Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

# Sentiment Analysis of Google Maps Reviews on Temple Tourism in Central Java Using IndoBERT Embeddings and BiLSTM

**Ranggi Praharaningtyas Aji[1], Primandani Arsi[2]**

[1]Information System Department, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia

[2]Informatics Department, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia

**Received:**
October 1, 2025
**Revised:**
April 11, 2026
**Accepted:**
May 30, 2026
**Published:**
June 22, 2026

Corresponding Author:
**Author Name*:**
Ranggi Praharaningtyas Aji
**Email*:**
ranggi.p.aji@amikompurwokerto.ac.id

DOI:
10.63158/journalisi.v8i3.1589



**Abstract.** The rapid growth of user-generated content provides valuable insights into tourists' perceptions of destinations. This study analyzes sentiment in Google Maps reviews of temple tourism destinations in Central Java using IndoBERT embeddings and a Bidirectional Long Short-Term Memory (BiLSTM) model. A total of 10,714 Indonesian-language reviews were collected through web scraping and processed through preprocessing, pseudo-labeling, embedding generation, and model training. To prevent data leakage, the dataset was divided into stratified training and testing sets, while Random OverSampling (ROS) was applied only to the training data. Since manually annotated labels were unavailable, sentiment categories were generated automatically using a pre-trained IndoBERT classifier. The BiLSTM model achieved 80.25% accuracy on the imbalanced dataset and approximately 95% accuracy against IndoBERT-generated pseudo-labels under balanced training conditions. Improvements in Macro F1-score and balanced accuracy indicate better recognition of minority classes. However, the results should be interpreted cautiously because pseudo-labeling and oversampling may affect performance. Overall, this exploratory study demonstrates the potential of IndoBERT and BiLSTM for Indonesian tourism sentiment analysis while highlighting the need for human-annotated data and stronger validation in future research.

**Keywords:** Sentiment analysis, Google Maps reviews, IndoBERT, BiLSTM, pseudo-labeled sentiment classification.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

## 1. INTRODUCTION

The tourism sector is one of the strategic pillars that plays a crucial role in promoting economic growth in many countries [1]. In Indonesia, tourism development constitutes an integral part of national development that is implemented systematically and in a well-planned manner [2]. The development of the tourism sector is carried out through the enhancement of tourism products and services as well as improvements in the quality of tourist attractions. Supporting tourism strategies include improving facilities and increasing visitor satisfaction [3]. Tourist attractions have become an important sector in supporting regional economic development, particularly in areas endowed with abundant natural and cultural resources [4] [5]. Among the various types of tourist destinations in Indonesia, historical tourism particularly temple tourism represents one of the cultural tourism attractions with significant historical value and international recognition as world heritage sites. Several temples located on the island of Java, particularly in the provinces of Central Java and the Special Region of Yogyakarta, have become major attractions for both domestic and international tourists. Temple tourism not only serves as a center for cultural preservation but also functions as an important component of cultural tourism that contributes to regional economic growth.

Social media has increasingly played a significant role as a platform for disseminating information about tourist destinations. Its presence enables travelers to share their experiences and opinions, thereby providing opportunities for destination managers to better understand tourist perceptions and improve service quality. One of the most widely used platforms is Google Maps, which allows users to access various types of information, including location details, images, and visitor reviews of tourist destinations. These reviews are not merely personal records of experiences but also contain valuable information regarding visitor satisfaction levels. Such information can serve as an important resource for evaluating and improving the quality of services provided by destination managers[6]. The rapid growth of user-generated opinions across digital platforms and social media has created a demand for automated systems capable of effectively identifying and classifying user reviews [7] [8].

Sentiment analysis is a text mining technique used to categorize opinions expressed in textual data into positive, negative, or neutral sentiments [9]. This method is particularly

useful for processing large volumes of textual data in a systematic and objective manner. Through sentiment analysis, tourism managers can easily understand public perceptions of the destinations they manage. However, the complexity of sentiment analysis increases when mixed sentiments occur, where visitors may praise architectural beauty while simultaneously criticizing facilities or accessibility within the same sentence[10]. The inability of models to recognize the dominant sentiment in such ambiguous data may reduce classification accuracy and lead to prediction errors. Therefore, more advanced computational approaches are required to address these challenges.

A study conducted by Norlina Mohd Sabri in 2024 applies the Support Vector Machine (SVM) algorithm to analyze tourism sentiment based on TikTok social media data. The study shows that SVM is capable of classifying user sentiment with reasonably good performance on short text data. However, the approach still relies on traditional feature representations and does not deeply capture semantic context. Furthermore, the study does not explicitly address class imbalance or the complexity of longer and more diverse language structures, such as those found in tourism reviews on the Google Maps platform[11]. The model also encounters difficulties in distinguishing neutral sentiment. These findings indicate that although conventional machine learning methods can perform sentiment classification, the resulting accuracy remains inconsistent and tends to be suboptimal. This limitation arises from the inability of traditional machine learning algorithms to capture semantic meaning and word order in text, particularly in the Indonesian language, which has complex structures and contextual nuances.

To address these limitations, deep learning approaches have increasingly been applied in sentiment analysis research. Deep learning methods have been shown to be more effective in handling complex textual data and nonlinear patterns, although they require greater computational resources [12]. One of the deep learning models widely used in sentiment analysis is Bidirectional Long Short-Term Memory (BiLSTM). BiLSTM is a neural network architecture that employs bidirectional LSTM units to capture both past and future contextual information simultaneously [13]. This model is capable of capturing contextual dependencies in sequential data and processing them through bidirectional layers to obtain richer contextual information [14]. As a result, the BiLSTM model demonstrates strong sensitivity in analyzing polarized public opinions through a context-aware approach [15].

Despite the growing number of sentiment analysis studies in the tourism domain, several important challenges remain insufficiently addressed, particularly in the context of Indonesian-language reviews. Previous studies have predominantly relied on conventional machine learning algorithms, which often struggle to capture contextual semantics and long-range word dependencies. In addition, issues such as class imbalance and the limited availability of manually annotated sentiment datasets continue to hinder the development of robust sentiment classification systems.

To overcome the scarcity of manually labeled data, pseudo-labeling has emerged as a practical weak-supervision strategy for large-scale sentiment analysis. In this approach, sentiment labels are automatically generated using a pre-trained model, enabling the construction of large datasets without the substantial time and cost required for manual annotation. Although pseudo-labeling facilitates scalable analysis of user-generated content, the resulting labels may inherit prediction biases from the pre-trained model and therefore require careful interpretation during model evaluation.

To address these challenges, this study proposes a sentiment analysis framework that integrates IndoBERT-based text representation with a Bidirectional Long Short-Term Memory (BiLSTM) model. Unlike previous studies, this research focuses specifically on Google Maps reviews of temple tourism destinations in Central Java and investigates the impact of class imbalance handling using Random OverSampling. The study utilizes IndoBERT-generated sentiment labels as supervision signals and evaluates how effectively BiLSTM can learn sentiment patterns from large-scale tourism review data.

Therefore, the main contributions of this study are threefold: (1) the development of a sentiment analysis pipeline combining IndoBERT embeddings and BiLSTM, (2) an empirical evaluation of class imbalance handling in tourism review data, and (3) an exploratory analysis of tourist sentiment based on large-scale Google Maps review data.

It should be noted that manually annotated sentiment labels were not available for this study. Consequently, the proposed framework relies on automatically generated pseudo-labels and should be regarded as an exploratory investigation rather than a fully validated sentiment classification system. The findings primarily provide insights into the feasibility of combining IndoBERT representations and BiLSTM architectures for large-scale tourism

review analysis while highlighting methodological considerations for future studies involving human-annotated datasets.

## 2. METHODS

This study follows a systematic research workflow starting from data collection to model evaluation. Figure 1 presents the overall research framework, which consists of several main stages, including data collection, preprocessing, data labeling, word embedding, data balancing, dataset splitting, BiLSTM model training, and model evaluation. This structured process ensures that the sentiment classification model is developed and assessed in a comprehensive and reliable manner.

It is important to note that the sentiment labels used in this study are generated automatically using a pre-trained IndoBERT model (pseudo-labeling), rather than manually annotated ground-truth labels. This approach enables large-scale data labeling but may introduce bias, as the resulting labels reflect the behavior of the pre-trained model rather than independent human judgment. Therefore, the findings of this study should be interpreted with this limitation in mind.

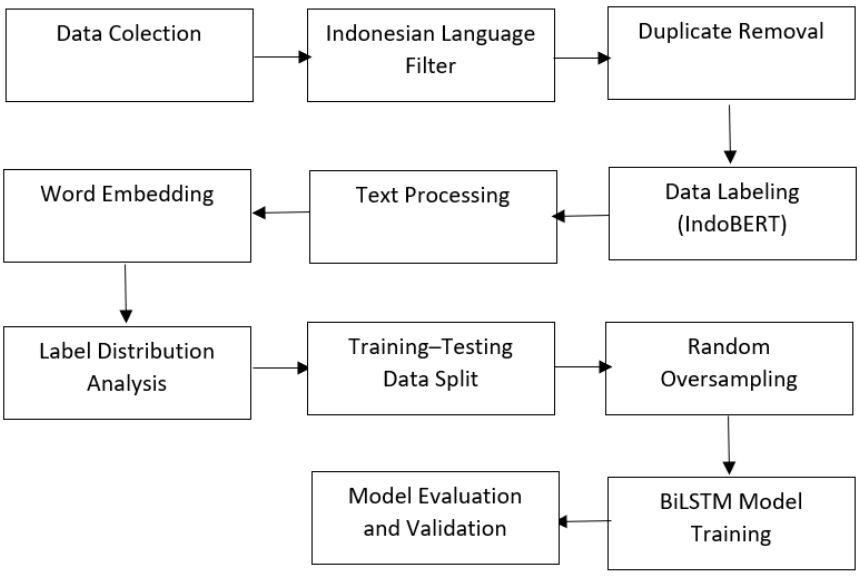


**Figure 1.** Research Methodology Workflow

Based on Figure 1, the research framework applied in this study is illustrated. The research process begins with the collection of user review data. The collected data are

then processed through a text preprocessing stage to prepare the dataset for the modeling process. The cleaned data are subsequently assigned sentiment labels automatically using IndoBERT and transformed into word embeddings. After the preprocessing and embedding stages, the dataset is first divided into training and testing sets using a stratified split approach. This step is performed prior to any data balancing procedure to prevent data leakage between the training and testing subsets.

Subsequently, data balancing is carried out using the Random OverSampling (ROS) technique, which is applied only to the training data to address class imbalance. The testing data are kept unchanged to ensure an unbiased evaluation of the model performance on unseen data. Finally, the study proceeds with BiLSTM model training using the prepared training data, followed by model evaluation to measure the performance of the proposed model.

### 2.1. Data Collection

The data used in this study were obtained from user reviews available on the Google Maps platform accessed through its website. The data collection process was conducted using a web scraping technique with the assistance of the Insta Data Scraper tool to extract reviews from Google Maps pages associated with temple tourism destinations in Central Java Province. The selected destinations include Borobudur Temple, Prambanan Temple, Plaosan Temple, Mendut Temple, and Gedong Songo Temple. The data collection was carried out on May 25, 2025, resulting in a total of 11,030 reviews, which were stored in CSV format for further processing and analysis.

### 2.2. Language Filtering

The initial dataset contained reviews written in several languages, including Indonesian, English, Malay, and Javanese. A language filtering process was performed using the langdetect library, in which only reviews identified as Indonesian (coded as “id”) were retained. This step is crucial because the model used in this study, IndoBERT, is specifically designed for the Indonesian language. Therefore, the model requires linguistically consistent input data to function optimally.

### 2.3. Duplicate Removal

This step involves identifying and removing identical text reviews within the dataset. Reviews that appear exactly the same often resulting from repeated submissions or automated postings are removed, and only the first occurrence of each review is retained. The primary objective of this process is to prevent bias and overfitting in the model. As a result, the dataset becomes more unique, clean, and of higher quality for training the sentiment analysis model.

### 2.4. Data Labeling

At this stage, sentiment labels were assigned automatically using a pre-trained IndoBERT-based sentiment classification model. The model employed in this study was **IndoBERTweet Sentiment Classification** developed for Indonesian-language sentiment analysis and publicly available through the Hugging Face repository. The model was selected because it has been pre-trained and fine-tuned on Indonesian textual data, enabling it to capture linguistic characteristics and contextual information specific to the Indonesian language. The labeling process began by tokenizing each review using the corresponding AutoTokenizer. The tokenized text was then passed to the pre-trained classifier to generate prediction scores (logits) for each sentiment category. The sentiment label of a review was determined using the argmax function, which selects the class with the highest prediction score. The generated labels were mapped into three sentiment categories: **Positive (0), Neutral (1), and Negative (2)**. The resulting labeled dataset was subsequently used as the target variable for the BiLSTM training process [16].

The use of pseudo-labeling enables large-scale sentiment annotation without requiring extensive manual labeling efforts. This approach is particularly useful when dealing with thousands of reviews, where manual annotation would be time-consuming and resource-intensive. In addition, pseudo-labeling provides a practical weak-supervision strategy for constructing sentiment datasets in low-resource settings. However, several limitations should be acknowledged. The sentiment labels generated in this study do not represent independently validated ground-truth annotations because no human-labeled reference dataset was available. Consequently, the assigned labels may inherit classification errors or biases originating from the pre-trained IndoBERT model. Furthermore, the evaluation results obtained in this study primarily reflect the BiLSTM model's ability to learn patterns

from IndoBERT-generated pseudo-labels rather than its capability to predict objectively verified human sentiment judgments. Therefore, the findings should be interpreted as an exploratory pseudo-label-based sentiment classification experiment rather than a fully validated sentiment analysis system. To ensure transparency and reproducibility, the pseudo-labeling process was applied consistently to all reviews after data cleaning and before the word embedding stage. The labeled dataset then served as the foundation for subsequent sentiment analysis and model evaluation.

### 2.5. Text Preprocessing

The preprocessing stage consists of several steps: (1) Cleaning, which involves removing special characters, URLs, emoticons, punctuation marks, and irrelevant numbers; (2) Case Folding, converting all text into lowercase letters to standardize the format; (3) Text Normalization, transforming informal words or slang into their standard forms according to the Indonesian dictionary (KBBI); (4) Tokenization, splitting text into individual words or tokens; (5) Stopword Removal, eliminating common words that do not carry significant meaning, such as “yang”, “di”, and “ke”, using the NLTK library, while retaining the word “tidak” because it conveys negative meaning; and (6) Stemming, converting affixed words into their root forms using the Sastrawi library for the Indonesian language[17] [18] [19].

### 2.6. Word Embedding

After preprocessing, the text data are transformed into numerical vectors using a word embedding technique. This study employs the IndoBERT (Indonesian BERT) model to generate vector representations with a dimension of 768. IndoBERT is selected because it has been specifically pre-trained on Indonesian language corpora, enabling it to capture contextual relationships and semantic meanings in Indonesian text more effectively than other word embedding models [20] [21].

### 2.7. Label Distribution Analysis

Label distribution analysis is a crucial stage for evaluating the balance of sample sizes across different sentiment classes within the dataset. This step aims to identify class imbalance, a condition in which the number of samples in one category (majority class) significantly exceeds those in other categories (minority classes). Such imbalance can be problematic because it may cause the model to become biased toward the majority class, resulting in misleadingly high accuracy while performing poorly in recognizing minority

classes. To prevent the model from ignoring minority samples, this imbalance must be addressed through techniques such as oversampling (increasing minority class samples) or undersampling (reducing majority class samples) [22].

### 2.8. Data Splitting

After the preprocessing and embedding stages, the dataset is first divided into training and testing sets with a ratio of 80:20. The splitting process is performed using a stratified split method to maintain the same class proportion in both subsets. This step is conducted prior to any data balancing procedure to prevent data leakage between the training and testing datasets. The training data are used for model learning, while the testing data are reserved for evaluating the model's performance on previously unseen data. Maintaining the original distribution in the testing set ensures that the evaluation results remain unbiased and reflective of real-world conditions[23].

### 2.9. Random OverSampling

To address the issue of class imbalance in the dataset, the Random OverSampling (ROS) technique is applied. This method works by randomly duplicating samples from the minority classes (neutral and negative) until their number becomes comparable to that of the majority class (positive). In this study, ROS is applied only to the training dataset after the data splitting stage, while the testing dataset remains unchanged. This approach is adopted to prevent data leakage, ensuring that duplicated samples do not appear in both training and testing subsets. The ROS technique is chosen because it is simple yet effective in balancing class distribution without requiring the generation of complex synthetic data. This data balancing process is essential to reduce model bias toward the majority class and to improve the model's ability to recognize all sentiment classes more fairly and accurately [24].

### 2.10. Model Training

The vector representations generated by IndoBERT are used as input to train the BiLSTM model. These embeddings are first adjusted into a format compatible with the LSTM architecture. The BiLSTM model employed in this study consists of two bidirectional LSTM layers and is complemented with Batch Normalization and Dropout layers to mitigate overfitting. The training process is conducted with a maximum limit of 50 epochs and monitored using an EarlyStopping mechanism to terminate the training when no further

improvement in validation loss is observed. Two training scenarios are implemented: one using the original imbalanced dataset and another using the oversampled dataset with a more balanced class distribution. The learning process is carried out iteratively and evaluated using a validation set, while performance changes are recorded to observe the model's progression during the training phase.

### 2.11. Evaluation and Validation

Model performance is evaluated using several metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). A confusion matrix is utilized to analyze detailed prediction outcomes, including correct and incorrect classifications for each sentiment class. The evaluation is conducted under two conditions: (1) using the original dataset with an imbalanced class distribution, and (2) using the balanced dataset after applying Random OverSampling. The comparison between these two conditions aims to demonstrate the effectiveness of the data balancing technique in improving the overall performance of the sentiment classification model.

## 3. RESULTS AND DISCUSSION

### 3.1. Dataset

The process of collecting tourism review data was carried out using a web scraping method on the Google Maps platform accessed through its web version. The scraping process was conducted using the Insta Data Scraper tool, and the extracted data were stored in CSV format. The data collection was performed on May 25, 2025, resulting in a total of 11,030 reviews. Examples of user reviews obtained from the scraping process are presented in Table 1. Table 1 presents sample reviews collected from Google Maps users. These reviews reflect a variety of visitor opinions and experiences, including positive, neutral, and negative sentiments.

**Table 1.** Sample of Google Maps Reviews

| No | Review Text (Indonesia) |
|---|---|
| 1 | Heran tapi juga kepo...bbrp kali datang diminta petugas bayar @10ribu,tapi selalu nggak diberi karcis bukti bayar nah kalo yg masuk ratusan orang |

| No | Review Text (Indonesia) |
|---|---|
| | tiap hari trus bukti lapor dan bayar ke pemerintahnya gimana… Nggak cuma saya ,orang2 yg masuk lain juga nggak dikasih karcis bukti bayar… |
| 2 | Candi Pawon, Pagoda Pawon dan CANDI MENDUT sama dengan Pagoda Borobudur. CANDI) saling terhubung dalam satu garis, artinya pada zaman dahulu merupakan jalur ziarah dari Pagoda Mentu ke Pagoda Borobudur melalui Pagoda Balu keindahan dan â€¦ |
| 3 | 1. Tiket untuk orang asing 20.500 sebenarnya 20.000 + 500 pajak. Tiket ini sebenarnya tiket gabungan, karena untuk dua candi, dan satu lagi candi Pawon yang letaknya tidak jauh. â€¦ |
| ..... | ..... |
| ..... | ..... |
| 11027 | Adem, tapi sayang jarang tempat berteduh kalau hujan.. masih pengembangan.. |
| 11028 | Sudah sering kesini sepertinya blom ada perubahan |
| 11029 | Harga oleh oleh disini murah, ga kaya di tempat wisata lain |
| 11030 | Lokasinya dekat parkir. Gak perlu jalan kaki jauh |

### 3.2. Language Filtering and Duplicate Removal

After the data were collected, a filtering process was performed to retain only reviews written in the Indonesian language and to remove duplicate entries. From the initial dataset obtained during the data collection stage, the language filtering and duplicate removal processes produced a cleaner dataset that was ready for further processing and analysis. Table 2 shows the output of the duplicate removal process. From the 10,792 reviews obtained after language filtering, 78 duplicate records were detected and removed, resulting in a final dataset of 10,714 reviews used for subsequent analysis.

**Table 2.** Dataset After Language Filtering and Duplicate Removal

| Category | Total Reviews |
|---|---|
| Initial Dataset (Before Duplicate Removal) | 10,792 |
| Removed Non-Indonesian Reviews | 78 |
| Final Cleaned Dataset | 10,714 |

### 3.3. Automatic Labeling

Sentiment labeling was performed automatically on tourism reviews that had passed the data cleaning stage using the IndoBERT model with a sequence classification architecture. IndoBERT is a pre-trained Indonesian language model capable of classifying text into three sentiment categories: positive, neutral, and negative. The labeling process begins with text tokenization using AutoTokenizer, followed by model inference without updating the model weights. Each review is converted into numerical representations in the form of input_ids and attention_mask, which are then processed by the model to generate logit values. The sentiment label is determined based on the highest logit value using the argmax method with the following label mapping: 0 (Positive), 1 (Neutral), and 2 (Negative). The prediction results indicate that the model is capable of distinguishing reviews based on the contextual meaning and opinion content embedded in the text.

**Table 3.** Sentiment Labeling Results Using IndoBERT

| Sentiment Text (Indonesia) | Labeling |
|---|---|
| Tempatnya bersih,lahan parkir luas dan udaranya sejuk. Cocok untuk rekreasi keluarga atau rombongan lainnya. Disini juga terdapat tokoh seperti wayang untuk di ajak berfoto | Positive |
| Jauh" dr kota medan mau lihat candi ini.. | Negative |
| utk yg suka foto, boleh mampir ke candi ini, karena bisa foto dari dekat dan tidak terlalu besar. ada toko sovenir patung kecil atau stupa kecil disampingnya, beli 1 ditawar harga yg bersahabat. Layak dikunjungi | Positive |
| Candinya masih dlm tahap pembangunan. Saat ini br sekitar 4 candi yg berdekatan. Sisanya tersebar sekitar wilayah tersebut. Jika selesai bisa jd kawasan candi besar sprt Prambanan dan Borobudur . Suasananya adem krn msh di kawasan dieng | Positive |
| Sayangnya nggk ada toilet dan mushallanya.....kebersihan area candi mantul.....petugasnya....ramah tamah....kerennn..hehe | Positive |
| Kompleks candi yang luasnya. Cocok untuk trekking ringan. | Positive |
| Masih satu kawasan dengan TWC Prambanan. ðŸ‘ â€¦ | Neutral |

Table 3 presents the results of sentiment labeling for the tourism reviews that have been classified. Reviews expressing visitor satisfaction with aspects such as cleanliness, comfort, and available facilities are categorized as positive sentiment. Reviews containing complaints related to accessibility issues or limited facilities are classified as negative sentiment, while reviews that provide general information without a clear emotional tone are categorized as neutral sentiment. All labeled data are subsequently used as the foundation for the next stage of sentiment analysis.

**Table 4.** Sentiment Label Distribution

| Label | Number of samples |
|---|---|
| Positive | 8.142 |
| Negative | 1.680 |
| Neutral | 892 |

Table 4 illustrates the frequency distribution of data across each sentiment category obtained from the labeling process. The positive class dominates the total number of reviews with 8,142 entries, followed by the negative class with 1,680 entries, and the neutral class with 892 entries. These figures indicate that the majority of visitors expressed favorable impressions of the tourism destinations under study. However, the noticeable disparity in the number of samples across sentiment classes reveals the presence of a class imbalance (imbalanced data) condition. Such imbalance requires special handling during the model training phase to ensure that the prediction results remain accurate and do not become biased toward the majority class.

### 3.4. Text Preprocessing

The preprocessing stage transforms raw data into clean and structured data suitable for further analysis. Examples of data transformation at each preprocessing stage are presented in Table 5.

**Table 5.** Text Preprocessing Results

| Processing | Samples Text (Indonesia) |
|---|---|
| Text | Heran tapi juga kepo…bbrp kali datang diminta petugas bayar @10ribu, tapi selalu nggak diberi karcis bukti bayar nah kalo yg masuk ratusan orang tiap hari trus bukti lapor dan bayar ke |

| | |
|---|---|
| | pemerintahnya gimana… Nggak cuma saya, orang2 yg masuk lain juga nggak dikasih karcis bukti bayar… |
| Cleaning | Heran tapi juga kepobbrp kali datang diminta petugas bayar ribu tapi selalu nggak diberi karcis bukti bayar nah kalo yg masuk ratusan orang tiap hari trus bukti lapor dan bayar ke pemerintahnya gimana nggak cuma saya orang yg masuk lain juga nggak dikasih karcis bukti bayar |
| Case Folding | heran tapi juga kepobbrp kali datang diminta petugas bayar ribu tapi selalu nggak diberi karcis bukti bayar nah kalo yg masuk ratusan orang tiap hari trus bukti lapor dan bayar ke pemerintahnya gimana nggak cuma saya orang yg masuk lain juga nggak dikasih karcis bukti bayar |
| Normalisasi | heran tetapi juga penasaran beberapa kali datang diminta petugas bayar ribu tapi selalu enggak diberi karcis bukti bayar nah kalau yang masuk ratusan orang tiap hari terus bukti lapor dan bayar ke pemerintahnya bagaimana enggak hanya saya orang yang masuk lain juga enggak dikasih karcis bukti bayar |
| Tokenisasi | heran, tetapi, juga, penasaran, beberapa, kali, datang, diminta, petugas, bayar, ribu, tapi, selalu, enggak, diberi, karcis, bukti, bayar, nah, kalau, yang, masuk, ratusan, orang, tiap, hari, terus, bukti, lapor, dan, bayar, ke, pemerintahnya, bagaimana, enggak, hanya, saya, orang, yang, masuk, lain, juga, enggak, dikasih, karcis, bukti, bayar |
| Stopword Removal | heran, juga, penasaran, beberapa, kali, datang, diminta, petugas, bayar, ribu, selalu, enggak, diberi, karcis, bukti, bayar, nah, kalau, masuk, ratusan, orang, tiap, hari, terus, bukti, lapor, bayar, pemerintahnya, bagaimana, enggak, orang, masuk, juga, enggak, dikasih, karcis, bukti, bayar |
| Stemming | heran, juga, penasaran, beberapa, kali, datang, minta, tugas, bayar, ribu, selalu, enggak, beri, karcis, bukti, bayar, nah, kalau, masuk, ratus, orang, tiap, hari, terus, bukti, lapor, bayar, perintah, bagaimana, enggak, orang, masuk, juga, enggak, kasih, karcis, bukti, bayar |

Table 5 presents the stages of text preprocessing applied to user review data before further analysis is conducted. In the original text stage, the data still contain informal words, abbreviations, punctuation marks, and numbers that may introduce noise. Therefore, a cleaning step is performed to remove punctuation, numbers, and unnecessary characters. Next, case folding is applied to convert all text into lowercase letters in order to standardize the format and avoid inconsistencies caused by capitalization. In the normalization stage, informal words and abbreviations are transformed into their standard forms so that the sentence meaning becomes clearer and more consistent. The tokenization stage then breaks sentences into individual word units to facilitate subsequent analysis. After that, stopword removal is performed to eliminate common words that do not significantly contribute to sentiment meaning. Finally, stemming is applied to convert words into their root forms, thereby reducing word variations and improving the effectiveness of the text classification process.

### 3.5. Word Embedding

At this stage, the word embedding process is performed to convert the review texts that have undergone preprocessing into numerical representations. This step is necessary because machine learning and deep learning algorithms can only process data in numerical form. For this purpose, the IndoBERT model (indolem/indobert-base-uncased) is utilized to generate vector representations for each review. These numerical representations are processed by the IndoBERT model through several Transformer layers to produce contextual embeddings, where each token is represented as a 768-dimensional numerical vector containing semantic information based on the context of the entire sentence. Since this study requires a single numerical representation for each review, all token embeddings within a sentence are combined using the mean pooling method. This process generates a single embedding vector that represents the overall semantic meaning of the review.

**Tabel 6.** Output Word Embedding

| Index | value 1 | value 2 | value 3 | value 4 | value 5 | value 6 | value 9 |
|---|---|---|---|---|---|---|---|
| 0 | 0.2473 | 0.2473 | 0.2473 | 0.2473 | 0.2473 | 0.2473 | 0.2473 |
| 1 | -0.2113 | -0.2113 | -0.2113 | -0.2113 | -0.2113 | -0.2113 | -0.2113 |
| 2 | 0.1615 | 0.1615 | 0.1615 | 0.1615 | 0.1615 | 0.1615 | 0.1615 |

| Index | value 1 | value 2 | value 3 | value 4 | value 5 | value 6 | value 9 |
|---|---|---|---|---|---|---|---|
| 3 | 0.1128 | 0.1128 | 0.1128 | 0.1128 | 0.1128 | 0.1128 | 0.1128 |
| 4 | -0.5510 | -0.5510 | -0.5510 | -0.5510 | -0.5510 | -0.5510 | -0.5510 |

Table 6 presents the results of the word embedding process using the IndoBERT model in the form of decimal numerical vectors. Each review is represented as a single vector containing both positive and negative values that computationally represent the meaning and contextual relationships of words. Although these numerical values cannot be interpreted directly, the embedding vectors store semantic information that differentiates each review. The embedding process produces a data shape of (10,714, 768), indicating that 10,714 reviews have been successfully converted into 768-dimensional vectors. These numerical representations are subsequently used as input features for the BiLSTM model, enabling the model to learn the relationship patterns between text embeddings and sentiment labels more effectively.

### 3.6. Label Distribution Analysis

The next stage involves label distribution analysis, which aims to analyze and visualize the distribution or proportion of data across each sentiment category (positive, negative, and neutral) within the entire cleaned review dataset. This step provides an overview of the class distribution and helps identify potential class imbalance within the dataset before the model training process.

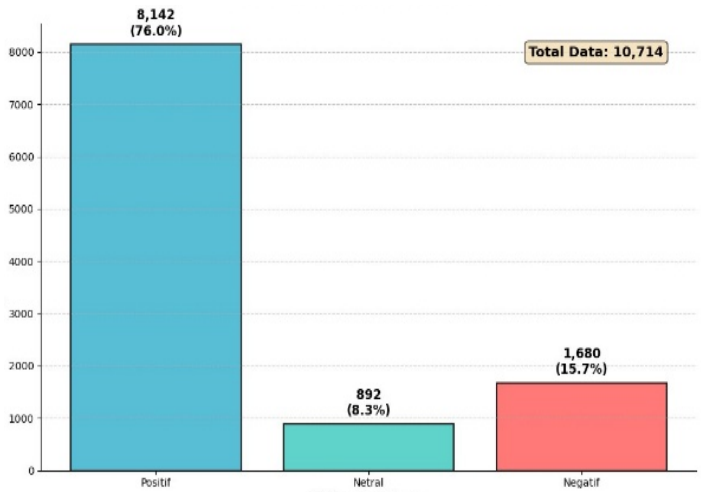


**Figure 2.** Distribution of Sentiment

Figure 2 illustrates the results of the sentiment label distribution analysis from a total of 10,714 labeled reviews. The data are divided into three categories: positive, negative, and neutral. The positive sentiment category dominates significantly with 8,142 reviews (76.0%), indicating that the majority of users expressed favorable impressions of the tourism destinations. The negative sentiment category consists of 1,680 reviews (15.7%), representing users who reported complaints or dissatisfaction. Meanwhile, the neutral category is the smallest group, with 892 reviews (8.3%). Neutral reviews generally contain factual information without a strong emotional tone. Overall, this sentiment distribution is imbalanced, as the positive class significantly exceeds the other two categories. Therefore, additional handling techniques, such as oversampling, are required to address this imbalance. This step is essential to ensure that the training model does not become biased toward the majority class before the data are used in the model training process.

### 3.7. Random Oversampling

The uneven distribution of sentiment labels was addressed using the Random OverSampling (ROS) technique. The initial dataset consisted of 10,714 reviews, with the following class distribution: Positive (8,142), Neutral (892), and Negative (1,680), indicating a significant class imbalance. After the data splitting stage, ROS was applied exclusively to the training dataset to avoid data leakage and to ensure a fair evaluation process. Deep learning models generally perform more effectively when trained on balanced datasets. Therefore, ROS was employed to increase the number of samples in the minority classes (Neutral and Negative) so that their sizes match the majority class within the training data. The oversampling process was conducted by randomly duplicating existing samples from the minority classes in the training set. Specifically, instances from the Neutral and Negative classes were randomly selected and replicated until each class reached the same number of samples as the majority class in the training subset. It is important to note that ROS does not generate new synthetic data, but simply replicates existing instances.

As a result, a balanced training dataset was obtained, where each sentiment class contains an equal number of samples. Meanwhile, the testing dataset remains unchanged and retains the original imbalanced distribution. This approach ensures that the model is trained on balanced data while still being evaluated on data that reflect real-world conditions. The embedding representations and corresponding labels in the training data

were updated to incorporate the duplicated samples, ensuring consistency in the input features used by the model. The visualization of the label distribution after applying ROS confirms that class balance is achieved within the training dataset, which is expected to support a more stable learning process and reduce bias toward the majority class.

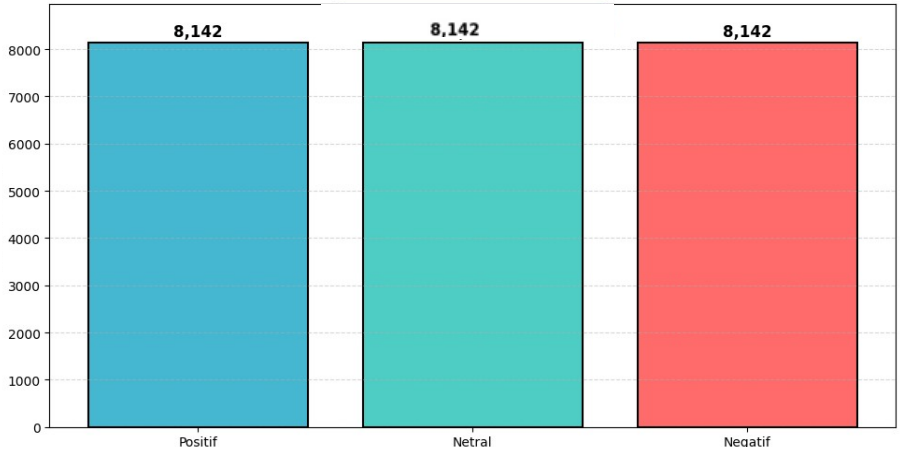


**Figure 3.** Label Distribution After Oversampling

Figure 3 presents the distribution profile of sentiment labels after the implementation of the Full Oversampling technique on the research dataset. A balanced condition has been achieved across all sentiment categories, with each class containing an equal number of reviews. The Negative, Neutral, and Positive categories each consist of 8,142 reviews, resulting in a total dataset size of 24,426 entries. The successful application of the Full Oversampling technique replicates samples from the minority classes until they match the size of the majority class, thereby effectively resolving the issue of data imbalance. Utilizing this balanced dataset as the training data for the BiLSTM model aims to optimize the learning process while minimizing bias caused by the dominance of a particular class.

### 3.8. Data Splitting

Data splitting was performed prior to the application of the Random OverSampling (ROS) technique to prevent data leakage. The dataset was divided into 80% training data and 20% testing data using a stratified split method. This approach ensures that the class distribution in both subsets remains consistent with the original imbalanced dataset.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

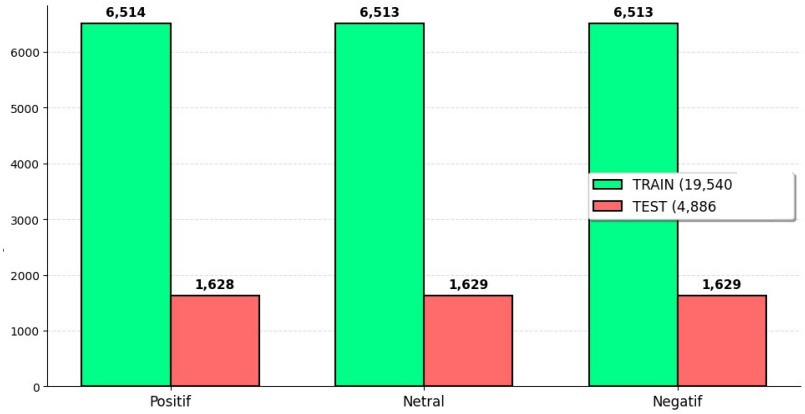


**Figure 4.** Data Splitting

Figure 4 presents the results of the data splitting process. The distribution of sentiment classes in both the training and testing subsets reflects the original data proportions, where the Positive class dominates, followed by Negative and Neutral classes. This indicates that the stratified splitting method successfully preserves the inherent class imbalance in both subsets. The training data are subsequently used for model learning and will later undergo a balancing process using ROS, while the testing data are kept unchanged to ensure an unbiased evaluation on unseen data. Maintaining the original distribution in the testing set is essential to reflect real-world conditions and to provide a reliable and unbiased assessment of model performance. This strategy also ensures that no duplicated samples from the oversampling process appear in the testing data, thereby strictly preventing data leakage.

### 3.9. BiLSTM Model Training

This stage aims to train the Bidirectional Long Short-Term Memory (BiLSTM) model, which is used to classify sentiment into three categories: positive, neutral, and negative. The data used in this stage consist of text reviews that have undergone a balancing process, ensuring that each class contains an equal number of samples. This strategy is intended to prevent bias during the model learning process. The review texts have been transformed into embedding vectors derived from semantic feature extraction using IndoBERT, with a dimensionality of 768. As a result, each review is represented with an input size of (1, 768), which is compatible with the input requirements of the BiLSTM model.

**Table 7.** BiLSTM Model Parameter Configuration

| BiLSTM Parameters | Values |
|---|---|
| Layer Type | Bidirectional LSTM |
| Number of Layers | 2 layers |
| Units (Layer 1) | 128 |
| Units (Layer 2) | 64 |
| Processing Direction | Bidirectional (forward–backward) |
| Input Shape | (1, 768) |
| Regularization (Dropout) | 0.4 and 0.5 |
| Normalization | Batch Normalization |

The BiLSTM model architecture consists of two sequential Bidirectional LSTM layers. The configuration of the model parameters is presented in Table 7. The first layer employs 128 units to capture sequential patterns from both directions forward and backward allowing the model to better understand the overall sentence context. Batch normalization is applied to stabilize the training process, while a dropout rate of 0.4 is used to reduce the risk of overfitting. The second layer utilizes 64 units, which function to refine and filter important information extracted from the previous layer. An additional dropout rate of 0.5 is applied to further enhance the model's generalization capability before proceeding to the classification stage.

The extracted features from the BiLSTM layers are then processed by dense layers to generate sentiment class predictions. The configuration of the dense layers is presented in Table 8. The first dense layer consists of 128 neurons with a ReLU activation function to capture nonlinear relationships among features, followed by batch normalization and a dropout rate of 0.5. The output layer comprises three neurons with a softmax activation function, which produces probability values for each sentiment class positive, neutral, and negative. The final prediction is determined based on the highest probability value.

**Table 8.** Dense Layer Parameters of the BiLSTM Model

| Dense Parameters | Values |
|---|---|
| Number of Dense Layers | 2 layers |
| Neurons (Dense 1) | 128 |

| Dense Parameters | Values |
|---|---|
| Activation (Dense 1) | ReLU |
| Normalization | Batch Normalization |
| Dropout | 0.5 |
| Output Neurons | 3 |
| Output Activation | Softmax |
| Output Classes | Positive, Neutral, Negative |

The BiLSTM model was subsequently trained using the balanced training dataset, while the testing data were used for validation. The training process was conducted for a maximum of 50 epochs with a batch size of 32. Model performance was monitored using accuracy and loss metrics, evaluated on both the training and validation datasets. In addition, an early stopping mechanism was implemented to prevent overfitting during the training process. A summary of the model training results for each epoch is presented in Table 9.

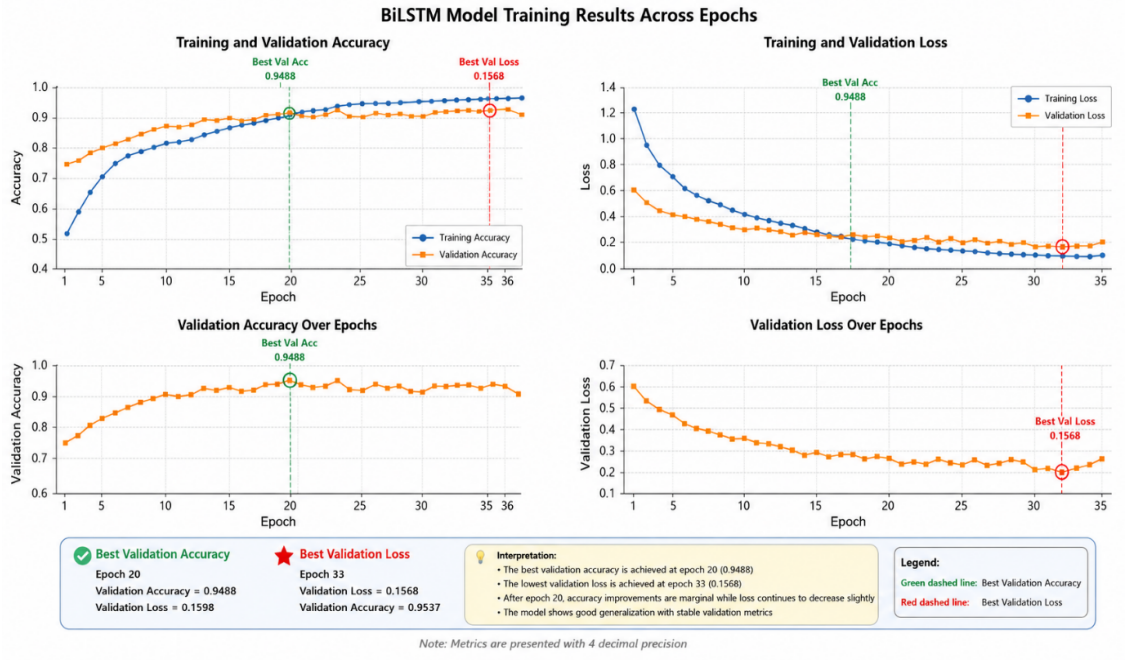


**Figure 5**. Training and validation performance of the BiLSTM model

Figure 5 illustrates the training and validation performance of the BiLSTM model across 36 epochs. Overall, both training accuracy and validation accuracy show a consistent upward trend, while training loss and validation loss decrease substantially during the

learning process, indicating that the model successfully learns meaningful patterns from the data. The highest validation accuracy was achieved at epoch 20 (94.88%) with a validation loss of 0.1598, suggesting strong generalization performance on unseen data. Although the lowest validation loss was observed at epoch 33 (0.1568), the corresponding validation accuracy (95.37%) improved only marginally compared to epoch 20. Furthermore, after epoch 20, fluctuations in validation accuracy and loss become more apparent, indicating diminishing performance gains and the potential onset of overfitting. Therefore, epoch 20 was selected as the optimal model checkpoint, as it provides the best balance between predictive performance and generalization capability. These results demonstrate that the combination of IndoBERT embeddings and the BiLSTM architecture is effective for sentiment classification of Indonesian tourism reviews while maintaining stable learning behavior throughout the training process.

### 3.10. Model Evaluation

Model evaluation was conducted under two conditions: (1) using imbalanced data and (2) using balanced data after the application of Random OverSampling. This comparison aims to assess the impact of data balancing on the performance of the BiLSTM model. Table 10 presents the evaluation results of the BiLSTM model on the imbalanced dataset.

**Table 10.** Evaluation Results of the BiLSTM Model on Imbalanced Data

| Evaluation Metrics | Values |
|---|---|
| Accuracy | 80.25% |
| Precision (Weighted) | 79.05% |
| Recall (Weighted) | 80.25% |
| F1-Score (Weighted) | 79.43% |
| F1-Score (Macro) | 61.73% |
| Balanced Accuracy | 59.70% |

Based on Table 10, the model achieves an accuracy of 80.25% on the imbalanced dataset. However, the Macro F1-Score of 61.73% and Balanced Accuracy of 59.70% indicate that the model performance is not evenly distributed across all classes. This suggests that the model is biased toward the majority class, resulting in weaker performance in recognizing minority classes.

**Table 11.** Evaluation Results of the BiLSTM Model on Balanced Data

| Evaluation Metrics | Values |
|---|---|
| Accuracy | 94.87% |
| Precision (Weighted) | 94.92% |
| Recall (Weighted) | 94.87% |
| F1-Score (Weighted) | 94.85% |
| F1-Score (Macro) | 94.63% |
| Balanced Accuracy | 94.58% |

Table 11 demonstrates a substantial improvement in model performance after the application of Random OverSampling. Most evaluation metrics exceed 94%, indicating that the model achieves highly consistent performance across all sentiment classes under the balanced training condition. The increase in Macro F1-Score from 61.73% to 94.63% suggests that the model is better able to recognize minority classes compared to the imbalanced scenario. However, this improvement may be influenced by the use of duplicated samples in the oversampling process, which can simplify the learning patterns within the training data. Therefore, these results should be interpreted with caution, as further discussed in Section 3.12.

A comparison between the imbalanced and balanced training scenarios reveals that the most substantial improvement occurs in the minority sentiment classes, namely Neutral and Negative. Although the overall accuracy increased from 80.25% to 94.87%, the improvement in Macro F1-score is considerably more informative because this metric evaluates all classes equally regardless of their frequencies. The Macro F1-score increased from 61.73% to 94.63%, indicating that the model became substantially more effective in recognizing minority-class samples after Random OverSampling was applied. Similarly, the balanced accuracy increased from 59.70% to 94.58%, demonstrating that the classifier achieved a more equitable performance across all sentiment categories. These findings suggest that the observed performance improvement is not solely attributable to better recognition of the majority class, but also reflects a significant enhancement in the model's ability to identify minority sentiment classes that were previously underrepresented in the training data.

### 3.11. ROC Analysis

Figure 6 presents the Receiver Operating Characteristic (ROC) curves for the three sentiment classes. The Area Under the Curve (AUC) values, which are close to 1.0 for all classes, indicate that the proposed BiLSTM model has strong discriminative capability in distinguishing positive, negative, and neutral sentiments. The ROC curves are positioned near the upper-left corner of the graph, reflecting high true positive rates and low false positive rates across different classification thresholds.

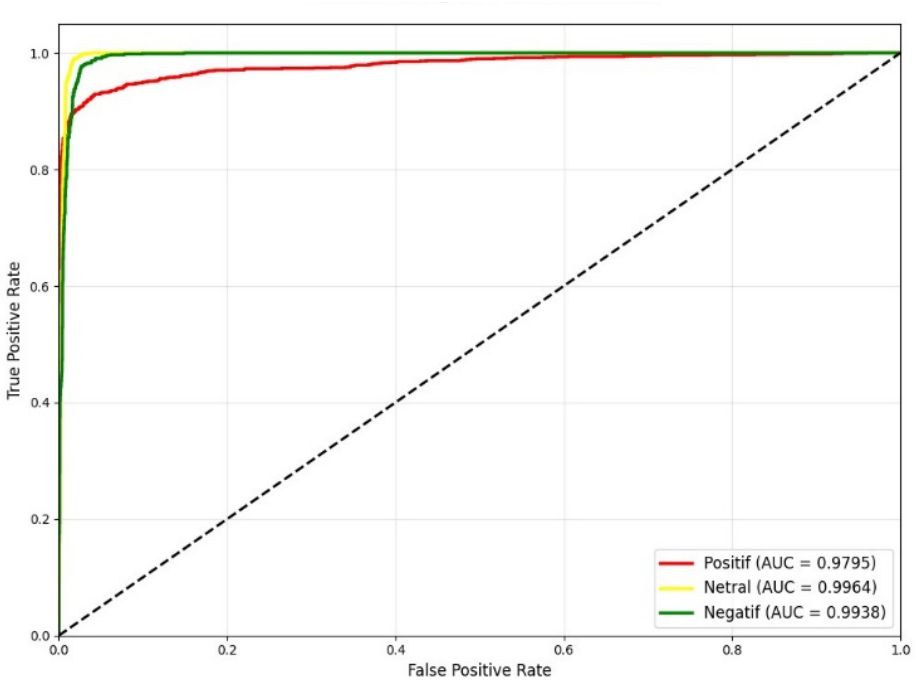


**Figure 6.** ROC Curve of the BiLSTM Model

These results are consistent with the high values of accuracy, precision, recall, and F1-score obtained under the balanced training condition. The findings suggest that the model can effectively separate sentiment categories and accurately identify sentiment patterns contained in tourism reviews. The strong ROC performance also indicates that the contextual representations generated by IndoBERT embeddings, combined with the bidirectional learning mechanism of BiLSTM, contribute positively to the sentiment classification process. Although the AUC values demonstrate excellent classification capability, the results should be interpreted together with the findings from the error analysis. Several prediction errors were still observed, particularly in the neutral class, where sentiment expressions tend to be less explicit and more context-dependent. Reviews containing mixed opinions or descriptive statements without strong emotional

indicators remain more difficult to classify accurately. Therefore, while the ROC analysis confirms the overall robustness of the proposed model, challenges remain in handling ambiguous sentiment expressions in tourism review data.

### 3.12. Prediction Distribution Analysis

Table 12 presents a comparison between the actual data distribution and the predicted results of the BiLSTM model for each sentiment category. The observations indicate that the number of predictions in each class is closely aligned with the actual values, with only a relatively small margin of difference. The Positive class shows a positive difference of 129 samples, while the Neutral and Negative classes exhibit negative differences, indicating a slight shift of predictions between classes. Despite this, the overall prediction distribution remains relatively consistent with the actual distribution, indicating stable class-wise prediction behavior, suggesting that the BiLSTM model is capable of maintaining consistent classification proportions without exhibiting excessive dominance toward any particular class.

**Table 12.** Comparison of Actual and Predicted Data

| Sentiment Class | Actual | Predicted | Difference |
|---|---|---|---|
| Positive | 1,628 | 1,499 | 129 |
| Neutral | 1,629 | 1,706 | -77 |
| Negative | 1,629 | 1,681 | -52 |

To gain deeper insights into the dominant words in each sentiment class, a WordCloud analysis was conducted. Figure 7 presents the visualization of the most frequent and prominent words in each sentiment category based on the model's predictions.


**Figure 7.** WordCloud of Dominant Words in Each Sentiment Class

Figure 7 illustrates distinct word patterns across each sentiment category. In the positive sentiment, the dominant words include "bagus" (good), "indah" (beautiful), "keren" (cool), "recommended," "menarik" (interesting), "cantik" (pretty), "asri" (pleasant), and "terawat" (well-maintained). These words reflect visitors' satisfaction and appreciation toward the tourism site. In contrast, the negative sentiment is characterized by dominant words such as "kotor" (dirty), "rusak" (damaged), "tidak terawat" (poorly maintained), "mahal" (expensive), "kecewa" (disappointed), "jelek" (bad), "kurang" (lacking), and "buruk" (poor), which indicate complaints and dissatisfaction from visitors. Meanwhile, the neutral sentiment includes dominant words such as "lokasi" (location), "sejarah" (history), "candi" (temple), "tempat" (place), "wisata" (tourism), "informasi" (information), "area," and "parkir" (parking). These words are generally descriptive and informational, without strong emotional connotations. This WordCloud analysis provides valuable insights for tourism managers to better understand aspects that are appreciated by visitors (from positive sentiment) as well as aspects that require improvement (from negative sentiment). For instance, the presence of words such as "kotor" and "tidak terawat" in the negative sentiment highlights the need to improve cleanliness and maintenance of tourism facilities. While these results indicate that the model is able to maintain relatively balanced prediction proportions across sentiment classes, a deeper interpretation of these findings, including their reliability and methodological implications, is discussed in the following section.

### 3.13. Error Analysis

To further understand the behavior of the proposed BiLSTM model, an error analysis was conducted by examining several correctly and incorrectly classified reviews from the testing dataset. This analysis is important because overall performance metrics alone cannot fully explain how the model handles different linguistic patterns and sentiment expressions. In tourism review sentiment analysis, misclassification often occurs when reviews contain ambiguous wording, mixed opinions, or limited emotional cues. Table 13 presents several examples of sentiment predictions generated by the model.

The examples presented in Table 13 indicate that the BiLSTM model performs well when sentiment expressions are explicit and strongly polarized. Reviews containing clearly positive words such as "indah" (beautiful), "nyaman" (comfortable), and "bersih" (clean) are generally classified correctly as positive sentiment. Similarly, reviews containing negative

expressions such as “kotor” (dirty) and “tidak terawat” (poorly maintained) are successfully identified as negative sentiment.

**Table 13.** Examples of Correctly and Incorrectly Classified Reviews

| Review Text | Actual label | Predicted label | Classification |
|---|---|---|---|
| Tempatnya sangat bersih, indah, dan nyaman untuk dikunjungi bersama keluarga. | Positive | Positive | Correct |
| Lokasi candi mudah dijangkau dan fasilitasnya cukup lengkap. | Positive | Positive | Correct |
| Tempatnya kotor dan beberapa fasilitas terlihat tidak terawat. | Negative | Negative | Correct |
| Masih satu kawasan dengan Prambanan dan dekat area parkir. | Neutral | Neutral | Correct |
| Candinya bagus, tetapi area parkir cukup sempit saat musim liburan. | Positive | Neutral | Incorrect |
| Tempatnya tidak terlalu ramai dan kondisi candi masih seperti sebelumnya. | Neutral | Negative | Incorrect |
| Pemandangannya indah, namun harga makanan di sekitar lokasi cukup mahal. | Neutral | Positive | Incorrect |
| Informasi sejarah yang tersedia cukup lengkap. | Neutral | Positive | Incorrect |

However, several misclassifications were observed, particularly in reviews belonging to the neutral category. This finding is consistent with previous sentiment analysis studies, which have reported that neutral sentiment is often the most difficult class to distinguish because it typically lacks strong emotional indicators. Unlike positive and negative

reviews, neutral reviews frequently contain descriptive or informational statements that may share lexical characteristics with other sentiment categories. For example, the review "Tempatnya tidak terlalu ramai dan kondisi candi masih seperti sebelumnya" was incorrectly classified as negative. Although the review mainly provides descriptive information, the presence of the phrase "tidak terlalu ramai" (not very crowded) may have been interpreted by the model as expressing dissatisfaction. Likewise, the review "Informasi sejarah yang tersedia cukup lengkap" was classified as positive instead of neutral because the phrase "cukup lengkap" (quite complete) carries a mild positive connotation.

Another source of error arises from reviews containing mixed sentiments. The review "Candinya bagus, tetapi area parkir cukup sempit saat musim liburan" combines both positive and negative opinions within a single sentence. In such cases, determining the dominant sentiment becomes challenging because the model must balance conflicting contextual signals. Similar ambiguity appears in reviews that simultaneously praise certain aspects of the destination while criticizing others. Furthermore, it is important to acknowledge that the sentiment labels used in this study were generated automatically through a pre-trained IndoBERT model. Consequently, some classification errors may originate not only from the BiLSTM model itself but also from potential inaccuracies in the pseudo-labeling process. Since the labels were not validated through manual human annotation, the actual sentiment expressed in some reviews may differ from the assigned labels.

The error analysis suggests that the proposed BiLSTM model is highly effective in recognizing reviews with clear sentiment polarity, while difficulties remain in handling neutral and mixed-sentiment reviews. Future studies should incorporate manually annotated datasets and explore attention-based or transformer-based architectures to better capture subtle contextual cues and complex sentiment expressions. The findings from the error analysis further support the importance of class balancing and contextual representation in tourism sentiment classification. The implications of these findings, together with the limitations introduced by pseudo-labeling and oversampling, are discussed in the following section.

### 3.14. Discussion

The experimental results demonstrate that the proposed BiLSTM model achieves substantially better performance when trained on the balanced dataset than on the original imbalanced dataset. The Macro F1-score increased from 61.73% to 94.63%, while the Balanced Accuracy improved from 59.70% to 94.58%. These improvements indicate that class imbalance significantly affects sentiment classification performance, particularly for minority classes. In the imbalanced scenario, the model tends to favor the majority positive class because positive reviews dominate the dataset. Consequently, the model exhibits weaker performance in recognizing neutral and negative reviews. After applying Random OverSampling, the training data become more evenly distributed, allowing the model to learn representative patterns from all sentiment categories and improving its ability to classify minority classes.

The strong performance of the proposed approach can also be attributed to the combination of IndoBERT embeddings and the BiLSTM architecture. IndoBERT provides contextualized word representations that capture semantic relationships within Indonesian text, while BiLSTM processes information from both forward and backward directions, enabling the model to better understand sentence context. This capability is particularly important in tourism reviews, where sentiment expressions are often influenced by surrounding words and contextual meaning. The high AUC values obtained across all sentiment categories further suggest that the model possesses strong discriminative capability in distinguishing positive, neutral, and negative reviews.

The error analysis presented in Section 3.13 reveals that most misclassifications occur in the neutral category. This finding is consistent with previous sentiment analysis studies, which report that neutral sentiment is generally more difficult to identify because it often lacks explicit emotional indicators. Many neutral reviews contain descriptive information about tourism destinations, facilities, or locations without expressing strong positive or negative opinions. In addition, some reviews contain mixed sentiments, where positive and negative opinions appear simultaneously within the same sentence. Such linguistic ambiguity increases classification difficulty and may lead to incorrect predictions even when overall model performance is high.

From an application perspective, the findings indicate that sentiment analysis based on Google Maps reviews can provide valuable insights for tourism destination managers. Positive sentiment patterns may help identify aspects that visitors appreciate, whereas negative sentiment patterns can reveal issues requiring improvement, such as facility maintenance, cleanliness, accessibility, or supporting services. Therefore, automated sentiment analysis has the potential to support evidence-based decision-making in tourism management and destination development.

Nevertheless, several limitations should be acknowledged. First, the sentiment labels used in this study are generated automatically through a pre-trained IndoBERT model (pseudo-labeling) rather than manually annotated by human experts. Consequently, the reported performance primarily reflects the model's ability to learn patterns from IndoBERT-generated labels and may not fully represent real-world sentiment classification performance. Second, although Random OverSampling was applied exclusively to the training data to prevent data leakage, the duplication of minority-class samples may still introduce bias and potentially inflate performance metrics. Furthermore, the absence of an independently annotated validation dataset limits the ability to assess the true quality of the generated labels.

Therefore, the reported results should be interpreted as an exploratory assessment and an upper-bound estimate under controlled experimental conditions rather than as a definitive measure of real-world sentiment classification performance. Future studies should incorporate manually annotated datasets, external validation procedures, and comparisons with transformer-only architectures to provide a more comprehensive evaluation of sentiment classification performance in Indonesian tourism reviews.

## 4. CONCLUSION

This study investigates the effectiveness of combining IndoBERT embeddings with a Bidirectional Long Short-Term Memory (BiLSTM) model for sentiment analysis of temple tourism reviews in Central Java. The experimental results demonstrate that handling class imbalance through Random OverSampling substantially improves classification performance, as reflected by the increase in Macro F1-score from 61.73% to 94.63% and Balanced Accuracy from 59.70% to 94.58%. These findings indicate that class balancing

plays an important role in improving the model's ability to recognize minority sentiment classes and achieve more consistent performance across all categories. Nevertheless, the results should be interpreted with caution. The sentiment labels used in this study were generated automatically using a pre-trained IndoBERT model through a pseudo-labeling approach rather than manually annotated ground-truth data. Consequently, the reported performance primarily reflects the model's ability to learn pseudo-labeled sentiment patterns and may not fully represent real-world sentiment classification performance. Furthermore, although Random OverSampling was applied only to the training data to prevent data leakage, the duplication of minority-class samples may still influence the learning process and evaluation outcomes. This study provides an exploratory framework for large-scale sentiment analysis of Indonesian tourism reviews and demonstrates the potential of integrating IndoBERT and BiLSTM for tourism-related text mining tasks. Future research should incorporate manually annotated datasets, perform external validation using independent review datasets, and compare the proposed approach with transformer-only architectures to obtain a more comprehensive and reliable assessment of sentiment classification performance.

## REFERENCES

[1] M. S. Viñán-Ludeña and L. M. de Campos, "Discovering a tourism destination with social media data: BERT-based sentiment analysis," *J. Hosp. Tour. Technol.*, vol. 13, no. 5, pp. 907–921, 2022, doi: 10.1108/JHTT-09-2021-0259.

[2] W. Xu, Z. Yao, Y. Ma, and Z. Li, "Understanding customer complaints from negative online hotel reviews: A BERT-based deep learning approach," *Int. J. Hosp. Manag.*, vol. 126, p. 104057, 2025, doi: 10.1016/j.ijhm.2024.104057.

[3] G. Sasongko, D. D. Kameo, V. N. Siwi, Y. Wahyudi, and A. D. Huruta, "The Effect of Service Quality and Heritage Tourism on Tourist Loyalty: The Case of Borobudur Temple," *Heritage*, vol. 8, no. 2, 2025, doi: 10.3390/heritage8020077.

[4] L. Qin and H. Zhang, "Impact of rich cultural tourism experience on tourist satisfactions and behavioral intentions toward Ningxia's cultural heritage: Moderation role of perceived cultural distance," *PLoS One*, vol. 20, no. 11 November, pp. 1–21, 2025, doi: 10.1371/journal.pone.0336220.

[5] J. A. Fernández Gallardo and R. Hernandez Rojas, "Impact of touristic sustainability on satisfaction with touristic services in a world heritage city. The case of the equestrian show in Córdoba (Spain)," *J. Cult. Herit. Manag. Sustain. Dev.*, vol. 15, no. 4, pp. 796–812, 2025, doi: 10.1108/JCHMSD-12-2023-0226.

[6] C. Zhang, Y.-X. Tian, and A.-Y. Hu, "Utilizing textual data from online reviews for daily tourism demand forecasting: A deep learning approach leveraging word embedding techniques," *Expert Syst. Appl.*, vol. 260, p. 125439, 2025, doi: 10.1016/j.eswa.2024.125439.

[7] R. N. Patil, Y. P. Singh, S. A. Rawandale, and S. Singh, "Improving Sentiment Classification on Restaurant Reviews Using Deep Learning Models," *Procedia Comput. Sci.*, vol. 235, pp. 3246–3256, 2024, doi: 10.1016/j.procs.2024.04.307.

[8] T. Anilsagar and S. A. S. S, "War strategy assisted Bi-LSTM for sentiment analysis of customer review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 40, no. 1, pp. 480–489, 2025, doi: 10.11591/ijeecs.v40.i1.pp480-489.

[9] K. Gao, J. Zhou, Y. Chi, and Y. Wen, "TourismNER: A Tourism Named Entity Recognition method based on entity boundary joint prediction," *Intell. Syst. with Appl.*, vol. 25, no. December 2024, p. 200475, 2025, doi: 10.1016/j.iswa.2025.200475.

[10] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl. Soft Comput.*, vol. 151, p. 111112, 2024, doi: 10.1016/j.asoc.2023.111112.

[11] N. M. Sabri, S. N. A. M. Subki, U. F. M. Bahrin, and M. Puteh, "Post Pandemic Tourism: Sentiment Analysis using Support Vector Machine Based on TikTok Data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 323–330, 2024, doi: 10.14569/IJACSA.2024.0150234.

[12] S. Yang, Q. Li, D. Jang, and J. Kim, "Deep learning mechanism and big data in hospitality and tourism: Developing personalized restaurant recommendation model to customer decision-making," *Int. J. Hosp. Manag.*, vol. 121, p. 103803, 2024, doi: 10.1016/j.ijhm.2024.103803.

[13] J. Chen, T. Zhang, Z. Yan, Z. Zheng, W. Zhang, and J. Zhang, "Attention-based BiLSTM with positional embeddings for fake review detection," *J. Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01130-9.

[14] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews," *Procedia Comput. Sci.*, vol. 245, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[15] D. Zhu, R. Jing, Q. Guo, D. Zhang, and F. Wan, "Sentiment analysis of tourism review text combined with bert-bilstm and attention mechanism," J. Com. Methods SE, vol. 2024. doi: 10.3233/JCM-247135.

[16] A. K. A. Ahamed, K. Lalitha, S. Saravanan, and S. Muthukumar, "Enhanced Deep Learning Based Non-Invasive Anomaly Detection of ECG Signals with Emphasis on Diabetes," *Int. J. Intell. Syst. Appl. Eng. IJISAE*, vol. 2023, no. 6s, pp. 284–294, 2023, [Online]. Available: www.ijisae.org

[17] Aachal Jakhotiya, Harshada Jain, Bhavik Jain, and Ms. Charmi Chaniyara, "Text Pre-Processing Techniques in Natural Language Processing: A Review," *Int. Res. J. Eng. Technol.*, vol. 9, no. 2, pp. 878–880, 2022.

[18] J. Sawicki, M. Ganzha, and M. Paprzycki, *The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques*, vol. 5, no. 3. 2023. doi: 10.1162/dint_a_00213.

[19] S. C. Necula, F. Dumitriu, and V. Greavu-Șerban, "A Systematic Literature Review on Using Natural Language Processing in Software Requirements Engineering," *Electron.*, vol. 13, no. 11, 2024, doi: 10.3390/electronics13112055.

[20] A. T. Riadi, F. Indriani, M. I. Mazdadi, M. R. Faisal, and R. Herteno, "Cross-Temporal Generalization of IndoBERT for Indonesian Hoax News Classification," *J. Tek. Inform.*, vol. 6, no. 5, pp. 5291–5304, 2025, doi: 10.52436/1.jutif.2025.6.5.4757.

[21] D. D. Purwanto, "Empirical Evaluation of IndoBERT and LSTM for Sentiment Analysis of Tourism Reviews : A Data-Driven Study on Kenjeran Park," *J. Tek. Inform.*, vol. 7, no. 1, pp. 463–474, 2026, doi: 10.52436/1.jutif.2026.7.1.4901.

[22] H. Setyawan, "Identification Of Plants And Its Use In Ancient Java : A Case Study Of The Ramayana And Kresnayana Reliefs Of Prambanan Temple," *Naditira Wirya*, vol. 16, no. 1, pp. 1–22, 2022, doi: 10.24832/nw.v16i1.498.

[23] A. Haque *et al.*, "Implication of Different Data Split Ratio on the Performance of Model in Price Prediction of Used Vehicles Using Regression Analysis," *Data Metadata*, vol. 3, 2024, doi: 10.56294/dm2024425.

[24] M. Hayaeian Shirvan, M. H. Moattar, and M. Hosseinzadeh, "Deep generative approaches for oversampling in imbalanced data classification problems: A comprehensive review and comparative analysis," *Appl. Soft Comput.*, vol. 170, p. 112677, 2025, doi: 10.1016/j.asoc.2024.112677.