Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

Comparison of RoBERTa and IndoBERT on Multi-Aspect Sentiment Analysis of Indonesian Hotel Reviews with Tuning Optimization

Rizky Ahsan Syarif*, Yuliant Sibaroni

School of Computing, Telkom University, Bandung, Indonesia
Email: [1,*]mrizkysyarif@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id
Correspondence Author Email: mrizkysyarif@telkomuniversity.ac.id
Submitted: **05/06/2025**; Accepted: **01/09/2025**; Published**: 02/09/2025**

Abstract−The hospitality industry heavily relies on online reviews as a crucial source of information that influences potential guests' decisions. However, conducting sentiment analysis on hotel reviews can be challenging due to the complexity of language and contextual diversity, especially in Indonesian. This study aims to develop and optimize a RoBERTa-based sentiment analysis model to improve the accuracy of sentiment classification in Indonesian hotel reviews, focusing on the aspects of facilities, cleanliness, location, price, and service. The methodology includes data collection through web scraping from the Traveloka platform, manual labeling, and text pre-processing. The RoBERTa model was trained and optimized using fine-tuning techniques and evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC. The results show that the optimized RoBERTa model achieves competitive performance, although the IndoBERT model with Bayesian Optimization demonstrates superior performance, particularly in terms of accuracy and efficiency in identifying positive and negative sentiments. This study is expected to contribute to the development of more effective and accurate aspect-based sentiment analysis (ABSA) for Indonesian-language hotel reviews. It also opens opportunities for applying NLP technology in the hospitality industry and across other review platforms, thereby improving sentiment analysis quality and assisting hotel managers in enhancing service and customer experience.

Keywords: Aspect-Based; Sentiment Analysis; RoBERTa; IndoBERT; Fine-Tuning

1. INTRODUCTION

In today's rapidly evolving digital era, online platforms like Traveloka have become essential resources for consumers to gather information and share experiences related to hospitality services. User reviews published on these platforms offer valuable insights into various aspects of hotels, such as their facilities, cleanliness, location, pricing, and service quality. However, the sheer volume and diversity of these reviews demand an effective analytical approach to accurately capture consumer sentiments and preferences. This is where Aspect-Based Sentiment Analysis (ABSA) becomes crucial, as it enables a deeper understanding of user opinions [1].

ABSA facilitates the identification of specific sentiments related to various aspects of hotel Serviees, offering a more comprehensive understanding of the customer experience. However, the use of pre-trained language models, such as IndoBERT and RoBERTa, in the Indonesian context presents several challenges. The differences in language structure and cultural nuances between Indonesian and English can impact the effectiveness of these models. Additionally, the class imbalance in the dataset, where certain sentiment categories are more prevalent, can introduce bias in the model's predictions. Consequently, it is crucial to adjust hyperparameters appropriately to enhance the model's performance in the ABSA task when applied to hotel reviews in Indonesia [2].

Previous research has shown that IndoBERT, as a variant of BERT specifically trained for Indonesian, has significant potential in ABSA tasks [1]. The customized IndoBERT model achieved an overall accuracy of 92.52%, with an F1 score for the positive class reaching 96.09%. The main advantage of IndoBERT lies in its superior ability to understand the context of the Indonesian language and filter sentiment more accurately than other models that are not specifically trained for this language. Although the model's performance on neutral and negative classes is still suboptimal, this is more influenced by the class imbalance challenge in the hotel review dataset [3]. On the other hand, although RoBERTa, a BERT extension with improved pre-training optimization, has demonstrated superior performance on various NLP tasks, this model has not been specifically applied to ABSA on Indonesian hotel reviews. This study aims to further explore the comparison between IndoBERT and RoBERTa in the context of ABSA for Indonesian hotel reviews, with a focus on hyperparameter optimization and tuning.

On the other hand, RoBERTa, an extension of BERT with enhanced pre-training optimization, has demonstrated superior performance across various NLP tasks. This model has been applied to emotion detection in Indonesian text, with hyperparameter tuning using the Bayesian Optimization method, achieving an accuracy of 83.64% [2]. However, this research has not specifically explored the application of RoBERTa in ABSA for Indonesian-language hotel reviews.

Moreover, several studies have suggested the integration of ABSA with a Zero-Shot Learning approach to efficiently analyze traveler reviews without requiring large annotated datasets. This method enables models to classify data without the need for specialized training, offering a potential solution to the data limitations often encountered in the context of ABSA for hotel reviews [4].

Hyperparameter tuning, especially the learning rate, plays a crucial role in optimizing the performance of the IndoBERT model. Previous studies have demonstrated that a learning rate of 2E-5 yields the best results in classifying Indonesian exam questions, achieving a validation accuracy of 97% and an F1 score of 97%. This suggests a strong potential for enhancing model performance in Aspect-Based Sentiment Analysis (ABSA) tasks on hotel reviews [5].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

These results underscore the significance of hyperparameter tuning in enhancing the performance of models in text classification tasks [2].

Sentiment analysis of Traveloka customer reviews was performed using SVM, Logistic Regression, and Naïve Bayes models. Although this study did not employ a transformer model, the findings offer valuable insights into customer sentiment towards Traveloka's services. These results can serve as a reference for understanding the context of hotel reviews on the platform [6].

Based on the current literature review, it is essential to compare the performance of IndoBERT and RoBERTa in the aspect-based sentiment analysis (ABSA) task for hotel reviews in Indonesia, considering appropriate hyperparameter adjustments. Enhancing the IndoBERT model for aspect-based sentiment analysis of Indonesian tourism-related user content could improve its performance in detecting sentiment for specific aspects, although challenges such as class imbalance and language variation still persist [7]. Furthermore, integrating Zero-Shot Learning into ABSA can enhance the model's effectiveness, even in the absence of extensive annotated data. This is especially relevant for hotel reviews in Indonesia, where data is often abundant yet imbalanced. By applying suitable hyperparameter adjustments, the model is expected to perform more effectively in extracting sentiment from hotel reviews on platforms like Traveloka. This research aims to address the gap in existing literature and contribute to the development of a more efficient ABSA model for analyzing Indonesian-language hotel reviews [4].

This study aims to make a substantial contribution to the advancement of a more effective ABSA model for analyzing Indonesian-language hotel reviews. With the growing popularity of platforms like Traveloka, conducting in-depth sentiment analysis of various aspects of hotels has become increasingly essential to better understand customer needs and preferences. Proper hyperparameter tuning can enhance model performance, enabling more accurate sentiment identification, particularly in cases involving ambiguous or unstructured text [5]. Furthermore, optimizing models like IndoBERT and RoBERTa is expected to improve sentiment recognition on imbalanced review data, as previous studies have shown that class imbalance remains a significant challenge in sentiment analysis [3]. Consequently, this research aims to address gaps in current ABSA methods and lay the foundation for more efficient and precise applications in hotel review analysis in Indonesia.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research introduces a system designed to analyze multi-dimensional sentiment in hotel reviews from Indonesia, utilizing optimization techniques to enhance model performance. The dataset, sourced from the Traveloka platform, is processed with the RoBERTa model, which undergoes fine-tuning to identify the most effective approach for optimal performance. The process involves several key stages, including data collection, selection, labeling, text preprocessing, tokenization, data splitting, model training, and evaluation. Additionally, various evaluation metrics are used to assess the performance of the optimized model in accurately capturing sentiment nuances. The Research Stages in this study can be seen in Figure 1.

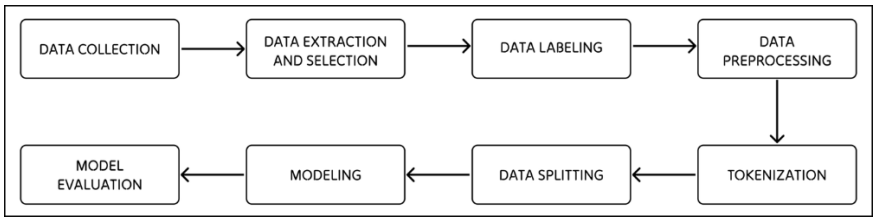

Figure 1. Research Stages

2.2 Data Collection

In the initial phase of this research, data collection was performed to gather the necessary dataset. This process employed web scraping techniques, which allow for the automatic extraction of data from websites, eliminating the need for manual data copying. The collected data consists of hotel reviews written in Indonesian from hotels located in Bandung City [8].

Web scraping is a method used to automatically retrieve information from websites without the need for manual copying. The primary goal is to convert web data into a structured format. This technique offers several advantages, including the ability to collect web data accurately and quickly. Additionally, it makes the data easy to analyze, minimizing the time spent searching for information [9].

2.3 Extraction and Aspect Selection

Aspect extraction in sentiment analysis involves identifying key terms that represent significant elements within the text. Recent approaches leverage natural language processing (NLP) techniques to automatically select relevant keywords, considering factors such as context, word frequency, and the relationships between words within the text being analyzed. Research has shown that combining statistical feature extraction methods with semantic

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

understanding enhances the accuracy of identifying the most important aspects, directly influencing the outcomes of sentiment analysis. This method allows for the discovery of more precise and pertinent aspects in the analyzed text [9].

Data selection plays a crucial role in data analysis. It involves choosing the most relevant data groups that provide the most valuable information from the available dataset [10]. For this study, the selected data groups include five main aspects: Facility, Cleanliness, Location, Price, and Service. These aspects were chosen for their significant contribution to the overall evaluation of hotel service quality.

2.4 Data Labeling

Data labeling is critical for developing accurate and reliable machine learning models, particularly when handling complex or domain-specific data that requires expert interpretation. Despite its time-consuming and costly nature, manual labeling remains the gold standard for ensuring high data quality. To optimize this process, integrating essential data management practices such as cleaning, standardization, and version control is crucial [12]. While semi-supervised and automated labeling techniques can significantly boost efficiency, manual labeling remains the benchmark, relying on professional expertise to minimize bias and enhance accuracy. Although manual labeling presents significant challenges in terms of time and costs, it remains indispensable for generating high-quality data necessary for training dependable machine learning models [13]. Furthermore, investing in manual labeling can lead to better-performing models and more accurate predictions in the long run.

2.5 Data Preprocessing

Data preprocessing plays a crucial role in data analysis and machine learning. It involves the transformation and cleaning of raw text into a structured format that is more suitable for analysis. This process not only enhances the quality and precision of the text but also ensures that it is properly formatted for efficient use in machine learning models [11]. The following stages are involved in text preprocessing:

2.5.1 Data Cleaning

Data cleaning involves the identification and removal of missing, invalid, or inconsistent values within a dataset, ensuring that the data utilized in the analysis is of high quality and reliable [12].

2.5.2 Case Folding

Case folding is a text preprocessing technique that transforms all uppercase characters into lowercase. During the case folding process, only accepted characters—specifically, 'a' to 'z'—are retained, while other characters are removed and treated as separators [13].

2.6 Tokenization

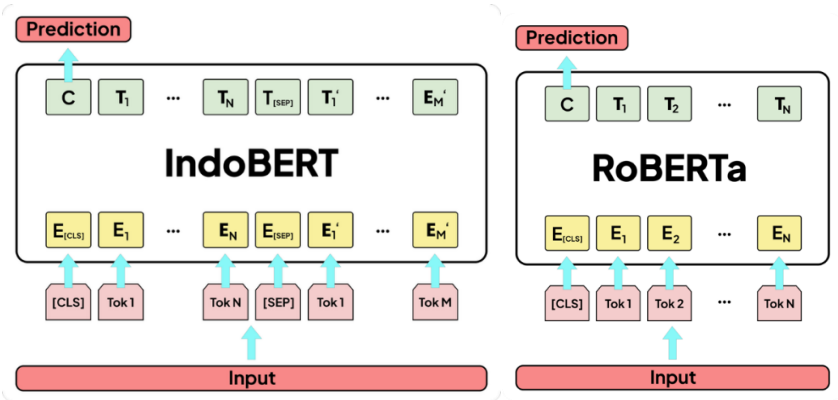

Figure 2. IndoBERT and RoBERTa Architecture

Based on Figure 2, the IndoBERT model begins tokenization using WordPiece, adding [CLS] tokens at the start and [SEP] tokens between sentences. It then converts each token into a vector by summing token embeddings, segment embeddings, and position embeddings, which are processed by the transformer layers to produce a contextual representation. RoBERTa, an extension of IndoBERT, employs Byte-Pair Encoding (BPE) and does not require [SEP] tokens. This makes RoBERTa more flexible when dealing with longer texts. Although the method for generating vectors remains the same as in BERT, RoBERTa is trained on a larger dataset and longer text sequences, which enhances its performance in various NLP tasks [14].

RoBERTa's tokenization approach offers notable advantages in handling long texts, making it more efficient for processing extended text sequences. In contrast to IndoBERT, RoBERTa is trained with a broader dataset and longer sequences, resulting in improved performance across a wide range of natural language processing tasks. These enhancements enable RoBERTa to better manage complex contexts and applications requiring the processing of large volumes of text, solidifying its position as a more powerful model in the field of natural language processing [14].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

2.7 Data Splitting

The process of dividing data into training and testing datasets is essential for evaluating the performance of the developed model on data that it has not previously encountered [15]. This division is crucial to ensure that the model not only memorizes patterns from the training data but also generalizes well, making accurate predictions on new, unseen data.

One common technique for splitting the data is K-Fold Cross-Validation. In this method, the input dataset is divided into K equally sized groups, known as folds. Each fold is used in turn as the testing dataset, while the remaining folds are used to train the model. This approach provides a more accurate evaluation of the model's performance across different subsets of the data, offering a comprehensive view of its ability to handle various data types [16].

2.8 Modeling

At the modeling stage, fine-tuning plays a critical role in refining a pre-trained machine learning model. This step involves adjusting the model's weights using new data that is more relevant and specific to the intended task. On the other hand, hyperparameter tuning aims to determine the optimal values for parameters that were not learned during the initial training phase, such as learning rate and batch size. Both of these processes are essential to ensure that the model not only performs well on the training data but also has the ability to generalize to unseen data. Recent studies have shown that hyperparameter tuning methods, including grid search and Bayesian optimization, significantly contribute to enhancing both the performance and efficiency of models [17].

In addition to fine-tuning and hyperparameter tuning, model comparison is also a crucial step in identifying the most suitable model for a given task. A recent study compared the RoBERTa (cahya/roberta-base-indonesian-522M) model with IndoBERT (indolem/indobert-base-uncased). Research on RoBERTa optimization through hyperparameter tuning for emotion detection revealed that the optimized RoBERTa model achieved an accuracy of 83.64% in detecting emotions in Indonesian text. While the IndoBERT model, which is specifically trained with Indonesian data, is known for its advantages, this paragraph only references studies related to RoBERTa [18].

2.9 Model Evaluation

In model evaluation, one of the primary tools used is the confusion matrix. It provides a summary of the model's prediction results by comparing the actual values with the predictions made on the test data. Typically, the confusion matrix is presented in a table format that displays the count of correct and incorrect predictions for each class. This matrix is essential for identifying model errors and understanding where the model's performance can be improved. By analyzing the confusion matrix, valuable insights can be gained about the model's accuracy and areas that need further refinement.

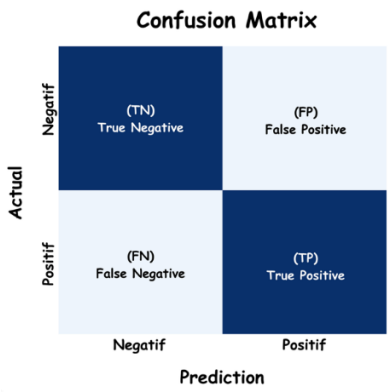

Figure 3. Confusion Matrix Graph

Based on the illustration in Figure 3, the confusion matrix consists of four main components: True Positive (TP), which represents the number of positive class samples correctly predicted as positive; True Negative (TN), which refers to the number of negative class samples correctly predicted as negative; False Positive (FP), which denotes the number of negative class samples incorrectly predicted as positive; and False Negative (FN), which indicates the number of positive class samples incorrectly predicted as negative. These four components serve as the foundation for calculating various evaluation metrics, such as accuracy, precision, recall, and F1-score [19]. By using the confusion matrix, we can determine how well the model distinguishes between different classes and understand the model's weaknesses in making predictions.

In model performance evaluation, key metrics include Accuracy, F1 Score, Precision, and Recall, which are used to assess the overall effectiveness of the model. Below are the formulas for calculating the model's performance evaluation metrics:

a. Accuracy represents the ratio of correct predictions to the total number of data points [20].

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

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

b. Precision measures how accurate the positive predictions made by the model are, compared to all the positive predictions made [20].

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

c. Recall indicates the model's ability to identify all positive samples in the data [21].

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

d. F1 Score is the harmonic mean of Precision and Recall, used to evaluate the balance between these two metrics [21].

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

3. RESULTS AND DISCUSSION

3.1 Data Collection

The data collection for this research was conducted using the web scraping technique with the Web Scraper extension in Google Chrome to extract hotel reviews and ratings from the Traveloka platform for hotels in Bandung. This automated method successfully gathered 305,875 reviews from various hotels, significantly reducing the time spent compared to manual data collection. Web scraping has proven to be an efficient and accurate approach for large-scale data collection and can be updated regularly [25]. In addition, reviews are collected randomly at specific intervals to avoid bias associated with specific periods and ensure a more balanced representation of consumer sentiment over time. The ratings collected were used to assess the sentiment of each review, adding valuable context to the analysis of customer opinions. This method also allows for continuous data extraction, ensuring the dataset remains up-to-date and relevant. Furthermore, web scraping enables the retrieval of diverse data from multiple sources, contributing to a more comprehensive analysis of consumer sentiment. The collected data is shown in Table 1.

Table 1. Hotel Reviews and Ratings

No	Reviews	Rating
1	Sangat menyenangkan menginap di hotel ini. Sangat rekomendasi untuk menginap lagi. Top deh.	10
2	menyenangkan menginap disini,dekat ke tempat yang hits di Bandung	8.8
3	Tempat yang sangat nyaman. Pemandangan bagus sekali. Bersih, ramah, indah.	8.5
...	...	...
305.874	Lumayan nyaman, cuma ada serangga di kamar mandi.	6.9
305.875	Hotel lama super gelap dan kotor 😫.	4.5

3.2 Extraction and Aspect Selection

Aspect extraction is carried out using a keyword matching technique to identify keywords associated with facilities, cleanliness, location, price, and service within the reviews. After identifying the relevant keywords, the reviews are categorized based on the similarity of the aspects for further analysis [9]. A single review may encompass up to five aspects, and each review is assigned to the appropriate group according to the aspects discussed. The results of aspect extraction are presented in Table 2.

Table 2. Aspect Extraction Results

No	Reviews	Rating	Hotel Aspects
1	Hotel paling oke lengkap fasilitasnya.	10	Fasilitas
2	Bersih, nyaman dan pelayanan sangat memuaskan	9.4	Kebersihan, Pelayanan
3	Lokasi strategis, Harga murah, Kamar nya Luas, cuma makanan nya kurang enak.	8.7	Fasilitas, Lokasi, Harga
4	Hotelnya nyaman, luas, bersih dan tempatnya strategis, pelayanannya ramah, oke banget	8,5	Fasilitas, Kebersihan, Lokasi, Pelayanan
5	Cuma menginap selama beberapa jam. Dengan harga yang cukup terjangkau and lokasi strategis, kamar sangat bersih dan nyaman, pelayanannya juga ramah	9.7	Fasilitas, Kebersihan, Lokasi, Harga, Pelayanan

3.3 Data Labeling

Manual labeling is performed by assigning two sentiment labels: positive and negative. These labels are applied to each aspect of the data [26]. To represent the sentiment for each aspect, five new columns are created. Each column corresponds to a different sentiment aspect. Positive values in these columns indicate positive sentiment, while

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

negative values represent negative sentiment. For example, the labeling results for the facility aspect are shown in Table 3.

Table 3. Data Labeling Results

No	Reviews	Rating	Sentiment
1	Overall fasilitas untuk family ok.	10	Positive
2	Hotel yang cocok banget buat anak kereta... Kebersihannya di tingkatkan lagi, soalnya dari aspek yang lain udah oke banget.	6.6	Positive
3	kamar terlalu kecil, untuk sholat kesulitan	5.7	Negative
4	Sprei bantal kamar mandi jorok, sarapan pagi kurang bervariasi. Cuma menginap selama beberapa jam.	4.5	Negative
5	Kamar bersih, nyaman, lokasi strategis	10	Positive

3.4 Data Preprocessing

Text preprocessing begins with removing special characters and URLs. Next, all text is converted to lowercase to ensure cleaner and more consistent data. This step plays an important role in reducing data dimensions, which improves overall sentiment analysis performance. The rating column, which was previously used as a sentiment indicator, is also removed as it is no longer needed after the data is labeled. Alphabetical characters and numbers are retained. The text preprocessing stages can be seen in Figure 4.

Table 4. Data Preprocessing Results

No	Reviews	Sentiment
1	makan paginya enak enak dan lengkap	Positive
2	lumyan lah sesuai harga cuma pelayanannya kurang sigap	Negative
3	pelayanan baik tapi makanan kurang enak	Positive
4	tidak ramah dan asking deposit terlalu mahal	Negative
5	murah ga serem oke lah tinggal resepsionisnya aja kurang ramah	Negative

3.5 Tokenization

During the tokenization process, the RoBERTa model is employed to transform text data into numerical vector representations by segmenting the text into tokens, which are subsequently converted into vector embeddings. Tokenization using RoBERTa has proven to be effective in generating distinct numerical representations for each token. The tokenization results using RoBERTa are presented in Table 5.

Table 5. Tokenization Results with RoBERTa

No	Review	Result	Sentiment
1	sangat rekomendasi hotel disini	[0, 19559, 4767, 6168, 13368, 2]	Positive
2	overall fasilitas untuk family ok	[0, 13831, 1415, 3628, 371, 41675, 9556, 2]	Positive
3	hotel dengan parkir yang sempit dan liftnya hanya satu	[0, 39397, 352, 10624, 292, 9579, 291, 23441, 327, 835, 543, 2]	Negative
4	kolam renangnya cocok buat ikan koi aja	[0, 47499, 12420, 327, 6517, 11090, 2882, 2857, 77, 13771, 2]	Negative
5	sprei bantal kamar mandi jorok sarapan pagi kurang bervariasi	[0, 1990, 27903, 41508, 6286, 10630, 348, 27186, 29029, 6157, 1769, 6637, 2]	Negative

3.6 Dataset Descriptive Statistics

Descriptive statistics is a crucial method in data analysis, used to systematically summarize and analyze the key characteristics of the data before progressing to further processing stages [1]. The application of descriptive statistics is essential for enhancing the validity and credibility of research, both during the initial exploratory phase and as a supportive tool for inferential analysis [1]. In this study, descriptive statistics are employed to illustrate the sentiment distribution (positive and negative) and the data quantity for each aspect, providing an overview of the data balance and the readiness of the dataset for training the aspect-based sentiment analysis model. This step ensures that any potential biases or imbalances in the dataset are identified early, allowing for necessary adjustments. Additionally, it helps in evaluating whether the dataset is sufficiently representative of the various aspects to be analyzed, thus ensuring the reliability of the subsequent analysis.

3.6.1 Data Distribution by Aspect

In addition to sentiment, the dataset is also categorized according to several aspects present in the reviews, including facilities, cleanliness, location, price, and service. The table below illustrates the data distribution for each aspect considered. These aspects are essential for a comprehensive analysis, as they provide valuable insights into the specific

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

factors influencing customer satisfaction. By segmenting the reviews into distinct categories, the study can identify which elements most significantly impact the overall sentiment. This approach enables a more focused understanding of the factors that shape customer experiences in hotel settings. The data distribution per aspect is shown in Table 6.

Table 6. Data Distribution by Aspect

Hotel Aspects	Total Data
Facilities	2.000
Cleanliness	2.000
Location	2.000
Price	2.000
Services	2.000
Total	10.000

The table above illustrates that the aspects of facility, cleanliness, location, price, and service are evenly distributed, with each aspect receiving 2,000 reviews, resulting in a total of 10,000 reviews in the dataset. There is no imbalance in the number of reviews for each aspect, and this distribution offers a clear understanding of the available data's proportion. As a result, the researcher can proceed to the next step without the need for additional preprocessing or data adjustments to address imbalances.

3.6.2 Sentiment Distribution

The review data is categorized into two primary groups: positive and negative. This analysis seeks to explore the sentiment trends within the user-submitted reviews. The distribution of sentiment per aspect is presented in Table 7.

Table 7. Sentiment Distribution Results

Hotel Aspect	Total Data	
	Positive	Negative
Facilities	1.000	1.000
Cleanliness	1.000	1.000
Location	1.305	695
Price	1.000	1.000
Services	1.000	1.000
Total	5.305	4.695

The sentiment distribution indicates that the review data is almost equally divided between positive and negative sentiments. However, there are 1305 positive sentiments related to the location aspect, primarily because there are no negative sentiments for the location aspect in the previously collected reviews. As a result, the model will predominantly receive positive feedback regarding this aspect.

3.7 Model Performance Analysis

In this study, RoBERTa and IndoBERT were used to analyze the sentiment of hotel reviews written in Indonesian. The evaluation was conducted across five key aspects: Facilities, Cleanliness, Location, Price, and Service. These aspects were selected based on their relevance to customer satisfaction. The model performance was assessed using several metrics, including accuracy, precision, recall, F1-score, and AUC. Below is the overall performance analysis of both models.

3.7.1 Model Evaluation Results

After training the model using the previously prepared hotel review dataset, the evaluation results for the RoBERTa and IndoBERT models, which were fine-tuned using Bayesian Optimization and Random Search methods, are presented in the table below. The evaluation includes metrics such as accuracy, precision, recall, F1-score, and AUC, along with the confusion matrix for both models. In this study, the data was split using 5-fold Cross Validation, where each fold used 80% of the data for training and 20% for testing. This method was chosen to ensure a more accurate model evaluation and minimize bias. The use of 5-fold Cross Validation demonstrates the effectiveness of this technique in evaluating the performance of classification models [27]. The comparison of the two models, RoBERTa and IndoBERT, highlights their strengths and weaknesses across different evaluation metrics, providing a clearer understanding of their suitability for the sentiment analysis task at hand. The details of these results will be explained in further depth. The results of the model evaluation are presented in Table 8.

Table 8. Model Evaluation Result

Aspect	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Facility	RoBERTa (Bayesian Optimization)	85.35%	85.35%	85.35%	85.35%	88.98%
	IndoBERT (Random Search)	95.65%	95.66%	95.65%	95.65%	96.90%

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

Aspect	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Cleanliness	RoBERTa (Bayesian Optimization)	95.15%	95.19%	95.15%	95.15%	96.70%
	IndoBERT (Bayesian Optimization)	97.45%	97.46%	97.45%	97.45%	89.24%
Location	RoBERTa (Random Search)	91.10%	91.13%	91.10%	91.12%	95.51%
	IndoBERT (Random Search)	96.95%	96.95%	96.95%	96.95%	98.46%
Price	RoBERTa (Random Search)	90.95%	91.06%	90.95%	90.94%	93.68%
	IndoBERT (Random Search)	96.30%	96.30%	96.30%	96.30%	97.20%
Service	RoBERTa (Random Search)	92.00%	92.14%	92.00%	91.99%	94.57%
	IndoBERT (Random Search)	96.40%	96.40%	96.40%	96.40%	97.28%

Based on the model evaluation results, both RoBERTa and IndoBERT demonstrated strong performance in analyzing hotel reviews in Indonesian. However, IndoBERT consistently outperformed RoBERTa in nearly all aspects tested, including Facilities, Cleanliness, Location, Price, and Service. IndoBERT showed superior performance across all metrics, with higher values for Accuracy, Precision, Recall, F1-Score, and AUC compared to RoBERTa, making it a more effective choice for sentiment analysis of hotel reviews. While RoBERTa produced competitive results, particularly in AUC, IndoBERT was better able to capture sentiment accurately across all aspects, making it the more superior model in this study.

3.7.2 Confusion Matrix

The confusion matrix illustrates the performance of the RoBERTa and IndoBERT models in classifying hotel reviews, showing the number of correct and incorrect predictions, for the application of the confusion matrix to each aspect, can be seen in Table 9 to Table 13.

a. Facilities Aspect

Table 9. Comparison of RoBERTa and IndoBERT Models Confusion Matrices on the Facilities Aspect

Algorithm	Actual	Prediction	
		Negative	Positive
RoBERTa	Negative	850	150
	Positive	143	857
IndoBERT	Negative	964	36
	Positive	51	949

b. Cleanliness Aspect

Table 10. Comparison of RoBERTa and IndoBERT Models Confusion Matrices on the Cleanliness Aspect

Algorithm	Actual	Prediction	
		Negative	Positive
RoBERTa	Negative	967	33
	Positive	64	936
IndoBERT	Negative	981	19
	Positive	32	968

c. Location Aspect

Table 11. Comparison of RoBERTa and IndoBERT Models Confusion Matrices on the Location Aspect

Algorithm	Actual	Prediction	
		Negative	Positive
RoBERTa	Negative	611	84
	Positive	94	1211
IndoBERT	Negative	666	29
	Positive	32	1273

d. Price Aspect

Table 12. Comparison of RoBERTa and IndoBERT Models Confusion Matrices on the Price Aspect

Algorithm	Actual	Prediction	
		Negative	Positive
RoBERTa	Negative	935	65
	Positive	116	884
IndoBERT	Negative	962	38
	Positive	36	964

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

e. Service Aspect

Table 13. Comparison of RoBERTa and IndoBERT Models Confusion Matrices on the Service Aspect

Algorithm	Actual	Prediction	
		Negative	Positive
RoBERTa	Negative	949	51
	Positive	109	891
IndoBERT	Negative	963	37
	Positive	35	965

3.7.3 RoBERTa Model Evaluation

The RoBERTa model used in this study has been optimized through two tuning methods: Bayesian Optimization and Random Search. The evaluation results reveal that, while RoBERTa provides competitive performance, IndoBERT outperforms it in several aspects, particularly in accuracy, precision, and recall. The evaluation was conducted using various metrics, including Accuracy, Precision, Recall, F1-Score, and AUC, to offer a comprehensive overview of the model's performance. For instance, in the "Facilities" aspect, RoBERTa with Bayesian Optimization achieved an accuracy of 85.35% and an AUC of 88.98%, indicating that the model performs reasonably well in identifying both positive and negative sentiments. However, challenges related to class imbalance between sentiment categories remain. The generated confusion matrix reveals that the model exhibits a higher rate of False Positives (FP) in certain aspects, suggesting that it tends to classify negative sentiment as positive in some cases.

3.7.4 IndoBERT Model Evaluation

Meanwhile, the IndoBERT model demonstrated superior performance across nearly all evaluated aspects. This model employs Random Search and Bayesian Optimization for tuning, and its evaluation results show higher accuracy compared to RoBERTa. For instance, in the Hygiene aspect, IndoBERT with Bayesian Optimization achieved an accuracy of 96.95% and an AUC of 97.28%. The evaluation encompassed various aspects, including Facility, Cleanliness, Location, and Service, all of which consistently showed high model performance. IndoBERT also outperformed RoBERTa in precision, recall, and F1-score. The confusion matrix results were more balanced, with a higher number of True Positives (TP) in almost all aspects, suggesting that IndoBERT is more effective at correctly identifying sentiment compared to RoBERTa.

3.7.5 Model Perfomance Comparison

When compared, IndoBERT consistently outperforms RoBERTa in nearly all aspects of sentiment analysis. This is evident from evaluation metrics such as precision, recall, and AUC, where IndoBERT demonstrates superior performance, particularly in identifying sentiments related to Facility, Cleanliness, and Service. While RoBERTa also yields promising results, especially in AUC, which highlights its capacity to distinguish between positive and negative sentiments, IndoBERT excels in more accurately capturing these sentiments. This is due to its specialized training on the Indonesian language, which provides it with a distinct advantage in understanding the local linguistic context.

These findings are further validated by the confusion matrix, which indicates that IndoBERT is more proficient in correctly predicting sentiments. For instance, in the Location Aspect, IndoBERT achieves a higher number of True Positives (TP) and fewer False Negatives (FN) compared to RoBERTa. This suggests that IndoBERT is more effective at identifying positive sentiments in reviews related to hotel locations, while RoBERTa tends to misclassify this category more frequently.

Thus, although RoBERTa delivers strong results, the IndoBERT model is more recommended for sentiment analysis of hotel reviews in Indonesia. It is particularly advantageous when dealing with more complex aspects that require a deeper understanding of the language. IndoBERT's proficiency in handling Indonesian language data and its superior evaluation results make it a more effective choice for this study.

3.8 Discussion

3.8.1 Perfomance of RoBERTa and IndoBERT Models

This study applies sentiment analysis to hotel reviews in Indonesia, utilizing two primary models: RoBERTa and IndoBERT. These models are optimized using two distinct tuning methods: Bayesian Optimization and Random Search. The evaluation results demonstrate that IndoBERT outperforms RoBERTa in nearly all tested aspects. Specifically, in the categories of Facility, Cleanliness, and Service, IndoBERT achieves higher precision, recall, and F1-score values compared to RoBERTa. Overall, IndoBERT, when optimized with Bayesian Optimization, produced more consistent results in detecting correct sentiments, showing higher True Positive (TP) rates and lower False Negative (FN) rates. These findings suggest that IndoBERT is more effective in accurately predicting both positive and negative sentiments.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

3.8.2 Effect of Tuning Methods on Model Perfomance

The Bayesian Optimization tuning method has been shown to deliver better results compared to Random Search on both models. IndoBERT, when optimized using Bayesian Optimization, achieved higher accuracy scores in the aspects of Cleanliness and Facilities, with respective values of 96.95% and 97.45%. Furthermore, RoBERTa, optimized with Bayesian Optimization, also demonstrated competitive results, showing higher AUC scores in several aspects, although it slightly lagged behind in precision and recall compared to IndoBERT. These results indicate that Bayesian Optimization is capable of identifying a more optimal combination of hyperparameters for these models.

3.8.3 Confusion Matrix Analysis for Each Aspect

By analyzing the Confusion Matrix, we were able to gain deeper insights into the classification errors made by both models. IndoBERT demonstrated higher rates of True Positives (TP) and False Negatives (FN). It performed relatively poorly in certain aspects, such as Facility, Cleanliness, and Service, which suggests that this model is more effective at identifying correct sentiments. In contrast, RoBERTa showed a higher frequency of False Positives (FP), particularly in the Service category. This indicates that RoBERTa tends to classify negative sentiments as positive in some cases. These findings suggest that IndoBERT has an advantage in maintaining a better balance between positive and negative classifications.

3.8.4 Discussion Result

Based on the findings discussed, it can be concluded that IndoBERT with Bayesian Optimization is a superior model for aspect-based sentiment analysis of hotel reviews in Indonesian. This model demonstrates higher accuracy and is more effective in identifying both positive and negative sentiments in reviews. While RoBERTa also yields good results, particularly in terms of AUC, IndoBERT proves to be more efficient in capturing sentiments accurately due to its specialized training on the Indonesian language and its ability to handle local language context. Therefore, IndoBERT is recommended for use in aspect-based sentiment analysis on platforms such as Traveloka.

4. CONCLUSION

Based on the comprehensive series of research conducted, ranging from data collection to model evaluation, it can be concluded that this study successfully developed and optimized the Aspect-Based Sentiment Analysis (ABSA) model for hotel reviews in the Indonesian language. The data, which were collected using web scraping techniques from the Traveloka platform, consisted of 305,875 hotel reviews in Bandung. These reviews were then extracted and analyzed based on five key aspects: facilities, cleanliness, location, price, and service. The RoBERTa and IndoBERT models were utilized and optimized using Bayesian Optimization and Random Search tuning techniques. The evaluation results revealed that IndoBERT, optimized with Bayesian Optimization, outperformed RoBERTa, particularly in terms of accuracy, precision, recall, and F1-score. It was notably more effective at capturing sentiment across nearly all aspects tested, especially cleanliness and facilities. Although RoBERTa produced competitive results, particularly in AUC, IndoBERT demonstrated superior performance in classifying positive and negative sentiments with fewer classification errors. This study also identified limitations stemming from data imbalance, particularly in the location aspect, which influenced the model's performance in sentiment identification for that aspect. Nevertheless, this research represents a significant advancement in the development of aspect-based sentiment analysis for Indonesian hotel reviews, leveraging cutting-edge technologies such as RoBERTa and IndoBERT. In the future, this research can be further enhanced through optimization of data processing and overcoming dataset limitations to achieve more accurate results. Additionally, it holds potential for application to other platforms that feature Indonesian-language customer reviews, thereby improving the quality of sentiment analysis in the hospitality industry.

REFERENCES

[1] N. K. Nissa and E. Yulianti, “Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[2] E. M. Pusung and I. N. Dewi, “Optimasi RoBERTa dengan Hyperparameter Tuning untuk Deteksi Emosi berbasis Teks,” *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 10, no. 3, pp. 240–248, Feb. 2025, doi: 10.25077/TEKNOSI.v10i3.2024.240-248.

[3] Y. A. Singgalen, “IndoBERT-Based Sentiment Analysis for Understanding Hotel Guests’ Preferences,” *Article in Journal of Computer System and Informatics*, vol. 6, no. 2, pp. 508–520, 2025, doi: 10.47065/josyc.v6i2.6864.

[4] I. Nawawi, K. F. Ilmawan, M. R. Maarif, and M. Syafrudin, “Exploring Tourist Experience through Online Reviews Using Aspect-Based Sentiment Analysis with Zero-Shot Learning for Hospitality Service Enhancement,” *Information (Switzerland)*, vol. 15, no. 8, Aug. 2024, doi: 10.3390/info15080499.

[5] F. Baharuddin and M. F. Naufal, “Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom’s Taxonomy,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, Oct. 2023, doi: 10.20473/jisebi.9.2.253-263.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 949−959
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7523

[6] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, "Sentiment analysis for customer review: Case study of Traveloka," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 682–690. doi: 10.1016/j.procs.2022.12.184.

[7] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 30–40, Feb. 2025, doi: 10.20473/jisebi.11.1.30-40.

[8] A. Lin, N. Livando, W. Chandra, G. Phan, and A. M. Husein, "Sentiment Analysis Of Hotel Reviews On Tripadvisor With LSTM And ELECTRA," *SinkrOn*, vol. 8, no. 2, pp. 733–740, Apr. 2023, doi: 10.33395/sinkron.v8i2.12234.

[9] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Enhancing machine learning-based sentiment analysis through feature extraction techniques," *PLoS One*, vol. 19, no. 2, Feb. 2024, doi: 10.1371/journal.pone.0294968.

[10] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.

[11] R. Merdiansah and A. Ali Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI*, vol. 7, no. 1, pp. 221–228, Mar. 2024, doi: https://doi.org/10.55338/jikomsi.v7i1.2895.

[12] Hizbul Izzi, Arief Setyanto, and Anggit Dwi Hartanto, "Optimalisasi Akurasi Algoritma Naïve Bayes Dengan Metode Syntetic Minority Oversampling Technique (Smote) Pada Data Numerik," *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, no. 1, pp. 217–227, Jan. 2025, doi: 10.29408/jit.v8i1.28340.

[13] F. A. Hizham and C. K. Murni, "Sentiment Analysis Based on Review of Puncak B29 Lumajang using Backpropagation Neural Network," Atlantis Press, 2024, pp. 210–214. doi: 10.2991/978-94-6463-346-7_39.

[14] Y. O. Sihombing, N. V. Situmorang, B. K. Negara, and J. M. Sutoyo, "Prediksi Sentimen Pada Teks Media Sosial Corporate University Menggunakan RoBERTa," *Prosiding PITNAS Widyaiswara*, vol. 1, no. 1, pp. 302–316, Sep. 2024, [Online]. Available: https://ejournal.iwi.or.id/ojs/index.php/pitnas2024/article/view/316/187

[15] Dian Agus Prawinata, Ani Dijah Rahajoe, and I Gede Susrama Mas Diyasa, "Analisis Sentimen Kendaraan Listrik Pada Twitter Menggunakan Metode Long Short Term Memory," *SABER : Jurnal Teknik Informatika, Sains dan Ilmu Komunikasi*, vol. 2, no. 1, pp. 300–313, Jan. 2024, doi: 10.59841/saber.v2i1.857.

[16] S. Chibuzor, "Smart City PM2.5 Air Pollution Modeling Techniques: Train-Test Data Split versus K-Fold Cross Validation Techniques," *Journal of Inventive Engineering and Technology (JIET*, vol. 2, no. 3, pp. 30–44, Mar. 2023, [Online]. Available: https://www.jiengtech.com/index.php/INDEX/article/viewFile/50/49

[17] J. A Ilemobayo *et al.*, "Hyperparameter Tuning in Machine Learning: A Comprehensive Review," *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 388–395, Jun. 2024, doi: 10.9734/jerr/2024/v26i61188.

[18] A. Alamsyah and N. D. Girawan, "Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model," *Big Data and Cognitive Computing*, vol. 7, no. 4, Dec. 2023, doi: 10.3390/bdcc7040168.

[19] A. Farhan AlShammari, "Implementation of Model Evaluation using Confusion Matrix in Python," *Int J Comput Appl*, vol. 186, no. 50, pp. 0975–8887, Nov. 2024, doi: 10.5120/ijca2024924236.

[20] A. R. Adinata, T. Rohana, K. A. Baihaqi, and S. Faisal, "Implementasi Algoritma Convolutional Neural Network dan YOLOV8 Untuk Klasifikasi Ras Kucing," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, pp. 1658–1667, Dec. 2024, doi: 10.47065/bits.v6i3.5913.

[21] H. Kaur and D. Kaur Sandhu, "Evaluating the Effectiveness of the Proposed System Using F1 Score, Recall, Accuracy, Precision and Loss Metrics Compared to Prior Techniques," *International Journal of Communication Networks and Information Security*, vol. 15, pp. 368–383, Sep. 2024, [Online]. Available: https://ijcnis.org/index.php/ijcnis/article/view/7168