ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 5, Issue 2, October 2025, pp. 178-189

Tourist Preference Analysis Based on Google Reviews Using the DBSCAN Method

Marita Qori'atunnadyah [a,1,*], Cahyasari Kartika Murni [a,2], Achmad Firman Choiri [a,3], Hadi Marianto [a,4], Muhammad Yazid [a,5]

[a] Informatics, Institut Teknologi dan Bisnis Widya Gama Lumajang, Jl. Gatot Subroto No. 4, Lumajang 67352, Indonesia
[1] maritaqori@gmail.com*; [2] cahyasarikartikamurni@gmail.com; [3] mr.choiri55@gmail.com; [4] hadimarianto9876@gmail.com;
[5] yazidassudais5@gmail.com
* Corresponding author

ARTICLE INFO

History
Submitted: September 3, 2025
Revised: October 13, 2025
Accepted: October 20, 2025

Keywords
Digital Tourism Analytics
Google Review
TF–IDF
DBSCAN
Lexicon-Based Sentiment Analysis

ABSTRACT

Tourism is a strategic sector contributing to regional economic growth. Although Lumajang Regency offers prominent natural destinations, data-based insights into tourist preferences remain limited. This study analyzed tourist preferences using Google Reviews through a text mining approach that integrated the density-based spatial clustering of applications with noise (DBSCAN) algorithm and lexicon-based sentiment analysis. Data were collected via web scraping from six major destinations, yielding 16,904 reviews, of which 9,800 contained analyzable text. The text data were preprocessed using the term frequency-inverse document frequency (TF–IDF) to generate numerical representations prior to clustering. Using DBSCAN with parameters $\varepsilon = 0.8$ and MinPts = 4, one main cluster comprising 9,353 reviews and 447 outliers was identified. The main cluster was dominated by keywords such as waterfall, beautiful, and scenery, emphasizing the visual appeal of Tumpak Sewu as Lumajang's tourism icon, while the outliers reflected reviews from international visitors and practical travel information. Sentiment analysis showed that most reviews were positive (68.0%), followed by neutral (24.1%) and negative (7.9%). These findings indicate a predominantly positive perception of Lumajang tourism, though accessibility and facilities require improvement. The study demonstrates the potential of digital review data for developing data-driven tourism management and promotion strategies.

1. Introduction

Tourism is a strategic sector in regional economic development because it contributes to increasing community income, preserving cultural values, and creating new jobs. Lumajang Regency has a variety of leading natural tourist destinations, such as Tumpak Sewu Waterfall, Ranu Klakah, and Puncak B29, which offer natural beauty and unique mountain panoramas. This significant potential can drive regional economic growth if managed sustainably and data driven. However, to date, the understanding of tourist preferences remains limited and insufficiently data driven. Therefore, more in-depth studies are needed to identify tourist patterns, needs, and interests in order to develop more effective and targeted tourism strategies.

p-ISSN 2798-253X
e-ISSN 2798-3153

The rise of digital technologies has made user reviews on online platforms like Google Maps have become a crucial source of information regarding tourist experiences and expectations. Various studies have shown that these reviews can be utilized in sentiment-based travel recommendation systems [1], identifying tourist archetypes [2], and designing promotional strategies based on emotional responses [3]. Therefore, employing digital review data opens opportunities for analyzing tourist preferences in greater depth and context. To support this analysis, methods capable of clustering unstructured and diverse review data are needed. One widely used method is density-based spatial clustering of applications with noise (DBSCAN), which is effective in grouping digital content based on proximity of meaning and location, while also detecting outliers [4]–[6]. DBSCAN's advantages lie in its ability to find clusters of arbitrary shape, its robustness to noise, and its lack of a predetermined number of clusters. Therefore, the application of DBSCAN is relevant in analyzing tourist patterns and preferences from digital review data, resulting in more comprehensive insights for tourism sector development [5], [6].

Similarly, previous researchers have extensively explored clustering and regional segmentation methods based on spatial and social data. For instance, k-means and c-means algorithms have been applied to regional grouping based on the teacher–student ratio [7], the human development index (HDI) [8], and spatial clustering related to tourism and the creative economy [9]. However, these methods require prior specification of the number of clusters and tend to perform less effectively when dealing with irregularly shaped or noisy data. To address these limitations, the DBSCAN algorithm offers a more flexible and robust approach, particularly for datasets containing high variability and noise. In the tourism sector, DBSCAN has been effectively applied to map domestic and international tourist visits [10] as well as to perform sentiment-based clustering of tourist reviews, such as for the Puncak B29 destination in Lumajang [11]. Therefore, the adoption of DBSCAN in this study is motivated by its capability to reveal natural groupings within unstructured tourism review data without requiring prior assumptions about the number of clusters.

As a crucial step in processing review text, the term frequency-inverse document frequency (TF-IDF) method is used to transform narrative information into a numerical representation. This technique evaluates the importance of words in a document relative to a set of other documents, allowing for text similarity measurements, keyword extraction, and the compilation of summaries of relevant information [12]–[14]. Despite its limitations in understanding semantic meaning, TF-IDF remains an effective baseline approach for processing unstructured digital reviews.

However, to date, limited research has specifically utilized Google Review content to identify tourist preference segmentation at the district level using the DBSCAN approach. This study aims to address this research gap. It focuses on tourist destinations in Lumajang Regency, where digital review data remain underexplored as an analytical resource for tourism development. The novelty of this research lies in integrating sentiment-based textual analysis with spatial clustering using DBSCAN. This approach uncovers latent patterns of tourist preferences derived directly from user-generated content. Unlike previous studies that primarily examined sentiment polarity or visit frequency, this study contributes a methodological framework for data-driven segmentation of tourist behavior. It combines semantic, spatial, and emotional dimensions. The findings are expected to provide actionable insights for developing adaptive and targeted tourism promotion and management strategies tailored to local characteristics.

2. Method

This research method analyzed tourist preferences based on Google reviews. It employed a text mining approach and the DBSCAN clustering algorithm. Research data were obtained from traveler reviews on the Google Review platform related to tourist destinations in Lumajang Regency. The data were collected using web scraping techniques with the Python programming language. The variables collected included the reviewer's name, review text content, star rating, upload time metadata, and the link (URL) of each review.

The research process in this study followed a structured and sequential approach. This ensures that each stage is systematically connected and supports the final analytical objectives. The overall research framework included data collection, preprocessing, TF–IDF computation, clustering using the DBSCAN algorithm, visualization, and sentiment analysis. Fig. 1 illustrates this framework. The flowchart provides a comprehensive overview of the workflow used to analyze tourist review data from Google Reviews in Lumajang Regency.

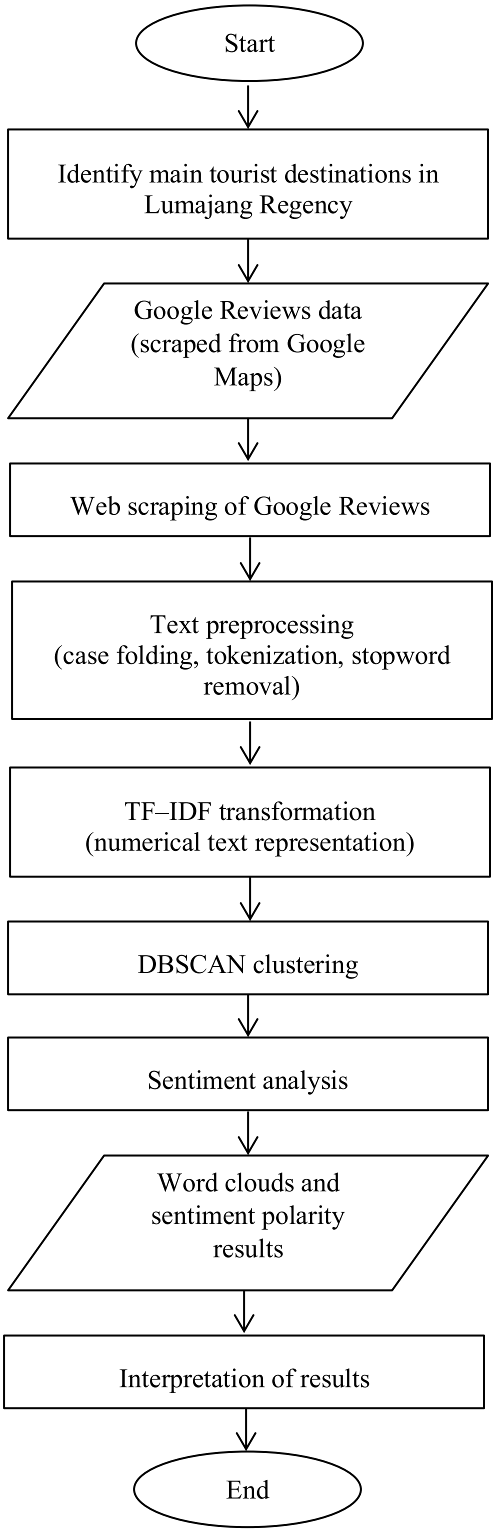
Start
Identify main tourist destinations in Lumajang Regency
Google Reviews data (scraped from Google Maps)
Web scraping of Google Reviews
Text preprocessing (case folding, tokenization, stopword removal)
TF–IDF transformation (numerical text representation)
DBSCAN clustering
Sentiment analysis
Word clouds and sentiment polarity results
Interpretation of results
End

Fig. 1 Research flowchart.

As illustrated in Fig. 1, the research begins with identifying major tourist destinations, followed by the extraction of Google Reviews to obtain textual and non-textual information. A total of 16,904 review records were collected, of which 9,800 contained analyzable text, while the remainder were star ratings without comments. All extracted data were stored in CSV format to facilitate the subsequent stages of preprocessing and analysis.

The data preprocessing stage was carried out to ensure the quality of the review text to be analyzed. This process began with text cleaning to remove irrelevant elements such as numbers, punctuation, URLs, and special characters. Next, case folding was performed to convert all letters into lowercase for uniformity. The text was then segmented into individual word units through tokenization, followed by stopword removal to eliminate common words that did not contribute meaningful information. The final step transforms the cleaned text into a numerical representation using the TF–IDF method, enabling the data to be further processed in clustering analysis.

TF–IDF is a statistical measure used to evaluate the importance of a word in a document relative to an entire corpus. It combines two components: Term frequency (TF) measures how frequently a word appears in a document, while inverse document frequency (IDF) assesses how rare or common the word is across documents [15]. In information retrieval, TF–IDF plays a crucial role in extracting relevant keywords, supporting the construction of effective queries in database searches [16]. It is also widely applied in text classification to weight terms and improve categorization accuracy [17], as well as to rank documents according to their relevance, thereby enhancing the efficiency of information retrieval systems [18].

The implementation of TF–IDF generally begins with data cleaning, which involves removing irrelevant words such as stopwords and performing text normalization through stemming or lemmatization techniques [19], [20]. After cleaning, the TF for each word is calculated as the ratio of the number of times a term appears to the total number of words in the document. IDF is then computed by taking the logarithm of the total number of documents divided by the number of documents that contain the term [21]. The TF–IDF score is obtained by multiplying the TF and IDF values, emphasizing terms that frequently appear in specific documents but are less common across the corpus [22]. Once the TF–IDF scores are obtained, the terms can be sorted based on their weights to identify the most representative keywords in each document [21]. Thus, TF–IDF serves as an effective foundational method in text processing because it extracts essential information from unstructured text while simultaneously improving the performance of analytical models in classification and information retrieval tasks.

Clustering, or cluster analysis, is a method used to identify meaningful groups within a dataset [23]. One of the most widely applied clustering algorithms is DBSCAN. This algorithm organizes data points based on density distribution, determined by two key parameters: epsilon (ε), which represents the neighborhood radius for determining proximity, and minimum points (MinPts), which specify the minimum number of data points required to form a dense region. A data point is classified as a core point when the number of neighboring points within ε (including the point itself) is greater than or equal to MinPts. Points that fall within the ε-radius of a core point but contain fewer neighbors than MinPts are defined as border points, while points that satisfy neither condition are categorized as noise or outliers [24].

The DBSCAN procedure begins with determining the appropriate values of ε and MinPts, followed by the selection of an initial point from the dataset. The distance between this point and all others is then computed using the Euclidean distance metric, expressed as in (1).

$$d_{ij} = \sqrt{\sum_{a}^{p}\left(x_{ia} - x_{ja}\right)^2} \quad (1)$$

where x_{ia} is the ath variable of object i $(i = 1, \dots, n; a = 1, \dots, p)$ and d_{ij} is the Euclidean distance value. In this study, Euclidean distance was selected because TF–IDF yields numerical vectors situated in a continuous multidimensional space, making Euclidean distance an effective measure for capturing differences in term importance across documents. This choice is consistent with previous research on DBSCAN-based text clustering, where Euclidean distance has proven to be a reliable metric for determining proximity between textual data points.

A cluster is formed when the number of points within ε exceeds the MinPts threshold, with the initial point serving as a core point. The algorithm continues by expanding the cluster through density-reachable points and iterates through the remaining data points until all objects have been evaluated. When a boundary point has no additional density-reachable neighbors, the process proceeds to the next point in the dataset. To ensure that the clustering results are meaningful and accurately reflect the underlying data structure, a validation process is required. Clustering validation may utilize internal or external indices. Internal indices assess the quality of the clustering structure without relying on external class labels, whereas external indices compare cluster assignments with predefined classifications. One widely used internal metric is the Silhouette coefficient [25], defined as in (2).

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \tag{2}$$

where $s(i)$ is the Silhouette coefficient value, $a(i)$ is the average distance between point i and all points in cluster A (the cluster to which point i belongs), and $b(i)$ is the average distance between point i and all points in clusters other than A. The Silhouette coefficient value ranges from -1 to 1, with values closer to 1 indicating good cluster quality, while values closer to -1 indicate clustering errors.

The results were visualized using word clouds to display the dominant words in each cluster based on TF-IDF weights. This technique helps identify key themes emerging within each traveler preference segment in a more intuitive and informative way. A word cloud is a visual representation of text data, where the size of each word indicates its frequency of occurrence or importance in a given context. This method is widely used in various fields, such as cloud manufacturing, social media analytics, and location-based services, as it simplifies complex information into a more understandable format. Its development has also advanced with the development of algorithms that can determine the size and placement of words based on relevance, as demonstrated in tag clouds in location-based services [26]. Furthermore, word clouds can be enriched with word embedding models, thus providing a multidimensional representation of concepts and their relationships within the text [27]. Thus, the use of word clouds serves not only as a visual illustration but also as an analytical tool that supports the understanding of keyword patterns in each cluster of traveler reviews.

As a complement, sentiment analysis was conducted to classify tourist reviews into positive, negative, and neutral categories. This process aims to provide an overview of tourists' emotional tendencies toward the destinations they visit. The procedure began with text preprocessing, following the same steps applied earlier to ensure that the review text is clean and standardized before analysis. After preprocessing, a lexicon-based approach was applied by utilizing a predefined sentiment dictionary to classify the polarity of each review. This method evaluates the emotional tone of the text by mapping individual words to sentiment scores, enabling the identification of positive, negative, or neutral sentiment across various contexts. The effectiveness of this approach depends significantly on the quality and suitability of the chosen lexicon, such as SentiWordNet, which assigns sentiment scores to words based on their semantic properties [28]. Established sentiment lexicons, including SentiWordNet and domain-specific lexicons such as those used in economics, provide structured frameworks for systematically assessing sentiment by mapping words to relevant sentiment scores [28], [29]. Compared with machine learning approaches, lexicon-based sentiment analysis offers advantages in generalizability and domain independence because it does not require model training and can be applied directly across different datasets [30]. Additionally, lexicon-based methods support fine-grained sentiment analysis, allowing for the capture of subtle nuances in emotional expression, which is particularly important when interpreting complex or context-dependent reviews [30]. After assigning sentiment scores to the text, the sentiment distribution within each cluster is evaluated to determine whether particular traveler groups tend to express predominantly positive, neutral, or negative opinions in their reviews.

3. Results and Discussion

This section presents the analysis of tourist review data obtained from Google Reviews for several natural tourist destinations in Lumajang Regency. The analysis was conducted through several stages, starting with data clustering using the DBSCAN algorithm and ending with lexicon-based sentiment classification. The results presented not only illustrate the cluster structure of tourist reviews but also the emotional tendencies reflected in the form of positive, neutral, and negative sentiment. Furthermore, the results of this analysis are interpreted to provide a more comprehensive understanding of tourist preferences and their implications for tourism management and development strategies in Lumajang Regency.

The data used in this study consisted of 16,904 traveler reviews obtained through the Google Reviews platform. Of these, 9,800 contained text that could be further analyzed, while the remaining 8,308 reviews contained only ratings without comments and were therefore not included in the text-based analysis. In general, the level of tourist satisfaction with tourist destinations in Lumajang Regency can be said to be very high. This is reflected in the average overall rating of 4.67 on a scale of 5, indicating that the majority of tourists gave positive assessments. These data distribution provides an initial overview of the image of tourist destinations in Lumajang Regency. Textual reviews will form the basis for analyzing tourist preferences, while the distribution of ratings provides a general indicator of visitor satisfaction levels.

The distribution of tourist review ratings, as illustrated in Fig. 2, shows that the majority of visitors gave very high ratings to tourist destinations in Lumajang Regency. Of the 16,904 total reviews, 13,574 reviews (80.3%) provided five-star ratings, confirming the dominance of positive experiences. Four-star ratings accounted for 2,159 reviews (12.8%), cumulatively representing over 93% of reviews in the high rating category. Meanwhile, reviews with a medium rating, namely three stars, totaled 560 reviews (3.3%). Low ratings were relatively few, with 166 reviews (1.0%) giving two stars and 445 reviews (2.6%) giving one star. This very small proportion of low ratings indicates that tourist complaints represent only a small portion of the total reviews. Overall, this rating distribution pattern reinforced previous findings that tourism in Lumajang Regency has a very positive image among tourists, with high levels of satisfaction with their travel experiences.

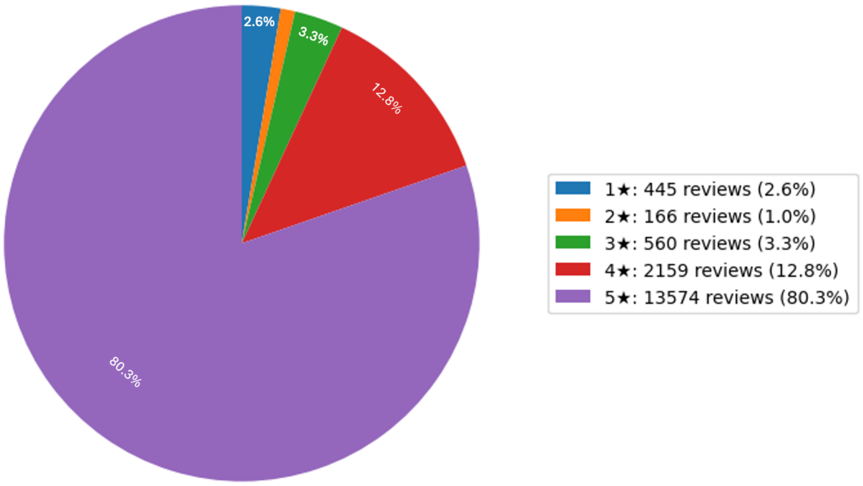

Fig. 2 Google review rating distribution.

The distribution of tourist reviews by destination, as presented in Fig. 3, shows a clear variation among major tourist attractions in Lumajang Regency. Out of the total 9,800 text-based reviews, Tumpak Sewu Waterfall stood out with the highest number of reviews, reaching 4,149 entries, reinforcing its position as the region’s primary tourism icon. Its breathtaking panorama makes it a major attraction for both domestic and international visitors. The second most reviewed destination was Puncak B29 with 2,184 reviews, followed by Selokambang Natural Bath with 1,274 reviews. Both sites are popular for their spectacular mountain views and family-friendly accessibility. On the other hand, destinations such as Ranu Regulo (830 reviews), Kapas Biru Waterfall (725 reviews), and Ranu Kumbolo (638 reviews) received fewer reviews. Although their numbers are smaller compared to Tumpak Sewu or B29, these attractions still offer unique appeal, especially for tourists interested in outdoor activities, hiking, and adventure tourism. Overall, this distribution highlights that tourism in Lumajang Regency is strongly centered around Tumpak Sewu, while other destinations enrich the variety by catering to more specific tourist segments.

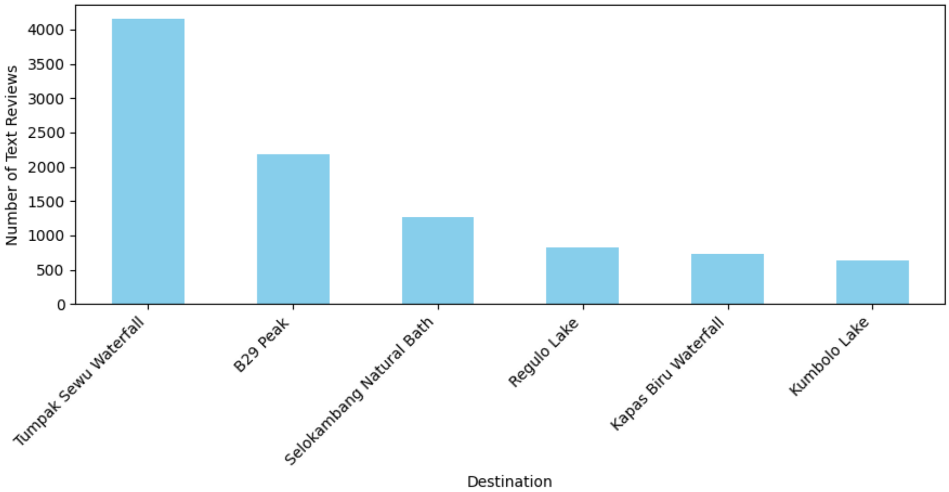

Fig. 3 Distribution of tourist reviews by destination.

The clustering process using the DBSCAN algorithm was performed on traveler review data collected from Google Reviews. Of the total data extracted, only 9,800 contained text reviews, while the remainder consisted of star ratings without comments and could not be analyzed further. All text data were then processed using TF-IDF before being clustered with DBSCAN. In this study, the variables employed in the DBSCAN clustering were the weighted term frequency values generated through the TF–IDF transformation. Each review was represented as a numerical feature vector, where each dimension corresponded to a unique term in the corpus and its TF–IDF weight indicated the importance of that term within the document. These TF–IDF features served as the input variables for DBSCAN to identify clusters of reviews with similar textual characteristics.

Based on a series of parameter experiments, the best configuration was obtained at an epsilon (ε) value of 0.8 and a minimum sample (MinPts) of 4, with a Silhouette score of 0.0116. Although the Silhouette score obtained was relatively low, this value is reasonable considering the characteristics of the dataset. The textual data derived from Google Reviews tends to be highly unstructured and sparse, which naturally leads to low cluster separation when represented through TF–IDF features. In this context, the use of the DBSCAN algorithm remains justified because it is designed to handle data with varying densities and to identify clusters of arbitrary shapes without requiring the number of clusters to be specified in advance. Unlike centroid-based methods such as k-means, DBSCAN can detect dense regions of similar reviews while isolating noise and outlier data points that do not belong to any cluster. Therefore, despite the small Silhouette value, DBSCAN provides a more flexible and robust approach for exploring patterns in heterogeneous text data such as user-generated reviews. The results of this clustering produced two main categories: one large cluster with 9,353 reviews and an outlier group with 447 reviews. Four reviews that were originally included in the small cluster were then combined into the main cluster to make the distribution more representative.

The primary focus of this study is on analyzing the textual content of tourist reviews, rather than the physical characteristics of the tourist attractions themselves. The analysis aims to uncover patterns of tourist perceptions, preferences, and emotional tendencies expressed in Google Review comments. Each cluster represents groups of reviews with similar linguistic and semantic characteristics, which indirectly reflect how visitors perceive different aspects of tourism experiences in Lumajang Regency. Therefore, while the visualization (e.g., word clouds) refers to specific destinations such as Tumpak Sewu or Ranu Klakah, the interpretation remains centered on the language and sentiment expressed by tourists, not on the physical attributes of the destinations.

The word cloud visualization, as illustrated in Fig. 4, represents the main cluster obtained from the DBSCAN results and contains the majority of tourist reviews. This cluster is characterized by the dominance of words such as “waterfall” (*air terjun*), “Tumpak Sewu”, “beautiful” (*indah*), “nice” (*bagus*), “view” (p*emandangan*), “panorama,” “down” (*turun*), “up” (*naik*), “cool” (*sejuk*), “access” (*akses*), “parking” (*parkir*), and “road” (*jalan*). This pattern indicated that most tourists emphasize natural beauty as the main attraction of their visit, particularly the panoramic view of Tumpak Sewu Waterfall, a tourism icon of Lumajang. Furthermore, words related to physical activity, such as “down” and “up,” indicated that the experience of navigating challenging paths was also considered part of the tourism value. Reviews related to accessibility, parking facilities, and area cleanliness also frequently appeared, indicating that comfort factors influence tourist perceptions.

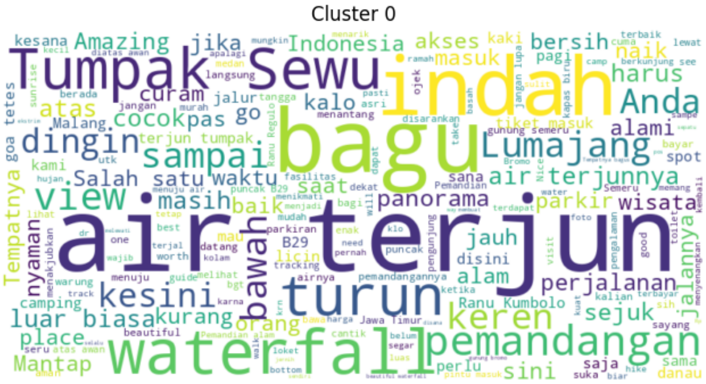

Fig. 4 Word cloud of tourist reviews in cluster 0.

The outlier cluster visualization, as shown in Fig. 5, represents reviews grouped under cluster –1, which exhibits a more diverse linguistic composition compared to the main cluster. Dominant words in this cluster included “waterfall” (*air terjun*), “guide,” “local,” “view,” “visit,” “place,” “worth,” “time,” “ojek,” “motorbike,” and “ticket.” The mix of English and Indonesian vocabulary within this cluster suggested that the reviews were written by both international travelers and local visitors, some of whom used a blended language style. As illustrated in Fig. 5, many reviews in this group also contain practical information, such as references to travel guides, transportation options, costs, or specific tips for accessing the destinations. The presence of terms like “B29” and “Lumajang” further indicated that several travelers discussed multiple destinations within a single review, leading to a lower density of similar content and resulting in their classification as outliers by the DBSCAN algorithm. This cluster captures contextual and comparative insights rather than shared emotional tones, distinguishing it from the dominant aesthetic appreciation found in the main cluster.

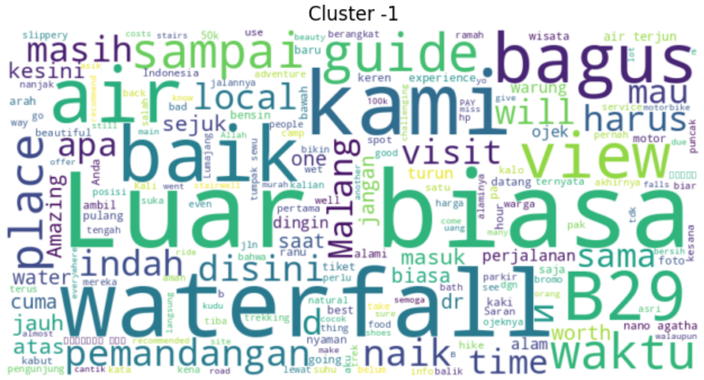

Fig. 5 Word cloud of tourist reviews in cluster -1.

Overall, the DBSCAN results showed that while most reviews fell into one large cluster reflecting tourists' admiration for the natural beauty and panoramic views of waterfalls, the outlier group still contained important information. This group highlights unique experiences, the need for practical information, and the presence of international travelers with different perspectives. Therefore, even though the Silhouette score value was relatively low due to the high dimension of text data and the dominance of one large cluster, this analysis still succeeded in revealing the segmentation of tourist preferences, which is useful for developing promotional strategies and managing tourist destinations in Lumajang Regency.

Following the clustering analysis, sentiment analysis was conducted to gain a deeper understanding of the emotional tendencies expressed in the reviews. Using a lexicon-based approach that incorporated lists of positive and negative words, including nonstandard spelling variations frequently found in user-generated content, the method evaluated the sentiment expressed toward various tourist destinations. By examining the lexical patterns within the reviews, this analysis identified overall sentiment tendencies whether positive, neutral, or negative thereby complementing the clustering results.

The distribution of sentiment polarity, as illustrated in Fig. 6, shows that the majority of reviews tended to be positive, accounting for 68.0% of the total. This percentage confirmed that most tourists reported satisfying experiences, particularly regarding the natural beauty and panoramic views of tourist destinations in Lumajang Regency. Positive reviews generally contain words such as "beautiful," "amazing," "extraordinary," "clean," "comfortable," "recommended," "cool," "friendly," "worth it," and "breathtaking," reflecting tourists' admiration for the scenic landscapes and the overall enjoyable atmosphere, especially at destinations like Tumpak Sewu Waterfall, Ranu Kumbolo, and other similar natural attractions. These expressions demonstrate a strong emotional response associated with pleasure, appreciation, and satisfaction.

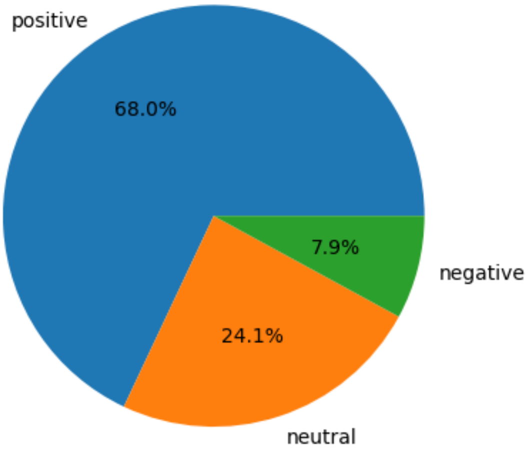

Fig. 6 Distribution of the sentiment.

A further 24.1% of reviews were categorized as neutral. Reviews in this category typically provide informative descriptions or convey experiences without strong emotional expression—for example, describing the condition of the route to the location, ticket prices, weather, or available facilities. Neutral reviews often contained objective or descriptive phrases such as "entrance ticket," "parking area," "footpath," "toilet," or "food stalls." While they do not explicitly express satisfaction or dissatisfaction, these reviews are essential for understanding contextual factors that influence tourists' experiences, such as accessibility and infrastructure conditions.

Meanwhile, only 7.9% of reviews were negative. Reviews in this category generally highlight challenging terrain, difficult accessibility, inadequate facilities, and complaints about additional costs, such as parking fees or motorcycle taxi services. Negative expressions identified in the dataset included words such as "dirty," "slippery," "crowded," "broken," "expensive," "far," "inconvenient," "disappointing," "unfriendly," and "not recommended." Although relatively small in proportion, negative reviews provide constructive feedback for destination managers, helping identify weaknesses that require attention particularly related to infrastructure, maintenance, and service quality improvements.

Overall, this sentiment distribution indicates that tourism in Lumajang Regency maintains a predominantly positive image among visitors. However, the existence of neutral and negative reviews still provides valuable insights that can serve as a basis for strategic evaluation, particularly regarding infrastructure development, accessibility enhancement, and the provision of clearer and more informative guidance for visitors. By combining the results of sentiment analysis and DBSCAN clustering, this study offers a comprehensive understanding of tourist preferences and perceptions toward natural attractions in Lumajang Regency, which can serve as an empirical foundation for developing adaptive, data-driven tourism management strategies.

4. Conclusion

This study successfully demonstrated how digital reviews via Google Reviews can be leveraged to understand tourist preferences for natural tourism destinations in Lumajang Regency. The analysis revealed that Tumpak Sewu Waterfall dominated as the destination with the most reviews, confirming its position as a key regional tourism icon. Through the application of the DBSCAN algorithm, the majority of reviews were grouped into one large cluster emphasizing the scenic beauty, accessibility, and experience of visiting Tumpak Sewu, while outliers shared perspectives from international tourists and practical travel information.

Sentiment analysis reinforced these findings, finding that 68.0% of reviews were positive, 24.1% were neutral, and only 7.9% were negative. This confirms the positive image of Lumajang tourism, although complaints about access, facilities, and additional costs still require management attention. Overall, the combination of clustering and sentiment analysis based on online reviews provides a more comprehensive picture of tourist preferences. The results of this study are expected to serve as a basis for local governments and tourism destination managers in designing promotional strategies, improving service quality, and developing data-driven tourism that is adaptive, targeted, and sustainable.

For future research, it is recommended that the analysis be expanded to include additional variables such as spatial data, seasonal visitation patterns, and tourist interactions on other social media platforms. Furthermore, more sophisticated machine learning approaches, such as word embedding or topic modeling, can be used to extract deeper semantic meaning from reviews. Thus, the analysis results can provide richer and more predictive insights, while strengthening the basis for data-driven decision-making in tourism development.

References

[1] S. Bairavel and M. Krishnamurthy, “User preference and reviews analysis with neural networks for travel recommender systems,” *Int. J. Eng. Res. Technol.*, vol. 13, no. 8, pp. 1896–1900, 2020, doi: 10.37624/ijert/13.8.2020.1896-1900.

[2] M. Sharmin, A. Sumy, Y.A. Parh, and S. Hossain, “Identifying and classifying traveler archetypes from Google Travel Reviews,” *Int. J. Stat. Appl*, vol. 11, no. 3, pp. 61–69, 2021, doi: 10.5923/j.statistics.20211103.02.

[3] L. Durmishi, A.C.M. Paredes, J.G. Sávoly, and G. Kovács, “Investigating the customer preference towards Michelin restaurants in Europe through Google Reviews,” *Ecocycles*, vol. 10, no. 1, pp. 115–123, 2024, doi: 10.19040/ecocycles.v10i1.434.

[4] H. Yaşar and M. albayrak, “Comparison of serial and parallel programming performance in outlier detection with DBSCAN algorithm,” *Bilecik Şeyh Edebali Üniversitesi Fen Bilim. Derg.*, vol. 7, no. 1, pp. 129–140, 2020, doi: 10.35193/bseufbd.649539.

[5] R. Benaya, Y. Sibaroni, and A.F. Ihsan, “Clustering content types and user roles based on tweet text using K-medoids partitioning based,” *J. Comput. Syst. Inform.*, vol. 4, no. 4, pp. 749–756, Aug. 2023, doi: 10.47065/josyc.v4i4.3751.

[6] O.O. Wijaya and Rushendra, “Analysis of Sulawesi earthquake data from 2019 to 2023 using DBSCAN clustering,” *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 8, no. 4, pp. 454–465, Aug. 2024, doi: 10.29207/resti.v8i4.5819.

[7] M. Qori’atunnadyah, “Pengelompokkan wilayah berdasarkan rasio guru-murid pada jenjang pendidikan menggunakan algoritma K-means,” *J. Inform. Dev.*, vol. 1, no. 1, pp. 33–38, Mar. 2023, doi: 10.30741/jid.v1i1.898.

[8] M. Qori’atunnadyah, “Fuzzy C-Means for regional clustering in East Java Province based on human development index indicators,” *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 16, no. 2, pp. 524–534, Dec. 2023, doi: 10.36456/jstat.vol16.no2.a8240.

[9] M. Qori’atunnadyah, “Metode C-Means untuk pengelompokkan kabupaten/kota Provinsi Jawa Timur berdasarkan indikator indeks pembangunan manusia (IPM),” *J. Inform. Dev.*, vol. 1, no. 2, pp. 51–58, Apr. 2023, doi: 10.30741/jid.v2i2.1013.

[10] M. Qori’atunnadyah, “Mapping of domestic and foreign tourist visits in East Java using the DBSCAN method,” *J. Pilar Nusa Mandiri*, vol. 21, no. 1, pp. 9–15, Mar. 2025, doi: 10.33480/pilar.v21i1.6073.

[11] F.A. Hizham, C.K. Murni, and M. Qori’atunnadyah, “Uji klasifikasi algoritma naïve Bayes classification dalam analisis sentimen ulasan Puncak B29 Lumajang,” *Progresif J. Ilm. Komput.*, vol. 20, no. 1, p. 361, 2024, doi: 10.35889/progresif.v20i1.1618.

[12] F. Lan, “Research on text similarity measurement hybrid algorithm with term semantic information and TF–IDF method,” *Adv. Multimed.*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/7923262.

[13] W. Zhuohao, W. Dong, and L. Qing, “Keyword extraction from scientific research projects based on SRP-TF–IDF,” *Chinese J. Electron.*, vol. 30, no. 4, pp. 652–657, Jul. 2021, doi: 10.1049/cje.2021.05.007.

[14] [D. Chaurasia, P.V.D. K, and M. Bhatta, “Enhancing text summarization through parallelization: A TF–IDF algorithm approach,” in *2024 2nd Int. Conf. Intell. Cyber Physical Syst. Internet of Things (ICoICI)*, Aug. 2024, pp. 1503–1508, doi: 10.1109/ICoICI62503.2024.10696641.

[15] N.S.M. Nafis and S. Awang, “An enhanced hybrid feature selection technique using term frequency–inverse document frequency and support vector machine–recursive feature elimination for sentiment classification,” *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[16] F. Lechtenberg, J. Farreres, A.-L. Galvan-Cara, A. Somoza-Tornos, A. Espuña, and M. Graells, “Information retrieval from scientific abstract and citation databases: A query-by-documents approach based on Monte-Carlo sampling,” *Expert Syst. Appl.*, vol. 199, Aug. 2022, Art. no 116967, doi: 10.1016/j.eswa.2022.116967.

[17] J. Attieh and J. Tekli, “Supervised term-category feature weighting for improved text classification,” *Knowledge-Based Syst.*, vol. 261, Feb. 2023, Art. no 110215, doi: 10.1016/j.knosys.2022.110215.

[18] S. Hao, C. Shi, L. Cao, Z. Niu, and P. Guo, “Learning deep relevance couplings for ad-hoc document retrieval,” *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no 115335, doi: 10.1016/j.eswa.2021.115335.

[19] N.P. Yunita, “Aplikasi pencarian hadis menggunakan vector space model dengan pembobotan TF–IDF dan Confix-Stripping stemmer,” *J. Teknol. Inf. Ilm. Komput.*, vol. 10, no. 3, pp. 665–676, Jul. 2023, doi: 10.25126/jtiik.20231036736.

[20] A.S. Bashir, A.A. Bichi, and A. Adamu, “Automatic construction of generic Hausa language stop words list using term frequency-inverse document frequency,” *J. Electr. Syst. Inf. Technol.*, vol. 11, no. 1, Dec. 2024, Art. no 58, doi: 10.1186/s43067-024-00187-5.

[21] A.F. Al Shammari, “Implementation of keyword extraction using term frequency–inverse document frequency (TF–IDF) in Python,” *Int. J. Comput. Appl.*, vol. 185, no. 35, pp. 9–14, Sep. 2023, doi: 10.5120/ijca2023923137.

[22] H.S. Lubis, M.K.M. Nasution, and A. Amalia, “Performance of term frequency–inverse document frequency and K-means in government service identification,” in *2024 4th Int. Conf. Sci. Inf. Technol. Smart Administration (ICSINTESA)*, Jul. 2024, pp. 772–777, doi: 10.1109/ICSINTESA62455.2024.10748106.

[23] P. Giordani, M.B. Ferraro, and F. Martella, *An Introduction to Clustering with R*, vol. 1. Singapore: Springer Singapore, 2020, doi: 10.1007/978-981-13-0553-5.

[24] M. Pietrzykowski, “Comparison of mini-models based on various clustering algorithms,” *Procedia Comput. Sci.*, vol. 176, pp. 3563–3570, 2020, doi: 10.1016/j.procs.2020.09.030.

[25] F. Batool and C. Hennig, “Clustering with the average silhouette width,” *Comput. Stat. Data Anal.*, vol. 158, Jun. 2021, Art. no 107190, doi: 10.1016/j.csda.2021.107190.

[26] H. Liu, X. Wang, Z. Wang, and Y. Cheng, “Does digitalization mitigate regional inequalities? Evidence from China,” *Geogr. Sustain.*, vol. 5, no. 1, pp. 52–63, Mar. 2024, doi: 10.1016/j.geosus.2023.09.007.

[27] P. Aceves and J.A. Evans, "Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science," *Organ. Sci.*, vol. 35, no. 3, pp. 788–814, May 2024, doi: 10.1287/orsc.2023.1686.

[28] M. Husnain, M.M.S. Missen, N. Akhtar, M. Coustaty, S. Mumtaz, and V.B.S. Prasath, "A systematic study on the role of SentiWordNet in opinion mining," *Front. Comput. Sci.*, vol. 15, no. 4, Aug. 2021, Art. no 154614, doi: 10.1007/s11704-019-9094-0.

[29] L. Barbaglia, S. Consoli, S. Manzan, L.T. Pezzoli, and E. Tosetti, "Sentiment analysis of economic text: A lexicon-based approach," *Econ. Inq.*, vol. 63, no. 1, pp. 125–143, Jan. 2025, doi: 10.1111/ecin.13264.

[30] A.M. van der Veen and E. Bleich, "The advantages of lexicon-based sentiment analysis in an age of machine learning," *PLoS One*, vol. 20, no. 1, Jan. 2025, Art. no e0313092, doi: 10.1371/journal.pone.0313092.