Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

Sentiment Analysis of Tourist Reviews at Waterfront City Pangururan Using Naive Bayes and TF-IDF Algorithm

Ichwanul Muslim Karo Karo[1,], Justaman Arifin Karo Karo[2], Selvia Djasmayena[3], Rizki Wahyudi [4]*

[1,4]Computer Science, Universitas Negeri Medan, Indonesia
[2]Politeknik Teknologi Kimia Industri, Indonesia
[3]Information System, STMIK Citra Mandiri Padangsidimpuan, Indonesia
Correspondence: E-mail: ichwanul@unimed.ac.id

ABSTRACT

This research aims to analyze the sentiment of tourist reviews on Waterfront City Pangururan tourist destination using Naive Bayes algorithm and TF-IDF method. Data was collected from 311 Google Maps reviews, then went through preprocessing stages such as data cleaning, tokenization, stopwords removal, and stemming. A classification model was built using Naive Bayes, with evaluation results showing an accuracy of 90.48% and a predominance of positive sentiments in the reviews. These results indicate that this approach is effective in automatically classifying travelers' opinions. The findings can be a valuable input for managers and local governments in improving the quality of tourism services and attractions.

ARTICLE INFO

Article History:
Submitted/Received
15 March 2025
First Revised 02 April 2025
Accepted 08 May 2025
First Available online
01 June 2025
Publication Date 01 June 2025

Keyword:
Naive bayes;
Sentiment analysis;
TF-IDF;
Tourist review;
Waterfront city Pangururan.

1. INTRODUCTION

Tourism has become Indonesia's leading sector in increasing the country's foreign exchange earnings (Putri et al., 2022). The importance of improving and developing tourism attractions is the key to attracting more tourists, both from within and outside the country. The comfort level of a tourist attraction is a key factor in accelerating the growth of the national tourism industry (Thaha & Aziz, 2020; Wijianto et al., 2024). Waterfront City Pangururan is a tourist destination that highlights the natural beauty of the Lake Toba region and has great potential to attract domestic and foreign tourists. With more and more tourists interested in visiting this destination, digital media such as social media platforms, online reviews, and blog sites have become the main data sources to understand visitors' opinions and experiences. Sentiment analysis of traveler reviews has proven to be very important in helping managers and local governments understand the visitor experience and identify aspects of services and facilities that need to be improved or maintained (Wijayanto et al., 2022).

In the era of information technology and the internet, access to tourist information and reviews is now faster, easier, and more accurate through online map platforms and digital services based on geographic information systems. This ease of access not only supports travelers' trip planning, but also has a positive impact on increasing the number of visits and becoming an effective promotional medium for tourist destinations (Puana et al., 2023). Batam City, as one of the cities in Indonesia that continues to develop its tourism potential, has challenges related to the lack of promotion and information disclosure regarding the quality of tourism objects in Batam. The development of information technology, such as online maps and Google Maps, allows tourists to plan routes and provide reviews and ratings on the places they visit (Rifa'i et al., 2021). The sentiments expressed in this review can help managers and local governments to improve the quality of tourism services and infrastructure.

Sentiment analysis is an important process that aims to automatically identify and infer sentiment tendencies and intensity in text (Kiran Kumar et al., 2022). It involves recognizing user opinions from text and classifying them into various sentiments or emotions, such as positive, negative, or neutral, to determine the user's attitude towards an object or entity (Jin et al., 2023; Saepudin et al., 2024). Sentiment analysis is a text data processing process that aims to determine a person's attitude or opinion towards a certain topic. One algorithm that is often applied in this analysis is Naive Bayes, because this algorithm is known to be simple but quite effective in classifying text based on the probability of words that appear. Naive Bayes works using the Bayesian theorem, assuming that each data feature used is independent of each other (Karo et al., 2022). In the context of Pangururan Waterfront City tourism, the use of this method allows grouping visitor reviews into categories such as positive, negative, or neutral, which is useful for understanding the level of visitor satisfaction.

This research aims to analyze the sentiment of Pangururan Waterfront City visitor reviews using the Naive Bayes and Term Frequency - Inverse Document Frequency (TF-IDF) algorithms. Pangururan Waterfront City is a waterfront development project located in Pangururan, the capital of Samosir Regency, North Sumatra Province, Indonesia. The project is part of the government's initiative in the development of national priority tourism areas, particularly Lake Toba as a super priority destination. It is hoped that the results of this study can provide an overview of the extent of tourist satisfaction with these attractions and provide recommendations based on public opinion data.

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699

Several previous studies have been referenced in the development of this study. Research by (Friadi & Kurniawan, 2024) used 1,140 reviews from Google Maps and compared the Naive Bayes method with Support Vector Machine (SVM). As a result, SVM showed higher accuracy (94%) than Naive Bayes (83%), demonstrating the effectiveness of the SVM model in classifying traveler sentiment. Furthermore, study by (Putu et al., 2021) analyzed 9,496 tweets with Naive Bayes for sentiment classification and Latent Dirichlet Allocation (LDA) for topic modeling. They found that the majority of positive sentiment came from praising beaches such as Gili Trawangan, while negative sentiment mostly discussed hygiene issues. Another research (Rifa'i et al., 2021) developed a web-based system using Naive Bayes. The process includes pre-processing, TF-IDF weighting, and evaluation through confusion matrix. Model accuracy ranged from 38% to 76%, with a predominance of positive sentiments (f1score 0.73). The system was also tested using the blackbox method and received an interface satisfaction score of 84.85% (excellent category). These three studies demonstrate how sentiment analysis can be used to evaluate tourist satisfaction and provide data-driven input for tourism sector development.

The structure of the article section contains several sections. The research methods section explains how the research was conducted, including the design, data, and analysis techniques. Results and discussion present the main findings and their interpretation, in relation to theory or previous research. The conclusion summarizes the results, addresses the research objectives, and makes suggestions or recommendations.

2. RESEARCH METHOD

2.1 Data Collection

In this study, data was obtained from Google Maps by looking at visitor reviews at Waterfront City Pangururan. Data retrieval using Google Chrome, Google Maps by utilizing Instant Data Scaper to retrieve the data. Then, the reviews are labeled. Data labeling serves to provide markers or categories to the raw data, so that models or algorithms can be trained and used to make better predictions or decisions. Each comment is given a sentiment label based on the feelings contained in the comment, which is categorized into Positive, Negative, Neutral.

2.3 Preprocessing data

Data preprocessing is a very important step in sentiment analysis (Karo Karo & Hendriyana, 2022). It helps clean, organize, and simplify text to make it easier for algorithms to process. Existing review data is processed with several preprocessing stages.

- **Data cleaning**: Remove unnecessary characters such as punctuation, numbers, and other symbols.
- **Tokenization**: Break down each comment into individual words (tokens).
- **Stopwords Removal**: Eliminate common words that do not provide significant meaning to the analysis and conjunctions.
- **Stemming**: Transforming words into their base form to simplify the analysis process.

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699

2.4. Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) is one of the feature extraction methods by weighting words in a text (Karo Karo et al., 2023). It considers two things: the frequency of occurrence of the word in the document (Term Frequency, TF) and how rarely the word appears in the whole document (Inverse Document Frequency, IDF). The concept is to give higher weights to words that appear frequently in certain documents, but rarely appear in other documents.

Mathematically, the TF-IDF formula for word *t* in document *d* is presented in equation [1]. Term Frequency (TF) serves to measure the frequency of word occurrence in a particular document. The more often a word *t* appears in document *d*. Sedangkan Inverse Document Frequency (IDF) measures how rarely a word t appears across documents. The less the word appears in other documents, the greater the weight.

$$TF - IDF(t, d) = TF(t, d) \text{ x } (IDF(t) \quad (1)$$

2.5. Naïve Bayes

Naïve Bayes is one of the methods used in text mining that is simple but has high accuracy in classifying (Pristiyono et al., 2021). Naïve Bayes has a short classification time that speeds up the sentiment analysis process. The naïve bayes classification method is a classification method based on probability and bayes theorem (bayes rule) with strong independence assumptions. Bayes' theorem can be expressed in the form of an equation [2]. $P(C \mid X)$ represents the probability that sentiment class *C* (e.g. positive or negative) is true, given feature *X* (words in the text). $P(X \mid C)$ represents the probability that feature *X* appears, given class *C*. $P(C)$ represents the probability of class C from the entire dataset. $P(X)$ represents the probability of feature *X* in entire dataset.

$$P\ (C \mid X) = \frac{P\ (X \mid C).P(C)}{P(X)} \quad (2)$$

3. RESULTS AND DISCUSSION

3.1. Dataset

The data taken is a set of google maps review data from waterfront city pangururan. From these reviews 311 reviews were obtained. Next is the labeling stage which is done manually. The text is read and analyzed by humans to determine the sentiment contained in it based on their interpretation of the text. Here are the results after the data were obtained and labeled (shown in **Table 1**).

Tabel 1. Dataset

VISITOR	COMMENT	LABEL
Debby	*Kalau kmu lagi berkunjung ke Samosir, sempetin deh ke sini. Soalnya ini salah satu spot wisata terbaru di Samosir.*	Positive
Muhamad Rozali	*Waterfront City Pangururan merupakan bangunan replika patung Solu Bolon, dimana artinya solu merupakan sampan*	Positive

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699

	kayu yang digunakan oleh masyarakat sekitar Danau Toba, dilokasi ini didapati pedagang kaki lima, jajanan , sewa sepeda santai,	
Subagio	*Waterfront City Pangururan merupakan sebuah kompleks wisata yang terletak di kota Pangururan, di Pulau Samosir, Danau Toba, Sumatera Utara, Indonesia. Berikut adalah beberapa informasi detail tentang Waterfront City Pangururan:*	Positive
Junnes	*Tempatnya bagus ya,kami suka dengan situasi pemandangan di sana*	Positive
Dini	*Beberapa spot wisata danau Toba yg mudah di jangkau dan murah ,*	Positive
Suryaningsih	*Dekat dg pasar, jajanan komplit, permainan anak" juga ada Ada beberapa Totem/ kayu ukir yg berdiri di dekat tugu Pangururan*	Positive

3.2. Feature Extraction Result

After going through the labeling and text preprocessing stages, then the text is extracted using the scikit-learn library which provides the TfidfVectorizer module. It is useful for calculating and converting documents into TF-IDF representation. Table 2 is an example of output from feature extraction. A TF-IDF value of 0 in a review indicates that the word is not available in the review. Of course, this is a sparse matrix. It has the potential to affect the performance of the algorithm.

Table 2. Output TF-IDF.

REVIEW	ANAK	BAGUS	…	WISATA	KAMI	SAYANG
Review 1	0.000000	0.000000	…	0.214997	0.000000	0.000000
Review 2	0.000000	0.000000	…	0.000000	0.000000	0.000000
Review 3	0.000000	0.000000	…	0.148860	0.000000	0.000000
Review 4	0.000000	0.224137	…	0.000000	0.224137	0.180832
Review 5	0.196289	0.000000	…	0.131457	0.000000	0.316730

3.3. Performance of Model

This research split the dataset into two parts: training data and testing data. 80% of the training data aims to build models with Naïve Bayes and TF-IDF algorithms. While the rest of the dataset is intended as test data. The process of testing performance, calculating values and identifying the classification of the naïve bayes algorithm against the machine that has been built using confusion matrix by calculating accuracy, precision, recall and F1-Score. **Figure 1**. is the result of the model with confusion matrix.

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699

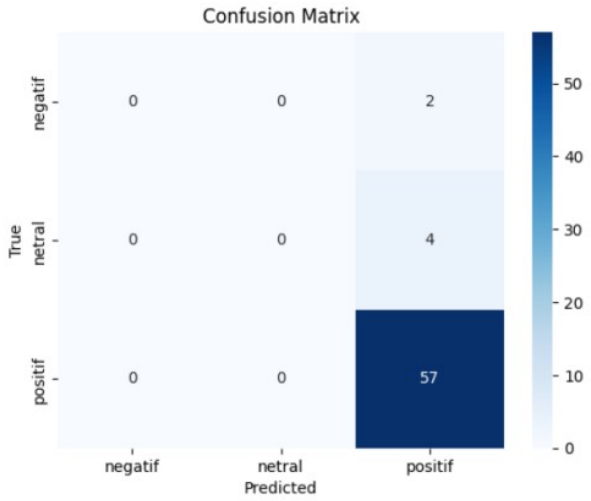

Figure 1. Confusion matrix of Naïve Bayes and TF-IDF model

Referring to the confusion matrix, the performance of Naïve Bayes and TF-IDF models in analyzing the sentiment of the reviews of Waterfront City Pangururan can be determined. Table 6 is the Naïve Bayes and TF-IDF perfomance report. The classification model evaluation results show that the model performs very well in recognizing the "positive" class, but fails in classifying the 'negative' and "neutral" classes. This can be seen from the evaluation metrics, where the precision, recall, and f1-score for the "positive" class reached 0.90, 1.00, and 0.95, respectively. In contrast, the "negative" and "neutral" classes scored 0 for all three metrics. This indicates that the model was not able to identify any examples from these two minority classes. The overall accuracy was 90% because the majority of the data was from the "positive" class (57 out of 63), so these results are biased.

Table 3. Classification report Naïve Bayes and TF-IDF

	precision	recall	f1-score	support
negative	0.00	0.00	0.00	2
netral	0.00	0.00	0.00	4
positive	0.90	1.00	0.95	57
accuracy			0.90	63
macro avg	0.30	0.33	0.32	63
weighted avg	0.82	0.90	0.86	63

The macro average, which calculates the average score of each class without considering the amount of data, shows low results with a precision of 0.30, recall of 0.33, and f1-score of 0.32. In contrast, the weighted average, which considers the proportion of each class, displays higher values of precision 0.82, recall 0.90, and f1-score 0.86, but this tends to be misleading as it is overly influenced by the dominance of the "positive" class.cFrom this analysis, it can be concluded that the model suffers from a strong bias towards the majority class and is unable to detect the minority class. To improve the performance of the model, it is suggested that data balancing (through oversampling techniques such as SMOTE), use of fairer evaluation metrics such as macro F1-score, and consider class weighting during model training. These steps can help improve classification accuracy for all classes in a more balanced manner**.**

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699

Effective tourism planning relies not only on infrastructure development, but also on a deep understanding of visitor perceptions and experiences. Through sentiment analysis of traveler reviews on platforms such as Google Maps, destination managers can identify satisfaction levels, complaints, and recurring themes in tourist visits. In Pangururan Waterfront City, for example, consistent positive sentiments on natural beauty, accessibility, and the presence of public facilities can be used as a focus for further promotion and development strategies. Conversely, negative reviews-even if they are few in number-can provide important insights into shortcomings in terms of cleanliness, comfort, or service. Utilizing this data allows local governments and tourism managers to develop more responsive, targeted, and sustainable policies, and improve the overall tourist experience.

4. CONCLUSION

This research shows that the Naive Bayes algorithm combined with the TF-IDF method can be used effectively in analyzing the sentiment of tourist reviews at Waterfront City Pangururan. The classification results show the dominance of positive sentiments with a fairly high accuracy rate of 90.48%. Preprocessing processes such as data cleaning, tokenization, stopwords removal, and stemming are instrumental in improving the performance of the classification model. These findings provide an overview of visitor satisfaction and can serve as a reference for managers to maintain and improve the quality of tourism services. In the future, this approach can be widely adopted in data-driven decision-making in the tourism sector.

5. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

6. REFERENCES

Friadi, J., dan Kurniawan, D. E. (2024). Analisis Sentimen Ulasan Wisatawan Terhadap Alun-Alun Kota Batam: Perbandingan Kinerja Metode Naive Bayes dan Support Vector Machine. Jurnal Sistem Informasi Bisnis, 14(4), 403–407.

Jin, Y., Cheng, K., Wang, X., and Cai, L. (2023). A Review of Text Sentiment Analysis Methods and Applications. Frontiers in Business, Economics and Management, 10(1), 58–64.

Karo, I. M. K., Fajari, M. Y., Fadhilah, N. U., and Wardani, W. Y. (2022). Benchmarking Naïve Bayes and ID3 Algorithm for Prediction Student Scholarship. IOP Conference Series: Materials Science and Engineering, 1232(1), 012002.

Karo Karo, I. M., dan Hendriyana, H. (2022). Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score. Jurnal Teknologi Terpadu, 8(2), 94–99.

Karo Karo, I. M., Romia, R., Dewi, S., and Fadilah, P. M. (2023). Hoax Detection on Indonesian Tweets using Naïve Bayes Classifier with TF-IDF. Journal of Information System Research (JOSH), 4(3), 914–919.

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699

Kiran Kumar, P., Jahna Tejaswi, N., Vasanthi, M. L., Srihitha, L. L., and Phanindra Kumar, B. (2022). Sentimental Analysis on Multi-domain Sentiment Dataset Using SVM and Naive Bayes Algorithm. Communications in Computer and Information Science, 1528 CCIS.

Pristiyono, Ritonga, M., Ihsan, M. A. Al, Anjar, A., and Rambe, F. H. (2021). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. IOP Conference Series: Materials Science and Engineering, 1088(1).

Puana, Z., Maryati, S., dan Koem, S. (2023). Pemetaan Objek Wisata Alam dan Sarana Prasarana Pendukung di Kabupaten Bone Bolango. Jurnal Pendidikan Geografi Undiksha, 11(3), 262–271.

Putri, F. R., Vhatika, I., Yanto, H., Zukhrufa, N., dan Panorama, M. (2022). Pengaruh Pariwisata Terhadap Petumbuhan Ekonomi di Indonesia Tahun 2019-2021. ILTIZAMAT: Journal of Economic Sharia Law and Business Studies, 2(1), 195–203.

Putu, N. L. P. M., Ahmad Zuli Amrullah, dan Ismarmiaty. (2021). Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(1), 123–131.

Rifa'i, A., Sujaini, H., dan Prawira, D. (2021). Sentiment Analysis Objek Wisata Kalimantan Barat Pada Google Maps Menggunakan Metode Naive Bayes. Jurnal Edukasi Dan Penelitian Informatika (JEPIN), 7(3), 400.

Saepudin, A., Aryanti, R., Fitriani, E., Royadi, R., dan Ardiansyah, D. (2024). Analisis Sentimen Pemanfaatan Artificial Intelligence di Dunia Pendidikan Menggunakan SVM Berbasis Particle Swarm Optimization. Computer Science (CO-SCIENCE), 4(1), 71–79.

Thaha, A. R., & Aziz, F. (2020). Penambangan Teks Pada Tujuan Wisata di Bandung Raya (Studi Kasus: Tangkuban Perahu dan Kawah Putih). Jurnal Sekretaris Dan Administrasi Bisnis, 4(2), 146–156.

Wijayanto, S., Prabowo, D. A., Kristiyanto, D. Y., dan Fathoni, M. Y. (2022). Analisis Sentimen Berbasis Aspek pada Layanan Hotel di Wilayah Kabupaten Banyumas dengan Word2Vec dan Random Forest. Jurnal Informatika: Jurnal Pengembangan IT, 8(1), 1–3.

Wijianto, W., Ulfah, I. F., dan Mustoffa, A. F. (2024). Pengaruh Tingkat Keamanan, Kenyamanan dan Produk Wisata terhadap Sustainable Tourism di Ponorogo. Journal of Indonesian Tourism, Hospitality and Recreation, 7(2), 137–144.

DOI: https://doi.org/10.17509/seict.v6i1.86210
p-ISSN 2774-1656 e-ISSN 2774-1699