



## Analyzing User-Generated Content in Yogyakarta Tourism Through Integrated TF-IDF, LDA, and Qualitative Descriptive Analysis

Riza Prapascatama Agusdin<sup>1\*</sup>, Shely Rizki Hardiana<sup>2</sup>, Yuseptia Angretnowati<sup>3</sup>,  
Dian Indri Purnamasari<sup>4</sup>, Rifki Indra Perwira<sup>1</sup>

<sup>1</sup>Department of Informatics, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia

<sup>2</sup>Department of Management, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia

<sup>3</sup>Department of Communication Science, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia

<sup>4</sup>Department of Accounting, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia

Received : September 11,  
2026

Revised : September 15,  
2026

Accepted : September 16,  
2026

Online : October 7, 2026

### Abstract

User-generated content shapes how tourists choose a destination, yet Indonesian destination managers rarely turn it into usable evidence. This study identifies the lexical and thematic structure of tourist reviews for five destinations in Yogyakarta, and explains that structure through the accounts of tourism stakeholders. A convergent mixed-methods design was applied to 15,000 Google Reviews, of which 7,963 survived preprocessing. TF-IDF weighted the distinguishing words, LDA grouped them into latent topics, and two focus group discussions with twenty participants explained why those topics mattered. The models yielded two to four topics per site. Malioboro proved the most fragmented because parking conduct and vendor pressure formed a topic separate from night atmosphere. Three concerns recurred everywhere: interaction with local actors, destination management, and price. Lexical, thematic, and qualitative evidence can therefore guide destination development together.

**Keywords** *UGC; TF-IDF; LDA; Qualitative Descriptive Analysis; Tourism Management*

## INTRODUCTION

Platforms that carry user-produced content, later termed User-Generated Content (UGC), have become an important determinant of how tourists form destination preferences (Gretzel & Yoo, 2008). UGC carries experiences, opinions, and recommendations, so it serves as both an information source and a decision-support factor (Nguyen & Tong, 2023). Tourists no longer act merely as consumers of that information, because they also produce it, and their production shapes destination image and travel intention (Santos, 2022). Since reviews come from peers rather than marketers, audiences read them as more credible than official promotion, and UGC has therefore become market intelligence about what tourists expect and how far a destination satisfies them (Aboalghanam et al., 2025; Yamagishi et al., 2024).

Turning that resource into managerial evidence still poses three unresolved problems in the Indonesian tourism context. First, prior UGC studies apply keyword weighting and topic modeling in isolation, so a study either ranks the terms that distinguish a destination or clusters those terms into topics, and few run both on a single preprocessed corpus so the outputs can be read against each other. Second, most treat a region or a platform as a single aggregate corpus, which dissolves the differences between a temple, a market, and a street corridor precisely when managers need those differences to allocate budget. Third, the analysis usually stops at the computational output, and ranked words cannot explain why tourists weigh an issue as they do,

### Copyright Holder:

© Riza, Shely, Yuseptia, Dian, & Rifki. (2026)

Corresponding author's email: rizapra@upnyk.ac.id

### This Article is Licensed Under:



yet very few Indonesian studies collect primary qualitative data that supplies the reasoning.

This study addresses the three problems together. It integrates Term Frequency-Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA), and qualitative descriptive analysis to identify the most representative words per destination, map the latent topics behind them, and explain both through primary qualitative evidence. Yogyakarta was chosen because it remains one of Indonesia's busiest tourism provinces, with 38,030,739 domestic tourist trips recorded in 2024 (BPS DIY, 2025), and that volume sustains a UGC ecosystem large enough for per-destination modeling. Five sites were then selected purposively against these criteria: (1) each appears consistently among the most visited attractions in the province, (2) each represents a different destination type so that the sample covers a heritage monument, a traditional market, an urban street corridor, a tourism village, and a living palace.

## LITERATURE REVIEW

### Qualitative Descriptive

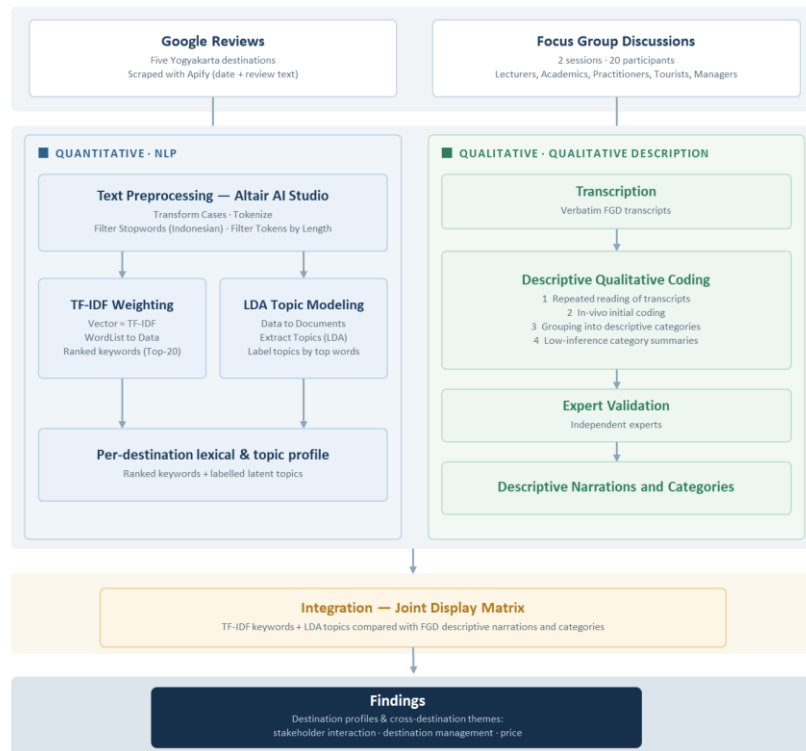
Qualitative descriptive (QD) research summarises events and experiences (Sandelowski, 2010). It differs from ethnography or grounded theory, which interpret data through existing theory, because QD reports experiences in language close to participants' and keeps the researcher close to the data (Hall & Liebenberg, 2024). Paired with computational output, it supplies reasoning that numbers cannot: a keyword ranking shows that a word matters, but only FGD data explains why. Guetterman et al. (2015) formalized this pairing as a joint display: a table in which readers trace agreement or contradiction row by row, while Moseholm and Fetters (2017) advise drawing meta-inferences only after analyzing both strands independently.

### TF-IDF and Latent Dirichlet Allocation (LDA)

TF-IDF remains foundational in information retrieval (Salton & Buckley, 1988), since it weights a term highest when that term concentrates in a small subset of documents rather than spreading evenly. Research by Ali et al. (2022) applied the same logic to tourist reviews by feeding TF-IDF vectors into a topic-sentiment pipeline. LDA takes another route, treating each document as a mixture of latent topics and each topic as a distribution over words, so it groups co-occurring terms instead of scoring them individually (Blei et al., 2003). Research by Sutherland et al. (2020) validated fourteen accommodation-review topics through expert review of the top-weighted words, while Wang et al. (2024) named their topics manually. Choosing the topic count therefore still rests on human judgment, which makes transparent reporting of that decision a methodological obligation.

## RESEARCH METHOD

This study integrates TF-IDF for keyword weighting, LDA for topic modeling, and descriptive qualitative analysis of FGD transcripts. The computational component draws on Google Reviews, while the qualitative component draws on FGDs with tourism experts, academics, active tourists, and destination managers. The design is convergent, so the team analyzed both independently before merging them at the interpretation stage (Creswell, 2022).



**Figure 1.** Research Workflow

## Data Sources

Review data for the five destinations came from Google Reviews and were collected with the Apify scraping tool. Each entry holds the review date and the original text. When a review was not written in Indonesian, the team retained the translation supplied by Google Reviews so its tokens (individual word units) could pass through the same workflow. Table 1 reports the corpus per destination, together with the collection period and the number of reviews carried into the analysis.

**Table 1.** Corpus Profile Per Destination

| Destination        | Total Review | Collection Period | Total Review after Pre-Processing |
|--------------------|--------------|-------------------|-----------------------------------|
| Prambanan Temple   | 3000         | 2025 - 2026       | 1435                              |
| Beringharjo Market | 3000         | 2025 - 2026       | 1488                              |
| Malioboro Street   | 3000         | 2025 - 2026       | 1506                              |
| Taman Sari Village | 3000         | 2025 - 2026       | 1947                              |
| Hadiningrat Palace | 3000         | 2025 - 2026       | 1587                              |

Qualitative data were collected through two FGD sessions, each with ten participants. Participants were purposively recruited based on these criteria: (1) direct involvement in the province's tourism sector and (2) willingness to openly discuss and be recorded on both favorable and unfavorable conditions. The discussion panels comprised two academics, six practitioners and destination managers, and twelve active tourists; each session lasted approximately 60 to 90

minutes. The discussions were guided by a semi-structured protocol covering four sets of questions aligned with the coding categories used in the subsequent analysis: participants' understanding and practices regarding sustainable tourism, how tourists select destinations, the image or character they associate with those destinations, and the role of digital technology in shaping the face of Indonesian tourism. Each session was recorded using digital devices, and professional transcribers prepared text documents for analysis.

### **Descriptive Qualitative Analysis**

The qualitative work summarises each phenomenon in participants' own words, without attaching it to an abstract theoretical framework (Doyle et al., 2020). The team worked manually and sorted the material into four areas of interest: sustainable tourism, tourist preferences and profiles, destination persona, and technology in Indonesian tourism. Coding ran in four steps. The researchers first read the transcripts repeatedly to build familiarity before assigning any code (Isangula et al., 2024). Then applied initial codes, favored in-vivo labels taken from the language of participants (Saldaña, 2021), and grouped similar codes into descriptive categories using worksheets. Finally, summarised each category in language close to the original data, following the low-inference principle at the core of the approach. To prevent single-interpreter bias, the team sent the categories to independent experts in sustainable tourism and revised them until those experts judged them defensible.

### **TF-IDF and LDA Analysis**

The team ran both computational analyses in Altair AI Studio, formerly RapidMiner. Each destination received a separate run, and within that run every individual review counted as one document. After the Retrieve and Nominal to Text operators, a single Process Documents from Data operator applied four operations in sequence: Transform Cases, Tokenize, Filter Stopwords (Indonesian), and Filter Tokens by Length. Setting the vector creation parameter to TF-IDF produced the weights, while WordList to Data converted the internal word list into a sortable table, so ranking that table revealed the vocabulary that best characterizes a destination. For LDA, the cleaned rows fed Data to Documents and then Extract Topics from Documents. The team trained several candidate models per destination until the topics proved interpretable, with no word recurring across them, and labeled each topic by its highest-probability words.

### **Data Integration and Ethics**

The team integrated the three result sets through a joint display matrix, an established technique for juxtaposing quantitative and qualitative findings (Guetterman et al., 2015). For each destination, the twenty highest-scoring TF-IDF keywords were placed alongside the LDA topics of that same site. The team then compared this profile against the descriptive categories drawn from the FGD material, recording in every row whether the two sources reinforced each other. All participants gave verbal consent after the team had explained the objectives, procedures, and risks of the study. Anonymization protected confidentiality at every stage, so participants are identified below only by code, and reviewers' identities received the same protection.

## **FINDINGS AND DISCUSSION**

This section reports the analysis of five destinations in the Special Region of Yogyakarta.

### **Focus Group Discussion Results**

Seven patterns recurred across the transcripts, and the four coding categories below

organize them.

### ***Tourist preferences and profiles***

Travel decisions proved social rather than individual. Most participants travel with family, a partner, or friends, and the group rather than the person sets the itinerary, since "the primary travel motivation is seeking experience and togetherness". Shared experience outweighs any appetite for novelty, so participants return willingly to familiar places when the company is right: "travel preferences depend not only on the location but also on whom one travels with". Three attributes settle the booking, namely price, attractiveness, and facilities. One participant named them directly: "price, facilities, and uniqueness are the deciding factors for tourist visits today".

### ***Technology in Indonesian Tourism***

Social media supplies inspiration and builds the initial shortlist, yet participants stop short of treating it as evidence, because conversations with close contacts and user reviews settle the final choice. Participants also reported a change in motive, describing "a shift in travel motivation toward showing one's presence on social media".

### ***Destination Persona***

Warm interaction with local hosts leaves a deeper impression than fine scenery offered by indifferent ones, and participants named encounters with traders, guides, and staff as the part of a visit they remember: "quality interaction with the local community or with guides is preferred over a beautiful destination that lacks friendliness". Promotion that overstates conditions on the ground then erodes both trust and the intention to return: "promotion of tourist sites often exaggerates, so expectation and reality do not match, and this reduces trust and makes tourists unlikely to come back".

### ***Sustainable Tourism***

Participants endorse sustainable practice in principle. It enters their choice only once the benefit to the environment and to the local community becomes visible on site, as one participant framed the goal: "sustainable tourism development through the preservation of cultural resources".

### **TF-IDF Results**

TF-IDF identifies which words set a destination apart within its own review corpus, and Table 2 follows that logic.

**Table 2.** Four highest-weighted terms per destination

| <b>Destination</b> | <b>Term and TF-IDF score</b>                                     |
|--------------------|------------------------------------------------------------------|
| Prambanan Temple   | candi 667.99; kuil 330.38; bagus 324.62; tiket 307.82            |
| Beringharjo Market | pasar 629.30; batik 496.84; murah 428.53; harga 409.50           |
| Malioboro Street   | jalan 587.49; jogja 456.55; ramai 363.95; malam 283.04           |
| Taman Sari Village | foto 611.79; guide 585.96; bagus 553.20; sejarah 525.68          |
| Hadiningrat Palace | keraton 579.40; sejarah 442.82; budaya 396.74; yogyakarta 382.55 |

\*Note. candi = temple, kuil = shrine, bagus = good, tiket = ticket, pasar = market, murah = cheap, harga = price, jalan = street, ramai = crowded, malam = night, foto = photo, sejarah = history, budaya = culture.

Each destination announces itself through its first term, so the leading words add little beyond confirmation. The ranks below them carry the working information. Ticketing at Prambanan, price at Beringharjo, night atmosphere at Malioboro, guiding at Taman Sari, and cultural learning at Hadiningrat Palace all sit in that second layer.

### LDA Topics

Topic modeling asks which words travel together inside the same review.

**Table 3.** Topic structure per destination

| Destination        | Topic Amount | Topic labels                                                                                                                                           |
|--------------------|--------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Prambanan Temple   | 2            | (1) Heritage site, Hindu mythology, and its management; (2) Visit experience and site conditions                                                       |
| Beringharjo Market | 2            | (1) Affordable traditional batik shopping; (2) Atmosphere and physical condition of the market                                                         |
| Malioboro Street   | 4            | (1) Transactions and selling pressure; (2) Walking experience and night atmosphere; (3) Parking and local actor behaviour; (4) Exploration of the area |
| Taman Sari Village | 2            | (1) Historic and scenic tourism; (2) Visit experience with guides                                                                                      |
| Hadiningrat Palace | 3            | (1) Dress code and rule compliance; (2) Access, tickets, and operations; (3) Javanese palace heritage                                                  |

At Prambanan, reviewers split the site into two registers. One treats the temple as a conservation object and names staff, restoration work, and Hindu mythology, while the other reports heat, distance, and cleanliness inside the complex. Queueing and ticketing govern that second register, so operational friction can outweigh admiration for the monument (Bucurescu, 2015).

At Beringharjo, the second topic contradicts the market's promotional image, because words for mud, dirt, and poultry stalls place the site in an ordinary trading landscape rather than a curated one. Shoppers accept that trade-off for low prices, yet they still grade hygiene established for destination satisfaction.

Four topics make Malioboro the most fragmented corpus here, and crowding explains that fragmentation because it works in two directions at once. Visitors seek the night atmosphere, then object to parking touts, smokers, and vendors who push a sale. Participants named overtourism as the main management problem, so Malioboro works as an experiential space whose value lies in atmosphere rather than in any single attraction.

At Taman Sari, photography and guiding dominate and reinforce one another, since guides narrate the water castle while the corridors and bathing pools supply the frames visitors want. Guide performance predicts satisfaction, and participants placed host interaction above scenery, so heritage value and photographic value rise together.

At the Hadiningrat Palace, visitors arrive for the Javanese court and its performances, then judge the day on rules of entry, opening hours, and parking. Culture draws them in, operations decide whether they leave satisfied, and dress rules act as a boundary most visitors accept once staff explains it.

### Joint Display and Cross-Destination Findings

The final step places the three strands side by side, one theme per row, and can trace agreement and addition across sources.

**Table 4.** Integration of computational results and qualitative context

| Theme                         | Computational evidence                                                                             | FGD context                               | Managerial implication                    |
|-------------------------------|----------------------------------------------------------------------------------------------------|-------------------------------------------|-------------------------------------------|
| Interaction with local actors | Guiding terms at Taman Sari; vendor pressure at Malioboro; courtiers and hospitality at the Palace | Friendly hosts outrank scenery            | Train frontline actors and certify guides |
| Destination management        | Hygiene and crowding at Beringharjo and Malioboro; access terms at Prambanan and the Palace        | Facilities decide the visit               | Repair hygiene, parking, and queue flow   |
| Price                         | Cheapness and price at Beringharjo and inside Palace ticketing                                     | Price sensitivity under economic pressure | Segment and position by price tier        |
| Alternative narratives        | Second- and third-ranked terms across all five corpora                                             | Expectation gaps reduce trust             | Promote what reviewers actually praise    |

Across the five joint displays, the review corpus and the FGD data converged on every theme. The qualitative material then explained why price, crowding, and host behavior weigh so heavily in tourist judgment.

### CONCLUSIONS

This study examined the words tourists use to describe five destinations in the Special Region of Yogyakarta and the reasons behind those words. Three concerns recurred across all destinations. Tourists value their interactions with local stakeholders, such as traders and guides; judge a destination by its accessibility, facilities, and cleanliness; and remain sensitive to price amid current economic pressure.

The study advances UGC research in tourism on the three counts identified in the Introduction. It sequences TF-IDF and LDA on a single preprocessed corpus, so the ranked vocabulary and latent topics can be read against each other rather than in separate studies. It models each destination separately, which reveals differences an aggregate regional corpus would hide, as the four-topic structure at Malioboro against the two-topic structure at Prambanan demonstrates. It then attaches primary qualitative data to the computational output through joint displays, and that attachment converts a ranking into an explanation managers can act on.

### LIMITATIONS & FURTHER RESEARCH

Four boundaries set during data collection limit these results, and each points to a next step. The corpus covers one province and one platform, so it carries the self-selection and rating biases of Google Reviews; widening it across regions and platforms such as TripAdvisor would test whether the three recurring concerns survive. Translation may have flattened idiomatic expressions, so Indonesian-specific preprocessing with stemming would sharpen the lexical

results. Topic counts rested on manual inspection, and each model was trained once, so later studies should fix the count with a coherence metric and report stability across random seeds. Twenty participants coded without inter-coder agreement cannot represent the full stakeholder population, so a larger panel with reported agreement and complete joint displays would make the integration transparent. Aspect-based sentiment analysis would also quantify polarity per topic, and a longitudinal design would show how destination image shifts over time.

## REFERENCES

- Aboalganam, K. M., AlFraihat, S. F., & Tarabieh, S. (2025). The Impact of User-Generated Content on Tourist Visit Intentions: The Mediating Role of Destination Imagery. *Administrative Sciences*, 15(4), 117. <https://doi.org/10.3390/admsci15040117>
- Ali, T., Omar, B., & Soulaïmane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. *MethodsX*, 9, 101894. <https://doi.org/10.1016/j.mex.2022.101894>.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- BPS DIY. (2025). *Perkembangan Pariwisata Daerah Istimewa Yogyakarta, Desember 2024*. BPS Provinsi D.I. Yogyakarta. <https://yogyakarta.bps.go.id/id/pressrelease/2025/02/03/1611/perkembangan-pariwisata-daerah-istimewa-yogyakarta-desember-2024.html>. (accessed 25 August 2026)
- Bucurescu, I. (2015). Managing tourism and cultural heritage in historic towns: Examples from Romania. *Journal of Heritage Tourism*, 10(3), 248–262. <https://doi.org/10.1080/1743873X.2014.968162>.
- Creswell, J. W. (2022). *A concise introduction to mixed methods research* (Second Edition). SAGE.
- Doyle, L., McCabe, C., Keogh, B., Brady, A., & McCann, M. (2020). An overview of the qualitative descriptive design within nursing research. *Journal of Research in Nursing*, 25(5), 443–455. <https://doi.org/10.1177/1744987119880234>.
- Gretzel, U., & Yoo, K. H. (2008). Use and Impact of Online Travel Reviews. In P. O'Connor, W. Höpken, & U. Gretzel (Eds.), *Information and Communication Technologies in Tourism 2008* (pp. 35–46). Springer Vienna. [https://doi.org/10.1007/978-3-211-77280-5\\_4](https://doi.org/10.1007/978-3-211-77280-5_4).
- Guetterman, T. C., Feters, M. D., & Creswell, J. W. (2015). Integrating Quantitative and Qualitative Results in Health Science Mixed Methods Research Through Joint Displays. *The Annals of Family Medicine*, 13(6), 554–561. <https://doi.org/10.1370/afm.1865>.
- Hall, S., & Liebenberg, L. (2024). Qualitative Description as an Introductory Method to Qualitative Research for Master's-Level Students and Research Trainees. *International Journal of Qualitative Methods*, 23, 16094069241242264. <https://doi.org/10.1177/16094069241242264>.
- Isangula, K. G., Kelly, S., & Wamoyi, J. (2024). Manual Qualitative Data Coding Using MS Word for Students and Early Career Researchers in Resource-Constrained Settings. *International Journal of Qualitative Methods*, 23, 16094069241299223. <https://doi.org/10.1177/16094069241299223>.
- Moseholm, E., & Feters, M. D. (2017). Conceptual models to guide integration during analysis in convergent mixed methods studies. *Methodological Innovations*, 10(2), 2059799117703118. <https://doi.org/10.1177/2059799117703118>.
- Nguyen, T. T. T., & Tong, S. (2023). The impact of user-generated content on intention to select a travel destination. *Journal of Marketing Analytics*, 11(3), 443–457. <https://doi.org/10.1057/s41270-022-00174-7>.
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4e ed.). SAGE.

- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).
- Sandelowski, M. (2010). What's in a name? Qualitative description revisited. *Research in Nursing & Health*, 33(1), 77–84. <https://doi.org/10.1002/nur.20362>.
- Santos, M. L. B. D. (2022). The “so-called” UGC: An updated definition of user-generated content in the age of social media. *Online Information Review*, 46(1), 95–113. <https://doi.org/10.1108/OIR-06-2020-0258>.
- Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020). Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. *Sustainability*, 12(5), 1821. <https://doi.org/10.3390/su12051821>.
- Wang, Z., Udomwong, P., Fu, J., & Onpium, P. (2024). Destination image analysis and marketing strategies in emerging panda tourism: A cross-cultural perspective. *Cogent Business & Management*, 11(1), 2364837. <https://doi.org/10.1080/23311975.2024.2364837>.
- Yamagishi, K., Canayong, D., Domingo, M., Maneja, K. N., Montolo, A., & Siton, A. (2024). User-generated content on Gen Z tourist visit intention: A stimulus-organism-response approach. *Journal of Hospitality and Tourism Insights*, 7(4), 1949–1973. <https://doi.org/10.1108/JHTI-02-2023-0091>.