Mapping problem using text mining to boost tourism industry: is it possible?

M Tamrin\textsuperscript{1} and L Septianasari\textsuperscript{2}

\textsuperscript{1} Department of Computer Science and Econometrics, Warsaw University of Life Science, Poland
\textsuperscript{2} Department of English Education, STKIP Muhammadiyah Bogor, Indonesia

Abstract. TripAdvisor has become a credential traveling platform for tourists worldwide to set travel plans. The widespread of big data in online platforms urges the use of text mining to benefit some sectors, including in the tourism industry. This study aimed to investigate the information extraction based on the online reviews on TripAdvisor for Gili Trawangan tourist destinations. The method used in this research was text mining with Support Vector Machine (SVM) to classify the online reviews that categorized into two classes, positive class and negative class. The results of information extraction show that the issue of horse cruelty, bad waste management, and ecosystem vulnerability dominated the negative sentiments. These negative sentiments need to be handled professionally by the tourism enterprise to boost the tourism industry in Gili Trawangan.

1. Introduction

Digital platforms have contributed to the growth of large scale data in both structured and unstructured data\cite{1}. The extensive distribution of information increases high interest in exploring and handling textual unstructured data with using new methods, like text mining, corpus-based computational linguistics, sentiment analysis, and others, which are very useful to gain new insights in managing business policymaking\cite{2, 3, 4, 5}. Virtual data grow rapidly during the digital era, so these unstructured data can be found on various platforms, such as social media, webpages, and electronic documents. Thus, the information spreading in large corpora could be extracted to be benefited by the enterprise or policymaker.

A website like TripAdvisor has an important role in collecting data which are related to tourism issue \cite{6}. Furthermore, these data can be used to increase tourism marketing\cite{7} since several tourists left digital footprints such as reviews and ratings to express their personal opinions about a place they had visited \cite{8, 9}. One of the strategic tourist destinations in Indonesia that have been visited by many domestic and foreign tourists is Gili Trawangan. The visit of foreign tourists to Gili Trawangan has been in the range of 2,600 to 3,000 people per day \cite{10, 11}. In North Lombok, tourism becomes a leading economic sector which accounts for 60\% of regional income\cite{11}. Gili Trawangan is one of three small island groups located beside Gili Air and Gili Meno in North Lombok, West Nusa Tenggara Province. The three small islands have tourism potential but the most famous of which is Gili Trawangan. This island has complete accommodation and facilities, such as restaurants, hotels, transportation, banks, ports, and many others. Due to its beauty, this island has a high potential to be the icon of tourism site in Indonesia, especially in West Nusa Tenggara. Hence, the reviews written by tourists about Gili Trawangan would help to improve the tourism development of this island to be more favorable site to visit.
Many studies discussed the role of online reviews on tourism sectors as to how the extracted information helps to boost the economic growth [12, 13, 14, 15]. Gretzel and Yoo investigated the use and impact of reviews given by the users of the traveling platform [16]. They argued that the reviews informed the other travelers about the accommodation of the places they wanted to visit so that they could manage their travel planning [16]. In addition, Dediu mentions that reviews in the online platform give significant impacts on the process of decision-making due to its content contribution [17]. However, the objectivity of the reviewers needs to be testified to avoid bias. On the other hand, Hlee, Lee, and Koo believed that online reviews had managerial implications, especially in hospitality and tourism [18]. The information which is available in online reviews can be used to support the marketing strategy of tourist enterprises. Moreover, after identifying 50 articles under the topic of online reviews in tourism and hospitality, Schuckert, Liu, and Law concluded that online reviews were related to online buying as to how satisfaction correlated with online management [19]. Their paper implied that online reviews in hospitality and tourism could be used to forecast the current or future trend in those fields and the extracted information of the online reviews could be a direction to make a decision. This argument was supported by Li et.al. [20] and Colladon et.al. [21] who argued that online reviews could be analyzed to forecast tourism demand.

Given the discussion of those previous studies, research in online reviews of tourism destination in Gili Trawangan is still limited. Meanwhile, the number of visitors in Gili Trawangan continues to increase from time to time. Thus, the digital footprints left by the tourists need to be analyzed to elevate the service, hospitality and tourism industry in Gili Trawangan. This manuscript, therefore, was designed to fill the gap of this issue. This research aimed to investigate the information extraction of online reviews in TripAdvisor, particularly the reviews written by foreigners in the case of tourism destination in Gili Trawangan. The investigation was focused on the sentiment analysis to what extent the online reviews of Gili Trawangan were addressed.

2. Research Method

The research method used in this project is text mining with Support Vector Machine (SVM) as a classification method and a supervised machine learning approach as the aids in extracting information from a text corpus. Text mining was used to collect the data from foreign tourists’ footprints on TripAdvisor. Text mining is a process of extracting the text automatically and semi-automatically with the aim of finding new information from a collection of unstructured data text [22, 23]. The process of text mining in the unstructured text requires several initial stages with the aim of preparing the transformation of an unstructured database to be structured database after being supervised by using certain rules [24]. R software was used to mining the corpus in which 298,736 tokens were scraped from TripAdvisor. After getting text data corpus, the next steps of text mining were administered to analyze the data. Those are preprocessing, representation, and knowledge discovery. The whole steps in text mining were illustrated below:

![Figure 1. Traditional Framework of Text Mining][25]

The figure above illustrates the process of data collecting and data analysis. Firstly, the text corpus was scrapped from TripAdvisor’s reviews. Secondly, we used stop word removal, spelling correction and stemming to cleaning the data during tokenization in preprocessing. Thirdly, Term Frequency-Inverse Document Frequency (TF-IDF) was carried out to convert data representation. Term frequency (TF) is defined as the number of occurrences of a word/term in a document in order to reduce the effect of
of words that often appear in the corpus and IDF gives higher weight according to the frequency of words appearing in the document [26, 23].Hu and Liu [25] introduced the most frequently used formula of TF-IDF as follows:

$$\text{tfidf}(w) = tf \times \log \frac{N}{df(w)} \quad (1)$$

Where:

- $tf$: term frequency
- $df(w)$: document frequency
- $N$: number of document in Corpus
- $\text{tfidf}(w)$: relative weight of the feature in the vector

This formula further was used during the representation stage. After doing TF-IDF procedure, the next procedural stage was Singular Value Decomposition (SVD). It aims to reduce the matrix dimension by finding the hidden structure and correlation of the matrix. Lastly, knowledge discovery was acquired by using Support Vector Machine (SVM) method. This last stage aimed to extract the information from online review data. SVM is a machine learning method that works on the principle of Structural Risk Minimization (SRM) by finding the best hyper-plants that functions as a separator of two classes on input space [27]. In addition, Christianini and Shawe-Taylor claimed that SVM could be applied to the classification of handwriting detection, object recognition, voice identification and others [27]. Thus, SVM was used in this study to discover knowledge or information extracted from online reviews.

3. Results and Discussion

This section elaborates on the result of data mining from text corpus and the stage in analyzing and extracting the data to be information. The details of result and discussion were elaborated in the following sub-sections.

3.1. Results

The corpus data in this research were scrapped from 2012 to 2020 and limited to only TripAdvisor’s reviews on Gili Trawangan tourist destinations. All reviews written in a language other than English were not included since we want to focus on the perspective of International visitors about their impression of Gili Trawangan during their visitation. Their review could be a benchmark of how well-prepared tourist destinations in Gili Trawangan to become one of the favorite international tourist destinations. In mining the text data, 1.290 reviews were collected to become the text corpus which then would be processed into the preprocessing phase and other stages.

3.1.1. Sentiment Class Labelling

After finishing the Preprocessing phase, then sentiment analysis was carried out for data labeling. The data labeling process was done automatically by calculating sentiment scores by using the lexicon dictionary. In general, sentiment analysis was used to classify textual documents into three classes of sentiments, namely positive, negative and neutral sentiments. The way to determine sentiment class is to count the number of positive words minus the number of negative words in each review or sentence. The sentences that have a score of more than zero ($s > 0$) will be classified into positive classes. On the other hand, the sentences that have zero score ($s = 0$) will be classified into neutral classes as well as sentences that have a score less than zero ($s < 0$) will be classified into negative classes. At this stage, R software was used to calculate the data. In addition, the sentiment score calculation formula used in the labeling process is as follows:

$$\text{score} = \text{the number of positive words} - \text{the number of negative words} \quad (2)$$
The results of labeling in the sentiment class were 726 negative sentiments and 2293 positive sentiments. The results of class labeling showed that the frequency of positive reviews was far more than negative ones. The manual process of sentiment class labeling based on lexicon dictionary can be seen on the table below:

| Classification | Score | Text |
|----------------|-------|------|
| Positive       | 4     | beautiful trawangan beautiful trawangan visit trawangan clean beach nice spot enjoyed visit trawangan circumnavigating walk snorkeling trip cool prepared higher prices compared mainland rubbish problem |
| Negative       | -1    | touristy deserted trawangan not hugely popular can crowded enjoyed visit trawangan circumnavigating walk snorkeling trip cool prepared higher prices compared mainland rubbish problem |

It can be seen in Table 1 that 4 positive words are detected in positive classification. Those words are <clean>, <nice>, and <beautiful> which appeared twice, therefore, the score for positive classification is 4. Besides, the score calculation of negative classification can be seen in Table 2.

| Text reviews                  | Positive words | Negative words |
|-------------------------------|----------------|----------------|
| touristy                      | not            |                |
| deserted trawangan            |                |                |
| hugely                         | popular        |                |
| can                            | crowded        |                |
| enjoyed                       |                |                |
| visit                         | trawangan      |                |
| circumnavigating              |                |                |
| walk                           |                |                |
| snorkeling                    |                |                |
| trip                           |凉爽            |                |
| cool                           |                |                |
| prepared                      |                |                |
| higher prices                 |                |                |
| compared                      |                |                |
| mainland                      | rubbish problem|                |

In accordance with formula (2), it is gained that the score of sentiment analysis in table 2 is:

\[
\text{score} = \text{the number of positive words} - \text{the number of negative words}
\]

\[
\text{score} = (3) - (4) = -1
\]

The final score obtained from the simulation calculation is <0, therefore, the result of the review classification is negative.

After analyzing all positive sentiment in text corpus, it can be known that the number of positive reviews is caused by at least these two things:

a. It shows that the tourists who came to Gili Trawangan had a good impression of the places they visited;
b. The calculation of high positive scores is due to the large number of reviewers who did not overtly give negative reviews, but negative reviews were written after the positive reviews were preceded. Therefore, during the labeling process, positive words dominated more than the negative words and the scoring results became positive.

3.1.2. Training and Testing Data

The training data were used to form a classifier model by machine learning algorithms and this model represented knowledge as a predictor of new data classes that had never existed. The greater the training data, the better the machine would be in learning data patterns. Testing data were used to measure the extent to which the classifier was managed to classify correctly. The data used for training data and testing data were data that already had class labels (positive and negative). As for the determination of the amount of training data and testing data, we used the Pareto principle within the comparison of 80%
for the training dataset and 20 % for the testing dataset (4:1). The comparison of the amount of training data and testing data can be seen in Table 3.

### Table 3. The comparison of Training and Testing Data

| Classification | Training Data (80%) | Testing Data (20%) | Total |
|----------------|---------------------|--------------------|-------|
| Positive       | 1834                | 459                | 2293  |
| Negative       | 581                 | 145                | 726   |
| Sum            | 2415                | 604                | 3019  |

The data in Table 3 were extracted from 1.290 English-written online reviews about Gili Trawangan tourist destinations. After being extracted on excel, the total number of data become 3019, with 2293 positive data and 726 negative data as presented in Table 3. These reviews were then used as the source of training and testing datasets without the disposal of twin data. The making of training and testing data was done randomly with the assumption that each review had the same opportunity to be used as testing or training data. Data randomization was done by considering the data class on each data review application because of the unbalanced data proportion between positive and negative classes. At last, 2415 items of training data and 604 items of testing data were attained.

#### 3.1.3. Data Classification in Support Vector Machine (SVM)

The classification process was carried out by studying the training data that had been formed based on positive and negative reviews. Furthermore, the pattern of the data was studied by using the SVM algorithm to detect data patterns in each class of training data. The results of training on the SVM algorithm were tested by using testing data to get the best accuracy value on the prediction of new data. This process is called machine learning. The Kernel methods (Linear, Polynomial, Radial Base Function (RBF), and Sigmoid) were used to find the accuracy value in the SVM method. The accuracy score in each kernel method can be seen as follows:

### Table 4. Accuracy score in Kernel method

| Kernel  | Accuracy |
|---------|----------|
| Sigmoid | 88,74%   |
| RBF     | 90,23%   |
| Polynomial | 76,82% |
| Linear  | 97,84%   |

Based on the comparison of the results of some Kernel methods in Table 4, it shows that the Linear Kernel method has the highest level of accuracy (97,84%) among other Kernel methods. Thus, the classification process in this study used Linear Kernel method with SVM algorithm. The confusion matrix was used to evaluate the best model. This study used cross-validation to testify machine performance, therefore, 5-fold cross-validation was formed. The experimental results of each machine learning using SVM can be seen as follows:

### Table 5. Machine learning accuracy score in SVM

| Machine Learning | Method Accuracy |
|------------------|-----------------|
| Machine 1        | 97,84%          |
| Machine 2        | 98,17%          |
| Machine 3        | 97,18%          |
| Machine 4        | 97,84%          |
| Machine 5        | 97,35%          |
In choosing the best machine learning, training and testing data were taken randomly from the total number of review data which already categorized based on its class. Therefore, all of the data have the same chance to be the testing data or the training data. This data collecting process aimed to form a new dataset: positive training data, negative training data, positive testing data, and negative testing data. From 5 machine learning experiments using the SVM method, machine 2 produced the highest accuracy (98.17%). The results of the calculation in the level of accuracy were obtained from the number of correctly classified testing data compared to the total of all the tested data as administered by using 5-fold cross-validation. Since machine 2 reached the highest accuracy, so it was chosen to analyze the confusion matrix. Thus, the confusion matrix in machine 2 contained the amount of testing data that was classified correctly and the amount of data that was classified incorrectly in the previous stage. Further explanations can be seen in Table 6 below.

| Prediction | Actual |          |
|------------|--------|----------|
|            | Negative | Positive |
| Negative   | 136     | 9        |
| Positive   | 2       | 457      |
| Class Recall | 98.55%  | 98%      |
| Accuracy   |         | 98.17%   |

Based on the table above, the predictive results of the confusion matrix in Machine 2 were gained after being tested by using SVM as follows:

a) From 138 data reviews in the negative class, 136 data reviews had been classified correctly while the remaining 2 reviews belonged to the positive class.

b) From 466 data reviews in the positive class, 457 reviews data had been classified correctly while the remaining 9 data reviews belonged to the negative class.

3.1.4 Data Visualization and Association

Data review visualization was carried out to extract information in the form of topics most frequently discussed by visitors in Gili Trawangan tourist destinations. Thus, the review text data could provide new yet substantial information. Moreover, the association between words could strengthen the information. Labeling class reviews between positive class and negative class were done by using Support Vector Machine (SVM).

3.1.4.1 Positive Class

Information extraction on positive reviews was done repeatedly to get accurate information on data patterns. Then, these were grouped into the positive class of visitor reviews on Gili Trawangan. The positive reviews were identified based on the frequency of words in all reviews that entered the positive class. Here is the visualization of the results of information extraction from visitor reviews that had been classified by using machine learning SVM method.
Based on the classification of positive reviews, it was obtained that the words which frequently appear are <beautiful>, <great>, <good>, and <love>. Those words semantically reflect positive sentiment. These words were then used as the basis to find out the associations among the words. Hence, the information could be obtained in the form of accurate positive sentiments. The table below shows the words association in the positive class.

Table 7. Words association in the positive class

| $beautiful$ | Assoc | $great$ | Assoc | $good$ | Assoc | $love$ | Assoc |
|-------------|-------|---------|-------|--------|-------|--------|-------|
| beach       | 0.22  | beach   | 0.18  | food   | 0.21  | lived  | 0.22  |
| changers    | 0.18  | snorkeling | 0.16 | amusing | 0.18 | climbing | 0.22 |
| cove        | 0.18  | night   | 0.16  | bal    | 0.18  | dissapoint | 0.22 |
| downloaded  | 0.18  | fun     | 0.16  | entertain | 0.18 | instructors | 0.22 |
| plays       | 0.18  | intan   | 0.16  | important | 0.18 | suset   | 0.22 |
| wandered    | 0.18  | operational | 0.16 | moment  | 0.18 | guesthouse | 0.17 |
| blue        | 0.17  | source  | 0.16  | qual   | 0.18  | compulsory | 0.17 |
| waters      | 0.16  | abondonned | 0.15 | retsuarnts | 0.18 | rocked   | 0.17 |
| aqua        | 0.16  | cockroaches | 0.15 | snorkeling | 0.18 |
| white       | 0.15  | rodents | 0.15  | skin   | 0.16  |

The association among the words in the positive class gave an insight of co-occurrence between one word and the others which indicated positive sentiment. Referring to Table 7, the word <beautiful> indicated a strong association with the word <beach>. From this pattern, it can be concluded that many visitors gave feedback or review about the beauty of the beach in Gili Trawangan, as well as reflected...
in the association between <beach> and <great>. The relationship between word associations gave a detailed description of the word connotation in the class of positive sentiment.

3.1.4.2. Negative Class

The information extraction on negative reviews was done repeatedly to get negative information or feedback from the most talked issue by the tourist about Gil Trawangan. Based on the labeling results, the number of visitors’ positive reviews was fewer than the number of visitors’ negative reviews. The results of extracting information in negative reviews were identified based on the frequency of word appearance in the review. Furthermore, it was also analyzed based on the relevance of words to topics that refer to negative sentiment.

![Most Frequent words in Negative Class](image)

**Figure 3.** Most frequent words in the negative class

According to the results of the negative review classification in Figure 3, it can be seen that several words like <horse>, <rubbish>, <bad>, and <crowded>, appear frequently in negative class as to reveal the negative sentiment. These words were then used as the main basis for finding the associations among the words. It was aimed to obtain more negative sentiment accurately. The result of word association analysis can be seen below:

| Words | Frequency |
|-------|-----------|
| not   | 949       |
| horse | 275       |
| beach | 236       |
| locals| 217       |
| boats | 211       |
| bicycle| 152    |
| Rubbish| 149     |
| tourist| 117      |
| bad   | 115       |
| waters| 113       |
| coral | 104       |
| crowded| 103      |

**Table 8.** Words association in the negative class

| $\text{horse}$ | Assoc | $\text{bad}$ | Assoc | $\text{rubbish}$ | Assoc | $\text{crowded}$ | Assoc | $\text{coral}$ | Assoc |
|---------------|-------|--------------|-------|------------------|-------|------------------|-------|--------------|-------|
| urge          | 0.32  | animals      | 0.31  | plastic          | 0.49  | surprised        | 0.25  | dead         | 0.67  |
| treatment     | 0.32  | pull         | 0.31  | bottles          | 0.37  | bomb             | 0.24  | beach        | 0.41  |
| abuse         | 0.31  | despicable   | 0.30  | aids             | 0.31  | brochures        | 0.24  | adequate     | 0.36  |
| heart         | 0.31  | harnesses    | 0.30  | band             | 0.31  | designed         | 0.24  | algae        | 0.36  |
| pull          | 0.31  | lifespan     | 0.30  | currents         | 0.31  | discovering      | 0.24  | kelp         | 0.36  |
| witnessed     | 0.31  | starved      | 0.30  | hair             | 0.31  | docking          | 0.24  | market       | 0.36  |
| walk          | 0.30  | tourist      | 0.27  | lose             | 0.31  | ecofriendly      | 0.24  | marts        | 0.36  |
| bones         | 0.30  | relentlessly | 0.26  | square           | 0.31  | explained        | 0.24  | riding       | 0.36  |
| bouncing      | 0.30  | contributing | 0.26  | tip              | 0.26  | highway          | 0.24  | soda         | 0.36  |
The word associations describe the close relations among words so that it forms a pattern of information from the word extraction. Referring to table 8, the word <rubbish> has a strong association with the word <plastic>. For this reason, it can be concluded that the type of waste that was mostly complained by visitors or tourists was plastic waste. The size of the association value determines how strong the relationship between words.

3.2. Discussion
The process of sentiment analysis of tourist reviews using SVM on unstructured text data was quite challenging. The writing errors and the use of uncommon terms in expressing opinions, including the use of unstandardized words, urged the writer to create a lexicon dictionary based on the scrapped data in the corpus to fix the errors and to standardize the mistakes in language use. This means that the review data written by tourists about the Gili Trawangan tourist destination on the TripAdviser cannot be fully understood in words due to typo, error, or mistake in writing the reviews. As a result of this, the lexicon dictionary was set to standardize all of the lexemes so it could meet the context and be understandable.

The results of this study were obtained by extracting online reviews posted from July 2012 until May 2020 with a total of 1290 reviews written in English. We only collected the reviews given by foreigners to avoid subjectivity in giving a review of the places they visited. Thus, the reviews written by locals in Bahasa Indonesia were excluded. The review writing pattern was identified after analyzing the data. Mostly, the reviews were preceded by positive feedbacks and then followed by negative feedback. This certainly could provide a different assessment of the labeling process which indeed refers to the calculation of word classes collected in the lexicon dictionary. It also implied that, aside from their satisfaction about their visit to Gili Trawangan, there was still an uncomfortable thing which need to be managed better by the tourism enterprise, stakeholders, or local government to maximize the tourism potential since tourism in small island always attracts tourists to come due to its beauty and exotic nature[28].

The rating scale of reviews data written on the TripAdvisor (5: Excellent, 4: Very good, 3: Average, 2: Poor, 1: Terrible) could not be used directly as class determination or labeling tool for machine learning testing. It was because there was no certainty that reviews rated in 3 to 5 could be determined wholly in the positive class. It was found in some reviews that several reviewers, who gave high ratings, not only wrote positive feedback in their reviews but also gave complaints or negative sentiment at the same time. Therefore, the labeling process and the use of the SVM method were very important as to be used in testing the determination of the data class (positive sentiment or negative sentiment).

The sentiment class of online reviews could be classified precisely with the Support Vector Machine (SVM) method[3, 4, 29]. In this research, the average accuracy between positive and negative classes was 98.1% based on the calculation of recall class values contained in the confusion matrix table, although a large amount of training and testing data used in testing machine learning largely determines the accuracy of a method. From the result of the classification, it could be extracted into new insight about the factors or sources of the problem (rubbish, crowded, dead coral, and plastics) which were being complained by the visitors. Even though the proportion of the total number of reviews that belong to the positive class is significantly greater than the negative review, however, the data of negative

| Shorse | Assoc | $bad | Assoc | $rubbish | Assoc | $crowded | Assoc | $coral | Assoc |
|--------|-------|------|-------|---------|-------|----------|-------|--------|-------|
| focus  | 0.30  | dehyd | 0.26  | dump    | 0.24  | indifferen | 0.24  | rock   | 0.35  |
| hip    | 0.30  | beaten| 0.26  | clean   | 0.22  | killed    | 0.24  | shoes  | 0.34  |
| hook   | 0.30  | urge  | 0.25  | bin     | 0.22  | level     | 0.24  | tons   | 0.32  |
| protruding | 0.30 | waters| 0.24  | corner  | 0.21  | observing | 0.24  | interested | 0.32 |
| roughly | 0.30 | appalled | 0.22 | littered | 0.21 | on board | 0.24 | soft | 0.32 |
| rub    | 0.30  | dire  | 0.22  | pile    | 0.21  | population | 0.24 | waters | 0.31 |
| rubbing | 0.30 | discomfort | 0.22 | metre  | 0.21 | resembling | 0.24 | broken | 0.31 |
| soaking| 0.30  | impact| 0.22  | pieces  | 0.21  | sustain   | 0.24 | | |

The results of this study were obtained by extracting online reviews posted from July 2012 until May 2020 with a total of 1290 reviews written in English. We only collected the reviews given by foreigners to avoid subjectivity in giving a review of the places they visited. Thus, the reviews written by locals in Bahasa Indonesia were excluded. The review writing pattern was identified after analyzing the data. Mostly, the reviews were preceded by positive feedbacks and then followed by negative feedback. This certainly could provide a different assessment of the labeling process which indeed refers to the calculation of word classes collected in the lexicon dictionary. It also implied that, aside from their satisfaction about their visit to Gili Trawangan, there was still an uncomfortable thing which need to be managed better by the tourism enterprise, stakeholders, or local government to maximize the tourism potential since tourism in small island always attracts tourists to come due to its beauty and exotic nature[28].

The rating scale of reviews data written on the TripAdvisor (5: Excellent, 4: Very good, 3: Average, 2: Poor, 1: Terrible) could not be used directly as class determination or labeling tool for machine learning testing. It was because there was no certainty that reviews rated in 3 to 5 could be determined wholly in the positive class. It was found in some reviews that several reviewers, who gave high ratings, not only wrote positive feedback in their reviews but also gave complaints or negative sentiment at the same time. Therefore, the labeling process and the use of the SVM method were very important as to be used in testing the determination of the data class (positive sentiment or negative sentiment).

The sentiment class of online reviews could be classified precisely with the Support Vector Machine (SVM) method[3, 4, 29]. In this research, the average accuracy between positive and negative classes was 98.1% based on the calculation of recall class values contained in the confusion matrix table, although a large amount of training and testing data used in testing machine learning largely determines the accuracy of a method. From the result of the classification, it could be extracted into new insight about the factors or sources of the problem (rubbish, crowded, dead coral, and plastics) which were being complained by the visitors. Even though the proportion of the total number of reviews that belong to the positive class is significantly greater than the negative review, however, the data of negative
reviews are still very important to be basis for policy making by both government and private stakeholders as managers of tourist destination areas to boost and maximize the tourism industry in Gili Trawangan.

The negative sentiment class provides space to extract new information about the factors that still become problems in the Gili Trawangan tourist area. Consequently, these problems should be investigated and handled carefully to avoid negative effects on the local tourism industry in the future. The high frequency of certain topics which has strong connotations in negative class, such as <rubbish>, <bad>, <crowded> and <horse>, become an interesting issue in Gili Trawangan that should be investigated further by the stakeholders. Rubbish becomes the most complained issue in the negative class. The word <rubbish> strongly associated with the words <plastic> (0.49) and <bottles> (0.37). This token is also associated with the word <littered> (0.21) and <bin> (0.22). It reveals the fact that the waste in Gili Trawangan needs to be managed thoughtfully. Based on the negative class words association appearing under the token <rubbish>, bottles and plastic waste seem to be the main problem of waste management in Gili Trawangan. Besides, the word association among <rubbish>, <bin>, and <littered> in online reviews seemed to urge the authorities to discipline all of the visitors and the locals in Gili Trawangan so as not to litter. They need to provide more bins in tourism sites. This opinion was supported by previous studies about waste management[30] and environmental damage[31] in Gili Trawangan. In 2016, around 11.94-ton rubbish was produced per day and it was predicted to double in 2024 [30]. This condition worsened by the fact that waste management was not optimized yet. The minimum availability of bin and garbage dump caused people to litter [31]. For this reason, this condition provoked the tourists to write negative sentiments about Gili Trawangan on their online reviews.

The word <horse> in negative class has strong association with the words <urge> (0.32), <treatment> (0.32) and <abuse> (0.31). It implies how the locals treated horses as a mode of transportation was adjudged to be excessive exploitation of animals and it caused reviewers to write negative reviews. From the perspective of some visitors, reflected in word associations of the negative class, this kind of treatment could be considered as animal abuse. This argument was also supported by the high percentage of the association between the word <bad> and the word <animals> which scored in 0.31. For the local people, it was a common thing to ride a horse and put a cart on it as a mode of transportation. In contrast, some tourists had a different point of view about this issue since it could be considered as harming the animal, especially by overburdening the horses. This condition could be a serious trigger for animal lovers, even more so when the animal protection, such as rest and shelter area or water and food supply, was not guaranteed[32]. Regarding this issue, Pinsky et.al.[33] pointed out that 93.9% of horse drivers in their pilot study did not have enough knowledge and awareness about animal welfare. Additionally, they stated that almost 50% of ponies were underweight and their owners would labor their horses despite their injury[33]. This condition could potentially damage tourism in Gili Trawangan. Thus, the stakeholders or the local government are suggested to watch over animal protection to avoid animal abuse. It is also recommended to held education programs about animal welfare and health for the local workers.

The other intriguing token in the negative class is <coral>. This word has high co-occurrence with the word <dead> (0.67) and <beach> (0.41) in text corpus of online reviews. The ecosystem damage in Gili Trawangan has been recognized since 2012[34] which was also detected in the online reviews of negative class. The tokens of reviews in negative class proves how this condition persuades the tourists to write negative reviews of Gili Trawangan. This affects their decision to visit Gili Trawangan again in the future due to such disappointment. Sauna and Hilman said that coral reefs become a vital tourism attraction for small island tourism, therefore, the coral or ecosystem damage would affect the sustainability of marine tourism and the growth of the tourism industry[34]. In line with it, the word <crowded> in review data indicated that there were types of tourists who did not like crowds and the explosion of visitors that lead to environmental exploitation and damage since the word crowded is not only addressed to humans but also boats. Based on the reviews of foreign tourists, it caused the high intensity of waste in tourist areas since they found much plastic waste in tourist spot. This issue
obviously needs to be managed by the tourism enterprises or local government. The stakeholders and local government are suggested to manage better policy so that the visitors do not litter and worsen the nature in tourism area.

However, the stakeholders should get informed with the tourist reviews to establish a sustainable tourism industry. In some cases, tourist visits are influenced by the reviews which can be a medium for travelers to observe the places they want to visit. Those online reviews, to some extent, influence the decision making regarding the traveling planning of the tourists. The tourism enterprises and all of the stakeholders should be more agile in capturing and extracting information from the feedbacks or reviews of the visitors, especially for the negative sentiments. Tourists’ negative sentiment could be used to manage the development and sustainability of the tourism industry since it could interfere with the number of visits in particular tourist attractions. Since marine tourism becomes a substantial means of economic development in the small island [35], protecting the ecosystem and paying attention to tourists’ reviews could be beneficial tools in boosting the tourism industry.

Tourism activities have become a critical point as the source of local income in small islands. In a small island like Gili Trawangan, tourism becomes the main economic value [36] thus managing the tourism sustainability in this island is necessary. Moreover, maintaining the natural balance in the coastal area also should be a shared priority because this is directly related to natural and human sustainability. The natural beauty of coastal ecotourism or marine tourism becomes a valuable selling point to boost the number of tourist visits. The quality of the landscape, the quality of the environment, and safety become the influential factors in tourism [37, 38, 39]. Besides, means of transportation and public facilities also become additional factors to attract tourists. Therefore, in the case of tourist reviews of Gili Trawangan tourism sites, we imply the local government and stakeholders to pay more attention to the natural sustainability to boost tourism sustainability in Gili Trawangan.

4. Conclusion

Online travel platform like TripAdvisor has been highly popular among travelers worldwide as a reference for finding tourist destinations. Online footprints left in the review column help the platform users to decide where to go and plan their journey. These reviews, however, also help the stakeholders or the tourism enterprise to develop their services as they can use the feedbacks to forecast the market demand. 1290 online reviews about Gili Trawangan had been scrapped from TripAdvisor to analyze the sentiment of Gili Trawangan Tourists. The information extraction in positive class shows that the tourists were mesmerized by the beauty of Gili Trawangan. Meanwhile, two main topics, environmental damage and animal abuse, became the most frequently criticized issues in the negative class. Most tokens in the negative class were dominated by working pony abuse, rubbish and littering, and dead coral. In this case, we identified that the foreign tourists were disappointed about how the horse drivers treated their working ponies and considered their act as animal cruelty. The online reviewers also got disturbed by the rubbish and littering habit. It polluted the environment and possibly killed the animals. They also felt disappointed because they found many dead corals on the beach and bleaching coral underwater. These negative sentiments certainly damage the image of Gili Trawangan as the lost paradise. Since tourism becomes an important economic resource in a small island like Gili Trawangan, it is important for tourism enterprises and the stakeholders to pay attention to both information extractions in positive class and negative class to boost the surrounding tourism industry. These online reviews significantly interfere with the number of tourists in Gili Trawangan. Therefore, we suggest the stakeholders pay more attention to animal protection, waste management, and ecosystem vulnerability in Gili Trawangan.

5. References

[1] M. Tamrin, “Big Data: A New Challenge for Tourism Sector in Nusa Tenggara Barat,” in The First Sustainable Tourism National Seminar STP Mataram, 2019, pp. 193–198.

[2] X. Tian, W. He, R. Tao, and A. Vasudeva, “Mining Online Hotel Reviews: A Case Study from Hotels in China,” in American Conference on Information Systems, 2016.
[3] J. B. Chrystal and S. Joseph, “Text Mining and Classification of Product Reviews Using Structured Support Vector Machine,” in *International Conference on Computer Science, Information Technology and Applications*, 2015, pp. 21–31.

[4] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of Sentimental Reviews Using Machine Learning Techniques,” in *The Third International Conference on Recent Trends in Computing 2015 (ICRTC-2015)*, 2015, pp. 821–829.

[5] Q. He, C. A. W. Glas, M. Kosinski, D. J. Stillwell, and B. P. Veldkamp, “Predicting Self-monitoring Skills Using Textual Post on Facebook,” *Comput. Human Behav.*, vol. 33, pp. 69–78, 2014.

[6] G. Chareyron, J. Da-Rugna, and T. Rainbault, “Big Data: A New Challenge for Tourism,” in *IEEE International Conference on Big Data*, 2014.

[7] J. Li, L. Xu, L. Tang, S. Wang, and L. Li, “Big Data in Tourism Research: A Literature Review,” *Tour. Manag.*, vol. 68, pp. 301–323, 2018.

[8] M. H. Salas-Olmedo, B. Moya-Gomez, J. C. Garcia-Palomares, and J. Gutierrez, “Tourists’ Digital Footprint in Cities: Comparing Big Data Sources,” *Tour. Manag.*, vol. 66, pp. 13–25, 2018.

[9] B. M. Jadav and V. B. Vaghela, “Sentiment Analysis Using Support Vector Machine Based on Feature Selection and Semantic Analysis,” *Int. J. Comput. Appl.*, vol. 146, no. 13, pp. 26–30, 2016.

[10] H. S. Halim, “Scrutinizing Coastal Ecotourism in Gili Trawangan, Indonesia,” *Int. J. Mar. Sci.*, vol. 7, no. 25, pp. 247–259, 2017.

[11] KEMENPAREKRAF, “Pariwisata Tambang PAD Terbesar Lombok Utara,” 2018. [Online]. Available: https://www.kemenparekraf.go.id/index.php/post/news-pariwisata-tambang-pad-terbesar-lombok-utara. [Accessed: 24 Jun-2020].

[12] C. Wu, H. Che, T. Y. Chan, and X. Lu, “The Economic Value of Online Review,” *Mark. Sci.*, vol. 34, no. 5, pp. 739–754, 2015.

[13] G. Jiang, J. Shang, W. Liu, X. Feng, and J. Lei, “Modeling the Dynamics of Online Review Life Cycle: Role of Social and Economic Moderations,” *Eur. J. Oper. Res.*, vol. 285, no. 1, pp. 360–379, 2020.

[14] W. Rashideh, “Blockchain Technology Framework: Current and Future Perspectives for the Tourism Industry,” *Tour. Manag.*, vol. 80, 2020.

[15] L. T. T. Tran, “Online Reviews and Purchase Intention: A Cosmopolitanism Perspective,” *Tour. Manag. Perspect.*, vol. 35, 2020.

[16] U. Gretzel and K. H. Yoo, “Use and Impact of Online Travel Reviews,” in *Information and Communication Technologies in Tourism*, 2008, no. January.

[17] L. Dediu, “Users’ Reviews on Tourism Sites: Their Influence on the Potential Tourists,” *Cactus Tour. J.*, vol. 13, no. 1, pp. 41–55, 2016.

[18] S. Hlee, H. Lee, and C. Koo, “Hospitality and Tourism Online Review Research: A Systematic Analysis and Heuristic-Systematic Model,” *Sustainability*, vol. 10, p. 1141, 2018.

[19] M. Schuckert, X. Liu, and R. Law, “Hospitality and Tourism Online Reviews: Recent Trends and Future Directions,” *J. Travel Tour. Mark.*, vol. 32, pp. 608–621, 2015.

[20] H. Li, M. Hu, and G. Li, “Forecasting tourism demand with multisource Big Data,” *Ann. Tour. Res.*, vol. 83, pp. 1–13, 2020.

[21] A. F. Colladon, B. Guardabascio, and R. Innarella, “Using Social Network and Semantic Analysis to Analyze Online Travel Forums and Forecast Tourism Demand,” *Decis. Support Syst.*, vol. 123, 2019.

[22] N. Guduru, “Text Mining with Support Vector Machines and Non-Negative Matrix Factorization Algorithms,” University of Rhode Island, 2006.

[23] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text Classification Algorithms: A Survey,” *Information*, vol. 10, no. 150, pp. 1–68, 2019.

[24] U. Y. Nahm and R. J. Mooney, “Text Mining with Information Extraction,” in *AAAI 2002 Spring*
Symposium on Mining Answers from Texts and Knowledge Bases, 2002, pp. 60–67.

[25] X. Hu and H. Liu, “Text Analytics in Social Media,” in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. New York: Springer, 2012, pp. 385–390.

[26] K. S. Jones, “A Statistical Interpretation of Term Specificity and Its Application in Retrieval,” J. Doc., vol. 60, no. 5, pp. 11–21, 2004.

[27] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press, 2013.

[28] F. Kurniawan, L. Adrianto, D. G. Bengen, and L. B. Prasetyo, “Vulnerability Assessment of Small Islands to Tourism: The Case of the Marine Tourism Park of the Gili Matra Islands, Indonesia,” Glob. Ecol. Conserv., vol. 6, pp. 308–326, 2016.

[29] Y. Liu, J.-W. Bi, and Z.-P. Fan, “A Method for Multi-class Sentiment Classification Based on An Improved One-VS-One (OVO) Strategy and the Support Vector Machine (SVM) Algorithm,” Inf. Sci., vol. 394–395, pp. 38–52, 2017.

[30] R. Ramli, “Solid Waste Management Analysis to Support A Sustainable Tourism Object of Gili Trawangan,” Gadjah Mada University, 2016.

[31] N. Budilesteri, M. Hutomo, and R. Ardiwijaja, “Permasalahan Lingkungan di Sempadan Pantai Taman Wisata Perairan Gili Trawangan, Nusa Tenggara Barat,” J. Kepariwisataan Indones., vol. 9, no. 1, pp. 91–107, 2014.

[32] WAP, “Your Guide to be Animal Friendly on Vacation,” World Animal Protection, Ottawa, 2014.

[33] T. C. Pinsky, I. K. Puja, J. Aleri, J. Hood, M. M. Sasadara, and T. Collins, “A Pilot Welfare Assessment of Working Ponies on Gili Trawangan, Indonesia,” Animals, vol. 9, no. 433, pp. 1–20, 2019.

[34] I. W. Suana and H. Ahyadi, “Mapping of Ecosystem Management Problems in Gili Meno, Gili Air and Gili Trawangan (Gili Matra) through Participative Approach,” J. Coast. Dev., vol. 16, no. 1, pp. 94–101, 2012.

[35] R. R. Croes, “A Paradigm Shift to A New Strategy for Small Island Economies: Embracing Demand Side Economics for Value Enhancement and Long Term Economic Stability,” Tour. Manag., vol. 27, pp. 453–465, 2006.

[36] S. Graci, “What hinders the path to sustainability? A study of barriers to sustainable tourism development in Gili Trawangan, Indonesia,” Pacific News, vol. 29, 2008.

[37] A. Krause and E. Weir, Ecotourism: management, development and impact. New York: Nova Science Publishers, 2011.

[38] M. C. Chien, “An empirical study on the effect of attractiveness of ecotourism destination on experiential value and revisit intention,” Appl. Ecol. Environ. Res., vol. 15, no. 2, pp. 43–53, 2016.

[39] M. R. Phillips, “Beach response to a total exclusion Barrage: Cardiff Bay, South Wales, UK,” J. Coast. Res., vol. 233, pp. 794–805, 2007.