Topic modeling on Natural Tourism Objects in West Bandung Regency Based on Reviews from Google Maps with the Latent Semantic Analysis (LSA) Method

I G N A Agni Prema N, Faqih Hamami*, and Ekky Novriza Alam

School of Industrial and System Engineering, Telkom University, Bandung, Indonesia
faqihhamami@telkomuniversity.ac.id

Abstract. The COVID-19 pandemic has prompted a transformation in Indonesia's tourism sector to accommodate changes in traveler preferences during the pandemic and to stabilize the tourism industry post-pandemic. West Bandung Regency, renowned for its abundant natural attractions, holds significant potential for developing tourist destinations. In support of this, the Tourism and Culture Office of West Java Province has devised a customer-centric "new normal strategy." This research employs topic modeling using latent semantic analysis (LSA) on traveler reviews from Google Maps. The study examined various scenarios and found that the optimal number of topics is 2, utilizing advanced stopword pre-processing and a bag of words (BoW) representation, resulting in a coherence score metric of 0.717. The choice of two topics is based on the highest coherence score metric. Opting for more topics would lead to redundancy, with additional topics merely reiterating discussions covered in the previous ones. The analysis of traveler reviews revealed keywords that frequently appear, such as accessibility, location, parking, tickets, beauty, nature, and facilities. The analysis of these keywords suggests that tourists frequently discuss topics concerning accessibility to attractions, facilities available, and the overall atmosphere experienced at these tourist spots. Based on these findings, recommendations are proposed, including enhancing accessibility, improving parking and payment systems, upgrading photo spots, and enhancing facilities at waterfalls and natural attractions in West Bandung Regency. By implementing the outcomes of this research, the government can take appropriate actions to improve the traveler experience, optimize natural tourist attractions, and promote West Bandung Regency as an attractive tourism destination post-pandemic.

Keywords: *data mining, latent semantic analysis, natural attractions, topic modeling, tourist reviews.*

1 Introduction

The COVID-19 pandemic significantly impacted Indonesia's tourism sector, causing a 90% drop in domestic tourist visits and prompting President Joko Widodo to empha-

A. Putro Suryotomo and H. Cahya Rustamaji (eds.), *Proceedings of the 2023 1st International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*,
Advances in Intelligent Systems Research 181,
https://doi.org/10.2991/978-94-6463-366-5_7

size the need for transformative measures in tourism [1, 2]. West Java Province, despite its untapped potential for tourism, remains underutilized, prompting the Governor's call for leveraging this potential to revive the sector [4]. Bandung Barat Regency, renowned for its natural waterfall attractions, suffered neglect during the pandemic due to reduced visits [5]. To address this, the Department of Tourism and Culture of West Java Province is formulating a "new normal strategy," prioritizing safety and comfort for traveler, also provide economic recovery for the province [5].

Utilizing tourist reviews on Google Maps for these attractions is crucial, and employing Latent Semantic Analysis (LSA) for topic modeling offers an effective means to swiftly process and understand these reviews, identifying visitor preferences to enhance the natural attractions [5]. Prior studies have showcased LSA's efficacy in analyzing large text data and deriving meaningful insights [6, 7, 8, 9]. However, the focus was primarily on accuracy, leaving room to extend its application to uncover tourist preferences regarding natural attractions in West Bandung Regency.

This research aims to leverage LSA to unveil concealed topics within tourist reviews, providing essential insights into their current preferences and needs. Such insights will be pivotal for the Department of Tourism and Culture of West Java Province, aiding in formulating strategies for tourism sector recovery and development. Understanding these preferences deeply will guide the government in targeted improvements, potentially making strategies in Bandung Barat Regency and West Java more effective, elevating tourism appeal, enhancing visitor experiences, and bolstering the local economy.

2 Method

2.1 Theoretical Review

Topic modeling, a technique within text mining, serves to dissect texts, unveiling interconnected themes, tracing topic evolution, and aiding in text summarization or further research development [10]. Among the popular methods for topic modeling identified by Zhang et al. [11], including Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF), LDA stands out as widely used for its probabilistic extraction of interrelated topics from textual data. Meanwhile, LSA employs principal component analysis to transform text into a matrix for mathematical scrutiny, and NMF dissects text documents into non-negative components representing internal document topics. This research opts for LSA, a method previously proven effective in analyzing review data, as it offers a friendly explanation of social science analysis within the general linear model [7]. However, LSA has limitations, including dependence on high-quality training data, incapability to capture multiple word meanings accurately, and constraints in expanding latent topic dimensions due to matrix rank limitations. These challenges emphasize the importance of considering task-specific requirements when selecting a topic modeling method for text analysis.

To gauge the quality of the model-generated topics, a coherence score serves as a metric, assessing the efficacy of the topic modeling technique. This score calculates semantic similarity among words within a topic, presuming that well-constructed topics will comprise semantically associated words. The significance of coherence score as an evaluation metric, highlighting its use in measuring topic model quality [12]. Several coherence score metrics such as C_v, C_umass, C_uci, and C_npmi evaluate topics by segmenting words into word pairs and computing word probabilities based on the formed corpus. These metrics subsequently aid in deriving confirmation measures and calculating coherence scores for the topics.

2.2 Knowledge Discovery in Databases (KDD)

This research adopts the Knowledge Discovery in Database (KDD) framework. According to Devedzic, KDD is a structured and holistic method for uncovering knowledge from data [13]. This framework involves a series of clearly defined stages, from project initiation to result evaluation. The stages in KDD include initiation, selection, pre-processing, transformation, data mining, and evaluation.

During the selection stage, data crawling is performed from Google Maps using the Python programming language and the Selenium library. The extracted data consists of Indonesian-language reviews of natural tourist attractions in the Bandung Barat Regency based on the list provided on the visitkbb.bandungbaratkab.go.id website. The crawled data is stored in a database tool, namely MongoDB, which is later exported into a CSV format for processing. The accumulated data from the crawling process amounts to 13,587 entries. The results obtained from this process are summarized in **Table 1**.

Table 1. Data scrapping results.

Id	Location	Datetime	Scrapped_at	Rating	Text
772bdxxx	Curug Malela	15/09/2022 11:58	15/02/2023 11:58	5	Nyampe 17:30 lnjut kebawah jam 18:00 mantap suasana 1 safar,..
8b066xxx	Curug Tilu Leuwi Opat	15/12/2022 12:24	15/02/2023 12:24	3	Lumayan. Tapi kurang klik karena airnya keruh.

In the pre-processing stage, scraped data is cleaned and standardized for consistency and ease of use. It involves steps like Data Filtering (sorting based on specific criteria), Data Cleaning (removing inconsistencies and irrelevant info), Stopword Removal (deleting meaningless standalone words), Stemming (converting words with affixes to base forms), and Tokenization (breaking data into smaller units or 'tokens'). Examples of applying the pre-processing stage to the data can be seen in **Table 2**.

Table 2. Application of pre-processing on raw data.

Step	Before	After
Data Filtering	Ada tempat peristirahatan. Jalan menuju Curug ini lumayan terjal, curug ini dijuluki sebagai Mini Niagara.	[Deleted because the review provided resembles Wikipedia]
Data Cleaning	(Diterjemahkan oleh Google) Cocok untuk kegiatan outdoor seperti hiking dan camping 👌 👌 (Asli) Perfect for outdoor activities like hiking and camping 👌 👌	cocok untuk kegiatan outdoor seperti hiking camping
Stemming	cocok kegiatan outdoor hiking camping	cocok untuk giat outdoor seperti hiking camping
Stopword Removal	cocok giat outdoor hiking camping	cocok giat outdoor hiking camping
Tokenization	cocok giat outdoor hiking camping	['cocok', 'giat', 'outdoor', 'hiking', 'camping']

The next stage is Transformation. Transformation is the process of converting text, images, graphics, time series, and others into numerical representations in the form of vectors. For text data, there are two types of transformations possible: using the bag of words method and TF-IDF. The bag of words is a way to represent a model as a vector by counting the frequency of word occurrences in a document [13]. Meanwhile, Term Frequency-Inverse Document Frequency or commonly known as TF-IDF is a method to represent data as a vector by calculating the weight of words in a document [15].

The data mining stage involves extracting information from the utilized data. For this research, the data mining stage utilizes the latent semantic analysis (LSA) method. Latent Semantic Analysis (LSA) is an analytical method used to extract and analyze the structure of a text collection [9]. By employing a mathematical technique called singular value decomposition (SVD), LSA can identify patterns of words that appear together in documents within a corpus. Mathematically, the equation for SVD can be written as follows:

$$A = U\Sigma V^T \quad (1)$$

$$U \in R^{(m\times k)}, V \in R^{(m\times k)} \quad (2)$$

The formula above contains several notations, each with its explanation. The notation V^T represents the transpose matrix of the column vector matrix V, which stores the word vector representations in the LSA vector space. By multiplying the matrices Σ and V^T, we can obtain a matrix that presents the probability of word occurrences within the topics identified by the LSA method. These topics in the LSA method consist of clusters of semantically related words, where the words within these clusters have a high correlation with each other. Consequently, the results from the LSA process offer a deeper understanding of the structure and hidden patterns within the text corpus. Information regarding word clusters and correlations unveiled by the LSA method can be used for various purposes such as text analysis, document classification, recommendation systems, and a more holistic understanding of content. Additionally, the vector-based approach in LSA can help tackle high-dimensional problems and provide more condensed and efficient word representations for various text analysis tasks and natural language processing.

The final stage in the KDD framework is evaluation, where the model is assessed to determine if it can be categorized as a good model or not. To evaluate a topic modeling process, this research employs the coherence score method. According to Morstatter and Liu [16], coherence score is an evaluation metric that measures the semantic relationship among the top words within a topic. This metric aids in identifying topics that can be semantically interpreted versus those that are statistical inference artifacts.

The coherence score, measured by Cohesion Value (CV), assesses the quality of topics based on their interpretability by humans, ranging from 0 to 1. A higher score implies that the topic is easier for humans to interpret. There are several steps to determine the coherence score using the CV metric.

1. Determining the keywords appearing within the topic, denoted as T.

$$T_n = \{T_{n,0}, T_{n,1} \dots T_{n,m}\} \tag{3}$$

2. Segmenting T, generating sets of subsets using sliding windows technique (S).

$$S = \{(T', T^*), T', T^*, T^* \subseteq T\} \tag{4}$$

3. Calculating the word probability values with NPMI, where P(T', T*) represents the probability of occurrence of sliding windows (T', T*), and ε is epsilon as a constant to avoid zero logarithms.

$$NPMI(T', T^*) = \frac{\log\frac{P(T',T^*)+\varepsilon}{P(T')P(T^*)}}{-\log(P(T',T)+\varepsilon)} \tag{5}$$

4. Computing cosine similarity using the NPMI score, where $\vec{v}$ is the topic word vector value, and $\vec{w}$ is the topic vector value.

$$\vec{v}_{x,k} = NPMI(T^n_{x,k}, T^n_{y,k}), \forall m \in \{1,2,\dots,N\} \tag{6}$$

5. Aggregating cosine similarity scores to determine the topic's coherence score.

$$s_{cos}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| \times |\vec{w}|} \tag{7}$$

After undergoing processing, out of the total 13,587 collected data in the selection stage, it was reduced to 3,093 clean data points that can be further utilized for model creation. To ensure the validity of this research's outcomes, several potential challenges when gathering data from online platforms like Google Maps need consideration. Firstly, the quality of online data is a crucial factor. Extracting data from tourist reviews on Google Maps holds risks such as spelling errors, irrelevant reviews, or inaccurate information. Therefore, testing and sorting of data are necessary to ensure the quality of the utilized data. Lastly, it's important to consider user biases in reviews, where personal experiences and individual perspectives may influence the analysis results.

3 Result and Discussion

3.1 Testing Scenarios

This study aims to explore the impact of stopwords and model representations on the topics derived through text analysis using Latent Semantic Analysis (LSA). Six scenarios are formed from various combinations of stopwords usage: No Stopword (NS), Stock Stopwords (SS), and Advance Stopwords (AS), coupled with two model representation types: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). All scenarios are conducted with a predefined number of topics (K) set at 10 based on practicality, ensuring a balanced level of detail in pinpointing document themes while still being manageable for interpretation and analysis [17]. The significance of stopwords in text analysis lies in their tendency to convey less meaningful information, such as conjunctions and articles, potentially disrupting the identification of more significant semantic patterns. NS scenario disregards stopwords, SS retains predetermined stopwords, while AS utilizes enriched stopwords to filter out additional irrelevant words. Additionally, the model representation plays a pivotal role in LSA, where different models can yield diverse outcomes. BoW, a simple yet effective method based on word frequency, and TF-IDF, which assesses word significance based on their frequency in documents and the entire corpus, are chosen as the comparison models. The testing model scenarios' combinations for this research are detailed in **Table 3**.

Table 3. Research testing scenarios.

Scenario	NS	SS	AS	BoW	TF-IDF
Scenario 1	✓	×	×	✓	×
Scenario 2	×	✓	×	✓	×
Scenario 3	×	×	✓	✓	×
Scenario 4	✓	×	×	×	✓
Scenario 5	×	✓	×	×	✓
Scenario 6	×	×	✓	×	✓

3.2 Experiment Results

Based on the testing results of the conducted scenarios, the coherence score for each number of topics in each scenario was obtained and can be seen at **Error! Reference source not found.**.

Table 4. Scenario testing results.

Number of Topics	S1	S2	S3	S4	S5	S6
2	0.689	0.542	**0.717**	**0.492**	**0.452**	0.452
3	**0.709**	**0.671**	0.680	0.459	0.367	0.416
4	0.596	0.615	0.672	0.429	0.421	**0.537**
5	0.577	0.557	0.624	0.414	0.402	0.470
6	0.554	0.616	0.643	0.398	0.450	0.503
7	0.502	0.567	0.613	0.350	0.410	0.413
8	0.533	0.552	0.625	0.369	0.422	0.421
9	0.541	0.609	0.572	0.417	0.387	0.433
10	0.538	0.509	0.581	0.366	0.408	0.438

Based on the results presented in Tables 4, it can be observed that scenario 3, involving the advanced stopword method and Bag of Words (BoW) model representation, achieved the highest coherence score of 0.717. From these findings, it can be concluded that the stopword removal process and the chosen model representation significantly impact the quality of the topics generated by the model.

3.3 Exploration of Model Results

Out of the 3,093 tourist reviews on natural waterfalls in the Bandung Barat Regency, a total of 2 topics were derived from modeling using advanced stopword removal and Bag of Words (BoW) model representation. The keywords appearing in these topics can be seen in **Error! Reference source not found.**.

Table 5. Keywords in selected topics.

Topic	Keywords
1	jalan, masuk, lokasi, parkir, tiket, indah, alam, foto, bayar, akses, motor, naik, hati, turun, pandang
2	jalan, foto, indah, alam, spot, pandang, bayar, sejuk, tiket, masuk, anak, fasilitas, makan, area, kolam

Deduced from the keywords, it's evident that tourists often discuss the accessibility of natural attractions, particularly entry routes, safety in accessing these spots, especially for two-wheeled vehicles, and the terrain encountered, indicated by keywords such as "jalan", "masuk", "lokasi", "motor", "naik", "hati", and "turun". Additionally, visitors frequently mention concerns about parking locations and ticket booths at waterfall attractions, reflected by the appearance of keywords like "lokasi", "parkir", "tiket", and "bayar". Furthermore, tourists often review the scenic beauty and available photo spots to enjoy these vistas, as indicated by keywords such as "foto", "indah", "alam", "spot", and "pandang". Lastly, with keywords like "facilities," "food," "area," and "pool," tourists discuss the amenities provided at waterfall attractions, including dining areas and swimming pools to ensure visitor safety. Interestingly, some words appear in both topics, like "jalan", "indah", "foto", etc. This occurrence might stem from imperfections in the model, causing some noise levels or ambiguity in topic determination and similarity concepts across topics might share significant words. Based on these keywords, recommendations can be made to improve and enhance the quality of the waterfall natural attractions in West Bandung Regency, as shown in **Error! Reference source not found.**.

Table 6. Recommendations for tourism object improvement strategies.

Keywords	Topic	Keywords
Jalan, Masuk, Motor, Naik, Turun	Increasing Accessibility of Natural Tourist Attractions	Improve road conditions to tourist attractions Improving road safety to natural tourist attractions
Lokasi, Parkir, Tiket, Bayar	Improvements to Parking Locations and Payment Systems	Improve the condition of inadequate parking lots Increase parking capacity Provide an easy and efficient ticket payment system Standardize prices to prevent fraud
Foto, Indah, Alam, Spot, Pandang	Photo Spot Improvement	Add photo spots at tourist attractions Repair existing photo spots Make regulations to maintain the beauty of tourist attractions
Fasilitas, Makan, Area, Kolam	Provision of Facilities	Improve the quality of public facilities at tourist attractions Provide a swimming area for tourists to swim Add dining facilities at tourist attractions

4 Conclusion

Based on the analysis conducted, it was found that the most effective pre-processing method and model representation for processing reviews data using latent semantic analysis (LSA) are the advanced stopword technique and the bag of words model. This was evidenced by the evaluation results using the coherence score metric, where the highest score was achieved with 2 topics in scenario 3, scoring 0.717, indicating the ease of topic interpretation by humans. From the interpretation of these topics, several recommendations can be suggested to improve and enhance the quality of natural waterfall attractions in the Bandung Barat Regency. Suggestions include improving road conditions and safety to these attractions, enhancing parking facilities, standardizing attraction prices, adding photo spots, improving existing photo spots, and upgrading the available facilities at these attractions. For future research on the same topic, it's recommended to broaden the data sources to bolster data quality and explore additional scenarios to gauge their influence on the model's outcomes.

References

1. kumparanNEWS. (2021, Mei 17). Ramai-ramai Tutup Tempat Wisata yang Tuai Polemik karena Sempat Buka saat Corona. From kumparan.com: https://kumparan.com/kumparannews/ramai-ramai-tutup-tempat-wisata-yang-tuai-polemik-karena-sempat-buka-saat-corona-1vkyETYw3CJ/full
2. Kompas. (2022, Februari 26). Merencanakan Strategi Sektor Pariwisata Menghadapi 2022. From kompas.com: https://www.kompas.id/baca/adv_post/merencanakan-strategi-sektor-pariwisata-menghadapi-2022
3. Rasti. (2021, Oktober 30). Masa Pandemi Bawa Perubahan Preferensi Wisatawan, Kini Lebih Inginkan Pariwisata Berkualitas. From mnews.co.id: https://mnews.co.id/read/fokus/masa-pandemi-bawa-perubahan-preferensi-wisatawan-kini-lebih-inginkan-pariwisata-berkualitas/
4. Antara. (2022, Februari 16). Strategi Jawa Barat Raih Target 40 Juta Wisatawan pada 2022. From tempo.co: https://travel.tempo.co/read/1561368/strategi-jawa-barat-raih-target-40-juta-wisatawan-pada-2022
5. Taufik, D. H. (2021). Perubahan Rencana Strategis Dinas Pariwisata dan Kebudayan Provinsi Jawa Barat Tahun 2018-2023. Bandung: Dinas Pariwisata dan Kebudayan Provinsi Jawa Barat.
6. Ali, T., Omar, B., & Soulaimane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. MethodsX, 9, 101894. https://doi.org/10.1016/j.mex.2022.101894
7. D. Valdez, A. C. Pickett, and P. Goodson, “Topic Modeling: Latent Semantic Analysis for the Social Sciences,” Soc Sci Q, vol. 99, no. 5, pp. 1665–1679, Nov. 2018, doi: 10.1111/ssqu.12528.
8. Syn, W. T., How, B. C., & Atoum, I. (2014). Using *Latent semantic analysis* to Identify Quality in Use (QU) Indicators from User Reviews. The International Conference on Artificial Intelligence and Pattern Recognition (AIPR2014), 143–151.
9. Sanjifa, Z. N., Sumpeno, S., & Suprapto, Y. K. (2019). Community Feedback Analysis Using *Latent semantic analysis* (LSA) To Support Smart Government. International Seminar on Intelligent Technology and Its Applications (ISITIA), 428-433.
10. Habibie, F., Kusumawardani, S. S., Prakasa, B., Putra, D. W., Widhiyanto, B. T., & Widyawan. (2017). Big Data Analytic for Estimation of Origin-Destination Matrix in Bus Rapid Transit System. *3rd International Conference on Science and Technology-Computer (ICST)*, 165-170.
11. Zhang, Y., Liu, Z., Chen, Y., & Ma, J. (2018). A Comparative Study of Latent Dirichlet Allocation and Non-Negative Matrix Factorization on Sentiment Analysis of Movie Reviews. *Journal of Convergence Information Technology*, 122-129.

12. Ajinaja, M., Adetunmbi, A. O., Ugwu, C. C., & Popoọla, O. S. (2022). Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling. *Iran Journal of Computer Science*.
13. Devedzic, Vladan. (2001). KNOWLEDGE DISCOVERY AND DATA MINING IN DATABASES. Handbook of Software Engineering and Knowledge Engineering, 615-637.
14. Brownlee, J. (2017, Oktober 9). A Gentle Introduction to the Bag-of-Words Model. From Machine Learning Mastery: https://machinelearningmastery.com/gentle-introduction-bag-words-model/
15. Saha, R. (2021, Januari 20). Understanding TF-IDF (Term Frequency-Inverse Document Frequency). From Geeks for Geeks: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/
16. Morstatter, Fred., Liu, Huan. (2018). In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics. Journal of Machine Learning Research 18, 1-32.
17. Naushan, Haaya. (2023). Topic Modeling with Latent Dirichlet Allocation. from Towards Data Science: https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8.
18. Bail, Chris (2019) Topic Modeling. From Sicss: https://sicss.io/2019/materials/day3-text-analysis/topic-modeling/rmarkdown/Topic_Modeling.html