JURNAL IPTEK
MEDIA KOMUNIKASI TEKNOLOGI
homepage URL : ejurnal.itats.ac.id/index.php/iptek

Lexicon-Based, Naïve Bayes, C4.5 for Analyzing Visitor Data Reviews as Recommendations for Priority Development of Labuan Bajo Tourism

Arnoldus Janssen Dahur[1], Defitroh Chen Sami'un [2]

Teknologi Rekayasa Komputer dan Jaringan, Politeknik Negeri Tanah Laut[1],
Sistem Informasi, Institut Filsafat dan Teknologi Kreatif Ledalero[2]

ARTICLE INFORMATION

Jurnal IPTEK – Volume 25
Number 2, December 2025

Page:
187 – 198
Date of issue:
December 30, 2025

DOI:
10.31284/j.iptek.2025.v29i2.8178

E-MAIL

arnoldusdahur@politala.ac.id

*Corresponding author:
Arnoldus Janssen Dahur
arnoldusdahur@politala.ac.id

PUBLISHER

LPPM- Adhi Tama Institute of Technology Surabaya
Address:
Jl. Arief Rachman Hakim No. 100, Surabaya 60117, Tel/Fax: 031-5997244

ABSTRACT

Labuan Bajo is one of Indonesia's national priority destinations, known for its natural beauty and unique ecosystem. However, several issues raised by tourists—such as complaints regarding prices, infrastructure, and environmental cleanliness—may affect its image and tourism sustainability. This study aims to analyze tourist perceptions of Labuan Bajo based on 7,000 reviews from TripAdvisor and Google Maps obtained through web crawling. Sentiment labeling was conducted using a Lexicon-Based approach, while classification was performed using the Naïve Bayes and C4.5 algorithms, both with and without the Synthetic Minority Oversampling Technique (SMOTE). The results showed 4,374 positive, 2,769 negative, and 1,804 neutral reviews. Based on the CRISP-DM method, Naïve Bayes achieved the highest accuracy of 88%, compared to 78% for C4.5. Dominant positive terms such as beautiful, stunning, and sustainable highlight Labuan Bajo's natural strengths, while negative terms like price, toilet, and trash indicate areas requiring improvement. The findings provide strategic recommendations to enhance tourism management and service quality toward sustainable tourism development.

Keywords: *Labuan Bajo; Lexicon Based; Naïve Bayes; C4.5; Priorities; SMOTE.*

ABSTRAK

Labuan Bajo merupakan salah satu destinasi prioritas nasional yang memiliki daya tarik alam dan keunikan ekosistem tinggi. Namun, sejumlah keluhan wisatawan terkait harga, infrastruktur, dan kebersihan lingkungan masih muncul dan berpotensi memengaruhi citra serta keberlanjutan pariwisata. Penelitian ini bertujuan untuk menganalisis persepsi wisatawan terhadap Labuan Bajo berdasarkan 7.000 ulasan dari TripAdvisor dan Google Maps yang diperoleh melalui *web crawling*. Pelabelan sentimen dilakukan menggunakan pendekatan *Lexicon-Based*, sedangkan klasifikasi menggunakan algoritma Naïve Bayes dan C4.5 dengan dan tanpa penerapan *Synthetic Minority Oversampling Technique* (SMOTE). Hasil menunjukkan 4.374 ulasan positif, 2.769 negatif, dan 1.804 netral. Berdasarkan metode CRISP-DM, Naïve Bayes mencapai akurasi tertinggi sebesar 88%, sementara C4.5 sebesar 78%. Kata positif dominan seperti *beautiful*, *stunning*, dan *sustainable* mencerminkan kekuatan daya tarik Labuan Bajo, sedangkan kata negatif seperti *price*, *toilet*, dan *trash* menunjukkan aspek yang perlu diperbaiki. Temuan ini memberikan rekomendasi strategis untuk peningkatan pengelolaan dan kualitas layanan pariwisata secara berkelanjutan.

Kata Kunci: *Labuan Bajo; Lexicon Based; Naïve Bayes; C.4.5; Prioritas; SMOTE.*

INTRODUCTION

Human beings, like social and emotional creatures, are inseparable from the need to relax, recreate, and enjoy themselves. One common way to relieve stress is through traveling, which has become increasingly popular among young people under the term "healing" [1]. Tourist destinations offer natural beauty that helps in reducing fatigue. Indonesia itself is rich in tourist attractions, especially since it is an archipelagic country. According to data from the Directorate General of General Administration—Ministry of Home Affairs, Indonesia has 17,504 islands, with 16,056 of them officially registered with the United Nations [2].

The abundance of islands in Indonesia provides numerous marine tourism destinations. One of the most well-known todays is Labuan Bajo. Located on Flores Island, West Manggarai Regency, East Nusa Tenggara Province, Labuan Bajo has been designated as one of Indonesia's Super Premium Destinations, as stated in the decree of the coordinating minister for Maritime Affairs and Natural Resources Number S-54/Menko/Maritim/VI/2016 [3]. Tourism brings significant benefits to both local communities and the Indonesian government, including improving the local economy, creating more job opportunities, increasing regional revenue through taxes, and providing many other advantages. Therefore, it is crucial to deliver high-quality services in tourism destinations to continuously enhance their positive image [4]. As a result, tourists are likely to give favorable reviews, which in turn will attract more visitors.

Tourist reviews are indeed important as an evaluation material for relevant stakeholders such as the tourism office or the tourism authority of Labuan Bajo under the Ministry of Tourism. Review data can be processed in various ways, one of which is by applying machine learning techniques. Machine learning methods are categorized into supervised learning and unsupervised learning. Among supervised learning algorithms are Naïve Bayes and the Decision Tree C4.5.

The Naïve Bayes algorithm is a probabilistic classification method that is simple yet effective, assuming independence among features. Its selection is based on its ability to handle large-scale data, its efficiency in training processes, and its strong performance in text-based data such as sentiment analysis. Meanwhile, the C4.5 algorithm is a decision tree method that constructs classification models based on information gain and gain ratio. It can handle both categorical and numerical data, as well as addressing issues of missing values. The strength of C4.5 lies in its easily interpretable results, presented in the form of a clear and structured decision tree. Consequently, C4.5 can generate more transparent classification models, facilitating decision-makers in evaluating tourist reviews.

Accordingly, this study is intended to evaluate tourist reviews of the Labuan Bajo destination using a machine learning approach—specifically the Naïve Bayes and C4.5 methods—combined with a lexicon-based sentiment labeling technique. The purpose of this research is to classify tourist opinions into positive, neutral, and negative sentiment categories and to formulate evidence-based recommendations for improving the effectiveness of destination management and the quality of tourism services in Labuan Bajo.

LITERATURE REVIEW

Research related to machine learning has been widely conducted, particularly in the context of tourism and the algorithms applied in this study. An analysis of review data from tourist destinations on Komodo Island and Rinca Island, which are part of the Labuan Bajo tourism area, was previously carried out by Singgalen in his research. The methods used included K-NN, Naïve Bayes, SVM, and Decision Tree, with results showing that SVM achieved the highest accuracy at 99.69%, outperforming the other methods [5].

Another study conducted by Rizal et al. analyzed visitor review data of Kejawanan Beach tourism using the Naïve Bayes algorithm. The results showed an accuracy rate of 78%, precision of 92%, recall of 80%, and an F1-score of 86%, indicating strong performance in sentiment classification. Based on the evaluation using the confusion matrix, the model consistently distinguished between positive and negative sentiments. The analysis concluded that most reviews provided by visitors to Kejawanan Beach were positive, reflecting a satisfying experience [6].

Further research analyzing visitor reviews of Madura tourism was conducted by Vina et al., employing the C4.5 algorithm on 100 news articles sourced from online media. The classification

results showed that the C4.5 algorithm achieved an accuracy of 76.5% across 10 testing iterations [7].

A study conducted at the Department of Informatics, UIN Sunan Gunung Djati Bandung, aimed to compare the Naïve Bayes Classifier and C4.5 algorithms in predicting the duration of students' study periods. The attributes used included student ID (NIM), name, gender, GPA, admission track, tahfidz (Qur'an memorization), school origin, and extracurricular activities during university studies. The findings revealed that the Naïve Bayes Classifier achieved an accuracy of approximately 88%, slightly higher than C4.5, which achieved an accuracy of around 87%. However, C4.5 processed data more quickly at 0.003 nanoseconds, compared to the Naïve Bayes Classifier, which required 12.7 nanoseconds [8].

Another study employing the Naïve Bayes algorithm was conducted by Munawaroh. The Naïve Bayes algorithm demonstrated competitive accuracy, as shown in a previous sentiment analysis of public opinion on COVID-19 vaccination on Twitter. In that study, Naïve Bayes achieved an accuracy rate of 94%, approaching the performance of SVM, which recorded the highest accuracy at 96.3% [9].

METHOD

The research method employed the Cross Industry Standard Process for Data Mining (CRISP-DM), which is divided into six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [10]. Figure 1 illustrates the stages undertaken in this study using the CRISP-DM technique.

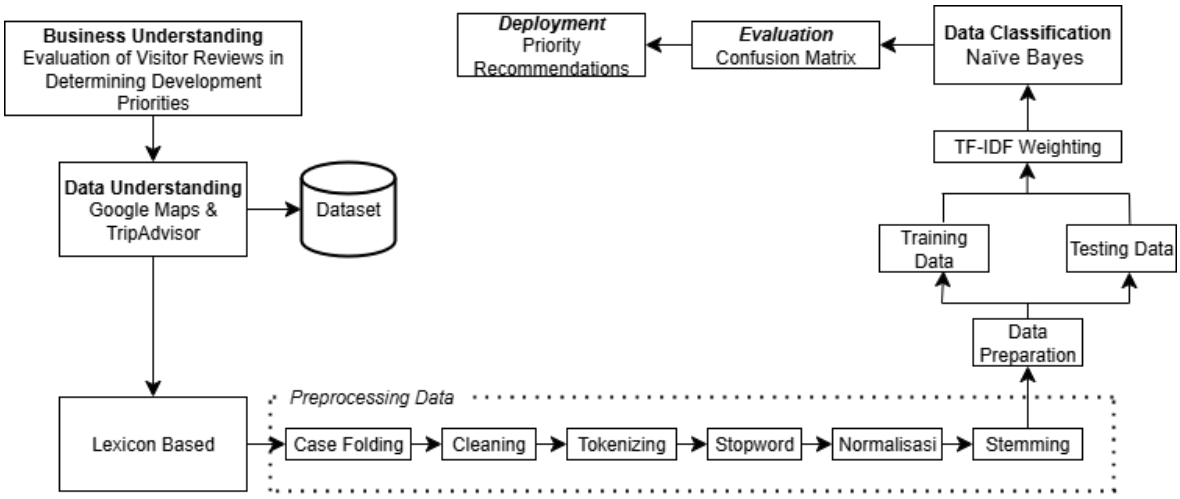

Figure 1. Research Stages

Business Understanding

This stage aims to find out the business objectives of this research. The objective of this research is to find out the implementation of machine learning algorithms, namely Naïve Bayes and C4.5, and the priority recommendations for Labuan Bajo tourism.

Data Understanding

This stage aims to find out the way of data collection, which was carried out by scraping and crawling from visitor reviews of Labuan Bajo tourism from TripAdvisor Maps. The extension used in the Chrome web browser is Data Scraper.

Data Preparation

This stage aims to change the data from the data collection so that it can be used when applying the algorithms. Data obtained from internet sources usually contain a lot of bias. Unstructured data is processed into structured data, also known as data preprocessing [11]. There are

several steps carried out in preprocessing data as follows: The first step is case folding which aims to change every letter in the data into lowercase. The second step is cleaning which aims to remove characters and emoticons that are not needed in data processing. Examples of characters such as [-!"#$%&'()*+,./:;<=>?@`{|}~]); and others. The third step is tokenizing which aims to separate each review in the data based on each word, because each word in data weighting has different values. The fourth step is stopword which aims to remove words that have no meaning, such as conjunctions yang, juga, dari, dia, kami, kamu, aku, saya, ini, itu, atau, dan, tersebut, pada, dengan, adalah, yaitu, ke. The fifth step is normalization which aims to change abbreviations or non-standard words to match the rules of KBBI. The sixth step is stemming which aims to change each word to match its root word. The library used in this stage is Sastrawi.

The data that has gone through the preprocessing stage is still in text form. Meanwhile, to process data using the Naïve Bayes and C4.5 algorithms, the data is converted into numerical form. Therefore, a method is needed that can be used to convert text data into numerical data. The commonly used method is Term Frequency-Inverse Document Frequency (TF-IDF). The calculation uses weighting on a word based on how often the word appears in the data.
The stages in calculating the weight values using TF-IDF are as follows [12].

1. Term Frequency (TF).
The first stage is calculating Term Frequency, which aims to calculate the number of word occurrences in the existing dataset [12]. The following formula is used to calculate it in equation (1).

$$tf_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d}, if tf_{t,d} > 0 \\ 0 \quad otherwise \end{cases} \quad \text{..... (1)}$$

with
tf_{td}: the number of occurrences of term t in document d

2. Document Frequency (DF).
The second stage is Document Frequency, which aims to find out the number in the dataset that contains term t. Usually, it is found in the dataset.

3. Inverse Document Frequency (IDF).
The third stage is Inverse Document Frequency (IDF), which aims to determine that the word that appears the least frequently in the document is the word that has the greatest weight. The formula used to calculate the IDF value can be found in equation (2).

$$idf_t = log_{10}(\frac{N}{df_t}) \quad \text{..... (2)}$$

with
N = the number contained in the text documents
df_t = the number of documents that contain term t

4. Term Frequency–Inverse Document Frequency (TF-IDF).
The fourth stage is Term Frequency–Inverse Document Frequency. The calculation of TF-IDF aims to calculate the weighting of words. The way to calculate the weight (w) of a sentence is by multiplying the term frequency (tf) and the inverse document frequency (idf). Equation (3) can be used to see how the weight (w) is calculated.

$$tfidf = tf_{t,d} \times idf_t \quad \text{.... (3)}$$

with
$tf_{t,d} = Term\ Frequency$
$idf_t = Inverse\ Document\ Frequency$

Modelling

This stage aims to process the data into the models used in this research, namely Naïve Bayes and C4.5. Both models are machine learning models for data classification.

a. Naïve Bayes

The Naïve Bayes classifier is a probabilistic machine learning model used for classification tasks. The core of this model is based on Bayes' Theorem, which is used to calculate the probability of a class based on the observed features [13]. Bayes' Theorem is a fundamental concept in probability theory used to calculate the posterior probability P(y | X), based on previously known values, namely the prior probability P(y), the likelihood P(X | y), and the evidence P(X) [14]. Mathematically, Bayes' Theorem is formulated as follows:

$$P(y|X) = \frac{P(X|y)*P(y)}{P(X)} \quad \text{…. (4)}$$

with:

- *P(y | X)* is the posterior probability, namely the probability of the target class *y* after considering the feature data *X*,
- *P(y)* is the prior probability of class *y*, namely the initial belief about the class before observing the data,
- *P(X | y)* is the likelihood, namely the probability of the occurrence of feature *X* given that the class is *y*,
- *P(X)* is the evidence, namely the overall probability of feature *X* regardless of the class.

b. C4.5

The C4.5 algorithm is a method used to form decision trees developed by Ross Quinlan. The main basis of this algorithm is building a decision tree based on the selection of attributes that have the main order or can be said to have the highest gain value based on entropy. Attributes function as the axis for attribute classification. At the stage of the C4.5 algorithm there are two main concepts, namely: the formation of the decision tree, and the creation of rules (model rules). The conditions that arise from the decision tree will determine the conditions in the if–then format [15].

The stages in applying the C4.5 algorithm are as follows:

1. Prepare training data

First, prepare the training data taken from the previous data (review data) and grouped into certain categories.

2. Calculate the parent node

Determine the parent node by calculating the gain value of each attribute. The attribute with the highest gain becomes the first parent node. Before calculating the gain, first calculate the entropy value using the predetermined formula as follows:

$$Entropy\ (S) = \sum_{i=1}^{n} - Pi\ log_n\ Pi \quad \text{….. (5)}$$

with:
S = the set of cases or data being analyzed
n = the number of partitions or divisions of set S
Pi = the proportion of the amount of data in partition Si to the total data in S

3. Calculate the gain value using the following formula

$$Gain\ (S, A) = entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S}\ X\ Entrophy\ (Si) \quad \text{….. (6)}$$

with:
S = the set of cases or data being analyzed
A = the attribute or feature being calculated
N = the number of partitions of attribute A
|Si| = the amount of data in partition Si compared to the total data in S
|S| = the total number of cases in S

4. Repeat steps 2 and 3 until all data have a class.
5. The process of splitting (partitioning) in the decision tree will stop if:
   - Every data item in one node has a similar category.

- There are no other attributes that can be used to split the data.
- The formed branch has no information (empty branch).

Evaluation

This stage aims to find out the level of accuracy, precision, recall, and F1-score. A confusion matrix is a table used to evaluate the performance of a classification model or prediction algorithm. The confusion matrix is a method used to represent the accuracy results of the model that has been created [16]. The confusion matrix table can be seen in Table 1.

Table 1. *Confusion Matriks*

Confusion Matriks	*Predicte Positive*	*Predicte Negative*
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positif (FP)*	*True Negative (TN)*

The confusion matrix used in this study includes the values of Accuracy, Precision, Recall, and F-Measure. Accuracy is the number of correct predictions out of all predicted data. Accuracy is a method commonly used to measure the goodness of a model in classification modeling. F-Measure is the ratio of the weighted average of Precision and Recall. F-Measure is a good metric to evaluate a model on imbalanced data. The calculation of the confusion matrix values can be carried out using the following formulas:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \; X \; 100\% \quad \text{.... (7)}$$

$$Precision = \frac{TP}{\sum TP+FP} \; X \; 100\% \quad \text{.... (8)}$$

$$Recall = \frac{TP}{TP+FN} \; X \; 100\% \quad \text{.... (9)}$$

$$F - Measure = \frac{2 \; x \; Precision \; X \; Recall}{Precision+Recall} \; X \; 100 \quad \text{.... (10)}$$

Deployment

The final stage is deployment, which aims to implement the model into the research object environment used in this study. The model that has been tested and evaluated will be applied and used to answer the predetermined project objectives. Deployment contains the evaluation results of the application of machine learning algorithms and provides recommendations in the form of a priority scale for the management of Labuan Bajo tourist attractions, which are easy to understand and can help in developing and increasing the number of tourist visitors.

RESULTS AND DISCUSSION

Data Collection

Data collection was carried out using a scraping technique sourced from Google Maps and TripAdvisor, with a free extension provided by the Google Chrome browser, namely Web Scraper. The total data collected amounted to 7,000 entries.

Data Labeling

Data labeling was carried out using the lexicon-based algorithm with the SentiWordNet dictionary. The advantage of using Lexicon-Based is that the data labeling process is performed automatically. Since SentiWordNet uses English for labeling, the data was first translated using the DeepL Translate library, version 1.3.0, before labeling. After labeling with the lexicon-based method, the data distribution obtained was: Positive 4,374, Negative 2,769, Neutral 1,804. Visually, the distribution of data after labeling can be seen in Figure 2.

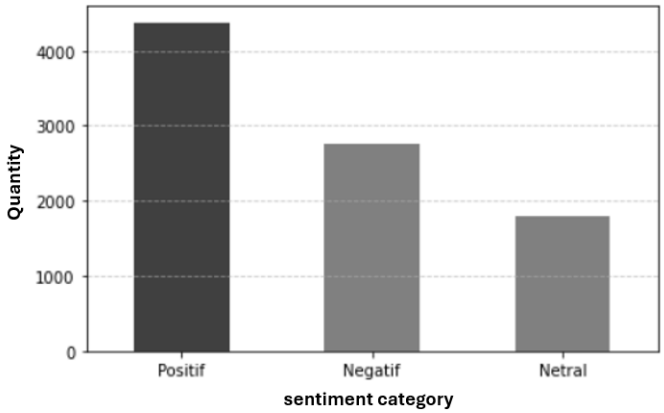

Figure 2. Data distribution after labeling.

Preprocessing Data

Data preprocessing aims to transform unstructured data into structured data. Data collection crawled from internet sources usually has shortcomings, namely the presence of bias. The obtained data still needs to undergo preprocessing so that it can be optimally used in data classification. Table 2 shows the preprocessing results according to the stages up to the final stage, namely stemming.

Table 2. Data Labeling Results

No	Ulasan	Preprocessing	Label
1	Seeing the dragons is an experience but the island itself is also very pretty and teeming with wildlife. Deer grace the beach in a scene that could be straight from a book. Snorkeling at Pink Beach is a must do. The pink comes from a reddish pink coral. Colourful fish galore...you will find nemo along with angel fish and many more. The surrounding islands reminiscent of Halong Bay in Vietnam. Highly recommend!	['komodo', 'alam', 'pulau', 'cantik', 'penuh', 'satwa', 'liar', 'rusa', 'hias', 'pantai', 'adegan', 'langsung', 'ambil', 'buku', 'snorkeling', 'merah', 'muda', 'beach', 'wajib', 'merah', 'muda', 'asal', 'karang', 'merah', 'muda', 'merah', 'ikan', 'berwarnawarni', 'limpah', 'temu', 'nemo', 'ikan', 'bidadari', 'pulaupulau', 'teluk', 'halong', 'vietnam', 'saran']	1
2	pulau komodo merupakan lokasi yang sangat terpencil di indonesia suami saya dan saya memiliki kesempatan luar biasa untuk mengunjungi taman nasional pulau komodo kami menghabiskan hari bepergian dengan kapal pesiar selam liveabord dan berhenti untuk mengunjungi pulau komodo merupakan hak istimewa untuk melihat bagian dunia yang begitu terpencil pemandangannya dan bisa mengamati komodo yang terkenal berkeliaran bebas benarbenar menarik dan pengalaman sekali seumur hidup	['pulau', 'komodo', 'lokasi', 'pencil', 'indonesia', 'suami', 'milik', 'sempat', 'unjung', 'taman', 'nasional', 'pulau', 'komodo', 'habis', 'pergi', 'kapal', 'pesiar', 'selam', 'liveabord', 'henti', 'unjung', 'pulau', 'komodo', 'hak', 'istimewa', 'dunia', 'pencil', 'pandang', 'amat', 'komodo', 'kenal', 'liar', 'bebas', 'benarbenar', 'tarik', 'alam', 'umur', 'hidup']	1
.			
.			
6999		'betapa', 'membuangbuang', 'tenaga', 'uang', 'sodok', 'batas', 'tongkat',	-1

No	Ulasan	Preprocessing	Label
	Betapa membuang-buang waktu, tenaga dan uang. Menyodok dan membatasi mereka dengan tongkat sehingga mereka dapat memaksa mereka berparade mengelilingi turis yang mengambil foto Instagram.	'paksa', 'parade', 'keliling', 'turis', 'ambil', 'foto', 'instagram']	
7000	Pantai pink tidak begitu merah muda seperti yang Anda lihat di internet	['pantai', 'pink', 'merah', 'muda', 'lihat', 'internet']	-1

Data Weighting

Data weighting aims to facilitate the algorithms used in conducting analysis related to text data so that it can be represented in numerical form. One of the methods for converting text data into numerical data is TF-IDF. Table 3 presents the results of word weighting from preprocessing and the word labeling table.

Table 3. Data Weighting Results

TF-IDF Dict	TF-IDF Vec
{'senang': 0.11897991059519214, 'langsung': 0.20539690515078743, 'komodo': 0.06462819529438314, 'pulau': 0.06558826022864064, 'lokasi': 0.2398139113574116, 'centra': 0.4077692868690886, 'kota': 0.27978646763851317, 'labu': 0.15801874676434, 'bajo': 0.15910928048796774, 'perahu': 0.10743409976973387, 'kesini': 0.4077692868690886}	[0.06462819529438314, 0.06558826022864064, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.10743409976973387, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.11897991059519214, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

Data Classification

Data classification aims to group data into certain categories based on the patterns or characteristics they possess so that the system can make predictions or decisions automatically. Since the algorithms used apply supervised learning, the data must first be trained before testing is conducted. In this study, the data exhibited imbalance, where positive data was more numerous than negative data. A total of 7,000 review data entries were collected in this study, but after the labeling process, only 2,769 reviews were categorized as negative sentiment. As a result, when evaluation was performed to determine the accuracy level, the results were still low. Therefore, the SMOTE technique was applied to address the imbalance issue. SMOTE is considered capable of performing classification, especially for minority data with very small quantities.

This condition creates data imbalance, making it necessary to apply the Synthetic Minority Oversampling Technique (SMOTE). The SMOTE technique is useful for handling unbalanced class distributions, particularly when the amount of data in the minority class is much smaller than in the majority class [4]. By applying SMOTE, the model is expected to improve its performance since the imbalance problem can be reduced, thereby minimizing bias toward the majority class. Furthermore, Table 4 presents a comparison of Accuracy, Precision, Recall, and F1-score values between the model without SMOTE and the model using the SMOTE technique.

Table 4. Data Classification Results

No		Tanpa *SMOTE*		*SMOTE*	
		NB	*C4.5*	*NB*	*C4.5*
1	*Accuracy*	0.78	0.73	0.89	0.78
2	*Precision*	0.99	0.78	0.91	0.83
3	*Recall*	0.31	0.82	0.88	0.78
4	*F1-score*	0.48	0.80	0.90	0.80

The application of the SMOTE technique has been proven to significantly improve the performance of the classification model. This can be seen from the accuracy results after balancing the data, where the Naïve Bayes algorithm increased its accuracy from 77% to 89%, while the C4.5 algorithm improved from 70% to 87%. This performance improvement shows that adding synthetic data to the minority class successfully reduced the model's bias toward the majority class, making the model more capable of recognizing negative sentiment patterns.
Based on the confusion matrix after applying SMOTE, predictions using Naïve Bayes produced 286 True Positive (TP) data, 40 False Positive (FP) data, 230 True Negative (TN) data, and 27 False Negative (FN) data. The visualization of the confusion matrix for each algorithm is presented in Figure 3 for Naïve Bayes and Figure 7 for C4.5.

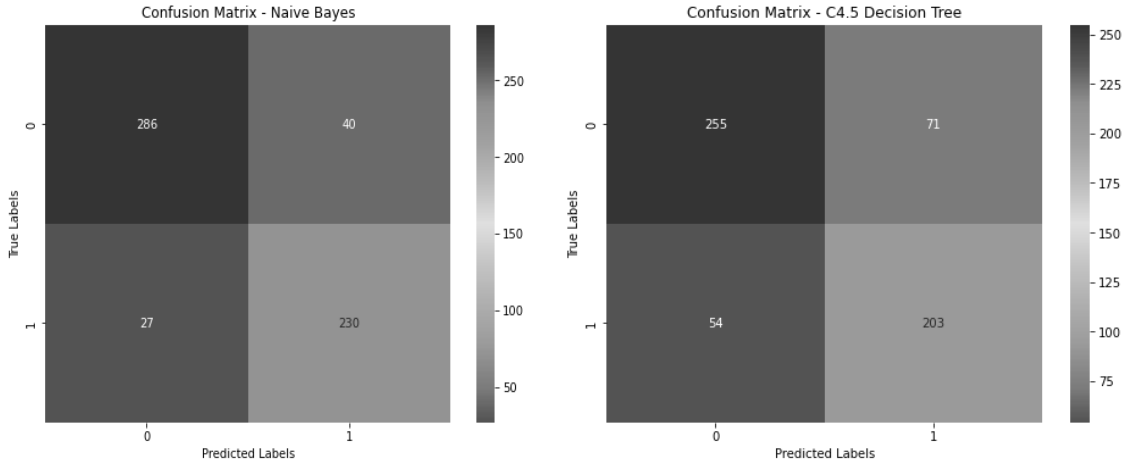

Figure 3. Naïve Bayes and C4.5 confusion matrix results

In addition to using accuracy, this study also evaluated the model with the Area Under the Curve (AUC) metric, which is more appropriate to apply in cases of imbalanced data. The AUC value describes the model's ability to distinguish between positive and negative sentiment classes based on the ROC curve, where the closer the value is to 1, the better the model's performance [12]. Based on the AUC value range according to Gorunescu, the performance of the Naïve Bayes and C4.5 algorithms after applying SMOTE falls within a good classification level, as shown in Table 5.

Table 5. AUC value range and classification level according to Gorunescu [17].

Auc Value	Classification Level
0,91-1,00	Perfect Classification
0,81-0,90	Good Classification
0,71-0,80	Fair Classification

Auc Value	Classification Level
0,61-0,70	Poor Classification
0,50-0,60	Failure

From the research results, the AUC ROC values for this study were 0.82 and 0.68 without using SMOTE, indicating good and fair classification levels. By applying SMOTE, the Naïve Bayes value was 0.94 and C4.5 was 0.79, which indicate very good and fair classification levels based on Table 5. Figure 4 shows the ROC AUC curve results using the Naïve Bayes and C4.5 methods, both without and with SMOTE.

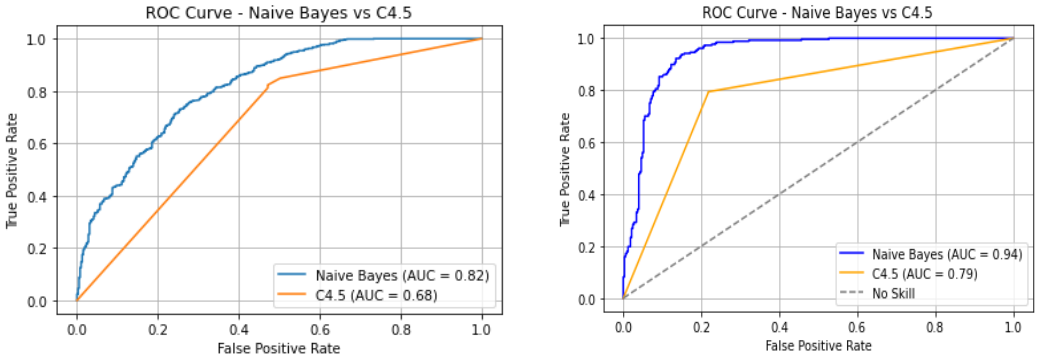

Figure 4. AUC ROC scores for Naïve Bayes and C4.5

Based on the classification results of 564 tourist reviews of Komodo Island and 364 reviews of Rinca Island using the k-NN, NBC, SVM, and DT algorithms within the CRISP-DM framework, the findings showed that the Support Vector Machine (SVM) algorithm delivered the best performance with an accuracy of 99.69%, precision of 100%, recall of 99.39%, f-measure of 99.69%, Area Under the Curve (AUC) of 100%, and a t-Test value of 0.958 [5]. Considering the characteristics of review data in text form, as well as previous studies that demonstrated good performance of Naïve Bayes and K-NN in sentiment classification tasks, the combination of these two algorithms was chosen as an alternative classification method that is efficient and accurate for analyzing public sentiment patterns. Based on the collected review data, the most frequently occurring words were grouped. Each group was divided into positive words and negative words, with the top 10 most frequently occurring words taken along with their frequency of occurrence. The most frequent positive words included indah (beautiful) (569), cantik (pretty) (326), lestari (sustainable) (322), jernih (clear) (249), eksotik (exotic) (239), Komodo (364), fosil (fossil) (283), habitat (347), snorkeling (269), and pantai (beach) (264). The dominant negative words were harga (price) (208), tiket (ticket) (151), kapal (ship) (100), jalan (road) (207), tangga (stairs) (106), toilet (98), sampah (trash) (72), panas (hot) (99), and licin (slippery) (85).

CONCLUSION

The analysis of 7,000 tourist reviews about Labuan Bajo collected from TripAdvisor and Google Maps using web crawling revealed 4,374 positive, 2,769 negative, and 1,804 neutral sentiments, showing an imbalance where positive feedback dominates. The study applied lexicon-based labeling and classification using the Naïve Bayes and C4.5 algorithms, with and without the SMOTE technique to handle data imbalance. Naïve Bayes achieved 77% accuracy before SMOTE and improved to 89% after its application, while C4.5 increased from 73% to 78%. This indicates that Naïve Bayes performs better for sentiment classification based on the CRISP-DM method. The most frequent positive words such as “beautiful,” “sustainable,” “clear,” “exotic,” “Komodo,” and “beach” highlight Labuan Bajo’s strengths in natural beauty and biodiversity. In contrast, negative terms like “price,” “ticket,” “boat,” “road,” “toilet,” and “trash” reveal issues related to costs, accessibility, and cleanliness. By maintaining its ecological appeal and addressing infrastructure and

service quality problems, Labuan Bajo can further enhance its image as a sustainable, comfortable, and visitor-friendly tourist destination.

BIBLIOGRAPHY

[1] N. Hikmah, N. K. Fauziyah, M. Septiani, and D. M. Lasari, "Healing Sebagai Strategi Coping Stress Melalui Pariwisata," *Indones. J. Tour. Leis.*, vol. 3, no. 2, pp. 113–124, 2022, doi: 10.36256/ijtl.v3i2.308.

[2] F. J. Amarrohman, M. Awaluddin, B. D. Yuwono, and A. Arifin, "Analisis Keberadaan Kepulauan Seribu Terhadap Batas Pengelolaan Laut Provinsi Dki Jakarta," *Elipsoida J. Geod. dan Geomatika*, vol. 3, no. 01, pp. 87–91, 2020, doi: 10.14710/elipsoida.2020.7754.

[3] H. L. L. Lada, "Komunikasi Pariwisata dalam Pengembangan Destinasi Wisata Premium Berbasis Pemberdayaan Masyarakat di Labuan Bajo," *Bull. Community Engagem.*, vol. 4, no. 3, pp. 57–67, 2024.

[4] Y. A. Singgalen, "Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1634–1646, 2022, doi: 10.47065/bits.v4i3.2766.

[5] Y. A. Singgalen, "Analisis Sentimen Pengunjung Pulau Komodo dan Pulau Rinca di Website Tripadvisor Berbasis CRISP-DM," vol. 4, no. 2, pp. 614–625, 2023, doi: 10.47065/josh.v4i2.2999.

[6] R. D. R. Apriliansyah, R. Astuti, W. Prihartono, and R. Hamonangan, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Pengunjung Di Pantai Kejawanan," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 1, 2025, doi: 10.23960/jitet.v13i1.5774.

[7] V. A. Savitri, M. Sa'id, H. Husni, and A. Muntasa, "A sentiment analysis of madura island tourism news using C4.5 algorithm," *J. Soft Comput. Explor.*, vol. 5, no. 1, pp. 9–17, 2024, doi: 10.52465/joscex.v5i1.258.

[8] Y. A. Gerhana, I. Fallah, W. B. Zulfikar, D. S. Maylawati, and M. A. Ramdhani, "Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period," *J. Phys. Conf. Ser.*, vol. 1280, no. 2, 2019, doi: 10.1088/1742-6596/1280/2/022022.

[9] K. Munawaroh and A. Alamsyah, "Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter," *J. Adv. Inf. Syst. Technol.*, vol. 4, no. 2, pp. 113–125, 2023, doi: 10.15294/jaist.v4i2.59493.

[10] S. Girendra Wardhani and A. Kurniawati, "Implementation of K-Nearest Neighbor Algorithm for Creditworthiness Analysis Using Methods Cross-Industry Standard Process for Data Mining (CRISP-DM)," *Int. Res. J. Adv. Eng. Sci.*, vol. 10, no. 1, pp. 152–157, 2025, [Online]. Available: https://archive.ics.uci.edu/dataset/144/statelog+german+credit

[11] D. P. Isnarwaty and I. Irhamah, "Text Clustering pada Akun TWITTER Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) dan K-Means," *J. Sains dan Seni ITS*, vol. 8, no. 2, pp. 2–9, 2020, doi: 10.12962/j23373520.v8i2.49094.

[12] D. Indra, J. Endro, and W. Amien, "Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction," vol. 01, pp. 1–9, 2023, doi: 10.21456/vol13iss1pp1-9.

[13] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[14] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Inf.*, vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[15] S. Dwiasnati and Y. Devianto, "Utilization of Prediction Data for Prospective Decision Customers Insurance Using the Classification Method of C.45 and Naive Bayes Algorithms,"

J. Phys. Conf. Ser., vol. 1179, no. 1, 2019, doi: 10.1088/1742-6596/1179/1/012023.

[16] M. Y. Aldean, P. Paradise, and N. A. Setya Nugraha, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac)," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 4, no. 2, pp. 64–72, 2022, doi: 10.20895/inista.v4i2.575.

[17] J. Prasetya, "Penerapan Klasifikasi Naive Bayes dengan Algoritma Random Oversampling dan Random Undersampling pada Data Tidak Seimbang Cervical Cancer Risk Factors," *Leibniz J. Mat.*, vol. 2, no. 2, pp. 11–22, 2022, doi: 10.59632/leibniz.v2i2.173.