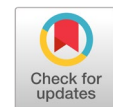


# Sentiment classification from reviews for tourism analytics



Nur Aliah Khairina Mohd Haris <sup>a,1</sup>, Sofianita Mutalib <sup>a,c,2,\*</sup>, Ariff Md Ab Malik <sup>b,c,3</sup>,  
Shuzlina Abdul-Rahman <sup>a,c,4</sup>, Siti Nur Kamaliah Kamarudin <sup>a,5</sup>

<sup>a</sup> School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, 40450 Shah Alam, Selangor

<sup>b</sup> Faculty of Business and Management, Universiti Teknologi MARA, Puncak Alam Campus, UiTM Selangor, 42300 Puncak Alam, Selangor

<sup>c</sup> Research Initiative Group Intelligent Systems, Universiti Teknologi MARA, 40450 Shah Alam, Selangor

<sup>1</sup> nur.aliahkhairina@gmail.com; <sup>2</sup> sofianita@uitm.edu.my; <sup>3</sup> ariff215@uitm.edu.my; <sup>4</sup> shuzlina@uitm.edu.my; <sup>5</sup> snkamaliah@uitm.edu.my

\* corresponding author

## ARTICLE INFO

### Article history

Selected paper from The 2022 5th International Symposium on Advanced Intelligent Informatics (SAIN'22), Yogyakarta (Virtually), September 14, 2022, <http://sain.ijain.org/2022/>. Peer-reviewed by SAIN'22 Scientific Committee and Editorial Team of IJAIN journal

Received July 4, 2022

Revised August 17, 2022

Accepted September 17, 2023

Available online March 31, 2023

### Keywords

Classification  
Sentiment analysis  
Text mining  
Tourism

## ABSTRACT

User-generated content is critical for tourism destination management as it could help them identify their customers' opinions and come up with solutions to upgrade their tourism organizations as it could help them identify customer opinions. There are many reviews on social media, and it is difficult for these organizations to analyze them manually. By applying sentiment classification, reviews can be classified into several classes and help ease decision-making. The reviews contain noisy contents, such as typos and emoticons, which could affect the accuracy of the classifiers. This study evaluates the reviews using Support Vector Machine and Random Forest models to identify a suitable classifier. The main phases in this study are data collection, preparation, labeling, and modeling. The reviews are labeled into three sentiments; positive, neutral, and negative. During pre-processing, steps such as removing the missing value, tokenization, case folding, stop words removal, stemming, and applying n-grams are performed. The result of this research is evaluated by looking at the performance of the models based on accuracy, where the result with the highest accuracy is chosen as the solution. In this study, data is data from TripAdvisor and Google reviews using web scraping tools. The findings show that the Support Vector Machine model with 5-fold cross-validation is the most suitable classifier with an accuracy of 67.97% compared to Naive Bayes with 61.33% accuracy and the Random Forest classifier with 63.55% accuracy. In conclusion, the result of this paper could provide important information in tourism besides determining the suitable algorithm to be used for Sentiment Analysis related to the tourism domain.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## 1. Introduction

Tourism is one of the fastest-growing sectors globally; it is a significant source of income for many countries. Through the years, tourism has evolved from the conventional to the implementation of technology in tourism. One key factor that has evolved along with the implementation of technology is feedback writing. Nowadays, tourists can write feedback on social media platforms. Nowadays, tourists can write feedback on social media platforms about their experiences visiting any tourist destination. These are accessible to everyone to read and are considered *user-generated content* (UGC) [1]. Public opinions, regarded as the most authentic contents, are the best source of feedback that businesses can use to upgrade their services [2]. Its influence on the tourism industry is growing [3]. The tourism industry is an industry that relies on feedback from its visitors [4]. Tourism organizations could not

afford to ignore UGC because it could help Tourism Destination. Management (TDM) better understand their visitors' preferences [5] [6].

Analysing these reviews could reveal the underlying information from UGC that could help the TDM in their decision-making process. Instead of manually analyze the reviews, the feedback can be analysed using the sentiment analysis method with machine learning classifiers. Sentiment analysis could detect the sentiments behind the reviews and assign them into respective categories. The implementation of sentiment analysis can be done in many ways and one of them is extracting the polarity of reviews by using several machine learning classifiers [2] [7] [8].

Sentiment analysis with the machine learning method has been used in many different domains such as classifying the informal Malay textual data using Decision Tree (J48), Support Vector Machine and Naive Bayes to evaluate the consumer feedback and spot patterns in social commerce [9]. In another study, Twitter data related to two top apparel international brands was compared and analyzed using Naive Bayes and lexicon dictionary to get the public opinions of the two brands [2]. In paper [10], airline passenger reviews were classified into 5 categories: plane condition, flight comfort, staff service, food and entertainment and price using the Bayes and Support Vector Machine method to understand the satisfaction levels of the passenger for these categories. Moreover, people's perceptions regarding vaccines in Indonesia on Twitter were captured in the first two weeks to predict the people's sentiment about it by using Support Vector Machine and Random Forest [11]. Majority of the previous research on sentiment analysis uses Naïve Bayes, Support Vector Machine and Random Forest as their Machine Learning classifiers.

Tourism destination organizations can analyse the reviews left by their visitors on social media to come up with decision-making strategies to improve their tourism destinations. However, the feedback is hardly used because the reviews contain noise which affects classification data analysis tasks [12] [13]. Noisy content on social media, such as reviews that contain abbreviations, informal language, and emoticons, needs to be pre-processed. This paper will identify and apply the appropriate machine learning algorithms for sentiment analysis on social media for tourism analytics, and the results will be compared along with designing a sentiment analysis dashboard for visualizing the reviews. The best model with selected k-fold validation and n-gram was selected based on its performance accuracy. The main contributions of this paper are as follows:

- This paper proposes a sentiment modeling for tourism analytics from reviews.
- A dictionary-based approach and several different parameters of machine learning models are compared to achieve the best model.
- The findings are represented using Microsoft Power BI to understand the results further.

Section 2 of this paper presents the related studies while Section 3 explains the methods that were used in this study. Section 4 presents the results and findings, and this paper is concluded in Section 5.

## 2. Related Works

Tourism destinations are any geographic location that attracts and caters to tourists as a source of revenue and is located within administrative regions where tourist attractions, public facilities, accessibility, and communities are interconnected [4] [14]. It is a combination of a complex and integrated portfolio of services offered by a destination that supplies a holiday experience that meets the needs of tourists [15]. In Malaysia, Taman Negara has become a key ecotourism destination with its diverse flora and fauna [1] [16] [17]. The national parks are located across several states, the largest being Taman Negara Pahang, which spans three provinces and contains a total area of 4, 343 square kilometers [1] [16]. Taman Negara Pahang is 130 million years old and one of the oldest tropical rainforests in the world. It is on the national agenda to accomplish the Sustainable Development Goals.

## 2.1. Sentiment Analysis

The computational analysis of text to determine people's opinions, appraisals, attitudes, emotions, and sentiment polarity (positive, negative, or neutral) towards entities, situations, events, and topics is known as sentiment analysis or opinion mining [7] [18]. It is an excellent tool for businesses to analyse opinions expressed by users on social media without explicitly asking any questions, as this approach often reflects their genuine thoughts [18] [19]. The increasing number of daily active users on social media makes it a potential data source for sentiment analysis as it could observe people's behaviors and opinions in textual form [11]. Sentiment analysis is widely used in many domains such as education, healthcare, politics, e-commerce and many more [20] [21] [22] [23]. The sentiment analysis method is applied to one of the famous domains, tourist reviews, to understand tourists' experiences, opinions, and emotions towards a tourism destination [24] [25]. It is better than the traditional way of sending questionnaires to get the visitors' feedback [25] [26].

To get data for the sentiment analysis, the reviews are extracted by web scraping and transformed into a machine-readable format for use during the classification process [27] [28] [29]. Thus, it ensures only critical parts of the text are kept, avoiding noisy data, such as the computational time, and ensuring the accuracy of the classifiers [30] [31]. Tokenization is used to break text documents into meaningful elements, and tokens are used in the next stage [23] [25] [29] [32] [33] [34]. The word separator characters, such as spaces, are identified and used as delimiters for the tokens [23] [29]. N-gram tokenization, such as unigram, bigram, and trigram, is applied to tokenize the reviews into terms [18] [35] [36]. Any Stop Word (words that often appear in a text but have no meaning) is discarded, and the Case folding task converts all the upper-case letters in the text into lower cases to uniformize the shape of the letters [29] [35] [37]. The Word Normalization technique consists of Stemming and Lemmatization, which can be applied to eliminate the word's prefix and suffix based on the root word [23] [25] [29] [34] [38].

Meanwhile, Term weighting is used to find the repetitive terms in a corpus to determine their significance [29] [32] [39]. A term is a word or phrase in a document used to understand the context of the whole corpus that has a weight associated with it. A high-level weight is given to the term with the highest occurrences in a document [29]. It also represents the class of a corpus, obtained from the result of n-gram tokenization [35]. The Term Frequency-Inverse Document Frequency (TF-IDF) process finds the term with the highest frequency and weight. The process then compares other terms based on the occurrence of the words in each document multiplied by the word that provides common or rare information in the total number of documents [32] [35].

## 2.2. Classification Method

Classification is a supervised learning task that applies training and testing processes for learning from experiences. Sentiment analysis is an example of supervised learning that uses classification methods to produce the analysis output based on three categories, which are "positive", "negative", and "neutral". Several classifiers are used for this purpose, such as Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). NB classifier is a simple probabilistic machine learning classifier and can be used to solve classification problems by assuming that the attributes are independent of each other. It can also classify a review's emotions and polarity classes for text classifications [2] [12] [32] [34] [35] [40]. The SVM classifier can solve binary classification problems [40] [41] with a representation of feature vectors by using the Bag-of-Words model [10]. RF can be used for regression and classification tasks and is known as an ensemble technique because it combines multiple models to make predictions.

NB is one of the predictive classifiers used in this study. It is the easiest classifier to be implemented in model development as it has low computational time and can accept many types of data. As it is generally a simple classifier, parameter Laplace correction is applied to the classifier. NB is a classifier that uses probability for its algorithm. If there is a class label where the record does not appear in the training set, NB will set the conditional probability as 0. This probability is calculated with other labels,

which leads to a misleading result; the Laplace correction fixes this problem by changing the probability from 0 to 1. Naive Bayes outperforms other algorithms by having the highest accuracy [8].

The SVM operator used in this study is the Support Vector Machine (LibSVM) because it supports polynomial attributes, which are the sentiment attributes, unlike the normal SVM operator. A few parameters are to be tuned in the SVM classifier to increase its accuracy. One of the parameters tuned is the type of SVM, and for this study, C-SVC is used as it supports the classification tasks. The kernel used as the linear kernel is claimed a preferred kernel for text classification problems. The C parameter is a penalty parameter that represents misclassifications in the model. This parameter is tuned to 0, 1, and 2 throughout the project. The problem with this classifier is that it needs the data to be balanced for it to predict the classes accurately. SVM is suitable to be used in tourism domain as it shows a high accuracy of 80.11% by using tourism dataset [24].

The last classifier is RF, an ensemble of random trees specified by the number of trees. The problem with this classifier is that the computational power is high automatically stops when the computer has insufficient memory. The parameter tuning done during RF is by changing the number of trees to 10, 50, and 100 trees. This parameter decides how many random trees will be generated for RF. RF is implemented because, in Comparison to an RF, the limitation of single tree does not generate great performance. RF will create many classification trees that are aggregated to obtain optimum accuracy. The advantages of RF are it has higher accuracy than a single tree, is a powerful algorithm, and can handle a large dataset with a high dimension. When RF is combined with feature extraction process, RF outperformed other algorithms based on accuracy, precision, recall and F1-score [27]. Although it has many good points, it also has its disadvantages. It is expensive to train an RF and hard to interpret, and the user does not have complete control for the RF model. The comparison table in Table 1, shows the three main studies using sentiment analysis approach. Generally, these papers applied Naive Bayes and SVM with additional KNN, J48 and Random Forest, with different best result from each paper. Therefore, for this study, SVM and Random Forest algorithms were applied for the sentiment classification.

Table 1. Reviews Studies Comparison

Previous Study	Setiowati & Setyorini [33]	Rus et al. [29]	Wadhe & Suratkar [34]
Source of Data	Traveloka	Traveloka	Tourism Website
Classifiers	Naïve Bayes SVM K-Nearest Neighbors J48	Naïve Bayes SVM K-Nearest Neighbors	Naïve Bayes SVM Random Forest
Best Result	Rule-based Evaluation for service-based • Precision: 0.97 • Recall: 1 • F-measure: 0.98 For sentiment value-based: • Precision: 0.99 • Recall: 1 • F-measure: 0.99	The best result is SVM: • Precision: 81% • F1: 81%	The best result is Random Forest + TFIDF Vectorization • Precision: 85% • Recall: 86% • F-measure: 84%

### 3. Method

This study includes data collection, pre-processing, data labeling, text vectorization, and modeling, as shown in Fig. 1. Subsequently, the methods applied for each phase are explained in this section.

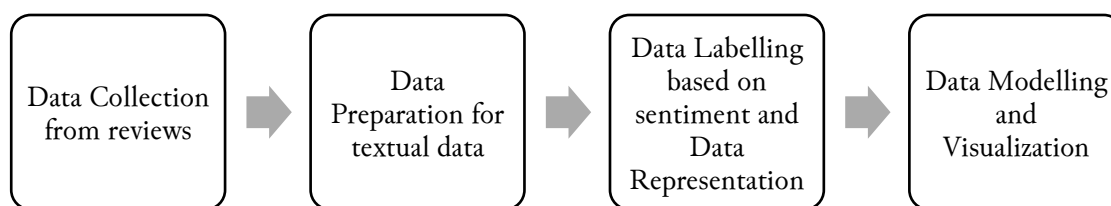


Fig. 1. Research Phases

### 3.1. Data Collection

The raw data is automatically collected from two UGC sources, TripAdvisor and Google Reviews, using Instant Data Scraper and SimpleScraper web scraping extensions. The keywords used to search on TripAdvisor are "Taman Negara", "National Park", "Endau Rompin", and "Gunung Ledang". The keywords used to search on Google Reviews are "Taman Negara", "Taman Negara Pulau Pinang", "Endau Rompin", "Taman Negara Canopy Walkway", "Mutiarataman Negara", and "Gunung Ledang". The total number of instances collected from TripAdvisor is 2,154. After analyzing the values inside the columns, the attribute columns obtained are:

- The user's URL
- Profile icon URL
- Name
- Number of times the user contributes to TripAdvisor
- Locations
- Review about the location
- The date of visit to location, and
- The date of writing the review

The three attributes of the collected data from Google Reviews:

- Name
- Date and review, and
- The rows that vary according to the location.

Since there are duplicates during the search process, the data collection is implemented on all sites and classified as the same location, even if they are from a different URL. The number of instances for all locations is 5,618. Total raw data collected on both websites are 7,772 instances, shown in Table 2.

Table 2. Reviews collected from TripAdvisor and Google Reviews

Location	Number of instances (TripAdvisor)	Number of instances (Google Reviews)
Taman Negara Pahang	336	1717
Taman Negara Pulau Pinang	815	1478
Taman Negara Endau-Rompin	27	580
Taman Negara Gunung Ledang	85	866
Mutiara Taman Negara	802	684
Taman Negara Canopy Walkway	89	293
	2,154	5,618

### 3.2. Data Pre-processing

The dataset for this project was pre-processed before modeling to ensure the optimum model performance. In this research, data pre-processing is done manually using Microsoft Excel and operators in RapidMiner. Noisy and missing values are removed as they have no importance and are irrelevant. A few pre-processing operators of RapidMiner extension are also used, such as Tokenize, Transform Cases, Filter Stopwords (English), Filter Tokens (by Length), Filter Stopwords (Dictionary), and Stem (Porter), as shown in Fig. 2. The study also tests Generate n-Grams (Terms) for the best combination of words and finds 2-gram is the best for application. For the Filter Tokens (by Length) operator, the characters are set to a minimum of two words and a maximum of 999 words in the parameter. This means it will remove words that are lesser than two characters and more than 999 characters, as they will have no meanings.

The text vectorization methods, Term Frequency (TF) and TF-IDF, are compared to determine the better accuracy method. The text vectorization method is chosen by selecting one with the highest accuracy for each classifier model. Out of the models with the highest accuracy, the method with a higher frequency is chosen for the last model. The accuracy displayed in the bar chart is the average accuracy obtained from 10 documents. The classifiers used in the models are NB, SVM, and RF, with cross-validation with 5-fold and 10-fold.

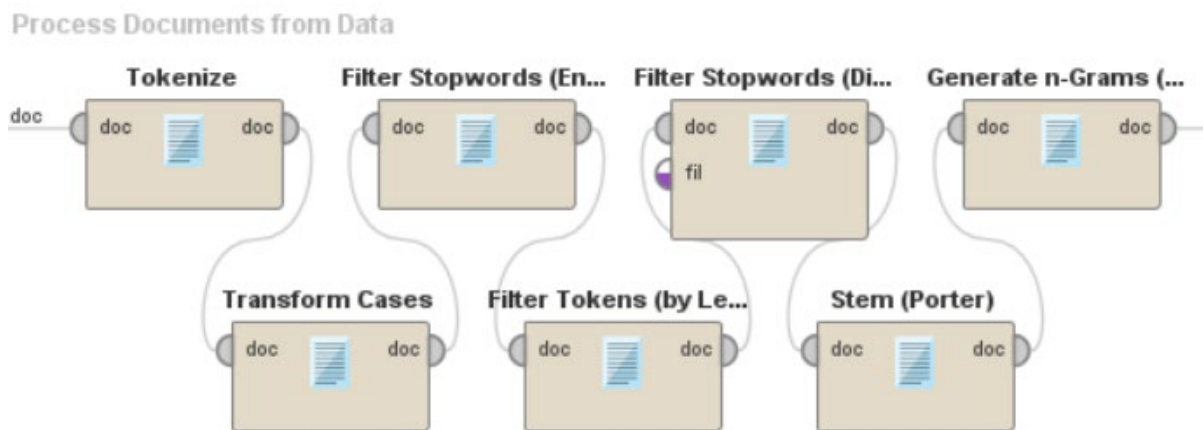


Fig. 2. Pre-processing Operators inside the Process Documents from Data Operator

### 3.3. Data Labelling and Text Vectorization for Data Representation

The review column is selected during the sentiment extraction process to generate the sentiment score used for sentiment classification. RapidMiner automatically calculates the score column for the sentiment polarity of a text, and the scoring string places the values on every word based on the availability of the words in the VADER sentiment dictionary. Any additional words are added to the Extracting Sentiment operator. The sentiment column, which is the class label, is obtained after the score column calculation using the rule set by the VADER documentation. If the score is higher than 0.05, the sentiment is considered positive. If the score is lower than -0.05, the sentiment is considered negative. The sentiment is considered neutral if the score is more than -0.05 and less than 0.05. Since the library's VADER score is already established, this study adopted the existing method. Further study will be planned to look into the labelling and extending the sentiment scoring method. The processed dataset has an unbalanced class label that labels 4,807 reviews as positive, 430 as neutral, and 560 as negative. The challenge encountered due to this condition is that any misclassified label may lead to inaccuracies. The minority label is treated as noise and disregarded by classifiers, such as SVM and RF. Thus, the class label is separated into ten new documents to overcome this problem. The sample operator is used on these documents to balance the data. Documents 1 to 9 each have 480 positive reviews, 43 neutrals, and 43 negatives. Document 10 has 487 positive reviews, 43 neutrals, and 56 negatives. Process Documents from Data (PDFP) operator can perform text vectorization. This research uses text vectorization schemas for this process, Term Frequency and TF-IDF, to determine the suitable method

for better accuracy. This study applies Term Frequency for the vectorization. However, the experiments done to achieve this will not be discussed in detail due to page limitations. Labeling the records shown in Fig.3.

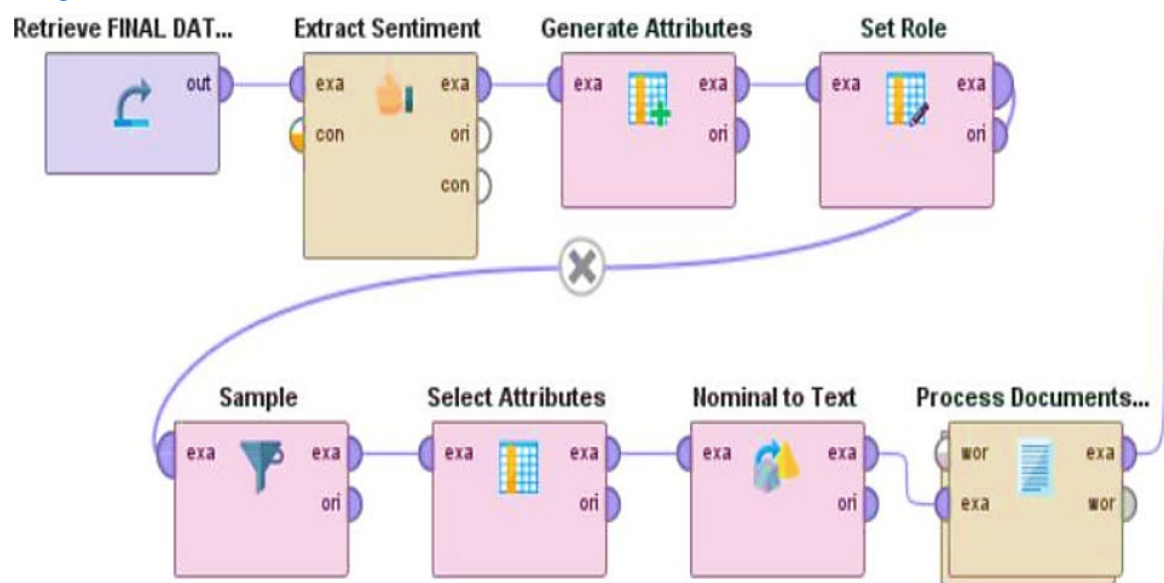


Fig. 3. Labeling the records using Sentiment score 'VADER' in RapidMiner

### 3.4. Modeling

Two supervised machine learning algorithms, Support Vector Machine (SVM) and Random Forest (RF), are applied. The main intention of this analysis is to determine the accuracy of each sentiment analysis technique to interpret the sentences from the customer reviews on Taman Negara Pahang. The model's accuracy could be derived from a confusion matrix that RapidMiner automatically generates. Other details such as the true positive and false positive could also be derived by it. Each classifier gives out different accuracy thus, comparisons between the model are done. The higher accuracy of the models indicates that the algorithm used for the model is suitable for the dataset and will generate a trustworthy result. The Comparison and analysis of the performance of these techniques in classifying this informal Malay textual data are modeled through 70% training and cross-validation. The k-fold cross-validation is used in this work; the 'k' number of groups from the dataset is split for testing and training. Since the dataset used in this project is split into ten different documents, the average accuracy for all ten is taken as a single accuracy. During the evaluation phase, the results and research findings were discussed according to the best performance of the supervised techniques, based on accuracy and further detailed Comparison was done based on True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) of the best classifier.

## 4. Results and Discussion

This section elaborates on the finding through SVM and RF classifiers for the sentiment classification of positive, neutral, and negative.

### 4.1. Modelling Experiments with SVM and RF Classifiers

SVM Classifier relies heavily on hyperparameters such as costing as  $c$  parameter. The  $c$  parameter affects the misclassification of the model, which is tuned in this experiment. The purpose of tuning is to select the best value of  $c$  for the final model. The  $c$  values that are changed in this experiment are 0, 1, and 2. Ten documents are tested in six types of models and the model accuracy is averaged. The highest accuracy is the SVM model that uses 1 as the  $c$  value; therefore value 1 is chosen and applied to the final SVM model. The confusion matrix for the table with the best result for SVM is shown in Table

3, Fig. 4 and Fig. 5, show that of two out of 3 classifiers, SVM and RF used in this experiment with higher accuracy when the TF method is used, Term Frequency is chosen for use in the last model.

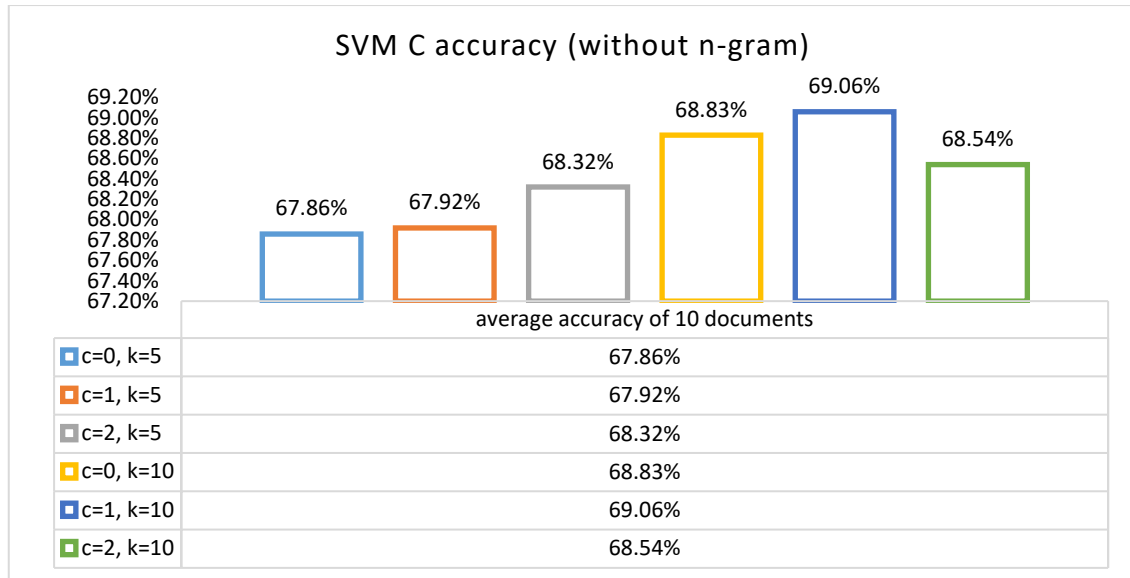


Fig. 4. Comparison of c Parameter for Support Vector Machine

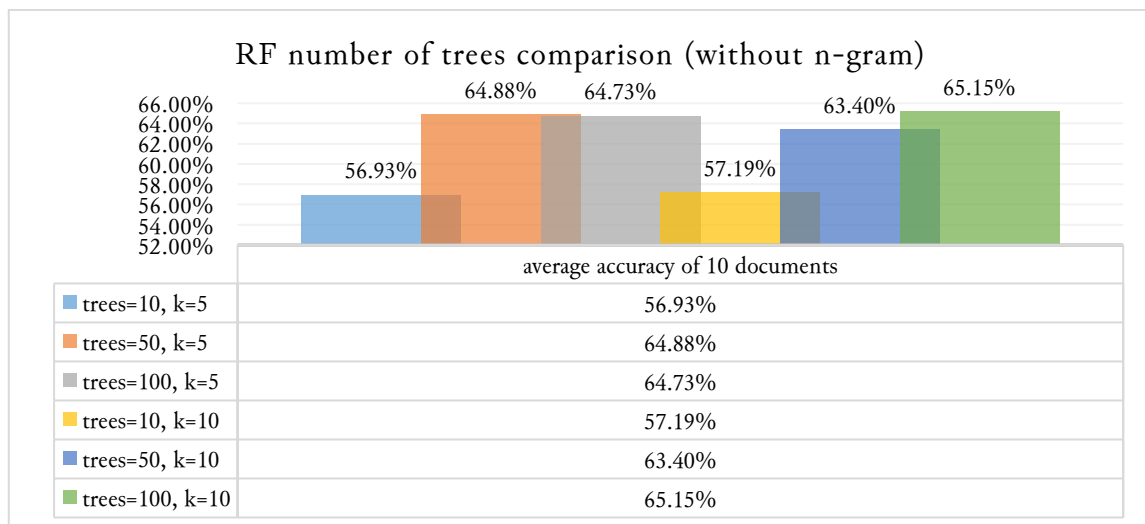


Fig. 5. Comparison of the Number of Trees for Random Forest in different k, number of folds; 5 and 10

Table 3. Reviews Studies Comparison

Document	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
1	89	49	227	49
2	89	52	236	52
3	92	40	224	40
4	87	51	225	51
5	92	46	230	46
6	91	47	229	47
7	90	45	225	45
8	113	16	242	16
9	97	47	241	47
10	111	36	258	36



The number of trees parameter measures the number of random trees generated for the RF model. The purpose of tuning the number of trees parameter in the RF classifier is to select the most optimum number of trees in the final model. The numbers of trees changed in this experiment are 10, 50, and 100, as shown in Fig. 4. The highest accuracy is recorded when the model uses 100; thus, 100 is chosen as the number of trees values for the last model.

#### 4.2. n-Gram Comparison

A sequence of words is created during the pre-processing phase using the n-gram operator. The type of n-gram is compared to measure which one generates better accuracy when applied during the pre-processing phase. The n-grams compared are 2-grams (bigrams) and 3-grams (trigrams). They are compared by changing the maximum length parameter at the Generate n-Grams (Terms) operator in RapidMiner. Fig. 5 illustrates four out of six classifier models recording high accuracy when 2-gram is applied. The models use SVM and RF as the classifier, and their accuracy ranges between 64% and 68%. This shows that 2-gram must be applied to the model for better accuracy.

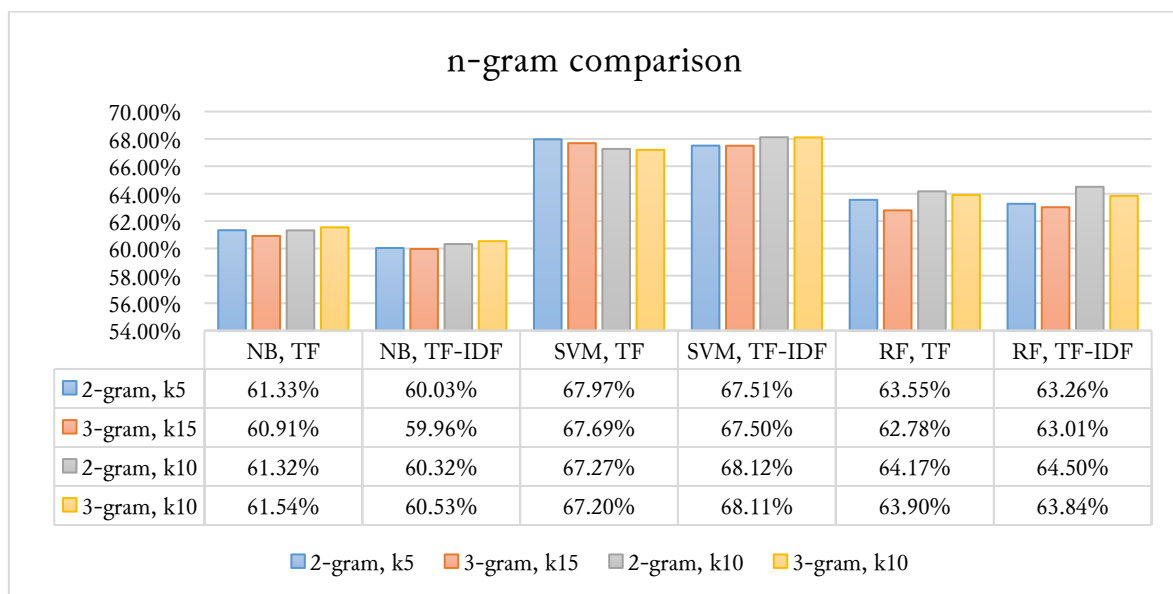


Fig. 6. n-Gram Comparison between Models

#### 4.3. Comparison of Results

The performance of the final model without n-gram in RF and SVM is compared based on 5-fold and 10-fold cross-validation. The model with 10-fold cross-validation appears to be the most accurate for classifying the sentiment of reviews for Taman Negara, with 69.06% using the SVM models and 65.15% for RF models. N-gram operator is added to the pre-processing phase and the model with the highest accuracy is chosen as the best model to be used with the Taman Negara dataset, with TF representation method. The 2-gram and 5-fold cross-validation model generally has higher accuracy than others. The RF model records an accuracy of 63.55% while the SVM model has an accuracy of 67.97%. In previous research, TF and n-gram are included in the pre-processing stage and SVM is the most efficient algorithm for classification modeling problem [11] [20] [21] [42]. These comparisons shows that the SVM algorithm is the most suitable to be used with the Taman Negara dataset in Comparison to RF. However, the limitation of this study is the processing machine, therefore, they need to be separated into 10 documents and compared the results.

#### 4.4. Dashboard

Fig. 7 illustrates the dashboard that visualizes the analytics features of the reviews written by Taman Negara visitors, such as the summary of the feedback, the word cloud, and the sentiments. It is a one-page dashboard with six charts, three slicers, and one card developed using Microsoft Power BI and deployed on the web. The positive sentiment score is 4,807 reviews, the neutral is 430 reviews, and the

negative is 560 reviews. These scores show that more tourists give significantly good reviews on Taman Negara from early 2020 till April 2020.

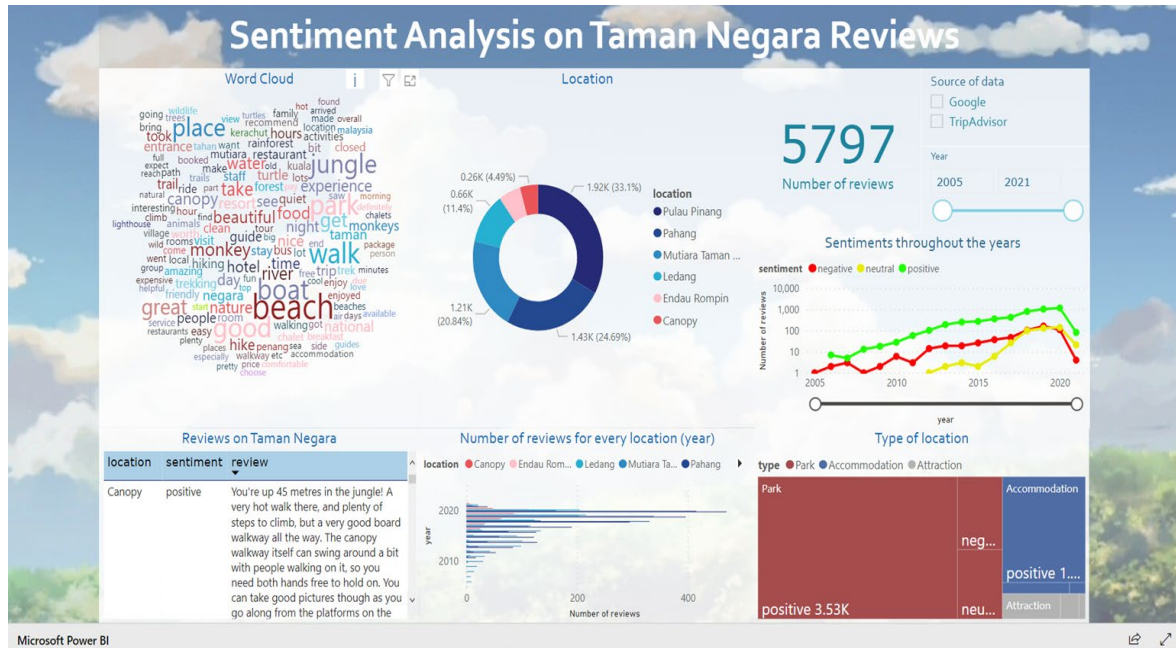


Fig. 7. Dashboard for Sentiment Analysis on Taman Negara Reviews

## 5. Conclusion

This study aims to evaluate the polarity of the tourism destination reviews by employing Support Vector Machine (SVM), and Random Forest (RF) classifiers and find a suitable classifier for the tourism dataset. The dataset used for this study is from the reviews of Taman Negara. The performances of these classifiers are compared to determine the most suitable classifier to be used in sentiment analysis for tourism reviews on social media. The SVM with a 5-fold cross-validation model is chosen as the most suitable classifier to be used with the tourism dataset. It outperforms other classifiers with a 67.97% accuracy, while RF with a 5-fold cross-validation has a 63.55% accuracy. The data visualization, a sentiment analysis dashboard, is produced to visualize the reviews from Taman Negara visitors. Among the limitations faced in this research are the inability to run several models due to inadequate computational power and a small scale number of records. Enhancements to be considered before continuing this project include hiring an expert human translator, running models on a machine with high computational power and collecting a larger dataset to avoid the issue of data imbalance. Improving the model by adding new classifiers are also recommended for future works.

## Acknowledgment

The authors would like to express the gratitude to Ministry of Higher Education, Malaysia for the research fund (FRGS/1/2021/ICT02/UITM/02/8) and School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support.

## Declarations

**Author contribution.** All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

**Funding statement.** This research is funded by the Ministry of Higher Education, Malaysia (FRGS/1/2021/ICT02/UITM/02/8).

**Conflict of interest.** The authors declare no conflict of interest.

**Additional information.** No additional information is available for this paper.

## References

- [1] S. Mutalib, A. H. Razali, S. N. K. Kamarudin, S. A. Halim, and S. Abdul-Rahman, "Prediction of Tourist Visit in Taman Negara Pahang, Malaysia using Regression Models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 12, pp. 746–754, 2021, doi: [10.14569/IJACSA.2021.0121292](https://doi.org/10.14569/IJACSA.2021.0121292).
- [2] A. Rasool, R. Tao, K. Marjan, and T. Naveed, "Twitter Sentiment Analysis: A Case Study for Apparel Brands," *J. Phys. Conf. Ser.*, vol. 1176, no. 2, p. 022015, Mar. 2019, doi: [10.1088/1742-6596/1176/2/022015](https://doi.org/10.1088/1742-6596/1176/2/022015).
- [3] "How User Generated Content (UGC) Has Transformed Travel," *Simple View*. Accessed Apr. 02, 2020. [Online]. Available at: <https://www.simplevieweurope.com/blog/read/2019/04/how-user-generated-content-ugc-has-transformed-travel-b106>.
- [4] R. S. Putra, R. Nurcahyo, and D. S. Gabriel, "Tourists Perception in Bali Using Social Media and Online Media Sentiment Analysis," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Dec. 2019, pp. 1–5, doi: [10.1109/ICETAS48360.2019.9117317](https://doi.org/10.1109/ICETAS48360.2019.9117317).
- [5] G. Akehurst, "User generated content: the use of blogs for tourism organisations and tourism consumers," *Serv. Bus.*, vol. 3, no. 1, pp. 51–61, Mar. 2009, doi: [10.1007/s11628-008-0054-2](https://doi.org/10.1007/s11628-008-0054-2).
- [6] N. Antonio, M. B. Correia, and F. P. Ribeiro, "Exploring User-Generated Content for Improving Destination Knowledge: The Case of Two World Heritage Cities," *Sustainability*, vol. 12, no. 22, p. 9654, Nov. 2020, doi: [10.3390/su12229654](https://doi.org/10.3390/su12229654).
- [7] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Context Enrichment Model Based Framework for Sentiment Analysis," in *Communications in Computer and Information Science*, vol. 1100, Springer, 2019, pp. 325–335, doi: [10.1007/978-981-15-0399-3\\_26](https://doi.org/10.1007/978-981-15-0399-3_26).
- [8] N. Sakinah Shaeali, A. Mohamed, and S. Mutalib, "Customer reviews analytics on food delivery services in social media: a review," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, p. 691, Dec. 2020, doi: [10.11591/ijai.v9.i4.pp691-699](https://doi.org/10.11591/ijai.v9.i4.pp691-699).
- [9] A. Nabihah, S. Mutalib, and A. M. A. Malik, "Sentiment Analysis for Informal Malay Text in Social Commerce," in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, Sep. 2021, pp. 1–6, doi: [10.1109/AiDAS53897.2021.9574436](https://doi.org/10.1109/AiDAS53897.2021.9574436).
- [10] A. Tegar Satria, Mustafid, and D. Mutiara Kusumo Nugraheni, "Implementation of Integrated Bayes Formula and Support Vector Machine for Analysing Airline's Passengers Review," in *E3S Web of Conferences*, Nov. 2020, vol. 202, p. 15004, doi: [10.1051/e3sconf/202020215004](https://doi.org/10.1051/e3sconf/202020215004).
- [11] I. M. Sumertajaya, Y. Angraini, J. R. Harahap, and A. Fitrianto, "Sentiment Analysis on Covid-19 Vaccination in Indonesia Using Support Vector Machine and Random Forest," *JUITA J. Inform.*, vol. 10, no. 1, p. 1, May 2022, doi: [10.30595/juita.v10i1.12394](https://doi.org/10.30595/juita.v10i1.12394).
- [12] M. Afzaal, M. Usman, and A. Fong, "Predictive aspect-based sentiment classification of online tourist reviews," *J. Inf. Sci.*, vol. 45, no. 3, pp. 341–363, Jun. 2019, doi: [10.1177/0165551518789872](https://doi.org/10.1177/0165551518789872).
- [13] S. Gupta and A. Gupta, "Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review," *Procedia Comput. Sci.*, vol. 161, pp. 466–474, Jan. 2019, doi: [10.1016/j.procs.2019.11.146](https://doi.org/10.1016/j.procs.2019.11.146).
- [14] N. Nasser, "Planning for Urban Heritage Places: Reconciling Conservation, Tourism, and Sustainable Development," *J. Plan. Lit.*, vol. 17, no. 4, pp. 467–479, May 2003, doi: [10.1177/0885412203017004001](https://doi.org/10.1177/0885412203017004001).
- [15] M. F. Cracolici and P. Nijkamp, "The attractiveness and competitiveness of tourist destinations: A study of Southern Italian regions," *Tour. Manag.*, vol. 30, no. 3, pp. 336–344, Jun. 2009, doi: [10.1016/j.tourman.2008.07.006](https://doi.org/10.1016/j.tourman.2008.07.006).
- [16] N. Backhaus, "Tourism and nature conservation in Malaysian National Parks," *Culture, Society, Environment - South Asian and South East Asian Studies*, 6, 2005. Accessed Apr. 02, 2020. [Online]. Available at : <https://www.zora.uzh.ch/id/eprint/2705/>.
- [17] A. Jambari *et al.*, "Quantifying species richness and composition of elusive rainforest mammals in Taman Negara National Park, Peninsular Malaysia," *Glob. Ecol. Conserv.*, vol. 18, p. e00607, Apr. 2019, doi: [10.1016/j.geco.2019.04.007](https://doi.org/10.1016/j.geco.2019.04.007).

[10.1016/j.gecco.2019.e00607](https://doi.org/10.1016/j.gecco.2019.e00607).

- [18] A. P. Kirilenko, S. O. Stepchenkova, H. Kim, and X. (Robert) Li, "Automated Sentiment Analysis in Tourism: Comparison of Approaches," *J. Travel Res.*, vol. 57, no. 8, pp. 1012–1025, Nov. 2018, doi: [10.1177/0047287517729757](https://doi.org/10.1177/0047287517729757).
- [19] G. Hu, P. Bhargava, S. Fuhrmann, S. Ellinger, and N. Spasojevic, "Analyzing Users' Sentiment Towards Popular Consumer Industries and Brands on Twitter," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov. 2017, vol. 2017–Novem, pp. 381–388, doi: [10.1109/ICDMW.2017.55](https://doi.org/10.1109/ICDMW.2017.55).
- [20] L. D. Utami and S. Masripah, "Comparison of classification algorithm on sentiment analysis of online learning reviews and distance education," *Techno Nusa Mandiri*, Sep. 15, 2021. Accessed Dec. 02, 2021. [Online]. Available at: <https://ejournal.nusamandiri.ac.id/index.php/techno/article/view/2715>.
- [21] I. Firmansyah, M. H. Asnawi, S. A. Hasanah, R. Novian, and A. A. Pravitasari, "A Comparison of Support Vector Machine and Naïve Bayes Classifier in Binary Sentiment Reviews for PeduliLindungi Application," in *2021 International Conference on Artificial Intelligence and Big Data Analytics*, Oct. 2021, pp. 140–145, doi: [10.1109/ICAIBDA53487.2021.9689771](https://doi.org/10.1109/ICAIBDA53487.2021.9689771).
- [22] R. Gull, U. Shoaib, S. Rasheed, W. Abid, and B. Zahoor, "Pre Processing of Twitter's Data for Opinion Mining in Political Context," in *Procedia Computer Science*, Jan. 2016, vol. 96, pp. 1560–1570, doi: [10.1016/j.procs.2016.08.203](https://doi.org/10.1016/j.procs.2016.08.203).
- [23] Stephenie, B. Warsito, and A. Prahutama, "Sentiment Analysis on Tokopedia Product Online Reviews Using Random Forest Method," *E3S Web of Conferences*, vol. 202. EDP Sciences, p. 16006, Nov. 10, 2020, doi: [10.1051/e3sconf/202020216006](https://doi.org/10.1051/e3sconf/202020216006).
- [24] S. Bruno, C. Yang, W. Tian, Z. Xie, and Y. Shao, "Exploring the characteristics of tourism industry by analyzing consumer review contents from social media: a case study of Bamako, Mali," *Geo-spatial Inf. Sci.*, vol. 22, no. 3, pp. 214–222, Jul. 2019, doi: [10.1080/10095020.2019.1649848](https://doi.org/10.1080/10095020.2019.1649848).
- [25] H. Irawan, G. Akmalia, and R. A. Masrury, "Mining Tourist's Perception toward Indonesia Tourism Destination Using Sentiment Analysis and Topic Modelling," in *Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things*, Sep. 2019, pp. 7–12, doi: [10.1145/3361821.3361829](https://doi.org/10.1145/3361821.3361829).
- [26] K. Kim, O. Park, S. Yun, and H. Yun, "What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management," in *Technological Forecasting and Social Change*, Oct. 2017, vol. 123, pp. 362–369, doi: [10.1016/j.techfore.2017.01.001](https://doi.org/10.1016/j.techfore.2017.01.001).
- [27] M. Afzaal, M. Usman, and A. Fong, "Tourism Mobile App With Aspect-Based Sentiment Classification Framework for Tourist Reviews," *IEEE Trans. Consum. Electron.*, vol. 65, no. 2, pp. 233–242, May 2019, doi: [10.1109/TCE.2019.2908944](https://doi.org/10.1109/TCE.2019.2908944).
- [28] D. T. Hermanto, M. Ziaurrahman, M. A. Bianto, and A. Setyanto, "Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier," *J. Phys. Conf. Ser.*, vol. 1140, no. 1, p. 012037, Dec. 2018, doi: [10.1088/1742-6596/1140/1/012037](https://doi.org/10.1088/1742-6596/1140/1/012037).
- [29] A. M. M. Rus, R. Annisa, I. Surjandari, and Zulkarnain, "Measuring Hotel Service Quality in Borobudur Temple Using Opinion Mining," in *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, Jul. 2019, pp. 1–5, doi: [10.1109/ICSSSM.2019.8887650](https://doi.org/10.1109/ICSSSM.2019.8887650).
- [30] W. Chen, Z. Xu, X. Zheng, Q. Yu, and Y. Luo, "Research on Sentiment Classification of Online Travel Review Text," *Appl. Sci.*, vol. 10, no. 15, p. 5275, Jul. 2020, doi: [10.3390/app10155275](https://doi.org/10.3390/app10155275).
- [31] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," in *Applied Soft Computing*, Jan. 2021, vol. 98, p. 106935, doi: [10.1016/j.asoc.2020.106935](https://doi.org/10.1016/j.asoc.2020.106935).
- [32] T. Kuhamane, N. Talmongkol, K. Chaisuriyakul, W. San-Um, N. Pongpisuttinun, and S. Pongyupinpanich, "Sentiment analysis of foreign tourists to Bangkok using data mining through online social network," in *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, Jul. 2017, pp. 1068–1073, doi: [10.1109/INDIN.2017.8104921](https://doi.org/10.1109/INDIN.2017.8104921).
- [33] Y. Setiowati and F. Setyorini, "Service Extraction and Sentiment Analysis to Indicate Hotel Service Quality

- in Yogyakarta based on User Opinion,” in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Nov. 2018, pp. 427–432, doi: [10.1109/ISRITI.2018.8864269](https://doi.org/10.1109/ISRITI.2018.8864269).
- [34] A. A. Wadhe and S. S. Suratkar, “Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques,” in *2020 International Conference on Industry 4.0 Technology (I4Tech)*, Feb. 2020, pp. 1–6, doi: [10.1109/I4Tech48345.2020.9102673](https://doi.org/10.1109/I4Tech48345.2020.9102673).
- [35] T. Widiyaningtyas, I. A. Elbaith Zaeni, and R. Al Farisi, “Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods,” in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5, doi: [10.1109/ICIC47613.2019.8985946](https://doi.org/10.1109/ICIC47613.2019.8985946).
- [36] P. H. D. Abd Samad, S. Mutalib, and S. Abdul-Rahman, “Analytics of stock market prices based on machine learning algorithms,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 2, p. 1050, Nov. 2019, doi: [10.11591/ijeecs.v16.i2.pp1050-1058](https://doi.org/10.11591/ijeecs.v16.i2.pp1050-1058).
- [37] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, “Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes,” in *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, Jul. 2019, pp. 49–54, doi: [10.1109/ICTS.2019.8850982](https://doi.org/10.1109/ICTS.2019.8850982).
- [38] “Stemming and Lemmatization in Python,” *Data Camp*. Accessed Apr. 02, 2021 [Online]. Available at: <https://www.datacamp.com/tutorial/stemming-lemmatization-python>.
- [39] Murni, T. Handhika, A. Fahrurozi, I. Sari, D. P. Lestari, and R. I. M. Zen, “Hybrid Method for Sentiment Analysis Using Homogeneous Ensemble Classifier,” in *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)*, Sep. 2019, pp. 232–236, doi: [10.1109/IC2IE47452.2019.8940896](https://doi.org/10.1109/IC2IE47452.2019.8940896).
- [40] A. P. Genoud, Y. Gao, G. M. Williams, and B. P. Thomas, “A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals,” in *Ecological Informatics*, Jul. 2020, vol. 58, p. 101090, doi: [10.1016/j.ecoinf.2020.101090](https://doi.org/10.1016/j.ecoinf.2020.101090).
- [41] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000, doi: [10.1017/CBO9780511801389](https://doi.org/10.1017/CBO9780511801389).
- [42] M. Sinoplu and E. B. Ceyhan, “Sentiment Analysis of Social Media Posts about Tourist Attractions: Black Sea Region Sample,” *Eur. J. Sci. Technol.*, vol. 36, no. 36, pp. 305–315, May 2022, doi: [10.31590/ejosat.1107640](https://doi.org/10.31590/ejosat.1107640).