International Journal of Advances in Data and Information Systems
Vol. 7, No. 1, April 2026, pp. 217~229
ISSN: 2721-3056, DOI: 10.59395/ijadis.v7i1.1483

Effectiveness Fine-Tuned Multilingual BERT Model for Sentiments Classification Toward Bali's Cultural Attractions

Nengah Widya Utami[1], Amna Binti Saad[2], Made Adi Paramartha Putra[3], I Gede Juliana Eka Putra[4]
[1,3,4] Faculty of Information Technology and Design, Primakara University, Indonesia
[2] Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Malaysia

Article Info

Article history:
Received Dec 06, 2025
Revised Jan 31, 2026
Accepted Feb 19, 2026

Keywords:
Multilingual BERT
Sentiments Classification
Machine Learning
Cultural Tourism
Tourist Review

ABSTRACT

This study examines the performance of a fine-tuned Multilingual BERT (mBERT) model for sentiment analysis of tourist reviews on Balinese cultural attractions. A multilingual dataset comprising 7,878 user-generated reviews from Google Maps and TripAdvisor was utilized to capture diverse linguistic expressions and visitor perspectives. The research methodology includes: (1) problem formulation and literature review; (2) dataset collection, preprocessing, and tokenization; (3) model training using mBERT as the baseline; (4) fine-tuning for domain adaptation; and (5) comparative evaluation with other Transformer models (XLM-Roberta and Distil-mBERT) and classical algorithms including Logistic Regression, Support Vector Machine, and Naïve Bayes. The results demonstrate a substantial improvement after fine-tuning. The baseline mBERT achieved 85.45% accuracy, while the fine-tuned model reached 92.13% accuracy with an AUC of 0.909, confirming the effectiveness of domain-specific adaptation. Although XLM-Roberta obtained slightly higher performance (93.15% accuracy, AUC 0.946), the fine-tuned mBERT showed stable and competitive results, making it the primary model of this study. Comparisons with classical methods further indicate that Transformer-based approaches provide more balanced and reliable sentiment classification. Sentiment distribution analysis reveals that tourist perceptions are predominantly positive, particularly regarding cultural authenticity and the quality of performances such as the Kecak and Fire Dance. Negative sentiments mainly relate to operational aspects, including crowd management, seating arrangements, and ticketing processes. Overall, this study provides empirical evidence that fine-tuned mBERT can effectively support data-driven evaluation of tourist experiences and deliver actionable insights for improving service quality and sustainability of Bali's cultural tourism.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Nengah Widya Utami,
Faculty of Information Technology and Design,
Primakara University,
Jalan Tukad Badung No. 135, Renon, Denpasar Selatan, Kota Denpasar, Bali 80226 Indonesia.
Email: widya@primakara.ac.id

1. INTRODUCTION

Bali is one of the most popular tourism destinations in Indonesia, known for its beautiful nature and strong cultural traditions. Tourism in Bali is not only about beaches or scenery, but also about cultural attractions such as traditional dances, ceremonies, arts, and local rituals. These attractions are based on Balinese culture, Hindu values, and the Tri Hita Karana philosophy, which teaches harmony between people, nature, and God. The government supports cultural tourism so that culture and tourism can grow together in a balanced and sustainable way. However, based on

the current study shows that cultural tourism has both good and bad effects. On the positive side, it gives income to local people, creates jobs, and supports community development. On the negative side, some cultural activities become commercialized or changed to fit tourist demands, which can reduce their authenticity. This situation shows that Bali needs to manage tourism carefully to protect its culture while still benefiting from tourism [1].

Visitors to tourist destinations today commonly use digital platforms to share their experiences, opinions, and personal impressions. Through online review sites, social media posts, travel blogs, and other forms of user-generated content, travelers freely express what they like, dislike, or find memorable during their visit. These platforms allow tourists to instantly document their experiences, often through photos, short videos, or brief comments that can reach a wide audience. Many visitors also rely on these digital spaces to recommend destinations to others or warn future travelers about potential issues, making their feedback even more influential. As a result, digital platforms have become a key channel for understanding tourist behavior and capturing authentic responses related to tourism destination [2], [3]. Bali cultural attractions such as temples, traditional villages, and heritage sites are frequently discussed by international and domestic tourists, generating extensive unstructured text data on platforms like TripAdvisor and Google Reviews. These insights can support stakeholders in developing sustainable tourism strategies, preserving cultural authenticity, and enhancing visitor satisfaction [4].

In recent years, sentiment analysis has emerged as a critical tool for analyzing large and complex textual datasets. However, these conventional approaches often rely on bag-of-words or static word embeddings, which limit their ability to capture contextual meaning, sarcasm, idiomatic expressions, and domain-specific terminology. Classical machine-learning approaches such as Naïve Bayes, KNN, SVM, and lexicon-based model face significant limitations, particularly when dealing with complex textual data and sentiment analysis, driving the shift toward advanced models like BERT [5], [6]. A primary weakness of these older methods lies in their reliance on traditional embeddings that assign a static vector representation to each word regardless of its context, which fails to capture the reality that word sense is context-dependent [6]. Furthermore, simple machine-learning models are frequently found to be ineffective for certain tasks, exhibiting suboptimal performance or signs of overfitting, even when efforts are made to increase data samples or training epochs [7].

BERT models offer substantial advantages in Natural Language Processing (NLP), primarily rooted in their architecture, which utilizes a bidirectional nature to read input simultaneously from both directions, enabling the creation of highly contextualized word embeddings that capture flexible semantics based on surrounding text. Furthermore, the Transformer architecture facilitates exceptional parallelizability, making these models proficient at processing extensive data swiftly [6], [8], [9], [10], [11]. Research employs various types of BERT and BERT-like models, including standard versions like Bert-base cased and Bert-base uncased [7]; specialized variants such as Multilingual BERT (mBERT), trained on over 100 languages for cross-lingual tasks [5], [6], [9], [12], [13], and DistilBERT for efficiency in resource-constrained environments [7], and optimized models like RoBERTa [7], [14], which often yield superior prediction metrics by utilizing robust pre-training methods along with domain-specific models like XLM-RoBERTa and Bangla BERT for resource-constrained languages [9].

Although many studies discuss the social, cultural, and economic impacts of tourism in Bali, very few have examined tourist sentiments using advanced NLP techniques. Existing research on sentiment analysis in Indonesian tourism mostly relies on traditional machine learning or lexicon-based approaches, which often fail to capture informal language, contextual meaning, and code-switching commonly found in online reviews. The multilingual nature of tourist comments mixing English, Indonesian, and Balinese terms creates additional challenges that are not well addressed in previous studies. There is also a lack of research applying transformer-based models, particularly mBERT, to analyze sentiments toward Bali's cultural attractions. Current models are not optimized for domain-specific vocabulary and cultural nuances, resulting in lower accuracy and limited insight for tourism stakeholders.

To fill this gap, this study uses a fine-tuned mBERT model to handle multilingual text more effectively and provide a more accurate, context-aware sentiment classification for cultural tourism attraction in Bali. Given these limitations, this study offers the following contributions:

1) We develop an efficient fine-tuned Multilingual BERT (mBERT) model specifically designed for multilingual tourist sentiment classification.
2) We analyze the sentiment distribution across various cultural attractions to identify patterns and differences in visitor perceptions.
3) We provide actionable insights and practical recommendations for tourism stakeholders based on the sentiment analysis results.

2. RESEARCH METHOD

The methodological workflow was structured into three main stages: (1) problem formulation and literature review, (2) dataset collection and preprocessing, and (3) model implementation and evaluation. These stages were carried out systematically to ensure that the model could effectively process multilingual input, learn domain-specific linguistic patterns, and generate reliable sentiment insights. The overall methodological framework adopted in this research is illustrated in Figure 1.

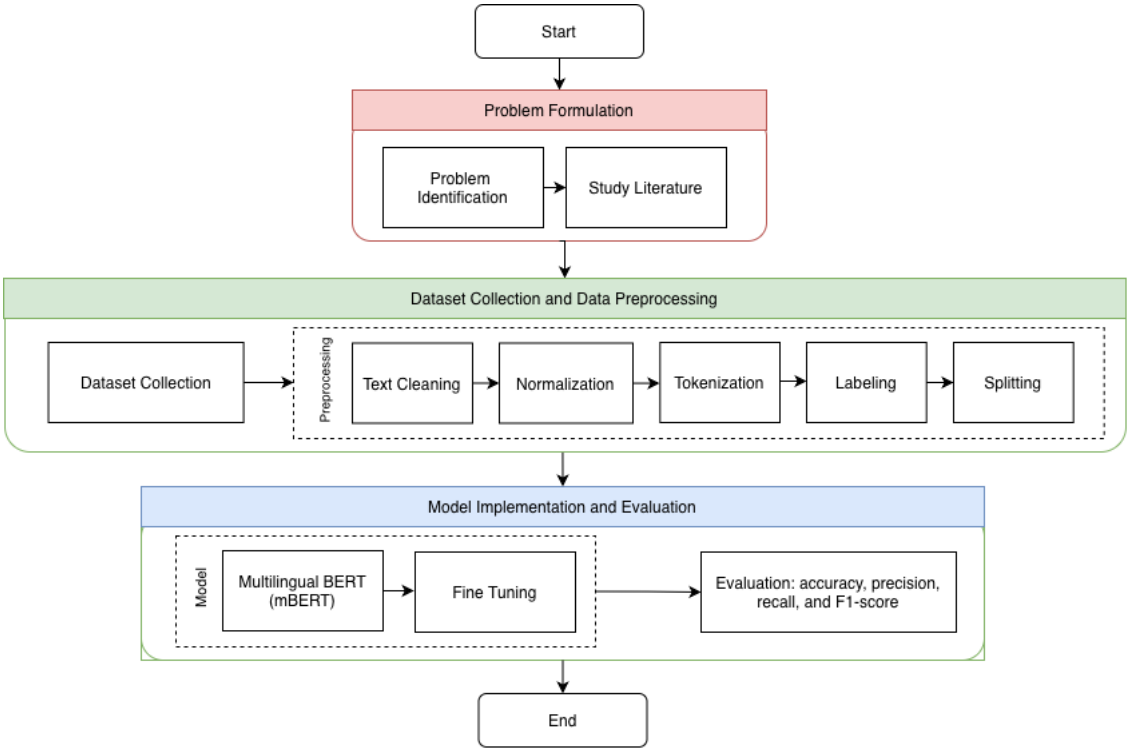


Figure 1. Overview of the methodology used in this research.

The research methodology consists of three main stages. First, the problem is formulated by identifying key challenges associated with multilingual tourist reviews such as informal expressions, mixed-language usage, and culturally specific terminology, supported through a comprehensive literature review to establish the research gap and justify the adoption of a transformer-based multilingual model. This review ensures a strong theoretical foundation for understanding previous approaches and highlights the need for a more context-aware sentiment analysis method. Second, dataset collection and preprocessing are conducted by gathering multilingual reviews from digital platforms, performing text cleaning and normalization, handling code-switching, tokenizing the text using the mBERT WordPiece tokenizer, assigning sentiment labels, and splitting the dataset into training, validation, and test sets to ensure reliable model development. Finally, the model implementation and evaluation stage involves fine-tuning the

Multilingual BERT model on the prepared dataset and assessing its performance using accuracy, precision, recall, F1-score, and ROC-AUC. In addition, a comparative evaluation is conducted by benchmarking the fine-tuned mBERT against other multilingual Transformer models, namely XLM-Roberta and Distil-mBERT, as well as classical machine learning algorithms including Logistic Regression, Support Vector Machine, and Naïve Bayes, in order to provide a comprehensive assessment of model effectiveness. The results of this workflow enable a deeper understanding of tourist perceptions and support the development of data-driven insights for tourism stakeholders.

2.1. Problem Formulation

This stage begins with identifying the core challenges present in multilingual tourist reviews, such as informal writing, mixed-language expressions, and culturally specific terminology. These characteristics make traditional sentiment analysis models insufficient. A comprehensive literature review is then conducted to analyze prior studies on cultural tourism, sentiment classification techniques, and transformer-based multilingual models. The review helps refine the research gap, justify the selection of mBERT, and define the methodological direction of the study.

2.2. Dataset Collection and Preprocessing

Tourist reviews related to Bali's cultural attractions are collected from various digital platforms containing user-generated content. The dataset collection process is carried out using a web-crawling technique to extract multilingual visitor reviews from cultural attraction pages on TripAdvisor and Google Maps. The data used in this study were collected through a web crawling process using Instant Data Scraper, a browser-based automated data extraction tool. Instant Data Scraper enables structured data collection from dynamic web pages without requiring manual scripting, making it suitable for large-scale review extraction from online platforms. Figure 2 is illustrated the dataset collection method in this study.

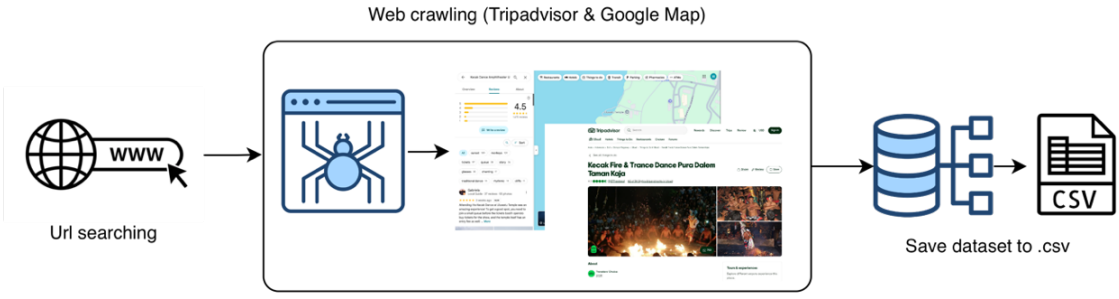


Figure 2. Dataset collection method using web crawling approach.

The crawling process was conducted by identifying review elements on the target web pages, including the review text and associated metadata. The tool automatically detected repeating data patterns and extracted the relevant content into a structured tabular format. The collected data were then exported in comma-separated values (CSV) format for further preprocessing and analysis. This approach allows for efficient and reproducible data acquisition while minimizing human intervention during the crawling process. However, since Instant Data Scraper relies on the structure of the target web pages, variations in page layout may affect the completeness of the extracted data. Despite this limitation, the tool proved effective for collecting multilingual user-generated reviews at scale.Table 1 summarizes the key attributes of the dataset, providing an overview of the information collected from tourist reviews of Balinese cultural attaraction.

Table 1. Dataset information.

No	Column	Description	Values
1	source	The platform from which the review was collected	Tripadvisor, Google Maps
2	lang	The language of the review	English, Indonesian, French, Japanese, German, Others
3	object	The attraction or activity being reviewed	Kecak Dance, Fire Dance, Barong Dance, Legong Dance,

No	Column	Description	Values
			Keris Dance, Sangyang Dance
4	location	The specific location of the attraction	Pecatu, Ubud, Sukawati, Batubulan, Kesiman
5	review	The textual content of the review, covering both positive and negative visitor experiences	Appreciation of performance, sunset views, traditional costumes, complaints about crowds, seating, safety
6	year	The year the review was written	2011–2025

The dataset used in this study was compiled by crawling tourist reviews from Tripadvisor and Google Maps for various Balinese cultural performances, including the Kecak Dance, Fire Dance, Barong Dance, Legong Dance, Keris Dance, and Sangyang Dance. Reviews were collected from multiple locations across Bali, such as Pecatu, Ubud, Sukawati, Batubulan, and Kesiman, spanning the years 2011 to 2025. During the collection process, care was taken to extract detailed and relevant information from each review, including the review text, language, source platform, attraction, location, and year. This resulted in a dataset that provides a comprehensive representation of visitor experiences and perceptions across different performances and locations. Table 2 presents a sample of the dataset used in this study, illustrating the structure and types of information collected from tourist reviews.

Table 2. Sample of datasets used in this study.

index	source	lang	object	location	review	year
0	Tripadvisor	English	Kecak and Fire Dance	Pecatu	I've been to Bali so many times, but this was ...	2025
1	Tripadvisor	English	Kecak and Fire Dance	Pecatu	The Kecak dance show was a horrible experience...	2025
2	Tripadvisor	English	Kecak and Fire Dance	Pecatu	Lovely way to see Bali and Hindu culture ...wi...	2025
3	Tripadvisor	English	Kecak and Fire Dance	Pecatu	If you like having a terrible time book ticket...	2025
4	Tripadvisor	English	Kecak and Fire Dance	Pecatu	We attended the show after many reviews online...	2025

The dataset has been made publicly available via Mendeley Data with DOI: 10.17632/5zzdy3f48k.1 [15]. Most reviews were collected via Google Maps, with notable contributions from Tripadvisor for certain performances like the Kecak and Fire Dance in Pecatu and Ubud. The largest number of reviews comes from Kecak Fire & Trance Dance Pura Dalem Taman Kaja (1,600 reviews), Kecak and Fire Dance in Pecatu (1,473 reviews), and Kecak Uluwatu (1,310 reviews). Other performances, such as Kecak Fire & Trance Dance Pura Batu Karu and Uma Dewi Kecak & Sanghyang Dance, have relatively few reviews, indicating lower visitor coverage in the dataset. The language distribution can be observed that the dataset is dominated by English, with a total of 901 reviews, making it the most frequently used language. Indonesian appears as the second most common language with 48 reviews, reflecting the local context of the data source. In addition, several other languages are present in much smaller proportions, including French (37), Swahili (18), Romanian (15), Italian (14), and Afrikaans (12). The remaining languages such as German, Spanish, Dutch, Catalan, Russian, Portuguese, Danish, Estonian, Norwegian, Slovenian, Somali, and Turkish contribute only a few instances.

The dataset exhibits a clear dominance of English-language reviews, which account for 62.23% (4,905 samples) of the total data. This is followed by Indonesian, representing 19.52% (1,538 samples), making it the second most frequent language in the corpus. Together, English and Indonesian comprise more than 81% of the dataset, indicating a strong concentration in these two languages. Several other languages appear with moderate to low frequencies, including Malay (4.51%), French (2.72%), and German (1.43%). Additional languages such as Japanese, Italian, Dutch, Spanish, and Russian contribute smaller proportions, each accounting for less than 2% of the total dataset. These languages can be categorized as low-resource within the context of this study. Table 3 presents the distribution of top 10 languages in the collected dataset.

Table 3. Distribution of languages int the collected dataset

index	language	language_name	count
0	en	English	4905
1	id	Indonesian	1538
2	ms	Malay	355

index	language	language_name	count
3	fr	French	214
4	de	German	113
5	ja	Japanese	106
6	it	Italian	91
7	nl	Dutch	85
8	es	Spanish	75
9	ru	Russian	71

The textual data from the reviews was preprocessed to ensure consistency and readiness for analysis. The review column was automatically detected, and any missing values were replaced with empty strings. The preprocessing steps included removing URLs, converting all text to lowercase, and normalizing whitespace by collapsing extra spaces, tabs, and line breaks. Punctuation was retained to preserve the meaning of the text, although it can be removed if necessary for specific analyses. After cleaning, empty strings were converted back to missing values (NaN) to verify that no entries were left without text. The cleaned text was stored in a new column, text_clean, which contained all reviews ready for downstream tasks such as sentiment analysis or keyword extraction. This process ensured that the dataset contained no missing review values after preprocessing, providing a reliable foundation for further analysis.

2.3. Model Implementation and Evaluation

The architecture used in this study follows the standard BERT pipeline as illustrated in the Figure 3, consisting of a pre-training stage, a fine-tuning stage, and a prediction stage. In the pre-training phase, BERT leverages a large corpus through a Transformer-based encoder architecture that learns bidirectional contextual representations. The model is initially exposed to diverse textual data using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), enabling it to acquire a deep understanding of language semantics and sentence relationships [16]. Following pre-training, domain-specific review data collected is used in the fine-tuning stage, where the model parameters are further adjusted for the sentiment classification task.

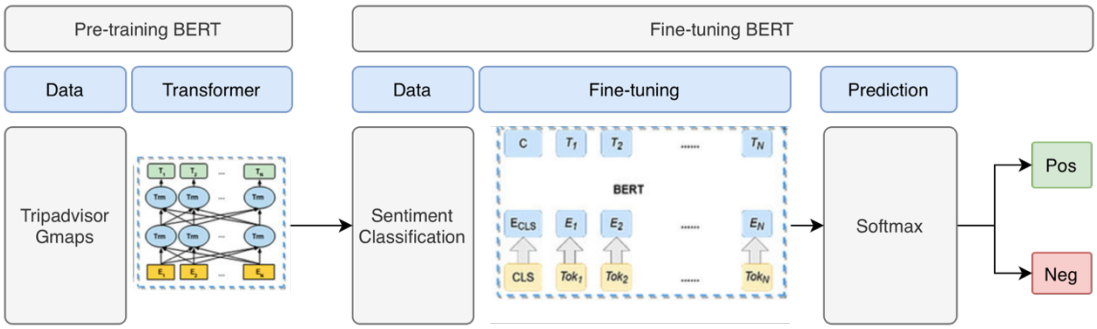


Figure 3. The architecture of BERT model.

During fine-tuning, each review is tokenized and fed into the BERT encoder, where the contextualized embeddings of all tokens are generated. The representation associated with the token, which captures the overall meaning of the input sequence, is forwarded to a classification layer. This layer, followed by a softmax function, assigns a sentiment label positive or negative based on the learned patterns within the tourism review corpus. As shown in the workflow, the model directly performs automatic sentiment labeling by mapping encoded input features to probability distributions over the sentiment classes. The integration of pre-training, domain-specific fine-tuning, and softmax-based prediction ensures that BERT can accurately classify user sentiments toward tourism destinations, leveraging both its general linguistic knowledge and its adaptation to the TripAdvisor and Google Maps datasets.

In this study, BERT is not only used as the underlying architecture but also serves as the primary model for automatic sentiment labeling. The fine-tuned BERT model processes each review through the representation and applies a classification layer followed by softmax to assign sentiment labels positive or negative based on learned contextual patterns. This integration ensures

that sentiment labeling benefits directly from BERT's bidirectional contextual understanding, resulting in more accurate and reliable classification outcomes. To illustrate the output generated by the BERT model, the results of the automatic sentiment labeling process are show in Table 4.

Table 4. Automatic sentiment labeling using BERT.

index	review	auto_label	confidence	label
0	I've been to Bali so many times, but this was ...	positive	0.932569	1
1	The Kecak dance show was a horrible experience...	negative	0.995116	0
2	Lovely way to see Bali and Hindu culture ...wi...	positive	0.941939	1
3	If you like having a terrible time book ticket...	negative	0.928469	0
4	We attended the show after many reviews online...	negative	0.995320	0

Figure 4 presents the overall distribution of sentiment labels generated through the automatic labeling process. The results indicate a strong dominance of positive sentiment, with 6,896 reviews, accounting for approximately 87.5% of the dataset. In contrast, 982 reviews, or about 12.5%, are classified as negative sentiment. This imbalanced distribution reflects a common characteristic of user-generated review data, where users are more likely to share feedback following positive experiences. The predominance of positive sentiment suggests generally favorable user perceptions toward the reviewed subject or service.The sentiment distribution show in Figure 4.

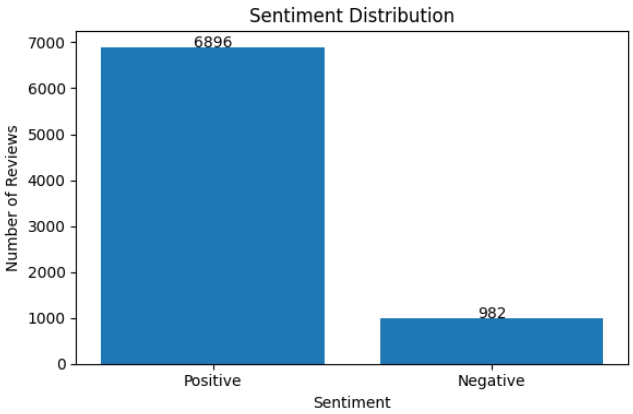
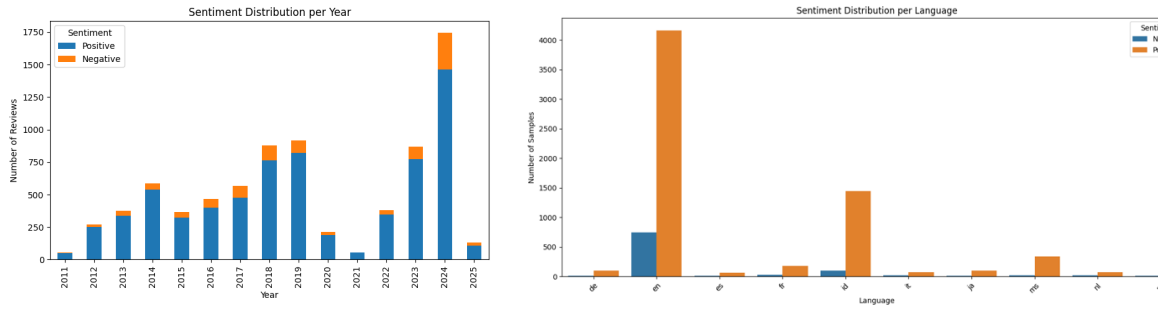


Figure 4. Distribution of sentiment label.

Figure 5 (a) illustrates the yearly distribution of positive and negative sentiments. Overall, positive sentiment consistently dominates across all years, indicating generally favorable user perceptions. A notable increase in review volume and positive sentiment is observed prior to 2020, reaching a peak around 2018–2019. During the COVID-19 pandemic period (2020–2022), a clear decline in the total number of reviews is evident. This reduction likely reflects decreased user activity and limited access to services during travel restrictions and social distancing measures. Although positive sentiment remains dominant in this period, a slight relative increase in negative sentiment can be observed, potentially indicating user dissatisfaction related to service disruptions, cancellations, or pandemic-related constraints. Figure 5 (b) presents the distribution of positive and negative sentiments across different languages. Positive sentiment dominates in all languages, with English contributing the largest number of positive reviews, followed by Indonesian. Negative sentiment appears in much smaller proportions across all languages, indicating generally favorable user perceptions regardless of language. These results suggest that the sentiment patterns are consistent across languages, although the volume of data varies significantly among them.



(a) (b)

Figure 5. Sentiment distribution
(a) Per year, (b) Per Language

The evaluation process was conducted to measure the performance of the sentiment classification model and to determine the effectiveness of the fine-tuning approach. Several standard evaluation metrics were employed, including accuracy, precision, and recall, each providing complementary insights into model behavior [17], [18]. Accuracy was used to assess the overall proportion of correctly predicted sentiments, while precision measured the model's ability to correctly identify positive or negative reviews without introducing excessive false positives. Recall, on the other hand, quantified the model's capability to retrieve all relevant sentiment instances, particularly important when dealing with imbalanced review distributions. The evaluation began with model accuracy, calculated using Equation 1, followed by precision and recall, estimated using Equations 2 and 3, respectively.

$$Accuracy\ (\%) = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}\ x\ 100 \qquad (1)$$

$$Precision\ (\%) = \frac{T_p}{T_p + F_p}\ x\ 100 \qquad (2)$$

$$Recall\ (\%) = \frac{T_p}{T_p + F_n}\ x\ 100 \qquad (3)$$

where T_p denotes true positive, T_n represent ture negative, followed by F_p and F_n which refers to false positive and false negative, respectively. In addition to accuracy, precision, and recall, we also include the F1-score, which is calculated based on the precision and recall values, as shown in Equation 4.

$$F1 - score\ (\%) = 2\ x \frac{Precision\ x\ Recall}{Precision + Recall} \qquad (4)$$

To complement the main evaluation metrics, the model's performance was further assessed using the confusion matrix and the ROC curve. The confusion matrix provides a clear summary of correct and incorrect predictions, allowing identification of whether the model tends to misclassify positive reviews as negative or vice versa. In addition, the ROC curve illustrates the balance between the true positive rate and the false positive rate across varying thresholds, with the AUC score indicating the model's overall discriminatory strength [18], [19]. Together, these analyses offer a concise yet informative view of the fine-tuned BERT model's ability to distinguish sentiment classes more effectively than the unfine-tuned baseline.

In this stage, this study not only applies fine-tuning mBERT model but also conducts a model comparison with classical machine learning models Logistic Regression (LR), Support Vector Machine (SVM), and Naïve Bayes. This comparison is intended to evaluate the performance of mBERT against traditional classification approaches using the same evaluation metrics, so that a clearer and more objective conclusion can be drawn regarding the effectiveness of the proposed deep learning model.

3. RESULTS AND DISCUSSION

This section presents the experimental findings and analytical insights derived from the proposed sentiment classification model. The results are examined from multiple perspectives, beginning with the performance evaluation of the fine-tuned multilingual BERT model compared to several baseline classifiers. Subsequently, the distribution of sentiment categories within the dataset is analyzed to identify key themes in tourist perceptions.

3.1. Model Performance

Before model training, the labeled dataset was split into three subsets: 70% for training, 15% for validation, and 15% for testing. After splitting, the text data was tokenized to convert raw text into numerical input suitable for the Transformer-based models. This process ensures that the models can effectively process and learn from the text data. This study evaluates the performance of sentiment classification models in a stepwise manner. Initially, Multilingual BERT (mBERT) is employed as the primary baseline model. Subsequently, the mBERT model is fine-tuned to enhance its performance on the target dataset. The fine-tuned mBERT is then compared with other multilingual Transformer-based models. To provide a broader perspective on model performance, comparisons were also made with other multilingual transformer models, namely XLM-Roberta [18] and DistilBERT (Distil-mBERT) [20], [21], [22]. Finally, the Transformer models are contrasted with classical machine learning approaches, including Logistic Regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB).

The baseline mBERT model demonstrates strong performance in multilingual sentiment classification, confirming its ability to capture contextual and semantic information across multiple languages. Even without extensive task-specific optimization, mBERT achieves reliable accuracy and balanced precision recall performance, indicating its effectiveness as a general-purpose multilingual sentiment model. The evaluation results show in Table 5.

Table 5. mBERT evaluation result

Metric	eval_loss	eval_model_preparation_time	eval_accuracy	eval_f1	eval_precision	eval_recall
Value	0.627908	0.005	0.854484	0.807523	0.768364	0.854484

The mBERT model demonstrates improved and strong performance on the sentiment classification task. A loss of 0.6279 indicates that the model has learned effectively with a relatively low and stable error rate. It achieves an accuracy of 0.854, meaning that approximately 85.4% of the samples were correctly classified. Looking at the class-specific metrics, the precision of 0.768 shows that the majority of positive predictions made by the model are correct, with a manageable level of false positives. The recall of 0.854 suggests that the model is highly capable of identifying positive instances, capturing most of the relevant samples with only a small portion being missed. The resulting F1-score of 0.808 reflects a solid balance between precision and recall, indicating consistent performance across evaluation aspects. These results confirm that mBERT is highly effective for text classification and provides reliable and stable performance, making it well-suited as the main model for this study.

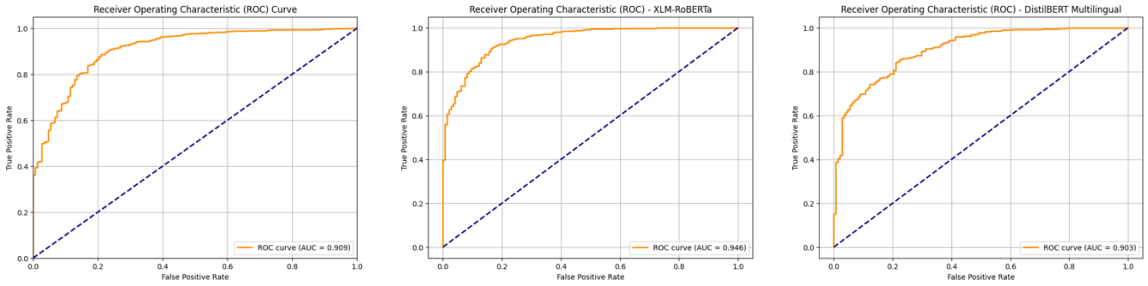
On this study, mBERT model was fine-tuned on the dataset according to the specifications outlined in the training configuration above. The model was trained for 5 epochs, using a batch size of 16 for both training and evaluation. Weight decay of 0.01 was applied to help regularize the model and prevent overfitting. During training, evaluation and checkpointing were performed at the end of each epoch, with the best model selected based on the F1-score to ensure optimal performance. A custom data collator and metric computation function were used to handle tokenization and performance evaluation. After fine-tuning, the model was evaluated on the test dataset, producing performance metrics that reflect the improvements from task-specific training. To provide a broader perspective on model performance, comparisons were also made with other multilingual transformer models XLM-R and DistilBERT. Table 6 are the results of the comparison between mBERT Baseline, mBERT Fine-Tuned, XLM-R, and Distil-mBERT.

Table 6. Result of comparison BERT model.

index	Model	eval_loss	eval_accuracy	eval_f1	eval_precision	eval_recall
0	mBERT Baseline	0.627908	0.854484	0.807523	0.768364	0.854484
1	mBERT Fine-Tuned	0.267835	0.921320	0.912993	0.915096	0.921320
2	XLM-R	0.323495	0.931472	0.929664	0.928629	0.931472
3	Distil-mBERT	0.248320	0.917090	0.910006	0.909677	0.917090

Based on Table 6, the comparison results show that all transformer-based models achieved strong performance, with the fine-tuned mBERT emerging as the primary focus of this study. The mBERT Baseline reached an accuracy of 85.45%, while fine-tuning significantly improved its performance to 92.13%, demonstrating the effectiveness of task-specific adaptation in enhancing multilingual sentiment classification. This substantial improvement confirms that the tuned mBERT is a reliable and well-optimized model for the dataset used in this research. As a comparison, XLM-R achieved the highest overall accuracy of 93.15% and an F1-score of 92.97%, while Distil-mBERT also performed competitively with 91.71% accuracy. Nevertheless, the fine-tuned mBERT remains the most relevant model for this study due to its strong and stable performance, efficient architecture, and consistency with the research objectives, making it an appropriate main model despite the slightly higher scores obtained by XLM-R.

The comparison of ROC curves shows that the fine-tuned mBERT model demonstrates very strong discriminative ability with an AUC of 0.909, confirming that the fine-tuning process successfully improved the model's capability to distinguish between positive and negative sentiment classes. This result highlights that tuned mBERT serves as a reliable main model for multilingual sentiment classification in this study. As a comparison, XLM-R achieved a slightly higher AUC of 0.946, while Distil-mBERT obtained an AUC of 0.903. Although XLM-R records the best overall score, the fine-tuned mBERT remains highly relevant due to its nearly comparable performance, stability, and efficiency, making it a well-justified primary model for this research. Figure 6 show the comparison of ROC this models.



(a) (b) (c)

Figure 6. ROC results
(a) ROC mBERT, (b) ROC XLM-Roberta, (c) ROC Distil-mBERT

As an additional evaluation, this study compares the fine-tuned transformer models with classical machine learning approaches, including Logistic Regression, SVM, and Naive Bayes show on Table 7. Among the classical models, SVM (Linear) achieved the best performance with an accuracy of 90.02% and an F1-score of 94.52%, followed by Logistic Regression with 88.49% accuracy and Naive Bayes with 87.48%. Although these models show strong recall, their overall performance remains below the transformer-based approaches.

Table 7. Result of comparison LR, SVM, NB

index	Model	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.8849	0.8838	1.0	0.9383
1	SVM (Linear)	0.9002	0.9097	0.9836	0.9452
2	Naive Bayes	0.8748	0.8748	1.0	0.9332

Compared to the classical methods, the fine-tuned mBERT (92.13%), Distil-mBERT (91.71%), and XLM-R (93.15%) demonstrate superior and more balanced results. These findings confirm that transformer models, particularly the fine-tuned mBERT as the main model, provide

better semantic representation for multilingual sentiment classification, making them more suitable for this research than traditional algorithms.

3.2. Sentiment Wordcloud

The sentiment distribution indicates that visitor perceptions are predominantly positive. Additionally, this study also generated wordcloud visualizations, as shown in Figure 7.

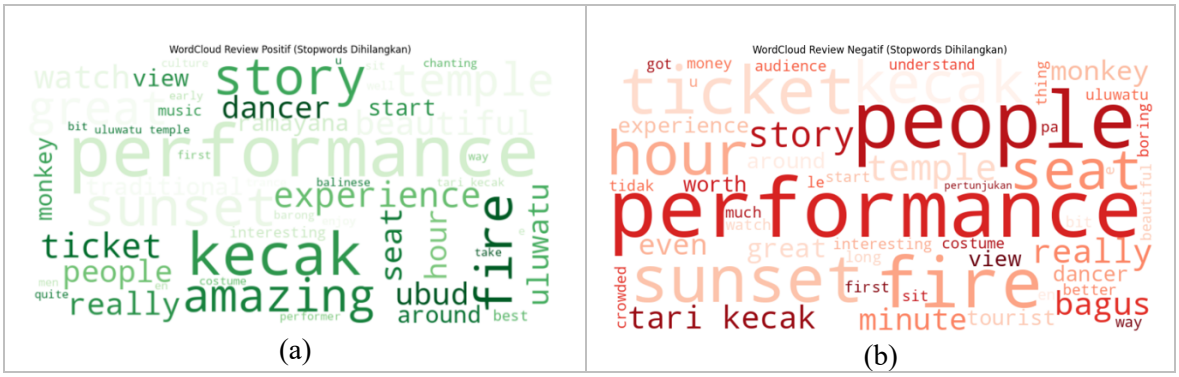


Figure 7. Wordcloud sentiment classification.
(a) Wordcloud positif, (b) Wordcloud negatif

The word cloud for positive reviews on Figure 7(a) highlights terms such as performance, story, kecak, amazing, and experience, indicating that visitors generally praise the show's artistic quality, atmosphere, and cultural value. In contrast, the negative review word cloud on Figure 7(b) shows frequently mentioned terms such as people, seat, ticket, and crowded, suggesting that dissatisfaction typically arises from issues related to crowd management, seating conditions, and ticketing rather than the performance itself. Together with the label distribution chart which shows a clear majority of positive sentiment these findings demonstrate that audience perception is largely favorable, with criticisms mostly centered on logistical challenges rather than the core cultural experience. The sentiment analysis reveals clear insights for tourism stakeholders. Visitors consistently express strong appreciation for the cultural and artistic elements of the Kecak and Fire Dance, emphasizing the quality of the performance, storyline, and traditional components. This affirms that the cultural core of the attraction is highly valued and should continue to be preserved and promoted. Meanwhile, negative sentiments mostly relate to operational issues such as seating, crowding, and ticketing rather than the performance itself. These findings indicate that logistical improvements are essential to enhancing the overall visitor experience.

Based on these insights, tourism stakeholders are advised to implement several practical improvements. First, crowd and seating management should be optimized by increasing seating capacity, improving queue systems, or offering reserved seating options to reduce discomfort. Second, ticketing processes may be enhanced through clearer information, streamlined online booking, and transparent pricing to minimize visitor frustration. Third, improving on-site facilities such as signage, access routes, and waiting areas can further elevate the experience. Lastly, emphasizing positive visitor feedback in promotional materials while addressing common operational complaints can help strengthen both satisfaction and reputation. By focusing on these actionable areas, stakeholders can enhance the overall sustainability and competitiveness of cultural tourism offerings.

4. CONCLUSION

This study confirms the effectiveness of the fine-tuned Multilingual BERT (mBERT) model for multilingual sentiment analysis of tourist reviews on Balinese cultural attractions. Using a dataset of 7,878 reviews collected from Google Maps and TripAdvisor, the research demonstrates that transformer-based models are capable of capturing linguistic nuances across diverse languages and expressing visitor perceptions with high reliability. The experimental design, which involved a 70:15:15 split for training, validation, and testing, followed by tokenization and systematic

evaluation, ensured robust and reproducible model assessment. The results highlight the importance of task-specific fine-tuning. The baseline mBERT achieved solid performance with 85.45% accuracy and an F1-score of 80.75%, indicating that pretrained multilingual representations already provide meaningful semantic understanding. However, after fine-tuning, mBERT improved substantially to 92.13% accuracy and an F1-score of 91.30%, confirming that domain adaptation is essential for recognizing sentiment patterns in tourism review data. ROC analysis further supports this finding, with the fine-tuned mBERT obtaining an AUC of 0.909, demonstrating strong discriminative capability between positive and negative sentiments. Comparative evaluation with other transformer models shows that XLM-R achieved the highest numerical performance (93.15% accuracy, AUC 0.946), while Distil-mBERT also performed competitively (91.71% accuracy, AUC 0.903). Nevertheless, the fine-tuned mBERT remains the primary and most relevant model for this study due to its stable performance, computational efficiency, and consistency with the research objectives. Additional comparison with classical machine learning methods reveals that although SVM, Logistic Regression, and Naïve Bayes achieved reasonably high scores, their performance was still below that of transformer-based models, confirming the superiority of contextual deep learning approaches for multilingual sentiment tasks.

The sentiment analysis results indicate that visitor perceptions toward Balinese cultural performances are predominantly positive, particularly regarding artistic quality, cultural authenticity, and immersive experience. Negative sentiments are mainly associated with operational aspects such as crowd management, seating, and ticketing systems. These findings provide practical implications for tourism stakeholders: improving service organization and visitor management while maintaining cultural integrity will enhance tourist satisfaction and support the sustainability of Bali's cultural tourism sector. This research demonstrates that the fine-tuned mBERT model offers an effective, reliable, and well-balanced solution for multilingual sentiment classification, bridging advanced NLP techniques with real-world tourism analytics and contributing valuable insights for data-driven destination management.

ACKNOWLEDGEMENTS

This research was supported by Directorate of Research and Community Services, Primakara University, Bali, Indonesia.

REFERENCES

[1] N. L. P. S. Widhiastuty and I. K. Murdana, "The Impacts of Bali cultural Tourism to the Local Balinese Community," *LACULTOUR: Journal of Language and Cultural Tourism*, vol. 3, no. 1, pp. 45–51, Jun. 2024, doi: 10.52352/lacultour.v3i1.1360.

[2] S. Yufei and D. Lei, "Research on aspect-level sentiment analysis of tourism review text based on joint learning," in *2023 IEEE International Conference on Control, Electronics and Computer Technology, ICCECT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1182–1186. doi: 10.1109/ICCECT57938.2023.10141071.

[3] F. Hastarita Rachman and B. Setya Rintyarna, "Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning," 2022.

[4] N. Widya Utami, N. Purnama, I. Made, and A. Prayoga, "Sentiment Analysis System Of Bali Tourism Using Naive Bayes Algorithm And Web Framework," *Informatika dan Sains*, vol. 14, p. 2024, 2024, doi: 10.54209/infosains.v14i03.

[5] Y. A. Singgalen, "Improved Sentiment Classification Using Multilingual BERT with Enhanced Performance Evaluation for Hotel Guest Review Analysis," *Journal of Computer System and Informatics (JoSYC)*, vol. 6, no. 2, pp. 508–520, 2025, doi: 10.47065/josyc.v6i2.6870.

[6] R. RICCIARDI and M. Manisera, "A Multilingual BERT-based classification of reviews for enhanced visitors' experience analysis," Jul. 31, 2024. doi: 10.21203/rs.3.rs-4495603/v1.

[7] N. Akhila, J. P. Sanjanasri, and K. P. Soman, "Comparative Study of Bert Models and Roberta in Transformer based Question Answering," in *2023 3rd International Conference on Intelligent Technologies, CONIT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/CONIT59222.2023.10205622.

[8] M. Rizky Yanuar and S. Shiramatsu, "Aspect Extraction for Tourist Spot Review in Indonesian Language using BERT."

[9] S. Islam, M. Jahidul Islam, M. Mahadi Hasan, S. M. Shahnewaz Mahmud Ayon, and S. Shabnam Hasan, "Bengali Social Media Post Sentiment Analysis using Deep Learning and BERT Model," in *2022 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ISIEA54517.2022.9873680.

[10] J. Xue and Y. Li, "Intelligent Tourism System Based on a BERT Model for the Cultural Relics Domain," in *2025 4th International Symposium on Computer Applications and Information Technology, ISCAIT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 195–199. doi: 10.1109/ISCAIT64916.2025.11010753.

[11] A. Chinnalagu, *Comparative Analysis of Fine-Tuned LLM, BERT and DL Models for Customer Sentiment Analysis*. IEEE, 2024.

[12] M. J. Shayegan and Y. Dastan, "A Multilingual Sentiment Analysis Model in Tourism," 2024. [Online]. Available: http://jac.ut.ac.ir

[13] R. Kumar. Dwivedi and A. Kr. Saxena, *Vietnamese Question Answering System from Multilingual BERT Models to Monolingual BERT Model*. Prof. Rakesh Kumar Dwivedi, Faculty of Engineering & Computing Sciences, Teerthanker Mahaveer University, 2020.

[14] N. Khurana and V. Prabhu, "Sentiment Analysis Using Transformer Models (BERT, ALBERT, RoBERTa, and DeBERTa) on a SMILE Twitter Dataset." [Online]. Available: http://t.co/GYKpsOb0W6

[15] N. W. Utami, "Dataset Review Bali's Cultural Attractions (Multilingual)," Mendeley Data.

[16] N. T. M. Trang and M. Shcherbakov, "Vietnamese question answering system from multilingual BERT models to monolingual BERT Model," in *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 201–205. doi: 10.1109/SMART50582.2020.9337155.

[17] T. Kasana and B. Kaur, "Evaluating Student Feedback: A Comparative Study of Traditional ML and DistilBERT Model," in *2025 5th International Conference on Intelligent Technologies, CONIT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/CONIT65521.2025.11167038.

[18] S. Imtiaz and E. K. Hashi, "Analyzing the Performance of Sentiment Analysis Using BERT, DistilBERT, RoBERTa and GPT-2," in *2024 6th International Conference on Sustainable Technologies for Industry 5.0, STI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/STI64222.2024.10951049.

[19] S. Riyadi, A. M. Masyhur, and A. D. Andriyani, "Deep Learning for Hate Speech Intensity Analysis: DistilBERT Classification Algorithm," in *2024 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ISIEA61920.2024.10607351.

[20] M. S. Arefin, M. Serajun Nabi, M. Touhami, Z. U. Rehman, M. Faizal Ahmad Fauzi, and H. Bin Abdul Karim, "Aspect-Based Sentiment Analysis for Product Reviews using DistilBERT," Institute of Electrical and Electronics Engineers (IEEE), Dec. 2025, pp. 1–6. doi: 10.1109/mecon67253.2025.11276893.

[21] S. R. Varghese, S. Juliet, and N. S. Athish, "Social Media Text Analysis for Disaster Management Using DistilBERT Model," in *Proceedings of 2024 International Conference on Science, Technology, Engineering and Management, ICSTEM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICSTEM61137.2024.10560620.

[22] J. M. Ruel Espiritu, R. J. E. Policarpio, A. C. B. Sabado, V. M. B. Sabado, and H. T. Sueno, "Classifying Fake News in Filipino Language using Improved DistilBERT Model," in *Proceedings - 11th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 320–325. doi: 10.1109/ICITACEE62763.2024.10762768.