Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

Empirical Evaluation of IndoBERT and LSTM for Sentiment Analysis of Tourism Reviews: A Data-Driven Study on Kenjeran Park

Devi Dwi Purwanto*[1]

[1]Department of Informatics, Widya Mandala Surabaya Catholic University, Indonesia

Email: [1]devi.dp@ukwms.ac.id

Received : Jun 17, 2025; Revised : Sep 11, 2025; Accepted : Sep 23, 2025; Published : Feb 15, 2026

Abstract

Tourism plays a pivotal role in Indonesia's economic and cultural landscape, contributing significantly to job creation, regional development, and international recognition. This study evaluates the performance of IndoBERT, a state-of-the-art Indonesian language model, and Long Short-Term Memory (LSTM) networks for sentiment classification of 2,560 Google reviews of Kenjeran Park in Surabaya, consisting of 54% positive, 28% neutral, and 18% negative sentiments. Preprocessing steps included slang replacement, stemming, stopword removal, and tokenization, with class imbalance addressed through weighted loss adjustments. IndoBERT was fine-tuned using contextual embeddings with a learning rate of 0.00005, while the LSTM model employed a 128-unit architecture trained over 150 epochs with the Adam optimizer. Experimental results show that IndoBERT achieved 87.50% accuracy, 0.7697 precision, 0.7643 recall, and 0.7643 F1-score, outperforming LSTM's 77.93% accuracy, 0.6826 precision, 0.6812 recall, and 0.6826 F1-score. This research establishes a comparative benchmark of transformer-based and RNN-based architectures for Indonesian tourism review sentiment analysis, introduces a domain-specific preprocessing pipeline with imbalance handling, and provides actionable insights for digital tourism analytics. Beyond its technical contributions, the study highlights the urgency of advancing robust natural language processing approaches for low-resource languages, thereby strengthening the field of informatics and supporting data-driven decision-making in the tourism sector.

Keywords : *Deep Learning, IndoBERT, LSTM, Sentiment Analysis,Urban Coastal Park, Tourism*

1. INTRODUCTION

Tourism is one of the sectors that has a major contribution to the economy. Tourist destinations such as Kenjeran Park in Surabaya are one of the popular destinations that attract tourists. Kenjeran Park that located in Surabaya is urban coastal parks which offer a unique combination of natural landscapes, recreational facilities, and cultural attractions within city limits, making them an important component of urban tourism development. As competition between destinations intensifies, understanding visitor perceptions through online reviews becomes critical for improving service quality and supporting sustainable tourism planning. [1], [2]

To respond to this need, visitor feedback and reviews play an essential role in monitoring service quality and guiding destination development. Reviews reflect diverse feelings and experiences, which can be further analyzed to understand public sentiment toward attractions. Sentiment analysis has therefore emerged as an effective tool to assess satisfaction levels and identify areas requiring improvement. Nevertheless, previous studies in Indonesia have mainly focused on elections, employing methods such as textblobs[3], using Naïve Bayes algorithm[4], and combine Naïve Bayes with SVM Classifier to predict sentiment from Twitter data. These approaches, however, remain dominated by lexicon-based and conventional machine learning techniques.

To be able to find out the sentiment of the review automatically, machine learning techniques such as IndoBERT and Long Short-Term Memory (LSTM) emerged, one of which was used in the research by Mandhasiya et al.[5]. BERT itself is a state of the art performance in NLP, IndoBERT is one of the BERT modes specifically for Indonesian.[6] IndoBERT is a transformer model based on Bidirectional Encoder Representations from Transformers (BERT) which is specifically for the Indonesian corpus.[7], [8] This model provides the ability to understand the context and meaning of words in a sentence better than other models.[9], [10], [11] Meanwhile, LSTM is an artificial neural network that can solve the problem of sequence in text, so it is very suitable for processing sequential text data.[12] In the research on sentiment analysis with LSTM and BERT on the election, the validation accuracy with the BERT model was 76.48%, while the LSTM model had a maximum accuracy of 87%.[13] Another study on sentiment analysis using LSTM on hospital services during the pandemic found an accuracy of 86%.[14] Another study on sentiment analysis using IndoBERT in the health sector found an accuracy of 96%.[15]

In the tourism domain, sentiment analysis has been widely explored ensemble multi-label models for hotel reviews in Vietnam outperform LSTM and CNN[16], hybrid BERT–LSTM frameworks improve POI recommendations[17], and combining sentiment with forecasting enhances tourist arrival prediction.[18] Comparative studies show that transformer-based models such as RoBERTa achieve the highest performance on large-scale review datasets,[19] while survey papers emphasize ongoing challenges in sentiment polarity and the growing role of deep learning.[20] Recently, multimodal approaches such as transforming reviews into AI-generated images with DistilBERT and CLIP demonstrate new possibilities for explainability by linking emotions with visual cues.[21]

Despite these advances, most studies on coastal tourism have focused on remote beach destinations that emphasize natural beauty, environmental sustainability, or economic impacts in suburban or rural areas. By contrast, research on urban coastal parks—such as Kenjeran Park—that integrate natural recreation with playgrounds and family-oriented facilities remains limited. This type of tourism is increasingly relevant in urban contexts where multifunctional public spaces are needed to combine leisure, cultural, and social dimensions. Exploring such destinations can provide valuable insights into tourist preferences, perceptions of comfort, aesthetic value, and socio-economic potential compared to conventional coastal areas.

Therefore, this study seeks to address the research gap by developing and implementing a sentiment analysis system for visitor reviews of Kenjeran Park in Surabaya using IndoBERT and LSTM. By analyzing sentiments in Indonesian-language tourist reviews, this research aims to generate deeper insights into visitor perceptions and satisfaction, while testing the effectiveness of IndoBERT for contextual understanding and LSTM for sequential text processing in complex review structures.

2. METHOD

This study uses a qualitative approach with a case study method that focuses on beach tourism located in the middle of the city and equipped with a playground. The research location was selected based on the criteria of a beach destination in an urban area, has a playground or family interactive space, and has active interaction between visitors and elements of the surrounding environment. Data collection began by scraping the Google review of Kenjeran Park located in Surabaya. Data analysis was carried out using the triangulation method to ensure the validity of the findings, and using thematic analysis to identify patterns related to destination sustainability. This study also considers the dynamics of urban development and how the existence of this beach-playground can support the transformation of sustainable urban tourism. Thus, this methodology not only aims to describe the phenomenon, but also evaluates and provides recommendations based on the principles of sustainable tourism for replication or development of similar models in other cities.

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

2.1. Research stage

The research framework is illustrated in Figure 1, consisting of six main stages: (1) data collection, (2) data preprocessing, (3) word embedding, (4) model design, (5) training, and (6) evaluation. To improve clarity, each stage is explained in detail in the following sub-sections. Figure 1 has also been revised to explicitly show the two parallel modeling paths: IndoBERT fine-tuning and LSTM training, both leading into the evaluation stage.

The dataset was collected from Google Reviews of Kenjeran Park, Surabaya. The scraping process was carried out using Python libraries Selenium and BeautifulSoup, which enabled automated retrieval of review text, star ratings, and metadata. To ensure data quality, the following filtering criteria were applied (1) only reviews written in Indonesian language were retained, (2) reviews shorter than five words were discarded, (3) duplicate reviews and those containing only emojis or URLs were excluded, (4) no personal identifiers (usernames or profile data) were stored to maintain ethical compliance.

In total, 3,200 raw reviews were collected, which after cleaning resulted in 2,560 usable reviews for modeling. Reviews were labeled into three sentiment classes according to the star ratings: positive (4–5 stars), neutral (3 stars), and negative (1–2 stars).

The preprocessing stage was carried out to clean and standardize the reviews before they were processed by the models. First, case folding was applied to convert all characters into lowercase, followed by the removal of punctuation marks, emoticons, and other non-alphanumeric symbols to reduce noise. Since many reviews contained informal Indonesian expressions, a slang replacement step was conducted using a custom-built slang dictionary in which common informal words such as “gue” were replaced with “saya”, and “nggak” was replaced with “tidak”. To further refine the text, stopword removal was performed using the Indonesian stopword list from Sastrawi, complemented with additional tourism-specific stopwords that were considered irrelevant for sentiment classification. Stemming was then applied using the Sastrawi stemmer to reduce each word to its root form, ensuring consistency in lexical representation. Finally, tokenization was performed to break down the cleaned text into individual tokens, with the Keras Tokenizer used for the LSTM pipeline and the WordPiece tokenizer applied for IndoBERT. Through this sequential preprocessing pipeline, the textual reviews were transformed into standardized and structured input suitable for embedding and model training.

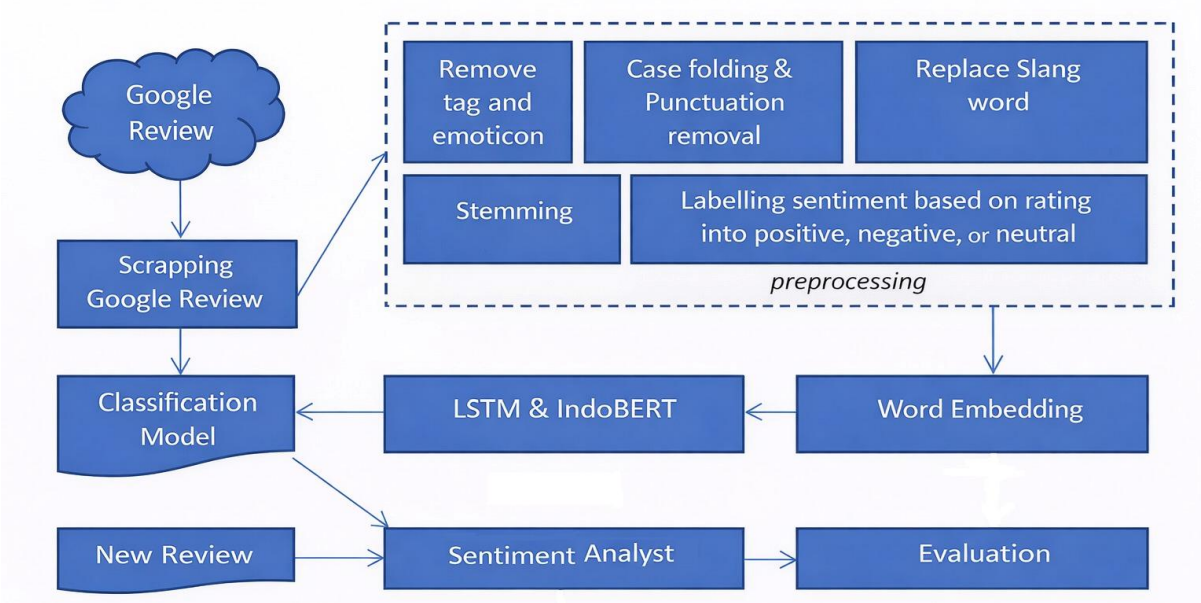

Figure 1. Research Stage

2.2. INDOBERT

This study will use IndoBERT-base, which is the basic model of IndoBERT developed based on training on an Indonesian language text corpus with a total of 5.5 billion words from various sources. This model has the ability to handle a variety of natural language processing (NLP) tasks. The architecture of this model consists of 12 transformer layers, each with 12 attention heads, and includes around 110 million parameters. There are several variants of IndoBERT-base, namely:**[22]**

a. INDOBERT-BASE-P1

This model utilizes a transfer learning approach and is trained using a very large and diverse dataset of Indonesian language texts, including news articles, Wikipedia content, and data from social media. With its broad data coverage, this model is capable of being applied to a variety of general NLP tasks.

b. INDOBERT-BASE-P2

Unlike the P1 variant, IndoBERT-base-P2 is trained on a more complex and diverse dataset, thus producing more precise output. This advantage makes IndoBERT-base-P2 more suitable for tasks that require deep language understanding, such as document classification and sentiment analysis.

Word embedding is a technique in natural language processing that aims to represent words in text form into numbers that can be understood by machine learning models. Bidirectional Encoder Representations from Transformers (BERT) is a word embedding technique that uses a transformer architecture that can capture word context so that it can better enable modeling of relationships between words in sentences. The first token of each sequence in the dataset is converted into a special classification token. In the embedding segment, when there are multiple sentences, the sentences are combined into a single sequence and distinguished using special tokens. Special tokens such as [CLS] at the beginning and [SEP] at the end of the sentence, each of which provides additional information for classification and sentence separation tasks.**[23]** In addition, positional encoding is used to preserve the order of words to vectorize the dataset. Within the BERT encoder transformer layer, the attentional head mechanism enables accurate connections between words in sentences in the dataset. This enables accurate understanding of the contextual semantic meaning of the same words in different contexts in a sentence.**[24]**

INDOBERT is a derivative of BERT which is specifically designed for processing Indonesian text in various types of tasks such as sentiment analysis, Q&A, text prediction, text generation, and text summarization.**[25]** INDOBERT shows better results than other models for classifying product and service review sentiment in vector form.**[26]** INDOBERT can determine more precise sentiment by considering the overall context of the sentence and is better compared to other word embedding techniques such as Word2Vec or GloVe.

2.3. LSTM

LSTM is a derivative method of Recurrent Neural Network (RNN) that can maintain long-term memory by training weights for sentiment analysis. LSTM stores separate memory cells in it that can be updated. The LSTM method has an architecture consisting of an input layer, a process layer, and an output layer.[27], [28] Each LSTM (Long Short-Term Memory) unit in Figure 2 has a main component in the form of a memory cell whose status at time t is stated as c_t. The process of reading and updating memory is controlled by a number of gates based on the sigmoid activation function, namely the input gate i_t, forget gate f_t, and output gate O_t. The calculation mechanism in LSTM runs as follows, at each time t, the model receives two types of external input, namely the previous hidden state h_{t-1} and the current input vector xt). The new hidden state ht is calculated based on the combination of these two inputs.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

When determining the state of a node in the hidden layer, all four components (input gates, output gates, forget gates, and input vector xt) simultaneously affect the cell state. In addition to external inputs, each gate also utilizes an internal source in the form of the previous cell state ct−1 originating from the cell block itself. The relationship between the cell state and each of these gates is called a peephole connection.[29]

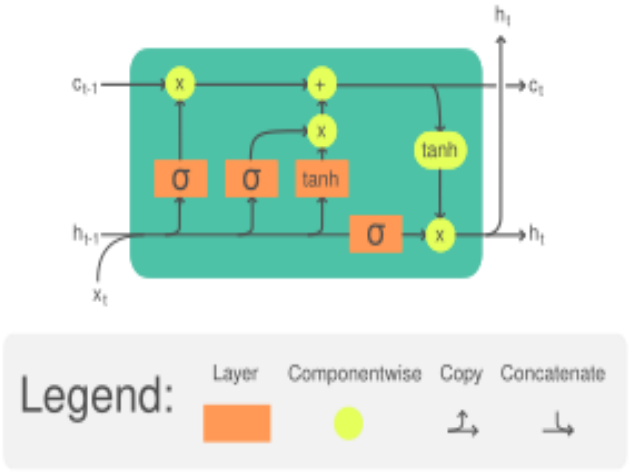

Figure 2. LSTM Architecture

Each gate plays a role in regulating the flow of information by deciding which values need to be updated or maintained through value transformations in the range 0-1. This process begins with the input gate that determines the new information added to the cell's memory state with equation 1. After that, the tanh layer produces a vector of candidate updates that will be added to the cell's memory state in equation 2. [30]

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i \quad (1)$$

$$C_t = \tanh(W_c[h_{t-1}, x_t] + b_c \quad (2)$$

The next process in the forget gate is to decide which part of the information needs to be deleted and which needs to be saved. This decision is generated by the sigmoid layer which maps the values between 0-1 in equation 3. In the memory state, it is possible for some of the values in the cell state to be deleted by multiplying the value by the result of the forget gate so that it produces a value close to 0 in equation 4.[31]

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (4)$$

At the output gate, it functions to determine the part of the cell state that is used as a hidden status in the next step. The process begins with sigmoid activation to select the part of the cell state that is the output in equation 5. From the results of equation 5, it will be continued with tanh activation to produce a new hidden status in equation 6.

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

In the deep learning model training process, there are several commonly used optimizer algorithms, such as Stochastic Gradient Descent (SGD), Adam, RMSProp, and so on. The Adam algorithm has the advantage of overcoming sparse gradient problems and is a development of the stochastic gradient descent method which is now widely adopted in various deep learning applications, including Natural Language Processing (NLP). In the Adam algorithm, m and v represent the first and second moment estimates of the gradient, while g indicates the gradient in the current mini-batch (Equations 9 and 10). Meanwhile, RMSProp is able to adaptively adjust the learning rate for each model parameter. The goal is to divide the learning rate by the exponential mean of the current gradient square, resulting in a more stable and efficient optimization process.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (7)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t \quad (8)$$

2.4. Evaluation

After getting the model and conducting a trial, it is necessary to evaluate the resulting sentiment classification. To measure model performance, accuracy, precision, recall, and F1-Score are used. [32] Accuracy value is an evaluation matrix used to measure how well a classification or prediction model can provide correct results or match existing data. Accuracy describes the percentage of success of the model in predicting the correct class or label.[33] The equation for calculating accuracy can be seen in equation 8. In addition, from the resulting confusion matrix, true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values will be obtained which will be used to calculate F1 score and precision. The equation for calculating precision can be seen in equation 10, equation 11 for calculating recall, while the F1 score can be seen in equation 12.

$$Accuracy = \frac{number\ of\ correct\ prediction}{Total\ of\ test\ data} \quad (9)$$

$$recall = \frac{TP}{TP+FN} \quad (10)$$

$$precision = \frac{TP}{TP+FP} \quad (11)$$

$$F1\ score = \frac{2.(Precision\ .recall)}{recall+precision} \quad (12)$$

3. RESULT

This section will discuss the description of the dataset used, data preprocessing, model formation, and model evaluation. Data preprocessing is an important process to remove noise in the data so that data complexity can be reduced, irrelevant word variations can be minimized, and feature representation becomes more consistent. This is very important to improve the performance of the sentiment analysis model.

3.1. Dataset

The dataset used in this study is the result of scraping tourist reviews from Kenjeran Park on the Google Review platform with a total of 2560 reviews collected. Each review category will be classified into three sentiment categories based on the star rating given by visitors. Ratings 1-2 will be given a negative label, rating 3 will be given a neutral rating, while ratings 4-5 will be given a positive rating. The distribution of data in each category can be seen in Table 1.

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

Table 1. Distribution of review dataset

Sentiment	Number of reviews	Percentage	Weight
Positive	1995	77.93%	1
Neutral	360	14.06%	5
Negative	205	8.01%	6
Total	2560	100%	

From the distribution of table 1, the number of each category is not balanced where the positive category is 77.93%. This will cause the resulting prediction to tend to be positive, to overcome this, a class weight adjustment will be carried out by giving a higher weight to classes other than positive. The comparison of these weights can be seen in the weight column in table 1.

3.2 Pre Process Data

From the review obtained through scrapping, data pre-processing will be carried out to ensure that the quality of the data entered into the model is free from noise. The pre-processing includes removing special characters, emoticons, and HTML tags, case folding and punctuation removal. The scrapping dataset uses non-standard Indonesian (slang words) so it is necessary to replace the slang words with standard forms. After replacing them with standard forms, the next step is to do stemming using Sastrawi. The next step is to label the target as positive, neutral, or negative according to the rating given. The results of this pre-processing will produce text that will be processed at the feature extraction stage. An illustration of the pre-processing can be seen in Figure 2.

Original text:	udh lama ga kesini dan bru bisa kesini lgi 2 taun yg lalu beda bgt pas masih rame? nya sekarang lebih ke terbengkalai
Teks after replacing slang word:	sudah lama tidak ke sini dan baru bisa ke sini lagi 2 tahun yang lalu beda banget pas masih ramai sekarang lebih terbengkalai
Teks after stemming:	Sudah lama tidak ke sini dan baru bisa ke sini lagi 2 tahun yang lalu beda banget pas masih ramai sekarang lebih bengkalai
Input for Word Embedding	[CLS] Sudah lama tidak ke sini dan baru bisa ke sini lagi 2 tahun yang lalu beda banget pas masih ramai sekarang lebih bengkalai [SEP]

Figure 2. Illustration of pre processing result

Based on the analysis in figure 2, we can see that (1)Ambiguity often appears in neutral reviews due to the vague use of everyday language. (2)Multi-aspect sentiment (containing both positive and negative aspects in a single sentence) makes it difficult for the model to determine the dominant polarity. (3) Sarcasm/Irony remains difficult for the model to capture, as positive expressions are sometimes used to satirize negative situations. (4) Context conflict occurs when strong words with negative or positive connotations dominate, even though the overall sentence is intended to be neutral. (5) This analysis shows that while the model's performance is quite good, challenges remain in reviews with multiple contexts, ambiguity, or sarcastic content.

3.3 Model

From the text results in example image 2, the next step is to do word embedding where each token between [cls] and [sep] will be converted into a 128-dimensional vector. The advantage of INDOBERT is that it uses pretrained contextual embedding so that each vector produced already has word semantics, overall sentence context, and relationships between words in the sentence, indoBERT used in this study uses indoBERT base p2. For IndoBERT using a learning rate of 0.00005. The next step after being converted into a vector in word embedding will be used as input in forming a model with LSTM. LSTM

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

is used to understand dependencies and capture context sequences because each sentence in the dataset has a meaning that is highly dependent on the order of words, especially travel review sentences which generally contain interrelated information in a long range of words. LSTM is able to capture the relationship between words that are far apart in a sentence through the internal memory mechanism (cell state) and 3 main LSTM gates.

Table 2. LSTM Model

Dimensi	**128**
Max sequence	Max_len
Hidden size	64
Dropout rate	0.5
Dense layer	3
Activation function	Softmax
Loss function	Sparse Categorical Crossentropy
Optimizer	Adam
Epochs	150
Batch size	32
Learning rate	0.001

From table 2, the sentiment classification model is built using Long Short-Term Memory (LSTM) based on Keras Sequential API. The initial stage begins with an embedding layer that functions to change each tokenized token into a 128-dimensional vector representation. The input_dim value is adjusted to the vocabulary size obtained from the tokenizer, while the max sequence adjusts the maximum input length (max_len) to ensure that each review has a uniform input length. This embedding process is important to change the discrete representation of words into a numeric form that can be processed by the next LSTM layer. After the embedding stage, the model applies an LSTM layer with 128 neuron units that function to read the sequence of words sequentially. LSTM is designed to capture temporal relationships between words, especially in travel review sentences that often contain time comparisons, changes in conditions, or complex emotional expressions. Because the return_sequences parameter is set to False, LSTM only produces output from the last step of the sequence, which represents the entire context of the review sentence. To reduce the risk of overfitting due to model complexity, a dropout layer of 0.5 is applied which randomly deactivates some neurons during the training process. In the final stage, the LSTM output is forwarded to a dense layer with 3 output neurons, each representing positive, neutral, and negative sentiment classes. Softmax activation is used to generate the probability of each class. The model is then compiled using the Adam optimizer, with a loss function using sparse_categorical_crossentropy which is suitable for multi-class classification with integer labels. The training process is carried out for 150 epochs with a batch size of 32, and involves validation data to monitor model performance periodically. These parameters are selected to obtain optimal convergence in learning sentiment representation from processed review data.

3.4 Evaluation Model

Model evaluation is carried out using random sampling of test data from the dataset of 20%. This evaluation is carried out to measure how well the model is able to classify the correct sentiment. Based on the evaluation results, the val_accuracy value is 77.93%, F1 Score 0.6826, Precision 0.6073, Recall 0.7793 and val_loss of 0.6725 using LSTM. Meanwhile, when using IndoBERT for 3 epochs, the accuracy results are 87.50%, F1 Score 0.7643, Precision 0.7697, and Recall 0.7607 and val_loss 0.6327.

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

From the comparison of the two models, it can be concluded that sentiment analysis with IndoBERT is better than LSTM. Figure 3 shows the confusion matrix of the IndoBERT model.

Based on the evaluation results, the model performance can be explained as follows. In the Negative class, the model achieved a precision value of 0.757, a recall of 0.683, and an F1-score of 0.718. For the Neutral class, the precision achieved was 0.615, a recall of 0.667, and an F1-score of 0.640. Meanwhile, in the Positive class, the model performed very well with a precision of 0.937, a recall of 0.933, and an F1-score of 0.935. When viewed from the macro average, the precision, recall, and F1-score values were 0.770, 0.761, and 0.764, respectively, indicating the average model performance in each class. Overall, based on the micro average, the model was able to achieve a precision, recall, and F1-score of 0.875, indicating consistent and fairly high performance overall.

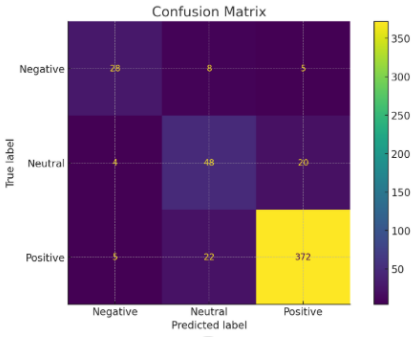

Figure 3. Confusion Matrix dengan Model IndoBERT

4 DISCUSSIONS

The results showed that IndoBERT outperformed LSTM in Indonesian sentiment analysis, with an accuracy of 87.50% and an F1-score of 0.7643. This finding also aligns with BERT's superior performance with Indonesian over classical methods in a case study of Google Apps reviews,[34] thus reinforcing the claim that the transformer approach, specifically IndoBERT, is more effective than traditional RNNs in handling Indonesian text data.

However, this study also found limitations in the neutral and negative classes, where the F1 and recall scores were relatively low. This can be explained by data imbalance, ambiguity in reviews, and inconsistent labels, such as sarcastic reviews being rated positively. This phenomenon emphasizes the importance of sarcasm detection through approaches based on hyperbole and swearword lexicons.[35], and a combination of linguistic features can help reduce misclassification due to irony.[36] Thus, sarcasm and misclassification are crucial challenges in developing sentiment analysis models for Indonesian.

The urgency of this research is crucial considering that public opinion data from social media, app reviews, and other online platforms are crucial sources of information for data-driven decision-making. This emphasizes the importance of improving model quality when dealing with ambiguously labeled text, making strategies such as domain-specific post-training[37], text augmentation, and the integration of a sarcasm detection module promising development directions for improving model performance. Overall, this research makes important contributions to the development of Indonesian-language NLP. First, it demonstrates the superiority of IndoBERT over LSTM in terms of accuracy and performance stability. Second, it highlights the model's limitations, particularly in the neutral and negative classes, indicating the need to address class imbalance and complex linguistic phenomena such as sarcasm and irony. Third, it emphasizes the relevance of sarcasm detection in reducing misclassification, in line with various recent studies. Fourth, it opens up opportunities for the application of advanced strategies such as hybrid models, text augmentation, and domain-specific post-training to

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

improve sentiment analysis performance in the future. Thus, this study not only provides empirical results, but also enriches the academic literature on Indonesian language sentiment analysis and emphasizes the importance of a transformer-based approach in the fields of informatics and computer science.

5 CONCLUSION

This study analyzed sentiment in online reviews of Kenjeran Park, Surabaya, by comparing LSTM and IndoBERT models. The results showed that IndoBERT achieved superior performance with an accuracy of 87.50% and an F1-score of 0.7643, outperforming LSTM which reached only 77.93% accuracy and an F1-score of 0.6826. The findings confirm IndoBERT's ability to capture contextual nuances in Indonesian reviews, including informal expressions and implicit sentiment, making it more effective for urban tourism sentiment analysis. This research contributes methodologically by providing empirical evidence of IndoBERT's superiority over RNN-based models in the tourism domain, and novel in applying Transformer-based approaches to Indonesian urban tourism reviews, an area still limited in academic studies.

For further research, several directions are recommended. Future studies should incorporate sarcasm and irony detection to address misclassification when ratings contradict review content, as well as conduct domain-adaptive pretraining of IndoBERT with tourism-specific corpora. Hybrid approaches that combine IndoBERT with other architectures such as CNN or Bi-LSTM could also be explored to improve robustness. Expanding the dataset to cover multiple urban tourism destinations would further strengthen generalizability and provide broader insights for tourism management and policy development.enhance robustness across sentiment classes. Finally, expanding the dataset to include reviews from multiple urban tourism destinations would allow cross-comparison and generalization of findings.

REFERENCES

[1] D. Arianto, "Aspect-based Sentiment Analysis on Indonesia's Tourism Destinations Based on Google Maps User Code-Mixed Reviews (Study Case: Borobudur and Prambanan Temples)".

[2] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 1, p. 79, Feb. 2023, doi: 10.22146/ijccs.77354.

[3] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J Big Data*, vol. 5, no. 1, p. 51, Dec. 2018, doi: 10.1186/s40537-018-0164-1.

[4] Department of Informatics,Widyatama University Bandung, Indonesia, M. Nur Habibi, and Sunjana, "Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis," *IJMECS*, vol. 11, no. 11, pp. 22–30, Nov. 2019, doi: 10.5815/ijmecs.2019.11.04.

[5] D. G. Mandhasiya, H. Murfi, A. Bustamam, and P. Anki, "Evaluation of Machine Learning Performance Based on BERT Data Representation with LSTM Model to Conduct Sentiment Analysis in Indonesian for Predicting Voices of Social Media Users in the 2024 Indonesia Presidential Election," in *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia: IEEE, Aug. 2022, pp. 441–446. doi: 10.1109/ICOIACT55506.2022.9972206.

[6] D. Sebastian, H. D. Purnomo, and I. Sembiring, "BERT for Natural Language Processing in Bahasa Indonesia," in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Bandung, Indonesia: IEEE, Dec. 2022, pp. 204–209. doi: 10.1109/ICICyTA57421.2022.10038230.

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, *arXiv*. doi: 10.48550/ARXIV.2011.00677.

[8] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models," *J. Ilm. Tek. Elektro Komput. Dan Inform*, vol. 9, no. 3, pp. 746–757, Aug. 2023, doi: 10.26555/jiteki.v9i3.26490.

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

[10] T. I. Z. M. Putra, S. Suprapto, and A. F. Bukhori, "Model Klasifikasi Berbasis Multiclass Classification dengan Kombinasi Indobert Embedding dan Long Short-Term Memory untuk Tweet Berbahasa Indonesia," *JISTED*, vol. 1, no. 1, pp. 1–28, Nov. 2022, doi: 10.35912/jisted.v1i1.1509.

[11] R. Merdiansah, S. Siska, and A. Ali Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *JIKOMSI*, vol. 7, no. 1, pp. 221–228, Mar. 2024, doi: 10.55338/jikomsi.v7i1.2895.

[12] T. Mou and H. Wang, "Online comments of tourist attractions combining artificial intelligence text mining model and attention mechanism," *Sci Rep*, vol. 15, no. 1, p. 1121, Jan. 2025, doi: 10.1038/s41598-025-85139-3.

[13] M. Khadapi and V. M. Pakpahan, "Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024," vol. 6, 2024.

[14] A. Rolangon, A. Weku, and G. A. Sandag, "Perbandingan Algoritma LSTM Untuk Analisis Sentimen Pengguna Twitter Terhadap Layanan Rumah Sakit Saat Pandemi Covid-19," *TeIKa*, vol. 13, no. 01, pp. 31–40, May 2023, doi: 10.36342/teika.v13i01.3063.

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

[16] V.-H. Nguyen, N. Nguyen, T.-H. Nguyen, Y.-N. Nguyen, M.-T. Dinh, and D. Doan, "Customer emotion detection and analytics in hotel and tourism services using multi-label classificational models based on ensemble learning," *Ann Oper Res*, Jan. 2025, doi: 10.1007/s10479-024-06434-2.

[17] A. Noorian, A. Harounabadi, and M. Hazratifard, "A sequential neural recommendation system exploiting BERT and LSTM on social media posts," *Complex Intell. Syst.*, vol. 10, no. 1, pp. 721–744, Feb. 2024, doi: 10.1007/s40747-023-01191-4.

[18] H. Laaroussi, F. Guerouate, and M. Sbihi, "Incorporating Deep Learning and Sentiment Analysis on Twitter Data to Improve Tourism Demand Forecasting," in *Digital Technologies and Applications*, vol. 669, S. Motahhir and B. Bossoufi, Eds., in Lecture Notes in Networks and Systems, vol. 669. , Cham: Springer Nature Switzerland, 2023, pp. 150–158. doi: 10.1007/978-3-031-29860-8_16.

[19] S. Srianan, A. Nanthaamornphong, and C. Phucharoen, "Advancing tourism sentiment analysis: a comparative evaluation of traditional machine learning, deep learning, and transformer models on imbalanced datasets," *Inf Technol Tourism*, Aug. 2025, doi: 10.1007/s40558-025-00336-0.

[20] H. M. U. Ali, Q. Farooq, A. Imran, and K. El Hindi, "A systematic literature review on sentiment analysis techniques, challenges, and future trends," *Knowl Inf Syst*, vol. 67, no. 5, pp. 3967–4034, May 2025, doi: 10.1007/s10115-025-02365-x.

[21] V. Calderón-Fajardo, I. Rodríguez-Rodríguez, and M. Puig-Cabrera, "From words to visuals: a transformer-based multi-modal framework for emotion-driven tourism analytics," *Inf Technol Tourism*, July 2025, doi: 10.1007/s40558-025-00334-2.

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

[23] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on*

Vol. 7, No. 1, February 2026, Page. 463-474
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.4901

Natural Language Processing, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[24] A. Zevana and D. Riana, "TEXT CLASSIFICATION USING INDOBERT FINE-TUNING MODELING WITH CONVOLUTIONAL NEURAL NETWORK AND BI-LSTM," *J. Tek. Inform. (JUTIF)*, vol. 4, no. 6, pp. 1605–1610, Jan. 2024, doi: 10.52436/1.jutif.2023.4.6.1650.

[25] S. Alaparthi and M. Mishra, "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey," 2020, *arXiv*. doi: 10.48550/ARXIV.2007.01127.

[26] P. F. Supriyadi and Y. Sibaroni, "Xiaomi Smartphone Sentiment Analysis on Twitter Social Media Using IndoBERT," vol. 10, no. 1, 2023, doi: doi.org/10.30865/jurikom.v10i1.5540.

[27] A. Kumar and R. Rastogi, "Attentional Recurrent Neural Networks for Sentence Classification," in *Innovations in Infrastructure*, vol. 757, D. Deb, V. E. Balas, and R. Dey, Eds., in Advances in Intelligent Systems and Computing, vol. 757. , Singapore: Springer Singapore, 2019, pp. 549–559. doi: 10.1007/978-981-13-1966-2_49.

[28] D. D. Purwanto, "Comparison of Premium Rice Price Prediction in East Java with ARIMA and LSTM (Case Study: National Food Agency Data)," 2024.

[29] Winda Kurnia Sari, D. P. Rini, Reza Firsandaya Malik, and Iman Saladin B. Azhar, "Multilabel Text Classification in News Articles Using Long-Term Memory with Word2Vec," *RESTI*, vol. 4, no. 2, pp. 276–285, Apr. 2020, doi: 10.29207/resti.v4i2.1655.

[30] Muhammad Ikram Kaer Sinapoy, Yuliant Sibaroni, and Sri Suryani Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 7, no. 3, pp. 657–662, June 2023, doi: 10.29207/resti.v7i3.4830.

[31] K. S. Witanto, N. A. Sanjaya Er, A. E. Karyawati, I. G. A. G. A. Kadyanan, I. K. G. Suhartana, and L. G. Astuti, "Implementasi LSTM Pada Analisis Sentimen Review Film Menggunakan Adam Dan RMSprop Optimizer," *JLK*, vol. 10, no. 4, p. 351, June 2022, doi: 10.24843/JLK.2022.v10.i04.p05.

[32] T. B. Rohman, D. D. Purwanto, and J. Santoso, "Sentiment Analysis Terhadap Review Rumah Makan di Surabaya Memanfaatkan Algoritma Random Forest".

[33] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM: Comparative Evaluation of SVM Kernels for Sentiment Classification in Fuel Price Increase Analysis," *MALCOM*, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

[34] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken Dw, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, Malang Indonesia: ACM, Sept. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.

[35] N. Arlim *et al.*, "Dictionary-based extraction of hyperbole and swear words for sarcasm detection in Indonesian Tweets," *Int. j. inf. tecnol.*, vol. 17, no. 5, pp. 2671–2678, June 2025, doi: 10.1007/s41870-024-02361-4.

[36] A. Kumar, S. R. Sangwan, A. K. Singh, and G. Wadhwa, "Hybrid Deep Learning Model for Sarcasm Detection in Indian Indigenous Language Using Word-Emoji Embeddings," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 5, pp. 1–20, May 2023, doi: 10.1145/3519299.

[37] N. P. I. Maharani, Y. Yustiawan, F. C. Rochim, and A. Purwarianti, "Domain-Specific Language Model Post-Training for Indonesian Financial NLP," 2023, *arXiv*. doi: 10.48550/ARXIV.2310.09736.