
Sentiment Analysis and Topic Modeling of Tourist Attractions in Gresik Regency Using the BERT Method

Salsabilla Putri Saharani¹, Wiyli Yustanti²

^{1,2}*Universitas Negeri Surabaya*

salsabilla.20012@mhs.unesa.ac.id, wilyliyustanti@unesa.ac.id

ABSTRACT

Tourism is one of the key sectors in national economic growth as well as a pillar of regional community welfare. Gresik Regency in East Java has considerable tourism potential, with more than 100 destinations covering religious tourism, natural attractions, and family recreation. However, tourist visit data from 2022–2024 shows a declining trend that requires an in-depth evaluation of visitor perceptions and experiences. This study aims to analyze public sentiment toward tourist destinations in Gresik Regency and identify the main topics of concern for tourists. The research data was collected from Twitter and Google Maps within the period of 2021–2024 using crawling techniques. Sentiment analysis was carried out with IndoBERT, while topic modeling was conducted using BERT. The results indicate that tourism reviews are dominated by positive sentiments highlighting the uniqueness of religious destinations, natural beauty, and family recreation atmosphere. However, negative sentiments were also found, emphasizing issues related to facilities, cleanliness, staff services, and accessibility to the sites. Topic modeling successfully grouped tourist opinions into coherent themes, and evaluation with coherence scores demonstrated good quality outcomes. The study concludes that although Gresik has strong tourism appeal, challenges in facility management, services, and digital promotion need to be addressed immediately. The integration of sentiment analysis and BERT-based topic modeling has proven effective in providing comprehensive insights into tourist perceptions and can serve as a basis for formulating regional tourism development strategies.

Keyword: Sentiment Analysis, BERT, Gresik, Tourism, Topic Modeling.

Article Info:

Article history:

Received September 20, 2025

Revised January 23, 2026

Accepted February 12, 2026

Corresponding Author

Salsabilla Putri Saharani

Universitas Negeri Surabaya

salsabilla.20012@mhs.unesa.ac.id

1. INTRODUCTION

A tourism destination is a regional area that possesses tourist attractions, facilities, accessibility, infrastructure, and a local community that are interconnected and share responsibility for realizing tourism in that region, with the aim of providing products and services for visitors [1]. Tourism is an important sector contributing to national economic growth, as it ranks third among the largest foreign exchange-generating sectors in Indonesia, including in regions such as Gresik Regency, East Java [2]. Gresik has more than 100 tourist destinations; more precisely, in 2022, the number of recorded attractions in Gresik reached 131, ranging from religious to natural tourism [3]. However, data from the official Gresik

Regency online tourism website (dakuwison.gresikkab.go.id) shows a decline in tourist visits during the 2022–2024 period, which has the potential to hinder tourism's contribution to local community welfare. This can be seen in the presence of historical religious tourism sites, such as the tombs of the Wali Songo namely Sunan Gresik (Maulana Malik Ibrahim) and Sunan Giri which attract thousands of pilgrims annually from various regions across Indonesia [4]. In the digital era, public opinion about tourism services is widely expressed through social media and review platforms. DataReportal reports that 41.2% of internet users aged 16–64 use the internet to share their opinions. Therefore, analyzing online opinions can be a strategic approach to understanding public perceptions and improving the quality of tourism services.

The main issue that arises is how to obtain a comprehensive understanding of public sentiment toward tourism in Gresik, as well as the specific issues most frequently highlighted by the public. This study formulates three key research questions: (1) what are the results of sentiment analysis on tourism in Gresik Regency using the BERT method, (2) what are the results of topic modeling on tourism reviews, and (3) what recommendations can be provided to stakeholders to improve the quality of tourist destinations. A study entitled “*The Impact of Tourism on Poverty Levels in Five ASEAN Countries*” revealed that investment, employment absorption, and financial inclusion generated by the tourism sector can reduce poverty levels in five ASEAN countries (Indonesia, Malaysia, Vietnam, the Philippines, and Thailand) [5].

Previous studies have applied Bidirectional Encoder Representations from Transformers (BERT) in sentiment analysis across various contexts, including reviews of tourist attractions in Yogyakarta, achieving an accuracy of more than 92%. To obtain meaningful insights from available data, sentiment analysis is required. Sentiment analysis is the process of automatically extracting, processing, and interpreting unstructured textual data to capture sentiment information contained in statements or opinions [6]. One method that can be used for sentiment analysis is BERT. The Bidirectional Encoder Representations from Transformers (BERT) model is a Natural Language Processing (NLP) framework. The BERT method has been proven to produce topics with better coherence compared to traditional methods such as LDA or NMF. Sentiment analysis using IndoBERT has achieved an accuracy of 92.84%, with a weighted average recall of 93%, precision of 92%, and an F1-score of 93% [7]. These studies demonstrate the effectiveness of BERT in text classification; however, they are generally limited to a single data source or have not yet combined sentiment analysis with topic modeling in the context of local tourism. In addition, BERT also has a variant specifically designed for topic modeling, known as BERTopic. BERTopic is a transformer-based BERT model that specializes in topic modeling using the c-TF-IDF approach [8].

This study proposes a BERT-based analytical approach to process tourism review data from Gresik Regency, collected from Twitter and Google Maps during the 2021–2024 period. Public sentiment is analyzed using IndoBERT, while topic modeling is conducted with BERT. The results are expected to identify public perceptions in the form of positive, negative, and neutral sentiments, while also uncovering the main themes underlying those opinions.

The novelty of this research lies in the integration of two analytical approaches—sentiment analysis and topic modeling—by leveraging the BERT model in the context of regional tourism. BERTopic is well known for its ability to generate coherent topics while maintaining competitiveness across various benchmark evaluations [9]. This approach not only maps public perceptions but also provides deeper thematic insights, which can be utilized by local governments and tourism stakeholders to develop strategies for improving service quality, facility management, and digital promotion. Thus, this study is expected to make a tangible contribution to data-driven tourism development efforts in Gresik Regency.

2. METHODS

This research is a quantitative study with a text mining approach in the field of Natural Language Processing (NLP), aiming to analyze sentiments and topics in tourism reviews of Gresik Regency. The research specifications follow the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, which consists of the stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment [10]. Data were collected through crawling techniques from two main sources, namely Twitter and Google Maps, within the 2021–2024 period and limited to reviews written in Indonesian. The data processing stage involved text preprocessing, including data cleaning, normalization, tokenization, and stopword removal to ensure input quality [11]. Sentiment analysis was performed using the IndoBERT model, a BERT variant optimized for the Indonesian language [12]. IndoBERT has proven capable of conducting sentiment analysis in Indonesian with effective transfer learning performance [13]. Meanwhile, topic modeling was applied using the BERT algorithm, which leverages c-TF-IDF representations to generate more coherent topics [14]. Model evaluation was conducted using accuracy metrics for sentiment classification and coherence score (Cv) to assess the quality of the generated topics [15]. Topic modeling is a textual data analysis methodology proven effective in identifying patterns and trends in media discussions on AI and related technologies [16]. Various topic modeling methods exist, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic [17]. One of the important steps in topic modeling is document embedding, a process in which the processed text is transformed into numerical vectors of a fixed length [18].

With this design, the study is expected to provide a comprehensive overview of public opinion as well as the main issues surrounding tourism in Gresik Regency.

3. RESULTS AND DISCUSSION

The research data were obtained from two main sources, namely Twitter and Google Maps, covering the period of 2021–2024. Through the crawling process, thousands of reviews related to tourist destinations in Gresik Regency were collected. After undergoing preprocessing stages (text cleaning, normalization, tokenization, and stopword removal), the amount of data available for analysis was reduced; however, the quality of the text improved, making it more suitable for processing in the subsequent stages. The number of data points from each source and their distribution can be seen in the following figure:

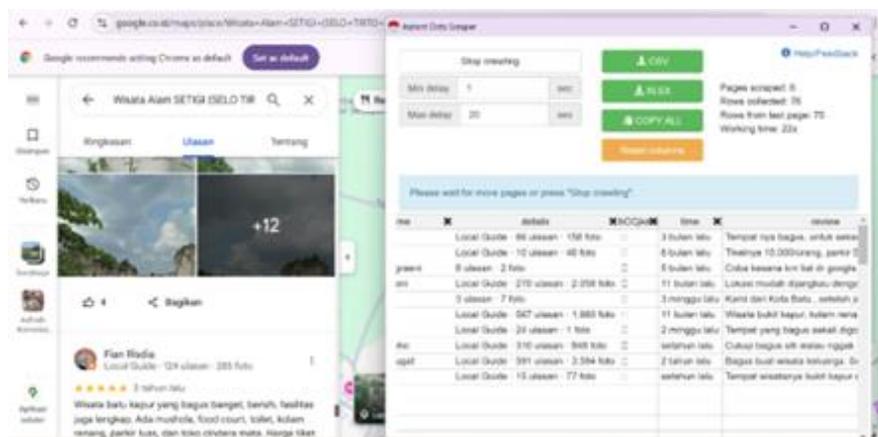


Figure 1. The Number Of Data Points

3.1 Sentiment Analysis with IndoBERT

The results of sentiment classification using IndoBERT show that tourism reviews of Gresik Regency are divided into three main categories: positive, negative, and neutral. In general, positive sentiment dominates the reviews, indicating that the public perceives Gresik as having attractive tourism potential. However, negative sentiments are also present, highlighting issues related to services, cleanliness, and the lack of facility maintenance. The percentage distribution of sentiments for each tourist attraction can be seen in the following table:

Table 1. Percentage Distribution of Tourist Attraction Sentiment

Num.	Name of Tourist Destination	Number of Sentiment Data		
		Netral	Positif	Negatif
1.	Makam Malik Ibrahim	161	134	17
2.	Makam Sunan Giri	251	104	43
3.	Malindo Swimming Pool	26	170	40
4.	Lontar Sewu Tourism Edu	17	191	45
5.	Selo Tirto Giri Tourism	22	193	47
6.	Gresik Tourism	1.026	1.058	317
Total		1.503	1.850	509

The results of sentiment analysis using the IndoBERT model show that tourism reviews of Gresik Regency are divided into three main categories: positive, negative, and neutral. Overall, reviews with positive sentiment are more dominant than the other two categories. The dominance of positive sentiment indicates that tourists perceive Gresik as having strong tourism appeal, both in terms of culture, religion, and natural recreation. This also reflects public appreciation of Gresik's tourism potential, which is considered attractive and worth revisiting.

Nevertheless, negative sentiments still appear in a significant proportion. Reviews with negative sentiment generally highlight several shortcomings experienced by tourists, particularly related to service quality, environmental cleanliness, and the lack of maintenance of tourism facilities. Some destinations even received criticism regarding accessibility conditions to the sites, which were considered inadequate. This illustrates that although the tourism appeal is relatively strong, managerial aspects and facility management still need improvement to ensure a more optimal visitor experience.

To provide a more comprehensive understanding of public perceptions, the analysis results are also complemented by word cloud visualizations displaying the most frequently occurring keywords in the reviews. For instance, in positive reviews of the Sunan Giri Tomb destination, the most prominent words include *"clean," "crowded,"* and *"comfortable,"* reflecting tourist satisfaction with the atmosphere and facilities. Conversely, negative reviews at the same location more frequently feature the words *"traffic jam"* and *"dirty,"* indicating complaints about traffic congestion around the tourist area as well as insufficient environmental cleanliness. The word cloud visualization provides visual evidence that tourist perceptions are strongly influenced by cleanliness and comfort. These two aspects serve as the primary

determinants in shaping both positive and negative impressions among visitors. Therefore, efforts to improve tourism in Gresik Regency should focus on facility maintenance, cleanliness management, and the provision of friendly and professional services.

The results of topic modeling on positive and negative sentiments regarding tourism in Gresik Regency are then visualized using a bar chart, as shown in the following figure:

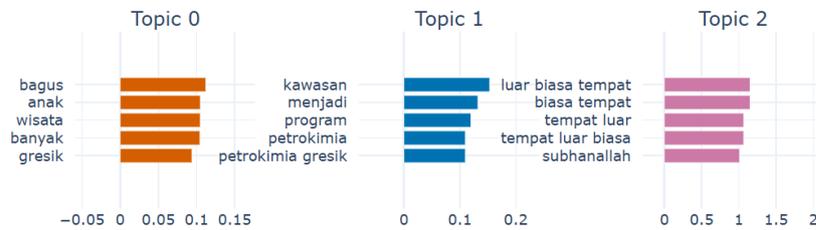


Figure 2. Positive Sentiment Modeling Topic Results

The predicted positive and negative sentiments in each document are visualized to facilitate understanding and to identify frequently occurring words. In this study, the visualization is presented using a WordCloud:



Figure 3. World Cloud Sentimen Negatif (Makan Malik Ibrahim)



Figure 4. World Cloud Sentimen Positif (Makam Malik Ibrahim)



Figure 5. World Cloud Sentimen Negatif (Makam oSunan Giri)



Figure 6. World Cloud Sentimen Positif (Makam Sunan Giri)



Figure 7. World Cloud Sentimen Negatif (Malindo Swimming Pool)



Figure 8. World Cloud Sentimen Positif (Malindo Swimming Pool)



Figure 9. World Cloud Sentimen Negatif (Lontar Sewu Tourism Edu)



Figure 10. World Cloud Sentimen Positif (Lontar Sewu Tourism Edu)

3.2 Topic Modeling with BERT

The results of topic modeling with BERT show that tourism reviews of Gresik Regency can be grouped into several major themes, which differ between positive and negative sentiments. In the positive sentiment category, the most dominant topic relates to the uniqueness of religious destinations, such as the Sunan Giri Tomb and the Malik Ibrahim Tomb, which are highly praised for their historical and spiritual value. In addition, topics concerning the natural beauty of several destinations, such as Selo Tirto and Lontar Sewu, also emerge, where tourists emphasize enjoyable recreational experiences, scenic views, and environments that support family activities. Another prominent topic is family recreation, with reviews highlighting playground facilities, swimming pools, and child-friendly spaces. This confirms that tourists are not only seeking religious values or natural panoramas but also comfort for family-oriented activities.

Meanwhile, in the negative sentiment category, the topics that arise reflect the challenges faced in Gresik’s tourism management. The first topic concerns facilities, including complaints about the cleanliness of public toilets, the availability of parking spaces, and poorly maintained supporting infrastructure. The second topic is staff service, where tourists criticize the lack of professionalism, friendliness, or readiness of officers to assist visitors. The third topic is accessibility to the sites, which is often considered difficult due to narrow, congested,

or poorly maintained roads. These three negative topics indicate that although Gresik's tourism attractions are appealing, the quality of management remains a key challenge for stakeholders.

To ensure the quality of the generated topics, an evaluation was conducted using the coherence score (Cv). The results indicate that the topics formed have good coherence, meaning that the words within each topic are semantically related and align with the context of the reviews. In addition, this study also carried out further validation through the word intrusion task and topic intrusion task. In the word intrusion task, respondents were asked to identify an irrelevant word from a set of words within one topic. The findings show that the majority of respondents were able to recognize the inconsistent word, suggesting that the topics formed are semantically consistent. In the topic intrusion task, respondents were asked to match review excerpts with the appropriate topic. The high level of agreement among respondents further reinforces that the topics generated by BERT accurately reflect the content of the reviews.

Overall, these findings demonstrate that the BERT approach is not only capable of categorizing tourist opinions into clear groups but also of producing in-depth insights into the factors that attract or hinder tourism in Gresik Regency. Positive topics can serve as a foundation for promotional strategies and tourism product development, while negative topics provide valuable input for policy evaluation and service improvement in the future.

3.3 Discussion of Findings

The findings of this study indicate that Gresik Regency possesses strong tourism appeal in terms of religion, culture, and nature. The dominance of positive sentiments in tourist reviews suggests that the public holds high appreciation for the available destinations. The uniqueness of religious sites such as the Sunan Giri Tomb and the Malik Ibrahim Tomb, along with the natural tourism potential of Lontar Sewu and Selo Tirto, serves as an important asset for regional tourism development. This aligns with the concept of sustainable tourism, which emphasizes cultural and environmental preservation as key assets to attract tourists in the long term.

Nevertheless, the significant presence of negative sentiments reveals serious challenges that need to be addressed promptly. The most frequently highlighted issues by tourists concern the lack of maintenance of public facilities, including the cleanliness of toilets, the condition of parking areas, and other supporting infrastructure. Furthermore, the professionalism and hospitality of staff are also critical aspects, as tourists consider the quality of interactions with personnel to greatly influence the overall travel experience. Another important factor is accessibility to the sites, with several destinations perceived as difficult to reach due to traffic congestion, narrow roads, or inadequate transportation facilities. These findings are consistent with Subarkah et al. (2022), who emphasize that service quality and facilities are key factors in maintaining tourist satisfaction and ensuring tourism sustainability.

The findings of this study also provide practical implications for the development of tourism strategies in Gresik Regency. First, local government and tourism managers need to prioritize the improvement of public facilities, particularly in terms of cleanliness and the comfort of amenities used by visitors. Second, enhancing the capacity of human resources in the tourism sector is essential, including training in hospitality, professionalism, and staff responsiveness in serving tourists. Third, digital promotion must be optimized, given that most tourists seek information and share experiences through social media and online review platforms. This effort will be more effective if combined with destination branding strategies that highlight Gresik's religious uniqueness and natural beauty.

By integrating sentiment analysis and topic modeling, this study offers a solid foundation for formulating data-driven policies. Local government and tourism stakeholders

can use these findings as a reference for conducting evaluations, designing improvement programs, and directing more targeted promotional efforts. In this way, tourism development in Gresik can become more adaptive to tourist needs while simultaneously contributing to the improvement of local community welfare.

CONCLUSION

This study has successfully addressed the research questions regarding sentiment analysis and topic modeling of tourism reviews in Gresik Regency using the BERT method. First, the sentiment analysis results with IndoBERT revealed that the reviews were dominated by positive sentiments, highlighting Gresik's strong potential as a tourism destination rooted in culture, religion, and nature. However, negative sentiments were also identified, exposing weaknesses in facilities, services, and accessibility. Second, the topic modeling results with BERTopic identified key themes in the reviews, including the uniqueness of religious destinations, natural beauty, family recreation atmosphere, as well as issues of cleanliness, service quality, and infrastructure. These findings demonstrate that the integration of sentiment analysis and topic modeling provides more comprehensive insights compared to conventional methods.

Theoretically, this study contributes to enriching the application of the BERT model in the context of regional tourism, particularly by simultaneously combining sentiment classification and topic mapping. This contribution underscores that deep learning-based approaches can serve as a reliable method for understanding public opinion derived from social media and online platforms. For future research, this approach has the potential to be further developed by incorporating data from more digital platforms, comparing performance with other algorithms, or expanding the scope of the study area. Thus, this research not only provides practical recommendations for tourism development in Gresik but also opens opportunities for further studies in public opinion analysis using NLP.

REFERENCES

- [1] Eddyono, F. (2021) *Pengelolaan Destinasi Wisata, Pengelolaan Destinasi Pariwisata*. Ponorogo: Uwais Inspirasi Indonesia.
- [2] Hasibuan, I.M. *et al.* (2023) "Kontribusi Sektor Pariwisata Terhadap Perekonomian Nasional," *urnal Masharif al-Syariah: Jurnal Ekonomi dan Perbankan Syariah*, 8(2), matter. 1200–1217. Available on: <https://journal.um-surabaya.ac.id/Mas/article/view/19280/pdf>.
- [3] Rinanda, H. M. (2022, 21 Januari). Gresik Ternyata Punya 131 Objek Wisata Lho, Yuk Kita Explore!. Retrieved September 6, 2025, from <https://www.detik.com/jatim/wisata/d-5908428/gresik-ternyata-punya-131-objek-wisata-lho-yuk-kita-eksplere>.
- [4] Kementerian Pariwisata dan Ekonomi Kreatif Republik Indonesia. (2022). *Pariwisata religi di Indonesia*. Jakarta: Kemenparekraf/Baparekraf RI. Accessed from <https://repository.upnjatim.ac.id/34306/2/20045010009-bab1.pdf>
- [5] Zainuri, Z. *et al.* (2021) "Dampak Pariwisata Terhadap Tingkat Kemiskinan di Lima Negara ASEAN," *e-Journal Ekonomi Bisnis dan Akuntansi*, 8(2), matter. 138. Available on: <https://doi.org/10.19184/ejeba.v8i2.26771>.
- [6] Suhendra, T., Intan, B., & Martadinata, A. T. (2024). Analisis Sentimen Pengguna

- Aplikasi Netflix Pada Ulasan Google Playstore Menggunakan Metode Naïve Bayes. *ESCAF*, 1011-1022.
- [7] Subarkah, M.Z. *et al.* (2022) “Analisis Sentimen Review Tempat Wisata Pada Data Online Travel Agency Di Yogyakarta Menggunakan Model Neural Network IndoBERTweet Fine Tuning,” *Seminar Nasional Official Statistics*, 2022(1), matter. 543–552. Available on: <https://doi.org/10.34123/semnasoffstat.v2022i1.1246>.
- [8] Aryani, D. *et al.* (2024) “Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles,” *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, 9(2), matter. 171–180. Available on: <https://doi.org/10.25139/inform.v9i2.8429>.
- [9] Gupta, P. *et al.* (2024) “Generative AI: A systematic review using topic modelling techniques,” *Data and Information Management*, 8(2), matter. 100066. Available on: <https://doi.org/10.1016/j.dim.2024.100066>.
- [10] Chapman, P. *et al.* (1999) “The CRISP-DM process model,” *The CRISP-DM Consortium*, 310(C), matter. 91. Available on: <http://www.crisp-dm.org/Process/index.htm>.
- [11] Liu, B. (2015) *Sentiment analysis: Mining opinions, sentiments, and emotions, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press. Available on: <https://doi.org/10.1017/CBO9781139084789>.
- [12] Koto, F. *et al.* (2020) “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*. arXiv, matter. 757–770. Available on: <https://doi.org/10.18653/v1/2020.coling-main.66>.
- [13] Imaduddin, H. *et al.* (2023) “Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach,” *International Journal of Advanced Computer Science and Applications*, 14(8), matter. 113–117. Available on: <https://doi.org/10.14569/IJACSA.2023.0140813>.
- [14] Grootendorst, M. (2022) “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” in. Available on: <http://arxiv.org/abs/2203.05794>.
- [15] Röder, M. *et al.* (tanpa tanggal) “Exploring the Space of Topic Coherence Measures,” *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015.
- [16] Wibowo Octavianto, A. *et al.* (2024) “Framing The Future: Exploring AI Narratives in Indonesian Online Media Using Topic Modelling,” *Jurnal Komunikasi Indonesia*, 13(2), matter. 172–194. Available on: <https://doi.org/10.7454/jkmi.v13i2.1245>.
- [17] Egger, R. dan Yu, J. (2022) “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Frontiers in Sociology*, 7(May), matter. 1–16. Available on: <https://doi.org/10.3389/fsoc.2022.886498>.
- [18] Ahn, Y. *et al.* (2022) “Dual embedding with input embedding and output embedding for better word representation,” *Indonesian Journal of Electrical Engineering and Computer Science*, 27(2), matter. 1091–1099. Available on: <https://doi.org/10.11591/ijeecs.v27.i2.pp1091-1099>.