Journal of
Information Systems Engineering
and Business Intelligence

Vol.11, No.1, February 2025
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content

Rifki Indra Perwira [1], Vynska Amalia Permadi [2]*, Dian Indri Purnamasari [3], Riza Prapascatama Agusdin [4]

[1)2)4)] *Informatics Department, Universitas Pembangunan Nasinonal Veteran Yogyakarta, Yogyakarta, Indonesia*
[1)] rifki@upnyk.ac.id, [2)] vynspermadi@upnyk.ac.id, [4)] rizapra@upnyk.ac.id

[3)] *Accounting Department, Universitas Pembangunan Nasinonal Veteran Yogyakarta, Yogyakarta, Indonesia*
[3)] dian_indri@upnyk.ac.id

Abstract

Background: Aspect-based sentiment analysis (ABSA) is essential in extracting meaningful insights from user-generated content (UGC) in various domains. In tourism, UGC such as Google Reviews offers essential feedback, but the challenges associated with processing in Indonesian language, including the unique linguistic characteristics, pose difficulties for automatic sentiment, and aspect detection. Recent advancements in transformer-based models, such as BERT, have shown great potential in addressing these challenges by providing context-aware embeddings.
Objective: This research aimed to fine-tune IndoBERT, a pre-trained Indonesian language model, to perform information extraction and key aspect detection from tourism-related UGC. The objective was to identify critical aspects of tourism reviews and classify their sentiments.
Methods: A dataset of 20,000 Google Reviews, focusing on 20 tourism destinations in DI Yogyakarta and Jawa Tengah, was collected and preprocessed. Multiple fine-tuning experiments were conducted, using a layer-freezing method by adjusting only the top layers of IndoBERT, while freezing others to determine the optimal configuration. The model's performance was evaluated based on validation loss, precision, recall, and F1-score in aspect detection and overall sentiment classification accuracy.
Results: The best-performing configuration involved freezing the last six layers and fine-tuning the top six layers of IndoBERT, yielding a validation loss of 0.324. The model achieved precision scores between 0.85 and 0.89 in aspect detection and an overall sentiment classification accuracy of 0.84. Error analysis revealed challenges in distinguishing neutral and negative sentiments and in handling reviews with multiple aspects or mixed sentiments.
Conclusion: The fine-tuned IndoBERT model effectively extracted key tourism aspects and classified sentiments from Indonesian UGC. While the model performed well in detecting strong sentiments, improvements are needed to handle neutral and mixed sentiments better. Future work will explore sentiment intensity analysis and aspect segmentation methods to enhance the model's performance.

Keywords: Aspect-Based Sentiment Analysis, Fine-tuning, IndoBERT, Sentiment Classification, Tourism Reviews, User-Generated Content

Article history: Received 31 October 2024, first decision 24 January 2025, accepted 14 February 2025, available online 28 March 2025

I. INTRODUCTION

User-generated content (UGC) is a key asset in the tourism sector, offering significant insights into the experiences, perspectives, and preferences of tourist. Previous research [1] has shown the potential of UGC in predicting tourism demand through methods such as social networks and semantic analysis, demonstrating its essential role in uncovering actionable feedback. Another report [2] further shows the capacity of UGC to capture customer needs by comparing large language models tailored for tourism. This underscores the significance of TripAdvisor in improving tourism services. However, the unstructured nature, varied linguistic expressions, and multilingual characteristics of UGC, particularly content in Indonesian, pose challenges for systematically analyzing sentiments and thematic aspects.

With the rapid growth, Indonesian tourism industry generates extensive UGC covering destinations, local activities, and accommodations. The lack of specialized research on analyzing Indonesian UGC to extract relevant insights has

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.11.1.30-40

left this potential underutilized, despite the abundance of data. As the volume of online reviews expands, developing efficient methods to analyze customer feedback becomes increasingly necessary, which can directly contribute to improved service quality and visitor satisfaction. Manually reviewing the large number of reviews generated on platforms like Google Maps is impractical, motivating the development of automated methods such as Natural language processing (NLP).

NLP methods, particularly sentiment analysis, are widely used in tourism to interpret large-scale textual data from online reviews [3], [4], [5]. Although traditional sentiment analysis effectively measures overall customer satisfaction, it often needs to capture the complexities of tourist experiences within a single review. This limitation is from the fact that standard sentiment analysis treats the entire reviews as a single entity [6], potentially neglecting complex experiences, such as positive feedback on cleanliness coupled with negative feedback on service quality. Therefore, the sentiment classification may not be effectively reflected by this simple sentiment analysis method, which limits its capacity to offer insightful information.

To address the limitations of traditional sentiment analysis, the Aspect-Based Sentiment Analysis (ABSA) has become a potential solution by segmenting reviews into several aspect categories. Since the sentiment will be categorized by aspect and cause an Aspect-Sentiment pair, ABSA might be a potential solution to provide businesses with a granular understanding and point toward areas that need improvement [7], [8]. This method is more favorable and provides a more non-biased sentiment classification, offering a detailed sentiment score regarding each aspect written in the review. However, a significant challenge in ABSA is the scarcity of annotated datasets that capture various aspects across domains, such as tourism reviews [9] The limitation hinders the development of robust models capable of accurately categorizing sentiment and extracting subtle information. A significant increase in the quantity and diversity of training data is needed for ABSA to provide accurate results [10].

The transferability of encoder models within transformer-based architectures [11] has been a promising method in sentiment analysis tasks. The state of the art of transformer-based architecture includes GPT-2 [12], BERT [13], and XLNet [14]. BERT has gained popularity due to its more manageable and compact design, improving contextual comprehension for several NLP tasks. The extensions, including RoBERTa [15] and ALBERT [16] have shown improved performance capabilities. BERT pre-trained models can be applied to a variety of tasks in different languages by fine-tuning [17], [18], [19]. For instance, previous research [20] has demonstrated that refined BERT performs exceptionally well in sentiment analysis, surpassing models such as LSTM and TextCNN, particularly in Vietnamese sentiment analysis [21].

BERT model has been adapted for Indonesian NLP tasks through the development of IndoBERT [22]. However, since the pre-training data does not entirely correlate with the needs of the tourism sector, it may be missing some of the specialized language patterns and terminology commonly used in tourism. Therefore, this research aimed to fine-tune IndoBERT, using a collection of Indonesian tourism reviews. The objective was to help the model adapt to the specialized domain, distinctive phrasing, and frequent sentiment expressions. A full assessment was also conducted to determine how well the fine-tuned model identifies various elements and emotions inside the reviews, providing more understanding regarding its effectiveness in generalizing unseen data.

II. Literature Review

Pre-trained language models, such as BERT [13] have become incredibly popular in NLP for various tasks including ABSA. The strength of BERT lies in its ability to generate context-aware embeddings, which help improve understanding accuracy by considering word context. This can be achieved through the transformer architecture and bidirectional pre-training, which allows BERT to learn from both directions of text, forward, and backward, to create better nuanced and precise word as well as context representations. The model has proven to succeed in several NLP task implementations [13], [24], [25]. In ABSA, BERT has demonstrated its robustness as a pre-trained model and adapted to several specific domains. As shown by several reports, BERT can be modified in both feature-based methods [18], [21] and fine-tuning methods on typical single-sentence inputs [28], [29], [30], with proven flexibility and effectiveness to be specific in sentiment analysis tasks.

IndoBERT [20], a variant of BERT explicitly tailored for the Indonesian language, has shown promising results in the context of the Indonesian language. This variant was trained on a vast and diverse Indonesian corpus, serving as an ideal candidate for tasks including Indonesian UGC, such as tourism reviews. Similar to BERT base architecture, IndoBERT follows the standard BERT-Base (uncased) configuration, consisting of 12 hidden layers with 768 dimensions each, 12 attention heads, and feed-forward hidden layers with 3,072 dimensions. However, IndoBERT was trained on a 31,923-size Indonesian WordPiece vocabulary, which was curated from over 220 million words aggregated sourced from three primary domains. These included Indonesian Wikipedia, which provided 74 million words covering general knowledge across various topics, news articles from sources such as Kompas, Tempo, and

Liputan6, contributing 55 million words with a formal and news-centric style, as well as the Indonesian Web Corpus, comprising 90 million words. This last source introduces a more casual and colloquial tone, accurately representing the informal language and everyday expressions often found in UGC.

The development of IndoBERT addresses the growing need for Indonesian-language NLP tools, as most pre-trained models are predominantly developed for high-resource languages such as English. IndoBERT enables a more accurate and nuanced analysis of Indonesian texts, crucial for tasks such as ABSA in tourism. According to previous research [32], pre-trained models including BERT and IndoBERT significantly improve performance on domain-specific tasks by transferring knowledge from large-scale corpora to specific applications like sentiment analysis. The success of IndoBERT is attributed to the ability to adapt to new tasks with minimal task-specific data, serving as a valuable tool for extracting insights from UGC in underrepresented languages like Indonesian.

III. METHODS

This research focused on fine-tuning IndoBERT for information extraction and key aspect detection from UGC related to tourism, specifically from Google Reviews. The method was structured into several phases, namely data collection, preprocessing, model fine-tuning, aspect, sentiment extraction, and performance evaluation, as shown in Figure 1.

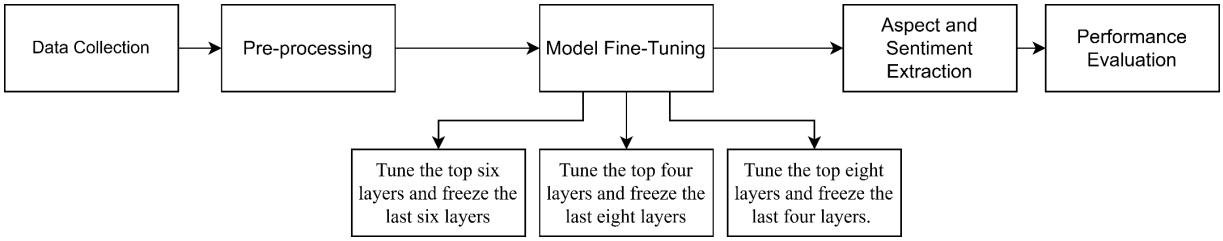

Fig. 1 Research Method Flow Diagram

A. Data Collection

The dataset for this research was collected from Google Reviews, targeting the top 20 tourism destinations in DI Yogyakarta and Jawa Tengah. These regions were selected due to cultural and tourism significance in Indonesia. DI Yogyakarta is home to world-renowned heritage sites such as Keraton Yogyakarta, Tamansari, and Museum Sonobudoyo. Meanwhile, Jawa Tengah offers diverse tourist attractions, ranging from historical cities to natural landscapes like the Dieng Plateau. These regions attracted many domestic tourist, serving as ideal sources for tourism-related UGC. The dataset focused on reviews written in Indonesian language, ensuring the research captured the sentiments and perspectives of local tourist.

TABLE 1
DATASET SAMPLE

Place	Rating	Review	Label
Keraton Ngayogyakarta Hadiningrat	4	*Menambah edukasi terkait adat istiadat dan budaya jawa khususnya pada daerah Yogyakarta*	[["Attractions and Activities: "positive"]]
Dieng Plateau	5	*Menawarkan pemandangan yg indah, ada candi, telaga, tempatnya dingin, pegunungan*	[["Attractions and Activities: "positive"]]
Museum Sonobudoyo	5	*Lokasi dekat malioboro dengan harga tiket masuk Rp10. 000. Museum nyaman dan ada fasilitas seperti toilet dan mushola. Museum berisikan tentang perkembangan budaya di Indonesia khususnya Jawa.*	[["Location and Accessibility: positive"], ["Amenities and Facilities: "positive"]]
Pantai Parangtritis	2	*Terlalu ramai pengunjung, banyak kereta kuda dan penjaja jasa photo, pasirnya hitam tidak seperti pantai di Wonosari Gunung Kidul yang berpasir putih. mempunyai panjang pantai yang luas bagus buat rombongan besar.*	[["Attractions and Activities: "negative"], ["Cleanliness: "negative"]]
Tamansari	4	*Tempatnya nyaman dan bersih, tapi akses menuju Tamansari agak sulit dijangkau dengan kendaraan umum.*	[["Cleanliness: "positive"], ["Location and Accessibility:"negative"]]

A total of 20,000 Google Reviews were collected, with 1,000 from each of the 20 destinations, as shown in Table 1. As this research targeted the information extraction in the ABSA task, the data was categorized by overall sentiment

(positive, neutral, and negative) as well as specific tourism-related aspects essential for evaluating service-related aspects. To ensure balanced representation across all sentiment categories, the dataset was designed to have an equal distribution of positive, neutral, and negative ratings for each aspect. This method ensured comprehensive sentiment coverage for each aspect, leading to a well-rounded dataset suitable for training the sentiment analysis model. Each review included the complete text, a star rating (1 to 5), and relevant metadata such as review date, location, and reviewer information, when available.

The selected aspects for analysis included Service Quality, which consisted of all reviews regarding feedback on customer service, staff friendliness, and responsiveness. Cleanliness focuses on the hygiene standards of facilities, accommodation, or public spaces. Location and accessibility covered convenience, access routes, and the general accessibility of the tourist site. Amenities and facilities showed the quality of available amenities including restrooms, parking, restaurants, and lodging options. Meanwhile, attractions and activities focused on the quality and diversity of tourist activities or attractions at the destination.

B. Pre-processing

The data collected were subjected to multiple preprocessing stages to ensure consistency, remove noise, and prepare for model training. Each phase improved data quality, effectively allowing IndoBERT to learn from the input. Initially, text normalization was performed by converting all characters to lowercase and deleting special characters, digits, and superfluous symbols, thereby standardizing the text. For example, a review *"Tempatnya sangat bagus!! Pemandangannya INDAH sekali. 100% recommended!!!"* was standardized to *"tempatnya sangat bagus pemandangannya indah sekali recommended."*

Tokenization was conducted using IndoBERT's WordPiece tokenizer, capable of handling Indonesian subwords effectively. An example sentence such as "*Pemandangannya sangat menakjubkan, pelayanannya baik."* was tokenized into *["pemandangan", "##nya", "sangat", "menakjubkan", ",", "pelayanan", "##nya", "baik"].* Stop words like *"dan"* (and), *"di"* (in/at), and *"yang"* (which/that) were removed to streamline the data, transforming phrases such as *"Pantainya sangat bersih dan nyaman"* into *"pantai sangat bersih nyaman".* This was followed by lemmatization, reducing words to their root forms for consistency across the dataset. For instance, words like *"menginap"* (stayed), *"penginapan"* (accommodation), and *"menginapnya"* (the stay) were lemmatised to their root form *"inap."* Longer reviews were segmented into smaller units, such as sentences or clauses, to handle reviews covering multiple aspects. An example review like *"Pemandangannya luar biasa, tetapi toiletnya kotor. Pelayanannya juga ramah."* was divided into three parts, namely *"Pemandangannya luar biasa," "Toiletnya kotor,"* and *"Pelayanannya juga ramah."* Subsequently, the dataset was divided into three subsets, namely training, validation, and test, with percentage allocations of 70%, 15%, and 15%, respectively.

C. Model Fine-Tuning and Hyperparameter

As the primary proposed method, several fine-tuning tests were carried out to maximize the model's performance on the given task by modifying the parameters while maintaining particular layers. Only the upper layers of IndoBERT would be adjusted in each experiment, and several last layers remained unchanged. This method followed the procedures in previous research [29]to reduce the possibility of losing previously learned information and successfully maintaining broad language comprehension by adjusting the higher layers to the target domain. Additionally, fine-tuning only the top layers helps the model capture domain-specific semantics, preserving general language representations in the lower layers [28][32]

In this research, three experimental configurations were assessed, namely (1) Tune the top six layers and freeze the last six layers, (2) Tune the top four layers and freeze the last eight layers, and (3) Tune the top eight layers and freeze the last four layers. Upon comparing results across all experiments, the model configuration achieving the lowest validation loss was selected as the final fine-tuned IndoBERT model. The learning parameters used in these experiments are fixed across all experiments, and the hyperparameter values are detailed in Table 2.

D. Aspect and Sentiment Extraction

A combined method was used in this research to ensure accurate aspect labeling. For initial aspect identification, IndoBERT was used to analyze context similarity and detect key phrases in reviews, automatically relating aspects to UGC. This process started by tokenizing each UGC review and representing the text as contextualized embeddings using IndoBERT pre-trained model. The embeddings captured the semantic meaning of phrases within the review, allowing the model to identify relationships between the text and predefined aspect categories. A predefined set of aspect keywords, such as 'toilet,' 'service,' 'cleanliness,' and 'access,' was established to serve as reference phrases for the aspect categories ('Amenities and Facilities,' 'Service Quality'). IndoBERT calculated cosine similarity between the embeddings of UGC phrases and reference keywords. A high similarity score indicated that a given phrase was

contextually related to a specific aspect category. In cases where multiple aspect matches occurred for a single phrase, IndoBERT performed ranking based on similarity scores and assigned the highest-ranking aspect to the phrase.

TABLE 2
LEARNING PARAMETERS

Hyperparameter	Values	Details
Learning Rate	3e-5	This parameter is used to control the rate of weight updates during training. A learning rate of 3e-5 was selected based on best practices in fine-tuning BERT-like models. The original BERT [11] found that a learning rate in the 2e-5 to 5e-5 was effective for fine-tuning transformer-based models, allowing for stable training without overshooting optimal weight updates. A lower learning rate ensures gradual weight updates, preventing catastrophic forgetting of pre-trained knowledge. The selection of 3e-5 was in line with previous findings [35], where fine-tuning BERT for text classification tasks showed that 3e-5 provided a balance between fast convergence and stable training.
Batch Size	32	This parameter selects the number of samples processed before updating the model's eights. A batch size of 32 was selected to balance computational cost and convergence stability. Smaller batch sizes like 16 can lead to higher variance in weight updates, potentially destabilizing the learning process. Meanwhile, large batch sizes like 64 may slow down convergence. Research on fine-tuning BERT models, including [36] RoBERTa, shows that a batch size of 32 is widely used for its effectiveness in transformer-based models.
Number of Epochs	5	This parameter selects the number of passes over the entire training dataset, with fine-tuning conducted for five epochs. In experiments with fewer epochs (3), the model showed underfitting, while extending training beyond five epochs (10) led to overfitting, where the validation loss increased despite continued improvements in training loss. Related research [37] reported the stability in fine-tuning BERT models, stating that 3-5 epochs were often optimal for fine-tuning, balancing learning with the risk of overfitting. The use of early stopping based on validation loss ensured that training did not proceed beyond the point of diminishing returns.
Dropout Rate	0.3	This parameter is used for regularisation method to prevent overfitting. A dropout rate of 0.3 was used to reduce the likelihood of overfitting. Previous research [38] introduced dropout as a regularisation method, showing that a rate between 0.2 and 0.5 prevented overfitting in large neural networks. The 0.3 was selected as it provided a balance, avoiding the risk of excessively reducing model capacity that could occur with higher dropout rates, such as 0.5. Meanwhile, the selected dropout rate provided sufficient regularisation to prevent overfitting on the training data.
Optimiser	AdamW	This parameter is used to adjust the model during training. The AdamW optimizer used an improved version, including weight decay, to counteract the problem of "overfitting by over-adaptation" in fine-tuning BERT-like models. Previous research [39] demonstrated that AdamW provided better generalization than Adam, particularly in fine-tuning large models like BERT. AdamW decoupling of weight decay and learning rate ensures that the model remains generalizable while adapting to the new task-specific dataset.

The automated process enabled IndoBERT to efficiently assign aspect labels to a large dataset of UGC reviews, significantly reducing the need for exhaustive manual labeling. However, to ensure accuracy and refine the model aspect assignments, manual annotation was conducted for a subset of the dataset. A total of three annotators independently labeled aspects and sentiments based on predefined guidelines, using the automated labels as references. Subsequently, inter-annotator agreement was calculated using Cohen's kappa, yielding a score of 0.82, which indicated strong agreement. Discrepancies were resolved through majority voting to ensure consistency. Furthermore, the sentiment score was determined using star ratings accompanying the reviews. Ratings of 4-5 were labeled as positive, while 3 and 1-2 represented neutral and negative, respectively. This structured method ensured that each review was correlated to relevant aspects and sentiments while maintaining the relationship with the context of the feedback.

E. Performance Evaluation

Performance evaluation explains that loss value will be evaluated to determine the best-fine-tuned method for extracting aspects and sentiments. However, strong generalization skills should be demonstrated by fine-tuning the model resilience in testing data. In this research, ABSA classification task evaluation used metrics, namely accuracy, recall, and F1-score, to measure how well the model could detect the aspect and the sentiment classification for unseen data.

IV. RESULTS

This section focuses on the results of adapting the IndoBERT and fine-tuning method to analyze sentiment in specific aspects of Indonesian tourism reviews, particularly those obtained from the research web scraping data collection from Google Reviews. Based on the results, the model fine-tuned overall loss performance is presented in the form of ABSA task of NLP, the evaluation of how well it classifies sentiment, accuracy in identifying different aspects, and key insights drawn from the review data performed using several classification metrics.

A. Model Performance

Several training experiments conducted in this research to enhance the IndoBERT model were used to assess fine-tune efficacy through validation loss evaluation. The results showed that the model top six layers adjusted while the bottom six remained fixed, yielding the lowest validation loss (0.324). Table 3 presents a comparative view of validation loss across all tested experiments.

TABLE 3
VALIDATION LOSS FOR DIFFERENT FINE-TUNING CONFIGURATIONS

Experiment Configuration	Validation Loss
Tune the top six layers and freeze the last six layers	0.324 (Best)
Tune the top four layers and freeze the last eight layers	0.368
Tune the top eight layers and freeze the last four layers	0.353

The lowest validation loss is 0.324, indicating strong generalization capabilities, while other configurations caused higher validation loss values. Training and validation loss curves (Figure 2) showed that training loss steadily decreased over the first few epochs, reaching convergence by epoch 5. Similarly, validation loss improved consistently, with minimal divergence between training and validation loss, confirming that the selected model generalized well to the validation set.

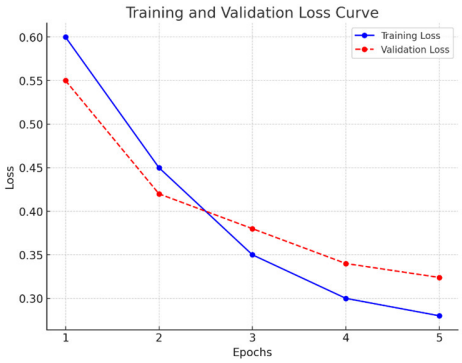

Fig. 2 Training and Validation Loss Curve

B. Model Performance on Aspect Detection

The fine-tuned IndoBERT model accurately detected critical aspects within the tourism reviews. The primary aspects analyzed, namely service quality, cleanliness, location, accessibility, amenities, facilities, attractions, and activities, were identified with high precision and recall. Table 4 summarises the model performance in aspect detection for each key aspect, including precision, recall, and F1-score. The model showed consistent performance across all aspects, with precision values ranging between 0.85 and 0.88, recall between 0.82 and 0.86, and F1 scores averaging approximately 0.85.

TABLE 4
ASPECT DETECTION

Aspects	Precision	Recall	F1-Score
Service Quality	0.87	0.84	0.85
Cleanliness	0.80	0.86	0.87
Location and Accessibility	0.85	0.83	0.84
Amenities and Facilities	0.86	0.82	0.84
Attractions and Activities	0.88	0.85	0.86

C. Model Performance on Sentiment Classification

Sentiment classification was performed for each detected aspect to categorize the sentiment as positive, neutral, or negative. The overall model performance for sentiment classification was evaluated using accuracy, precision, recall, and F1 scores provided in Table 5. The model achieved an overall sentiment classification accuracy of 0.84, with high performance on positive and negative sentiments, where precision and recall exceeded 0.85.

TABLE 5
SENTIMENT DETECTION

Sentiments	Accuracy	Precision	Recall	F1-Score
Positive	0.88	0.89	0.87	0.88
Neutral	0.81	0.83	0.80	0.81
Negative	0.84	0.85	0.83	0.84

D. Aspect-Specific Sentiment Insights

The fine-tuned IndoBERT model provided detailed insights into the sentiments associated with specific aspects across the tourism destinations. The distribution of sentiment for each aspect is summarized in Table 6. These results showed that cleanliness and attractions including activities received the highest positive sentiments, with 70% and 68% of reviews expressing satisfaction in these areas, respectively. Meanwhile, service quality, amenities, and facilities had a more balanced distribution of positive, neutral, and negative sentiments, showing areas where improvements could be made to enhance tourist experiences.

TABLE 6
ASPECT SENTIMENT INSIGHTS

Aspects	Positive (%)	Neutral (%)	Negative (%)
Service Quality	60	20	20
Cleanliness	70	15	15
Location and Accessibility	65	18	17
Amenities and Facilities	55	25	20
Attractions and Activities	68	20	12

V. DISCUSSION

The experiments determined that the optimal configuration included freezing the bottom six layers of the model and fine-tuning the top six layers. This setup achieved the lowest validation loss at 0.324 and the highest precision, recall, and F1 scores. The outcome was in line with previous research on transformer-based models, which indicated that fine-tuning only the upper layers while keeping the lower ones frozen allowed the model to retain foundational language knowledge and adapt effectively to the specific domain. According to previous research [40], a similar strategy was applied to multilingual BERT (m-BERT) for Indonesian ABSA. The results showed an 8% increase in F1-score by selectively leveraging BERT pre-trained layers rather than adjusting the entire model. This shows the advantage of partial layer adaptation for domain-specific ABSA tasks.

The experiment, which included freezing only the last four layers, showed a slightly higher validation loss of 0.353. Although this method allowed the model to incorporate domain-specific semantics through the top layers, freezing fewer levels could cause some overwriting of pre-trained knowledge, increasing the risk of overfitting. These results are supported by previous research [32], where the lowest layers of pre-trained models retained essential language knowledge, improving generalization across tasks.

Freezing several layers, as in the experiment where eight layers were frozen, caused a higher validation loss of 0.368. In this case, the model's inability to adjust enough layers likely limited its capacity to adapt fully to the nuances of the tourism domain. Fine-tuning limited layers can cause underfitting, where the model fails to capture the domain-specific information effectively. Similarly, freezing excessive layers during fine-tuning reduced the model's ability to adapt to new tasks [35].

The results indicated that fine-tuned IndoBERT achieved precision scores between 0.85 and 0.88 in aspect detection. This showed that the model successfully extracted key aspects such as service quality, cleanliness, location, and accessibility, as well as amenities and facilities from the reviews. Generally, ABSA includes identifying aspect categories within text and associating with sentiments [7]. The high F1-scores achieved in this research for both aspect detection and sentiment classification showed that the fine-tuned IndoBERT model met these criteria effectively. This high performance could be attributed to the context-aware embeddings provided by the transformer-based architecture of IndoBERT, which allowed capturing the relationship between words more effectively than traditional models. The

model achieved an overall accuracy of 0.84 in sentiment classification and exceptionally high performance in positive and negative sentiment detection. The slightly lower performance for neutral sentiment (F1-score: 0.81) can be explained by the ambiguity often present in neutral reviews, where the language used is more difficult to classify. This shows that there is space for development by addressing representations of sentiment, although IndoBERT excels at detecting strong opinions.

Detailed error analysis showed many issues that the model faced. One of the most significant concerns was the tendency to conflate neutral sentiments with negative ones, particularly in evaluations with mild criticism or factual comments that were not blatantly hostile. A review like "*Pelayanan cukup cepat, tapi tidak ada yang istimewa*." ("Service was fast enough, but nothing special.") could be misinterpreted as negative rather than neutral. This error was most likely caused by the model emphasizing crucial aspects of the review and misinterpreting minor displeasure as a negative attitude. As reported in previous research [7], typical sentiment analysis methods struggle to capture the nuances of neutral sentiments, which are more ambiguous than expressing positive or negative thoughts. Furthermore, implementing a sentiment intensity analysis could aid in distinguishing between low- and high-intensity comments, allowing the model to discover better critical evaluations without being overly negative.

Another significant challenge includes overlapping aspects within a single review. Specifically, some reviews contained mixed sentiments for different aspects, such as *"Pemandangannya sangat indah, tapi toiletnya kotor."* ("The view is beautiful, but the toilets are dirty."). In this case, the model occasionally transferred sentiment from one aspect to another, leading to incorrect classifications. The issue potentially stemmed from the complexity of detecting multiple aspects within a single review. Moreover, ABSA has been found difficult when reviews contain multiple aspects with contrasting sentiments, as the model must provide a significant difference [9]. A potential solution could be to use aspect segmentation techniques to clarify where one aspect ends and another begins, allowing the model to focus on each aspect individually and avoid sentiment "leakage".

Mixed sentiment errors can lead to contradictory conclusions in tourism feedback systems, especially when a review contains both positive and negative sentiments for different aspects. For instance, when a review praises the location but criticizes the amenities, failing to separate these sentiments has the potential to cause an overall neutral classification that obscures actionable insights. This can delay necessary improvements or misallocate resources to address incorrect issues, thereby impacting customer satisfaction.

Incorporating external tourism-specific datasets is capable of significantly improving the adaptability of IndoBERT by exposing the model to different linguistic styles, colloquialisms, and sentiment variations commonly found in tourism blogs, online reviews, and booking platforms. For example, *"Tempatnya bagus, tapi stafnya kurang ramah."* ("The place is nice, but the staff is not friendly."), where the overall sentiment was neither purely positive nor negative. In these cases, the model tended to classify the sentiment as neutral or even incorrectly because of the difficulty in handling mixed sentiments. Furthermore, the model found it challenging to manage subjective reviews, where expectations could vary widely. *"Penginapannya bagus, tapi tidak sesuai dengan harga yang dibayar."* ("The accommodation was reasonable, but not worth the price paid.") was classified as negative. This was because the model prioritized the price critique over the positive note about the accommodation. Future improvements might include multi-label classification methods, allowing each aspect and its sentiment to be categorized independently. The method can help the model address reviews with mixed or conflicting sentiments.

Although this research has not been applied in a live tourism analytics system, the results provide valuable insights that support the development of practical recommendations for real-world implementation. The identified challenges and model performance results guide the following evidence-based suggestions for future integration. Pilot research in partnership with local tourism boards could be conducted to transition this research into a tourism feedback system. A controlled environment would allow for performance monitoring and feedback collection, informing necessary refinements in sentiment tracking dashboards. Furthermore, sentiment-tracking dashboards can be implemented to visualize trends by aspect such as service quality and amenities. These dashboards enable quick identification of improvement areas, helping stakeholders address customer concerns more effectively. The training set can be enriched with additional datasets from travel forums, blogs, and booking platforms to address the challenge of mixed or conflicting sentiments. Incorporating diverse linguistic expressions from these sources ensures the model handles domain-specific nuances and improves adaptability.

VI. CONCLUSIONS

In conclusion, this research successfully fine-tuned the IndoBERT model for ABSA on Indonesian-language UGC from Google Reviews, showing the value of UGC as a rich source of authentic customer feedback. Further analysis was conducted to develop an automated method for extracting and evaluating essential elements as well as feelings from tourism reviews, addressing the industry's requirement for more efficient and accurate feedback interpretation.

The fine-tuning process included multiple experimental configurations, exploring different frozen and fine-tuned layer combinations to determine the best performance. The best architecture, with the lower six layers frozen and the upper six layers fine-tuned, had the lowest validation loss of 0.324 and performed well in aspect identification and sentiment classification tasks. This improved algorithm effectively identified important aspects of tourism reviews, such as service quality, cleanliness, location, amenities, and attractions. The model was highly accurate, with precision scores between 85% and 89%, classifying overall sentiment (positive, negative, or neutral) with an accuracy of 84%. Although IndoBERT excelled at identifying strong positive and negative feelings, there was difficulty with more nuanced, mixed, or uncertain sentiments, as indicated by the lower F1-score of 81% for neutral emotions.

Despite the significant performance, error analysis showed several shortcomings, such as misclassifying neutral and negative sentiments, overlapping aspects, as well as managing evaluations with mixed or highly subjective sentiments. These results suggested further refinement, particularly in interpreting subtle sentiment expressions and processing reviews with multiple or conflicting sentiments. Moreover, future research should address these limitations by incorporating sentiment intensity analysis, aspect segmentation, and multi-label classification methods. These improvements would enhance the model ability to capture nuanced sentiments and properly show the complexity of user feedback in tourism reviews.

Author Contributions: *Rifki Indra Perwira:* Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision *Vynska Amalia Permadi:* Investigation, Data Curation, Writing - Original Draft. *Dian Indri Purnamasari:* Investigation, Data Curation. *Riza Prapascatama Agusdin:* Methodology, Writing - Review & Editing.

All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Directorate General of Research and Strengthening of Research and Development (DRTPM) of the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) of Indonesia under Penelitian Fundamental - Reguler Penelitian Kompetitif Nasional [Fundamental Research - National Competitive Regular Research] program Grand number 080/E5/PG.02.00.PL/2024.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The dataset used in this study was collected from Google Reviews, specifically targeting reviews from 20 popular tourism destinations in DI Yogyakarta and Jawa Tengah. Due to the platform's privacy considerations and data-sharing restrictions, the entire dataset cannot be publicly available. However, to support reproducibility and transparency, this paper includes detailed descriptions of the dataset and representative data samples in Table 1. These samples and descriptions provide sufficient insight into the structure and nature of the data to facilitate understanding and potential replication of the analysis.

Informed Consent: There were no human subjects.

Institutional Review Board Statement: Not applicable.

Animal Subjects: There were no animal subjects.

ORCID:
Rifki Indra Perwira: https://orcid.org/0000-0002-6476-0573
Vynska Amalia Permadi: https://orcid.org/0000-0002-8631-5366
Dian Indri Purnamasari: https://orcid.org/0000-0002-0058-3408
Riza Prapascatama Agusdin: https://orcid.org/0009-0008-9356-4484

REFERENCES

[1] A. Fronzetti Colladon, B. Guardabascio, and R. Innarella, "Using social network and semantic analysis to analyze online travel forums and forecast tourism demand," *Decis Support Syst*, vol. 123, p. 113075, Aug. 2019, doi: 10.1016/j.dss.2019.113075.

[2] S. Barandoni, F. Chiarello, L. Cascone, E. Marrale, and S. Puccio, "Automating Customer Needs Analysis: A Comparative Study of Large Language Models in the Travel Industry," Apr. 2024.

[3] N. Raghunathan and K. Saravanakumar, "Challenges and Issues in Sentiment Analysis: A Comprehensive Survey," *IEEE Access*, vol. 11, pp. 69626–69642, 2023, doi: 10.1109/ACCESS.2023.3293041.
[4] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Trans Knowl Data Eng*, vol. 35, no. 11, pp. 11019–11038, Nov. 2023, doi: 10.1109/TKDE.2022.3230975.
[5] A. Jain, A. Bansal, and S. Tomar, "Aspect-Based Sentiment Analysis of Online Reviews for Business Intelligence," *International Journal of Information Technologies and Systems Approach*, vol. 15, no. 3, pp. 1–21, Aug. 2022, doi: 10.4018/IJITSA.307029.
[6] N. Raghunathan and K. Saravanakumar, "Challenges and Issues in Sentiment Analysis: A Comprehensive Survey," *IEEE Access*, vol. 11, pp. 69626–69642, 2023, doi: 10.1109/ACCESS.2023.3293041.
[7] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Trans Knowl Data Eng*, vol. 35, no. 11, pp. 11019–11038, Nov. 2023, doi: 10.1109/TKDE.2022.3230975.
[8] A. Jain, A. Bansal, and S. Tomar, "Aspect-Based Sentiment Analysis of Online Reviews for Business Intelligence," *International Journal of Information Technologies and Systems Approach*, vol. 15, no. 3, pp. 1–21, Aug. 2022, doi: 10.4018/IJITSA.307029.
[9] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 6279–6284. doi: 10.18653/v1/D19-1654.
[10] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," *IEEE Trans Affect Comput*, vol. 13, no. 2, pp. 845–863, Apr. 2022, doi: 10.1109/TAFFC.2020.2970399.
[11] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762
[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.
[13] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Proceedings of the Conference*, 2019, pp. 4171–4186. [Online]. Available: https://github.com/tensorflow/tensor2tensor
[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019. [Online]. Available: https://github.com/zihangdai/xlnet
[15] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[16] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[17] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020. [Online]. Available: https://github.com/annisanurulazhar/absa-playground
[18] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, and V. Basile, "ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets," in *CEUR Workshop Proceedings*, 2019. [Online]. Available: https://github.com/google-research/
[19] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine, and M. Boeker, "GottBERT: a pure German Language Model," 2020. [Online]. Available: https://tblock.github.io/10kGNAD
[20] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," 2020. [Online]. Available: https://github.
[21] Q. Thai Nguyen, T. Linh Nguyen, N. Hoang Luong, and Q. Hung Ngo, "Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews," in *Proceedings - 2020 7th NAFOSTED Conference on Information and Computer Science, NICS 2020,* 2020. [Online]. Available: https://github.com/google-research/bert
[22] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.00677
[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[24] E. Yulianti, N. Pangestu, and M. A. Jiwanggi, "Enhanced TextRank using weighted word embedding for text summarization," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, p. 5472, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5472-5482.
[25] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, p. 5641, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.
[26] M. Polignano, V. Basile, P. Basile, M. de Gemmis, and G. Semeraro, "AlBERTo: Modeling Italian Social Media Language with BERT," *Italian Journal of Computational Linguistics*, vol. 5, no. 2, pp. 11–31, Dec. 2019, doi: 10.4000/ijcol.472.
[27] Q. T. Nguyen, T. L. Nguyen, N. H. Luong, and Q. H. Ngo, "Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, IEEE, Nov. 2020, pp. 302–307. doi: 10.1109/NICS51282.2020.9335899.
[28] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 1, p. 79, Feb. 2023, doi: 10.22146/ijccs.77354.
[29] F. Said and L. P. Manik, "Aspect-Based Sentiment Analysis on Indonesian Presidential Election Using Deep Learning," *Paradigma - Jurnal Komputer dan Informatika*, vol. 24, no. 2, pp. 160–167, Sep. 2022, doi: 10.31294/paradigma.v24i2.1415.
[30] A. Tiwari, K. Tewari, S. Dawar, A. Singh, and N. Rathee, "Comparative Analysis on Aspect-based Sentiment using BERT," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Feb. 2023, pp. 723–727. doi: 10.1109/ICCMC56507.2023.10084294.
[31] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," May 2019.
[32] M. E. Peters *et al.*, "Deep contextualized word representations," Feb. 2018.

[33] F. Said and L. P. Manik, "Aspect-Based Sentiment Analysis on Indonesian Presidential Election Using Deep Learning," *Paradigma - Jurnal Komputer dan Informatika*, vol. 24, no. 2, pp. 160–167, Sep. 2022, doi: 10.31294/paradigma.v24i2.1415.
[34] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 1, p. 79, Feb. 2023, doi: 10.22146/ijccs.77354.
[35] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," May 2019.
[36] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019. [Online]. Available: https://github.com/pytorch/fairseq
[37] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines," Jun. 2020.
[38] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," 2014.
[39] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Nov. 2017.
[40] A. N. Azhar and M. L. Khodra, "Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ICAICTA49861.2020.9428882.