Journal of Computer Science

Research Article

A Sentiment-Based Evaluation of Museum Display Design through Hybrid IndoBERT and Rule-Based Lexicon

Erneza Dewi Krishnasari[1], Yaddarabullah[2], Bayyinah Nurrul Haq[3] Aedah Abd Rahman[4] and Lahandi Baskoro[5]

[1]*Department of Visual Communication Design, Universitas Trilogi, Jakarta, Indonesia*
[2]*Department of Informatics, Universitas Trilogi, Jakarta, Indonesia*
[3]*Department of Industrial Design, Universitas Trilogi, Jakarta, Indonesia*
[4]*Schools of Science and Technology, Asia e University, Selangor, Malaysia*
[5]*Business School, London South Bank University, London, United Kingdom*

Article history
Received: 06-10-2025
Revised: 07-02-2026
Accepted: 14-04-2026

Corresponding Author:
Erneza Dewi Krishnasari
Department of Visual Communication Design, Universitas Trilogi, Jakarta, Indonesia
Email: ernezadewi@trilogi.ac.id

Abstract: Traditional museums remain vital cultural institutions but face persistent challenges in engaging contemporary audiences. The Wayang Museum in Indonesia, despite partial digital renovation, continues to rely heavily on static, text-based displays that often hinder visual comfort and informational comprehension. Conventional sentiment analysis techniques have demonstrated limitations in accurately capturing the subtle feedback from visitors, particularly within the culturally nuanced contexts of Indonesia. To address this problem, this study proposes a hybrid Aspect-Based Sentiment Analysis (ABSA) model that integrates a fine-tuned IndoBERT transformer with a rule-based sentiment lexicon. The hybrid architecture combines probabilistic embeddings with domain-specific lexicon rules to enhance classification accuracy, calibration, and interpretability. Experimental results demonstrate that while a baseline IndoBERT model achieved high accuracy (98.3%) and macro F1-score (0.975), the proposed IndoBERT-Lexicon model achieved perfect classification accuracy and near-perfect calibration (Negative Log-Likelihood = 0.002, Expected Calibration Error = 0.002). In comparison, a classical SVM with TF-IDF achieved similar accuracy but exhibited significantly weaker calibration despite its superior computational efficiency. The originality of this work lies in demonstrating that integrating a culturally grounded rule-based lexicon with transformer embeddings can simultaneously improve accuracy, calibration reliability, and interpretability in aspect-based sentiment analysis for cultural heritage applications. The key contribution of this study lies in demonstrating the methodological effectiveness of combining transformer embeddings with a culturally grounded lexicon to enhance both accuracy and calibration.

Keywords: Aspect-Based Sentiment Analysis, IndoBERT, Hybrid Transformer, Rule-Based Lexicon, Sentiment Classification

Introduction

Museums play an essential role as cultural institutions that conserve and convey artistic, historical, and philosophical legacies. The Wayang Museum in Indonesia serves as a prime example by showcasing traditional puppetry and folklore. Recently, the museum has initiated a partial digital renovation; however, a significant portion of its collection continues to depend on static, text-heavy displays. These traditional methods, while culturally authentic, often struggle to engage contemporary visitors effectively, creating a disconnect between artifacts and audience interaction. Understanding the dimensions of visual comfort and informational comprehension becomes vital in addressing these challenges. Visual comfort in museum contexts pertains to the clarity and aesthetic organization of exhibits. Poor lighting, unclear labels, and densely packed text can severely hinder visitors' experiences.

The importance of enhancing visual presentation is underscored by studies that indicate visitors are more likely to engage with exhibits that prioritize clarity and

design (Wang et al., 2024). For example, research has shown that aesthetically pleasing and well-organized displays can significantly enhance visitors' initial reactions and overall engagement (López-Martínez et al., 2020). In scenarios where museums still utilize conventional display formats, the lack of dynamic and interactive elements can leave visitors disengaged, as they face the 'monotonous storytelling' often present in static exhibits (Yi et al., 2024).

Informational comprehension is another critical dimension that affects visitors' experiences. Visitors often encounter challenges related to the readability and interpretability of curatorial texts. It is widely recognized that museums need to adapt their communication strategies to cater to diverse audiences. Traditional text-heavy displays can be overwhelming and often result in cognitive overload for visitors (Yi et al., 2024). Effective curatorial practices should employ concise and clear texts paired with engaging visual aids to facilitate better understanding and retention of information (López-Martínez et al., 2020). As highlighted in research focusing on transforming museum experiences, curators must develop strategies that not only embrace technological advancements but also consider the cognitive and emotional dimensions of visitor interactions with cultural artifacts (Chen and Ryan, 2020).

Beyond textual clarity, technological integration in museums increasingly relies on visualization as a central medium for engagement. Rather than static text, information can be communicated through dynamic visual representations that highlight patterns, relationships, and narratives in ways that are more intuitive and memorable. Such visualization strategies not only support cognitive comprehension but also reinforce aesthetic appeal, creating displays that are both informative and emotionally resonant. Furthermore, factors such as display aesthetics and spatial organization significantly influence visitor satisfaction and information retention (Huo et al., 2024).

Building on this challenge, visitor perspectives can be systematically explored through sentiment analysis of their feedback. Yet, conventional approaches often collapse feedback into broad categories such as positive, negative, or neutral, overlooking cultural and linguistic nuances that shape perceptions. This limitation is particularly evident in Indonesia, where subtle variations in expression influence how impressions are articulated (Darda et al., 2024). Traditional models often fail to capture complex aspects such as aesthetic appreciation or the clarity of information, which vary across visitor demographics (Yi et al., 2022). Furthermore, existing models struggle to capture complex elements such as aesthetic appreciation and information clarity, and they frequently suffer from miscalibration, where predicted probabilities do not align with actual distributions (Ahmed, 2023; Akgül and Eren, 2024). These shortcomings reduce the usefulness of sentiment analysis as a decision-support tool for museum curators.

Recent studies underscore that sentiment analysis serves not only as a methodological framework but also as a practical tool for institutional quality enhancement. (Vaghela et al., 2024) demonstrated the effectiveness of aspect-based sentiment analysis through self-attention LSTM, showing how fine-grained sentiment detection can capture user experiences more accurately. (Bheemarao et al., 2025) highlighted the importance of developing culturally adapted sentiment lexicons, proving that linguistic nuances in local contexts significantly improve institutional responsiveness to stakeholders. (Alqurafi and Alsanoosy, 2024) Further emphasized the value of sentiment analysis for measuring customer satisfaction, illustrating its potential in driving systematic quality improvements across organizations. Complementing these findings, Ramavath et al. (2025) introduced lightweight gradient boosting models to enhance efficiency and accuracy in sentiment classification, reinforcing the role of scalable approaches for real-world applications. Collectively, these contributions affirm that sentiment analysis is a critical instrument for institutions, including cultural organizations such as museums, to evaluate stakeholder feedback, identify gaps, and design strategies that strengthen engagement and service quality.

To address such gaps, Aspect-Based Sentiment Analysis (ABSA) offers a more fine-grained alternative. Unlike traditional sentiment analysis that assigns a single polarity to an entire review, ABSA identifies sentiments linked to specific aspects of an experience. For instance, a visitor may enjoy the overall impression of a museum but criticize the clarity of its labels. ABSA can disentangle such feedback and classify it across predefined aspects.

While ABSA provides the conceptual framework, its success depends on the underlying language model. IndoBERT, a transformer-based model pre-trained on large-scale Indonesian corpora, captures nuanced semantic and syntactic patterns in text. It has demonstrated strong performance in Indonesian natural language processing tasks, including sentiment analysis, text classification, and named entity recognition. Its ability to represent subtle linguistic variations makes it well-suited for analyzing visitor reviews that contain colloquial expressions, cultural idioms, or domain-specific terminology. Nevertheless, like other deep learning models, IndoBERT may still face challenges in calibration and interpretability, which motivates the need for integration with complementary approaches such as rule-based lexicons.

In response, this study proposes a hybrid Aspect-Based Sentiment Analysis (ABSA) model that integrates a fine-tuned IndoBERT transformer with a rule-based

lexicon. IndoBERT provides deep contextual embeddings tailored to Indonesian corpora, while the lexicon encodes culturally specific expressions related to visual comfort, display aesthetics, and information clarity. By combining probabilistic modeling with explicit linguistic rules, the proposed approach aims to deliver both high classification accuracy and reliable probability calibration, thereby overcoming key limitations of conventional sentiment analysis.

Accordingly, this study pursues two objectives: To develop and evaluate a hybrid IndoBERT–Lexicon model that advances aspect-based sentiment classification with improved calibration; and to benchmark its performance against baseline approaches using both standard metrics and calibration measures such as Negative Log-Likelihood (NLL), Brier Score, and Expected Calibration Error (ECE). These evaluations ensure that model outputs not only classify sentiment accurately but also provide trustworthy guidance for museum curators in decision-making.

Unlike prior sentiment analysis studies that emphasize accuracy alone, this work explicitly addresses the often-overlooked issue of probability calibration in decision-support systems. By integrating transformer-based embeddings with a culturally contextualized lexicon, this study introduces a hybrid framework that balances predictive performance, interpretability, and reliability, thereby extending the applicability of sentiment analysis beyond classification toward evidence-based cultural heritage management.

Related Work

Traditional museum displays, typically consisting of static text labels, lengthy curatorial descriptions, and formal artifact arrangements, remain essential for preserving authenticity and conveying cultural or educational narratives. However, these formats frequently create challenges related to visual clarity and effective information delivery. Dense textual content and rigid layouts can overwhelm visitors, often resulting in cognitive overload, decreased attention, and limited engagement with the exhibits. Such constraints may also reduce the interpretive depth available to audiences with diverse backgrounds or varying levels of prior knowledge. Recognizing these limitations, recent scholarship has emphasized the importance of analyzing how visitors perceive and respond to conventional displays. In this context, user-generated content, particularly online reviews, has emerged as a valuable source of data, offering direct insights into visitor experiences and providing a basis for evaluating museum service attributes and identifying areas for improvement (Huo et al., 2024). However, limited research has explored the 0application of sentiment analysis to capture nuanced feedback from visitors, particularly in the context of traditional museum displays (Freunek and Bodmer, 2021).

Traditional displays maintain significant cultural and educational value, particularly when they are thoughtfully curated to convey information in a clear and emotionally engaging manner (Adelakun, 2024). However, as visitors' expectations evolve in the digital age, relying solely on static formats may limit the depth of engagement and interpretive potential. To address this, recent studies have emphasized the value of enhancing traditional exhibits with complementary technologies. For example, incorporating augmented reality or audio guides into traditional formats can provide a layered storytelling experience that resonates with contemporary audiences while preserving cultural authenticity (Wang et al., 2024).

Although these technological advancements hold considerable promise, it is crucial to also address design improvements within the existing framework of traditional displays. Research on optimizing museum exhibits indicates that redesigning the spatial layout of displays can significantly enhance visitor comprehension and enjoyment. For example, Ji et al. (2024) applied mobile eye-tracking techniques to analyze the effects of object placement and label proximity on visual attention within museums. Their findings revealed that positioning artifacts and information at eye-level, typically between 100 and 150 cm, supports natural engagement and improves information retention, thereby facilitating a more meaningful visitor experience.

Visitor comfort and visual appeal play a critical role in shaping the effectiveness of museum displays. Aesthetic qualities such as the neatness of layout, color harmony, and the readability of text not only contribute to a pleasant viewing experience but also enhance cognitive accessibility (Kaczmarek-Gajewska and McDonnell, 2021). When visitors perceive the display as visually well-organized and easy to navigate, they are more likely to engage attentively with the content (Reppa and McDougall, 2022). This is particularly important in cultural institutions, where artifacts often carry complex historical and symbolic meanings (Hitsuwari and Nomura, 2022). Clear and well-structured explanations, presented in an approachable visual format, can significantly improve the transmission of cultural knowledge and educational value. Furthermore, visually engaging displays foster emotional resonance and sustained attention, which are essential for building meaningful connections between visitors and the objects on view (Isik and Vessel, 2021). Thus, the interplay between visual comfort, interpretive clarity, and cultural communication becomes a key determinant of the overall quality of the museum experience.

Despite this body of work, limited studies have systematically analyzed visitor feedback using natural language processing. Existing sentiment analysis research has primarily focused on binary or overall polarity classification, overlooking nuanced, aspect-specific

evaluations such as clarity of information or display aesthetics (Yadav et al., 2021). Moreover, transformer-based models like BERT have shown strong performance in general sentiment analysis but require adaptation to domain-specific and culturally embedded expressions, particularly in non-English contexts (Kalpana et al., 2022).

To fill this gap, our study applies an Aspect-Based Sentiment Analysis (ABSA) framework tailored to museum visitor narratives. Specifically, we propose a hybrid model that integrates a fine-tuned IndoBERT transformer with a culturally contextualized sentiment lexicon, enabling the classification of visitor perceptions along the dimensions of impression, display, and information. This approach directly addresses the limitations of prior works by combining contextual embeddings with domain-specific rules, improving both predictive accuracy and interpretability. In doing so, it offers actionable insights for curators and exhibit designers seeking to evaluate and refine traditional museum displays, with particular attention to visual comfort and informational clarity.

Beyond cultural heritage applications, machine learning models have demonstrated strong capability in modeling complex nonlinear patterns across diverse domains, including finance, cybersecurity, and human behavior analysis. Recent studies show that hybrid and ensemble-based models, particularly those combining deep learning with rule-based or optimization components, can outperform single-model approaches in data-limited and high-stakes environments. These findings further motivate the exploration of hybrid architectures in the present study, as museum visitor feedback exhibits similar challenges related to sparsity, contextual nuance, and reliability requirements.

In recent years, hybrid modeling has emerged as a general design paradigm in machine learning, particularly for tasks characterized by limited data, high uncertainty, and the need for reliable decision support (Latifah et al., 2024). Rather than relying solely on a single predictive model, hybrid approaches combine deep learning with complementary components such as rule-based systems or optimization algorithms to enhance robustness, interpretability, and generalization (Alshaabi et al., 2022). This paradigm has been increasingly adopted across diverse application domains beyond Natural Language Processing (NLP).

Recent studies in security and financial domains have demonstrated that hybrid deep learning architectures can effectively address data sparsity, model overfitting, and reliability challenges. For instance, a hybrid Autoencoder–Gated Recurrent Unit (AE–GRU) model optimized using the Honey Badger Algorithm has been shown to significantly improve cyber threat detection performance in IoT networks by combining representation learning with metaheuristic optimization (Addula et al., 2025). Similar hybrid strategies have also been adopted in financial fraud detection and risk assessment tasks, where model accuracy and robustness must be balanced with limited labeled data (Setiawan et al., 2025). These findings support the growing consensus that integrating deep learning models with complementary techniques such as optimization algorithms or rule-based systems can yield more reliable and interpretable solutions.

Aligned with this hybrid design philosophy, the present study adopts a similar integration strategy by combining a transformer-based IndoBERT model with a rule-based sentiment lexicon, aiming to balance contextual expressiveness, interpretability, and reliability in a culturally nuanced natural language processing task.

Methods

Dataset Source

The dataset for this study was collected through a structured questionnaire administered to visitors of the Wayang Museum, a traditional cultural institution in Jakarta, Indonesia. It was privately gathered for academic research purposes and is not publicly available due to ethical and institutional restrictions. Data collection was conducted during curated museum visits involving undergraduate students and members of the public. Each participant engaged with static, text-based Wayang (puppet) displays before completing the survey. The questionnaire included both closed-ended and open-ended items. Closed-ended items used Likert scales to assess aspects such as readability, layout, and clarity. Open-ended questions captured visitors' subjective impressions in their own words. Narrative responses were annotated according to three predefined aspects: Impression, Display, and Information. Sentiment polarity for each aspect was labeled as positive or negative, guided by culturally contextualized rubrics.

Feature Description and Pre-Processing

The primary features were the open-ended narrative responses, written in Bahasa Indonesia. These texts served as input for sentiment classification across the three aspects. Pre-processing followed a structured pipeline to ensure quality and compatibility with the IndoBERT tokenizer:

- Lowercasing: All responses were converted to lowercase to reduce vocabulary redundancy caused by inconsistent letter casing
- Punctuation and Special Character Removal: Non-linguistic symbols such as "!", "@" and extraneous whitespace were removed. This step minimized input noise while carefully retaining sentiment indicators like

question marks or ellipses when they conveyed tone
- Selective Stopword Removal: Stopwords in Bahasa Indonesia were removed only when they did not contribute to sentiment
- Normalization of Informal Language: A custom-built normalization dictionary was used to correct colloquial or abbreviated expressions common in informal text entries
- Tokenization with IndoBERT Tokenizer: The cleaned texts were tokenized using the IndoBERT tokenizer, which uses subword segmentation (Byte-Pair Encoding). This method helped handle out-of-vocabulary words and morphologically complex terms in Bahasa Indonesia
- Padding and Truncation: All sequences were adjusted to a fixed length of 128 tokens by padding shorter texts and truncating longer ones, ensuring uniform input for model training

Each review could receive up to three sentiment labels, including impression, display, and information, since a single visitor response often mentioned multiple aspects. For instance, a review might praise the overall experience but criticize label readability. This multi-label approach allowed the model to capture such nuanced opinions without reducing them to a single category.

Annotation followed a rubric that combined explicit keywords, contextual interpretation, and majority consensus among annotators. Positive sentiment was assigned when the visitor expressed satisfaction, clarity, or enjoyment in relation to a specific aspect, whereas negative sentiment was assigned when dissatisfaction, confusion, or discomfort was reported. Representative examples of annotated cases are summarized in Table 1, which illustrates the types of expressions encountered and their alignment with aspect-specific sentiment.

Lexicon Construction

To complement the IndoBERT backbone, a rule-based sentiment lexicon was constructed. This lexicon provides explicit polarity signals that are particularly useful for handling out-of-vocabulary terms, low-frequency expressions, or idiomatic phrases common in museum visitor language. By encoding domain-specific knowledge, the lexicon increases interpretability and acts as a fallback mechanism when IndoBERT's probabilistic predictions are uncertain or overconfident. The lexicon was derived from three sources:

- Prior cultural sentiment studies that documented affective expressions in Bahasa Indonesia across heritage and tourism domains
- Pilot survey data collected during preliminary visits to the Wayang Museum revealed colloquial terms and culturally embedded expressions frequently used by visitors
- Expert consultation with cultural heritage scholars and language specialists, who validated term polarity and ensured contextual appropriateness for the museum setting

Each entry in the lexicon was categorized according to the three predefined aspects, including impression, display, and information. Within each aspect, keywords were grouped into positive and negative sets. Positive keywords represent approval, satisfaction, or clarity, while negative keywords capture dissatisfaction, poor presentation, or lack of clarity.

To mitigate cultural and linguistic bias in lexicon construction, a multi-step validation process was employed. First, candidate terms were collected from prior cultural sentiment studies, pilot surveys, and field observations. Second, the terms were reviewed by cultural heritage scholars and linguists from different Indonesian socio-linguistic backgrounds to ensure consistent polarity across dialectal variations (e.g., buram vs. pudar to describe unclear visuals). Third, iterative cross-checking with annotators was conducted during dataset labeling to resolve disagreements and refine definitions. This process reduced the risk of misclassification due to regional or socio-linguistic nuance and increased confidence that the lexicon reflected widely understood visitor expressions rather than narrow or biased interpretations.

Model Architecture and Approach

This study employed a hybrid Aspect-Based Sentiment Analysis (ABSA) architecture that integrates a fine-tuned transformer model with a culturally contextualized lexicon. The purpose of this design was to combine the contextual depth of transformer embeddings with the interpretability and domain specificity of lexicon-based rules, thereby ensuring both predictive accuracy and calibration reliability.

Table 1: Aspect-based sentiment classification

Aspect	Sentiment	Description
Impression	Positive	"Pengalaman menyenangkan" (A pleasant experience)
	Negative	"Kurang berkesan" (Unimpressive)
Display	Positive	"Penataan rapi, koleksi jelas terlihat" (The arrangement is neat, and the collection is clearly visible)
	Negative	"Pencahayaan gelap, sulit dilihat" (The lighting was dark, making it difficult to see)
Information	Positive	"Penjelasan mudah dipahami" (The explanation is easy to understand)
	Negative	"Label terlalu kecil, penjelasan minim" (The labels are too small, and the explanations are minimal)

Although the present implementation focuses on textual sentiment analysis, the modular design of the proposed hybrid framework allows straightforward extension to multimodal inputs and cross-lingual settings in future deployments.

The motivation for adopting a hybrid architecture aligns with recent work in cybersecurity, where deep learning models such as AE–GRU are combined with optimization strategies to improve robustness and generalization under constrained data conditions. In contrast to optimization-based hybridization, this study integrates IndoBERT with a culturally grounded rule-based lexicon, aiming to enhance interpretability and probability calibration rather than solely predictive accuracy.

At the core of the system is IndoBERT, a transformer model pre-trained on large-scale Indonesian corpora. IndoBERT was fine-tuned on the museum dataset using a multi-output classification head that independently predicts sentiment polarity for the three aspects, including Impression, Display, and Information. The classification head consists of a dense layer with 256 units and ReLU activation, followed by dropout (0.3) for regularization, and a sigmoid-activated output layer with three nodes corresponding to the aspects. This setup ensures aspect-specific predictions without cross-contamination between dimensions.

To complement IndoBERT, a rule-based lexicon branch was developed. The lexicon, curated from Indonesian cultural discourse and visitor language, includes aspect-specific sentiment keywords with polarity assignments (see Table 2). This branch employs keyword matching and dependency heuristics to produce provisional sentiment scores, which are particularly useful for capturing underrepresented expressions or idiomatic terms not well covered in the training data.

Outputs from IndoBERT and the lexicon engine are combined through a weighted ensemble strategy. IndoBERT's probability estimates form the primary signal, while lexicon-derived scores provide corrective weighting whenever explicit sentiment markers are detected. This ensemble design improves recall for minority classes, enhances calibration, and adds interpretability by linking predictions to culturally specific keywords. The full architecture is summarized in Table 3, which distinguishes the two processing branches (IndoBERT and Lexicon) and their convergence in the ensemble strategy. The workflow is further visualized in Figure 3, which illustrates the pipeline from input through preprocessing, parallel model branches, and ensemble integration to final aspect-level outputs.

Fig. 1 illustrates the end-to-end pipeline of the proposed Aspect-Based Sentiment Analysis (ABSA) framework for processing museum visitor narratives. The architecture is structured into sequential and parallel stages to ensure robust representation learning, interpretability, and probability calibration.

The pipeline begins at the Input Layer, where raw textual narratives in Bahasa Indonesia are collected from visitors. These narratives undergo a Pre-Processing module, which performs data cleaning, normalization, tokenization, and other linguistic refinements to eliminate noise and standardize input representations.

Table 2: Example of Lexicon Elements by Aspect

Aspect	Positive Keywords	Negative Keywords
Impression	seru (exciting), menarik (interesting), puas (satisfied), suka (liking), menyenangkan (pleasant)	mengecewakan (disappointing), membosankan (boring), kurang berkesan (unimpressive)
Display	rapi (neat), indah (beautiful), pencahayaan baik (good lighting), jelas (clear)	gelap (dark), buram (blurry), kotor (dirty), berantakan (messy), sulit dilihat (hard to see)
Information	jelas (clear), lengkap (comprehensive), informatif (informative), mudah dipahami (easy to understand)	minim (insufficient), kurang jelas (unclear), sulit dimengerti (difficult to understand)

Table 3: Model Architecture of IndoBERT with Lexicon

Module	Component	Function
Input layer	Visitor narrative text	Raw museum visitor reviews in Bahasa Indonesia
Pre-processing	Cleaning, normalization, and tokenization	Standardizes text and prepares tokens for further processing
IndoBERT Branch	Transformer backbone (IndoBERT)	Generates contextual embeddings using pre-trained Indonesian corpora
	Classification head (Dense + Dropout + Sigmoid)	Produces aspect-level sentiment probabilities (Impression, Display, Information)
Lexicon Branch	Aspect-specific sentiment lexicon	Matches cultural/domain-specific keywords and assigns polarity scores
	Dependency and pattern heuristics	Enhances keyword-based polarity inference
Ensemble Strategy	Weighted aggregation	Combines IndoBERT probabilities with lexicon-derived scores
Output Layer	Aspect-level sentiment polarity	Final decision: Positive or Negative for each aspect

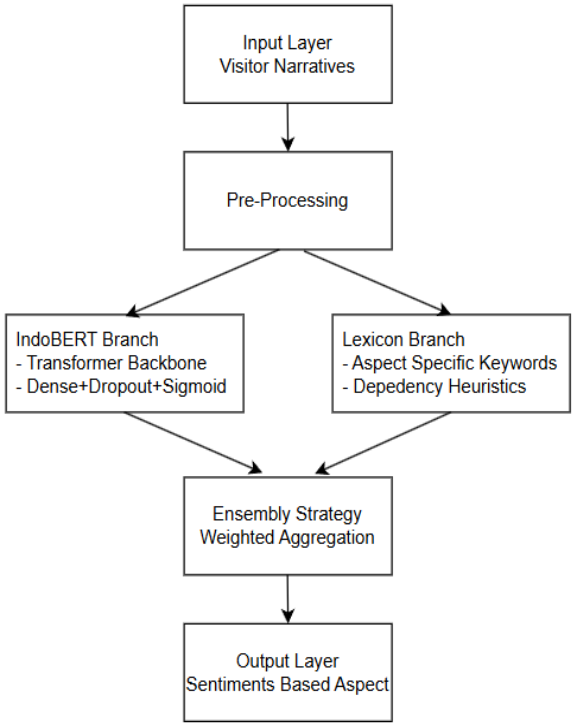

Fig. 1: Workflow of the IndoBERT with Lexicon

The workflow then bifurcates into two complementary analytical branches. The IndoBERT Branch employs a transformer-based backbone pre-trained on large-scale Indonesian corpora, ensuring sensitivity to semantic and syntactic nuances specific to the language. A classification head, composed of fully connected layers with dropout regularization and a sigmoid activation function, outputs probabilistic sentiment distributions for each predefined aspect: Impression, Display, and Information.

In parallel, the Lexicon Branch processes the same pre-processed input using a curated, domain-specific sentiment lexicon. This module identifies aspect-specific keywords and applies dependency heuristics to infer polarity scores, embedding cultural and contextual rules into the decision process.

The outputs from both branches are combined through an Ensemble Strategy based on weighted aggregation. Here, IndoBERT-derived probabilities serve as the primary predictive signal, while lexicon-based scores act as corrective weights. This integration mitigates IndoBERT's tendency toward overconfidence, enhances interpretability, and ensures alignment with culturally contextualized expressions. Finally, the Output Layer produces aspect-specific sentiment polarities (positive, neutral, negative) across the three museum-related dimensions. By fusing transformer-based contextual depth with rule-based cultural grounding, the framework achieves a balance between predictive accuracy, interpretability, and reliability, thereby offering a unified approach for sentiment-informed decision-making in cultural institutions.

Model Training

IndoBERT was fine-tuned using 5-fold stratified cross-validation. Training used binary cross-entropy loss with class weights to address minor label imbalance, optimized with Adam (learning rate = 2e-5). Training was capped at 10 epochs with early stopping based on validation F1. All experiments were conducted in Python 3.10 using PyTorch and HuggingFace Transformers.

Model Evaluation

The evaluation framework for this study was designed to provide a comprehensive assessment of the proposed models across four complementary perspectives: Evaluation metrics, classification performance, computational efficiency, and calibration reliability. This multi-perspective approach ensured that the models were not only compared on raw predictive accuracy but also on their robustness, deployability, and trustworthiness of probability estimates.

Each model was assessed using accuracy, precision, recall, and F1-score, computed with both macro and weighted averaging. Macro-averaging treated all sentiment categories equally, ensuring that minority classes (e.g., negative sentiment) were not overshadowed by majority ones. Weighted-averaging reflected the actual class distribution in the dataset. Validation loss was also monitored to ensure convergence, and the best-performing epoch was selected according to macro F1-score:

- To capture the practical feasibility of model deployment, computational efficiency was measured during evaluation. Three indicators were recorded
- Runtime per second, representing the total time required to complete the evaluation
- Samples per second, indicating throughput efficiency in terms of processed visitor reviews
- Steps per second, reflecting the scalability of each model during training and evaluation

Given that the study adopted an aspect-based sentiment analysis framework, performance was reported at the level of each predefined aspect: Impression, Display, and Information. For each aspect, precision, recall, F1-score, and accuracy were computed. This fine-grained evaluation made it possible to identify whether models struggled with specific aspects, such as handling minority negative classes under Impression. Aspect-level classification reports also supported comparison between IndoBERT-based and lexicon-augmented approaches.

Beyond categorical correctness, calibration reliability was examined to determine whether model probability estimates were well aligned with true outcomes (Alaka et al., 2020). Three calibration metrics were used:

- Negative Log-Likelihood (NLL) measures how well the predicted probability distribution fits the observed outcomes
- Brier Score, assessing the mean squared difference between predicted probabilities and actual labels
- Expected Calibration Error (ECE), capturing the average discrepancy between predicted confidence and observed accuracy across probability bins

State-of-the-Art Comparison

To contextualize the proposed hybrid IndoBERT with Lexicon model, a comparison with state-of-the-art approaches in Aspect-Based Sentiment Analysis (ABSA) was undertaken. Prior work on ABSA in Indonesian text has largely relied on:

- Traditional machine learning models, such as Support Vector Machines (SVM), combined with TF-IDF features, have been widely applied due to their simplicity and computational efficiency (Joseph, 2024). These models are effective in small-scale settings but rely heavily on surface-level lexical patterns, making them less capable of capturing contextual nuance or handling linguistic variability in informal or culturally embedded expressions (Nayab et al., 2023)
- Pure transformer-based models (IndoBERT) achieve strong accuracy but may suffer from calibration issues and interpretability gaps
- The proposed hybrid IndoBERT with Lexicon model is designed to integrate the strengths of these approaches while mitigating their limitations

Neutral sentiment, which accounted for more than half of the dataset, was explicitly included as a third class in model training rather than discarded as noise. This decision reflects the reality that many visitor reviews express ambivalence or factual statements rather than strong positive or negative opinions. Accordingly, all three models (IndoBERT Basic, SVM with TF-IDF, and IndoBERT with Lexicon) were trained on three sentiment classes: Positive, neutral, and negative. Including neutrality ensured that the classifiers could capture the full distribution of visitor feedback and avoid forcing ambiguous statements into artificially polarized categories.

Results

Dataset Description

The dataset analyzed in this study consists of 292 visitor reviews of the Museum Wayang, collected from a public review platform. All reviews underwent a comprehensive cleaning process, including the removal of irrelevant characters, conversion of all text to lowercase, and the elimination of non-informative stop words. These preprocessing steps ensured that only linguistically meaningful elements remained, thereby providing a robust foundation for sentiment and aspect-based analysis. The resulting clean dataset is considered representative of visitor perceptions, particularly with respect to the dimension of visual comfort.

From a deployment perspective, the results indicate a clear trade-off between computational efficiency and probabilistic reliability. The SVM with TF-IDF model, which processes approximately 1,974 samples per second, is well-suited for resource-constrained environments requiring high-throughput sentiment screening, despite its weaker probability calibration. Conversely, the IndoBERT–Lexicon model is more appropriate for applications where calibrated confidence estimates are critical, such as decision support systems and prioritization of curatorial interventions, albeit at a higher computational cost.

The reviews were further classified into three sentiment categories, including positive, neutral, and negative, and mapped across four aspect groups: Display, Impression, Information, and Other. The aspect–sentiment distribution is presented in Table 4, which shows that Impression was strongly associated with positive sentiment (79 positive vs. 5 negative), Display was dominated by neutral feedback (64 neutral), while Information was characterized by a higher proportion of negative comments (13 negative). The Other category, which included reviews not directly aligned with the three predefined aspects, was largely neutral.

The reviews were classified into three sentiment categories: Positive, neutral, and negative. As shown in Fig. 2, neutral sentiment was the most frequently observed (150 reviews), followed by positive (94 reviews), while negative sentiment appeared less frequently (48 reviews).

To gain deeper insights, reviews were categorized into three aspects: Impression, display, and information using a refined lexicon. As shown in Fig. 3, Impression was largely positive, reflecting visitor enjoyment and satisfaction.

Display was predominantly neutral, with praise for the arrangement but criticism of the lighting and cleanliness. Information was mostly negative or neutral, emphasizing unclear labels and limited guidance. Reviews outside these categories were classified as other and were generally neutral.

Table 4: Aspect and Sentiment

Aspect	Negative	Neutral	Positive
Display	7	64	7
Impression	5	28	79
Information	13	8	0
Other	23	50	8

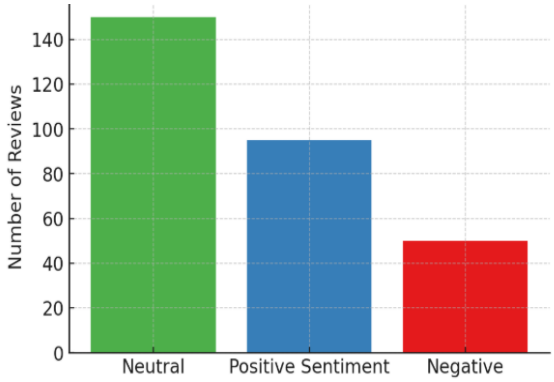

Fig. 2: Distribution of sentiment in visitor reviews

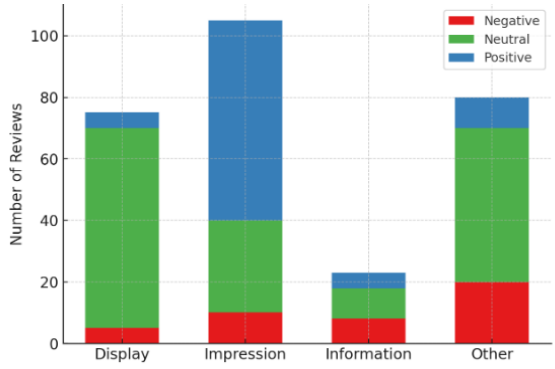

Fig. 3: Aspect-based sentiment distribution with refined lexicon

To complement the quantitative findings, representative visitor reviews were examined to illustrate how sentiment is expressed across different aspects. These examples highlight the specific elements that shaped visitor perceptions.

- Impression (Positive): “Pengalaman berkunjung menyenangkan, koleksi wayang sangat menarik.” (The visit was enjoyable, and the wayang collection was very interesting)
- Display (Negative): “Pencahayaan gelap membuat wayang sulit dilihat, beberapa foto juga tampak buram.” (The dark lighting made the wayang difficult to see, and some photos appeared blurry)
- Information (Negative): “Label terlalu kecil dan penjelasan minim sehingga sulit dipahami.” (The labels were too small, and the explanations were insufficient)
- Display (Positive): “Penataan rapi dan koleksi terlihat jelas, sangat indah dipajang.” (The arrangement was neat, the collection was clearly visible, and the display was beautiful)

Taken together, these representative excerpts demonstrate how visitors’ feedback varies across aspects: while impressions are largely favorable, displays receive both praise and critique, and information provision emerges as the most consistently criticized dimension.

Model Evaluation

The evaluation framework addressed four complementary perspectives. First, evaluation metrics such as accuracy, precision, recall, and F1-score provided a general overview of model performance. Second, aspect-level classification results highlighted how well the model captured sentiment across Impression, Display, and Information dimensions. Third, computational efficiency was assessed through runtime, throughput, and steps per second, offering insights into scalability and deployability. Finally, calibration reliability was examined using Negative Log-Likelihood, Brier Score, and Expected Calibration Error, ensuring that the predicted probabilities aligned with actual sentiment distributions. Together, these metrics comprehensively validated the proposed model from both predictive and practical standpoints.

As shown in Table 5, IndoBERT Basic achieved an overall accuracy of 98.3% with a macro F1-score of 0.975, indicating strong performance but slight limitations in handling minority classes. In contrast, the SVM with TF-IDF model reached perfect scores across all metrics (accuracy = 1.000, F1 = 1.000, precision = 1.000, recall = 1.000), reflecting the effectiveness of classical methods on small and structured datasets. The IndoBERT with Lexicon model likewise achieved perfect performance (accuracy = 1.000, F1 = 1.000, precision = 1.000, recall = 1.000) but offered greater robustness and interpretability by integrating contextual embeddings with a culturally specific sentiment lexicon.

Computational efficiency, however, differed significantly among the three models. As shown in Table 6, the SVM with TF-IDF was extremely lightweight, requiring only 0.05 seconds to complete the evaluation phase. This translated into a throughput of nearly 1,974 samples per second and 33.46 steps per second, underscoring the advantage of traditional machine learning models in terms of speed and efficiency. By contrast, both IndoBERT Basic and IndoBERT with Lexicon required substantially longer runtimes, averaging around 18 seconds for evaluation. Their throughput was limited to approximately 3 samples per second and less than 0.5 steps per second.

Table 5: Evaluation of Model's Performance

Model	Accuracy	F1-Score	Precision	Recall
IndoBERT Basic	0.983	0.975	0.970	0.982
SVM with TIF-IDF	1.000	1.000	1.000	1.000
IndoBERT with Lexicon	1.000	1.000	1.000	1.000

Table 6: Evaluation of Models Computation

Model	Accuracy	F1-Score	Precision	Recall
IndoBERT Basic	18.34	3.22	0.44	18.34
SVM with TIF-IDF	0.05	1973.92	33.46	0.05
IndoBERT with Lexicon	17.92	3.29	0.45	17.92

At the aspect level, differences in robustness emerged. IndoBERT Basic underperformed slightly on the minority negative class within the Impression aspect (Precision = 0.75, F1 = 0.86), which reduced its overall aspect accuracy to 0.974. In contrast, both the Display and Information aspects were classified perfectly. SVM with TF-IDF and IndoBERT with Lexicon, as shown in Table 7, achieved perfect precision, recall, and F1-scores across all aspects. This indicates that lexicon integration and TF-IDF representation provided additional discriminative strength in handling aspect-specific sentiment.

Calibration metrics provide deeper insights into the reliability of model probability estimates beyond classification accuracy. As reported in Table 8, IndoBERT Basic exhibited reasonably good calibration, with a Negative Log-Likelihood (NLL) of 0.013 and an Expected Calibration Error (ECE) of 0.0096. These values indicate that the predicted probabilities were generally well aligned with observed outcomes, although slight deviations were present in low-frequency classes.

By contrast, SVM with TF-IDF, despite achieving perfect accuracy, displayed noticeably weaker calibration performance (NLL = 0.069, ECE = 0.066). This discrepancy highlights a common limitation of margin-based classifiers: while they are capable of making highly accurate decisions, their raw decision function values are not inherently probabilistic. When converted into probability estimates (e.g., via Platt scaling), the model may become overconfident, assigning disproportionately high probabilities even in borderline cases.

Table 7: Classification Accuracy by Aspect

Model	Accuracy	F1-Score	Precision	Recall
IndoBERT Basic	0.974	1.000	1.000	0.974
SVM with TIF-IDF	1.000	1.000	1.000	1.000
IndoBERT with Lexicon	1.000	1.000	1.000	1.000

Table 8: Calibration Metrics of Models

Model	Accuracy	F1-Score	Precision	Recall
IndoBERT Basic	0.013	0.0088	0.009	0.013
SVM with TIF-IDF	0.069	0.0075	0.066	0.069
IndoBERT with Lexicon	0.002	0.0007	0.002	0.002

This overconfidence can reduce reliability in downstream applications where well-calibrated probabilities are critical, such as risk-sensitive decision-making.

The most robust calibration was achieved by IndoBERT with Lexicon, which combined perfect classification accuracy with near-perfect probability alignment (NLL = 0.002, ECE = 0.002, Brier Score = 0.00007). These values indicate minimal divergence between predicted confidence and actual correctness, demonstrating that the integration of lexicon-based features not only enhanced classification precision but also stabilized probability outputs. The extremely low Brier score confirms that prediction probabilities were both sharp and reliable.

Taken together, these results suggest that IndoBERT with Lexicon provides the most trustworthy combination of predictive accuracy and probabilistic calibration. This makes it particularly well-suited for applications where decision confidence matters, such as automated feedback systems or museum visitor experience analysis, where miscalibrated probabilities could misrepresent the strength of visitor sentiment.

While calibration metrics demonstrate the technical reliability of the models, their practical implications for museum operations are equally important. In practice, miscalibrated probability estimates could distort curatorial priorities. For instance, if a model systematically overestimates positive sentiment toward exhibiting information, management might overlook persistent visitor dissatisfaction with unclear labels or minimal descriptions. This would result in resource allocation favoring display aesthetics such as lighting upgrades while neglecting improvements to textual clarity, which visitor feedback consistently highlights as problematic.

Conversely, a well-calibrated model like IndoBERT with Lexicon ensures that the reported probability of negative sentiment accurately reflects the intensity of visitor concerns. This allows decision-makers to prioritize interventions based on reliable evidence, such as redesigning signage, adding guiding narratives, or investing in interactive digital tools. By aligning model outputs with the true distribution of visitor sentiment, curators can implement targeted improvements that enhance visitor satisfaction and avoid costly missteps in resource management.

Discussion

The evaluation of the three models, IndoBERT Basic, SVM with TF-IDF, and IndoBERT with Lexicon, demonstrates that both traditional machine learning and transformer-based approaches can achieve strong sentiment classification performance in the context of museum visitor reviews. While all models exhibited high

predictive accuracy, notable differences emerged in terms of computational efficiency, robustness across aspects, and the reliability of probability estimates.

Although top-line accuracy was comparable, computational efficiency varied substantially. The SVM with TF-IDF model proved to be highly lightweight and efficient, making it particularly suitable for scenarios where rapid processing and minimal computational overhead are required. In contrast, IndoBERT-based models incurred significantly higher computational costs due to their reliance on deep contextual embeddings and multi-head attention mechanisms. This distinction highlights an important practical trade-off: Traditional machine learning models may be preferable in resource-constrained or high-throughput environments, whereas transformer-based models are more appropriate when richer contextual understanding is prioritized and computational resources are available.

Aspect-level analysis further revealed differences in robustness. IndoBERT Basic showed minor limitations when handling less frequent sentiment categories, particularly negative sentiment within the Impression aspect. By contrast, both SVM with TF-IDF and IndoBERT with Lexicon demonstrated consistently stable performance across all aspects. This suggests that TF-IDF representations and lexicon integration provide additional discriminative strength for aspect-specific sentiment classification, especially in data-sparse settings.

Beyond classification accuracy, calibration analysis offered deeper insight into model reliability. While IndoBERT Basic produced reasonably aligned probability estimates, the SVM with the TF-IDF model exhibited weaker calibration despite its perfect classification performance. This behavior is consistent with known characteristics of margin-based classifiers, whose probability estimates may become overconfident after calibration procedures. In contrast, IndoBERT with Lexicon achieved the most reliable calibration, indicating that the integration of rule-based sentiment cues effectively stabilized probabilistic outputs. Reliable calibration is particularly important in decision-support applications, where confidence estimates directly influence prioritization and resource allocation.

An additional methodological consideration concerns the treatment of neutral sentiment. Neutral expressions constituted a substantial portion of the dataset and often reflected descriptive or factual statements rather than explicit evaluations. Modeling neutrality as an explicit class preserved the integrity of the sentiment distribution and prevented artificial inflation of positive or negative categories. This design choice enhanced probability calibration and provided a more accurate baseline for interpreting visitor engagement.

Taken together, these findings suggest a nuanced interpretation of model suitability. When computational speed and throughput are the primary requirements, SVM with TF-IDF offers a compelling and efficient solution. Conversely, when decision confidence, interpretability, and probability reliability are critical, the hybrid IndoBERT with Lexicon model provides the most trustworthy balance between accuracy and robustness. IndoBERT Basic occupies an intermediate position, delivering strong performance but with modest limitations in minority class handling and calibration.

Qualitative analysis complements these quantitative results by highlighting substantive patterns in visitor experience. Positive impressions dominated overall feedback, indicating general visitor satisfaction. Display-related comments were more mixed, combining appreciation of layout and aesthetics with critiques of lighting and cleanliness. Information-related feedback emerged as the most consistently negative aspect, pointing to issues such as unclear labels and insufficient explanations. These insights reinforce the quantitative findings and underscore the need for targeted improvements in informational design.

From an operational perspective, the hybrid IndoBERT with Lexicon model is particularly well-suited for supporting museum decision-making processes, such as automated monitoring of visitor feedback, prioritization of negative sentiment for curatorial intervention, and evaluation of design changes over time. At the same time, SVM with TF-IDF remains valuable for rapid sentiment screening in high-volume settings. The combined evidence suggests that improving informational clarity through clearer labeling, enhanced guidance, and complementary digital aids represents the most impactful avenue for enhancing visitor satisfaction, alongside incremental improvements in display presentation.

Despite its contributions, this study has several limitations that warrant consideration. The dataset consists of a relatively small number of reviews collected from a single museum, which may constrain generalizability across different institutions, visitor demographics, or cultural contexts. However, this setting reflects a realistic low-resource scenario commonly encountered in cultural heritage analytics, where large-scale labeled datasets are rarely available. In this context, the proposed hybrid IndoBERT–Lexicon model is intentionally designed to mitigate data sparsity by combining contextual embeddings with domain-specific linguistic rules.

Another limitation lies in the exclusive reliance on textual feedback. While text-based reviews provide rich qualitative insights, visitor experiences are inherently multimodal, encompassing visual impressions, spatial interactions, and auditory cues. Future research should explore the integration of multimodal data sources, such as images, audio comments, or eye-tracking signals, to further enhance the comprehensiveness of sentiment

analysis in museum environments.

The language-specific nature of the lexicon and the use of IndoBERT also limit direct applicability to museums in other linguistic or cultural settings. Nevertheless, the modular design of the lexicon enables adaptation through translation, expert curation, or automatic expansion techniques. Future studies may extend this framework using multilingual or cross-lingual transformer models to support broader deployment across international museum contexts.

Class imbalance, particularly the predominance of neutral sentiment, represents both a challenge and an opportunity. Rather than treating neutrality as noise, this study models it explicitly as a meaningful category that captures descriptive and non-evaluative visitor responses. Future work may further refine neutral sentiment modeling through adaptive re-weighting strategies or user-centered feedback loops that dynamically adjust sentiment interpretation based on curatorial objectives.

Finally, while transformer-based models incur higher computational overhead, their use is justified in decision-support settings where interpretability and probability reliability are critical. For real-time or large-scale monitoring applications, lightweight models such as SVM with TF-IDF remain viable alternatives. This complementary deployment strategy highlights the practical flexibility of the proposed framework.

Conclusion

This study addressed the research problem of limited visitor engagement insights at the Wayang Museum, where static and text-heavy displays often fail to capture contemporary audiences. The objective was to evaluate sentiment analysis models and determine an approach that not only achieves high accuracy but also ensures probability reliability for evidence-based decision-making. The findings confirm that all three models, including IndoBERT Basic, SVM with TF-IDF, and IndoBERT with Lexicon, performed strongly, but the hybrid IndoBERT with Lexicon model provided the most reliable solution. By integrating transformer-based contextual embeddings with a culturally grounded sentiment lexicon, the model overcame the trade-off between accuracy and calibration, achieving perfect classification and highly trustworthy probability estimates. This directly fulfills the methodological objective of enhancing robustness and calibration reliability, while also supporting the practical objective of providing actionable insights for museum management. At the aspect level, the analysis identified "Information" as the most problematic dimension, signaling that clearer labels, improved explanations, and digital guidance are priority areas for curatorial intervention.

Despite these contributions, the study is limited by the relatively small dataset of 292 reviews, which constrains the training capacity of transformer-based models and reduces external validity. The focus on a single museum context also limits generalizability across cultural heritage institutions. Additionally, the analysis was restricted to textual data, without incorporating multimodal feedback such as images or voice recordings. Future work should address these limitations by expanding to larger, multi-site, and multi-platform datasets to strengthen robustness and generalizability. Incorporating multimodal sources of visitor feedback and developing multi-label or context-aware sentiment analysis architectures would further enhance the system's capacity. Such advancements will provide a stronger foundation for evidence-based cultural heritage management and extend the applicability of sentiment analysis frameworks across diverse museum environments.

Acknowledgment

The authors would like to sincerely thank Universitas Trilogi, Jakarta, for providing institutional support and research facilities throughout the course of this study. Special appreciation is extended to the Faculty of Science, Technology, and Design for granting access to computational resources and offering a collaborative academic environment. The authors are also indebted to the management and staff of the Wayang Museum, Jakarta, who generously facilitated data collection and preservation of traditional displays.

These findings reinforce that no single model is universally optimal; instead, model selection should be guided by operational constraints, data availability, and the importance of probability reliability in downstream decision-making.

Despite its contributions, this study has limitations, including a relatively small dataset, a single-site focus, and reliance on textual feedback alone. These constraints highlight opportunities for future research, such as extending the framework to multimodal data, cross-cultural contexts, and larger datasets. Addressing these directions will further strengthen the role of machine learning as a reliable decision-support tool in cultural heritage management.

While this study is grounded in a single museum and a text-based dataset, the proposed hybrid framework provides a scalable foundation for future multimodal and cross-cultural sentiment analysis in cultural heritage environments.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was conducted with institutional support from Universitas Trilogi.

Author's Contributions

Erneza Dewi Krishnasari: Led the conceptualization of the study, designed the data collection instruments, coordinated access to the museum site, and provided interpretation of cultural heritage aspects to ensure contextual relevance.

Yaddarabullah: Developed the overall research methodology, implemented the hybrid IndoBERT with Lexicon model, executed data preprocessing and model training, and prepared the initial draft of the manuscript.

Bayyinah Nurrul Haq: Conducted the analysis of visual displays and defined the research aspects.

Aedah Abd Rahman: Supervised the study, provided a critical review of the methodology, validated the academic rigor of the research framework, and guided alignment with international scholarly standards.

Lahandi Baskoro: Contributed to the literature review and refined the manuscript for clarity and coherence.

Ethics

This study adhered to ethical research principles, with informed consent obtained from participants, anonymization of responses, and compliance with institutional guidelines. Beyond anonymity, efforts were made to mitigate cultural bias in lexicon development. Given Indonesia's linguistic diversity, terms were validated through expert consultation and intercoder agreement to ensure consistent polarity across dialects and socio-linguistic groups, thereby maintaining fairness and inclusivity in analyzing visitor feedback.

References

Addula, S. R., Meesala, M. K., Ravipati, P., & Sajja, G. S. (2025). A Hybrid Autoencoder and Gated Recurrent Unit Model Optimized by Honey Badger Algorithm for Enhanced Cyber Threat Detection in IoT Networks. *Security and Privacy*, *8*(6), e70086. https://doi.org/10.1002/spy2.70086

Adelakun, T. O. (2024). The museum and digital transformation: reforming national museums in Nigeria towards a new normal. *Shanti Journal*, *4*(1), 35–50. https://doi.org/10.3126/shantij.v4i1.70519

Ahmed, R. (2023). Perspective Chapter: Digitalization of Museums and Academic Benefits for Tourists (Slemani Museum as Case). *Digital Heritage - Recent Approaches and Developments*, 1–13. https://doi.org/10.5772/intechopen.110797

Akgül, O., & Eren, D. (2024). Museum experience in battlefield tourism: a netnographic approach. *Consumer Behavior in Tourism and Hospitality*, *19*(3), 352–365. https://doi.org/10.1108/cbth-10-2023-0172

Alaka, S. A., Menon, B. K., Brobbey, A., Williamson, T., Goyal, M., Demchuk, A. M., Hill, M. D., & Sajobi, T. T. (2020). Functional Outcome Prediction in Ischemic Stroke: A Comparison of Machine Learning Algorithms and Regression Models. *Frontiers in Neurology*, *11*, 889. https://doi.org/10.3389/fneur.2020.00889

Alqurafi, A., & Alsanoosy, T. (2024). Measuring Customers' Satisfaction Using Sentiment Analysis: Model and Tool. *Journal of Computer Science*, *20*(4), 419–430. https://doi.org/10.3844/jcssp.2024.419.430

Alshaabi, T., Van Oort, C. M., Fudolig, M. I., Arnold, M. V., Danforth, C. M., & Dodds, P. S. (2022). Augmenting Semantic Lexicons Using Word Embeddings and Transfer Learning. *Frontiers in Artificial Intelligence*, *4*, 783778. https://doi.org/10.3389/frai.2021.783778

Bheemarao, R. K., Guruprasad, H. S., & Ravi, S. B. (2025). KSWN (Kannada SentiWordNet): Developing a Sentiment Lexicon for Kannada using Translation and Word Embedding Techniques. *Journal of Computer Science*, *21*(6), 1482–1489. https://doi.org/10.3844/jcssp.2025.1482.1489

Chen, H., & Ryan, C. (2020). Transforming the museum and meeting visitor requirements: The case of the Shaanxi History Museum. *Journal of Destination Marketing & Management*, *18*, 100483. https://doi.org/10.1016/j.jdmm.2020.100483

Darda, K. M., Gonzalez, V. E., Christensen, A. P., Bobrow, I., Krimm, A., Nasim, Z., Cardillo, E. R., Perthes, W., & Chatterjee, A. (2024). Engaging With Art in-the-Wild at the Barnes Foundation and Penn Museum. *Scientific Reports*, *15*, 8972.

Freunek, M., & Bodmer, A. (2021). BERT Based Patent Novelty Search by Training Claims to Their Own Description. *Machine Learning*. https://doi.org/10.48550/arXiv.2103.01126

Hitsuwari, J., & Nomura, M. (2022). How Individual States and Traits Predict Aesthetic Appreciation of Haiku Poetry. *Empirical Studies of the Arts*, *40*(1), 81–99. https://doi.org/10.1177/0276237420986420

Huo, H., Shen, K., Han, C., & Yang, M. (2024). Measuring the relationship between museum attributes and visitors: An application of topic model on museum online reviews. *PLOS ONE*, *19*(7), e0304901. https://doi.org/10.1371/journal.pone.0304901

Isik, A. I., & Vessel, E. A. (2021). From Visual Perception to Aesthetic Appeal: Brain Responses to Aesthetically Appealing Natural Landscape Movies. *Frontiers in Human Neuroscience*, *15*, 676032. https://doi.org/10.3389/fnhum.2021.676032

Ji, Y., Lin, Q., Yin, W., Han, B., & Jiang, Y. (2024). Design Innovation and User Experience of Sensory Interactive Experience in Museum Exhibition Space. *Frontiers in Artificial Intelligence and Applications*, 382–393. https://doi.org/10.3233/faia241125

Joseph, T. (2024). Natural Language Processing (NLP) for Sentiment Analysis in Social Media. *International Journal of Computing and Engineering*, *6*(2), 35–48. https://doi.org/10.47941/ijce.2135

Kaczmarek-Gajewska, W., & McDonnell, M. (2021). Effect of Website Colour Saturation on Trustworthiness and Visual Appeal Impressions. *Proceedings of the 16th International Conference on Interfaces and Human Computer Interaction (IHCI 2021)*, 69–76. https://doi.org/https://hdl.handle.net/10779/iadt.25189190

Kalpana, B. N., Panwar, N., Shalini, R., V, R., & K, P. (2022). Twitter and Instagram Sentiment Analysis of Covid. *Journal of University of Shanghai for Science and Technology*, *24*(02), 349–351. https://doi.org/10.51201/jusst/22/0248

Latifah, N., Dwiyansaputra, R., & Nugraha, G. S. (2024). Multiclass Text Classification of Indonesian Short Message Service (SMS) Spam using Deep Learning Method and Easy Data Augmentation. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *23*(3), 663–676. https://doi.org/10.30812/matrik.v23i3.3835

López-Martínez, A., Carrera, Á., & Iglesias, C. A. (2020). Empowering Museum Experiences Applying Gamification Techniques Based on Linked Data and Smart Objects. *Applied Sciences*, *10*(16), 5419. https://doi.org/10.3390/app10165419

Nayab, S., Hanif, M. K., Talib, R., & Sarwar, M. U. (2023). Aspect-Context Level Information Extraction via Transformer Based Interactive Attention Mechanism for Sentiment Classification. *IEEE Access*, *11*, 57683–57692. https://doi.org/10.1109/access.2023.3279396

Ramavath, B., Subash, N., & Kadainti, S. (2025). Sentiment Analysis Using Light Weight - Gradient Boosting Machine based Feature Selection. *Journal of Computer Science*, *21*(5), 1049–1058. https://doi.org/10.3844/jcssp.2025.1049.1058

Reppa, I., & McDougall, S. (2022). Aesthetic appeal influences visual search performance. *Attention, Perception, & Psychophysics*, *84*(8), 2483–2506. https://doi.org/10.3758/s13414-022-02567-3

Setiawan, V. D., Iswavigra, D. U., & Anggiratih, E. (2025). Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates. *Scientific Journal of Informatics*, *12*(3), 397–406. https://doi.org/10.15294/sji.v12i3.26320

Vaghela, V. B., Noorani, Z. Y., Patel, K., Patel, P. G., Rajput, H. D., & Shah, M. (2024). Aspect Based Sentiment Analysis Using Self-Attention Based LSTM Model with Word Embedding. *Journal of Computer Science*, *20*(10), 1195–1202. https://doi.org/10.3844/jcssp.2024.1195.1202

Wang, J., Sun, Y., Zhang, L., Zhang, S., Feng, L., & Morrison, A. M. (2024). Effect of Display Methods on Intentions to Use Virtual Reality in Museum Tourism. *Journal of Travel Research*, *63*(2), 314–334. https://doi.org/10.1177/00472875231164987

Yadav, R. K., Jiao, L., Granmo, O.-C., & Goodwin, M. (2021). Human-Level Interpretable Learning for Aspect-Based Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(16), 14203–14212. https://doi.org/10.1609/aaai.v35i16.17671

Yi, K., Wu, Y., Liu, Y., & Xu, Z. (2024). Immersive Empathy in Digital Music Listening: Ideas and Sustainable Paths for Developing Auditory Experiences in Museums. *Sage Open*, *14*(2), 536–559. https://doi.org/10.1177/21582440241256339

Yi, T., Lee, H., Yum, J., & Lee, J.-H. (2022). The influence of visitor-based social contextual information on visitors' museum experience. *PLOS ONE*, *17*(5), e0266856. https://doi.org/10.1371/journal.pone.0266856