



Sentiment Analysis of Raja Ampat Tourism Destination Using CRISP-DM: SVM, NBC, DT, and k-NN Algorithm

Astuti Kusumawicitra Laturiuw¹, Yerik Afrianto Singgalen^{2*}

^{1, 2*}Tourism Department, Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia.

Email: ¹astuti.kusumawicitra@atmajaya.ac.id, ^{2*}yerik.afrianto@atmajaya.ac.id

Abstract

This study presents a sentiment analysis of tourists' opinions on Raja Ampat Tourism Destination using data mining techniques. The study collected data from Tripadvisor and processed it through sentiment classification. The algorithms used in the analysis were Support Vector Machine, Naive Bayes Classifier, Decision Tree, and k-Nearest Neighbor. The study followed the Cross-Industry Standard for Data Mining methodology and went through several stages such as business and data comprehension, data preparation and cleaning, feature selection, modeling, model evaluation, result presentation, deployment, and maintenance. The study's findings revealed that visitors generally had positive opinions about Raja Ampat's tourism attractions, particularly cultural diversity, and undersea beauty. The Decision Tree algorithm showed the highest accuracy value of 99.12%, precision of 98.96%, recall of 99.34%, AUC of 0.991, and f-measure of 99.13%. SVM also had excellent performance with an accuracy value of 100%, precision of 100%, recall of 100%, AUC of 1.000, and f-measure of 100%. The study concludes that Decision Tree and SVM, with the assistance of SMOTE operators, are the best algorithms for sentiment analysis in this context.

Keywords: Sentiment Analysis; CRIPS-DM; SVM; NBC; DT; k-NN; Raja Ampat

1. INTRODUCTION

Raja Ampat Regency is one of the well-liked tourism destinations in Eastern Indonesia, specifically Southwest Papua Province. [1] asserts that Raja Ampat Regency's growing tourism industry also fosters the area's economic growth by offering housing and transportation services. [2] explains how tourist visitor activities improve the socioeconomic circumstances of nearby populations. Tourism may benefit local economies, social welfare, the preservation of regional cultural heritage, and environmental conservation. However, a proper investigation into how tourists view Raja Ampat is imperative. Based on this, the study evaluated Raja Ampat tourists' perceptions using a data mining methodology. We used a sentiment analysis approach through Support Vector



Machine (SVM), Naïve Bayes Classifier (NBC), Decision Tree (DT), and k-Nearest Neighbor (kNN) to assess the accuracy, precision, and recall of datasets taken from the Tripadvisor website.

Promoting tourist places through digital technology is one of the factors that encourages visitors to Raja Ampat Regency. [3] argues that using social media and other digital platforms is one of the measures to promote particular specialty tourism locations. [4] claimed efficiently increasing sales volume through information technology in promoting travel destinations and small and medium enterprises' products. This demonstrates how vital and successful information technology is in destination marketing for boosting revenue and tourist numbers. Raja Ampat Regency's tourism spots are geographically dispersed among several intriguing locales. However, they are not connected to any Internet networks. As a result, given the circumstances surrounding Raja Ampat's tourism spots, the information about those locations is only partially complete. To be known to local tourism stakeholders and receive support from the marketing process, tourists must visit authentic destinations and upload their photos to their respective social media accounts.

The growth of Raja Ampat tourism requires the support of the local population and information technology. [5] Designing a tourism information system that details attractions, accessibility, lodging, and amenities demonstrates community support and engagement in boosting Raja Ampat tourism. Meanwhile, [6] addressed how different community-based strategies might encourage support and involvement from the local population. This demonstrates how the community's form and support for the growth of Raja Ampat's tourist destinations are expressed in various behaviors, such as active community involvement and initiatives to upload posts about tourism-related activities willingly. Social media users will become more interested in online content on Raja Ampat tourist sites. Additionally, social media data will be used to inspire trips to Raja Ampat and be considered when booking vacations.

Creating tourism information systems enables software designers to include unique features that allow tourists to share their travel experiences and evaluate the lodging and transportation they utilize while on vacation. [7] demonstrates that Booking.com and TripAdvisor are the two internet platforms travelers use the most frequently to plan their trips, book lodging and transportation, and then rate the goods and services they use while traveling. [8] contends that it is also thought that travelers use the Tripadvisor site to benefit from the advantages provided by Tripadvisor managers, including fame or personal branding. Visitors to Raja Ampat destinations review different lodging options and modes of transportation while on their tours, which is relevant to the development of Raja Ampat tourism. These reviews are a resource for other travelers and are considered when choosing

where to travel. Therefore, the appropriate and pertinent methodology must be used to investigate tourist sentiment thoroughly.

It is crucial to consider Tripadvisor's popularity in the context of its use as a platform that encourages passengers to publish evaluations of travel-related products and services. According to this study, tourist impressions of Raja Ampat's tourist destinations should be analyzed using data mining. Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), Decision Tree (DT), and k-Nearest Neighbor (kNN) algorithms must be used in data mining research to analyze visitor sentiment toward Raja Ampat tourist destinations to promote Raja Ampat tourism and increase visitor contentment. Raja Ampat's tourism offerings could be enhanced by learning trends or insights about how tourists view the region. The study's findings can be used to improve marketing strategies and uphold the level of the vacation experience. The data mining technique and the SVM, NBC, DT, and k-NN algorithms also provide more accurate and efficient decision-making in Raja Ampat tourism management, allowing for the sustainable development of tourist attractions.

Tourism is one of Indonesia's key economic areas, particularly in Raja Ampat. The success of tourism is significantly influenced by the level of service and the visitor experience. Research examining visitor sentiment is essential to determine the attitudes and opinions of tourists about the Raja Ampat tourist destinations in Papua. Using sentiment analysis, it is possible to decide on the factors that affect visitor happiness and the advantages and disadvantages of the products and services offered in the tourism industry. By enhancing visitor experiences and services, Raja Ampat tourism management decision-makers will be able to attract more visitors and have them spend more money. This study focuses on using datasets from the Tripadvisor website together with Support Vector Machine (SVM), Naive Bayes Classifier (NBC), Decision Tree (DT), and k-Nearest Neighbor (kNN) algorithms by the data mining procedure because of the urgency of the research issues. Also, this research initiative was conducted based on previous research related to sentiment analysis in the tourism sector [9]–[15].

2. METHODS

Popular data mining techniques include the CRISP-DM (Cross-Industry Standard Process for Data Mining) method, which involves several stages in the data processing process. The stages of the CRISP-DM approach are as follows: business understanding: this step entails determining the goal of the data processing procedure, looking over the data sources, and comprehending the surroundings and any limits; data understanding: the data that will be processed must be gathered and learned during this stage. Understanding the data's structure, traits, and quality falls under this category; data preparation: this stage comprises combining data from many sources; data cleaning: data cleaning entails removing

flaws or inconsistencies that could compromise the data mining process; feature selection: during this step, the most pertinent characteristics or variables are chosen to analyze data by the goal; modeling: selecting the best model and processing technique for the data is done at this stage. Algorithm selection, parameter setting, and model training all fall under this category; model evaluation: At this stage, the developed model's performance is assessed; results presentation: In this stage, the data analysis's findings are presented visually or in reports that aid in decision-making; deployment: this phase entails applying the model or data analysis results to resolve issues or discover new information; maintenance: during this phase, the model or system that has been developed is maintained to guarantee its long-term efficacy and quality. Meanwhile, an overview of research methods can be seen in the following figure.

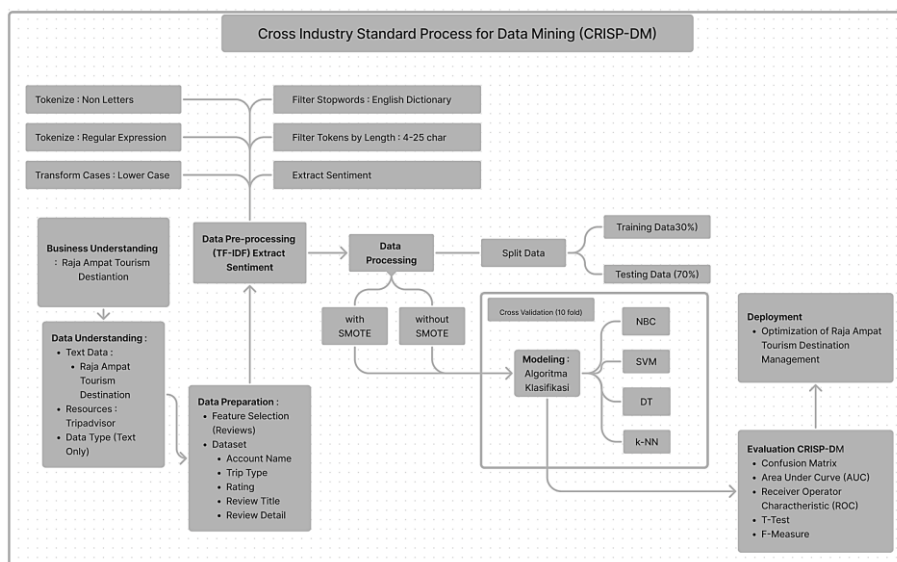


Figure 1. Sentiment Analysis of Raja Ampat Tourism Destination Using CRISP-DM Method

Figure 1 shows the CRISP-DM implementation aligned with the research environment for investigating visitor perceptions of Raja Ampat tourism attractions. Business Understanding: the objective of this stage is to comprehend the rationale behind the data analysis procedure, which entails applying the NBC, SVM, DT, and k-NN algorithms to analyze Tripadvisor data to learn how travelers feel about the Raja Ampat tourism area. Data Understanding: in this stage, Tripadvisor data will be gathered and examined to comprehend its composition, features, and level of accuracy. This process is part of finding the relevant data fields, such as the review text, rating, and review date. Data Preparation: in this phase, the data will be cleaned, irrelevant fields will be removed, and the data will

be transformed into an analysis-ready format. Data cleaning: at this point, the data will be free of mistakes, such as misspellings or inconsistent data formatting, that could skew the analysis's results.

The Webharvy program scraps information about visitor reviews of Raja Ampat tourist attractions from the Tripadvisor website. 596 review data were present before data pre-processing. Tokenizing non-letters, tokenizing regular expressions, changing cases to lowercase, filtering stopwords using English and dictionary language, and filtering tokens by length are the stages in the pre-processing of the review data. There are 335 review data obtained based on the outcomes of pre-processing, and they were prepared to be processed using the NBC, SVM, DT, and k-NN algorithms in the Rapidminer application. The results of pre-processing review data with Rapidminer are shown in the image below, which was obtained through a scraping procedure utilizing Webharvy.

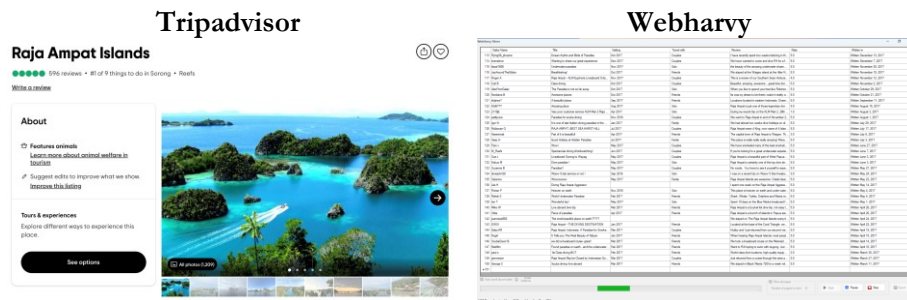


Figure 2. Scraping Data from Tripadvisor Using Webharvy

Figure 2 shows data collection from the Tripadvisor website using the Webharvy app. There are several reasons to use review data from the Tripadvisor Website: First, Tripadvisor has a system in place for verifying user reviews after they are uploaded; second, admins will remove reviews that don't match or don't pass verification; third, the verification process is carried out through the email associated with a Tripadvisor website user, indicating that the review data must comply with the website administrator's requirements; Fourth, the selection of reviews for sentiment analysis is based on the completeness of the information provided, such as the account name, rating, review title, review content, visitation time, review time, category, or type of trip. The Rapidminer application's NBC, SVM, DT, and k-NN algorithms will classify reviews that fit these criteria.

In the feature selection stages, the most pertinent features or variables to be employed in the analysis of the NBC, SVM, DT, and k-NN algorithms will be chosen at this stage. Modeling: Using the prepared data as a training set, sentiment analysis models will be built using the NBC, SVM, DT, and k-NN algorithms.

Evaluation of the Model: in this step, the sentiment analysis models' effectiveness will be assessed to decide which algorithm yields the best outcomes. The data analysis outcomes will be visually presented or made available in reports during this stage, which aids in decision-making. Deployment: at this stage, sentiment analysis models will be applied to learn more about how travelers feel about the Raja Ampat tourism location. Maintenance: the sentiment analysis models will be maintained at this stage to guarantee their long-term usefulness and quality, as shown in the figure below.

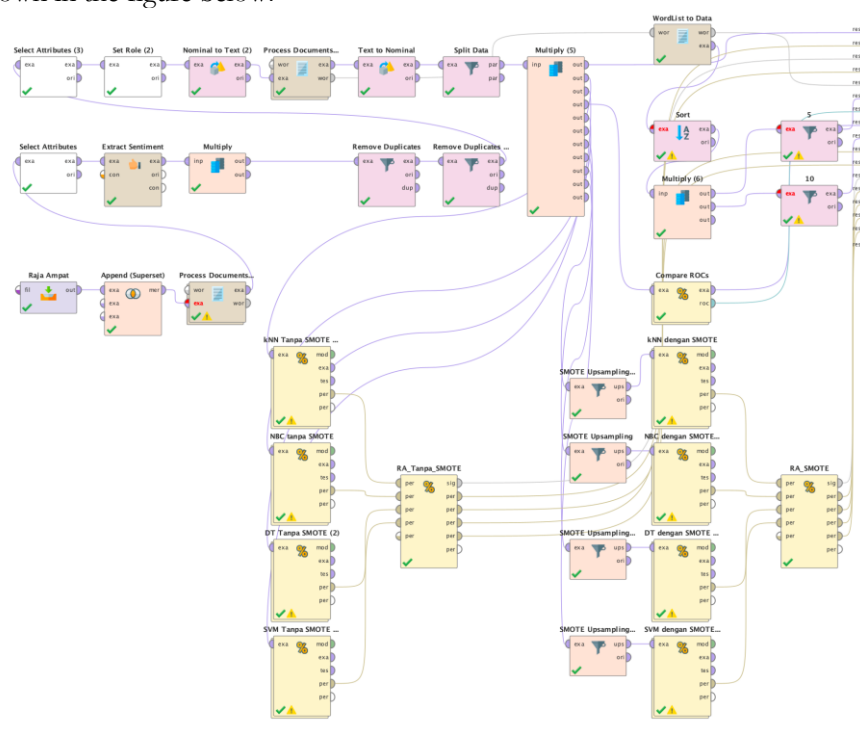


Figure 3. Implementation of CRISP-DM in Rapidminer

Figure 3 is an implementation of CRISP-DM using the Rapidminer application. The well-known software platform RapidMiner, which provides several advantages for data analysts and businesses, is used for applying the CRISP-DM technique [16]. Analysts may easily navigate between the many steps of the data analysis process because of the user-friendly interface it offers [17]. It also provides various data mining tools and techniques to examine organized, semi-structured, and unstructured data, take multiple data kinds, making integration with other platforms and data sources simple [18]. Rapidminer offers automated modeling and visualization solutions that simplify the modeling and free analysts to concentrate on the findings and insights [19]. The CRISP-DM technique can be

used with the competent and potent tool RapidMiner. It could assist businesses in deriving valuable insights from their data to enhance swift decision-making.

Comparing the CRISP-DM approach to other data analysis frameworks, there are several benefits. Initially, it is a thorough and systematic process that directs analysts through every phase of the data analysis process [11]. Second, it is adaptable and can be used for different data analysis tasks, making it appropriate for many sectors and areas [12]. Thirdly, it emphasizes the importance of comprehending the analysis's business goals to ensure that conclusions are pertinent and valuable [13]. Fourthly, it focuses on collaboration between team members and stakeholders, promoting openness and communication throughout the analysis process [14]. Finally, it emphasizes the significance of upkeep and supervision, which keeps the analysis current and valuable [20]. In general, the CRISP-DM method supports data-driven decision-making and helps secure the success of data analysis projects.

3. RESULTS AND DISCUSSION

The topic in this study emphasizes the context of Raja Ampat tourist sites connected to customer reviews of goods and services obtained during a trip to Raja Ampat Regency, Southwest Papua Province, Indonesia.

3.1 Extract Sentiment of Raja Ampat Destination Visitors

The natural and cultural features of Raja Ampat Regency have the potential to draw both domestic and international tourists. [21] demonstrates how the Raja Ampat Regency's natural and cultural resources have a wide range of potential and can be promoted as a tourist destination. In addition, [22] provides evidence that residents of Raja Ampat Regency have traditional sporting activities that may be marketed as one of the cultural attractions and in-depth local knowledge that encourage tourism. Meanwhile, [23] demonstrates how business actors have been motivated by the growth of information technology to promote goods and services via websites to various digital platforms. This illustrates how information technology, cultural resources, and natural resources contribute to the growth of Raja Ampat tourism. Therefore, those who use technology can access details about Raja Ampat's accommodations, amenities, and modes of transportation.

Raja Ampat tourism is becoming more well-known for stakeholder assistance in promoting various tourist sites via information technology. [1] demonstrates the rise in domestic and international visitors to Raja Ampat from 2007 to 2019. This reflects a growth in the number of tourists visiting different tourist attractions in the Raja Ampat Regency, Southwest Papua Province, Indonesia, every year. The following figure shows the information regarding the number of visitors.

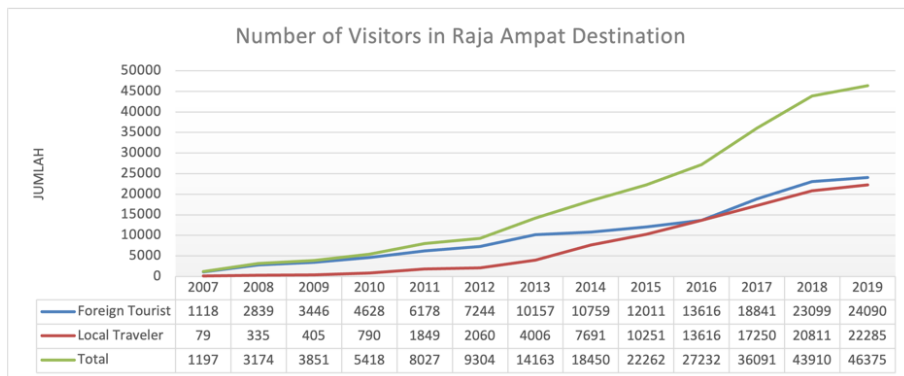


Figure 4. Number of Visitors in Raja Ampat Destination

Source : *Dinas Kebudayaan dan Pariwisata Kabupaten Raja Ampat 2022* [1]

Data in figure 4 show increased visitors to Raja Ampat-related locations between 2007 and 2019. The regularity of visitors to Raja Ampat shows that there are factors, such as attractions, accessibility, lodging, and facilities, that encourage the urge to visit the destination. According to the *Dinas Kebudayaan dan Pariwisata Kabupaten Raja Ampat 2022*, there were 20 homestays in 2009, 43 in 2013, and 200 in 2021. This demonstrates how the community supports and collaborates with other stakeholders while participating in the development of Raja Ampat. This exhibits the community's involvement, support, and cooperation with the growth of Raja Ampat's tourist destinations and the collaboration of its different stakeholders, as demonstrated below.

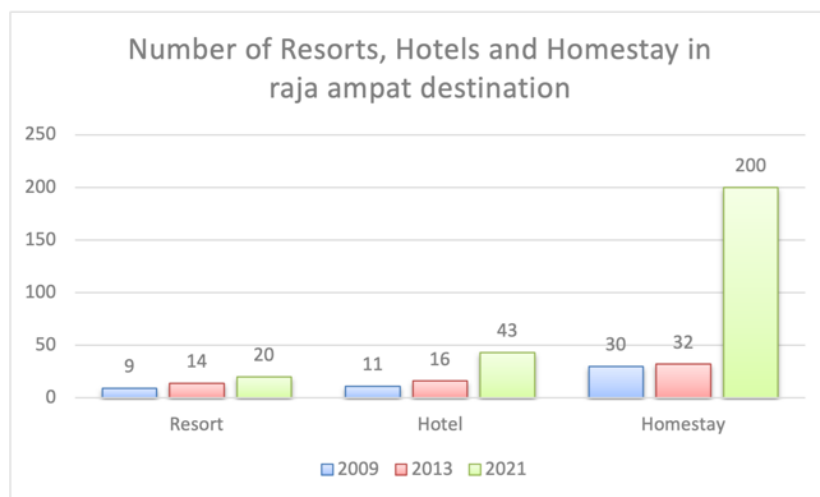


Figure 5. Number Of Resorts, Hotels, and Homestay In Raja Ampat Destination

Source : *Dinas Kebudayaan dan Pariwisata Kabupaten Raja Ampat 2022* [1]

Figure 5 depicts the growth of the hotel industry in Raja Ampat destinations between 2009 and 2021. Expanding the tourism industry stimulates corporate players to create business management information systems that may streamline production procedures, sell goods and services, and tap into a larger market [23]. In addition, [24] also demonstrates that the government, non-governmental organizations (NGOs), and other private parties support the development of homestay housing services in Raja Ampat. Meanwhile, [25] stressed that Raja Ampat's growing tourism industry financially benefits local communities. This demonstrates how Raja Ampat's local economy has increased due to the expansion of tourism. However, each visitor perceives the level of accommodation services during the tour differently. Based on information from tourist reviews, it is evident that comments made concerning the caliber of goods and services received when visiting Raja Ampat are connected to maritime tourism activities, as shown in the accompanying table.

Table 1. Negative and Positive Classification Score

Review	Scoring String	Classification
They say: "With our overall approach we hope to minimize our impact on the environment and maximize conservation efforts"	hope (0.49) drag (-0.23) destroying (-0.67) low (-0.28) dreadful (-0.49)	Score: -1,43589743589744 Negativity: 1,92307692307692
This doesn't really apply if you drag your kayaks across the coral, destroying it whilst trying to paddle out at low tide, as I have just witnessed! A dreadful contradiction!	contradiction (-0.26)	Positivity: 0,487179487179487 Uncovered Token: 53 Total Tokens: 59
We stayed some days even if in the low season. Great snorkeling spots it is the heaven for divers, highly recommend the hidden bay as it is a gem even if it not possible to swim for the presence of crocodiles.	low (-0.28) great (0.79) heaven (0.59) recommend (0.38)	Score: 1,48717948717949 Negativity: 0,282051282051282 Positivity: 1,76923076923077 Uncovered Token: 39 Total Tokens: 43

Source: Data Processing (Extract Sentiment in Rapidminer)

The method used to determine the rating for each traveler review is shown in Table 1. The computation is based on the word count, total token count, and specified weights; thus, it is clear that the more negative meaningful words in traveler evaluations, the greater the risk that the review will be identified as having a negative attitude. The results of the traveler sentiment classification are generated as recommendations for the development of goods and services in tourist places, according to several research. [26] utilizes the Support Vector Machine (SVM) method to identify preferences for goods and services when traveling, using the sentiment approach of travelers visiting Ubud tourism destinations. On the other hand, [27] demonstrates how the branding of popular tourist spots is examined using the lexicon and pivot methodologies. This shows that sentiment analysis methods can be applied with classification algorithms like SVM, NBC, DT, and k-NN that are compatible with the characteristic of the dataset.

There are many advantages to the method used to assess visitor mood in Raja Ampat Regency. The level of visitor satisfaction with the attractions and services can be determined using this strategy. Additionally, sentiment analysis can assist the government and stakeholders in making strategic choices regarding the growth of the local tourism industry. The outcomes of the sentiment analysis can also be utilized to raise awareness of central tourist locations, boost the quality of services, and make suggestions for upgrading amenities and infrastructure. Therefore, the tourist sentiment analysis approach can enhance tourism and bolster the Raja Ampat region's economy.

The sentiment analysis method has emerged as a crucial tool for assessing viewpoints and viewpoints from various sources, including social media. The benefit of this method is that it can reveal the feelings and thoughts hidden in the text's phrases. Using the sentiment analysis approach, organizations can get valuable information about client impressions and how to enhance their offerings. However, this method's flaw is that it needs help understanding the context and meaning of a text. Additionally, irony, humor, and slang that are frequently used on social media are commonly missed by sentiment analysis. To evaluate the results, users of the sentiment analysis approach must integrate it with other data analysis and consider contextual aspects.

3.2 Confusion Matrix of SVM, NBC, DT, and k-NN: Accuracy, Precision, Recall, ROC

The quality of goods and services in Raja Ampat tourist locations can be assessed using traveler sentiment analysis using Naive Bayes Classifier (NBC), Decision Tree (DT), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM) algorithms. This technique determines whether traveler evaluations posted on social media, blogs, and other online platforms are favorable, unfavorable, or neutral. This enables decision-makers to comprehend customers' viewpoints and

assess the caliber of goods and services provided by Raja Ampat tourism spots. Additionally, suggestions for enhancements and more successful marketing tactics can be made using the findings of this sentiment analysis. To maintain the accuracy and context of the analysis results, keep in mind that the results of this sentiment analysis are not dependable and still require human judgment. As a result, it is essential to employ the NBC, DT, k-NN, and SVM algorithms carefully and by the evaluation objectives to be met when using them as a tool for analyzing traveler sentiment.

TripAdvisor is a famous online platform travelers use to provide reviews and ratings of travel destinations. Using TripAdvisor as a data source to analyze traveler sentiment towards Raja Ampat destinations has the advantage of evaluating opinions and opinions from various traveler sources and points of view. Using data collected from TripAdvisor reviews, positive, negative, and neutral sentiments towards Raja Ampat tourist destinations and the services and products offered can be identified. Decision-makers can use the information from this analysis of traveler sentiment to enhance the caliber of their services and goods and develop more successful marketing plans. However, caution should be exercised when using TripAdvisor as a traveler sentiment analysis data source. The resulting information might not accurately reflect how all tourists view Raja Ampat destinations and might be impacted by biased rating subjectivity, linguistic restrictions, and a dearth of reviews. As a result, the data needs to be carefully examined and validated before the analytical results can be believed and used as a foundation for decision-making.

Confusion matrix analysis is a crucial assessment technique for assessing the effectiveness of sentiment classification algorithms like Naive Bayes Classifier (NBC), Decision Tree (DT), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM). Confusion matrix analysis can compare algorithm classification results with actual training and test data classes through calculations. The number of classification mistakes, including real positives and negatives and false positives and negatives, can now be counted. Decision-makers can decide whether the algorithm is suitable for the sentiment evaluation of Raja Ampat destinations by understanding the accuracy and error rate of the sentiment classification method. As a result, it is crucial to conduct a confusion matrix analysis on the sentiment classification algorithms NBC, DT, k-NN, and SVM to ensure that the sentiment analysis outcomes can be relied upon and used as a foundation for better decision-making.

One oversampling method used to address class imbalances in training data for sentiment classification algorithms is SMOTE (Synthetic Minority Over-sampling Technique). The amount of training data used in the NBC algorithm for sentiment analysis distinguishes between SMOTE operators and no SMOTE operators. Because training data was unbalanced and inadequately representative of those

minority classes, NBC's algorithm would have performed poorly in sentiment classification among those minority classes without SMOTE operators. Training data for minority classes can be synthetically generated by replicating or producing new data based on pre-existing training data using SMOTE operators. In this scenario, SMOTE operators will have more synthetic data to balance the proportion of each class's data and increase the accuracy of sentiment categorization in underrepresented classes. As a result, it is crucial to use SMOTE operators in the NBC algorithm for sentiment analysis to maximize the algorithm's efficacy in classifying sentiment, particularly in minority classes where there is typically a dearth of data, as shown in the figure 6 below.

Accuracy and f-measure of NBC without SMOTE Operator

accuracy: 99.14% +/- 1.10% (micro average: 99.15%)

	true Negatif	true Positif	class precision
pred. Negatif	10	1	90.91%
pred. Positif	3	455	99.34%
class recall	76.92%	99.78%	

f_measure: 99.56% +/- 0.57% (micro average: 99.56%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	10	1	90.91%
pred. Positif	3	455	99.34%
class recall	76.92%	99.78%	

Accuracy and f-measure of NBC using SMOTE Operator

accuracy: 99.89% +/- 0.35% (micro average: 99.89%)

	true Negatif	true Positif	class precision
pred. Negatif	456	1	99.78%
pred. Positif	0	455	100.00%
class recall	100.00%	99.78%	

f_measure: 99.89% +/- 0.35% (micro average: 99.89%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	456	1	99.78%
pred. Positif	0	455	100.00%
class recall	100.00%	99.78%	

Figure 6. Confusion Matrix of NBC Algorithm using SMOTE and without SMOTE Operator

Figure 6 shows that using SMOTE (Synthetic Minority Over-sampling Technique) operators on the Naive Bayes Classifier (NBC) algorithm for traveler sentiment analysis has several benefits. One of the main benefits of using SMOTE operators is to address the problem of training data imbalance in minority classes in sentiment analysis. Regarding traveler sentiment analysis, the minority class usually refers to negative sentiment towards tourist destinations or services. By using

SMOTE operators, training data in minority classes can be synthetically generated to make the amount of data in each class more balanced. This will improve NBC's algorithm's ability to classify traveler sentiment, especially among minority classes often inadequately represented in training data. In addition, with a more balanced amount of data in each category, NBC's algorithm can also improve accuracy and precision in classifying overall traveler sentiment. Therefore, using SMOTE operators on NBC's algorithm improves the quality of traveler sentiment analysis and provides more accurate and representative information about the tourist experience in a particular destination. It also shows the different results of DT, k-NN, and SVM algorithms, as shown in figure 7 below.

Accuracy DT using SMOTE Operator			
accuracy: 99.12% +/- 1.13% (micro average: 99.12%)			
	true Negatif	true Positif	class precision
pred. Negatif	451	3	99.34%
pred. Positif	5	453	98.91%
class recall	98.90%	99.34%	

Accuracy k-NN using SMOTE Operator			
accuracy: 50.22% +/- 0.78% (micro average: 50.22%)			
	true Negatif	true Positif	class precision
pred. Negatif	456	454	50.11%
pred. Positif	0	2	100.00%
class recall	100.00%	0.44%	

Accuracy SVM using SMOTE Operator			
accuracy: 100.00% +/- 0.00% (micro average: 100.00%)			
	true Negatif	true Positif	class precision
pred. Negatif	456	0	100.00%
pred. Positif	0	456	100.00%
class recall	100.00%	100.00%	

Figure 7. Accuracy of DT, k-NN, and SVM Algorithm using SMOTE Operator

The accuracy values for the DT, k-NN, and SVM algorithms when the SMOTE operator is used are shown in Figure 7. Additionally, each algorithm's t-test results differ before and after the SMOTE operator. When comparing the performance of algorithms for traveler sentiment analysis, pairwise t-tests between NBC, DT, k-NN, and SVM sentiment classification algorithms with and without using SMOTE operators are crucial. A paired t-test can be used to compare statistically two data sets from the same group collected under two distinct situations. This group refers to the data collection used for traveler sentiment research. The two other requirements, however, deal with the presence or absence of SMOTE operators on each classification algorithm. The pairwise t-test will provide important information in evaluating whether using SMOTE operators in the

classification algorithm offers a significant difference in the algorithm's performance in classifying traveler sentiment, as shown in Figure 8 below.

Pairwise t-test of NBC, DT, k-NN, SVM without SMOTE Operator

A	B	C	D	E
	0.977 +/- 0.016	0.991 +/- 0.011	0.989 +/- 0.015	0.989 +/- 0.015
0.977 +/- 0.016		0.025	0.080	0.081
0.991 +/- 0.011			0.728	0.723
0.989 +/- 0.015				0.995
0.989 +/- 0.015				

Pairwise t-test of NBC, DT, k-NN, SVM with SMOTE Operator

A	B	C	D	E
	0.502 +/- 0.008	0.999 +/- 0.003	0.991 +/- 0.011	1.000 +/- 0.000
0.502 +/- 0.008		0.000	0.000	0.000
0.999 +/- 0.003			0.055	0.331
0.991 +/- 0.011				0.025
1.000 +/- 0.000				

Figure 8. Pairwise t-Test of NBC, DT, k-NN, SVM with and without SMOTE Operator

Figure 8 displays the outcomes of a paired t-test that contrasted two instances of the algorithms' accuracy, precision, recall, and f-measure score. The results of the paired t-test will be vital in determining whether the operator's use of SMOTE offers a significant difference in the performance of algorithms for categorizing traveler sentiment. Pairwise t-tests with and without using SMOTE operators are essential to evaluate the performance of the NBC, DT, k-NN, and SVM algorithms for traveler sentiment analysis. In addition, the Receiver Operating Characteristics (ROC) of NBC, DT, k-NN, and SVM Algorithms are shown below.

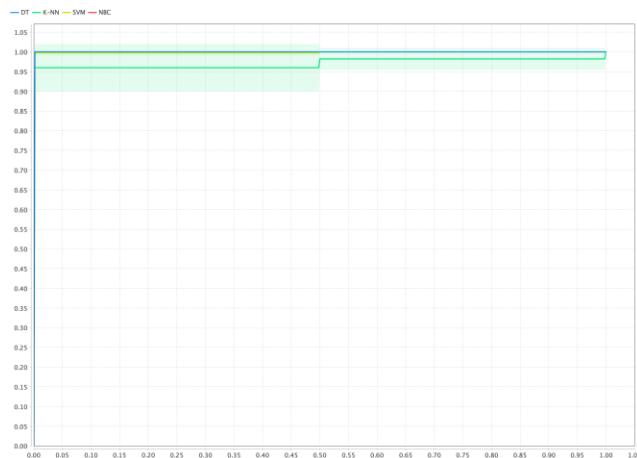


Figure 9. Receiver Operating Characteristics (ROC) of NBC, DT, k-NN, and SVM Algorithm

Figure 9 displays the outcome of Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) calculations relating to the NBC, DT, k-NN, and SVM algorithms with and without operators. SMOTE is a crucial stage in assessing how well algorithms for sentiment analysis of travelers perform. The True Positive Rate (TPR) and False Positive Rate (FPR) of classification are compared using the ROC curve (FPR). AUC can also be used to demonstrate a model's overall success. In the context of traveler sentiment analysis, the effectiveness of algorithms in differentiating positive and negative emotions from traveler reviews can be evaluated using ROC and AUC. Each algorithm's classification methods utilizing and without SMOTE operators will be examined using ROC and AUC testing. With higher AUC values, the system performs better at classifying traveler sentiment in ROC and AUC tests. Comparing the ROC and AUC of the NBC, DT, k-NN, and SVM algorithms with and without the operator SMOTE is thus a critical step in evaluating algorithm performance for traveler sentiment analysis.

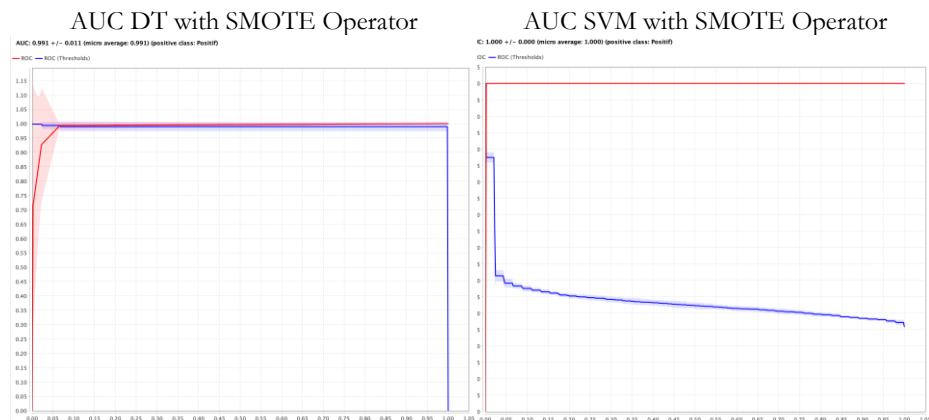


Figure 10. Area Under the Curve (AUC) of DT and SVM Algorithm

Figure 10 demonstrates that the DT and SVM algorithms provide better Area Under the Curve (AUC) results than the NBC and k-NN techniques. Thus, it shows that when sentiment analysis is performed using the Decision Tree (DT) and Support Vector Machine (SVM) algorithms, the values for the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are increased. The test findings demonstrate that the SMOTE-added DT and SVM algorithms have greater ROC and AUC values than their non-SMOTE-added counterparts. It shows that DT and SVM algorithms perform better when SMOTE is used in sentiment categorization on unbalanced data. Therefore, SMOTE in DT and SVM algorithms can be utilized as a substitute to solve the sentiment classification issue with unbalanced data.

4. CONCLUSION

The study's findings show that tourists had overwhelmingly positive perceptions of Raja Ampat's tourist attractions, with a focus on the region's cultural richness and underwater beauty. It is evident that DT and SVM, with the help of SMOTE operators, have the best algorithm performance based on the results of sentiment analysis using the NBC, DT, k-NN, and SVM classification methods. The values of the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are increased when Synthetic Minority Over-sampling Technique (SMOTE) operators are used. The test findings demonstrate that the SMOTE-added DT and SVM algorithms have greater ROC and AUC values than their non-SMOTE-added counterparts. This illustrates how DT and SVM algorithms can better perform sentiment categorization on unbalanced data when SMOTE is used. Therefore, SMOTE in DT and SVM algorithms can be utilized as a substitute to solve the sentiment classification issue with unbalanced data.

ACKNOWLEDGEMENT

Thanks to the Tourism Department, Faculty of Business Administration and Communication, *Lembaga Penelitian dan Pengabdian Masyarakat* (LPPM), and the Atma Jaya Catholic University of Indonesia due to the support for this research and publication.

REFERENCES

- [1] K. M. Kasikoen, "Pengaruh Kegiatan Pariwisata Terhadap Pengembangan Wilayah Studi Kasus Kabupaten Raja Ampat," *KENDALI Econ. Soc. Humanit.*, vol. 1, no. 2, pp. 109–115, 2022, doi: <https://doi.org/10.58738/kendali.v1i2.89>.
- [2] A. G. Gustiarini, "Karakteristik Penginapan Kampung Wisata Arborek Kabupaten Raja Ampat Provinsi Papua Barat," *Barista J. Kaji. Bhs. dan Pariwisata*, vol. 9, no. 2, pp. 37–46, 2022, doi: [10.34013/barista.v9i02.648](https://doi.org/10.34013/barista.v9i02.648).
- [3] M. F. Hasa, R. Soekarta, T. H. I. Alam, R. D. P. Rawi, M. R. Setyawan, and S. N. Kayatun, "Pemanfaatan Sosial Media sebagai Wadah Promosi Wisata Kebun Buah Naga di Raja Ampat," *J. Pengabd. Dharma Laksana Mengabdikan untuk Negeri*, vol. 4, no. 2, pp. 277–282, 2022, doi: <https://dx.doi.org/10.32493/j.pdl.v4i2.18246>.
- [4] M. P. Saptono, E. Murniyasih, L. Suryani, and R. Faisal, "Pengembangan Aplikasi Pariwisata Berbasis Android Dalam Upaya Peningkatan Aksesibilitas UKM (Desa Wisata) Kampung PAM di Kabupaten Raja Ampat Provinsi Papua Barat," *J. Dedication to Papua Community*, vol. 3, no. 1, pp. 220–230, 2020.
- [5] M. S. Rumetna and T. N. Lina, "Sistem Informasi Kampung Wisata Arborek Dengan Metode Waterfall," *Inf. Syst. Educ. Prof.*, vol. 5, no. 1, pp.

- 31–40, 2020, doi: <https://doi.org/10.51211/isbi.v5i1>.
- [6] E. Tanati, W. Wahyudi, and A. S. Sinery, “Tingkat Partisipasi Masyarakat Lokal dalam Pengelolaan Ekowisata di Kampung Saporkren Distrik Waigeo Selatan Kabupaten Raja Ampat,” *J. Sumberd. Akuatik Indopasifik*, vol. 4, no. 2, pp. 193–202, 2020, doi: 10.46252/jsai-fpik-unipa.2020.vol.4.no.2.118.
- [7] P. Rita, R. Ramos, M. T. Borges-Tiago, and D. Rodrigues, “Impact of the rating system on sentiment and tone of voice: A Booking.com and TripAdvisor comparison study,” *Int. J. Hosp. Manag.*, vol. 104, no. 2, pp. 1–12, 2022, doi: 10.1016/j.ijhm.2022.103245.
- [8] Y. Harwani, “Constructing Brand Personality from the TripAdvisor Online Reviews,” *Int. J. Bus. Manag. Technol.*, vol. 5, no. 4, pp. 152–157, 2021.
- [9] Y. A. Singgalen, “Analisis Sentimen Pengunjung Pulau Komodo dan Pulau Rinca di Website Tripadvisor Berbasis CRISP-DM,” *J. Inf. Syst. Res.*, vol. 4, no. 2, pp. 614–625, 2023, doi: 10.47065/josh.v4i2.2999.
- [10] Y. A. Singgalen, “Analisis Perilaku Wisatawan Berdasarkan Data Ulasan di Website Tripadvisor Menggunakan CRISP-DM: Wisata Minat Khusus Pendakian Gunung Rinjani dan Gunung Bromo,” *J. Comput. Syst. Informatics*, vol. 4, no. 2, pp. 326–338, 2023, doi: 10.47065/josyc.v4i2.3042.
- [11] Y. A. Singgalen, “Analisis Sentimen Wisatawan terhadap Taman Nasional Bunaken dan Top 10 Hotel Rekomendasi Tripadvisor Menggunakan Algoritma SVM dan DT berbasis CRISP-DM,” *J. Comput. Syst. Informatics*, vol. 4, no. 2, pp. 367–379, 2023, doi: 10.47065/josyc.v4i2.3092.
- [12] H. J. Christanto and Y. A. Singgalen, “Sentiment Analysis on Customer Perception towards Products and Services of Restaurant in Labuan Bajo,” *J. Inf. Syst. Informatics*, vol. 4, no. 3, pp. 511–523, 2022, doi: 10.51519/journalisi.v4i3.276.
- [13] Y. A. Singgalen, “Analisis Sentimen Wisatawan Melalui Data Ulasan Candi Borobudur di Tripadvisor Menggunakan Algoritma Naïve Bayes Classifier,” *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, p. 1343–1352, 2022, doi: 10.47065/bits.v4i3.2486.
- [14] 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.
- [15] Y. A. Singgalen, “Analisis Sentimen dan Pemodelan Topik dalam Optimalisasi Pemasaran Destinasi Pariwisata Prioritas di Indonesia,” *J. Inf. Syst. Informatics*, vol. 4, no. 1, pp. 459–470, 2021.
- [16] E. Widodo, A. T. Sasongko, and A. Z. Kamalia, “Klasifikasi Barang Paling Laku (PARETO) Indomaret Untung Suropati 35 (T3M1) Menggunakan Rapidminer dengan Metode Naive Bayes,” *J. Teknol. Pelita Bangsa*, vol. 13, no. 4, pp. 231–234, 2022.

- [17] M. Nasution, A. A. Ritonga, and A. P. Juledi, "Implementasi Rapidminer dalam Mengklasifikasi Indeks Demokrasi," *J. Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 99–106, 2022.
- [18] Auliya, T. Suprapti, and G. Dwiletari, "Analisis Algoritma Klasifikasi Neural Network pada Penderita Penyakit Kanker Payudara," *J. Ilm. Betrik*, vol. 14, no. 1, pp. 175–183, 2023.
- [19] M. R. Qisthiano, P. A. Prayesy, and I. Ruswita, "Penerapan Algoritma Decision Tree dalam Klasifikasi Data Prediksi Kelulusan Mahasiswa," *G-Tech J. Teknol. Terap.*, vol. 7, no. 1, pp. 21–28, 2022, doi: <https://doi.org/10.33379/gtech.v7i1.1850>.
- [20] A. Fadli, M. I. Zulfa, and Y. Ramadhani, "Performance Comparison of Data Mining Classification Algorithms for Early Warning System of Students Graduation Timeliness," *J. Teknol. dan Sist. Komput.*, vol. 6, no. 4, pp. 158–163, 2018, doi: 10.14710/jtsiskom.6.4.2018.158-163.
- [21] L. G. Komiter and P. Rahardjo, "Studi Postensi Destinasi Wisata Desa Saporkren Kabupaten Raja Ampat," *STUPA Sains, Teknol. Urban, Perancangan, Arsit.*, vol. 5, no. 1, pp. 365–380, 2023, doi: 10.24912/stupa.v5i1.22693.
- [22] B. Bamburo, L. Pratama, and S. Anwar, "Permainan Olahraga Tradisional Raja Ampat Papua Barat," *UNIMUDA Sport*, vol. 3, no. 2, pp. 1–12, 2022.
- [23] F. G. J. Rupilele, "Rancang Bangun Sistem Informasi Tour dan Travel Berbasis Web Pada Raja Ampat Tom's Adventure," *J. Jendela Ilmu*, vol. 2, no. 2, pp. 58–63, 2021, doi: 10.34124/ji.v2i2.114.
- [24] Selna Adesetiani, L. M. Kolopaking, and E. Eriyatno, "Strategi Pengembangan Usaha Wisata Homestay Berbasis Komunitas di Kabupaten Raja Ampat, Provinsi Papua Barat," *Sodality J. Sociol. Pedesaan*, vol. 9, no. 3, pp. 1–23, 2021, doi: 10.22500/9202136386.
- [25] E. Tanati, W. Wahyudi, and A. S. Sinery, "Tingkat Partisipasi Masyarakat Lokal dalam Pengelolaan Ekowisata di Kampung Saporkren Distrik Waigeo Selatan Kabupaten Raja Ampat," *J. Sumberd. Akuatik Indopasifik*, vol. 4, no. 1, pp. 193–202, 2020, doi: 10.46252/jsai-fpik-unipa.2020.vol.4.no.2.118.
- [26] I. W. B. Suryawan, N. W. Utami, and K. Q. Fredlina, "Analisis Sentimen Review Wisatawan Pada Objek Wisata Ubud Menggunakan Algoritma Support Vector Machine," *JINTEKS (Jurnal Inform. Teknol. dan Sains)*, vol. 5, no. 1, pp. 133–140, 2023.
- [27] Larasati, S. Nabilla, and E. Haryanto, "Sentiment Analysis untuk Review Destinasi Wisata Unggulan Gunung Kidul Menggunakan Lexicon dan Pivot," *Indones. J. Bus.*, vol. 5, no. 2, pp. 102–111, 2022.