Sentiment Analysis Of Tourist Reviews Using K-Nearest Neighbors Algorithm And Support Vector Machine

Anita Wulan Sari1*, Teguh Iman Hermanto2, Meriska Defriani3,
1,2,3) Sekolah Tinggi Teknologi Wastukancana, Purwakarta, Indonesia
1)anitawulan04@wastukancana.ac.id, 2)teguhiman@wastukancana.ac.id, 3)meriska@wastukancana.ac.id.

Submitted : May 23, 2023 | Accepted : Jun 15, 2023 | Published : Jul 1, 2023

Abstract: After Indonesia was awarded as a country with extraordinary natural charm, many foreign tourists came to Indonesia. According to the records of the Central Bureau of Statistics for 2020, approximately 5.47 million foreign tourists entered Indonesia. With the large number of foreign tourist visits, the need for tourist attractions is increasing, but finding information is now not difficult. One source of information for finding reviews of tourist attractions is TripAdvisor. On this website, there is a lot of information or reviews about various tourist attractions. However, the number of reviews makes tourists confused about identifying the quality of tourist attractions to be visited, so sentiment analysis needs to be done. Sentiment analysis itself is a technique to extract, identify, and understand sentiments or opinions contained in a text. In this research, two classification methods will be used in sentiment analysis techniques, namely K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM). Besides that, the object of this research will be to focus on the most popular tourist attractions in Indonesia according to Trip Advisor, namely Waterbom Bali, Mandala Suci Wenara Wana, Teras Sawah Tegalalang, Pura Tanah Lot, and Pura Luhur Uluwatu. The purpose of the research is to find out the results of accurate sentiment analysis for the five tourist attractions and compare the two algorithms used. and after testing, it was found that the Support Vector Machine algorithm is superior to the K-Nearest Neighbors algorithm.

Keywords: Sentiment Analysis; Tourist Spot Reviews; TripAdvisor; K-Nearest Neighbors; Support Vector Machine;

INTRODUCTION

Indonesia is a country that has an extraordinary natural charm that makes it attractive to both local and foreign tourists. Reporting from the British website money.co.uk with the title "Natural Beauty Report," which was released on February 7, 2022, Indonesia was named the most beautiful country with the most natural panoramas, beating New Zealand, Colombia, Japan, and France.

Indonesia has been named the most beautiful country, inviting many foreign tourists to visit Indonesia. This is supported by facts from the Indonesian central statistics agency, with the number of foreign tourist visits to Indonesia for the January–December 2022 period reaching 5.47 million visits, up 251.28% compared to January–December 2021, which was only 1.55 million visits. The more foreign tourists visit Indonesia, the more the country's foreign exchange resources in the tourism sector will increase (Somantri & Dairoh, 2019) therefore, the development and management of tourist attractions...
must be carried out, both by the government and the managers of tourist attractions, so that they remain an attraction for foreign tourists.

In addition, the number of foreign tourists visiting has increased the need for tourist information, and with the development of technology today, tourists can easily find information about tourist attractions through reviews on social media or digital platforms, one of which is TripAdvisor. TripAdvisor is a website-based digital platform developed specifically to provide travel guidance through information on hotel and restaurant accommodations and transportation services to the location of the destination that tourists want (Singgalen, 2022). Prospective tourists can find references to tourist attractions based on the experiences of tourists who have visited these attractions by uploading reviews from these tourists (Arifiyanti et al., 2022).

The TripAdvisor website displays many reviews and opinions about tourist destinations. These reviews and opinions have important value for visitors or tourists in determining which tourist attractions to visit. However, the number of reviews that are too many and diverse makes it difficult for visitors to evaluate the quality of the tourist attractions. In this case, sentiment analysis of tourist attraction reviews can be used to help visitors choose tourist attractions that suit their needs. In addition, with sentiment analysis, we will get quick feedback from tourists by utilizing the reviews they have given so that the stakeholders involved can take action on what steps to take in the future in order to improve the quality of tourist attractions (Pati & Umar, 2022). Sentiment analysis itself is a technique for extracting, identifying, and understanding sentiments or opinions contained in a text (Atimi & Pratama, 2022). This technique is often used in natural language processing (NLP) to obtain information from text. With this technique, visitors can find out the opinion or tendency of opinion on an issue or object by a person, whether it tends towards a negative or positive opinion (Purba, 2023).

In this research, two classification methods will be used in sentiment analysis techniques, namely K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM). The KNN method is a classification method based on the concept of nearest neighbors (Irawan et al., 2022), while SVM is a classification method based on the concept of hyperplane separation (Rivanie et al., 2021). In addition, this research will focus on tourist destinations on the island of Bali, with the object of research being the five most popular tourist attractions in Indonesia according to the TripAdvisor site. The purpose of this research is to provide information about the quality of the five tourist attractions in Bali using visitor sentiment and to determine the accuracy level of the comparison of the two methods tested.

**LITERATURE REVIEW**

Marchenda, Dian, and Bayu conducted sentiment analysis on reviews of applications using support vector machine and Nave Bayes. This study uses the Allo Bank application subject whose data is obtained from the Google Play Store review, and the results state that the Support Vector Machine method gets an accuracy value of 94.29% and the Nave Bayes Classifier method gets an accuracy value of 93.97%, so for the comparison results, the Support Vector Machine algorithm produces an accuracy value that is superior to using the Nave Bayes Classifier algorithm (Madjid et al., 2023).

Pavithaa, et al. Conducting sentiment analysis related to movie recommendations, this study conducted sentiment analysis on movie reviews by comparing the Naive Bayes (NB) classifier and Support Vector Machine (SVM) algorithms. The result is that SVM accuracy gets 98.63%, while the NB accuracy score is 97.33%. Thus, SVM exceeds NB and proves to be more suitable for sentiment analysis (Pavitha et al., 2022).

Using the Naive Bayes Classifier and K-Nearest Neighbor Methods, Pati and Umar's research on "Sentiment Analysis of Visitor Comments on Lake Weekuri Tourist Attractions" found that the accuracy rate for the K-Nearest Neighbor algorithm is a respectable 76.53% while that of the Naive Bayes Method is only 73.47%. (Pati & Umar, 2022).

Sentiment Analysis of Indonesian Digital Payment Customer Satisfaction Towards GOPAY, DANA, and ShopeePay was the focus of research by Maharani and Triayudi. Results of testing using the Nave Bayes and K-Nearest Neighbor methods showed that the K-Nearest Neighbor method obtained an accuracy value on GOPAY of 32.68%, DANA of 35.31%, and ShopeePay of 33.12%, while the Nave Bayes method obtained an accuracy value on GOPAY of 27.15%, DANA of 29.86%, and ShopeePay of 30.40%. Therefore, it can be said that using the K-Nearest Neighbor method in sentiment analysis**
pertaining to digital payments in Indonesia yields superior accuracy results than using the Nave Bayes method. (Maharani & Triayudi, 2022).

Then the research of Abdul Mohaimin et al. Conducting sentiment analysis of airline reviews with the methods used, namely Nave Bayes and Support Vector Machine. The results show that in the case of airline reviews, SVM provides much better accuracy results (82%), while the Nave Bayes algorithm is only 76% (Rahat et al., 2020).

Then came the study by Josen, Josen, and others. The classification results for the study, "Analysis of Sentiment Classification Reviews on E-Commerce Shopee Based on Word Cloud with Naive Bayes and K-Nearest Neighbor Methods," demonstrate that the KNN method performs better, with a classification accuracy value of 92.8%, compared to the Naive Bayes method's accuracy value of 91.4%. (Limbong et al., 2022).

Furthermore, Ziedhan Alifio Dieksona et al. Conducting sentiment analysis of customer reviews with a case study of the Traveloka application, this research uses three classification methods: support vector model (SVM), logistic regression, and Nave Bayes. And the results show that SVM has better accuracy in determining the sentiment of tweets about Traveloka (Diekson et al., 2023).

Finally, research conducted by Rivanda and Aris with the research title “The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia”. The results showed the influence of fake accounts, which can reduce the performance of sentiment classification. Experimental results with both algorithms also prove that the Support Vector Machine algorithm has better performance than the Nave Bayes algorithm, with the highest accuracy value of 80.6% (Pratama & Tjahyanto, 2021).

METHOD

This research requires a research methodology consisting of steps of activities and processes used in organized and systematic research. The research stages describe the flow of research conducted from the beginning until the research ends (Rambe et al., 2023). For this reason, the stages of this research use the framework shown in Figure 1.

fig. 1 General Research Method

Literature Study

The literature study stage is the process of collecting data to be used as reference material that supports research. Literature studies are obtained through journals and articles related to sentiment analysis, classification, tourist attractions, K-nearest neighbor, and support vector machines.

*name of corresponding author
Data Collection
The data used in this research is tourist attraction review data obtained from the Trip Advisor website. The data collection process is carried out using scraping techniques with additional web scraper applications from Google Chrome, and the reviews used are in Indonesian. The data collection stage is further divided into three stages. The first stage is scraping. Scraping data is assisted by using an additional application from Google Chrome, namely the web scraper. The second stage is downloading the dataset.

Text Preprocessing
The text preprocessing stage is divided into several processes, namely: labeling, which is the process of giving sentiment values to data based on sentiment classes, which are divided into two classes (Rahayu et al., 2022), namely positive and negative. A score of 1-2 is stated as a negative class with a negative label, and a score of 4-5 is a positive sentiment with a positive code, while the score of 3 is discarded because it is feared that it will affect the sentiment results. Then Case folding is done to homogenize each word contained in the review into lowercase letters. Then cleansing to remove characters or symbols, emoticons, and url links that have no informational relationship. Then tokenizing is done to separate sentences into words. after that normalization to change non-standard words into standard words. Then filtering is done using stop removal to remove unnecessary words. Finally, stemming is done to change words that have affixes into basic words (Putra et al., 2020).

Feature Extraction
In the feature extraction process, the data is converted into a matrix and weighted for each word using the TF-IDF algorithm. The results of the TF-IDF value weighting will show the similarity between documents in the dataset.

Classification
At this stage, the data that has undergone text preprocessing will be divided into training data and testing data. The training data will be classified using the K-Nearest Neighbor and Support Vector Machine methods. Once the training data model has been obtained, the model will be tested using the testing data to get the classification results. Classification is done using Jupyter tools and the Python programming language.

Evaluation
In this evaluation stage, the comparison results of the classification of review data between the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) algorithms using a confusion matrix can be seen. The confusion matrix helps in comparing the classification result data between the two algorithms. In the evaluation process, it can produce values by measuring accuracy, precision, recall, and the F1-Score.

RESULT
Data Collection
The research process was carried out by retrieving the necessary data, namely using scraping techniques through the TripAdvisor website by adding a web scraper extension. Scraping was carried out on February 26–28, 2023. The following is the web scraping process using Webscraper.io to extract the dataset of reviews about tourist attractions in Bali (Simarmata & Phanie, 2023):
1. Open Webscraper.io and create "Scraper".
2. Specify the URL of the website you want to scrape.
3. Create a "selector" to specify the part of the web page you want to extract data from.
4. Specify the columns you want to extract and save them in Excel format.
5. Run "Scraper" and wait for the webscraping process to finish.

The results of the collected review data are then presented in Table 1 below.
Table 1. Scraping Data Result

<table>
<thead>
<tr>
<th>Tourist Spot</th>
<th>Total Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterbom Bali</td>
<td>500</td>
</tr>
<tr>
<td>Mandala Suci Wenara</td>
<td>500</td>
</tr>
<tr>
<td>Tegalalang Rice Terrace</td>
<td>500</td>
</tr>
<tr>
<td>Tanah Lot Temple</td>
<td>500</td>
</tr>
<tr>
<td>Uluwatu Temple</td>
<td>500</td>
</tr>
</tbody>
</table>

Text Preprocessing

Labeling is carried out to group each review data using Microsoft Excel by giving class attributes to each review data containing positive and negative sentences. For data that contains neutral sentences, duplicates and contains promotional statements, it is deleted by the author. Table 2 and Figures 2 to 6 exhibit the findings of the number of positive and negative data that have been tagged on each tourist destination dataset.

Table 2. Labeling Result

<table>
<thead>
<tr>
<th>Tourist Spot</th>
<th>Sum of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Waterbom Bali</td>
<td>76</td>
</tr>
<tr>
<td>Mandala Suci Wenara</td>
<td>130</td>
</tr>
<tr>
<td>Tegalalang Rice Terrace</td>
<td>179</td>
</tr>
<tr>
<td>Tanah Lot Temple</td>
<td>59</td>
</tr>
<tr>
<td>Uluwatu Temple</td>
<td>115</td>
</tr>
<tr>
<td><strong>Total Data</strong></td>
<td><strong>559</strong></td>
</tr>
</tbody>
</table>

![fig. 2 Rating waterbom bali](image1)

![fig. 3 Rating mandala suci wenara](image2)

![fig. 4 Rating tegalalang rice](image3)

![fig. 5 Rating tanah lot temple](image4)

![fig. 6 Rating uluwatu temple](image5)

To clean the data, there are several libraries needed, namely NLTK (Natural Language Toolkit): NLTK is one of the most popular natural language processing libraries. NLTK provides various functions for cleaning, tokenization, stemming, lemmatization, word indexing, unnecessary word removal, and many more. The following is an image of the text preprocessing results on the five datasets.

*name of corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
In the tourist destinations Waterbomb Bali, Mandal Suci Wenara Wana, Tegalalang Rice Terraces, Tanah Lot Temple, and Uluwatu Noble Temple displayed in the figure below, there are a number of dominant words that frequently appear.
Feature Extraction

The next feature extraction process uses TF-IDF. The library needed for the feature extraction process is Scikit-Learn: Scikit-learn is a powerful machine learning library that also has a text processing module. This library provides text processing algorithms such as document vectorization, TF-IDF (Term Frequency-Inverse Document Frequency), and theme modeling with Latent Dirichlet Allocation (LDA). The following source code is used to perform TF-IDF feature selection on the dataset.

```
fig. 17 code TF-IDF
```
Classification

After the feature extraction process, proceed to the classification stage using K-NN and SVM. The following source code is used to perform classification on the dataset.

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors=3)
classifier.fit(X_train, y_train)
knn_pred = classifier.predict(X_test)
```

To see how accurately the K-Nearest Neighbor and Support Vector Machine algorithms that have been implemented on the dataset can be calculated with the confusion matrix. The calculation will be explained in more detail in the evaluation.

Evaluation

In this evaluation section, testing is done using the confusion matrix method. In the evaluation process, it can produce values by measuring accuracy, precision, recall, and F1-Score of the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) algorithms.

K-Nearest Neighbor (K-NN)

In terms of testing using machine learning, we later discovered that the K-Nearest Neighbor (K-NN) method's accuracy on the Waterbom dataset was 67%. The accuracy, precision, recall, and F1-score values for testing with the Waterbom dataset are shown in Figure 20's confusion matrix.

```
precision  recall  f1-score  support
negatif    0.75    0.30    0.43    10
positif    0.65    0.93    0.76    14
```

On the Mandala Suci Wenara dataset, the test accuracy obtained using the K-Nearest Neighbor (K-NN) approach is 60%. Figure 21 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

```
from sklearn.svm import SVC
classifier = SVC(kernel = 'linear', random_state = 0)
classifier.fit(X_train, y_train)
svm_pred = classifier.predict(X_test)
```

*name of corresponding author
On the Tegalalang Rice Terrace dataset, the test accuracy result utilizing the K-Nearest Neighbor (K-NN) approach is 69%. Figure 22 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

On the Tanah Lot Temple dataset, the test accuracy obtained using the K-Nearest Neighbor (K-NN) approach is 71%. Figure 23 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.
On the Uluwatu Temple dataset, the test accuracy obtained using the K-Nearest Neighbor (K-NN) approach is 81%. Figure 24 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

Support Vector Machine (SVM)

We also performed testing using the Support Vector Machine algorithm, employing steps identical to the K-Nearest Neighbor method, and we discovered that testing using this method on the Waterbom dataset had a value of 88%. The confusion matrix, accuracy, precision, recall, and F1-score figures for testing with the Waterbom dataset are shown in Figure 25.

On the Mandala Suci Wenara dataset, the test accuracy obtained using the support vector machine approach is 83%. Figure 26 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

*name of corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
The Tegalalang Rice Terrace dataset's test accuracy score using the Support Vector Machine approach is 89%. Figure 27 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

![Confusion Matrix of SVM method on the Tegalalang Rice Terrace dataset](image1)

On the Tanah Lot Temple dataset, the test accuracy obtained using the support vector machine approach is 88%. Figure 28 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

![Confusion Matrix of SVM method on the Tanah Lot Temple dataset](image2)

On the Uluwatu Temple dataset, the test accuracy obtained using the support vector machine approach is 95%. Figure 29 shows the confusion matrix as well as the accuracy, precision, recall, and F1-score values connected to testing with the dataset.

![Confusion Matrix of SVM method on the Uluwatu Temple dataset](image3)

*name of corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
To see more clearly the comparison of accuracy results between the K-Nearest Neighbor and Support Vector Machine methods, see Figure 3 below:

![Comparison of Accuracy Results of K-NN and SVM Methods](image)

**DISCUSSIONS**

Based on the test accuracy results that have been obtained, it can be concluded that the Support Vector Machine classification method produces a higher accuracy value compared to the K-Nearest Neighbors classification method. In testing using the Support Vector Machine method, the accuracy value obtained at Waterbom Bali is 88%, Mandala Suci Wenara Wana is 83%, Tegalalang Terraces is 89%, Tanah Lot Temple is 88%, and Uluwatu Temple is 95%. While the K-Nearest Neighbors method obtained an accuracy value at Waterbom Bali of 67%, Mandala Suci Wenara Wana by 60%, Tegalalang Terraces by 69%, Tanah Lot Temple by 71%, and Uluwatu Temple by 81%. For further research, it is recommended to add or compare with other classification methods such as decision trees, random forest classifiers, logistic regression, or other classification methods.

**CONCLUSION**

The sentiment analysis process in this research is carried out using the K-Nearest Neighbors algorithm and the Support Vector Machine. The analysis process begins with scraping tourist attraction review data on the TripAdvisor website with the help of a web scraper application, then entering the text preprocessing stage before the data is case-folded, cleansed, tokenized, normalized, stemmed, and filtered. The data is first labeled manually. After that, it enters the word weighting stage using TF-IDF, and the weighting results will be entered into two algorithm models, namely K-Nearest Neighbors and Support Vector Machine. Sentiment classification uses accuracy parameters against the five tourist attractions as a comparison, and evaluation results using the confusion matrix show the Support Vector Machine method gets a higher accuracy value against all tourist attractions that are the object of this study compared to using the K-Nearest Neighbor (K-NN) method.

**REFERENCES**


*name of corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.