



Visitor Perceptions of Post-Fire Museum X in Online Reviews: Sentiment Analysis and Topic Modelling

Muhammad Fadli Fakhirin¹, Rahmi², Muhamad Prabu Wibowo³

^{1,2,3} Universitas Indonesia, Depok, Indonesia

Corresponding author email: fadli.fakhirin@gmail.com

Info Article

Received: 27 July 2026

Revised: 3 September 2026

Accepted: 10 September 2026

Online Version: 21 September 2026

Abstract

In the digital era, online reviews have increasingly become a valuable source of user-generated language data for understanding users' experiences, evaluations, and opinions. This study examines public perceptions of Museum X after its post-fire reopening through Google Maps online reviews as the data source. The corpus comprised 1,822 Indonesian-language reviews published from 15 October 2024–1 June 2026. The study combines lexicon-based sentiment analysis using the Indonesian Sentiment Lexicon (InSet) and BERTopic topic modelling to identify lexical polarity and dominant museum visitor experience issues. The results show that positive lexical sentiment was the largest category (46.54%), followed by negative (32.38%), no coverage (17.29%), and neutral (3.79%). The mean rating was high at 4.69 out of 5. A manual validation on a stratified sample ($n = 249$) yields 81.5% agreement ($\kappa = 0.72$) and confirms that no-coverage reviews are predominantly sentiment-bearing. BERTopic produced five regular topics and an outlier cluster, with the dominant topic concerning ticketing, entry, services, collections, and price (820 reviews), while the other topics captured historical-cultural collections, general positive impressions and comfort, family education, and history learning. Combining lexical sentiment, rating, and topic context shows that Museum X's educational and historical value that visitors express in their reviews remains prominent, while ticketing and entry management require attention. The study demonstrates a transparent, coverage-aware workflow for service evaluation from unsolicited online feedback.

Keywords: Online reviews; Lexicon-based sentiment analysis; Topic modelling; BERTopic; Museum visitor experience.

This is open access article under the [CC-BY](#) licence



INTRODUCTION

In the digital era, online reviews have become an important source of unsolicited feedback through which the public describes experiences and advises prospective users. Unlike structured questionnaires, reviews are produced without predetermined response categories and may combine narrative description, emotional reaction, and numerical ratings (Alexander et al., 2018; Packer & Ballantyne, 2016). For cultural institutions, this combination provides a valuable basis for understanding what visitors appreciate, what problems they encounter, and which aspects of an experience require attention. Online reviews can therefore bridge the gap between institutional assumptions and genuine visitor concerns, surfacing operational issues that could otherwise remain unnoticed (Drivas et al., 2025; King et al., 2023).

For museums, such feedback matters both for routine service development and after periods of closure caused by disruptive events. Institutional recovery is usually tracked through indicators such as

visitor numbers or budgets, while the returning public's experience of renewed exhibitions, services, and facilities is seldom examined on a comparable scale, even though destination perceptions can shift across successive phases of disruption (Liu et al., 2024). The museum examined here, referred to as Museum X, is one of several museums in Indonesia. Its Google Maps online reviews were authored in the period following a post-fire reopening. This study does not model the recovery process itself. Rather, it treats these reviews as a naturally occurring record from which to profile visitor sentiment and the themes visitors raise, with the reopening period serving as the corpus's temporal setting.

The use of user-generated data in museum management has moved from reading individual comments toward large-scale computational analysis. Riva & Agostino (2022) identified dimensions of cultural heritage, personal experience, and services in TripAdvisor reviews, and Huo et al. (2024) linked museum attributes to visitor experience and satisfaction through topic modelling. More recent work has paired sentiment with topic modelling directly, such as Saoualih et al. (2025), who used sentiment-guided BERTopic to surface experiential gaps in heritage museums, and Candrea et al. (2025), who combined sentiment analysis and Latent Dirichlet Allocation (LDA) topic modelling to compare visitor groups. Across these studies, appreciation and complaint routinely co-occur within the same body of reviews, yet how a museum's overall rating and its lexical negativity diverge is rarely made explicit.

Sentiment analysis and topic modelling are complementary tools for this purpose. Sentiment analysis detects and extracts subjective information from text and assigns a polarity that summarises how visitors evaluate what they describe (Wankhade et al., 2022). Topic modelling, by contrast, uncovers latent themes that recur across a corpus by identifying groups of words that co-occur, thereby revealing what visitors discuss (Ogunleye et al., 2025). Polarity alone indicates how favourably visitors respond, not what they respond to. Hence, topic modelling supplies the missing subject matter.

This study adopts a lexicon-based approach to sentiment analysis because polarity is derived transparently by matching words with specified weights (Taboada et al., 2011). In the Indonesian context, the Indonesian Sentiment Lexicon (InSet) is used, which provides 3,609 positive and 6,609 negative entries, weighted from -5 to $+5$ (Koto & Rahmanningtyas, 2017). However, lexicon-based scoring is domain-dependent because general-purpose dictionaries exhibit limited cross-domain validity and questionable domain independence (Chan et al., 2021), and terms outside the lexicon cannot contribute to the score at all (Wankhade et al., 2022). Because InSet was compiled from microblog language, its coverage of museum-review vocabulary is not guaranteed. Therefore, a score of zero can arise from two distinct situations: when positive and negative weights cancel each other out, or when no lexicon entry matches the text. The first is a genuine neutral, while the second reflects a limit of dictionary coverage. Since there is no objectively neutral text and the zero point must itself be read against a baseline rather than taken at face value (van der Veen & Bleich, 2025), conflating these two cases would let coverage gaps masquerade as neutrality. Separating them yields a coverage-aware labelling that makes explicit how much of the sentiment distribution rests on actual lexicon matches.

Topic modelling here uses the BERTopic approach. BERTopic forms document representations using a transformer-based language model, clusters those representations, and then produces representative words through class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) (Grootendorst, 2022). An embedding-based approach suits short reviews whose wording varies, but its output still requires evaluation of coherence, diversity, membership probability, and the proportion of outliers. A high outlier ratio in particular leaves many reviews unassigned, so the resulting topics describe only the reviews that cluster cleanly rather than the corpus as a whole.

Prior Indonesian studies have analysed reviews of tourism and heritage sites using sentiment analysis and topic modelling. Suprina et al. (2025) applied opinion mining and text-network analysis to Google Maps and TripAdvisor reviews of Jakarta attractions, including the National Monument and Kota Tua, and classified reviews as positive, negative, or neutral. Tanebeth (2026) compared the performance of sentiment analysis using machine-learning and transformer models on 1,110 Google Maps reviews of the National Monument. The study most closely related to the present research methodologically is Asyiah & Aktavia (2025), who combined BERTopic with Indonesian lexicon-based sentiment analysis of Google Maps reviews, scoring each review by summing the positive and negative lexicon matches and labelling a

score of zero as neutral. Their study, however, concerns reviews of a university rather than a museum.

In these and comparable studies, reviews containing no lexicon-matched terms are either assigned a class by a supervised model or, in lexicon-based approaches, assigned to the neutral category. The present study departs from this practice by treating lexicon non-coverage as a distinct outcome, separate from substantive neutrality. This distinction is consequential, as the validation reported here shows that the majority of no-coverage reviews are sentiment-bearing to a human reader. This coverage-aware treatment has not, to the authors' knowledge, been applied to Indonesian museum reviews.

Methodologically, the study contributes a transparent, coverage-aware workflow for lexicon-based sentiment analysis of naturalistic, informal Indonesian text. Rather than treating an unmatched review as neutral, it distinguishes genuine neutrality from a lack of lexical coverage, turning the gap between a lexicon and a new register into a measurable quantity rather than a hidden assumption. This coverage-aware procedure is transferable to other lexicon-based analyses of short, informal evaluative text, where the same distinction is needed for a valid reading of sentiment.

This study tackles three questions in sequence. First, how are lexical sentiment and InSet coverage distributed across visitor reviews of Museum X? Second, what themes are dominant across these reviews? Third, how does sentiment vary within each theme, and what does this imply for service priorities? The corresponding objectives are to describe the sentiment and coverage profile of the reviews, to identify their thematic structure, with the topic model's coherence, diversity, and outlier ratio reported as validation metrics rather than as findings in their own right, and to derive service-evaluation implications from the Google Maps online reviews. Practically, the results can help the museum distinguish educational and experiential strengths from operational problems that warrant follow-up.

RESEARCH METHODS

This study employed a computational text-mining approach to examine visitor perceptions expressed in Indonesian-language Google Maps online reviews of Museum X. The data consist of publicly posted reviews, which were analysed along two subsequently integrated paths: lexicon-based sentiment analysis using InSet and topic modelling using BERTopic. The following subsections present the research design, the time and place of research, the research target, the research procedure, the instruments and data collection techniques, and the data analysis techniques used in this study.

Research Design

This study used a descriptive computational text-mining design, with secondary data in the form of online reviews. This design was chosen in order to map the distribution of lexical polarity and topic structure rather than to test cause-and-effect relationships. The unit of analysis is the Google Maps online reviews dated from 15 October 2024 to 1 June 2026. The start date follows the reopening of Museum X to the public after the fire. The analysis integrated lexicon-based sentiment analysis using InSet, BERTopic topic modelling, and ratings used as an explicit comparative indicator of the text.

Research Target/Subject

The research target is the perceptions expressed in Google Maps reviews by visitors after the museum reopened. The data were collected by extracting the reviews from the Google Maps page using Apify on 11 July 2026. The Apify actor used for extracting the online reviews is Google Maps Reviews Scraper (version 0.0.515), which yielded 5,226 reviews. From these extraction results, all records meeting the selection criteria were retained: original-language metadata with the value 'id', non-empty review text, a date within the study period, and a non-duplicate combination of raw text, date, and rating. However, the original-language metadata is imperfect, as some Indonesian-language reviews may be misclassified. Such reviews were therefore excluded by the language filter, meaning that the analysed corpus represents a conservative subset of the available reviews and may under-represent code-mixed or otherwise misclassified Indonesian-language reviews.

Reviewer names, profile links, and account identifiers were not used in the analysis or in reporting. The institution's name was pseudonymised as Museum X, and representative reviews are not quoted verbatim to reduce traceability. This practice follows a contextual approach to internet research ethics that weighs privacy expectations, risk, and the potential for re-identification rather than the public accessibility of the data alone (Panel on Research Ethics, 2022).

Research Procedure

The research procedure was carried out sequentially. First, the data were read from the extraction spreadsheet, relevant column names were identified, dates were parsed, empty text was removed, language was selected, and duplicates were deleted. Second, the review text was normalised and prepared in two copies: `text_for_sentiment` for lexicon matching and `text_for_bertopic` for topic modelling. Third, InSet was loaded without morphological normalisation, and its words and phrases were matched against the reviews' n-grams. Fourth, sentiment scores were computed using negation, diminisher, and coverage rules. Fifth, BERTopic formed the clusters and the word representations. Finally, topic labels, ratings, coverage, and sentiment were aggregated in order to answer the three research questions.

All processing was carried out in the researcher's computing environment, using R as the primary software environment and Python via the reticulate package. The R packages used were `readxl`, `dplyr`, `tidyr`, `tidytext`, `stringr`, `stringi`, `ggplot2`, `reticulate`, `purrr`, `lubridate`, `readr`, `stopwords`, `forcats`, and `tibble`. BERTopic was implemented in Python using the packages `bertopic`, `sentence-transformers`, `umap-learn`, `hdbscan`, `scikit-learn`, `gensim`, `pandas`, `numpy`, `plotly`, `kaleido`, `openpyxl`, and `gensim`.

Instruments and Data Collection Techniques

The data instruments consist of the Google Maps online review spreadsheet, two InSet lexicon files for positive and negative entries, and an R script for the analysis. The variables used include review text, date, rating, language, sentiment score, number of lexicon matches, coverage status, sentiment label, topic number, topic probability, and representative words.

All extracted reviews were preprocessed. The review text was converted to UTF-8, the museum name was pseudonymised, characters were transliterated to Latin-ASCII, selected emojis were mapped to emotion tokens, letters were converted to lowercase, URLs and mentions were cleaned, hashtags were retained as words, non-standard forms were mapped through a review normalisation dictionary, excessive letter repetition was simplified, and non-alphabetic characters were removed. Examples of the mapping include 'gak' to 'tidak', 'bgt' to 'banget', 'antri' to 'antre', and variants of 'musholla' to 'mushola'. Negators were retained on the sentiment path but removed from the BERTopic path together with stopwords so that they would not become topic markers. The preprocessing yielded 1,822 reviews for the sentiment analysis. One document became empty after topic-specific preprocessing, so BERTopic used 1,821 documents.

Table 1. Data Selection and Filtering Stages from Extraction to Final Corpus

Stage	n retained	n removed
Extracted from Google Maps with Apify	5,226	0
Removed text-free reviews	2,507	2,719
Removed non-indonesian reviews	1,822	685
Removed after preprocessing	1,822	0
Removed after topic modelling preprocessing	1,821	1

Data Analysis Technique

A total of 1,822 reviews were used in the sentiment analysis. Sentiment was calculated with the positive and negative InSet lists. Entries that were exactly identical within the same polarity were deduplicated. Where a single key appeared in both polarities, the entry was removed in order to avoid ambiguity. Function words and single-token stopwords were pruned, including 'kurang' because that word is processed as an operator. Root words and derived words available in InSet therefore remain separate

entries with their original weights. This decision prevents artificial merging and the new redundancy that could arise if the entire dictionary were reduced to a single form.

Reviews were tokenised and formed into n-grams up to the maximum length of an InSet phrase, capped at four tokens. Matching uses a longest-phrase, non-overlapping strategy in which longer phrases are selected before single tokens that occupy the same position. The collocation 'salah satu' is protected so that the word 'salah' is not automatically counted as negative. The negators 'tidak', 'tak', 'bukan', and 'belum' occurring one or two tokens earlier invert the weight of an entry. The operator 'kurang' at the same distance turns a positive weight into a negative one. These limited contextual rules are consistent with the principle that lexicon-based analysis must take account of the polarity shifts caused by negation (Taboada et al., 2011).

The sentiment score is defined as the sum of the weights of all matches after adjustment. Where at least one match exists, a sentiment score greater than 0 is labelled positive, a sentiment score less than 0 is labelled negative, and a sentiment score equal to 0 is labelled neutral. Documents without a match are labelled 'no coverage' rather than neutral. Sentiment density is computed as the score divided by the number of matches for covered documents.

To assess the automated sentiment labels, one author manually coded a stratified random sample of 249 reviews (60 positive, 60 negative, 60 no-coverage, and all 69 labelled neutral by the classifier) into positive, negative, or neutral, blind to the classifier output and the star rating. Because "no coverage" reflects the absence of lexicon matches rather than a perceptible sentiment, human coding was restricted to polarity. The no-coverage set is analysed by comparing the classifier's abstention with the human polarity of those reviews. Agreement on covered reviews is reported as overall accuracy, per-class precision, recall, and F1, and Cohen's κ . Sampling was stratified by classifier label to ensure sufficient minority-class cases, so per-class precision is representative of each predicted class while recall is conditional on the validation sample.

For the topic modelling, the topic text was processed using multilingual BERTopic with `calculate_probabilities = TRUE`. On the basis of 1,821 documents, `min_topic_size` and `min_cluster_size` were set adaptively at $\text{floor}(1,821 \times 0.02) = 36$. UMAP used `n_neighbors = 15`, `n_components = 5`, `min_dist = 0`, `metric = cosine`, and `random_state = 42`. HDBSCAN used `metric = euclidean`, `cluster_selection_method = eom`, and `prediction_data = TRUE`. CountVectorizer used unigrams (1,1), `min_df = 1`, `max_df = 1.0`, and a combined stopwords list of Indonesian, English, and generic terms. BERTopic then formed the topic representations with c-TF-IDF.

Model quality was assessed through the number of topics, the outlier ratio, the mean topic probability, `c_v` coherence, `c_npmi` coherence, and the diversity of the top ten words. Coherence was used to assess the relatedness of the words within a topic (Röder et al., 2015), while diversity was computed as the proportion of unique words among all top words. Using both avoids assessing the model based on a single measure (Dieng et al., 2020). These values are treated as comparative diagnostics and are read alongside keywords, cluster size, and probability rather than as absolute thresholds of quality.

Each document was assigned a dominant topic and a membership probability. Sentiment was then aggregated by topic in order to compute the number of covered reviews, the mean rating, the mean sentiment score for covered documents, and the proportions of positive, neutral, negative, and no-coverage reviews. Topic -1 was treated as a heterogeneous outlier group rather than as a substantive topic. Thematic names were assigned analytically from the top words and the corpus context, without altering the numerical labels produced by the model.

RESULTS AND DISCUSSION

Corpus Characteristics

The final corpus consists of 1,822 Indonesian-language reviews. All of the reviews carry a rating, with a mean of 4.69 out of 5, and 1,489 reviews (81.72%) giving five stars. The dominance of high ratings indicates a very positive explicit evaluation of the visit. One document became empty after negators and

topic stopwords were removed, so 1,821 documents entered BERTopic. This difference of one document explains why the denominators for the sentiment analysis and the topic modelling are not the same.

Table 2. Characteristics of the Research Corpus

Indicator	Value
Review period	15 October 2024–1 June 2026
Reviews for sentiment analysis	1,822
Reviews with InSet coverage	1,507 (82.71%)
Documents for BERTopic	1,821
Mean rating	4.69 out of 5
Five-star ratings	1,489 (81.72%)

In general, the findings display two layers of evaluation. At the explicit layer, ratings are very high, while at the textual layer, the reviews contain a mixture of appreciation, criticism, and expressions that InSet cannot assess. Separating these layers matters because even a five-star review may contain notes about tickets, price, or access, while short reviews such as certain expressions of praise may have no dictionary match at all. The following results answer the research questions while maintaining this distinction between indicators.

The Distribution of Lexical Sentiment

Across all 1,822 reviews, the positive category is the largest group, with 848 reviews (46.54%). The negative category covers 590 reviews (32.38%), followed by the no coverage category with 315 reviews (17.29%), and the neutral category with 69 reviews (3.79%). Because 315 reviews have no InSet match at all, the neutral proportion must not be combined with no coverage. Among the 1,507 covered reviews, the composition becomes 56.27% positive, 39.15% negative, and 4.58% neutral. These results indicate that textual perception tends to be positive, but that roughly one in six reviews lies beyond the reach of the dictionary.

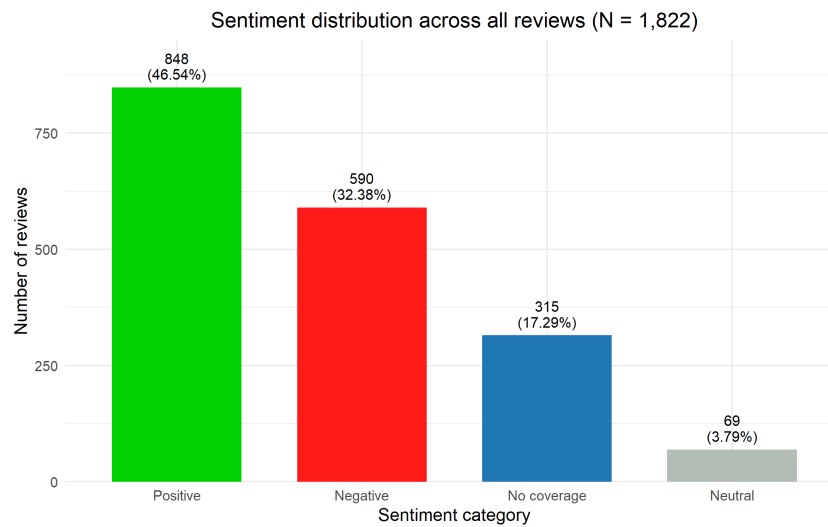


Figure 1. Sentiment distribution across all reviews.

Table 3. Sentiment Distribution and InSet Coverage

Category	n	% of all reviews	% of covered reviews
Positive	848	46.54	56.27
Neutral	69	3.79	4.58
Negative	590	32.38	39.15
No coverage	315	17.29	—

Category	n	% of all reviews	% of covered reviews
Total	1,822	100.00	100.00

On the 189 covered reviews from the stratified sampling, the classifier agreed with the human coder in 81.5% of cases (Cohen's $\kappa = 0.72$, substantial agreement). Per-class performance was high for positive (precision 0.95, recall 0.77) and neutral (precision 0.90, recall 0.86), but lower in precision for negative (0.58, recall 0.81). The lexicon assigns a negative label to some holistically positive reviews that contain negative terms, consistent with the gap between summed lexical polarity and overall evaluation. Two results bear directly on the design. First, among the sampled no-coverage reviews, 78.3% carried sentiment to the human coder (65.0% positive, 13.3% negative), confirming that no-coverage cannot be equated with neutrality. Second, of the sampled five-star reviews the classifier labelled negative, 68.8% were also read as non-positive by the human coder (50.0% negative, 18.8% neutral), indicating that the five-star/negative divergence is substantially genuine rather than an artefact of the method.

Table 4. Human vs. Classifier Agreement, Covered Reviews (n = 189)

Human (rows) / Classifier (columns)	Positive	Neutral	Negative	Total
Positive	57	2	15	74
Neutral	0	62	10	72
Negative	3	5	35	43
Total	60	69	60	189

The cross-tabulation of rating and sentiment in Table 4 makes it clear that the lexical category is not a substitute for the rating. Among the 1,489 five-star reviews, 716 were classified as positive, 437 as negative, 59 as neutral, and 277 as no coverage. In other words, 29.35% of the five-star reviews still carry a negative lexical score. This mismatch does not automatically indicate an error because a single review may give a high overall assessment while mentioning a particular problem, and the InSet negative weights may be stronger than the positive weights within the same text. This divergence between a positive overall rating and a negative textual polarity is not unique to the present corpus. Valdivia et al. (2019) likewise found that visitors tend to rate their visits highly, even though their texts still contain negative sentences and aspects, so the numerical score need not represent the sentiment of every sentence in a review. The pattern nevertheless confirms that the interpretation of sentiment must always be placed alongside rating and topic.

Table 5. Cross-Tabulation of Review Rating and Sentiment

Rating	Positive	Neutral	Negative	No Coverage	Total
1	11	2	24	5	42
2	9	1	14	3	27
3	20	0	28	6	54
4	92	7	87	24	210
5	716	59	437	277	1,489
Total	848	69	590	315	1,822

The larger positive proportion is consistent with the tendency of museum reviews to express appreciation, even though criticism remains concentrated on particular features (Gao & Yu, 2024). This result is also consistent with the view that museum reviews combine collection and service dimensions (Alexander et al., 2018; Riva & Agostino, 2022). At Museum X, the high ratings indicate the strength of the overall experience, while 590 negative reviews signal that operational evaluation must not be ignored. The no-coverage category prevents 315 reviews from being wrongly positioned as neutral and makes the limits of InSet visible in quantitative terms.

Leaving the lexicon un-normalised helps to preserve the traceability of the original weights, but it also makes matching dependent on the availability of an identical surface form. Normalisation of the reviews reduces variation in slang and spelling, while derived words that are not listed in InSet are not

forced into a root form. This choice is more conservative and may lower coverage, but no new scores arise from merging derived and root entries that may carry different weights. Because InSet originates from microblogs, the 17.29% no-coverage figure also serves as an empirical measure of the limits of domain transfer to museum reviews.

Dominant Topics in the Visitor Experience

BERTopic produced six groups which is five regular topics and one outlier group (−1). A total of 286 documents fell into the outlier group, giving an outlier ratio of 15.71%, while 1,535 documents fall within the regular topics. The mean topic probability across all documents is 0.666. If only the regular topics are counted, the mean rises to 0.741. The outlier mean is only 0.261. This difference supports the decision not to assign a substantive label to group −1. The presence of outliers is an inherent consequence of BERTopic's HDBSCAN-based clustering step, which labels documents outside dense clusters as noise (Grootendorst, 2022).

Table 6. Evaluation of BERTopic Model Quality

Indicator	Value
Documents for BERTopic	1,821
Total topics, including outliers	6
Regular topics	5
Outlier ratio	15.71%
Mean probability, all documents	0.666
Mean probability, regular topics	0.741
c_v coherence	0.501
c_npmi coherence	0.047
Top-10 word diversity	0.700

A c_v coherence of 0.501 and a c_npmi of 0.047 indicate that within-topic word associations are moderate (c_v) to low (c_npmi), as is common for short, lexically overlapping user-generated reviews. Topic quality is therefore corroborated primarily through the top terms and the diversity score (0.700) rather than treated as high on coherence alone. Röder et al. (2015) position coherence as an approach to approximating judgements of interpretability, while a diversity of 0.700 means that 35 of the 50 top words are unique. This combination indicates that the topics are not entirely repetitive, even though the words ‘anak’, ‘bagus’, and ‘sejarah’ appear across groups.

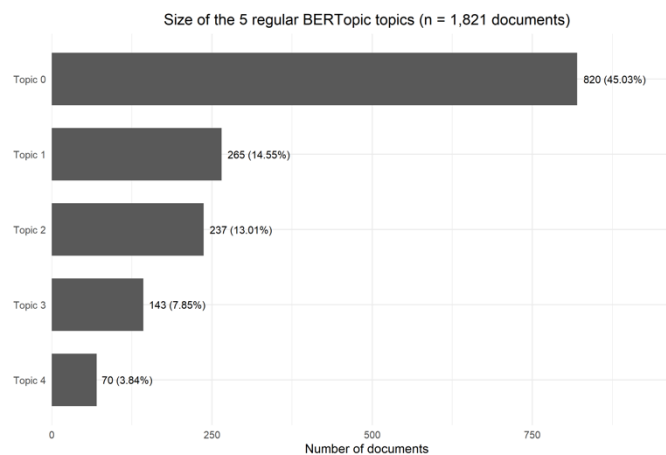


Figure 2. Size of the five regular BERTopic topics.

Topic 0 is the largest group, with 820 reviews, or 45.03% of the 1,821 documents. The words ‘tiket’, ‘anak’, ‘masuk’, ‘beli’, ‘imersifa’, ‘online’, and ‘harga’ show that this group combines the entry and purchase processes, immersive services, collections, and general evaluation. Therefore, it is best to read it

not as a single coherent issue, but as a heterogeneous operational cluster. Its representative reviews span ticket purchase and pricing, entry and queueing, access to immersive services, and general appreciation. The sensitivity analysis verifies this. At finer clustering settings, Topic 0 splits into several smaller topics, albeit at the cost of lower coherence and greater instability, which is why the more parsimonious five-topic solution is retained. Interpretations of Topic 0 are accordingly framed at the level of this operational cluster rather than any single service. Its mean rating of 4.50 is the least positive among the regular topics, making this the largest theme at the centre of operational evaluation. Although it is heterogeneous, the vocabulary proximity of ticket–entry–purchase–online–price suggests that friction occurring before or at the start of the museum experience can shape the overall assessment of a visit.

Topic 1 covers 265 reviews (14.55%) concerning history, statues, culture, collections, exhibitions, objects, and eras. With a mean rating of 4.86, this group confirms the strength of Museum X as a cultural heritage space. Topic 2 contains 237 reviews (13.01%) with expressions such as ‘bagus’, ‘keren’, ‘ramah’, ‘bersih’, ‘seru’, ‘menyenangkan’, and ‘nyaman’, with a rating of 4.91, making it a cluster of general appreciation and comfort. Topic 3 consists of 143 reviews (7.85%) about children, education, family, leisure, and knowledge, and holds the highest rating at 4.92. The smaller Topic 4, with 70 reviews (3.84%), focuses on learning history, gaining insight, cultural introduction, and children, with a rating of 4.90.

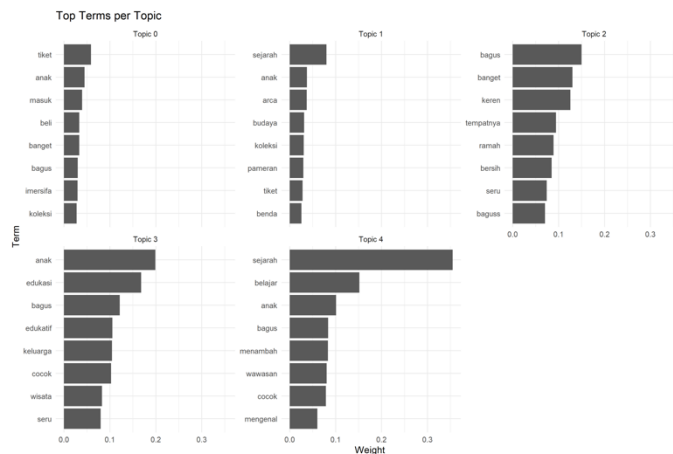


Figure 3. Top eight words in each regular topic.

Table 7. Summary of the Regular BERTopic Topics

Topic	Value	n (%)	Rating	Probability
0	Tickets, entry process, purchase, immersive services, collections, and price	820 (45.03)	4.50	0.688
1	Historical collections, statues, culture, exhibitions, and objects	265 (14.55)	4.86	0.750
2	General appreciation, friendliness, cleanliness, enjoyment, and comfort	237 (13.01)	4.91	0.815
3	Children's education, family, learning trips, and knowledge	143 (7.85)	4.92	0.887
4	Learning history, broadening insight, and cultural introduction	70 (3.84)	4.90	0.777

These results reveal a layered structure of experience. The operational theme of tickets and the entry process dominates the volume of conversation, while the collection–history and general appreciation topics form the next substantive layer. The two education topics are smaller but carry very high ratings. This pattern is in line with the findings of Riva & Agostino (2022) concerning the interconnection of cultural heritage, personal experience, and service, and with Huo et al. (2024), who showed that museum

attributes can appear with differing intensity in reviews. For Museum X, this means that collection and educational value are strengths, but this experience is constrained by an operational bottleneck that is discussed far more frequently.

Robustness Checks of the Topic Solution

To assess the stability of the topic solution, the model was re-fitted under alternative parameter settings with the document embeddings held fixed, so that only the clustering parameters varied. The number of topics is sensitive to these settings, ranging from five (at 2–3% minimum topic size) to nineteen (at 1%). However, the dominant ticketing or operational topic persisted across all configurations, and coherence remained in the same low-to-moderate range throughout. The selected setting (minimum topic size = 2% of the corpus, `n_neighbors` = 15) yields the most parsimonious stable solution and the lowest outlier ratio of the settings tested, while finer settings fragment the corpus into more, less stable topics at equal or lower coherence. The five-topic solution is therefore retained as the most stable and parsimonious representation, with the understanding that its dominant theme, on which the findings and recommendations rest, is invariant across settings, and that finer granularity separates the heterogeneous dominant topic only at the cost of coherence and stability.

Table 8. Topic Solution Under Alternative Parameter Settings

Setting	Regular topics	Outlier ratio	c_v	c_npmi	Dominant topic present
1% min size (nn 15)	19	0.345	0.501	−0.051	Yes
2% min size (nn 15) — selected	5	0.157	0.501	0.047	Yes
3% min size (nn 15)	5	0.185	0.500	0.047	Yes
2% min size (nn 10)	9	0.333	0.544	0.055	Yes

Sentiment Patterns within Each Topic

Integrating sentiment and topic shows that label proportions and mean scores can place emphasis differently. Topic 0 has 367 positive reviews (44.76%) and 346 negative reviews (42.20%), yet the mean score across its 746 covered reviews is −0.336. The number of positive documents is thus slightly higher, but the weight of negative words within the covered documents is strong enough to lower the mean. This topic also has the least positive rating among the regular topics. That combination makes tickets, purchases, entry access, price, and immersive services the area that most consistently requires attention.

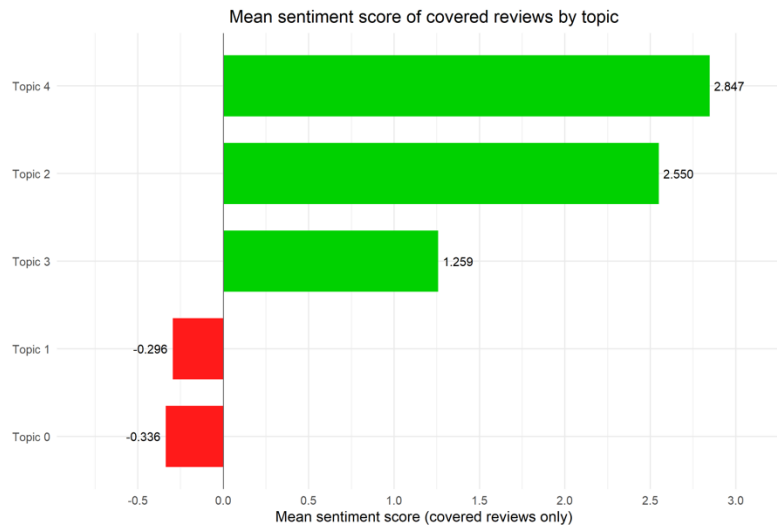


Figure 4. Mean sentiment score of covered reviews by topic.

Table 9. Sentiment Distribution by Topic

Topic	Positive, n (%)	Neutral, n (%)	Negative, n (%)	No coverage, n (%)	Mean sentiment density	Mean score*
-1	115 (40.21%)	15 (5.24%)	73 (25.52%)	83 (29.02%)	0.79	0.975
0	367 (44.76%)	33 (4.02%)	346 (42.20%)	74 (9.02%)	0.08	-0.336
1	137 (51.70%)	12 (4.53%)	98 (36.98%)	18 (6.79%)	0.41	-0.296
2	107 (45.15%)	0 (0.00%)	33 (13.92%)	97 (40.93%)	1.26	2.550
3	74 (51.75%)	5 (3.50%)	33 (23.08%)	31 (21.68%)	0.98	1.259
4	48 (68.57%)	4 (5.71%)	7 (10.00%)	11 (15.71%)	1.73	2.847

*The mean score is computed only for reviews that have at least one InSet match.

Topic 1 also displays a difference between label frequency and weight intensity. Of all of its documents, 51.70% are labelled positive and 36.98% negative, yet the mean score across its 247 covered reviews is -0.296. A rating of 4.86 indicates that the topic of collections, statues, culture, and history remains well appreciated. Because historical terms can appear alongside descriptions of conflict, damage, or negative events, the lexical score must not be equated with rejection of the collection. This finding explains why a high rating can accompany a slightly negative mean score: the lexicon sums the strength of individual words, whereas the rating summarises the overall evaluation of the visit.

Topic 2 is a cluster of general appreciation, with a mean score of 2.550 and a rating of 4.91. Positive reviews make up 45.15% and negative reviews only 13.92%, but 40.93% have no coverage, the highest proportion among the regular topics. Spelling variants and short expressions of praise, such as 'baguss' and 'kereen', may explain part of this non-coverage, because the lexicon was not normalised and letter repetition was simplified to two characters only. The size of the no-coverage group within this topic therefore cannot be read as an absence of attitude because at the same time the rating and the representative words in fact display strong appreciation.

Topics 3 and 4 give the most consistently positive signal. Topic 3 has 51.75% positive, 23.08% negative, a mean score of 1.259, and a rating of 4.92. Topic 4 has 68.57% positive, 10.00% negative, the highest mean score at 2.847, and a rating of 4.90. Both groups reinforce the museum's role as a space for family education, historical introduction, and the broadening of insight, reflecting the historical and educational value expressed in the reviews. The shared words 'anak', 'bagus', and 'sejarah' do not make the two identical. In Topic 3, the context is closer to family leisure and education, whereas Topic 4 focuses more on the learning process and cultural introduction.

Outlier group -1 contains 286 reviews with a low mean probability. Its composition is 40.21% positive, 25.52% negative, 5.24% neutral, and 29.02% no coverage. Although the mean score of the covered reviews is positive (0.975) and the mean rating is 4.70, its words mix general impressions, children, entry processes, collections, and history. This group is not used to draw thematic conclusions because HDBSCAN did not place it stably within any of the regular topics.

The themes of education and history learning show the strongest alignment between high ratings and positive scores. The collection history theme continues to be rated highly, yet the intensity of negative words demands contextual reading. The theme of tickets and the entry process shows the sharpest mix of sentiment, along with a relatively low rating. This synthesis supports the argument of Alexander et al. (2018) that the museum experience is shaped simultaneously by curatorial elements and by everyday service.

Implications

Theoretically, this study shows that rating, sentiment label, score, and topic represent different layers of the visitor experience. The rating captures a global evaluation; the label indicates the direction of the summed weights; the mean score indicates the intensity of the weights across covered reviews; and the topic explains the object of evaluation. Misalignment between the layers is not noise to be eliminated but

information about the structure of the reviews. This layered framework is relevant to visitor studies that view the museum experience as a multidimensional construct (Packer & Ballantyne, 2016).

Methodologically, the no-coverage category needs to be retained in lexicon-based reporting. If the 315 reviews without a match were placed in the neutral category, the neutral proportion would rise from 3.79% to 21.08%, which would be misleading. That separation also makes the domain limits of InSet visible. Not normalising the lexicon preserves the original curated weights of words and phrases, while normalising the reviews increases the likelihood of matching variants that can be explicitly mapped. This procedure prioritises auditability over an unmeasured expansion of coverage.

In practical terms, because Topic 0 is the largest cluster and carries the least positive rating and a slightly negative mean score, it warrants managerial attention. The corpus supports the observation that friction around ticketing, entry, and pricing recurs. Specific operational responses, such as clarifying the distinction between regular and immersive tickets, displaying slot availability, or simplifying the purchase-and-entry flow, are candidate directions for investigation that would require further visitor or institutional evidence and are not established by the topic-level analysis alone. This finding is consistent with Huo et al. (2024) whose topic-model analysis of museum reviews showed that peripheral services and operational arrangements shape visitor experience and satisfaction.

The second priority is to maintain and expand the strengths in children's, family, and history education reflected in the reviews grouped in Topics 3 and 4. Family programmes can be linked to the collection narrative of Topic 1 through visit routes, accompanying materials, or thematic learning agendas. Topic 2 indicates the importance of friendliness, cleanliness, enjoyment, and comfort as reinforcers of experience. Service strategy therefore needs to combine operational smoothness with the quality of interpretation, rather than treating them as separate agendas.

Limitations

This study has several limitations. First, it is bounded by the nature of Google Maps online reviews, which originate from visitors who choose to write and therefore do not represent the whole visitor population. The reviews are also limited to metadata where `originalLanguage = 'id'`, and the period ends on 1 June 2026. Ratings that are heavily skewed towards five stars also limit the range of explicit evaluations. Because the data do not derive from a probability design, percentages are used to describe the corpus rather than to estimate the opinions of all visitors to Museum X.

Second, the sentiment analysis depends on exact matching with InSet and on limited contextual rules. The system does not yet capture irony, long-distance negation, relations between clauses, or terms whose meaning shifts within a museum context. The decision not to normalise the lexicon preserves the original weights but results in some morphological and spelling variants remaining uncovered. Sentiment figures must therefore be understood as lexical polarity rather than as a direct psychological measurement of satisfaction.

Third, BERTopic produced a single topic structure from the configuration specified by the script. Topic 0 remains broad because it mixes tickets, entry processes, immersive services, collections, and general evaluation, while 15.71% of the documents became outliers. Further research can compare time periods, museums, or other platforms and assess the consistency of topics across periods. Such comparison can clarify whether ticketing problems are temporary to the reopening phase or constitute a persistent service pattern.

This study also uses a single sentiment lexicon and a single preprocessing pipeline. InSet was selected for its transparent, word-level weights, its coverage of Indonesian, and its surface-form matching, which enables traceability of those weights to the final score. Systematically varying the lexicon and the preprocessing strategy is beyond the scope of the present study and is left to future work.

CONCLUSION

Based on the sentiment analysis, this study found that reviews of Museum X tend to be lexically positive, yet cannot be assessed in their entirety by InSet. Of the 1,822 reviews, 848 (46.54%) are positive,

590 (32.38%) negative, 69 (3.79%) neutral, and 315 (17.29%) without coverage. Among the covered reviews, the proportion of positive reviews reaches 56.27%. A mean rating of 4.69 indicates a strong overall evaluation, while the presence of 437 five-star reviews labelled as negative confirms that the rating and lexical polarity measure different layers.

The topic modelling with BERTopic on 1,821 documents revealed five regular topics and one outlier group with a ratio of 15.71%. The dominant topic covers tickets, entry processes, purchases, immersive services, collections, and price, while the remaining topics reflect historical-cultural collections, appreciation and comfort, family education, and history learning. Diagnostics of c_v 0.501, c_{npmi} 0.047, diversity 0.700, and a mean probability of 0.666 support the interpretation, while acknowledging the limits of short reviews and the outlier group.

By integrating the results of the sentiment analysis and topic modelling, the study showed that family education and history learning have the strongest alignment between high ratings and positive scores. Conversely, the theme of tickets and the entry process is the largest, has the lowest rating among the regular topics, and carries a slightly negative mean score. The analysis points to ticketing, immersive-service slots, pricing, online purchase, and entry flow as the operational areas most in need of attention.

REFERENCES

- Alexander, V. D., Blank, G., & Hale, S. A. (2018). TripAdvisor Reviews of London Museums: A New Approach to Understanding Visitors. *Museum International*, 70(1–2), 154–165. <https://doi.org/10.1111/muse.12200>
- Asyiah, N., & Aktavia, W. (2025). Penerapan BERTopic dan Analisis Sentimen Leksikal Pada Ulasan Relevan di Google Maps Mengenai Universitas Pamulang. *Journal of Information System Research (JOSH)*, 6(4), 2129–2138. <https://doi.org/10.47065/josh.v6i4.8007>
- Candrea, A. N., Ciobanu, E., Nechita, F., Brătucu, G., Coman, E., Şchiopu, C., & Alexandrescu, M. B. (2025). How Do Visitors to Mountain Museums Think? A Cross-Country Perspective on the Sentiments Decoded from TripAdvisor Reviews. *Electronics (Switzerland)*, 14(8). <https://doi.org/10.3390/electronics14081637>
- Chan, C. H., Bajjalieh, J., Auvi, L., Wessler, H., Althaus, S., Welbers, K., Van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using ‘off-the-shelf’ dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, 3(1), 1–27. <https://doi.org/10.5117/CCR2021.1.001.CHAN>
- Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8, 439–453. <https://doi.org/10.1162/tacl>
- Drivas, I. C., Vraimaki, E., & Lazaridis, N. (2025). I Can’t Get No Satisfaction? From Reviews to Actionable Insights: Text Data Analytics for Utilizing Online Feedback. *Digital*, 5(3). <https://doi.org/10.3390/digital5030035>
- Gao, B., & Yu, S. (2024). Upgrading museum experience: Insights into offline visitor perceptions through social media trends. *Emerging Trends in Drugs, Addictions, and Health*, 4. <https://doi.org/10.1016/j.etdah.2023.100137>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. <http://arxiv.org/abs/2203.05794>
- 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 July). <https://doi.org/10.1371/journal.pone.0304901>

- King, E., Smith, M. P., Wilson, P. F., Stott, J., & Williams, M. A. (2023). Creating Meaningful Museums: A Model for Museum Exhibition User Experience. *Visitor Studies*, 26(1), 59–81. <https://doi.org/10.1080/10645578.2022.2129944>
- Koto, F., & Rahmaningtyas, G. Y. (2017). 2017 International Conference on Asian Language Processing. *21st International Conference on Asian Language Processing (IALP)*, 391–394.
- Liu, P., Zeng, Z., Wang, H., Zhang, H., Zhang, J., & Liu, Z. (2024). Crisis-resistant tourism markets in the pandemic recovery: Evidence from the Confucius Temple Scenic Area. *Tourism Management Perspectives*, 51. <https://doi.org/10.1016/j.tmp.2024.101221>
- Ogunleye, B., Lancho Barrantes, B. S., & Zakariyyah, K. I. (2025). Topic modelling through the bibliometrics lens and its technique. *Artificial Intelligence Review*, 58(3). <https://doi.org/10.1007/s10462-024-11011-x>
- Packer, J., & Ballantyne, R. (2016). Conceptualizing the Visitor Experience: A Review of Literature and Development of a Multifaceted Model. In *Visitor Studies* (Vol. 19, Number 2, pp. 128–143). Routledge. <https://doi.org/10.1080/10645578.2016.1144023>
- Panel on Research Ethics. (2022). *Does Research Using Social Media Platforms Require Research Ethics Board Review?: Guidance to support the implementation of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2*. www.ethics.gc.ca
- Riva, P., & Agostino, D. (2022). Latent dimensions of museum experience: assessing cross-cultural perspectives of visitors from tripadvisor reviews. *Museum Management and Curatorship*, 37(6), 616–640. <https://doi.org/10.1080/09647775.2022.2073560>
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. <https://doi.org/10.1145/2684822.2685324>
- Saoualih, A., Shen, S., Safaa, L., & Su, Y. (2025). A thematic investigation of tourist experiential gaps in Moroccan cultural heritage museums using a sentiment-guided BERTopic text mining approach. *Tourism Recreation Research*. <https://doi.org/10.1080/02508281.2025.2538250>
- Suprina, R., Mustika, A., Rahmanita, M., Rachman, A. F., & Christiyanto, W. W. (2025). An Analysis of Jakarta's Destination Image Through Opinion Mining on Tourists' Digital Reviews. *KnE Social Sciences*, 10(29), 248–259. <https://doi.org/10.18502/kss.v10i29.20269>
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), 267–307. https://doi.org/10.1162/COLI_a_00049
- Tanebeth, R. D. (2026). Perbandingan Kinerja Analisis Sentimen Ulasan Monumen Nasional (Monas) Menggunakan Model Deep Learning BERT dan Klasifikasi Machine Learning Multinomial Naïve Bayes. *Jurnal Komputer Dan Informatika*, 14(1), 1–8. <https://doi.org/10.35508/jicon.v14i1.27042>
- Valdivia, A., Hrabova, E., Chaturvedi, I., Luzón, M. V., Troiano, L., Cambria, E., & Herrera, F. (2019). Inconsistencies on TripAdvisor reviews: A unified index between users and Sentiment Analysis Methods. *Neurocomputing*, 353, 3–16. <https://doi.org/10.1016/j.neucom.2018.09.096>
- van der Veen, A. M., & Bleich, E. (2025). The advantages of lexicon-based sentiment analysis in an age of machine learning. *PLoS ONE*, 20(1 January). <https://doi.org/10.1371/journal.pone.0313092>
- Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7), 5731–5780. <https://doi.org/10.1007/s10462-022-10144-1>