Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1653~1668
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.10709

Recency, frequency, quality: novel feature from sentiment analysis for clustering and ranking in tourism big data analytics

Ni Wayan Sumartini Saraswati[1], I Ketut Gede Darma Putra[2], Made Sudarma[3], I Made Sukarsa[2], I Gusti Ayu Agung Mas Aristamy[4]
[1]Master of Informatics, Indonesian Institute of Business and Technology in Denpasar Bali, Denpasar, Indonesia
[2]Department of Information Technology, Faculty of Engineering, Udayana University in Denpasar Bali, Denpasar, Indonesia
[3]Department of Engineering Science, Faculty of Engineering, Udayana University in Denpasar Bali, Denpasar, Indonesia
[4]Department of Informatics, Faculty of Technology and Informatics, Indonesian Institute of Business and Technology in Denpasar Bali, Denpasar, Indonesia

Article Info

Article history:

Received May 19, 2025
Revised Feb 23, 2026
Accepted Mar 5, 2026

Keywords:

Big data analytics
Clustering
New feature
Recency, frequency, quality
Sentiment analysis

ABSTRACT

Understanding tourist perceptions has been a key benefit of sentiment analysis in tourism data. However, its outcomes can be further utilized to gain insights into the characteristics of tourist attractions and hotels. This study aims to develop a new feature, called recency, frequency, quality (RFQ), derived from sentiment analysis results to cluster and rank tourist attractions and hotels in Bali. RFQ consists of three components: review recency, review frequency, and review quality. These dimensions reflect the recentness of reviews, the popularity based on the number of reviews, and the review quality measured by the ratio of positive to negative sentiment polarity. Using big data analytics through clustering and ranking, the study finds that the quality of tourist attractions and hotels is primarily concentrated in Badung and Gianyar regencies. More tourist attractions are found in the silver cluster than in the gold, indicating the need to enhance quality. In the hotel sector, the diamond cluster dominates among star-rated hotels, suggesting overall high quality. Budget hotels show fairly good quality, with most falling under the gold cluster.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Ni Wayan Sumartini Saraswati
Master of Informatics, Indonesian Institute of Business and Technology in Denpasar Bali
Tukad Pakerisan Street No. 97, Denpasar, Bali, Indonesia
Email: sumartini.saraswati@instiki.ac.id

1. INTRODUCTION

Sentiment analysis is a technique within natural language processing (NLP) that is used to analyze and identify the sentiment or emotion contained in a given text [1]. Sentiments can be classified as positive, negative, or neutral. More broadly, sentiment analysis helps identify feelings, emotions, or opinions in text, with applications such as analyzing customer feedback [2], brand monitoring [3], market research [4], reputation management [5], political analysis [6], [7], product development [8], [9], and efforts to enhance customer service [10]. Generally, sentiment analysis is conducted using approaches such as lexicon-based methods, machine learning, hybrid method, or deep learning techniques [11], [12].

Previous research applied sentiment analysis to tourism big data to explore travelers' perceptions of Bali's attractions and hotels. We developed a novel hybrid sentiment analysis method called LeALSVM [13], which has been proven to perform exceptionally well in analyzing unlabeled big data. Subsequent studies utilizing the LeALSVM method were conducted to explore tourist perceptions of temple attractions in Bali, as well as to understand the strengths and challenges of hotels in Bali along with the characteristics of their guests.

The insights gained from these studies contributed to the formulation of priority scales for stakeholders to maintain Bali's tourism image in support of sustainable tourism. However, it would be unfortunate if the outcomes of sentiment analysis on tourism big data were limited solely to understanding tourist perceptions of attractions and hotels in Bali. It is worth exploring whether these results can also generate other valuable insights that may further contribute to the development of Bali's tourism sector.

Various studies have explored the development of recommendation systems from sentiment analysis within the tourism industry. A study conducted in China [14] analyzed reviews of 5A-rated tourist attractions using the ROST tool to calculate a new multi-dimensional ranking method and proposes an online evaluation ranking based on PLTSs and the IDOCRIW-COCOSO model. In this research the percentage of sentiment orientation for ranking is composed of the percentage of positive, neutral, and negative sentiments. Although this study conducted rankings based on sentiment analysis results, the rankings do not reflect the recency of reviews and the number of reviews received. Another study utilized aspect-based sentiment analysis to compute sentiment orientation and applied it within MCDM and IHF-TOPSIS methods to rank tourist destinations [15]. In this study, key alternative features were extracted by using tokenization and POS tagging on online reviews. A fuzzy evaluation matrix on several alternatives with several different significant features was obtained according to the subjective scoring from related domain experts. Finally, the ranking of alternatives was determined by the fuzzy PROMETHEE method. Although it has the advantage of carrying out rankings based on the review aspects of tourist attractions, the same ranking does not take into account the recency of reviews and the frequency of reviews which describe the popularity of tourist attractions. Research on hotel recommendation systems [16] employed the BERT model for sentiment analysis based on aspects such as cleanliness and service. The results were then combined with textual features (Word2Vec and TF-IDF) and classified using the random forest algorithm and cosine similarity with fuzzy logic. This research succeeded in ranking and recommending hotels only based on the suitability of the aspect categories but also did not consider the recency, frequency and quality of the reviews of the hotel.

In light of these prior studies that ranked the tourist attractions and hotel, this research seeks to examine how sentiment analysis derived from tourism big data can yield deeper insights into the characteristics of tourist attractions and hotels in Bali, as well as inform their ranking. To this end, it is essential to develop a set of features constructed from the sentiment analysis outcomes. As these features are based on in visitor perceptions, the review data must be quantified to serve as input variables for clustering and ranking models. The clustering of tourist attractions and hotels aims to generate insights into the typologies formed based on the perceived quality of these destinations and accommodations, as expressed by tourists visiting Bali.

Key contributions: this study aims to develop a novel feature derived from review polarity to address the aforementioned issue. We refer to this new feature as the recency, frequency, quality (RFQ) feature. To demonstrate that this feature is capable of providing meaningful insights into the characteristics and ranking of tourist attractions and hotels in Bali, we employed the k-means clustering method. K-means has been shown to be effective in producing reliable clustering models [17]–[19]. Moreover, the clustering model is intended to generate insights into the characteristics of tourist attractions and hotels based on tourist perceptions, which constitutes the subsequent objective of this research. To support empirical validation in the development of this new feature, clustering and ranking were carried out on TripAdvisor review big data.

The clustering and ranking of tourist attractions serve as essential components in optimizing destination management, marketing strategies, and development planning. Clustering facilitates the identification of potential and the distribution of destination strengths, while ranking offers valuable insights for both tourists and destination managers regarding the relative popularity and quality of attractions. High-ranking destinations may be prioritized for continued support and promotion, whereas those with untapped potential but lower rankings can be strategically targeted for enhancement. Collectively, these approaches inform strategic decision-making, improve destination competitiveness, and foster sustainable and inclusive tourism development.

2. METHOD

The big data analytics model developed for clustering and ranking based on sentiment analysis is illustrated in Figure 1. The big data utilized in this study were derived from the list of tourist attractions and hotels obtained from the Bali Provincial Tourism Office. The dataset comprises 381,278 records. The RFQ feature was constructed from the results of sentiment analysis using the LeALSVM method [13]. This method is a hybrid of lexicon-based sentiment analysis and active learning-support vector machine (AL-SVM) to address the problem of preparing training data for big data. The Lexicon method is used to prepare training data for the AL-SVM model. LeALSVM is a pool-based AL method that allows the model to use a small amount of training data but obtains good classification accuracy results. This method divides unlabelled data into several groups and executes them sequentially. The LeALSVM method has been proven to provide good classification performance, achieving an accuracy of 95.8%. This method performed case folding, tokenization,

stopword removal, and lemmatization on the review data before applying the sentiment analysis model. The RFQ feature was subsequently employed as an input variable for the clustering model using the k-means algorithm and for the ranking model. In the final stage, insights into Bali's tourism landscape were generated based on the outputs of both models.

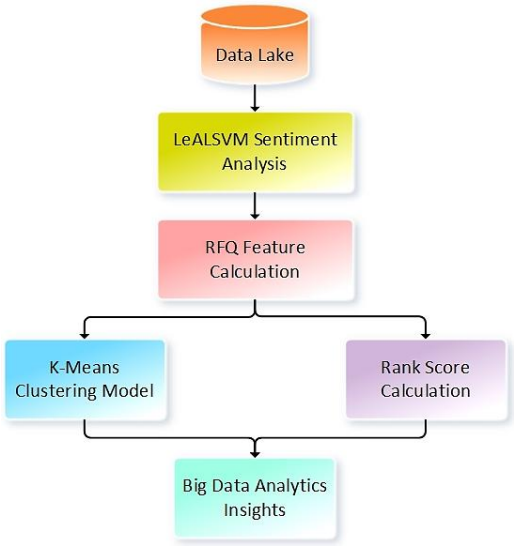


Figure 1. Big data analytics flow

The clustering and ranking model based on review data was constructed using variables developed through the RFQ feature approach. A summary description of the RFQ features is presented in Table 1. The review recency reflects the recentness of reviews. The review frequency representing the level of popularity of each entity. Meanwhile, the review quality representing the perceived quality and satisfaction levels expressed by tourists. The main goal of clustering is to find patterns or groups that share similar characteristics. The features selected must meaningfully represent differences between objects. The quality of clustering results is greatly influenced by feature relevance. The theory states that representative features improve the interpretability of clusters. Recency, based on consumer behavior theory recency, frequency, monetary (RFM) model, suggests that the more recent an interaction, the higher the relevance or activity of the object. By incorporating recency, clusters can distinguish between active and inactive objects, thus helping to understand the temporal dynamics of behavior. In this research, recency also has a similar meaning, only the difference lies in the fact that recent reviews represent hotel objects or tourist attractions that have more recent interest (current popularity). Frequency describes the intensity of an activity or interaction. In user/customer behavior theory RFM, frequency is an indicator of user engagement and loyalty. In RFQ, the difference is that frequency represents how popular the tourist attraction or hotel is. Recency, frequency, amount (RFA) is an adaptation of RFM theory for contexts where "monetary" is irrelevant or is more appropriately replaced by other indicators such as "average value" for example average expenditure. Quality is considered a crucial dimension influencing perception and satisfaction. By adding the quality dimension, clusters differentiate not only the intensity and duration of interactions but also the perceived value/assessment of those interactions, resulting in richer and more meaningful cluster results. This is the third differentiator of RFQ features from RFM and RFA.

Table 1. A summary description of the RFQ features

Features	Features description
RECENCY	The review recency was derived by calculating the number of days between the most recent review date in the dataset and the latest review date for each tourist attraction and hotel.
FREQUENCY	The review frequency was constructed by calculating the frequency of review occurrences for each object relative to the average number of reviews received by all tourist attractions and hotels.
QUALITY	The review quality was calculated by determining the ratio of positive reviews to negative reviews for each tourist attraction and hotel.

For each variable used in the clustering and ranking process, standardization or z-score normalization was applied using a standard scaler in order to ensure that all variables have the same value range [20], [21]. The standard scaler is a scaling technique in which the values within a column are transformed to exhibit the

characteristics of a standard Gaussian distribution. It standardizes features by subtracting the mean and scaling to unit variance [22]. Unit variance implies dividing all values by the standard deviation. This operation is performed independently for each feature. The primary reason for using feature scaling in clustering is because clustering algorithms are highly sensitive to differences in scale between features. Further reasons include avoiding the dominance of large-scale features, improving the accuracy of distance calculations, and facilitating faster convergence of centroid calculations.

PCA is particularly effective for visualizing and exploring high-dimensional data sets, or data with many features, because it can easily identify trends, patterns, or outliers. The data we used had limited features, resulting in a small dimensionality. This consideration led us to avoid using PCA in our RFQ feature analysis.

Two embedding-based clustering methods in tourism, simple yet efficient scalable multi-view tensor clustering and embedding-induced graph refinement clustering network, have the potential to cluster tourist destinations. However, clustering is based on geographic proximity, similarity of reviews, and similarity of facilities. This differs from the clustering in this study, which grouped based on tourist perceptions using the RFQ feature.

The clustering and ranking models for star-rated hotels and budget hotels were developed separately to obtain more accurate and context-specific insights. Specifically, clustering for both star-rated and budget hotels was performed using the variables of number of rooms, tourist-assigned ratings, review frequency, review recency, and review quality. In contrast, clustering for tourist attractions was based on rating and the RFQ variables, as presented in Table 2.

Table 2. Variables used for clustering

Object	Variables for clustering
Tourist destinations	Rating and RFQ
Star hotels and budget hotels	Rating, RFQ, and number of rooms

2.1. The clustering model

The methods employed to determine the optimal number of clusters for tourist attractions and hotels include the elbow method, silhouette coefficient, and silhouette score. These variables were calculated based on clustering variations ranging from two to six cluster centers. Additional considerations were made by interpreting the silhouette analysis plots for each clustering variation. Certain cluster quantities were deemed suboptimal for specific datasets due to the presence of clusters with below-average silhouette scores and significant fluctuations in silhouette plot sizes. Consequently, a more nuanced and ambivalent silhouette analysis was required to ensure a reliable determination of the optimal number of clusters.

The k-means clustering model was evaluated using the silhouette score. The silhouette score measures how well each data point fits within its assigned cluster compared to the nearest neighbouring cluster [23]. This score ranges from -1 to 1, with higher values indicating better-defined clustering structures [24]. A silhouette coefficient close to +1 suggests that the sample is far from neigh boring clusters, indicating appropriate clustering. A value of 0 implies that the sample lies on or very near the decision boundary between two adjacent clusters, while negative values indicate that the sample may have been incorrectly assigned to a cluster [25].

2.2. The rank score model

The rank score was developed by machine learning model. A higher rank score indicates a better perceived quality of tourist attractions and hotels from the perspective of visitors. Calibration of ranking weights using regression models such as linear regression showed a low R-square value of 0.3543 for star hotels. Researchers assumed that the RFQ and ranking data were not time series, making this model less suitable for use. Therefore, the ranking approach used random forest regression, which was not demonstrated for solving time series prediction. Furthermore, random forest regression can capture complex relationships and indicate feature importance, or the contribution of each feature, which can be used as an alternative weighting approach for ranking.

The ranking model development stage began with normalization for review RFQ and rating features, so that all features were within the same range (0-1). The review rating feature served as an output feature. The recency feature, in particular, was transformed into a negative form before entering the model. This was done because conceptually, high quality, and frequency values are associated with low recency values. In other words, this transformation implements a negative correlation between recency and the target variable. After the random forest regression model was completed, it was evaluated using R-square, RMSE, MAE, and MSE, and shows an R-squared of 0.9243 for star hotels. To get the contribution of each feature, model feature importances is called as shown by Table 3.

Table 3. Feature contribution weight

Domain	Feature importances		
	R	F	Q
Star hotel	0.32	0.16	0.52
Budget hotel	0.32	0.24	0.45
Tourist attraction	0.29	0.31	0.40

3. RESULTS AND DISCUSSION

The clustering process for tourist attractions, star-rated hotels, and budget hotels was conducted independently for each respective domain group. Based on the analysis using the elbow method, silhouette coefficient, and silhouette score, two clusters were identified for tourist attractions, three clusters for star-rated hotels, and three clusters for budget hotels. The following sections present the clustering analysis for each of these domain groups.

3.1. Clustering of tourist attractions

The elbow graph Figure 2 indicates an optimal point at K=3; however, the silhouette coefficient analysis Figures 3(a)-(e) suggests that K=2 Figure 3(a) yields a more balanced cluster distribution. Although the silhouette scores for K=2 and K=3 Table 4 do not differ significantly, the more proportional cluster structure observed at K=2 makes it the most appropriate choice for grouping the tourist attraction data.

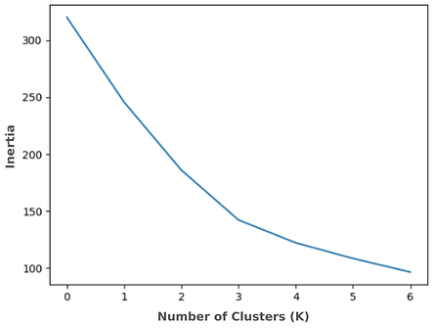


Figure 2. Elbow k-means graph for tourist attraction

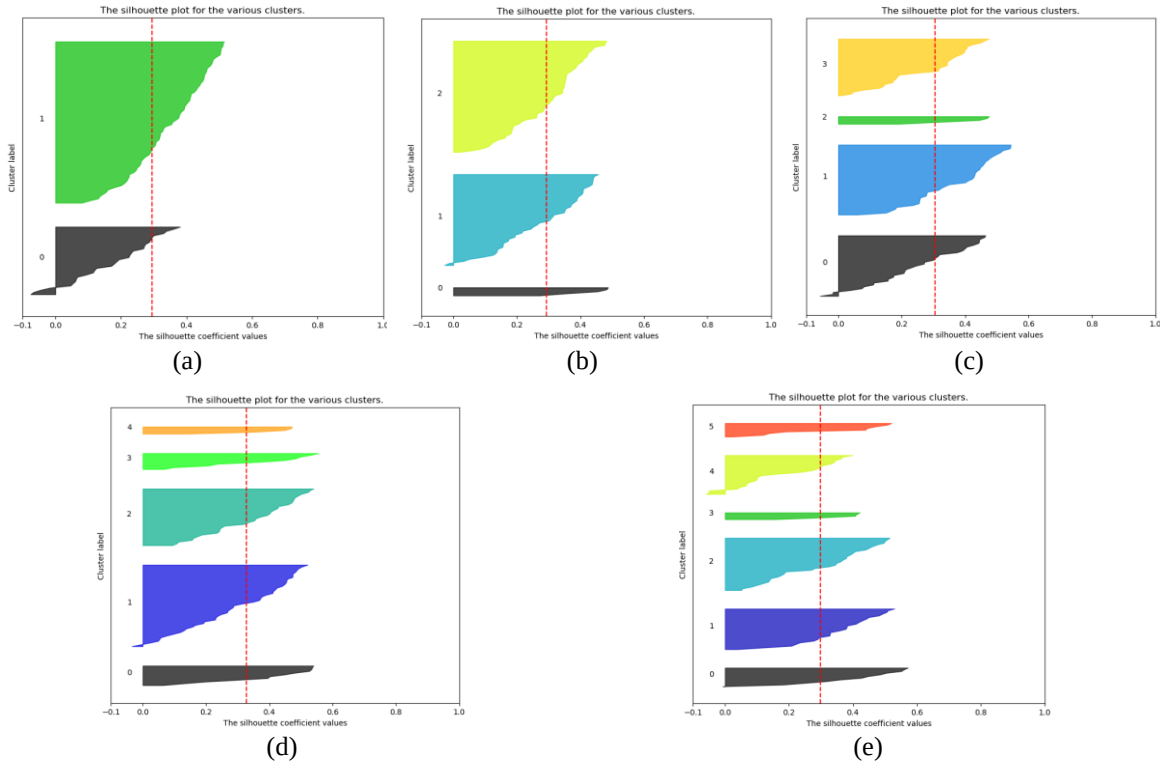


Figure 3. Silhouette coefficient of tourist attraction with; (a) K=2, (b) K=3, (c) K=4, (d) K=5, and (e) K=6

Table 4. Silhouette score tourist attractions

Number of tourist attraction clusters	Silhouette score
K2	0.294305952
K3	0.293640539
K4	0.30861775
K5	0.327275462
K6	0.307059733

Based on the boxplot analysis Figures 4(a)-(d), tourist attractions in Bali are categorized into two clusters: gold (cluster 1) and silver (cluster 0). The gold cluster is characterized by higher ratings, more recent and frequent reviews, and better review quality. In contrast, the silver cluster exhibits lower performance across all of these indicators.

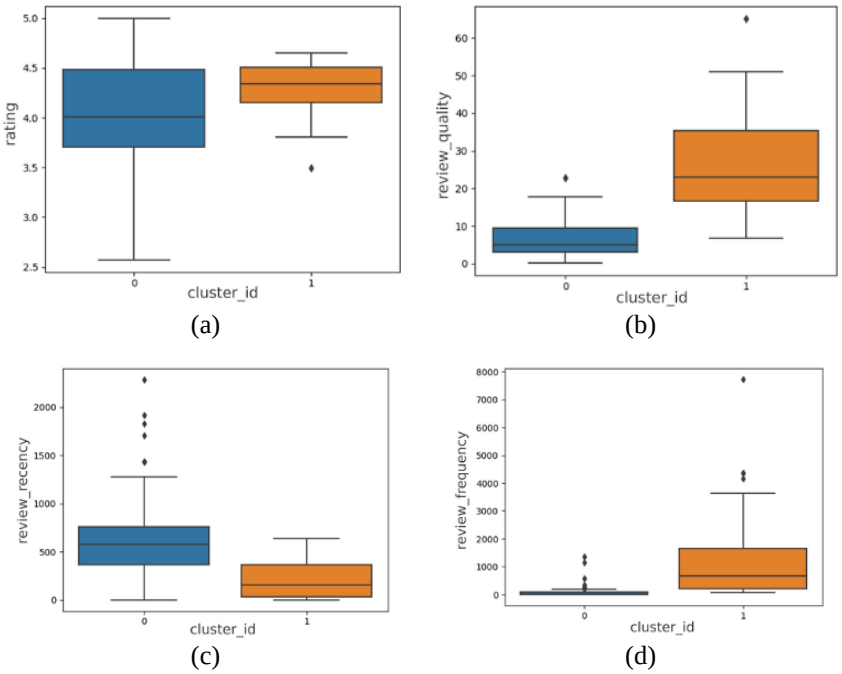


Figure 4. Cluster analysis of; (a) rating, (b) review quality, (c) review recency, and (d) review frequency of tourist attractions

3.2. Clustering of star hotels

Based on the elbow graph Figure 5, the initial optimal point was identified at K=4. However, the silhouette coefficient analysis Figures 6(a)–(e) indicates that K=3 Figure 6(b) produces a more balanced cluster distribution, avoiding the presence of disproportionately small clusters as observed at K=4. Although the silhouette scores for K=3 and K=4, Table 5 are relatively similar, the improved cluster balance makes K=3 the optimal choice for clustering star-rated hotels.

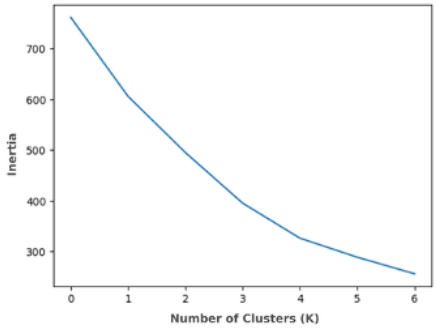


Figure 5. Elbow k-means chart for star hotels

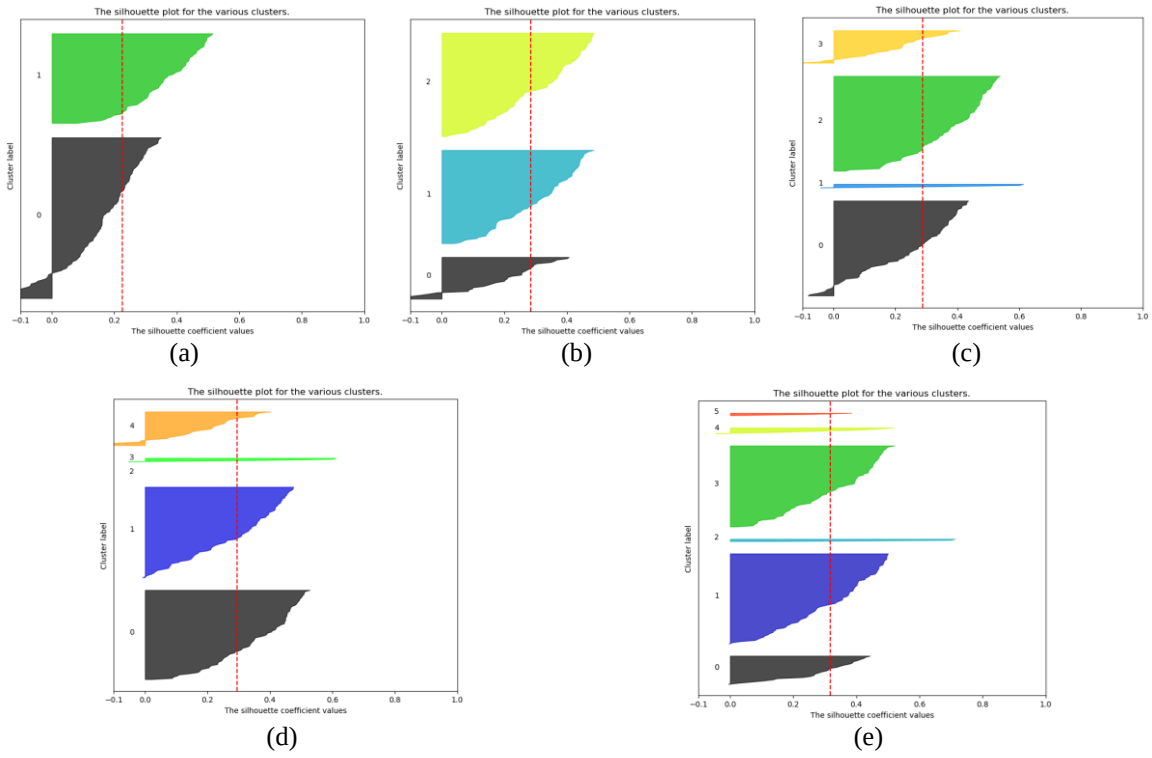


Figure 6. Silhouette coefficient of star hotels with (a) K=2, (b) K=3, (c) K=4, (d) K=5, and (e) K=6

Table 5. Star hotel score silhouette

Number of star hotel clusters	Silhouette score
K2	0.225097506
K3	0.284424787
K4	0.29941797
K5	0.29706672
K6	0.315589175

Based on the boxplot analysis Figures 7(a)–(e) (in Appendix), star hotels in Bali are grouped into 3 clusters: diamond (cluster 2), gold (cluster 0), and silver (cluster 1). The diamond cluster ranks highest ratings and review quality, while gold cluster leads in review frequency and number of rooms. The silver cluster demonstrates the lowest performance across nearly all indicators (ratings, review quality, review recency). In contrast, the gold cluster records the most recent reviews, with the diamond cluster occupying a middle position.

3.3. Clustering of budget hotels

Based on the elbow graph Figure 8, the optimal point was identified at K=3. Further analysis using the silhouette coefficient graph Figures 9(a)–(e) indicates that K=3 Figure 9(b) provides a more balanced cluster distribution compared to K=2 and K=4, although its silhouette score is slightly lower than that of K=4 Table 6. Taking into account both cluster balance and validity, K=3 was selected as the optimal number of clusters for budget hotels.

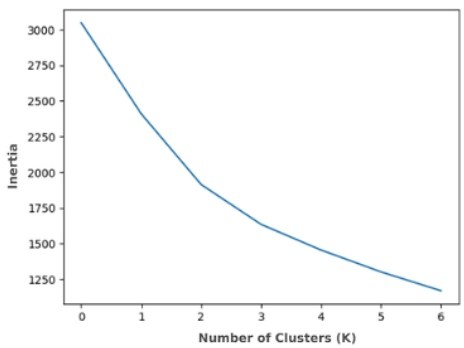


Figure 8. Elbow graph of k-means clustering for budget hotels

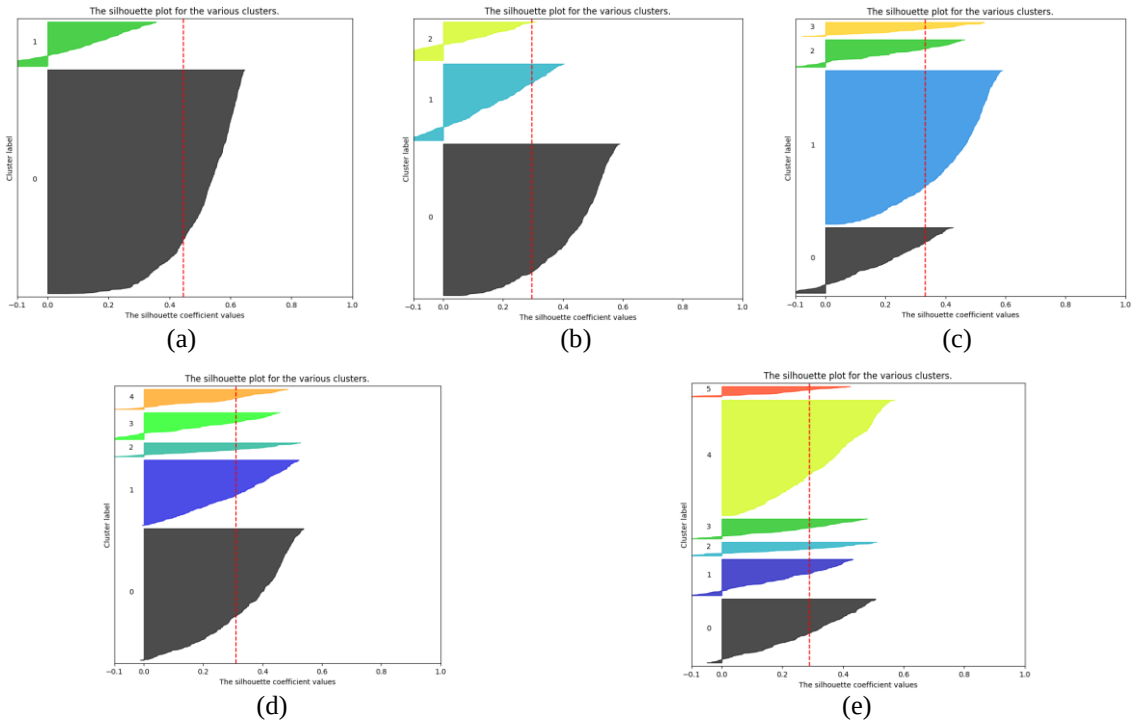


Figure 9. Silhouette coefficient of budget hotels with; (a) K=2, (b) K=3, (c) K=4, (d) K=5, and (e) K=6

Table 6. Budget hotels silhouette score

Number of budget hotel clusters	Silhouette score
K2	0.444721418
K3	0.296521713
K4	0.33028122
K5	0.303642742
K6	0.29610088

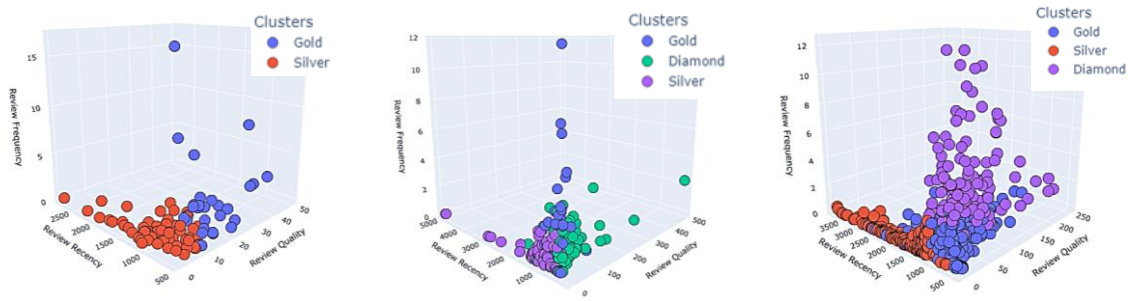
The clustering results for budget hotels produced three groups, categorized as silver (cluster 0), gold (cluster 2), and diamond (cluster 1), based on the boxplot analysis Figures 10(a)-(e) (in Appendix). The Silver cluster is characterized by the lowest ratings, review quality, and review frequency, as well as the oldest review recency. In contrast, the diamond cluster demonstrates the best performance across all of these indicators. The gold cluster falls between the two, with characteristics closer to the silver cluster in terms of the number of rooms but showing better performance in other aspects.

3.4. Clustering method comparison

The clustering results using k-means are visualized in a 3D plot based on the RFQ features to better illustrate cluster distribution and the influence of these features Figures 11(a)–(c). Interestingly, star hotels in the Diamond cluster do not exhibit the highest review frequency, which contrasts with budget hotels in the same cluster.

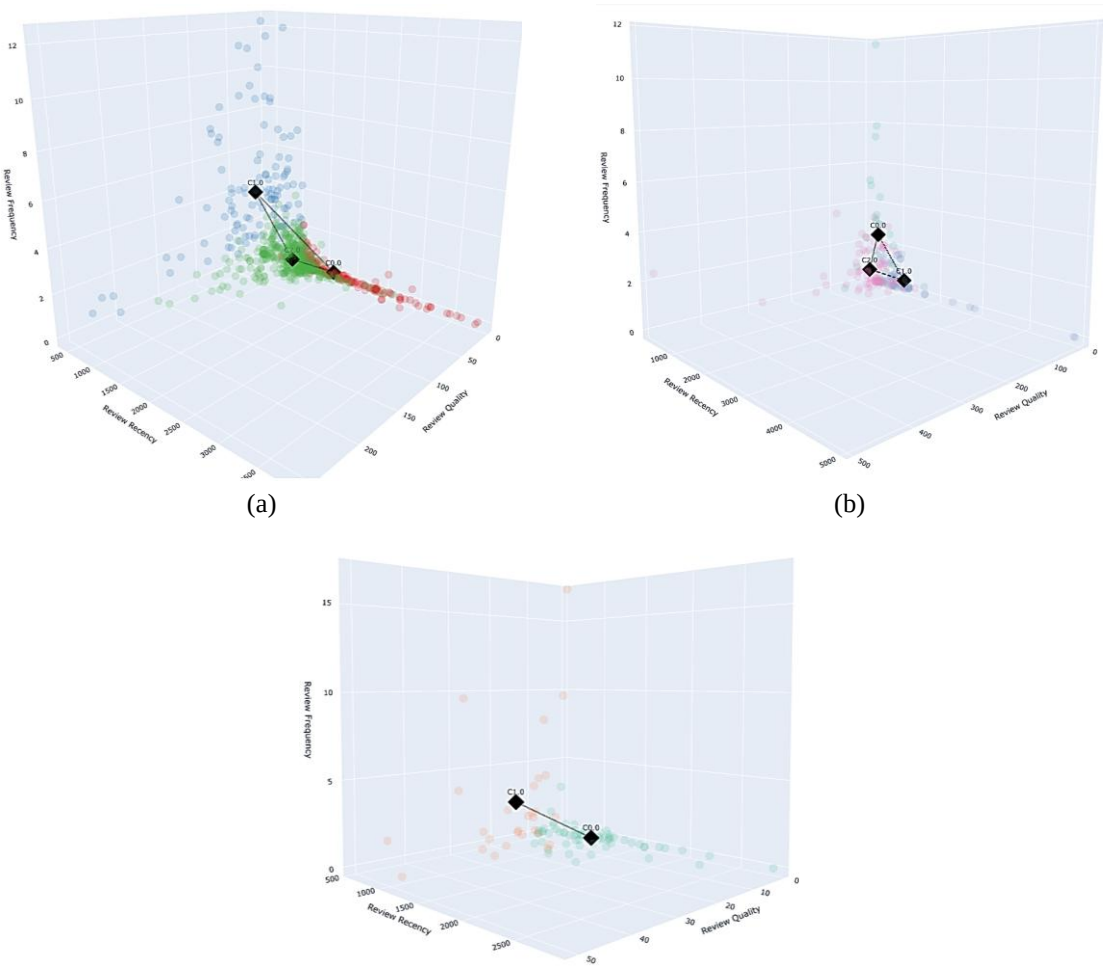
Figures 12(a)–(c) and Table 7 present the cluster centroids and distances between cluster, indicating the quality of separation. In budget hotels, the largest distance between clusters is shown between clusters 0 and 1, which indicates a fairly strong separation, while the smallest distance is found in clusters 1 and 2, so these two clusters are quite close. In star hotels, the shortest distance is between clusters 0 and 2, indicating their close proximity, while the distances between clusters 0–1 and 1–2 are quite large. In tourist attractions, only two clusters are identified (distance 376.71), suggesting moderate separation without further comparison.

K-means outperforms other clustering methods because it is fast, simple, and effective for large datasets with clear, spherical clusters. To validate its use in implementing the RFQ feature for clustering tourist attractions and hotels, we compared its results with those of DBSCAN and Hierarchical Clustering. The comparison results are presented in 3D graphs, in Figures 13(a) and (b) and 14(a)-(c). DBSCAN proves overly sensitive to outliers (cluster –1), resulting in unbalanced clustering results. This method, using RFQ features, can potentially lose the characteristics of the groups formed based on the purpose of this feature. Meanwhile, the results of the hierarchical clustering method provide clustering results similar to the k-means method, with a better k-means silhouette score. This analysis is the reason for choosing k-means for implementing RFQ feature clustering in this study.



(a) (b) (c)

Figure 11. 3D plot of k-means results for RFQ; (a) tourist attractions, (b) star hotels, and (c) budget hotels



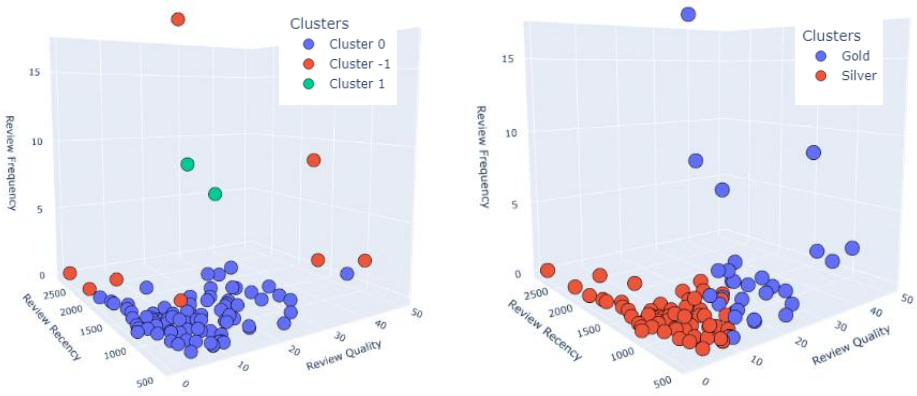
(a) (b)

(c)

Figure 12. Cluster centroid visualization; (a) budget hotel, (b) star hotel, and (c) tourist attraction

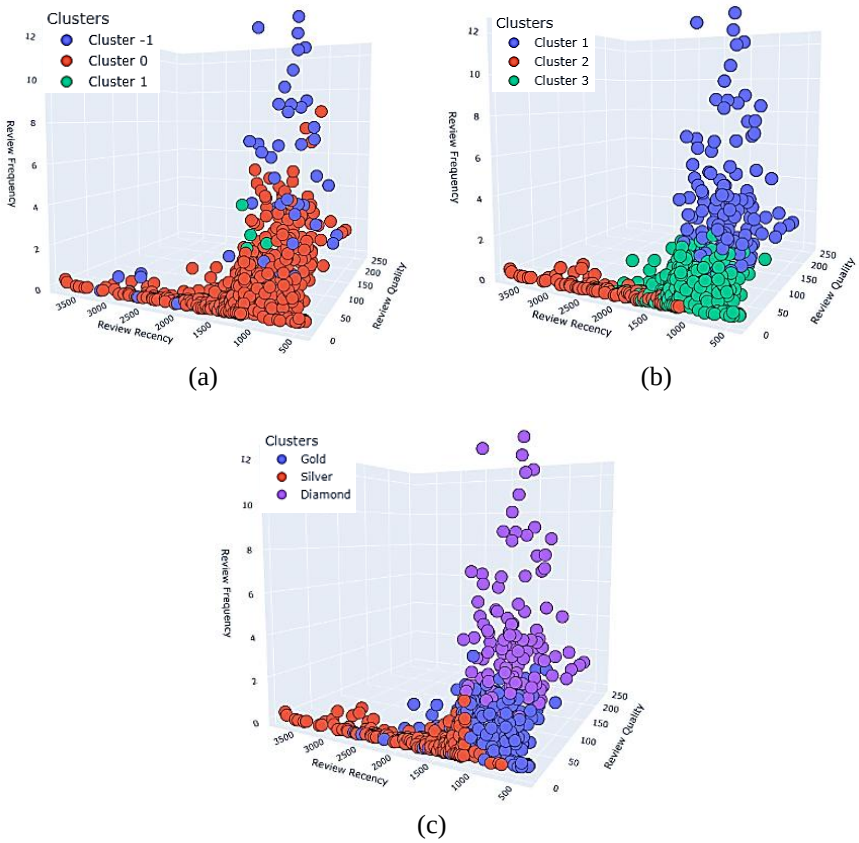
Table 7. Cluster centroid distance

Domain	Cluster pair	Distance
Budget hotel	0–1	872.31
	0–2	552.84
	1–2	319.88
Star hotel	0–1	600.15
	0–2	84.33
	1–2	523.64
Tourist attraction	0–1	376.71



(a) (b)

Figure 13. RFQ plot for tourist attraction; (a) DBScan (K=2) and (b) k-means (K=3)



(a) (b)

(c)

Figure 14. RFQ plot budget hotel; (a) DBScan (K=2), (b) hierarchical clustering (K=3), and (c) k-means (K=3)

3.5. RFQ feature analysis: ANOVA test, ablation studies, and correlation analysis

ANOVA tests whether there are significant differences between clusters on certain features/variables. Table 8 presents the ANOVA results for the RFQ features. Statistical validation using ANOVA showed that all RFQ variables showed significant differences between clusters (p-value <0.05). This indicates that cluster formation was not random but rather reflected a real separation based on review characteristics. Therefore, the clustering results can be considered statistically valid.

Ablation studies are used to determine the contribution of each RFQ component individually to clustering accuracy. Table 9 shows that each individual feature is sufficient to robustly capture the cluster pattern. This could be due to the combination of features is not synergistic, meaning the clusters from each

feature stand alone and do not support each other. This could indicate that the features capture different behavioral dimensions, rather than complementary dimensions. Although the combined features yield a slightly lower silhouette score, they produce more meaningful, comprehensive, and applicable clusters. Considering the modest decrease in silhouette score, the combination of RFQ features remains an option for gaining insights into hotel and tourist attraction clusters.

Table 8. ANOVA test result on RFQ

Domain	p-value		
	F	R	Q
Star hotel	8.01×10^{-169}	5.64×10^{-75}	1.27×10^{-49}
Budget hotel	1.42×10^{-12}	3.54×10^{-11}	1.92×10^{-8}
Tourist attraction	5.42×10^{-7}	8.78×10^{-5}	7.89×10^{-11}

Table 9. Ablation studies result for RFQ

Domain	Silhouette score						
	R	F	Q	RF	RQ	FQ	RFQ
Star hotel	0.388	0.354	0.384	0.307	0.340	0.315	0.284
Budget hotel	0.409	0.408	0.361	0.349	0.315	0.353	0.297
Tourist attraction	0.357	0.448	0.437	0.288	0.285	0.401	0.294

Figures 15(a)–(c) present the correlation analysis of RFQ variables. Across all domains, a consistent pattern emerges: frequency and recency, as well as quality and recency, are negatively correlated, indicating that the more frequently or better the reviews, the shorter the interval between reviews tends to be.

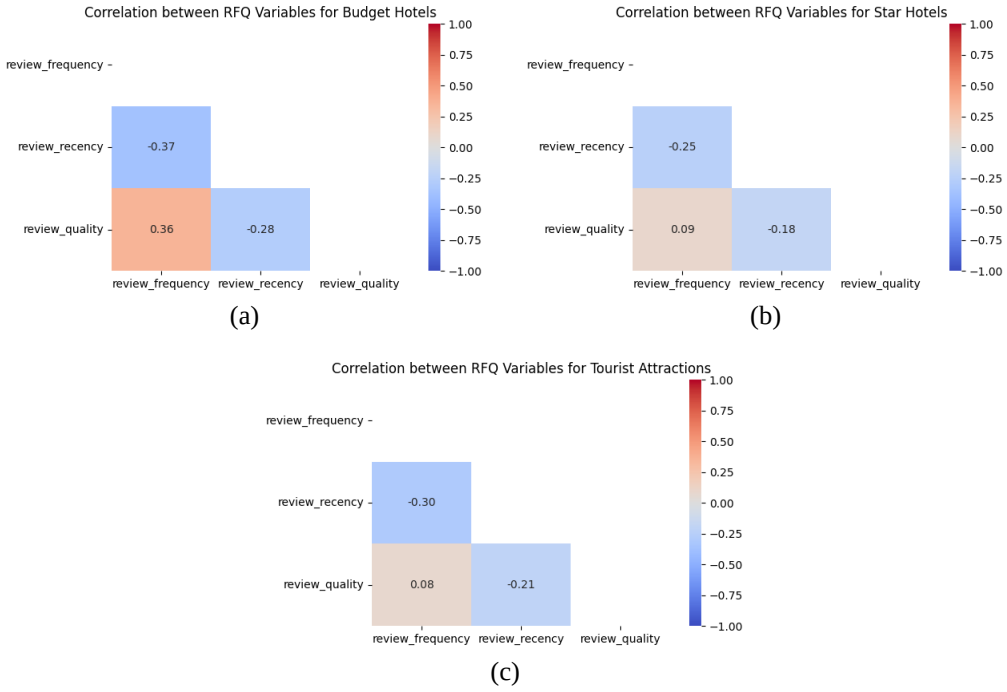


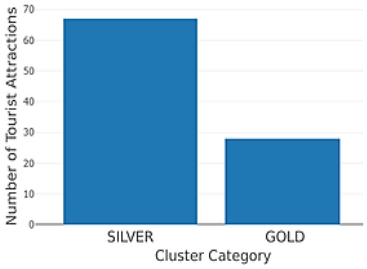
Figure 15. Correlation diagram for RFQ; (a) budget hotel, (b) star hotel, and (c) tourist attraction

3.6. Clustering results insight

As shown in Figures 16(a)-(c), approximately two-thirds of tourist attractions in Bali fall under the silver cluster (67 destinations), while the gold cluster comprises only 28 destinations. For star hotels, the distribution includes 83 hotels in the diamond cluster, 34 in gold, and 75 in silver. Meanwhile, budget hotels are predominantly represented in the gold cluster (478 hotels), followed by diamond (122) and silver (242). This indicates a dominance of the mid-tier segment in terms of quality within the budget hotel sector in Bali.

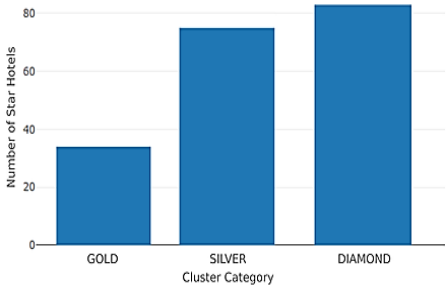
Insights derived from the ranking results in the big data analytics reveal that, among the top 20 tourist attractions in Bali, Gianyar Regency dominates with 8 destinations, followed by Badung (6), Buleleng (3), and one each in Karangasem, Bangli, and Tabanan. This indicates that tourism development remains concentrated

in Gianyar and Badung. Therefore, efforts to strengthen tourist attractions in other regencies should be intensified. In the category of the top 20 star hotels, Badung accounts for 12 hotels, followed by Gianyar with 7, and Denpasar with 1, highlighting the concentration of high-end hotel development in Badung (particularly Nusa Dua) and Gianyar (especially Ubud). As for budget hotels, Badung again leads with 10 hotels, followed by Gianyar with 8, and Karangasem and Buleleng with 1 each. This suggests that budget hotel development also remains focused in Badung and Gianyar, both in terms of quantity and quality.



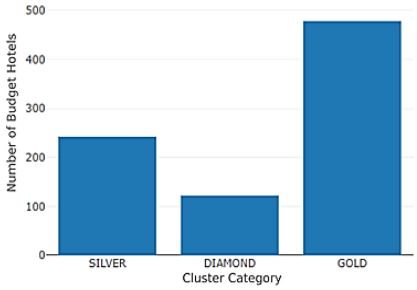
From left to right: i) Silver: 67 tourist attraction and ii) Gold: 28 tourist attraction

(a)



From left to right: i) Gold: 34 hotels, ii) Silver: 75 hotels, and iii) Diamond: 83 hotels

(b)



From left to right: i) Silver: 242 hotels, ii) Diamond: 122 hotels, and iii) Gold: 478 hotels

(c)

Figure 16. Graph of number of; (a) tourist attractions, (b) star hotels, and (c) budget hotels in each cluster

The clustering model using RFQ features can be considered optimal, validated through cost analysis (elbow method) and silhouette score, both standard measures of clustering performance. This is further supported by statistical tests and correlation analysis. In other words, the clustering model developed in this study effectively represents the formation of groups for tourist attractions and hotels, reflecting the tourism characteristics of Bali. The effectiveness of the RFQ feature is evidenced by the alignment between clustering outcomes and RFQ-based rankings. Specifically, in the highest-performing clusters also appear among the top rankings, indicating strong consistency between clustering structure and ranking insight. Specific strategies directly related to the RFQ findings can be implemented based on the value of each RFQ variable. For example, if the low value is the review recency, then policies that increase the novelty and consistency of tourists in providing reviews are needed.

A low silhouette score, despite seemingly good clustering, may be due to the unbalanced distribution and scale of RFQ variables, which complicates effective distance mapping. A low silhouette score does not necessarily imply poor clustering, as shown by the successful grouping of data through boxplots, 3D plot, and feature analysis. The clusters produced in this study remain meaningful and useful, despite the low silhouette score. This study has limitations, as the data were sourced only from TripAdvisor, which may introduce bias. Future research could consider multiple tourism platforms to analyze this RFQ feature. Further development may include integrating geospatial data for location-based insights, real-time sentiment tracking for dynamic RFQ updates, and policy implications with expansion to other sectors (restaurants and cultural events). Another limitation of this study is the inability to conduct sensitivity analysis and temporal trend analysis to see whether the RFQ can track increases or decreases. This is because there is no official source stating the ranking of hotels and tourist attractions that can be used as parameters, while the ranking formed from rating data does not represent review recency, so this trend analysis cannot be carried out.

4. CONCLUSION

The RFQ feature, which represents a novelty in this study, has proven effective for clustering tourist attractions and hotels, as well as for calculating ranking indices of these entities. By employing the RFQ feature, it is possible to assess the quality from three perspectives: reviews recency, number of reviews reflecting popularity, and review quality determined by the ratio of positive to negative sentiment polarity. Insights from big data analytics through clustering and ranking reveal that the development of tourist attractions and hotels, in terms of quality, remains concentrated in the regencies of Badung and Gianyar. Therefore, efforts should be made to improve the quality of attractions and hotels in other regencies across Bali in order to achieve a more balanced tourism industry and equitable economic growth. Based on the clustering results, tourist attractions in the Silver group outnumber those in the gold group, highlighting the need for strategic improvement in quality. Among star hotels, the diamond cluster holds the highest number, indicating that the overall quality of star-rated accommodations in Bali is already high. In contrast, budget hotels are generally of fair quality, with the majority falling within the gold cluster. The successful application of the RFQ feature in this study is expected to serve as a reference for future research that requires clustering based on sentiment analysis polarity, or as a basis for further development of the feature to fit more advanced modeling needs.

FUNDING INFORMATION

This research was conducted using personal funds provided by the authors, without any financial support from external parties. All stages of the research process, including planning, data collection, analysis, and manuscript preparation, were fully self-funded by the authors. Therefore, no sponsorship, grants, or financial assistance from government agencies, private organizations, or other institutions were involved in supporting this study.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ni Wayan Sumartini Saraswati	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
I Ketut Gede Darma Putra		✓								✓		✓		
Made Sudarma		✓								✓		✓		
I Made Sukarsa		✓								✓		✓		
I Gusti Ayu Agung Mas Aristamy						✓				✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data sharing is not applicable to this article. The data generated during this research is not publicly available due to the sensitive nature of the research, which could impact Bali's tourism image if used inappropriately.

REFERENCES

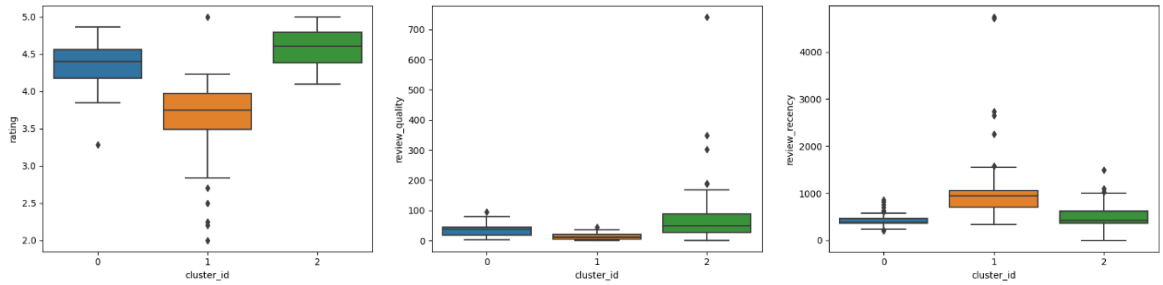
[1] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 6, p. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.

[2] A. Chamekh, M. Mahfoudh, and G. Forestier, "Sentiment Analysis Based on Deep Learning in E-Commerce," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2022, pp. 498–507, doi: 10.1007/978-3-031-10986-7_40.

[3] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022, doi: 10.1007/s10462-022-10144-1.

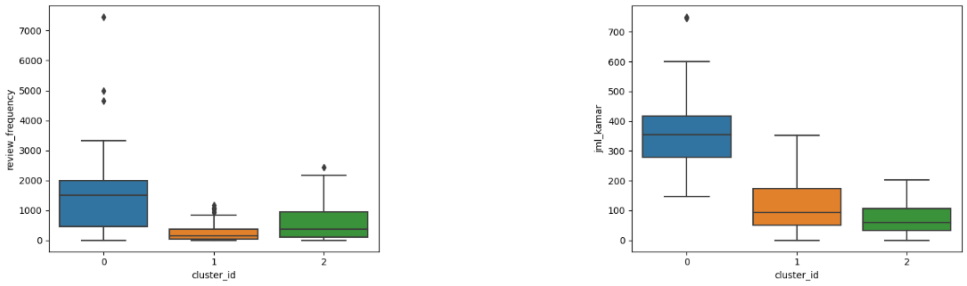
[4] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 108–132, 2023, doi: 10.1109/TAFFC.2020.3038167.
[5] F. Pollák, P. Dorčák, and P. Markovič, "Corporate reputation of family-owned businesses: Parent companies vs. their brands," *Information (Switzerland)*, vol. 12, no. 2, pp. 1–16, 2021, doi: 10.3390/info12020089.
[6] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021, doi: 10.1016/j.knosys.2021.107134.
[7] T. A. Al-Qablan, M. H. M. Noor, M. A. Al-Betar, and A. T. Khader, "A survey on sentiment analysis and its applications," *Neural Computing and Applications*, vol. 35, no. 29, pp. 21567–21601, 2023, doi: 10.1007/s00521-023-08941-y.
[8] H.-B. Yan and Z. Li, "Review of sentiment analysis: An emotional product development view," *Frontiers of Engineering Management*, vol. 9, no. 4, pp. 592–609, Dec. 2022, doi: 10.1007/s42524-022-0227-z.
[9] M. Giannakis, R. Dubey, S. Yan, K. Spanaki, and T. Papadopoulos, "Social media and sensemaking patterns in new product development: demystifying the customer sentiment," *Annals of Operations Research*, vol. 308, no. 1–2, pp. 145–175, 2022, doi: 10.1007/s10479-020-03775-6.
[10] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, doi: 10.1109/ACCESS.2023.3307308.
[11] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Applied Sciences (Switzerland)*, vol. 13, no. 7, 2023, doi: 10.3390/app13074550.
[12] R. Srivastava, P. K. Bharti, and P. Verma, "Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 71–77, 2022, doi: 10.14569/IJACSA.2022.0130312.
[13] N. W. S. Saraswati, I. K. G. D. Putra, M. Sudarma, and I. M. Sukarsa, "Enhance Sentiment Analysis in Big Data Tourism Using Hybrid Lexicon and Active Learning Support Vector Machine," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 1–12, 2024, doi: 10.11591/eei.v13i5.7807.
[14] Y. Luo, X. Zhang, Y. Qin, Z. Yang, and Y. Liang, "Tourism Attraction Selection with Sentiment Analysis of Online Reviews Based on Probabilistic Linguistic Term Sets and the IDOCRIW-COCOSO Model," *International Journal of Fuzzy Systems*, vol. 23, no. 1, pp. 295–308, 2021, doi: 10.1007/s40815-020-00969-9.
[15] Y. Qin, X. Wang, and Z. Xu, "Ranking Tourist Attractions through Online Reviews: A Novel Method with Intuitionistic and Hesitant Fuzzy Information Based on Sentiment Analysis," *International Journal of Fuzzy Systems*, vol. 24, no. 2, pp. 755–777, 2022, doi: 10.1007/s40815-021-01131-9.
[16] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Applied Soft Computing*, vol. 98, p. 106935, 2021, doi: 10.1016/j.asoc.2020.106935.
[17] R. C. Ripan *et al.*, "A Data-Driven Heart Disease Prediction Model Through K-Means Clustering-Based Anomaly Detection," *SN Computer Science*, vol. 2, no. 2, pp. 1–12, 2021, doi: 10.1007/s42979-021-00518-7.
[18] D. M. Saputra, D. Saputra, and L. D. Oswari, "Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method," in *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, 2020, pp. 341–346, doi: 10.2991/aisr.k.200424.051.
[19] R. W. S. Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 1, p. 32, 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.
[20] L. Ismail and H. Materwala, "Comparative Analysis of Machine Learning Models for Diabetes Mellitus Type 2 Prediction," in *Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020*, 2020, pp. 527–533, doi: 10.1109/CSCI51800.2020.00095.
[21] A. Omar and T. Abd El-Hafeez, "Optimizing epileptic seizure recognition performance with feature scaling and dropout layers," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2835–2852, 2024, doi: 10.1007/s00521-023-09204-6.
[22] S. Patil and S. Bhosale, "Hyperparameter Tuning Based Performance Analysis of Machine Learning Approaches for Prediction of Cardiac Complications," in *Advances in Intelligent Systems and Computing Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020)*, 2021, pp. 605–617, doi: 10.1007/978-3-030-73689-7_58.
[23] A. Hussein, F. K. Ahmad, and S. S. Kamaruddin, "Cluster Analysis on Covid-19 Outbreak Sentiments from Twitter Data using K-means Algorithm," *Journal of System and Management Sciences*, vol. 11, no. 4, pp. 167–189, 2021, doi: 10.33168/JSMS.2021.0409.
[24] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, 2020, pp. 747–748, doi: 10.1109/DSAA49011.2020.00096.
[25] M. Lugner *et al.*, "Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study," *Diabetologia*, vol. 64, no. 9, pp. 1973–1981, 2021, doi: 10.1007/s00125-021-05485-5.

APPENDIX



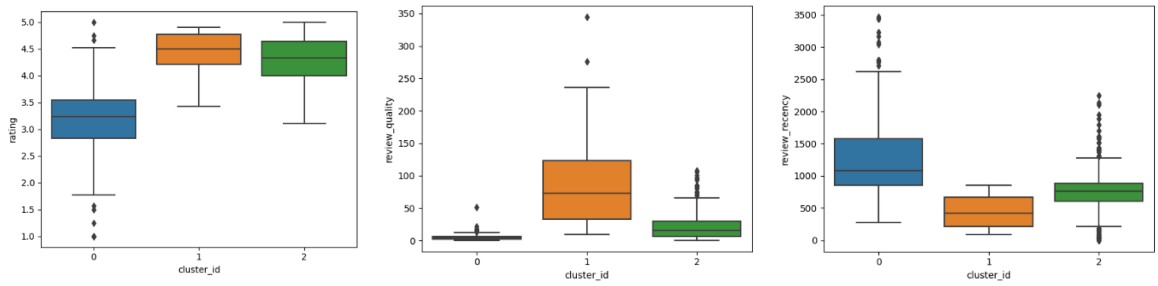
(a) (b) (c)

Figure 7. Cluster analysis of; (a) rating, (b) review quality, and (c) review recency

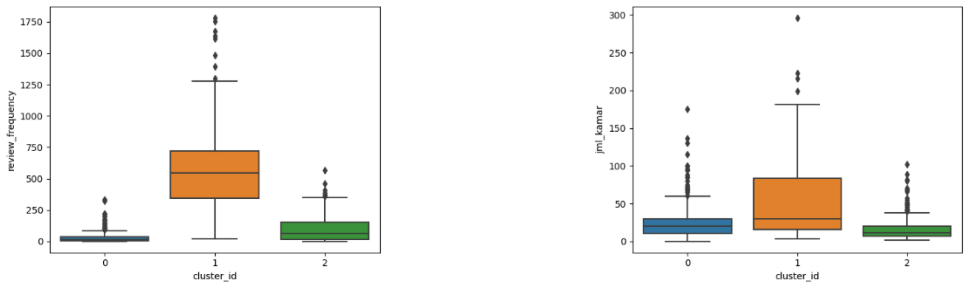


(d) (e)

Figure 7. Cluster analysis of; (d) review frequency and (e) number of rooms in star hotels *(continued)*



(a) (b) (c)



(d) (e)

Figure 10. Cluster analysis of; (a) rating, (b) review quality, (c) review recency, (d) review frequency, and (e) number of rooms in budget hotels

BIOGRAPHIES OF AUTHORS

Ni Wayan Sumartini Saraswati is associate professor at Indonesian Institute of Business and Technology. She holds a Doctoral degree from engineering sciences doctoral program at Udayana University in 2024. She obtained his M.T. (Master of Engineering) degree from Udayana University, Indonesia, in 2011. She received Bachelor's degree in engineering from Telkom University in 2003. Her research interests are big data analytics, machine learning and business intelligence system. She can be contacted at email: sumartini.saraswati@instiki.ac.id.

I Ketut Gede Darma Putra hold a Doctoral degree from Gajah Mada University, Indonesia, in 2007. He also obtained his M.T. degree from Gajah Mada University, Indonesia, in 2000. He received his S.Kom. degree in informatics engineering from the Institute of Ten November Technology Surabaya, Indonesia, in 1997 and now he is a lecturer in the Department of Electrical Engineering and Information Technology, Udayana University Bali, Indonesia. He is currently a professor of information technology science at the Faculty of Engineering, Udayana University, since 2014. His research interests are biometrics, image processing, data mining, and soft computing. He can be contacted at email: ikgdarmaputra@unud.ac.id.

Made Sudarma holds a Doctorate from Udayana University, Indonesia, in 2012. He also holds a Master of Applied Science (M.A.Sc.) from SITE-OU: School of Information Technology and Engineering, Ottawa University Canada in 2000. During his studies at SITE-OU, he was an assistant professor and a member of the research team of the built-in self-testing compaction generator field VLSI technology. He is also a professor of information technology science at the Electrical Engineering Study Program, Faculty of Engineering, Udayana University at Udayana University since 2019. His research includes internet and web applications, cloud computing, artificial intelligence, data warehousing and data mining, computer graphics, and virtual reality, as the author of books and as a reviewer in international and national journals. In addition, he also completed vocational education (IPU., ASEAN Eng) and is active in academic activities, and he also work as an information technology consultant in local government. He can be contacted at email: msudarma@unud.ac.id.

I Made Sukarsa hold a Doctoral degree from Udayana University, Indonesia, in 2019. He also obtained his M.T. degree from Gajah Mada University, Indonesia, in 2005. He received his S.T. degree in informatics engineering from the Gajah Mada University, Indonesia, in 2000. He is currently a professor of information technology science at the Faculty of Engineering, Udayana University. Currently actively teaching and conducting research on IT governance, dialog models on chatbot engines, data warehouses, and system integration. He can be contacted at email: sukarsa@unud.ac.id.

I Gusti Ayu Agung Mas Aristamy is an Informatics lecturer at Institut Bisnis dan Teknologi Indonesia (INSTIKI). She hold her Master degree from Information System Master program at Institut Teknologi Sepuluh Nopember (ITS) in 2018. She obtained her Bachelor Degree (S.TI) from Udayana University in 2016. Her research interests in field of: information system, augmented reality, and human computer interaction. She can be contacted at email: agungmas.aristamy@instiki.ac.id.