



2026 1st ISIBER



#313 (1571232430): Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA

Detail Information Please visit our official website <https://isiber.unsia.ac.id/>

#313 ([1571232430](#)): *Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA*

Hide details

BIBTeX

	Drag to change order	Author name	Author affiliation (edit for paper)	Author email	Email	Delete
Authors	⋮	Tegar Mandiri	Universitas Budi Luhur, Indonesia	2212500165@student.budiluhur.ac.id	✉	✂
	⋮	Safitri Juanita	Universitas Budi Luhur, Indonesia	safitri.juanita@budiluhur.ac.id	✉	✂



Paper title *Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA* Only the chairs can edit

Conference and track **2026 International Seminar on Intelligent Business and Edge-Computing Research (ISIBER) - 5 - Information System**

Abstract Only the chairs can edit Customer reviews on digital platforms contain valuable insights into guest experiences, but their...

Keywords customer feedback; hospitality industry; hotel reviews; latent dirichlet allocation; topic modeling Only the chairs can edit

Topics Big Data Platforms and Technologies; Big Data and Machine Learning Applications and Experiences

Similarity On [Dec 8, 2025](#), iThenticate computed a similarity score of [3](#) for the review manuscript.

Personal notes



Roles

You are the creator and an author for this paper.
You have authored an accepted paper in this conference.

Status

finalmanuscript

Copyright

IEEE; IEEE completed [Dec 31, 2025](#)

Registration

[Tegar Mandiri](#) has [registered and paid](#) for [Batch1:RS-1](#)

Presented

by Tegar Mandiri in session 2-7: *Parallel Session 2-7* chaired by [Catur Nugroho](#) from Thu, February 26, 2026 13:00 WIB until 16:00 (4th paper) in Room 7 (15 min.)

Review manuscript

Final (accepted) manuscript / PDF must be submitted to PDF eXpress

Presentation

Auxiliary files



up to 6 pages



Program

Time (Jakarta)	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9
Thursday, February 26									
07:15 am-09:15 am	1-1: <i>Parallel Session 1-1</i>	1-2: <i>Parallel Session 1-2</i>	1-3: <i>Parallel Session 1-3</i>	1-4: <i>Parallel Session 1-4</i>	1-5: <i>Parallel Session 1-5</i>	1-6: <i>Parallel Session 1-6</i>	1-7: <i>Parallel Session 1-7</i>	1-8: <i>Parallel Session 1-8</i>	1-9: <i>Parallel Session 1-9</i>
09:15 am-12:25 pm	OC: <i>Opening Ceremony + Plenary Speaker</i>								
12:25 pm-01:00 pm	LB: <i>Lunch Break</i>								
01:00 pm-04:00 pm	2-1: <i>Parallel Session 2-1</i>	2-2: <i>Parallel Session 2-2</i>	2-3: <i>Parallel Session 2-3</i>	2-4: <i>Parallel Session 2-4</i>	2-5: <i>Parallel Session 2-5</i>	2-6: <i>Parallel Session 2-6</i>	2-7: <i>Parallel Session 2-7</i>	2-8: <i>Parallel Session 2-8</i>	2-9: <i>Parallel Session 2-9</i>
04:00 pm-04:15 pm	CC: <i>Closing</i>								
04:15 pm-04:30 pm	EoC: <i>The END of CONFERENCE</i>								

Application of SVM and SMOTE for Early Prediction of Heart Disease Based on Medical Record Data

Emir Ahmad Reza (Binus University, Indonesia); Thooyibah Thooyibah (Bina Nusantara University, Indonesia)

A Data-Driven Hybrid STR-LSTM Model with Adaptive Meta-Learning for Short-Term Electricity Load Forecasting

Desri Kristina Silalahi (Telkom University, Indonesia & National Taiwan University of Science and Technology, Taiwan); Yung-Yao Chen and Chao-Lung Yang (National Taiwan University of Science and Technology, Taiwan)

Classification of Phishing Mail Detection Using Hybrid Approach of Fusion Model

Sakib Ur Rahman and Taukir Ahmed (UITS, Bangladesh); Mehadi Hasan Foysal (BUP, Bangladesh); Talukder Abdullah Al Talha (NSU, Bangladesh); Nahid Hasan Shakil (CSU, Australia); Md. Mushfiqur Rahman (Samarkand State University, Uzbekistan); Roni Chandra Nath (BUP, Bangladesh); Md Rifat Hossain (UITS, Bangladesh); Md. Mahfujul Alam (UIU, Bangladesh)

Regional Segmentation of Motorcycle Accessory Purchases Using Optimized K-Means Clustering

Neisha Aurelia, Callista Tanjung and Ivan Diryana Sudirman (Bina Nusantara University, Indonesia)

LSTM-Based Hourly Sales Forecasting in SME Food and Beverage Retail

Moch Azis Munawar and Ivan Diryana Sudirman (Bina Nusantara University, Indonesia)

Performance Analysis of Time Series Models to Predict Sales Activity of SMEs

Justine Christian Susanto and Maureen Letitia Wiratama (Bina Nusantara University, Indonesia); Patrick Christian Nathaniel (Bina Nusantara University, Indonesia); Ivan Diryana Sudirman (Bina Nusantara University, Indonesia)

Harmonizing Sentiments: Grab Apps User Reviews using Machine Learning Model

Evaristus Didik Madyatmadja (Bina Nusantara University & School of Information Systems, Indonesia); Beverly Dennis, Irawan Subandi, Jennyfer Dermawan Yaputri, Kana Toya, Rayner Livano Kalyana and Daffa Farras Putra Tarigan (Bina Nusantara University, Indonesia)

Enhanced NGINX Log Analysis Using Retrieval-Augmented Generation and Re-Ranking Mechanism with Large Language Model

Beny Ibrani (Bina Nusantara University, Indonesia); Karto Iskandar (BINA NUSANTARA University, Indonesia)

Data-Driven Cognitive Level Profiling Using Spectral Clustering in Higher Education

Nuril Kusumawardani (Binus University, Indonesia); Yuliana Lisanti (Bina Nusantara University, Indonesia); Istiani Istiani (Bina Nusantara University, Indonesia); Veronica Veronica (Bina Nusantara University, Indonesia); Chen Kang Lee (Universiti Tunku Abdul Rahman, Malaysia)

Reciprocal Interaction Modeling and Community-Based Influence Measurement in Large Twitter Networks Using the Leiden and Louvain Algorithms

Handrizal Handrizal, Ainul Hizriadi, Briyan Yehezkiel and Muhammad Nurhadi Alfayyadh (Universitas Sumatera Utara, Indonesia)

Evaluating CVIM Performance in Convenience Goods Through Transformer-Based Text Analytics

Jefdi Andi and Ivan Diryana Sudirman (Bina Nusantara University, Indonesia)

Quantifying the Zone of Tolerance from Review Sentiment Distributions Using GMM and BERTopic

Ivan Diryana Sudirman (Bina Nusantara University, Indonesia); Patah Herwanto (Universitas Ekuitas, Indonesia)

2-7: Parallel Session 2-7

Room 7

Chair: Catur Nugroho (Universitas Siber Asia, Indonesia)

Feature Envy Extraction in Kotlin Using Custom Detekt Rules

Farid Akbar Rahmat Firdaus, Adhistya Erna Permanasari and Teguh Bharata Adji (Universitas Gadjah Mada, Indonesia); Argo Wibowo (Universitas Gadjah Mada, Indonesia & Universitas Kristen Duta Wacana, Indonesia)

Implementation of Market Basket Analysis Using Apriori Algorithm to Determine Product Associations of Wonogiren Batik

Nur Fitriana Shalihah (Universitas Telkom, Indonesia); Muhammad Lulu Latif Usman and Annisaa Utami (Telkom University, Indonesia)

Short-Term Hyperlocal PM2.5 Forecasting Based on Meteorological Conditions Using GRU Architecture

Given Putra (Universitas Tarumanagara, Indonesia); Tony Tony (Tarumanagara University, Indonesia)

Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA

Tegar Mandiri and Safitri Juanita (Universitas Budi Luhur, Indonesia)

Comparative Analysis of MobileNet Architectures for Apple Disease Classification

M Aldiki Febriantono (Bina Nusantara University, Indonesia & BINUS, Indonesia)

Implementation of Artificial Intelligence In A Computer-Aided Diagnosis System For Disease Detection from Medical Images

Mochammad Rafi Al Rizkan (Bina Nusantara University, Indonesia); Natasha Audrey (Binus University, Indonesia); Suzanna Suzanna (Bina Nusantara University, Indonesia)

Malicious Web Page Detection: A Comparative Analysis of Machine Learning with Geospatial Features

M. Dja'far Karzein Hafidz Yuntanozra, Nur Aulia Pertiwi, Muhammad Faki Raihan, Friza Wisono and Ilham Mulya Cahya Satria (Politeknik Siber dan Sandi Negara, Indonesia)

Crime Risk Estimation with Spatially Misaligned Administrative Data: Evidence from Proportional Small Area Estimation in Indonesia

Avrilla Eta Cahya Wardani (Universitas Gadjah Mada, Indonesia); Rudy Hartanto (Gadjah Mada University & Electrical Engineering and Information Technology Departmen, Faculty of

Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA

Tegar Mandiri
Department of Information System
Universitas Budi Luhur
Jakarta, Indonesia
2212500165@student.budiluhur.ac.id

Safitri Juanita
Department of Information System
Universitas Budi Luhur
Jakarta, Indonesia
safitri.juanita@budiluhur.ac.id

Abstract—Customer reviews on digital platforms contain valuable insights into guest experiences, but their unstructured nature makes large-scale manual analysis impractical. This study aims to identify service-quality themes expressed in guest feedback for highly rated hotels in Jakarta by applying a topic modelling framework based on Latent Dirichlet Allocation (LDA). The dataset consists of 12,798 Indonesian-language reviews of highly rated properties, collected primarily from Google Maps and its partner platforms. The text was processed through cleaning, normalization, stemming, and stopword removal, after which a coherence-based evaluation was used to select the optimal model. A seven-topic configuration ($C_v=0.6065$) provided the most coherent representation of customer feedback. The resulting topics reveal two dominant patterns: positive reviews consistently highlight staff friendliness and breakfast quality as key drivers of guest satisfaction, while negative reviews frequently refer to late check-ins, administrative issues, and miscommunication. These findings produce detailed, data-driven indicators of service performance and offer practical guidance for hotel operators seeking to prioritize operational improvements. Overall, the results demonstrate the effectiveness of LDA in extracting actionable service-quality insights in the Jakarta hospitality sector.

Keywords—customer feedback, hospitality industry, hotel reviews, latent dirichlet allocation, topic modeling.

I. INTRODUCTION

The hospitality sector in Jakarta continues to grow, supported by increasing business travel, tourism activities, and the wider use of digital information channels [1]. Tourism is an important contributor to Indonesia's economic growth, with Jakarta functioning as a key destination for both domestic and international travelers. This situation reflects the diverse accommodation needs in Jakarta, covering different price ranges and service levels. As competition becomes more intense, hotels are increasingly required to better understand guest expectations and to distinguish their services in a competitive market.

In this context, customer reviews published on online platforms have an important role in shaping hotel reputation and supporting managerial decision-making. These reviews contain experience-based information that reflects guests' perceptions of service quality, facilities, cleanliness, location, and value for money more comprehensively than traditional survey methods [2]. The large volume of such reviews allows hotels to capture a wide range of guest experiences across different contexts and conditions. However, the unstructured and large-scale nature of online reviews makes manual analysis inefficient and vulnerable to subjective interpretation, especially for hotels that receive a high volume of feedback across multiple platforms, highlighting the need for systematic and data-driven analysis approaches.

Beyond traditional manual inspection, computational text mining techniques provide scalable approaches for analyzing large volumes of unstructured hotel reviews and identifying recurring patterns in guest feedback [3]. Among these approaches, topic modeling methods such as Latent Dirichlet Allocation (LDA) are used to uncover latent themes in text by representing each document as a mixture of topics and each topic as a distribution over words. This representation enables a more detailed understanding of the issues discussed by customers, rather than relying solely on overall rating scores [4], [5]. LDA has been applied across various domains, including healthcare discussions [5], online misinformation [6], and tourism-related content [7], demonstrating its usefulness in revealing underlying concerns, attitudes, and experience dimensions that are not explicitly labeled in the data.

In the hospitality domain, LDA-based topic modeling has been widely applied to analyze online hotel reviews and travel-related narratives. These studies commonly identify topics related to service, cleanliness, room comfort, food and beverage, price, and location, which are then associated with guest satisfaction or rating outcomes [8]. Several works further combine LDA with sentiment analysis or supervised learning models to evaluate customer satisfaction levels or predict hotel ratings, for example by using LDA to extract topic features and subsequently applying Support Vector Machine classifiers for sentiment prediction [9]. In addition, some studies adopt a knowledge discovery in databases (KDD) framework that integrates LDA-based topic modeling and sentiment analysis to systematically examine customer experience patterns in user-generated hotel reviews [10].

However, most existing studies rely on international datasets, multilingual corpora, or data from global hotel chains, with limited attention given to localized settings. As a result, findings from these studies may not fully capture the characteristics of specific urban destinations. Research that explicitly focuses on high-rated hotels within a single city remains limited. This gap leads to an incomplete understanding of how topic structures reflect local guest expectations, cultural nuances, and service quality standards in cities such as Jakarta.

In the Indonesian context, tourism and hospitality research has increasingly highlighted the role of digitalization and data-driven evaluation. However, the analytical use of online hotel reviews, particularly through topic modeling techniques, remains relatively limited. Jakarta generates a large volume of hotel reviews across multiple online platforms, providing a rich source of customer feedback. Despite this potential, only a small number of studies systematically apply topic modeling to examine how guests evaluate high-rated urban hotels within the local cultural and economic setting.

This gap raises important questions regarding which service-quality dimensions dominate customer feedback in high-performing Jakarta hotels and which aspects, such as staff hospitality, cleanliness, breakfast quality, check-in procedures, location, or shared facilities, most strongly shape guest satisfaction. To address this gap, the present study applies an LDA-based topic modeling framework to 12,798 Indonesian-language reviews of high-rated hotels in Jakarta, with the aim of discovering fine-grained service-quality themes that can support data-driven decision-making for hotel operators seeking to sustain or enhance their service performance.

To answer these questions, an analytical approach is required that can manage large volumes of unstructured review text while retaining the diversity of guest experiences. Topic modeling, particularly Latent Dirichlet Allocation (LDA), provides a suitable approach by representing each review as a mixture of latent topics and each topic as a distribution of words. This approach allows service-related dimensions to emerge directly from guest narratives rather than being defined in advance by researchers. By applying LDA to high-rated hotel reviews in Jakarta, this study seeks to reveal the underlying structure of customer feedback and to produce interpretable themes that relate specific aspects of the stay, such as staff interactions, room conditions, and facilities, to overall perceptions of service quality.

This study contributes in three main ways. First, it compiles a dataset of Indonesian-language reviews from highly rated hotels in Jakarta, which serves as a basis for examining guest experience patterns in an urban hospitality context. Second, it generates detailed service-quality insights using LDA-based topic modeling, helping to clarify how guests perceive different aspects of hotel services. Third, the study presents a structured topic modeling workflow that supports data-driven evaluation of hotel operations and highlights the practical use of LDA for decision-making in Jakarta’s hospitality sector.

II. METHODOLOGY

This section describes the proposed framework, consisting of sequential topic modeling stages applied to high-rated hotel customer reviews from Jakarta. The workflow was implemented in Python using the Gensim library for LDA based topic modeling. The research flow is shown in Figure 1.

A. Data Preparation

The dataset used in this study was constructed from hotel search results in Jakarta retrieved from the Google Maps platform in 2025. The search results provided a list of hotels along with their ratings and customer reviews aggregated from several partner platforms, including Tripadvisor, Trip.com, ZenHotels.com, and all.accor.com. A total of 13 hotels in the DKI Jakarta area were selected based on three main criteria: a minimum guest rating of 4.5, a price range between IDR 350,000 and IDR 750,000 per night, and the availability of seven essential facilities representing the upper-middle service segment (free Wi-Fi access, free parking area, swimming pool, air conditioning, child-friendly facilities, fitness center, and wheelchair access).

During the data preparation stage, customer reviews associated with the selected hotels were crawled, yielding a collection consisting predominantly of Indonesian-language review text. An initial validation and cleaning process was

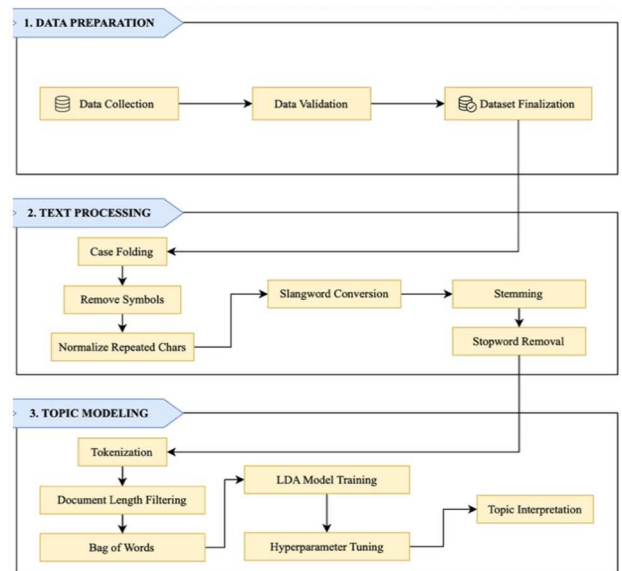


Fig. 1. A Proposed framework for Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA Topic Modeling

systematically conducted to remove duplicate entries, reviews without textual content, and records considered irrelevant to the guest experience. Following this preparation process, the dataset comprised 12,798 review records that were deemed suitable for subsequent text processing and topic modeling analysis. Table I presents sample reviews from the dataset to illustrate the raw data characteristics prior to the text preprocessing stage.

B. Text Processing

The text processing stage converts raw reviews into a clean and standardized corpus before further modeling. This process consists of two main steps, namely text cleaning and text normalization, which are applied sequentially to all documents generated during the data preparation stage. In the text cleaning stage, all characters are converted to lowercase (case folding), followed by the removal of symbols such as HTML tags, URLs, numbers, punctuation marks, and non-alphabetic characters, as well as the normalization of excess spaces to ensure consistent text representation. In addition, excessive letter repetition in expressive words is normalized by limiting the number of consecutive characters, for example “mahaaaaa” (expensive), “baguuuus” (good), or “lamaaaa” (slow), to produce more standardized and reliably processable word forms.

The text normalization stage focuses on aligning lexical forms and reducing unnecessary word variations [11]. First, slang and non-standard words are converted into their formal equivalents using a domain-specific slang dictionary developed for hotel reviews, ensuring consistent lexical mapping. Second, each token is processed using an Indonesian stemmer to reduce words to their base forms, allowing morphological variations (e.g., prefixes, suffixes, or insertions) to be treated as a single term in the analysis. Third, a hospitality-specific stopword list is applied to remove high-frequency words with low semantic value, thereby enhancing the relevance of the corpus to service quality analysis [11]. Overall, these steps produce a cleaner and more representative corpus, which is subsequently used for feature engineering

and topic modeling. This series of steps produces a cleaner, more compact, and representative corpus, which is then used as input in the feature engineering and topic modeling stages.

C. Topic Modeling

The topic modeling stage is designed to uncover latent thematic structures within the corpus of hotel customer reviews through the LDA approach [12]. The effective application of topic models to online customer reviews necessitates specific data transformation steps, including tokenization, document length filtering, and term pruning, to ensure interpretability and quality [8]. The preprocessed reviews are then converted into a Bag-of-Words (BoW) representation using a dictionary constructed from the entire corpus. In text classification, the BoW method records the number of occurrences of each term in the bag created for each document, disregarding word order and grammatical structure [13]. This BoW corpus, together with its associated dictionary, serves as the primary input for the subsequent LDA modeling process.

The LDA model is trained using the Gensim library, assuming that each document is composed of a mixture of latent topics and that each topic is characterized by a probabilistic distribution over words [14]. To determine a suitable number of topics, multiple model configurations are explored over a predefined range of topic counts, and each configuration is evaluated using a coherence-based metric to assess the semantic consistency and interpretability of the resulting topics [3]. Based on this evaluation, a topic configuration that balances coherence and thematic separation is selected as the baseline model. After this stage, hyperparameters such as α and η are further tuned through a grid search [15] to obtain a more stable and semantically robust topic structure. The final model is then interpreted by examining the most representative terms associated with each topic and by analyzing the emerging thematic patterns, which are subsequently aligned with the dimensions of hotel service quality identified in the literature and the domain context.

The thematic structures identified through this topic modeling procedure form the basis for the subsequent Results and Discussion section. In this section, the experimental findings, topic configurations, and interpretations will be presented and analyzed in comprehensive detail. The discussion focuses on explaining the meaning and relevance of the extracted topics in relation to hotel service quality. These results are used to provide a clearer understanding of customer perceptions derived from the analyzed review data.

III. RESULTS & DISCUSSION

This section presents the experimental results obtained from applying the proposed topic modeling framework to the hotel review dataset and systematically discusses the implications of the identified topics for assessing service quality in Jakarta's hospitality sector. The section is structured to cover corpus reduction and document characteristics, coherence evaluation, model hyperparameter optimization, intertopic distance visualization, and a consolidated topic summary, discussion, and implications.

A. Corpus Reduction And Document Characteristics

The initial dataset consisted of 12,798 hotel customer reviews collected from 13 high-rated properties in Jakarta, as

TABLE I. EXAMPLE OF CUSTOMER REVIEWS FROM ONE OF THE HIGHEST-RATED HOTELS IN JAKARTA

This accommodation truly offers comfort with its strategic location in the heart of the city, providing easy access to various attractions and business centers. In addition, the staff are always friendly and quick to serve the needs of guests, creating a pleasant and warm atmosphere. The facilities provided are comprehensive and varied, ranging from a mini soccer field perfect for leisurely sports activities, a swimming pool for relaxation, to a gym equipped with modern equipment to support your fitness routine. To start the day, the breakfast options available are quite adequate, with a variety of menu choices to suit your tastes and nutritional needs. Despite offering many conveniences and comforts, the prices are still quite affordable, making it an ideal choice for travelers and business people who want comfort without having to spend a lot of money. *(Penginapan ini benar-benar menawarkan kenyamanan dengan lokasinya yang strategis di jantung kota, memberikan kemudahan akses ke berbagai tempat menarik dan pusat bisnis. Selain itu, para stafnya selalu ramah dan cepat dalam melayani kebutuhan para tamu, sehingga menciptakan suasana yang menyenangkan dan hangat. Fasilitas yang disediakan sangat lengkap dan bervariasi, mulai dari lapangan sepak bola mini yang cocok untuk mengisi waktu luang dengan aktivitas olah raga, kolam renang untuk bersantai, hingga gym yang dilengkapi dengan peralatan modern untuk menunjang rutinitas kebugaran. Untuk mengawali hari, pilihan sarapan yang tersedia cukup memadai, dengan beragam pilihan menu yang sesuai dengan selera dan kebutuhan nutrisi Anda. Meski menawarkan banyak kemudahan dan kenyamanan, namun harga yang ditawarkan masih cukup terjangkau sehingga menjadi pilihan ideal bagi para pelancong dan pebisnis yang menginginkan kenyamanan tanpa harus mengeluarkan banyak uang.)*

*Italic text represents the original Indonesian review

detailed in the methodology section. To ensure that only reviews with sufficient lexical content were included in the topic modeling process, a minimum document length threshold of 10 words was applied. This filtering step reduced the corpus to 6,341 reviews, effectively removing extremely short or non-informative entries. This procedure helps ensure that the remaining documents provide meaningful textual information for subsequent topic modeling analysis.

The statistical distribution of document lengths in the filtered corpus is detailed in Table II. The dataset has a mean length of 22.37 words, a median of 16 words, and a maximum of 249 words per review. The interquartile range, with the 25th and 75th percentiles at 13 and 24 words respectively, indicates that the majority of reviews are of moderate length. This characteristic ensures adequate lexical richness while preventing disproportionately long documents from dominating the term distribution. Consequently, the corpus provides a balanced basis for reliable topic extraction, with sufficient variability to capture diverse guest experiences without being skewed by sparse or trivial texts.

B. Coherence Evaluation

To identify an appropriate number of topics for the LDA model, multiple configurations were trained with topic counts ranging from 1 to 15 and evaluated using the coherence score

TABLE II. STATISTICAL SUMMARY OF DOCUMENT LENGTHS AFTER APPLYING A MINIMUM THRESHOLD OF 10 WORDS

Statistic	Value
Count	6,341
Mean	22.37
Std	18.45
Min	10
25%	13
50% (Median)	16
75%	24
Max	249

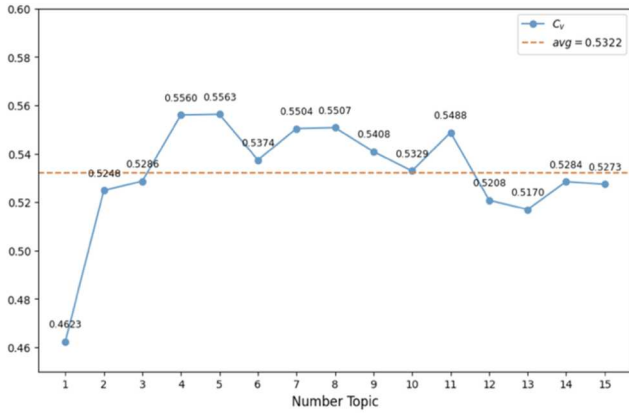


Fig. 2. Coherence scores of LDA models across 1–15 topics, with a dotted line representing the average coherence value.

C_v as the primary indicator of semantic interpretability. The coherence curve in Fig. 2 shows a marked increase in C_v between one and four topics, followed by fluctuating but generally stable values, with topic counts from 4 to 11 remaining above the overall average coherence threshold of 0.5322. Within this range, several configurations achieve competitive coherence levels; however, the seven-topic setting is selected as the baseline because it provides a favorable trade-off between high coherence and sufficient granularity to capture the diversity of service-related themes present in the review corpus.

C. Hyperparameter Optimization

The baseline LDA model, initialized with default hyperparameter settings, achieved a coherence score $C_v = 0.5504$, providing a reasonable but improvable representation of the review corpus, as indicated by its position on the coherence curve in Fig. 2. To enhance topic quality, a grid search was conducted over the α and η parameters with candidate values {0.01, 0.1, symmetric, auto} for each, resulting in sixteen model configurations that were evaluated using the same coherence metric. Among these candidates, the configuration with $\alpha = \text{auto}$ and $\eta = \text{auto}$ produced the highest coherence score, reaching $C_v = 0.6065$, and was therefore selected as the final model for subsequent analysis. All topic-level coherence values reported in Table III correspond to this optimized configuration, reflecting the semantic quality of



Fig. 3. Intertopic distance map of the baseline LDA model with default hyperparameters and coherence score ($C_v = 0.5504$).

TABLE III. THE COHERENCE SCORES FOR EACH TOPIC BASED ON HOSPITALITY REVIEWS IN JAKARTA

Topic	Coherence Score
Topic 1	0.5001
Topic 2	0.6080
Topic 3	0.6479
Topic 4	0.6639
Topic 5	0.6499
Topic 6	0.6881
Topic 7	0.7223

topics under the best-performing model rather than the baseline.

D. Intertopic Distance Visualization

To further assess the quality of the learned topics beyond scalar coherence values, intertopic distance maps were examined to analyze how topics are positioned relative to each other in the latent semantic space. Fig. 3 shows the intertopic distance visualization for the baseline LDA model, trained with default hyperparameter settings and a coherence score of $C_v = 0.5504$, where several topic clusters are positioned relatively close to one another, indicating limited separation among certain themes. After hyperparameter optimization with $\alpha = \text{auto}$ and $\eta = \text{auto}$, the best-performing model ($C_v = 0.6065$) exhibits a noticeably different structure, as illustrated in Fig. 4, with topic clusters that are more compact and clearly separated, suggesting improved thematic distinctiveness.

These visualizations provide an intuitive complement to the coherence evaluation by revealing how topics are distributed and grouped in a two-dimensional projection of the underlying probability space. The increased separation among clusters in the optimized model indicates that the final configuration captures more discriminative patterns in hotel customer feedback, thereby providing a more robust basis for topic summarization and interpretation.

E. Topic Summary, Discussion, And Implications

Based on the optimized LDA configuration, the final model produces seven coherent topics that capture the dominant themes present in the hotel customer reviews. Table

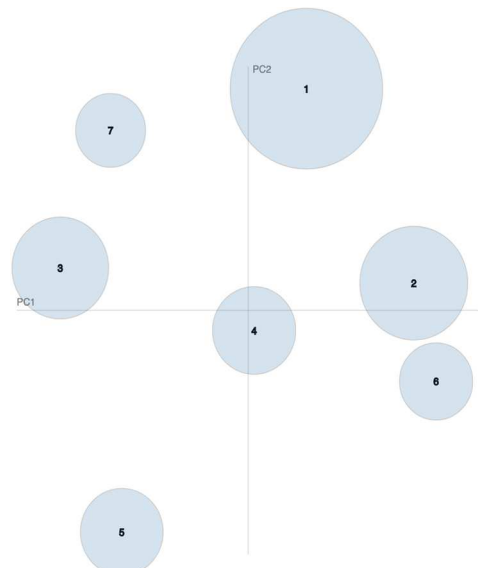


Fig. 4. Intertopic distance map of the optimized LDA model with ($\alpha = \{\text{auto}\}$), ($\eta = \{\text{auto}\}$), and coherence score ($C_v = 0.6065$).

4 summarizes these topics by reporting, for each topic, its proportion of tokens in the corpus, the most representative terms, and a concise interpretation of the underlying service-related dimension. Token share values indicate that Topic 1 accounts for the largest portion of the corpus, followed by Topics 2 and 3. This distribution suggests that certain aspects of the guest experience receive substantially more attention in customer feedback than others, highlighting the relative salience of different service quality dimensions.

Beyond the distribution patterns, the extracted topics reveal several dimensions of hotel service quality reflected in customer reviews. Among these, staff hospitality and breakfast experience (Topic 1) emerge as the most dominant theme based on token share, highlighting the central role of interpersonal service and food quality in shaping positive guest perceptions. In contrast, room cleanliness and bathroom conditions (Topic 2) frequently appear as sources of dissatisfaction, indicating that shortcomings in basic comfort and hygiene can substantially affect the overall stay experience. The remaining topics capture complementary aspects of the guest experience, including location and price-value perception, service responsiveness, check-in procedures, room amenities, and shared facilities. Taken together, the full set of topics provides a concise yet comprehensive overview of the service quality dimensions that characterize customer experiences in Jakarta's hotel market.

From a managerial perspective, the word clusters in Topic 1 indicate that hotel staff performance and breakfast quality are consistently emphasized in customer perceptions of service quality. Accordingly, hotels should prioritize continuous staff training to maintain and enhance customer service skills, while systematically monitoring breakfast quality as a key component of daily service evaluation. In addition, the topic modeling results identified several other topics, particularly Topics 2, 4, and 5, which collectively reflect recurring customer feedback related to cleanliness, service responsiveness, and the check-in process. These findings underscore the need for systematic operational improvements, including stricter housekeeping standards,

clearer front-desk procedures, and improved coordination during peak service hours.

The themes captured in Topics 3, 6, and 7 further highlight the importance of location, room amenities, and shared facilities in shaping guests' value perceptions, implying that hotels in competitive urban markets like Jakarta must balance functional convenience with experiential features to remain attractive. Collectively, these findings demonstrate how topic modeling can translate large volumes of unstructured reviews into actionable service quality insights that align with established dimensions in the hospitality literature while reflecting the specific priorities of local customers.

The final topics indicate that staff hospitality and breakfast experience are among the most frequently emphasized positive aspects in guest feedback. In contrast, issues related to room cleanliness, bathroom condition, check-in processes, and administrative handling consistently emerge as common sources of complaint. Location and accessibility, room amenities, and family-oriented shared facilities also appear as relevant dimensions in guests' evaluations, suggesting that both functional and experiential aspects contribute to perceived service quality. Taken together, these findings provide practical insights for hotel operators to reflect on service areas related to front-desk procedures, housekeeping standards, and food and beverage services, while continuing to maintain strengths associated with staff friendliness and facility provision.

Methodologically, the proposed framework illustrates how topic modeling can be used to complement sentiment analysis by producing more fine-grained and interpretable themes that are aligned with operational service attributes in the hospitality domain. Future work may extend this approach by incorporating reviews from multiple cities, integrating sentiment information within each topic, or comparing alternative topic modeling techniques to further enhance the depth and generalizability of insights derived from large-scale textual feedback.

TABLE IV. TOPIC SUMMARY WITH TOKEN SHARE, KEY TERMS, AND TOPIC INTERPRETATION

Topic Interpretation	Tokens %	Terms and Weights
Staff service quality & breakfast experience (Topic 1)	34.1% of tokens	service (<i>layan</i>): 0.0714, friendly (<i>ramah</i>): 0.0624, good (<i>bagus</i>): 0.0552, eat (<i>makan</i>): 0.0551, staff (<i>staf</i>): 0.0503, tasty (<i>enak</i>): 0.0480, breakfast (<i>sarap</i>): 0.0460, help (<i>bantu</i>): 0.0263, provide (<i>kasih</i>): 0.0167, menu (<i>menu</i>): 0.0162
Room cleanliness, bathroom condition, and comfort issues (Topic 2)	17.1% of tokens	shower (<i>mandi</i>): 0.0529, ac (<i>ac</i>): 0.0518, cigarette (<i>rokok</i>): 0.0336, cold (<i>dingin</i>): 0.0330, odor (<i>bau</i>): 0.0303, hot (<i>panas</i>): 0.0202, parking (<i>parkir</i>): 0.0196, sleep (<i>tidur</i>): 0.0193, towel (<i>handuk</i>): 0.0190, toilet (<i>toilet</i>): 0.0186
Location, accessibility, and price-value perception (Topic 3)	13.7% of tokens	location (<i>lokasi</i>): 0.0836, city (<i>kota</i>): 0.0447, strategic (<i>strategis</i>): 0.0437, mall (<i>mall</i>): 0.0343, price (<i>harga</i>): 0.0337, center (<i>pusat</i>): 0.0328, easy (<i>mudah</i>): 0.0228, reach (<i>jangkau</i>): 0.0217, access (<i>akses</i>): 0.0197, good (<i>bagus</i>): 0.0182
Service responsiveness and guest support issues (Topic 4)	10.1% of tokens	night (<i>malam</i>): 0.0409, full (<i>penuh</i>): 0.0165, need (<i>butuh</i>): 0.0148, really (<i>sungguh</i>): 0.0139, slow (<i>lambat</i>): 0.0134, expect (<i>harap</i>): 0.0133, handle (<i>laku</i>): 0.0122, contact (<i>hubung</i>): 0.0120, food (<i>makan</i>): 0.0119, price (<i>harga</i>): 0.0117
Check-in process, waiting time, front-desk handling (Topic 5)	10% of tokens	hour (<i>jam</i>): 0.0436, check (<i>periksa</i>): 0.0376, receptionist (<i>resepsionis</i>): 0.0250, wait (<i>tunggu</i>): 0.0242, elevator (<i>lift</i>): 0.0221, guest (<i>tamu</i>): 0.0168, bring (<i>bawa</i>): 0.0156, luggage (<i>barang</i>): 0.0127, pick (<i>ambil</i>): 0.0121, morning (<i>pagi</i>): 0.0119
Room amenities and in-room facility availability (Topic 6)	7.8% of tokens	available (<i>sedia</i>): 0.0430, breakfast (<i>sarap</i>): 0.0328, free (<i>gratis</i>): 0.0247, food (<i>makan</i>): 0.0228, television (<i>televisi</i>): 0.0226, complete (<i>lengkap</i>): 0.0208, laundry (<i>laundry</i>): 0.0187, bathroom (<i>mandi</i>): 0.0183, wash (<i>cuci</i>): 0.0181, kitchen (<i>dapur</i>): 0.0179
Hotel shared spaces and family activity areas (Topic 7)	7.2% of tokens	pool (<i>kolam</i>): 0.1424, swimming (<i>renang</i>): 0.1349, child (<i>anak</i>): 0.0737, kunningan (<i>kunningan</i>): 0.0342, gym (<i>gym</i>): 0.0248, fitness (<i>bugar</i>): 0.0223, garden (<i>taman</i>): 0.0188, spacious (<i>luas</i>): 0.0187, family (<i>keluarga</i>): 0.0185, breakfast (<i>sarap</i>): 0.0168

*Italic text represents the original Indonesian review

IV. CONCLUSION

This study applies the LDA model to 12,798 reviews from highly rated hotels in Jakarta and identifies seven service-quality themes that shape overall guest experiences. The extracted topics reveal key dimensions of customer perception, including staff service quality and breakfast experience; room cleanliness, bathroom condition, and comfort issues; location and price-value considerations; service responsiveness and guest support; check-in procedures and front-desk performance; room amenities; and shared hotel facilities. The results indicate that staff friendliness and breakfast quality are consistently associated with positive guest evaluations, while issues related to cleanliness, bathroom condition, responsiveness, and check-in processes frequently appear in negative feedback. Overall, these results illustrate the capability of LDA to convert large volumes of unstructured review text into interpretable insights that support service quality evaluation in Jakarta's hospitality sector. Future research may integrate sentiment analysis to quantify topic-level guest attitudes and expand the dataset to broader geographical contexts for comparative analysis.

REFERENCES

- [1] D. Fauzi and P. A. Sumirat, "Evaluation of Indonesian Tourism Marketing towards Quality Tourism," *Int. J. Magistravitae Manag.*, vol. 1, no. 2, pp. 117–129, 2023, doi: 10.33019/ijomm.v1i2.22.
- [2] S. Wei and H. Kim, "Online Customer Reviews and Satisfaction with an Upscale Hotel: A Case Study of Atlantis, The Palm in Dubai," *Information*, vol. 13, no. 3, p. 150, 2022, [Online]. Available: <https://ijomm.ubb.ac.id/index.php/ijomm/article/view/22/10>.
- [3] C. Andreou, N. Tsapatsoulis, and V. Anastasopoulou, "A dataset of hotel reviews for aspect-based sentiment analysis and topic modeling," in *2023 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP) 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP 2023)*, 2023, pp. 1–9, doi: 10.1109/SMAP59435.2023.10255219.
- [4] A. Afaq, L. Gaur, and G. Singh, "A Latent Dirichlet Allocation Technique for Opinion Mining of Online Reviews of Global Chain Hotels," in *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, 2022, pp. 201–206, doi: 10.1109/ICIEM54221.2022.9853114.
- [5] S. Juanita, D. Purwitasari, I. K. E. Purnama, and M. H. Purnomo, "Identifying Precautionary Measures for High-Risk Disease from Doctor's Answer Text Using LDA," in *2021 International Electronics Symposium (IES)*, 2021, pp. 41–46, [Online]. Available: <https://ieeexplore.ieee.org/document/9593997/>.
- [6] Y. Li, X. Li, Y. Zhai, D. Wang, and C. Hon, "Misinformation Features Detection in Weibo: Unsupervised Learning, Latent Dirichlet Allocation, and Network Structure," *IEEE Access*, vol. 12, pp. 166977–166987, 2024, doi: 10.1109/ACCESS.2024.3494015.
- [7] T. Ali, B. Omar, and K. Soulaimeane, "Analyzing tourism reviews using an LDA topic-based sentiment analysis approach," *MethodsX*, vol. 9, no. 101894, pp. 1–10, 2022, doi: 10.1016/j.mex.2022.101894.
- [8] Y. Nakamura and N. Oomiya, "Applying the topic model to hotel reviews of internet sites and analyzing their results," *J. Glob. Tour. Res.*, vol. 6, no. 1, pp. 15–20, 2021, [Online]. Available: https://www.jstage.jst.go.jp/article/jgtr/6/1/6_15/_pdf-char/ja.
- [9] E. Erniyati, P. Harsani, M. Mulyati, and L. D. Fahriza, "Topic Modeling LDA and SVM in Sentiment Analysis of Hotel Reviews," *Komputasi J. Ilm. Komput. dan Mat.*, vol. 20, no. 2, pp. 93–100, 2023, [Online]. Available: <https://journal.unpak.ac.id/index.php/komputasi/article/view/7604>.
- [10] Y. A. Singgalan, "Understanding Hotel Customer Experience through User-Generated Reviews using Knowledge Discovery in Databases (KDD)," *J. Comput. Syst. Informatics*, vol. 6, no. 1, pp. 48–63, 2024, doi: 10.47065/josyc.v6i1.6014.
- [11] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, and O. Patlan, "Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications," in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, 2020, pp. 187–191, doi: 10.1109/PICST51311.2020.9467919.
- [12] S. Ying, "Guests' Aesthetic experience with lifestyle hotels: An application of LDA topic modelling analysis," *Heliyon*, vol. 10, no. e35894, pp. 1–12, 2024, doi: 10.1016/j.heliyon.2024.e35894.
- [13] W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An Overview of Bag of Words: Importance, Implementation, Applications, and Challenges," in *Proceedings of the 5th International Engineering Conference, IEC 2019*, 2019, pp. 200–204, doi: 10.1109/IEC47844.2019.8950616.
- [14] L. Wang *et al.*, "GENSIM: Generating Robotic Simulation Tasks Via Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2024, pp. 1–36, [Online]. Available: <https://openreview.net/forum?id=OI3RoHoWAN>.
- [15] H. Alibrahim and S. A. Ludwig, "Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization," in *2021 IEEE Congress on Evolutionary Computation (CEC)*, 2021, pp. 1551–1559, doi: 10.1109/CEC45853.2021.9504761.

Certificate

This Certificate of Appreciation is proudly presented to

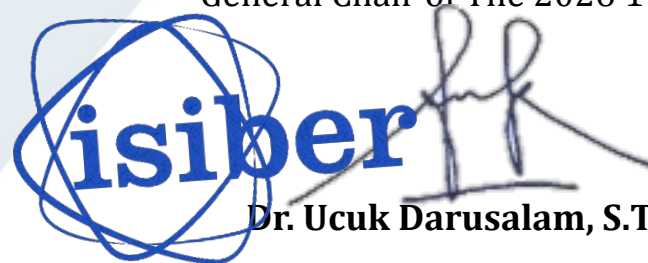
Tegar Mandiri and Safitri Juanita (Universitas Budi Luhur, Indonesia)

as the AUTHORS of the paper entitled:

Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA

at The 2026 1st International Seminar on Intelligent Business and Edge-Computing Research (ISIBER)
held on 26 February 2026 at Universitas Siber Asia, Jakarta, Indonesia.

General Chair of The 2026 1st ISIBER,

The image shows the ISIBER logo, which consists of the word "isiber" in a blue, lowercase, sans-serif font. To the right of the logo is a handwritten signature in blue ink. The signature appears to be "Ucuk Darusalam".

Dr. Ucuk Darusalam, S.T., M.T.

Crt.1571232430



 26 February
2026



isiber.unsia.ac.id