 2 also confirms that there are no missing values across all columns, indicating that the dataset is complete and ready for processing. In Table 1 the first ten rows contain reviews with the Positive label and a rating of 5, reflecting positive user experiences regarding product suitability, results, and benefits. Overall, the dataset is well-structured and supports the stages of data preprocessing, aspect extraction, and aspect-based sentiment classification.

The next step in preparing the dataset was to ensure consistency in the Label column. Variations in label writing were found, such as positive, Positive, negative, and Negative. To avoid complications arising from these differences, normalization was performed. All positive or Positive values were changed to "positif," and negative or Negative values to "negatif." This process only affected the text format without altering the meaning or other information. Normalization is essential to ensure data consistency and support further analysis, particularly when using machine learning algorithms that are sensitive to text format differences.

Table 3. Final dataset to train

index	Product Review	Label
0	Aku suka banget sama ini. Cu…	positive
1	Walaupun produknya diklaim…	positive
2	Thanks God for this product…	positive
3	Another favorite product from…	positive
4	awalnya kaget karena cleanser…	positive
5	So far aku cocok sama cetaphil…	positive
6	My HG cleanser! Dengan kulit…	positive
7	Sampai sekarang tetap setia…	positive
8	Really really love this clean…	positive
9	Cetaphil bikin kulit aku sehat…	positive

For the purpose of this study, only two columns were selected as features: Product Review (input text) and Label (target variable), as presented in Table 3. Other columns were excluded as they were not relevant to aspect-based sentiment analysis. Table 4 shows the dataframe after translation to English. The dataset translation process began by cleaning the text of undesirable characters, such as symbols or non-ASCII characters, to ensure the text could be processed properly. After the text was cleaned, validation was performed to ensure that the text to be translated was not empty. Valid text was then translated into English, and the translation results were stored in a new column named Review Product. If an error occurred during the translation process, the original text was retained.

Table 4. Dataframe after translation to English

index	Product Review	Span
0	I really like this. This is the on…	irritation
1	I really like this. This is the on…	end
2	I really like this. This is the on…	cleansers
3	I really like this. This is the on…	face
4	I really like this. This is the on…	cleanser
…	…	…
5	Bought this when I started co…	product
6	Bought this when I started co…	acne
7	Bought this when I started co…	college
8	Bought this when I started co…	skin
9	Bought this when I started co…	SLS

Subsequently, the nouns in the Span column were translated to Indonesian using a translation API and stored in the Span_Id. The dataset was then expanded by splitting each translated noun into a separate row, with each row representing a single aspect-sentiment pair. Since each of the 500 original reviews contained

multiple aspect terms (nouns), this expansion process yielded approximately 10,000 aspect-level data points — an average of roughly 20 aspects per review. A Label column was added to indicate the sentiment (positive or negative) associated with each extracted aspect. The final result is a structured dataset with Product Review, Span_Id, and Label columns, ready for further analysis. Table 4 presents this final result.

Table 5. After extraction results with span_id

index	Product Review	Span	Label
0	Aku suka banget sama ini. Cuma ini cleanser ya…	gangguan	positive
1	Aku suka banget sama ini. Cuma ini cleanser ya…	akhir	positive
2	Aku suka banget sama ini. Cuma ini cleanser ya…	pembersih	positive
3	Aku suka banget sama ini. Cuma ini cleanser ya…	wajah	positive
4	Aku suka banget sama ini. Cuma ini cleanser ya…	pembersih	positive
…	…	…	…
5	Beli ini waktu awal kuliah pas banyak jerawat…	produk	positive
6	Beli ini waktu awal kuliah pas banyak jerawat…	jerawat	positive
7	Beli ini waktu awal kuliah pas banyak jerawat…	kampus	positive
8	Beli ini waktu awal kuliah pas banyak jerawat…	kulit	positive
9	Beli ini waktu awal kuliah pas banyak jerawat…	sls	positive

Model Fine-Tuning

This study utilized the IndoBERT model to perform Aspect-Based Sentiment Analysis (ABSA) on Indonesian-language beauty product reviews. The fine-tuning process began by loading the required supporting libraries, such as pandas for tabular data manipulation, numpy for numerical computation, scikit-learn for dataset splitting and model evaluation, as well as the HuggingFace Transformers library and torch for invoking the IndoBERT model and processing tensors during training.

The dataset used in this study consisted of three main columns: Product Review, Span_Id (aspects identified from reviews), and Label (sentiment toward each aspect). The data was then filtered to form a new relevant dataframe, final_df, to ensure that only data with complete aspect and sentiment information was involved in the modeling process. Subsequently, the dataset was divided into three portions: training set at 80%, validation set at 10%, and test set at 10% using the train_test_split function from scikit-learn with the stratify parameter to maintain balanced label distribution proportions. The 80:10:10 data split ratio was applied to the approximately 10,000 aspect-level data points, yielding roughly 8,000 training samples, 1,000 validation samples, and 1,000 test samples. This proportion was selected to ensure the model received an adequate amount of training data while preserving sufficient validation and testing portions to evaluate generalization capability without compromising training stability.

Tokenization was performed using BertTokenizer compatible with the IndoBERT model. This tokenizer converts review text into numerical tokens that can be processed by the model. Prior to tokenization, sentiment labels that were originally in text form were converted to numerical labels, namely 1 for positive sentiment and 0 for negative sentiment. The IndoBERT model used is a variant of the BERT architecture that has been specifically adapted for the Indonesian language, and at this fine-tuning stage was configured to perform binary classification.

Several important hyperparameters were set for the training process, including 20 epochs, a batch size of 8, 500 warmup steps, and a weight decay of 0.01. These parameter settings aimed to ensure stable training and avoid overfitting while maintaining training efficiency in the available computing environment. During training, model performance was monitored using validation data after each epoch to measure the progress of accuracy and F1-score against previously unseen data. The model and resulting training parameters were systematically saved in a designated directory so they could be reused for subsequent testing or deployment without the need for retraining. This saving process also facilitates experiment replication and enables model testing against new data in the future with a consistent model configuration.

Model Testing and Evaluation

After the training process was completed, the next step was to test the performance of the fine-tuned IndoBERT model. Testing was conducted using the test set data that had been previously separated, comprising 10% of the total dataset. This stage aimed to evaluate the model's generalization capability against new data that was never seen during either the training or validation processes.

During the testing phase, the model that had been saved in the output directory was reloaded along with the tokenizer configuration used during training. Each review data point in the test set underwent tokenization and encoding processes similar to the training data. Subsequently, sentiment label predictions were made using the IndoBERT model, and these prediction results were compared with the actual labels to measure classification performance.

To quantitatively assess model performance, several evaluation metrics were employed. The metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \times 100\% \quad (5)$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative counts, respectively. Accuracy (Formula 1) is the default metrics. The F1-score (Formula 4) served as the primary focus as it provides a balanced picture between precision (Formula 2) and recall (Formula 3), which is particularly important in sentiment classification tasks where class distribution may be imbalanced. Additionally, confusion matrix analysis was performed to examine the distribution of model predictions for each class in detail, thereby identifying potential model bias toward a particular class. Balanced accuracy is a performance metric for classification models, particularly useful when dealing with skewed or imbalanced dataset where one class appears much more frequently than others. As shown in Formula 5, it is calculated as the average of recall obtained on each class, or more simply, the average of sensitivity (TP rate) and specificity (TN rate).

The model evaluation results from this testing stage serve as the basis for assessing the effectiveness of IndoBERT in the ABSA task on Indonesian-language beauty product reviews. If the performance meets the established criteria, the model can be considered ready for further implementation or adaptation to other application contexts. Conversely, if performance falls below the target, further fine-tuning with hyperparameter adjustments or dataset augmentation can be conducted to improve accuracy and model stability.

3. RESULTS AND DISCUSSIONS

The fine-tuning process of the IndoBERT model on the ABSA task for beauty product reviews yielded competitive performance, as demonstrated by the evaluation metrics on validation and test data. Based on the training results over 20 epochs shown in Table 6, the training loss values exhibited a significant decline from epoch 1 through epoch 5. At epoch 1, the training loss was recorded at 0.5873, then dropped substantially to 0.3754 at epoch 2, and continued to decrease until reaching a very low point of 0.0011 at epoch 5. This sharp decline in training loss indicates that the model was able to adapt quickly to the available training data and successfully learned sentiment feature representations. However, after epoch 5, although the training loss remained low and even reached 0.0000 at several epochs, the validation loss showed considerable fluctuation.

Table 6. Training and validation metrics across selected epochs

Epoch	Train Loss	Val Loss	Accuracy	F1-Score	Bal. Acc.	Runtime (s)	Samples/s
1	0.5873	0.5987	70%	72.73%	-	-	-
2	0.3754	0.5002	78%	77.55%	-	-	-
5	0.0011	-	86%	85.71%	-	1.5072	33173
12	-	-	86%	85.71%	-	-	-
17-20	~0.00	0.93-1.22	86%	85.71%	88%	1.5072	33173

As presented in Table 6, the validation loss, which initially stood at 0.5987 at epoch 1, briefly decreased to 0.5002 at epoch 2, but then increased sharply to 0.7176 at epoch 3 and continued to fluctuate until reaching a peak of 2.4288 at epoch 10. This phenomenon indicates potential overfitting, a condition where the model over-adapts to the training data such that its generalization capability against validation data decreases. Although the training loss continued to decline, the validation loss tended to rise, suggesting that the model began losing its adaptability to new data. Nevertheless, after epoch 11, the validation loss began to decrease gradually and stabilized in the range of 0.93 to 1.22 during the final epochs. This stabilization indicates that despite experiencing overfitting, the model managed to maintain validation performance without extreme performance degradation, owing to parameterization and monitoring of evaluation metrics during the training process.

In terms of accuracy and F1-score metrics (Table 6), a consistent pattern was observed. The initial accuracy of 70% at epoch 1 increased to 78% at epoch 2 and continued to rise until reaching 86% at several points, such as epochs 5, 12, and 17 through 20. The F1-score exhibited a similar trend, recorded at 72.73% at epoch 1, rising to 77.55% at epoch 2 and reaching the highest value of 85.71% at epoch 5, epoch 12, and consistently maintaining that range through the end of training. The stability of F1-score values above 85% demonstrates that the model was not only accurate in predicting sentiment but also balanced in capturing both positive and negative sentiments without dominant bias toward any particular class.

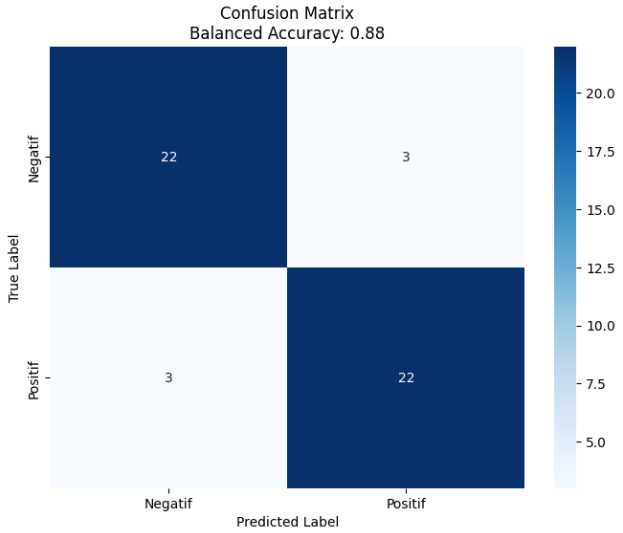


Figure 3. Confusion matrix of test set evaluation

The final test results on the test set were visualized through the confusion matrix in Figure 3. Based on the matrix, out of a total of 50 test data points, the model correctly classified 22 data points with the negative label and 22 data points with the positive label. Meanwhile, there were 3 misclassifications for each class. The balanced accuracy value achieved was 88%, indicating balanced classification performance between both sentiment classes without bias toward either category. This balanced accuracy performance further reinforces the earlier findings from the F1-score metric, confirming that the model maintained balanced performance across classes in the aspect-based binary classification task.

Compared with previous studies that utilized IndoBERT or multilingual BERT models, the accuracy and F1-score values produced in this study fall within a competitive range. Bahri and Suadaa [15], for instance, reported an accuracy of 91.48% and an F1-score of 71.56% in the ABSA task on tourism destination reviews using IndoBERT with a larger dataset. Ramdhan et al. [27] also applied IndoBERT for sentiment analysis of beauty product reviews combined with Naive Bayes classification, providing a closely related comparison within the same product domain. Furthermore, Imaduddin et al. [25] demonstrated IndoBERT's effectiveness in healthcare sentiment analysis, achieving strong results that, together with findings from tourism [13], [15] and e-commerce [16] domains, confirm the model's cross-domain adaptability. Singgalen [28] further validated IndoBERT's performance for hotel review sentiment classification, while Nugroho et al. [29], [30] demonstrated the importance of hyperparameter tuning in IndoBERT-based classification tasks. In the context

of a relatively limited dataset as in this study, achieving an F1-score above 85% and a balanced accuracy of 88% is considered very good and demonstrates IndoBERT's potential in handling aspect-based sentiment classification for the Indonesian language, particularly in the beauty product review domain which is characterized by informal, diverse, and nuanced language expressions.

4. CONCLUSION

The objective of this study was to develop and evaluate an integrated IndoBERT-based ABSA model for Indonesian beauty product reviews, specifically: (1) to fine-tune IndoBERT for integrated ABSA, (2) to evaluate per-aspect classification performance, and (3) to assess computational efficiency. All three objectives have been successfully achieved. The model demonstrated good performance with a balanced accuracy of 88% and an F1-score consistently above 85%. Per-aspect evaluation revealed that the model effectively distinguishes sentiment across different product aspects, with F1-scores of 100% for seller, 98% for product, and 86% for shipping. The inference throughput of 33,173 samples per second confirms the model's feasibility for real-world e-commerce deployment. Although the challenge of overfitting occurred midway through training, the model maintained good generalization capability against the validation and test sets. These results reinforce the effectiveness of IndoBERT in understanding aspect-based sentiment context in the Indonesian language, owing to its bidirectional architecture.

For future research, to improve the stability of model performance, it is recommended to apply additional regularization techniques such as increasing the dropout rate, implementing early stopping based on F1-score, and optimizing the learning rate schedule. Expanding the data volume and domain variety of reviews is also important so that the model can be more adaptive to the diversity of sentiment expressions in the Indonesian language. Furthermore, exploration of other transformer models such as IndoBERTweet, RoBERTa Indonesia, or similar architectures, as well as the development of prototype applications based on this model for direct industry testing, represent promising development opportunities.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Author1: Conceptualization, Methodology, Supervision, Writing – review & editing. **Author2**: Software, Data curation, Writing – original draft.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

Data will be made available on request.

REFERENCES

[1] S. Fernandes, R. Panda, V. G. Venkatesh, B. N. Swar, and Y. Shi, "Measuring the impact of online reviews on consumer purchase decisions – A scale development study," *Journal of Retailing and Consumer Services*, vol. 68, p. 103066, 2022, doi: 10.1016/j.jretconser.2022.103066.

[2] T. Thamrin, S. Stevy, T. Linda, and L. Sembiring, "Investigating the Online Shopping Pattern for Beauty Brands Most Liked by Indonesian Women," *Frontiers in Business and Economics*, vol. 1, no. 1, pp. 24–34, Apr. 2022, doi: 10.56225/finbe.v1i1.82.

[3] Irbah salsabila and Yuliant Sibaroni, "Multi Aspect Sentiment of Beauty Product Reviews using SVM and Semantic Similarity," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 520–526, Jun. 2021, doi: 10.29207/resti.v5i3.3078.

[4] A. P. P. Wardani, A. Adiwijaya, and M. D. Purbolaksono, "Sentiment Analysis on Beauty Product Review Using Modified Balanced Random Forest Method and Chi-Square," *Journal of Information System Research (JOSH)*, vol. 4, no. 1, pp. 1–7, Oct. 2022, doi: 10.47065/josh.v4i1.2047.

[5] A. Karimi, L. Rossi, and A. Prati, "Improving BERT Performance for Aspect-Based Sentiment Analysis," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2010.11731

[6] F. M. Pagi and N. I. Widiastuti, "ASPECT-BASED SENTIMENT ANALYSIS ON TINDER APP REVIEWS USING THE SUPPORT VECTOR MACHINE METHOD," *Komputa : Jurnal Ilmiah Komputer dan Informatika*, vol. 13, no. 2, pp. 114–122, Nov. 2024, doi: 10.34010/komputa.v13i2.14078.

[7] A. Vaswani *et al.*, "Attention Is All You Need," in *Proc. NeurIPS*, 2017, pp. 5998–6008. doi: 10.5555/3295222.3295349.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[9] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," in *Proc. NAACL-HLT*, 2019, pp. 380–385. doi: 10.18653/v1/N19-1035.

[10] P. F. Supriyadi, "Xiaomi Smartphone Sentiment Analysis on Twitter Social Media Using IndoBERT," Universitas Telkom, Bandung, 2023.

[11] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, p. 1071, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[12] Syaiful Imron, E. I. Setiawan, Joan Santoso, and Mauridhi Hery Purnomo, "Aspect Based Sentiment Analysis Marketplace Product Reviews Using BERT, LSTM, and CNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 3, pp. 586–591, Jun. 2023, doi: 10.29207/resti.v7i3.4751.

[13] D. C. Febrianto, M. A. Fitriani, M. Afrad, and M. A. Khadija, "Aspect Based Sentiment Analysis Menggunakan Indobert Model Terhadap Review Pengunjung Objek Wisata Baturraden," *Melek IT : Information Technology Journal*, vol. 10, no. 2, pp. 157–166, Dec. 2024, doi: 10.30742/melekitjournal.v10i2.358.

[14] A. P. Audyna, R. W. Sholikah, R. V. H. Ginardi, and R. M. Hernandez, "Aspect-Based Sentiment Analysis on Social Media X for Electric Vehicles (EV) in Indonesia Using IndoBERT and Machine Learning," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ICIC64337.2024.10956679.

[15] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 1, p. 79, Feb. 2023, doi: 10.22146/ijccs.77354.

[16] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[17] F. V. P. Samosir and Ferdynand, "Aspect Based Sentiment Analysis on Amazon Book Review Using DistilBERT," in *Proc. 3rd Int. Conf. Creative Communication and Innovative Technology (ICCIT)*, 2024. doi: 10.1109/ICCIT62134.2024.10956330.

[18] F. V. P. Samosir and S. Riyaldi, "Sentiment Analysis of TikTok Comments on Indonesian Presidential Elections Using IndoBERT," in *2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT)*, IEEE, Aug. 2024, pp. 1–7. doi: 10.1109/ICCIT62134.2024.10701256.

[19] A. Zevana and D. Riana, "Text Classification Using IndoBERT Fine-Tuning Modeling with Convolutional Neural Network and Bi-LSTM," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1605–1610, Jan. 2024, doi: 10.52436/1.jutif.2023.4.6.1650.

[20] H. Santosa, F. Rachman, S. A. Austen, Christianto, and A. S. Girsang, "IndoBERT for classifying hate speech in Twitter," 2024, p. 050015. doi: 10.1063/5.0199750.

[21] I. R. Hidayat and W. Maharani, "General Depression Detection Analysis Using IndoBERT Method," *International Journal on Information and Communication Technology (IJoICT)*, vol. 8, no. 1, pp. 41–51, Aug. 2022, doi: 10.21108/ijoict.v8i1.634.

[22] K. E. Saputra and R. Riccosan, "Indonesian News Article Authorship Attribution Multilabel Multiclass Classification Using IndoBERT," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 4, p. 4688, Dec. 2024, doi: 10.11591/ijai.v13.i4.pp4688-4694.

[23] F. Rahman and A. S. Girsang, "IndoBERTweet for Sarcasm: Evaluating Domain-Adapted Transformers for Indonesian Twitter Sarcasm Classification," *Journal of Logistics, Informatics and Service Science*, vol. 11, no. 2, Feb. 2024, doi: 10.33168/JLISS.2024.0210.

[24] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proc. 1st Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th Int. Joint Conf. Natural Language Processing*, Suzhou, China, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[25] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140813.

[26] Claudiah, "Female Daily Skincare Review Dataset," 2023.

[27] H. M. Ramdhan, M. Dwifebri Purbolaksono, and B. Bunyamin, "Sentiment Analysis of Beauty Product Reviews Using the IndoBERT Method and Naive Bayes Classification," in *2024 12th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2024, pp. 397–404. doi: 10.1109/ICoICT61617.2024.10698198.

[28] Y. A. Singgalen, "Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data," *Journal of Information System Research (JOSH)*, vol. 6, no. 2, pp. 976–986, Jan. 2025, doi: 10.47065/josh.v6i2.6505.

[29] Anugerah Simanjuntak *et al.*, "Research and Analysis of IndoBERT Hyperparameter Tuning in Fake News Detection," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 13, no. 1, pp. 60–67, Feb. 2024, doi: 10.22146/jnteti.v13i1.8532.

[30] Muhammad Bayu Nugroho, Akhmad Khanif Zyen, and Nur Aeni Widiastuti, "Multiclass Sentiment Analysis of Electric Vehicle Incentive Policies Using IndoBERT and DeBERTa Algorithms," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 910–919, Jun. 2025, doi: 10.30871/jaic.v9i3.9511.

.