Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine-Tuned IndoBERT

Aulia Pinkan Maretta [1]*, Allsela Meiriza [2]*
* Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya
09031182227020@student.unsri.ac.id [1], allsela@unsri.ac.id [2]

Article Info

Article history:

Received 2025-08-12
Revised 2025-08-31
Accepted 2025-09-04

Keyword:

Aspect-Based Sentiment Analysis,
Fine-Tuning,
Google Review,
IndoBERT.

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) has become a crucial approach for extracting detailed opinions from user-generated content, especially in the healthcare domain. This study analyzes public sentiment toward hospital services in Indonesia using IndoBERT, fine-tuned on 2.448 reviews collected from Google Reviews and Instagram. Sentiment labels were automatically assigned with a pre-trained Indonesian RoBERTa classifier, while aspect extraction was performed through a lexicon-based approach covering five service dimensions: Facilities, Staff Competence, Empathy and Communication, Reliability and Responsiveness, and Cost and Affordability. To address class imbalance, the IndoBERT model was optimized using class weight adjustments. The results demonstrate strong performance, achieving an overall accuracy of 96%. In terms of sentiment classification, the model obtained F1-scores of 89% for negative, 83% for neutral, and 99% for positive sentiment, with a macro-average F1 of 90%. By aspect, Facilities (82.24%) and Empathy & Communication (91.71%) received the highest positive sentiment, while Cost & Affordability recorded the highest proportion of negative sentiment (25%). These findings underscore the effectiveness of IndoBERT-based ABSA in capturing nuanced public perceptions and highlight its potential as a decision-support tool for hospitals to enhance service quality and patient satisfaction in Indonesia.

I. INTRODUCTION

Aspect-Based Sentiment Analysis (ABSA) has become one of the important approaches in natural language processing (NLP) to explore public perceptions of an entity based on specific aspects. Early research on ABSA has largely focused on explicit sentences, where aspects and opinions can be clearly identified through keywords or phrases that directly indicate sentiment toward a specific aspect [1], [2]. However, as the need to understand more complex opinions has grown, ABSA has begun to address implicit cases, where aspects or opinions are not always expressed directly [3], [4].

In the context of healthcare services, sentiment analysis plays a crucial role in evaluating the quality of hospital services based on public feedback. For example, [5] analyzed 1,931 reviews of hospital services in Palangka Raya using various machine learning models (KNN, Logistic Regression, and Decision Tree), and detected emotions using the NRC Lexicon. The results showed that 65.3% of reviews had positive sentiment, and dominant emotions such as anticipation and positivity reinforced the public's expectations and satisfaction with hospital services. [6] compared the performance of LSTM, BiLSTM, and GRU in classifying Twitter users' sentiment toward hospital services during the pandemic and found that the BiLSTM model provided the best accuracy at 86%.

Another study by [7] utilizing the BERT model showed that the transformer-based approach is highly effective in processing Indonesian-language review texts, particularly in the healthcare domain. Research by [8] also contributes in a similar context by applying the Naïve Bayes Classifier method to public reviews of hospital services in Malang. Using an 80:20 training and testing data split and testing with

K-Fold Cross Validation, the study achieved an accuracy of up to 90%

Similar findings are supported by [9] which used IndoBERT and achieved an accuracy of 96% on health app review data such as Alodokter and Halodoc. However, most previous studies still focus on general sentiment classification and have not fully integrated ABSA techniques in the context of local hospital services in Indonesia. Meanwhile, the ABSA approach can identify service dimensions such as staff competence, facilities, costs, and medical staff empathy in greater depth. For example, a study by [10] found that the majority of positive reviews from the public about Khalishah Hospital were related to the quality of medical staff and facilities, while administrative aspects remained the main complaint.

The ABSA approach has also been used in an international context. [3] Through the OptiASAR model, it combines BiLSTM-GRU and NER-BERT architectures to analyze Yelp-based healthcare reviews. This model demonstrates its superiority in identifying service aspects with precision and relevance for decision-making in the medical field.

In Indonesia, ABSA research remains limited compared to sentiment analysis in general. One significant study is by [11], which integrates Conditional Random Field (CRF) and Weighted Average Ensemble (revies) to extract aspects and sentiments from hospital reviews on Google Maps. The results showed that SVM and Naïve Bayes models in combination with ensembles are capable of providing competitive classification performance. Meanwhile, research by [12] demonstrated that fine-tuning Multilingual BERT on Indonesian hotel reviews with sentence-pair classification provides improved performance for ABSA tasks.

Although various methods have been developed, approaches to simultaneously extract implicit aspects and opinions, particularly in the context of hospitals in Indonesia, remain limited. Indirect opinions often appear in public reviews. For example, a sentence like “The doctor is really busy, long wait” contains negative sentiment toward responsiveness, even though it doesn't mention it explicitly. This limitation is further emphasized in [4] which underscores the importance of addressing ambiguous and implicit expressions in ABSA through the integration of advanced techniques. To effectively overcome these challenges in the Indonesian healthcare context, a transformer-based model that is specifically trained on the Indonesian language is required.

Among the various transformer-based models, IndoBERT is considered the most suitable for this task because it is pre-trained on a large-scale Indonesian corpus, allowing it to better capture the linguistic characteristics and contextual nuances of the Indonesian language compared to multilingual models such as mBERT [13]. Recent empirical studies also demonstrate that IndoBERT consistently outperforms mBERT and XLM-R in sentiment classification tasks involving Indonesian public service reviews, achieving an F1-score of 0.882 [14]. Given that hospital reviews are predominantly expressed in semi-formal to formal Indonesian, IndoBERT provides a more contextually appropriate foundation for this study.

This study applies an Aspect-Based Sentiment Analysis approach using IndoBERT to analyze user sentiment toward hospital services in Indonesia, focusing on key aspects of healthcare quality. The approach combines lexicon-based aspect extraction, sentiment labeling with RoBERTa, and sentiment classification through IndoBERT fine-tuning. In addition, the study compares non-transformer machine learning models such as SVM and Random Forest with transformer-based approaches (IndoBERT), followed by model performance evaluation using precision, recall, and F1-score metrics. By leveraging review data collected from Google Reviews and Instagram, this study provides a detailed, aspect-specific understanding of public perceptions of healthcare services based on real-world data.

II. Methodology

This study adopts a Natural Language Processing (NLP) and Aspect-Based Sentiment Analysis (ABSA) approach that aims to extract and evaluate public opinion on hospital services based on digital reviews. The methodological stages are systematically arranged as follows:

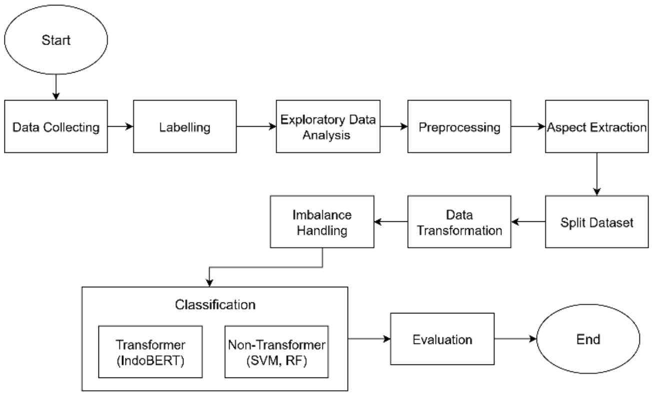


Figure 1. Research flow

A. Data Collecting

The initial stage of this study was the data acquisition process, which focused on public reviews from two social media platforms, namely Google Review [11] and Instagram [15]. Data collection was carried out on July 11, 2025, from reviews of XYZ Hospital in Palembang using the following methods:

1. Google Review: Data was collected using the Instant Data Scraper application, which extracted review text, user names, as well as the date and number of feedback provided.
2. Instagram: Data was obtained through web scraping, by entering the URL of Instagram posts related to RS XYZ that had user comments.

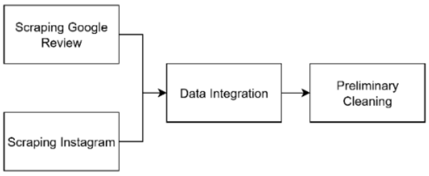


Figure 2. Flow of Data Collecting

B. Labelling

Sentiment labels are automatically determined using the pre-trained model w11wo/indonesian-roberta-base-sentiment-classifier, which is derived from the RoBERTa architecture and has been adapted to the Indonesian corpus as done in study [16]. This model is capable of classifying text into three sentiment categories: positive, neutral, and negative, and was chosen for its high performance in Indonesian emotion classification tasks. The labeling process was performed on all collected raw data, resulting in a dataset ready for exploration and further model training.

C. Exploratory Data Analysis

This step is taken to understand the distribution and characteristics of the data before it enters the modeling stage. EDA is super important in helping researchers get some early insights into the structure and patterns in the data, as well as identifying potential issues like label imbalance [17]. In this study, EDA includes several key visualizations, namely the distribution of reviews by year, sentiment distribution, and review length distribution per sentiment category.

D. Preprocessing

The preprocessing stage is carried out to improve the quality of text representation before it is entered into the model.

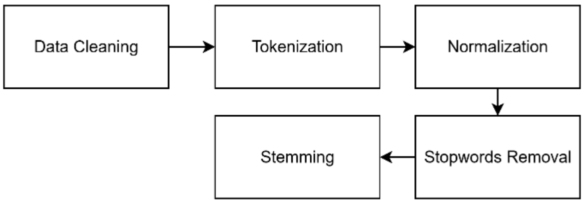


Figure 3. Flow of Preprocessing

1) *Data cleaning*

This step involves several operations, including (a) converting all letters to lowercase, (b) removing URLs and HTML tags, (c) eliminating non-alphabetic characters, (d) trimming excess spaces, and (e) removing reviews that become empty after cleaning. These procedures aim to reduce noise and ensure that the input to the model is clean and consistent [18]

2) *Tokenization*

Tokenization is the process of splitting text into smaller units, usually words or subwords, which serve as the basic elements for analysis [5].

3) *Normalization*

Normalization standardizes words to a consistent form, such as converting contractions to their full form or correcting spelling errors [17].

4) *Stopwords Removal*

Stopwords are common words (e.g., “and”, “the”, “is”) that carry little semantic meaning. Removing them reduces dimensionality and focuses the model on meaningful content [7].

5) *Stemming*

Stemming reduces words to their base or root form (e.g., “running” into “run”), which helps consolidate different forms of the same word and reduces vocabulary size [19].

E. Aspect Extraction

This study compiled a domain-specific lexicon, with five hospital service aspects identified from the literature [20], [21], [22], namely:

1. Facilities
2. Reliability & Responsiveness
3. Staff Competence
4. Empathy & Communication
5. Cost & Affordability

For each aspect, a list of keywords was extracted from the dataset, reflecting the actual terms and expressions used by patients in their reviews. Aspects in each review were determined by matching these dataset-derived keywords to the aspect dictionary. This lexicon-based approach enables automatic categorization of reviews according to domain-specific aspects, without requiring manual annotation, ensuring efficiency in identifying the key service dimensions mentioned by patients.

F. Split Data

The cleaned and labeled dataset was divided into training, validation, and test sets using a stratified split to maintain the proportional distribution of sentiment labels across subsets. Specifically, 70% of the data was used for training, 15% for validation, and 15% for testing [23]. The clean_text column was used as the feature, while the mapped sentiment label served as the target. The validation set was employed to monitor model performance at each epoch and to determine the best model configuration based on accuracy.

G. Data Transformation

Text review data is converted into numerical representations using IndoBERT tokenizer-based tokenization, which performs subword tokenization in accordance with the transformer architecture. This process produces an input format that includes input_ids, attention_mask, and token_type_ids.

H. Imbalance Handling

The distribution of labels in the training dataset exhibited a significant imbalance, with positive reviews being the dominant class. To mitigate this issue, a class weight adjustment strategy was incorporated into the loss function, specifically by assigning higher weights to minority classes within the Cross-Entropy Loss calculation [24]. This adjustment aimed to enhance the model's sensitivity to minority classes and reduce bias toward the majority class.

I. Classification

This study compares baseline machine learning models with IndoBERT, a transformer-based deep learning model, to evaluate performance differences.

1) Classification with Non-Transformer

1.1) Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method primarily used for classification. It aims to find the optimal hyperplane that separates classes while maximizing the margin. SVM can be linear, for linearly separable data, or non-linear, using kernel functions to handle complex data distributions [25] [26] .

1.2) Random Forest

Random Forest is an ensemble method based on bagging (bootstrap and aggregating). The training set is sampled using random sampling with replacement to build the desired number of trees. Combining bagging with random feature selection produces an uncorrelated forest of decision trees, improving generalization and reducing overfitting [27].

2) Classification with Transformer

IndoBERT is a monolingual BERT-based language model tailored for Indonesian. It was pre-trained on a large-scale Indonesian corpus and achieved state-of-the-art results across multiple NLP tasks in the IndoLEM benchmark including sentiment analysis, named-entity recognition, and morpho-syntactic analysis outperforming multilingual BERT and other baselines [28].

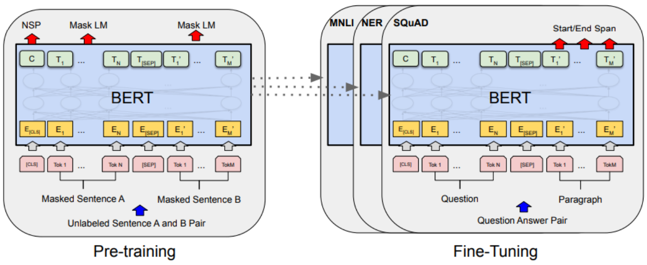


Figure 4. IndoBERT Pre-Training dan Fine-Tuning Stages

The model used in this study is indobenchmark/indobert-base-p1, which is the base version of the IndoBERT model that has been pre-trained on a large Indonesian-language corpus [29]. Fine-tuning is performed to allow the model to adjust its internal parameters to a specific dataset labeled with positive, neutral, and negative sentiments [30].

The training process was conducted using the Trainer API from the Hugging Face Transformers library with the following settings: 6 epochs, a learning rate of 1e-5, a batch size of 8 for both training and evaluation, and the Adam optimizer with a weight decay of 0.01. The model was evaluated and saved at the end of each epoch, and the best model was loaded based on accuracy. Logging was performed every 10 steps, and only the two most recent models were retained.

J. Evaluation (ABSA)

TABLE I
CONFUSION MATRIX

Class	Prediction Positive	Prediction Neutral	Prediction Negative
Actual Positive	TP	FNt1	FNg1
Actual Neutral	FP2	TNt	FNg2
Actual Negative	FP1	FNt2	TNg

Model performance evaluation is conducted using a classification report, which includes metrics such as accuracy, precision, recall, and F1-score. These metrics are calculated based on the values in the confusion matrix and are formulated as follows [17], [31]:

1. *Accuracy:* Measures the proportion of correct predictions out of all predictions made by the model, calculated using (1).

$$\frac{TP + TNg + TNt}{TP + FNt2 + .. + FNt1 + TNg} \tag{1}$$

2. *Precision:* Measures how accurate the model's positive predictions are by calculating the ratio of the number of correct positive predictions (true positives) to all predictions declared positive by the model, calculated using (2).

$$Positive = \frac{TP}{TP + FP1 + FP2}$$
$$Neutral = \frac{TNt}{TNt + FNt1 + FNt2}$$
$$Negative = \frac{TNg}{TNg + FNg1 + FNg2} \tag{2}$$

3. *Recall:* Measures the model's ability to identify all positive instances in the data, calculated using (3*).*

$$Positive = \frac{TP}{FNg1 + FNt1 + TP}$$
$$Neutral = \frac{TNg}{FP1 + FNt2 + TNg}$$

$$Negative = \frac{TNt}{FNg2 + FP2 + TNt} \quad (3)$$

4. *F1-Score:* A calculation that represents the balance between precision and recall. If the False Negative (FN) and False Positive (FP) values are not balanced, it is more advisable to use F1-score than accuracy. F1-score is calculated using (4*).*

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

The evaluation is performed on test data that was not involved in the training to ensure the objectivity of the results.

III. Results and discussion

This section presents the results and analysis of the research, with a focus on the practical application of the methodology used. This can be accomplished through a straightforward presentation of research data. Additionally, this section includes various visual elements such as explanations, images, tables, and other relevant visualizations.

A. Data Collecting

The initial stage of this study was the collection of user reviews from two main platforms, namely Google Review and Instagram. A total of 2.448 reviews were collected, covering user comments related to services, facilities, and experiences at XYZ Hospital in Palembang.

The main attributes collected are shown in Table II below:

TABLE II
Data attributes

Atributes	Description
Username	The identity of the reviewer on Google Review. This may include the full name or display name of the Google account.
Time	A timestamp indicating when the review was posted. Expressed in a relative format (e.g., “2 weeks ago,” “5 days ago”).
Text	The full text of the user's review, which includes opinions, complaints, praise, or suggestions related to the hospital's services.
Response	The official response from the hospital (XYZ) to the user's review, which typically includes expressions of appreciation or clarification regarding the issues raised.

B. Labelling

The labeling process is carried out using the text-classification pipeline from the transformers library, where each line of text is analyzed and cut to a maximum of 512 tokens to avoid processing errors. The sentiment label prediction results are then stored in a new column named sentiment.

TABLE III
Labelling Result

Text	Sentiment
Perawat ramah dan fasilitas lengkap	Positive
Pelayanan ya sangat cepat tempat ya bersih dan nyaman dokter & perawat sangat ramah tamah	Positive
Kalau ada minus, saya kasih minus. Perawat judes, keluarga saya bayar disini bukan gratis!!!	Negative
Pelayanan sangat buruk, petugasnya pada songong2, ga ada ramah ramahnya	Negative
Pelayanan bagus, rumah sakitnya bersih	Positive

C. Exploratory Data Analysis

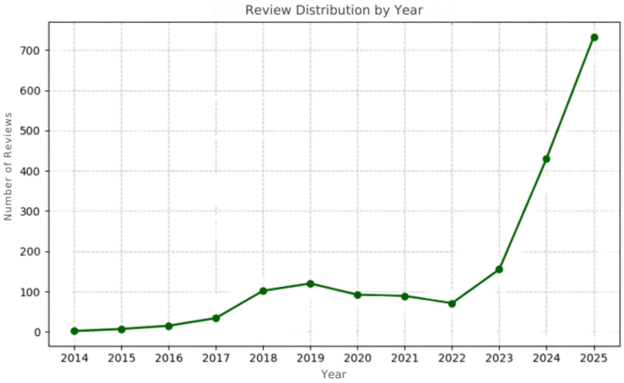


Figure 5. Distribution of the number of reviews by year.

The above figure shows the distribution of user reviews by year. The data indicate a steady increase in reviews from 2014 to 2019, followed by fluctuations in 2020–2022. A significant surge occurred in 2024, with the highest number of reviews recorded in 2025, exceeding 700 entries.

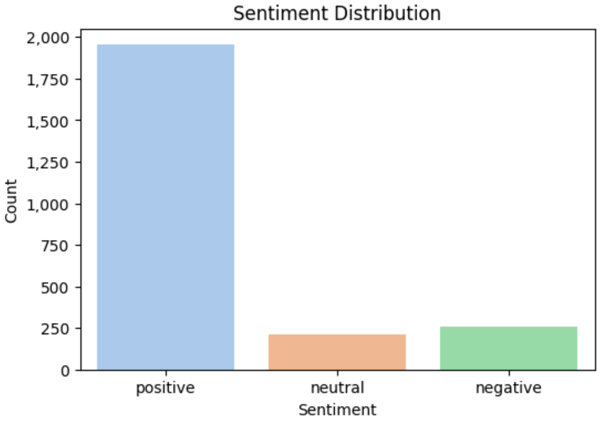


Figure 6. Sentiment Distribution

The above figure shows the distribution of sentiment labels across the dataset. The results reveal a strong imbalance, with positive sentiment dominating at 80.6%, followed by negative sentiment at 10.6%, and neutral sentiment at 8.7%.

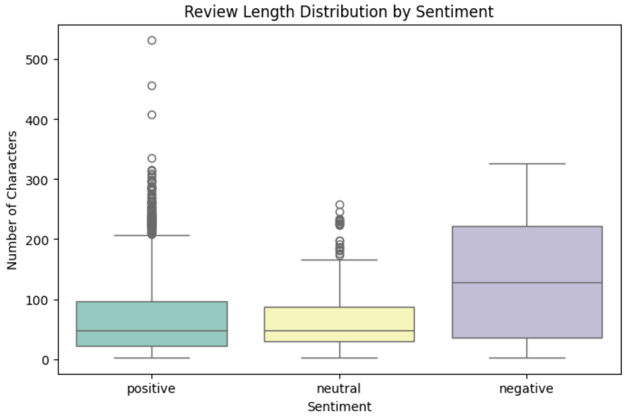


Figure 7. Review Length Distribution by Sentiment

The above figure shows the review length distribution by sentiment category. Negative reviews are generally longer and more varied than positive and neutral reviews, indicating that users tend to provide more detailed explanations when expressing complaints. In contrast, neutral reviews are usually short and concise, while positive reviews vary, with some long entries reflecting more descriptive appreciation.

D. *Preprocessing*

1) *Data Cleaning*

The amount of data before cleaning was 2,448 reviews. After data cleaning, 2.421 reviews remained usable. The difference before and after cleaning is shown in Table IV.

TABLE IV
CLEANING RESULT

Text	Clean_Text
Pelayanan ya sangat cepat tempat ya bersih dan nyaman dokter & perawat sangat ramah tamah	pelayanan ya sangat cepat tempat ya bersih dan nyaman dokter perawat sangat ramah tamah
Kalau ada minus, saya kasih minus. Perawat judes, keluarga saya bayar disini bukan gratis!!!	kalau ada minus saya kasih minus perawat judes keluarga saya bayar disini bukan gratis

2) *Tokenization*

After data cleaning, tokenization is performed to split text into meaningful units (tokens) for further processing. Table V shows examples before and after tokenization.

TABLE V
TOKENIZATION RESULT

Before Tokenization	After Tokenization
pelayanan ya sangat cepat tempat ya bersih dan nyaman dokter perawat sangat ramah tamah	['pelayanan', 'ya', 'sangat', 'cepat', 'tempat', 'ya', 'bersih', 'dan', 'nyaman', 'dokter', 'perawat', 'sangat', 'ramah', 'tamah']
kalau ada minus saya kasih minus perawat judes keluarga saya bayar disini bukan gratis	['kalau', 'ada', 'minus', 'saya', 'kasih', 'minus', 'perawat', 'judes', 'keluarga', 'saya', 'bayar', 'disini', 'bukan', 'gratis']

3) *Normalization*

After tokenization, normalization standardizes text by converting words to lowercase, removing punctuation, and handling informal expressions. Table VI shows examples before and after normalization.

TABLE VI
NORMALIZATION RESULT

Before Normalization	After Normalization
['pelayanan', 'ya', 'sangat', 'cepat', 'tempat', 'ya', 'bersih', 'dan', 'nyaman', 'dokter', 'perawat', 'sangat', 'ramah', 'tamah']	['pelayanan', 'ya', 'sangat', 'cepat', 'tempat', 'ya', 'bersih', 'dan', 'nyaman', 'dokter', 'perawat', 'sangat', 'ramah', 'tamah']
['kalau', 'ada', 'minus', 'saya', 'kasih', 'minus', 'perawat', 'judes', 'keluarga', 'saya', 'bayar', 'disini', 'bukan', 'gratis']	['kalau', 'ada', 'minus', 'saya', 'kasih', 'minus', 'perawat', 'judes', 'keluarga', 'saya', 'bayar', 'disini', 'bukan', 'gratis']

4) *Stopwords Removal*

After normalization, stopwords removal is applied to eliminate common words (e.g., *dan, ya, saya*) that do not carry significant meaning for sentiment analysis. Table VII presents examples before and after stopwords removal.

TABLE VII
STOPWORDS REMOVAL RESULT

Before Stopwords	After Stopwords
['pelayanan', 'ya', 'sangat', 'cepat', 'tempat', 'ya', 'bersih', 'dan', 'nyaman', 'dokter',	['pelayanan', 'cepat', 'tempat', 'bersih', 'nyaman', 'dokter', 'perawat', 'ramah', 'tamah']

'perawat', 'sangat', 'ramah', 'tamah']	
['kalau', 'ada', 'minus', 'saya', 'kasih', 'minus', 'perawat', 'judes', 'keluarga', 'saya', 'bayar', 'disini', 'bukan', 'gratis']	['minus', 'kasih', 'minus', 'perawat', 'judes', 'keluarga', 'bayar', 'disini', 'gratis']

5) *Stemming*

After stopwords removal, stemming is applied to reduce words to their root forms, ensuring that different inflections of a word are treated as the same token. Table VIII shows examples before and after stemming.

TABLE VIII
STEMMING RESULT

Before Stemming	After Stemming
['pelayanan', 'cepat', 'tempat', 'bersih', 'nyaman', 'dokter', 'perawat', 'ramah', 'tamah']	layan cepat tempat bersih nyaman dokter perawat ramah tamah
['minus', 'kasih', 'minus', 'perawat', 'judes', 'keluarga', 'bayar', 'disini', 'gratis']	minus kasih minus perawat judes keluarga bayar disini gratis

E. *Aspect Extraction*

Aspect extraction was applied to identify key features mentioned in user reviews. Each review may contain one or more aspects reflecting opinions on hospital services.

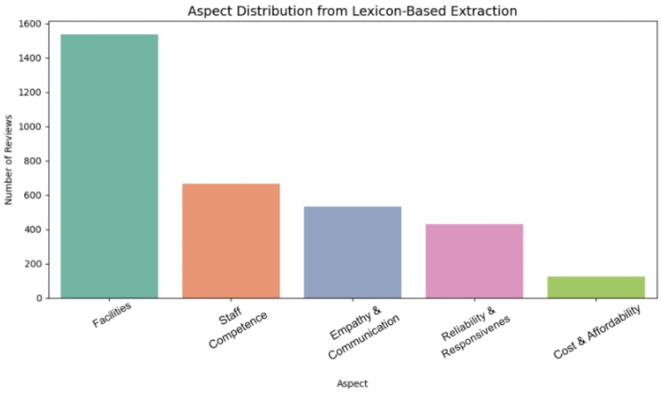


Figure 8. Aspect Distribution from Lexicon-Based Extraction

TABLE IX
DATA DISTRIBUTION BY ASPECT

Aspect	Count
Facilities	1537
Staff Competence	665
Empathy & Communication	531
Reliability & Responsiveness	431
Cost & Affordability	124

F. *Data Transformation*

The dataset was split into training, validation, and test sets. The training set contained 1,366 positive reviews, 181 negative reviews, and 147 neutral reviews. The validation set contained 363 reviews, and the test set contained 293 positive reviews, 39 negative reviews, and 32 neutral reviews.

Each review text is converted into tokens using the indobenchmark/indobert-base-p1 tokenizer. Tokenization produces two main vectors: input_ids and attention_mask, each with a fixed dimension of 128 tokens. input_ids represents the token IDs in the IndoBERT dictionary, while attention_mask marks important tokens (1) and padding (0).

G. *Imbalance Handling*

To address class imbalance in sentiment labels, class weights are calculated based on the frequency of labels from the training data. The calculated weights are then applied to the CrossEntropyLoss function to give a greater penalty for classification errors in minority classes. The final weights used are shown in Table X:

TABLE X
CLASS WEIGHTS FOR EACH SENTIMENT LABEL

Sentiment	Class Weight
Negative	3,13
Neutral	3,83
Positive	0,41

H. *Classification*

1) *Classification with Non-Transformer*

TABLE XI
CLASSIFICATION REPORT

Classification Report	SVM		
	Precision	Recall	F-1 Score
Negative	0.77	0.69	0.73
Neutral	0.91	0.31	0.47
Positive	0.91	0.98	0.94
Accuracy	0.89		
	Random Forest		
Negative	0.90	0.46	0.61
Neutral	0.88	0.22	0.35
Positive	0.87	0.99	0.93
Accuracy	0.87		

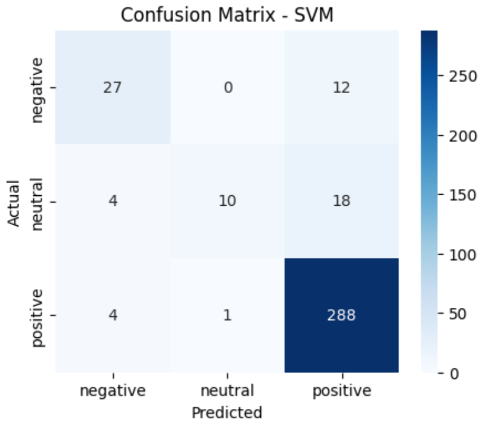


Figure 9. Confusion Matrix SVM

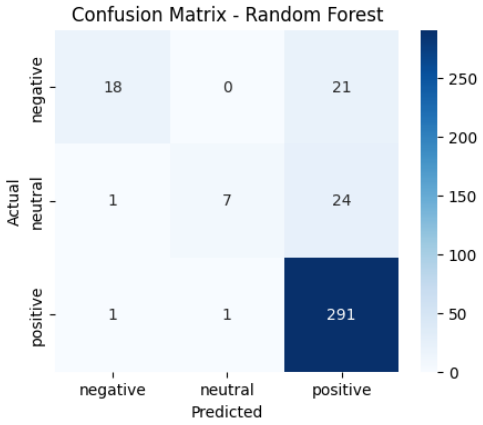


Figure 10. Confusion Matrix Random Forest

2) *Classification with Transformer*

TABLE XII
CLASSIFICATION REPORT

	Precision	Recall	F1-Score
Negative	0.85	0.92	0.89
Neutral	0.92	0.75	0.83
Positive	0.98	0.99	0.99
Accuracy			0.96
Macro avg	0.92	0.89	0.90
Weighted avg	0.96	0.96	0.96

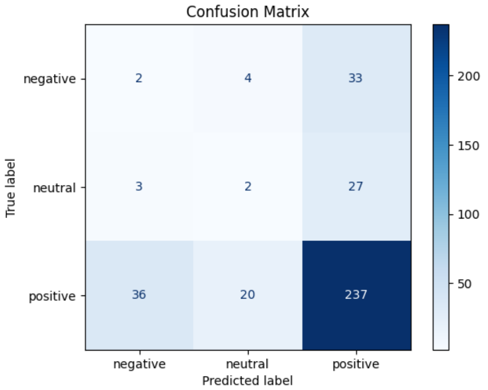
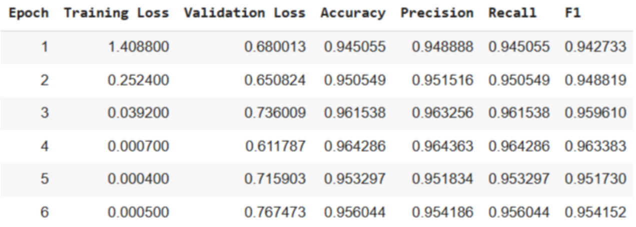


Figure 11. Confusion Matrix IndoBERT



Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	1.408800	0.680013	0.945055	0.948888	0.945055	0.942733
2	0.252400	0.650824	0.950549	0.951516	0.950549	0.948819
3	0.039200	0.736009	0.961538	0.963256	0.961538	0.959610
4	0.000700	0.611787	0.964286	0.964363	0.964286	0.963383
5	0.000400	0.715903	0.953297	0.951834	0.953297	0.951730
6	0.000500	0.767473	0.956044	0.954186	0.956044	0.954152

Figure 12. Training History

I. *Evaluation (ABSA)*

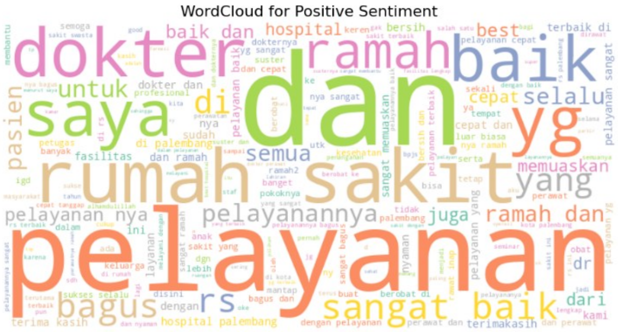


Figure 13. WordCloud for Positive Sentiment

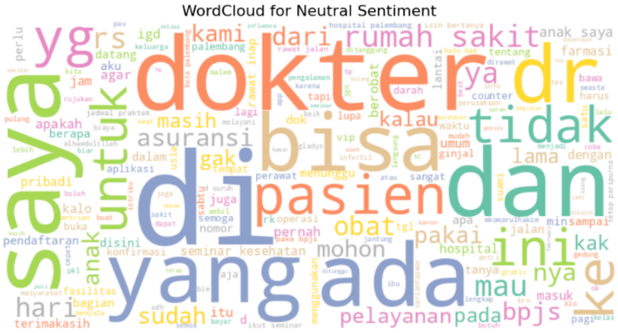


Figure 14. WordCloud for Neutral Sentiment

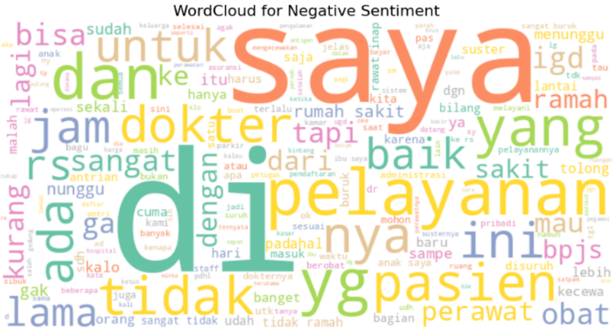


Figure 15. WordCloud for Negative Sentiment

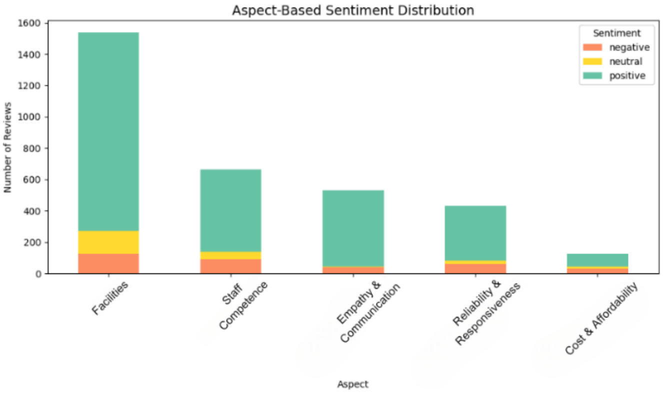


Figure 16. Sentiment Distribution by Aspect

TABLE XIII
SENTIMENT DISTRIBUTION BY ASPECT (IN %)

Aspect	Negative (%)	Neutral (%)	Positive (%)
Facilities	8.33	9.43	82.24
Staff & Competence	13.53	7.52	78.95
Empathy & Communication	7.72	0.56	91.71
Reliability & Responsiveness	14.39	4.87	80.74
Cost & Affordability	25	9.68	65.32

J. *Discussion*

Based on the visualization in Figure 16 and the detailed values in Table XIII, several important insights emerged across different service aspects. The Facilities aspect was the most frequently mentioned, with 82.24% positive sentiment, indicating overall satisfaction with hospital infrastructure, treatment room conditions, cleanliness, and comfort. Strategically, this highlights that continuous investment in facilities and hygiene standards is essential to maintaining patient trust and ensuring competitive service quality.

The Staff Competence aspect received 78.95% positive sentiment, reflecting strong patient appreciation for the professionalism and expertise of medical personnel. This suggests that hospitals should sustain and expand training programs, promote skill development, and implement competency-based evaluations to ensure consistent service quality.

The Empathy and Communication aspect achieved the highest positive sentiment rate at 91.71%, underscoring the value patients place on interpersonal interactions. Nevertheless, the 7.72% negative and 0.56% neutral reviews suggest gaps in communication and emotional support. From a strategic perspective, hospitals could implement targeted training in patient-centered communication, improve consultation protocols, and establish feedback loops that reinforce empathy as a service standard.

Sentiment in the Reliability and Responsiveness aspect was more varied, with 80.74% positive, 4.87% neutral, and 14.39% negative reviews. Negative comments were primarily related to long waiting times and delays, reflecting operational inefficiencies. This variability signals a strategic need to optimize scheduling systems, reduce administrative bottlenecks, and increase staffing efficiency. Improving responsiveness not only enhances patient satisfaction but also strengthens perceptions of organizational reliability.

The Cost and Affordability aspect, though less frequently mentioned, showed 65.32% positive sentiment but 25% negative sentiment. While most patients perceive service costs as reasonable, a notable portion still considers affordability an issue, suggesting the need for transparent pricing and consideration of flexible payment options for certain patient groups.

Turning to the classification models, the baseline methods (SVM and Random Forest) demonstrated varying performance in sentiment classification. SVM achieved an accuracy of 89%, performing well on the positive class (F1-score 94%) but relatively weak in detecting neutral reviews (F1-score 47%). Random Forest reached an accuracy of 87%, showing good performance in the positive class (F1-score 93%), but similar to SVM, its ability to recognize neutral sentiment was limited (F1-score 35%).

In contrast, the fine-tuned IndoBERT model significantly outperformed the baselines, achieving the highest accuracy of 96% and a macro-average F1-score of 90%. IndoBERT demonstrated superior performance across all classes, with particularly outstanding results in the positive class, achieving a Precision of 98%, Recall of 99%, and F1-score of 99%. For the negative class, the model reached a Precision of 85%, Recall of 92%, and F1-score of 89%, indicating strong capability in identifying negative sentiments despite class imbalance. Meanwhile, in the neutral class, IndoBERT obtained a Precision of 92%, Recall of 75%, and F1-score of 83%, showing that although neutral sentiment remains more challenging due to its inherent ambiguity, the model still managed to perform better than traditional machine learning approaches.

These results highlight IndoBERT's effectiveness in capturing the complexity and nuances of the Indonesian language across different sentiment categories, though several limitations should be acknowledged. The use of automatically generated sentiment labels from a pre-trained Indonesian RoBERTa classifier with length truncation may have introduced label noise and domain drift when applied to hospital-specific language. The dataset itself is drawn from a single hospital, which constrains the extent to which the findings can be generalized to other regions or service contexts. Moreover, even with the application of class-weighted loss to address imbalance, the neutral class remained relatively difficult to classify, suggesting residual ambiguity in mixed-polarity narratives.

IV. Conclusion

This study applied Aspect-Based Sentiment Analysis (ABSA) using IndoBERT to analyze 2.448 user reviews of a hospital in Indonesia. Sentiments were automatically generated by RoBERTa for initial labeling, while aspect extraction was performed using a lexicon-based approach across five service dimensions: Facilities, Staff Competence, Empathy & Communication, Reliability & Responsiveness, and Cost & Affordability. Class imbalance was addressed using weighted CrossEntropyLoss.

Results showed that IndoBERT outperformed non-transformer baselines, achieving 96% accuracy and a macro-average F1 of 90%. Aspect-level analysis highlighted high positive sentiment for Facilities (82%) and Empathy & Communication (92%), strong appreciation for Staff Competence (79%), mixed responses for Reliability & Responsiveness (81% positive, 14% negative), and concerns in Cost & Affordability (65% positive, 25% negative), suggesting a need for transparent pricing and flexible payment options.

For future work, it is expected that manual verification of sentiment labels, expanding the dataset to include multiple hospitals, and exploring hybrid or model-based aspect extraction methods could further enhance the accuracy and comprehensiveness of ABSA, enabling more effective insights for hospital service improvement.

References

[1] X. Li, L. Chen, B. Chen, and X. Ge, "BERT-BiLSTM-Attention model for sentiment analysis on Chinese stock reviews," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns-2024-1847.

[2] A. Bansal and N. Kumar, "Aspect-Based Sentiment Analysis Using Attribute Extraction of Hospital Reviews," *New Gener Comput*, vol. 40, no. 4, pp. 941–960, Dec. 2022, doi: 10.1007/s00354-021-00141-3.

[3] D. Singh, S. Shivprakash Barve, and A. Krishna Dwivedi, "OptiASAR: Optimized Aspect-Based Sentiment Analysis of Reviews With BiLSTM-GRU and NER-BERT in Healthcare Decision-Making," *IEEE Access*, vol. 13, pp. 47459–47474, 2025, doi: 10.1109/ACCESS.2025.3549303.

[4] Suhariyanto, R. Sarno, C. Fatichah, and R. Abdullah, "Aspect-based sentiment analysis: natural language understanding for implicit review," Dec. 01, 2024, *Institute of Advanced Engineering and Science*. doi: 10.11591/ijece.v14i6.pp6711-6722.

[5] A. C. T Angel and V. H. Pranatawijaya, "Analisis Sentimen dan Emosi dari Ulasan Google Maps untuk Layanan Rumah Sakit di Palangka Raya Menggunakan Machine Learning," 2024.

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

[7] A. Sri Widagdo *et al.*, "Analisis sentimen terhadap pelayanan Kesehatan berdasarkan ulasan Google Maps menggunakan BERT".

[8] T. Dzulkarnain, D. E. Ratnawati, and B. Rahayudi, "Penggunaan Metode Naïve Bayes Classifier pada Analisis Sentimen Penilaian Masyarakat Terhadap Pelayanan Rumah Sakit di Malang," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 993–1000, Oct. 2024, doi: 10.25126/jtiik.2024117979.

[9] H. Imaduddin, F. Yusfida A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach." [Online]. Available: www.ijacsa.thesai.org

[10] A. F. RIZKI, W. Prihartono, and F. Rohman, "Analisis Sentimen Ulasan Google Maps Rumah Sakit Khalishah Di Cirebon Dengan Algoritma Naive Bayes," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6309.

[11] E. I. Setiawan, P. Tjendika, J. Santoso, F. X. Ferdinandus, Gunawan, and K. Fujisawa, "Aspect-Based Sentiment Analysis of Healthcare Reviews from Indonesian Hospitals based on Weighted Average Ensemble," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1579–1596, Dec. 2024, doi: 10.47738/jads.v5i4.328.

[12] A. N. Azhar and M. L. Khodra, *Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis*. IEEE, 2020.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Online. [Online]. Available: https://huggingface.co/

[14] Dhendra and V. Gayuh Utomo, "Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews," *Jurnal Transformatika*, vol. 23, no. 1, pp. 86–95, Jul. 2025, doi: 10.26623/transformatika.v23i1.12095.

[15] A. A. Azhari, Y. Sibaroni, and S. S. Prasetiyowati, "Detection of Indonesian Hate Speech in the Comments Column of Indone-sian Artists' Instagram Using the RoBERTa Method," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 764–773, Aug. 2023, doi: 10.29100/jipi.v8i3.3898.

[16] L. Simanihuruk and H. Suparwito, "Long Short-Term Memory and Bidirectional Long Short-Term Memory Algorithms for Sentiment Analysis of Skintific Product Reviews," *ITM Web of Conferences*, vol. 71, p. 01016, 2025, doi: 10.1051/itmconf/20257101016.

[17] F. Ayu, D. Aryanti, A. Luthfiarta, D. Adiwinata, and I. Soeroso, "Aspect-Based Sentiment Analysis with LDA and IndoBERT Algorithm on Mental Health App: Riliv," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[18] V. Vinardo and I. Wasito, "Two-Stage Sentiment Analysis on Indonesian Online News Using Lexicon-Based," *sinkron*, vol. 8, no. 4, pp. 2109–2119, Oct. 2023, doi: 10.33395/sinkron.v8i4.12769.

[19] R. Randy Suryono, "Sentiment Classification of Indonesian-Language Roblox Reviews Using IndoBERT with SMOTE Optimization," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[20] Mahfudhoh and I. Muslimin, "Pengaruh Kualitas Pelayanan Terhadap Kepuasan Pasien Pada Rumah Sakit Umum Daerah Kota Cilegon," *Jurnal Ilmiah Manajemen*, vol. Vol. 8 No. 1, 2020, Apr. 2020.

[21] M. C. Huwae *et al.*, "Analysis of Inpatient Care Service Quality in Hospitals Based on National Standards: A Literature Review," *Healthy Tadulako Journal (Jurnal Kesehatan Tadulako*, vol. 11, no. 1, 2025.

[22] A. Diaz Pratama, K. P. Legawa, N. Waya, and I. Bernarto, "The Effect Of Hospital Service Management Quality On Patient Satisfaction At Siloam Lippo Village Hospital Building B Karawaci," 2024.

[23] F. Farias, T. Ludermir, and C. Bastos-Filho, "Similarity Based Stratified Splitting: an approach to train better classifiers," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.06099

[24] H. T. Madabushi, E. Kochkina, and M. Castelle, “Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data,” Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.11563

[25] S. Rahmawati, A. Wibowo, and A. F. N. Masruriyah, “Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN,” *Jurnal RESTI*, vol. 8, no. 5, pp. 607–614, Oct. 2024, doi: 10.29207/resti.v8i5.5980.

[26] H. Hikmayanti, A. F. Nurmasruriyah, A. Fauzi, N. Nurjanah, and A. Nur Rani, “Performance Comparison of Support Vector Machine Algorithm and Logistic Regression Algorithm,” *International Journal of Artificial Intelegence Research*, vol. 7, no. 1, p. 1, 2023, doi: 10.29099/ijair.v7i1.1.1114.

[27] A. Nurpiana and A. W. Wijayanto, “Comparison of Models for Classification of Learning Achievement of Middle School Students in Indonesia in 2019 using the Support Vector Machine Algorithm, Conditional Inference Trees, and Random Forest,” *Jurnal Matematika, Statistika dan Komputasi*, vol. 18, no. 3, pp. 447–455, May 2022, doi: 10.20956/j.v18i3.19208.

[28] 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.

[29] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, “Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 1, p. 677, Apr. 2025, doi: 10.11591/ijeecs.v38.i1.pp677-688.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 2019, [Online]. Available: http://arxiv.org/abs/1810.04805

[31] 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*, no. Volume 15 Issue 04 Year 2023, Dec. 2023, [Online]. Available: https://ijcnis.org