IPTEK The Journal of Technology and Science, 31(2), 2020 (e/pISSN:2088-2033, 0853-4098)

DOI: 10.12962/j20882033.v31i2.6338

ORIGINAL RESEARCH

OPINION ANALYSIS OF TRAVELER BASED ON TOURISM SITE REVIEW USING SENTIMENT ANALYSIS

Siti Azza Amira[1] | Mohammad Isa Irawan*[2]

[1]Department of Technology Management, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

[2]Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Correspondence

*Mohammad Isa Irawan, Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Email: mii@its.ac.id

Present Address

MIPA Tower, Kampus ITS Sukolilo, Jl. Raya ITS, Surabaya 60111, Indonesia

Abstract

Technology development nowadays makes it easier for people to access information. One of them is to find information regarding a place. Many prospective visitors would read reviews from people who have visited a place to find out how they rate a place. Opinion on other people's reviews is very influential in influencing others' decisions in assessing a place they want to visit. Opinion analysis can be done by conducting a sentiment analysis of hotel customer reviews. The data used are traveler reviews of hotels in East Java on the Tripadvisor site. Traveler reviews data was taken by crawling on tourist sites, and the unstructured reviews data would be a preprocessing and weighted term from reviews using the TF-IDF method. The classification process is done using the support vector machine method to find opinions from traveler reviews, which are positive or negative. Based on the classification results, hotels that have the most positive sentiments in Surabaya are Harris Hotel Gubeng and Pop! Hotel Gubeng with the same number of reviews, 252 reviews. In comparison, hotels with the most positive sentiment in Malang are Harris Hotel Malang with 311 reviews. The opinion analysis results are expected to help the hotel manager evaluate and develop to increase the number of tourist visits.

KEYWORDS:

Classification, Opinion Analysis, Sentiment Analysis, Support Vector Machine, Tourism Site Review

1 | INTRODUCTION

Nowadays, internet development is very rapid. Internet technology has been used in many fields such as the media to get various kinds of information, media to run businesses (selling online or shopping online), media to get entertainment (music, movies, magazines), to access social media like Instagram, Twitter, and Facebook. Today many people use the internet to make it easier for them to get information about hotels. The increasing number of hotels that can be visited while on vacation would confuse people to determine which hotel is suitable for them. Many sites provide information about hotels in various countries. One of them is Tripadvisor. Based on data released by Skift in 2013 by looking at similar web statistics in seeing visitor traffic during

October 2013, TripAdvisor is ranked two of the ten online travel sites in the world that are most accessed and utilized by tourists with a total of 48.5 million visitors. On these sites, people can find a lot of information about hotels, restaurants, and tourist attractions in various countries and reviews from people who have visited the place. On TripAdvisor, people who have seen a hotel, restaurant, or tourist place can write their opinions about the site. Comments from other people are usually used as a reference for potential customers, whether they should visit that place or not. Opinion from other people is very influential in influencing others' decisions in assessing a place they want to see.

Indonesia is one of the famous countries for natural beauty and cultural beauty. Many tourist attractions in Indonesia that can be visited by local tourists and foreign tourists. One of the provinces in Indonesia that can be visited for a vacation in East Java. Many cities/regencies in East Java can be used as tourist destinations because of the lots of natural beauty such as beaches, mountains, and waterfalls widely available in several regions. Besides that, there are many tourist attractions, artificial tours, and cultural tours in East Java. Even according to the head of the East Java Culture and Tourism Office. In 2017 the number of local tourist visits to East Java was 65,623,535. This number increased by 13.01% from 2016, where the number of local tourist visits to East Java was 58,068,493. While the number of foreign tourist visits to East Java in 2017 as many as 690,509 increased by 11.62% compared to 2016, the number of foreign tourist visits was 618,651. Judging from many tourists who are interested in visiting East Java, it also opens opportunities for hotel managers to attract customers and make them stay at the hotel they manage. Thus, hotel managers should pay more attention to reviews of people who have visited the hotel because it can maintain and increase the number of visits.

Based on research from Chory et al.[1] with the research title "Sentiment Analysis on User Satisfaction Level of Mobile Data Services Using Support Vector Machine (SVM) Algorithm" and research from Ayu and Sarno[2] with the research title "Sentiment Detection of Comment Titles in Booking.com Using Probabilistic Latent Semantic Analysis" to analyze customer data reviews on the booking.com site using the PLSA method, it is necessary to analyze the comments of users to see how public satisfaction in using services.

In the review feature on various sites, they didn't show which reviews are positive (reviews with good opinions) only or negative (reviews with wrong opinions) only on websites, so sentiment analysis can be done to find out how traveler reviews on the sites. Traveler reviews on various sites are written based on what they feel when writing the reviews, so it can be said to be an honest review from the traveler. Sentiment analysis is a study of how to analyze opinions, sentiments, evaluations, judgments, attitudes, and emotions of an entity where they can be products, services, individuals, issues, events, organizations, and topics[3]. Sentiment analysis can be applied to several different types of levels, either it's text in the form of documents or sentences. Reviews from travelers are influenced by emotions (sentiments) to be classified, and that polarity can be determined positive or negative[4].

Many methods can be used to classify text. One method that can be used is Support Vector Machine (SVM). Support Vector Machine method is one of many methods that can be used to classify spatial data. This method has been widely applied to solve various kinds of problems in many fields, like gene expression analysis, finance, weather, and the medical field. In sentiment analysis, the support vector machine method's implementation is widely used because it can provide better results than similar classification methods such as Artificial Neural Network (ANN), especially in finding solutions. After all, SVM can find a globally optimal solution[5–8].

Based on the background described in this study, sentiment analysis was conducted to find out opinions from reviews written by hotel customers on TripAdvisor. Also, the review would be analyzed based on the aspects where the review can be categorized. The three aspects are used based on the TripAdvisor site: location, cleanliness, and service. The data used is review data from Tripadvisor obtained by crawling using the scrapy library in python. Term Frequency-Inverse Document Frequency (TF-IDF) method is used to find the weight values of a word and Support Vector Machine method to determine opinion sentiment

2 | LITERATURE REVIEW

2.1 | TripAdvisor

Tripadvisor is one of the largest travel sites in the world headquartered in Needham, Massachusetts. Figure 1 shows the website of Tripadvisor. This site provides tourist attractions, hotels, restaurants, and flights and can help tourists book their tours. This site offers a comment feature where tourists can provide reviews of tourist attractions, hotels, and restaurants that have been

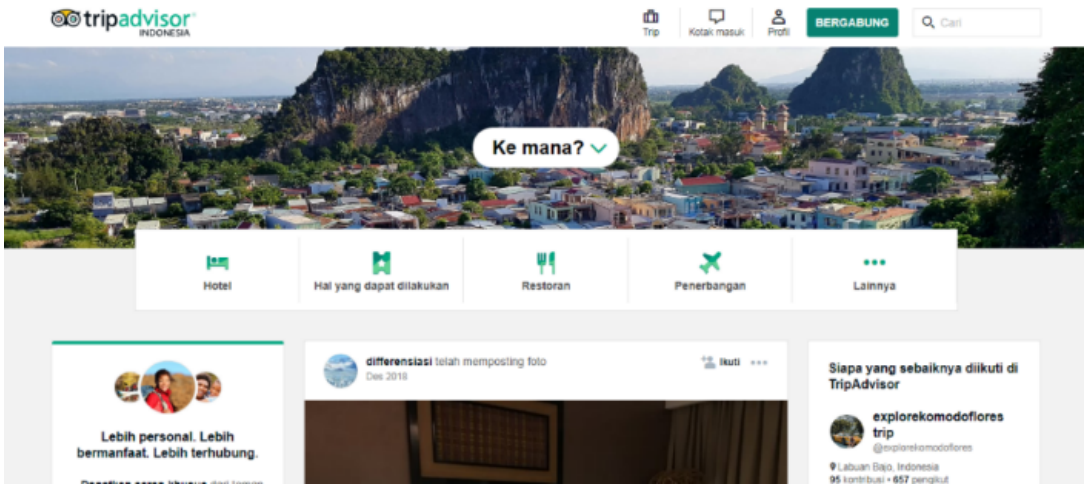

FIGURE 1 TripAdvisor website (www.tripadvisor.co.id).

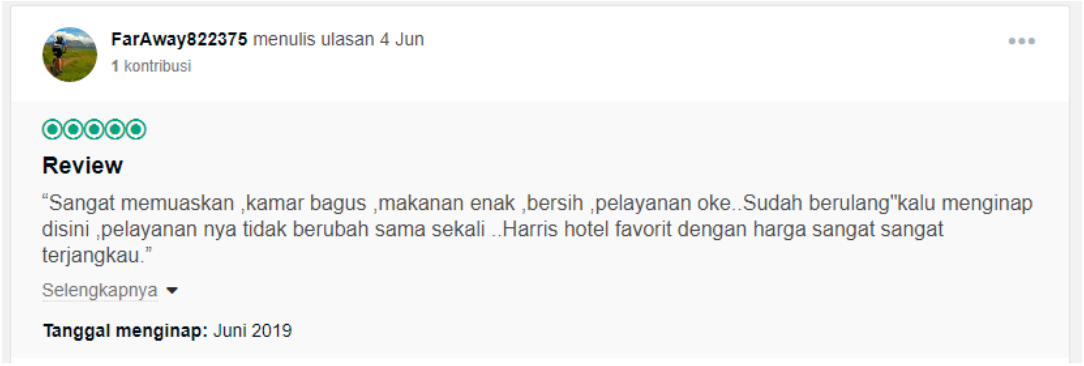

FIGURE 2 Review in TripAdvisor website.

visited to share experiences with various people worldwide, with 315 million active and inactive reviewers. Figure 2 shows a snapshot of the customer's review page. There is also a feature to compare flight prices, provide a link for tourists to book travel packages and find hotel prices

2.2 | Sentiment Analysis

Sentiment analysis is a branch of science from text mining, natural language programs, and artificial intelligence that learns how to analyze opinions, sentiments, evaluations, attitudes, judgments, and emotions of an entity that can be a product, service, organization, individual, issue network, events, and topics[9, 10]. Sentiment analysis is also called opinion mining, which is useful in managing natural languages, text mining, and computational linguistics and aims to determine opinions on a particular topic, where such behavior can indicate judgment and reasons and trends[11].

Sentiment analysis is mostly used to conduct analysis or to be able to assess public opinion, both opinions that refer to the likes or dislikes of goods or services. This sentiment means subjective information and has positive and negative polarity values where this polarity value can be used as a parameter to determine a decision[4].

2.3 | Text Minning

Text mining is a process to find patterns in the form of information or knowledge in a document or source that was previously not visible to become a desired pattern for a particular purpose[12]. Text mining is often used to help analyze information, assist the decision-making process, and manage information in the form of text in large numbers. The data would be processed with

various methods such as classification, clustering, sentiment analysis, etc. Text mining is in data mining but has different process stages and more stages than data mining. This is because of text mining processes data in the text whose characteristics are more complex than ordinary or structured data. Therefore, in text mining, several initial steps are needed to prepare to be changed to become structured[4].

2.4 | Text Pre-processing

In text mining, there is an initial stage before processing data, namely preprocessing. The Pre-processing process aims to process data that is initially still in the form of text that is changed first to fit the required format. In Text preprocessing several steps can be done, namely case folding, tokenizing, filtering, and stemming.

2.5 | TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a method used to calculate each word's weight that has been extracted. TF-IDF is used to calculate common words in information retrieval. In the TF-IDF weighting model, this method integrates the model term frequency (TF) and inverse document frequency (IDF), which is the term frequency (TF) process to calculate the number of occurrences of a word in a document/text and inverse document frequency (IDF) to calculate terms in various documents/texts which are considered as general terms and are considered not important[13–15].

In TF-IDF Methods, first, count the term frequency tft, d where t is the term in document d, which shows the term t in document d. This would affect the term's weight, which would be higher when many terms appear in one document[16]. The value of tf would be calculated using the weighting term frequency ($W_t f$), with the formula in Equation 1.

$$W_{tf_{t,d}} = \begin{cases} (1 + log_{10} tf_{t,d}, & \text{if } tf_{td} > 0 \\ 0, & \text{if } tf_{td} = 0) \end{cases} \tag{1}$$

Many words that appear in documents are generally the value of the term frequency of words that are not important. To avoid weighting on non-essential words, we use document frequency weighting to count the number of documents containing the term value. In a document, the emergence of a term that exists in most documents can result in a unique term search process interrupted. Inverse Document Frequency (IDF) is useful for reducing the weight of a term if the term's appearance is spread across all documents. The formula of the inverse document frequency is shown in Equation 2.

$$idf_t = log_{10} \frac{N}{df_t} \tag{2}$$

Furthermore, weighting TF-IDF is done by multiplying the results of document frequency with the inverse document frequency. The formula is shown in Equation 3.

$$W_{t,d} = W tf_{t,d} \times idf_t \tag{3}$$

2.6 | Support Vector Machine

Support Vector Machine is one method for classifying. The learning method used is guided. This algorithm was created by Vladimir Vapnik, which functions in analyzing data and recognizing patterns to get classification results. The algorithm support vector machine has a simple concept where the best hyperplane or boundary line is calculated, separating the two classes. Thus, the data enters into a category or another category[17, 18]. Examples of class separation in support vector machine are shown in Figure 3.

In Figure 3, it has been shown that there are two classes, namely +1, which shows positive classes, and -1, indicating negative classes. The hyperplane margin calculation is used to get the best hyperplane line and look for the maximum point. In dividing two classes with the best hyperplane, it can be defined in Equation (4).

$$(w \cdot x_i) + b = 0 \tag{4}$$

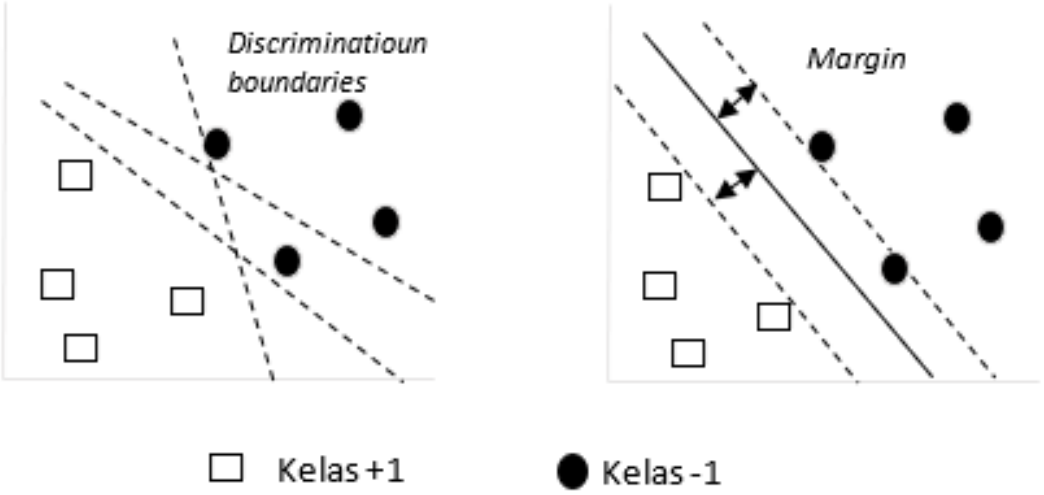

FIGURE 3 Hyperlane Support Vector Machine.

In pattern x_i, which is included in class -1, it can be formulated in Equation 5.

$$(w \cdot x_i + b) \leq 1, y_i = -1 \tag{5}$$

In pattern x_i, which is included in class +1, it can be formulated as in Equation 6.

$$w \cdot x_i + b \geq 1, y_i = 1 \tag{6}$$

Finding the largest margin value is done by maximizing the distance between the hyperplane and its closest point. This is obtained by formula:

$$\frac{1}{\|w\|} \tag{7}$$

The classification process problem is that most sample data are not linearly separated, so if a linear support vector machine is used, the results obtained are not optimal and result in poor classification results. Linear support vector machines can be changed to non-linear support vector machines using adding kernel functions. This method works by mapping input data to a higher dimensional feature space. It is expected that the input data from the mapping to the feature space would be linearly separated so that the optimal hyperplane can be searched[19]. Here are some of the frequently used kernel functions. Linear Kernel,

$$K(x_i, x_d) = X_i^T \cdot X_j \tag{8}$$

Polynomial Kernel,

$$K(x_i, x_d) = (X_i^T X_j + 1)^d, y > 0 \tag{9}$$

Radial Basis Function (RBF) Kernel,

$$K(x_i, x_d) = exp(-y\|X_i - X_j\|^2), y > 0 \tag{10}$$

Sigmoid Kernel,

$$K(x_i, x_d) = tanyX_i^T X_j + r \tag{11}$$

Three algorithms can be used in processing support vector machine training data, namely Quadratic Programming, Sequential Minimal Optimization, and Sequential Training. In its use, we must pay attention to the advantages and disadvantages of each algorithm. Quadratic Programming is a formulation process that can provide numerical analysis results with a complex algorithm and takes a long time. Sequential Minimal Optimization is the development of Quadratic Programming, where this algorithm can only provide small optimization. Whereas Sequential Training is a simple algorithm that doesn't take much time. Sequential training steps are as follows[20, 21]. First, we initialized against α_i=0 and other parameters, such as α, γ, C, and εvalues. The α_i = alpha i is used to searching support support-vector. The γ= Gamma constants to control speed. The C = variable slack. The ε= epsilon is used to find error values

Second, we calculated the Hessian matrix obtained from multiplication between the polynomial kernel and y, a vector worth 1 and -1. The calculation follows Equation 12.

$$D_{ij} = y_i y_j (K(x_i, x_j) + \lambda^2) \tag{12}$$

Third, we performed an iteration for each iteration initialized, then calculate the E_i value, which can be calculated using equation 13.

$$E_i = \Sigma_{j=1}^{n} \alpha_i D_{ij} \tag{13}$$

Fourth, we calculated $\delta\alpha_i$ values, which can be calculated using Equation 14.

$$\delta\alpha_i = min\ (max\ [\gamma(1 - E_i), -\alpha_i], C - \alpha_i) \tag{14}$$

Fifth, we updated the value of α_i using using Equation 15.

$$\alpha_{i+1} = \alpha_i + \delta\alpha_i \tag{15}$$

Return to the third step and do it repeatedly until you get the maximum iteration or $max(|\delta\alpha_i|) < \epsilon$. From the above process, we would get a support vector value (SV), where the value of SV=($\alpha > thresholdSV$). After that, the value of bias b is calculated by using Equation 16.

$$b = -\frac{1}{2}\left(\Sigma_{i=1}^{N} \alpha_i y_i K(x_i, x^-)\right) + \Sigma_{i=1}^{N} \alpha_i y_i K(x_i, x^+) \tag{16}$$

Sixth, we calculated the function f(x) to determine the classification results in a particular sentiment class with equation 17.

$$f(x) = \Sigma_{i=1}^{m\alpha_i} y_i K(x_i, x) + b \tag{17}$$

If the function is positive, then the document would be classified in the positive sentiment class. In contrast, if the function is negative, it would be classified in the negative sentiment class.

3 | MATERIAL AND METHOD

This study was carried out in four sequential processes. They are data collection, data pre-processing, feature extraction, and sentiment analysis. Figure 4 shows the processes.

3.1 | Literature Study

A literature study is done by looking for literature sources related to sentiment analysis, text mining, support vector machine method, and TF-IDF method to gain knowledge and support the research. The literature used can be sourced from books, journals, previous thesis.

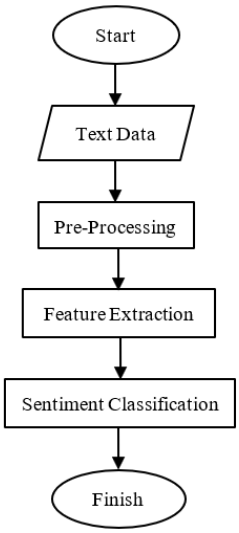

FIGURE 4 Research methodology.

FIGURE 5 Aspect category on TripAdvisor.

3.2 | Data Collection

At this stage, data would be collected to be used in conducting research. The data used is data related to the TripAdvisor site reviews with restrictions only on data hotel in East Java. The process of retrieving data using the scrapy library in python and reviews would be saved in a CSV file. The data obtained would be separated again based on the point, so when there are two sentences in a review, then the data would be counted as two documents. After that, the data that has been obtained would be divided into training data and data testing. The training data would be labeled, and the manual categorization of aspects would be doing. As shown in Figure 5, the category aspect is location, cleanliness, and service.

3.3 | Data Preprocessing

After crawling data in the form of reviews from people on the TripAdvisor site, the data would then proceed to the preprocessing stage. At this stage, there are several steps taken, namely case folding, tokenizing, filtering, and stemming.

Case folding. At this stage, the inconsistent reviews data that has been obtained would be changed into all lowercase letters, and characters other than letters would be removed, such as numbers and punctuation marks, so that the final result of the document/text is already small and all the punctuation marks are gone.

Tokenizing. At this stage, reviews that have become lowercase letters, and the marking is gone would be tokenized to be broken down into several parts to separate into words.

Filtering. The results from the tokenizing stage would be filtered at this stage, and words that are considered not important would be deleted.

Stemming. After filtering, the words would be trimmed if there is a prefix, suffix, or insertion to become a necessary word. the result from this stage would be carried out to the TF-IDF method for weighting before entering the classification stage.

3.4 | Feature Extraction

Furthermore, we can do the feature extraction process or say that "weighted term." At this stage, we would weigh the essential words obtained from the results of the preprocessing process. Feature extraction is carried out using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The weighting process is done by utilizing a module that is in python, which is by using the scikit-learn module.

TABLE 1 Review data from Tripadvisor.

Doc	Traveler Review
1.	Staff hotel ramah dan sigap membantu saat pertama kali tiba di hotel, Sarapan variatif dan rasanya enak , Harganya relatif lebih murah.
2.	Tapi akan lebih bagus lagi seandainya disekitar hotel ada minimarket sehingga tidak menyulitkan jika ingin belanja.
3.	Hotel bisnis yang sangat oke disini staffnya sangat ramah membantu dari mulai masuk sudah disambut makanan di restoran juga enak enak kamar nya nyaman.
4.	Pelayanan menyenangkan, Ketika masuk disambut dengan hangat dan langsung diberikan minuman.
5.	Mungkin perlu di tambah semacam minimarket di area restoran dan kolam renang supaya fasilitas lebih seru dan tidak membuat bosan.
6.	Lokasinya strategis di tengah kota.

TABLE 2 Aspect category labeling.

Doc	Traveler Review	Aspect
1.	Staff hotel ramah dan sigap membantu saat pertama kali tiba di hotel, Sarapan variatif dan rasanya enak , Harganya relatif lebih murah.	Service
2.	Tapi akan lebih bagus lagi seandainya disekitar hotel ada minimarket sehingga tidak menyulitkan jika ingin belanja.	Location
3.	Hotel bisnis yang sangat oke disini staffnya sangat ramah membantu dari mulai masuk sudah disambut makanan di restoran juga enak enak kamar nya nyaman	Service
4.	Pelayanan menyenangkan, Ketika masuk disambut dengan hangat dan langsung diberikan minuman.	Service
5.	Mungkin perlu di tambah semacam minimarket di area restoran dan kolam renang supaya fasilitas lebih seru dan tidak membuat bosan.	Service
6.	Lokasinya strategis di tengah kota.	Location

3.5 | Sentiment Classification

After the topic has been generated from each of the following reviews, each review's sentiment would be determined to find out the positive or negative opinions. The determination of sentiment in this study would be conducted using the support vector machine method.

The support vector machine method would find the best hyperplane or dividing line, which would divide the two classes and classify the data based on the closeness of the word on the side of the line.

4 | RESULTS AND DISCUSSION

4.1 | Data Collection

The review data in this study were obtained from crawling using scrapy library in python. The data taken is hotel customer reviews in East Java on the TripAdvisor site. The data then divided into training data and testing data. For training data, 300 reviews were used with 150 positive label reviews and 150 negative label reviews from various hotels in East Java. While for testing data using review data from 5 hotels in Surabaya and five hotels in Malang, the most popular was the rating on the TripAdvisor site. Examples of raw data crawling results from TripAdvisor sites can be seen in Table 1 .

After that, the data that has been obtained would be divided into training data and data testing. The training data would be labeled, and the manual categorization of aspects would be doing. Table 2 shows an example of labeling process's result.

4.2 | Data Preprocessing

After crawling data in the form of reviews from people on the TripAdvisor site, the data would then proceed to the preprocessing stage. At this stage, there are several steps taken, namely case folding, tokenizing, filtering, and stemming. The preprocessing phase is done by using the Natural Language Toolkit (NLTK) in python. The stages are documents whose sentences have been divided into each word and become basic words. The results at this stage would be used for the next process.

Case folding. In this process, we changed reviews into all lowercase letters and characters other than letters would be removed, such as numbers and punctuation marks. Table 3 shows the example of case folding on a list of reviews.

TABLE 3 Review data after case-folding process.

Doc	Review
1.	staff hotel ramah dan sigap membantu saat pertama kali tiba di hotel sarapan variatif dan rasanya enak harganya relatif lebih murah
2.	tapi akan lebih bagus lagi seandainya disekitar hotel ada minimarket sehingga tidak menyulitkan jika ingin belanja
3.	hotel bisnis yang sangat oke disini staffnya sangat ramah membantu dari mulai masuk sudah disambut makanan di restoran juga enak enak kamar nya nyaman
4.	pelayanan menyenangkan Ketika masuk disambut dengan hangat dan langsung diberikan minuman
5.	mungkin perlu di tambah semacam minimarket di area restoran dan kolam renang supaya fasilitas lebih seru dan tidak membuat bosan
6.	lokasinya strategis di tengah kota

TABLE 4 Review data after tokenizing process.

Doc	Review
1.	['staff', 'hotel', 'ramah', 'dan', 'sigap', 'membantu', 'saat', 'pertama', 'kali', 'tiba', 'di', 'hotel', 'sarapan', 'variatif', 'dan', 'rasanya', 'enak', 'harganya', 'relatif', 'lebih', 'murah']
2.	['tapi', 'akan', 'lebih', 'bagus', 'lagi', 'seandainya', 'disekitar', 'hotel', 'ada', 'minimarket', 'sehingga', 'tidak', 'menyulitkan', 'jika', 'ingin', 'belanja']
3.	['hotel', 'bisnis', 'yang', 'sangat', 'oke', 'disini', 'staffnya', 'sangat', 'ramah', 'membantu', 'dari', 'mulai', 'masuk', 'sudah', 'disambut', 'makanan', 'di', 'restoran', 'juga', 'enak', 'enak', 'kamar', 'nya', 'nyaman']
4.	['pelayanan', 'menyenangkan', 'Ketika', 'masuk', 'disambut', 'dengan', 'hangat', 'dan', 'langsung', 'diberikan', 'minuman']
5.	['mungkin', 'perlu', 'di', 'tambah', 'semacam', 'minimarket', 'di', 'area', 'restoran', 'dan', 'kolam', 'renang', 'supaya', 'fasilitas', 'lebih', 'seru', 'dan', 'tidak', 'membuat', 'bosan']
6.	['lokasinya', 'strategis', 'di', 'tengah', 'kota']

TABLE 5 Review data after filtering process.

Doc	Review
1.	['hotel', 'ramah', 'sigap', 'membantu', 'pertama', 'hotel', 'sarapan', 'variatif', 'rasanya', 'enak', 'harganya', 'relatif', 'murah']
2.	['lebih', 'bagus', 'hotel', 'minimarket', 'menyulitkan', 'belanja']
3.	['hotel', 'hotel', 'bisnis', 'ramah', 'membantu', 'makanan', 'restoran', 'enak', 'enak', 'kamar', 'nyaman']
4.	['pelayanan', 'menyenangkan', 'hangat', 'minuman']
5.	['tambah', 'semacam', 'minimarket', 'area', 'restoran', 'kolam', 'renang', 'fasilitas', 'seru', 'membuat', 'bosan']
6.	['lokasinya', 'strategis', 'tengah', 'kota']

Tokenizing. In this process, we broke data down into several parts so that it would separate into words. Table 4 shows the example of tokenizing on a list of reviews.

Filtering.. In this process we filtered the data and deleted insignificant words. Table 5 shows the example of filtering on a list of reviews.

Stemming.. In this process we trimmed the words if there is a prefix, suffix, or insertion. This process transforms each word into its basic form. Table 6 shows the example of stemming on a list of reviews.

4.3 | Feature Extraction

After Pre-processing, data would be carried out to the TF-IDF method for weighting before entering the classification stage. First, calculate the value of TF, DF, and IDF as shown in Table 7 . Furthermore, weighting TF-IDF is done by multiplying the results of document frequency with the inverse document frequency. Table 8 shows the results.

4.4 | Sentiment Classification

Table 9 shows the sentiment classification results of hotel customers in Surabaya and Malang. Based on Table 9 , hotels in Surabaya with the most positive sentiment are Harris Hotel Gubeng and Pop! Hotel Gubeng with the same number of reviews, 252 reviews. Then, the Malang hotels that have the most negative sentiment is Grand Darmo Suite with 46 reviews.

Based on Table 10 , hotels in Malang with the most positive sentiment are Harris Hotel Malang with 311 reviews. Then, hotels in Malang that have the most negative sentiment is Harris Hotel Malang with 79 reviews.

TABLE 6 Review data after stemming process.

Doc	Review
1.	['hotel', 'ramah', 'sigap', 'bantu', 'pertama', 'hotel', 'sarapan', 'variatif', 'rasa', 'enak', 'harga, 'relatif', 'murah']
2.	['lebih', 'bagus', 'hotel', 'minimarket', 'sulit', 'belanja']
3.	[''hotel', 'bisnis', 'ramah', 'bantu', 'makanan', 'restoran', 'enak', 'enak', 'kamar', 'nyaman']
4.	['layan', 'senang', 'hangat', 'minuman']
5.	['tambah', 'macam', 'minimarket', 'area', 'restoran', 'kolam', 'renang', 'fasilitas', 'seru', 'buat', 'bosan']
6.	['lokasi', 'strategis', 'tengah', 'kota']

TABLE 7 TF, DF, and IDF values.

Term	TF-IDF						DF	IDF
	D1	D2	D3	D4	D5	D6		
hotel	2	1	2	0	0	0	4	0,352
ramah	1	0	1	0	0	0	3	0,477
sigap	1	0	0	0	0	0	1	0,954
bantu	1	0	1	0	0	0	3	0,477
pertama	1	0	0	0	0	0	1	0,954
sarapan	1	0	0	0	0	0	1	0,954
variatif	1	0	0	0	0	0	1	0,954
rasa	1	0	0	0	0	0	1	0,954
enak	1	0	2	0	0	0	4	0,352
harga	1	0	0	0	0	0	1	0,954
relatif	1	0	0	0	0	0	1	0,954
murah	1	0	0	0	0	0	1	0,954
lebih	0	1	0	0	0	0	1	0,954
bagus	0	1	0	0	0	1	3	0,477
minimarket	0	1	0	0	1	0	2	0,653
…	…	…	…	…	…	…	…	…
Kota	0	0	0	0	0	1	2	0,653

TABLE 8 TF-IDF value.

Term	TF-IDF					
	D1	D2	D3	D4	D5	D6
hotel	0.704	0.352	0.704	0.000	0.000	0.000
ramah	0.477	0.000	0.477	0.000	0.000	0.000
sigap	0.954	0.000	0.000	0.000	0.000	0.000
bantu	0.477	0.000	0.477	0.000	0.000	0.000
pertama	0.954	0.000	0.000	0.000	0.000	0.000
sarapan	0.954	0.000	0.000	0.000	0.000	0.000
variatif	0.954	0.000	0.000	0.000	0.000	0.000
rasa	0.954	0.000	0.000	0.000	0.000	0.000
enak	0.352	0.000	0.704	0.000	0.000	0.000
harga	0.954	0.000	0.000	0.000	0.000	0.000
relatif	0.954	0.000	0.000	0.000	0.000	0.000
murah	0.954	0.000	0.000	0.000	0.000	0.000
lebih	0.000	0.954	0.000	0.000	0.000	0.000
bagus	0.000	0.477	0.000	0.000	0.000	0.477
minimarket	0.000	0.653	0.000	0.000	0.653	0.000
…	…	…	…	…	…	…
kota	0.000	0.000	0.000	0.000	0.000	0.653

TABLE 9 Sentiment classification of hotel reviews in Surabaya.

Hotel	Class	#Review
Grand Darmo Suite Surabaya	Positive	177
	Negative	46
Harris Hotel Gubeng Surabaya	Positive	252
	Negative	18
POP! Hotel Gubeng Surabaya	Positive	252
	Negative	22
Primebiz Hotel Surabaya	Positive	172
	Negative	38
Swiss-Bellin	Positive	227
Tunjungan	Negative	44
Surabaya		

TABLE 10 Sentiment classification of hotel reviews in Malang.

Hotel	Class	#Review
The 101 OJ Hotel Malang	Positive	190
	Negative	3
Harris Hotel Malang	Positive	311
	Negative	79
Kartanegara Premium Guest	Positive	157
House Malang	Negative	23
Hotel Santika Premiere Malang	Positive	160
	Negative	51
Hotel Tugu Malang	Positive	79
	Negative	31

4.5 | Sentiment Classification Based on Aspect

The next step is to analyze by looking at the sentiment results based on the reviews' aspect categories. Then categories of aspects that have the most positive sentiments and negative sentiments can be known. There are three categories of aspects used, location, cleanliness, and service.

Based on Table 11 we could see that at Grand Darmo Suite Surabaya, the aspect that has the most positive sentiment is the service aspect, with 89 reviews or 39.9% of the total reviews. At Harris Hotel Gubeng Surabaya, the aspect that has the most positive sentiment is the service aspect, with 133 reviews or 49.3% of the total reviews. At POP! Gubeng Hotel Surabaya, the aspect that gets the highest positive sentiment, is the service aspect with 135 reviews or 49.3% of the total reviews. Furthermore,at Primebiz Hotel Surabaya, the aspect that has the most positive sentiment is the service aspect, with 88 reviews or 41.9% of the total reviews. Finally, at Swiss-Bellin Tunjungan Surabaya, the aspect that has the most positive sentiment is the service aspect, with 120 reviews or 44.3% of the total reviews. Based on these results, the Service aspect is the most aspect reviewed by customers at five hotels in Surabaya. The service aspect is the most satisfying aspect of hotel customers in Surabaya.

TABLE 11 Sentiment classification of hotel reviews in Surabaya based on aspect.

Hotel	Aspect	Class	#Review	%
Grand Darmo Suite Surabaya	Location	Positive	49	22%
		Negative	13	5.8%
	Cleanliness	Positive	39	17.5%
		Negative	5	2.2%
	Service	Positive	89	39.9%
		Negative	28	12.6%
Harris Hotel Gubeng Surabaya	Location	Positive	49	18.1%
		Negative	4	1.5%
	Cleanliness	Positive	70	25.9%
		Negative	3	1.1%
	Service	Positive	133	49.3%
		Negative	11	4.1%
POP! Hotel Gubeng Surabaya	Location	Positive	38	13.9%
		Negative	4	1.5%
	Cleanliness	Positive	79	28.8%
		Negative	4	1.5%
	Service	Positive	135	49.3%
		Negative	14	5.1%
Primebiz Hotel Surabaya	Location	Positive	35	16.7%
		Negative	7	3.3%
	Cleanliness	Positive	49	23.3%
		Negative	12	5.7%
	Service	Positive	88	41.9%
		Negative	19	9%
Swiss-Bellin Tunjungan Surabaya	Location	Positive	68	25.1%
		Negative	9	3.3%
	Cleanliness	Positive	39	14.4%
		Negative	4	1.5%
	Service	Positive	120	44.3%
		Negative	31	11.4%

TABLE 12 Sentiment classification of hotel reviews in Malang based on aspect.

Hotel	Aspect	Class	#Review	%
The 101 OJ Hotel Malang	Location	Positive	21	10.9%
		Negative	0	0%
	Cleanliness	Positive	55	28.5%
		Negative	0	0%
	Service	Positive	114	59.1%
		Negative	3	1.6%
Harris Hotel Malang	Location	Positive	50	12.8%
		Negative	27	6.9%
	Cleanliness	Positive	77	19.7%
		Negative	15	3.8%
	Service	Positive	184	47.2%
		Negative	37	9.5%
Kartanegara Premium Guest House Malang	Location	Positive	45	25%
		Negative	2	1.1%
	Cleanliness	Positive	49	27.2%
		Negative	3	1.7%
	Service	Positive	63	35%
		Negative	18	10%
Santika Premiere Hotel Malang	Location	Positive	30	14.2%
		Negative	10	4.7%
	Cleanliness	Positive	28	13.3%
		Negative	3	1.4%
	Service	Positive	102	48.3%
		Negative	38	18%
Tugu Hotel Malang	Location	Positive	22	20%
		Negative	2	1.8%
	Cleanliness	Positive	13	11.8%
		Negative	0	0%
	Service	Positive	44	40%
		Negative	29	26.4%

Based on Table 12 , we could see that at the 101 OJ Hotel Malang, the spect that has the most positive sentiment, is service aspect with 114 reviews or 59,1% of the total reviews. At Harris Hotel Malang, the aspect that has the most positive sentiment is the service aspect, with 184 reviews or 47,2% of the total reviews. At Kartanegara Premium Guest House Malang, the Aspect that has the most positive sentiment is the service aspect, with 63 reviews or 35% of the total reviews. At Santika Premiere Hotel Malang, the aspect that has the most positive sentiment is the service aspect, with 102 reviews or 48,3% of the total reviews. Finally, at Tugu Hotel Malang, the aspect that has the most positive sentiment is the service aspect, with 44 reviews or 40% of the total reviews. Based on these results, the Service aspect is the most aspect reviewed by customers at five hotels in Malang. The service aspect is the most satisfying aspect of hotel customers in Malang.

4.6 | Evaluation

The success rate of the classifier can be measured by evaluating. In general, a confusion matrix is used to measure evaluation and done by measuring the value of accuracy, precision, recall, and F1-Score. Table 13 shows the result of the measurements.

TABLE 13 Accuracy, precission, recall, and F1-score of data in Surabaya.

Hotel	Acc.	Precision	Recall	F1-Score
Grand Darmo Suite	88%	98%	88%	98%
Harris Hotel	93%	95%	98%	97%
POP! Hotel	94%	96%	97%	97%
Primebiz Hotel	90%	95%	94%	94%
Swiss-Bellin	89%	96%	90%	93%

TABLE 14 Accuracy, precision, recall, and F1-score of hotel in Malang.

Hotel	Acc.	Precision	Recall	F1-Score
Grand Darmo Suite	99%	99%	100%	99%
Harris Hotel	87%	90%	93%	92%
POP! Hotel	92%	92%	97%	95%
Primebiz Hotel	80%	98%	80%	83%
Swiss-Bellin	83%	98%	81%	89%

5 | CONCLUSION

The support vector machine method combined with TF-IDF can solve problems in sentiment classification. This is evidenced by the ability of the TF-IDF method to give a weight value to a word and the ability of the Support vector machine method to provide labels in each review, which are positive reviews and negative reviews. The service aspect is the most aspect reviewed by hotel customers in Surabaya and Malang. The service aspect is the most satisfying aspect of hotel customers both in Surabaya and Malang.

On the accuracy test, the highest accuracy value on hotel review data in Surabaya is POP! Gubeng Hotel Surabaya with an accuracy rate of 94%, while the highest accuracy value on hotel reviews data in Malang is The 101 OJ Malang Hotel, with an accuracy rate of 99%. With this value of accuracy, it means the classifier used has worked well in classifying reviews. Sentiment Analysis of hotel reviews can be done to find out how the opinions of hotel customers. It can contribute to hotel managers in improving service to increase the number of visits.

References

1. Chory RN, Nasrun M, Setianingsih C. Sentiment Analysis on User Satisfaction level of Mobile Data Services Using Support Vector Machine (SVM) Algorithm. In: IEEE International Conference on Internet of Things and Intelligence System (IOTAIS) Institute of Electrical and Electronics Engineers Inc.; 2018. p. 194–200.
2. Khotimah DAK, Sarno R. Sentiment Detection of Comment Titles in Booking.com Using Probabilistic Latent Semantic Analysis. In: Proceeding on 6th International Conference on Information and Communication Technology (ICoICT) Bandung, Indonesia: Institute of Electrical and Electronics Engineers Inc.; 2018. p. 514–519.
3. Liu B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 2012;5(1):1–167.
4. Indriati I, Ridok A. Sentiment Analysis For Review Mobile Application Using Neighbor Method Weighted K-Nearest Neighbor (NWKNN). Journal of Environmental Engineering and Sustainable Technology 2016;03(01):23–32.
5. Peng JX, Ferguson S, Rafferty K, Stewart V. A sequential algorithm for sparse support vector classifiers. Pattern Recognition 2013;46(4):1195–1208.
6. Nazemi A, Dehghan M. A neural network method for solving support vector classification problems. Neurocomputing 2015;152(1):369–376.
7. Javidi MM, Zarisfi Kermani F. Utilizing the advantages of both global and local search strategies for finding a small subset of features in a two-stage method. Applied Intelligence 2018;48(10):3502–3522.
8. Zuo W, Wu X, Lin L, Zhang L, Yang MH. Learning support correlation filters for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 2019;41(5):1158–1172.
9. Liu B. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing 2010;2:627–666.
10. Setiowati Y, Djunaidy A, Siahaan DO. Pair Extraction of Aspect and Implicit Opinion Word based on its Co-occurrence in Corpus of Bahasa Indonesia. In: 2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019 Institute of Electrical and Electronics Engineers Inc.; 2019. p. 73–78.
11. Basari ASH, Hussin B, Ananta IGP, Zeniarja J. Opinion Mining of Movie Review Using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering 2013;53:453–462. http://dx.doi.org/10.1016/j.proeng.2013.02.059.
12. Mustafa A, Akbar A, Sultan A. Knowledge Discovery using Text Mining : A Programmable Implementation on Information Extraction and Categorization. International Journal of Multimedia and Ubiquitos Engineering 2009;4(2):183–188.
13. Vapnik V, Izmailov R. Rethinking statistical learning theory: learning using statistical invariants. Machine Learning 2019 mar;108(3):381–423.

14. Rameshbhai CJ, Paulose J. Opinion mining on newspaper headlines using SVM and NLP. International Journal of Electrical and Computer Engineering 2019;9(3):2152–2163.

15. Brown AD, Kachura JR. Natural Language Processing of Radiology Reports in Patients With Hepatocellular Carcinoma to Predict Radiology Resource Utilization. Journal of the American College of Radiology 2019;16(6):840–844.

16. Luqyana WA, Imam Cholissodin, Perdana RS. Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN 2018;2(11):4704–4713.

17. Vapnik V, Izmailov R. Knowledge transfer in SVM and neural networks. Annals of Mathematics and Artificial Intelligence 2017;81(1):3–19.

18. Karim R, Kundu AK. Computational analysis to reduce classification cost keeping high accuracy of the sparser LPSVM. International Journal of Machine Learning and Computing 2019;9(6):728–733.

19. Shandra EN, Setiawan BD, Sari YA. Klasifikasi Pola Sidik Bibir untuk Menentukan Jenis Kelamin Manusia dengan Metode Gray Level Co-Occurrence Matrix dan Support Vector Machine. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 2019;3(3):2753–2760. http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4786.

20. Pereira DR, Pisani RJ, De Souza AN, Papa JP. An Ensemble-Based Stacked Sequential Learning Algorithm for Remote Sensing Imagery Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2017;10(4):1525–1541.

21. Nóbrega JP, Oliveira ALI. A sequential learning method with Kalman filter and extreme learning machine for regression and time series forecasting. Neurocomputing 2019;337(1):235–250.

How to cite this article: Amira S.A., Irawan M.I., (2020), Opinion Analysis of Traveler Based on Tourism Site Review Using Sentiment Analysis, *IPTEK The Journal of Technology and Science*, *31(2):223–235*.