Heliyon 8 (2022) e10894

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning

Natt Leelawat [a,b], Sirawit Jariyapongpaiboon [a], Arnon Promjun [a], Samit Boonyarak [c], Kumpol Saengtabtim [a], Ampan Laosunthara [b], Alfan Kurnia Yudha [a], Jing Tang [b,c,*]

[a] Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
[b] Disaster and Risk Management Information Systems Research Unit, Chulalongkorn University, Bangkok, Thailand
[c] International School of Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

ARTICLE INFO

Keywords:
COVID19
Machine learning
Sentiment analysis
Tweet
Tourism
Thailand

ABSTRACT

The coronavirus disease 2019 (COVID-19) pandemic has severely affected Thailand's economy, which relies heavily on tourism. In this study, we labeled the sentiment and intention classes of English-language tweets related to tourism in Bangkok, Chiang Mai, and Phuket. Then, the accuracy of three machine learning algorithms (decision tree, random forest, and support vector machine) in predicting the sentiments and intentions of the tweets was investigated. The support vector machine algorithm provided the best results for sentiment analysis, with a maximum accuracy of 77.4%. In the intention analysis, the random forest algorithm achieved an accuracy of 95.4%. In a subsequent preliminary qualitative content analysis, the top 10 words found in each sentiment and intention class were gathered to provide insights and suggestions to help increase tourism in Thailand. The results of this study suggest that to help restore tourism in Thailand, tourist destinations, natural attractions, restaurants, and nightlife should be promoted. In addition, the two main concerns of tourists to Thailand should be addressed: COVID-19 and current political tensions.

1. Introduction

The coronavirus disease 2019 (COVID-19) outbreak began on December 31, 2019 in Wuhan, China [1]. The first case in Thailand occurred on January 13, 2020. In late March 2020, Thailand mandated the closure of shopping malls, educational institutions, offices, tourist locations, and all other non-essential venues and services. Lockdowns and curfews were implemented in Thailand to control the spread of the disease. However, their effect on the Thai economy was severe because Thailand relies heavily on tourism and exports [2].

During the implementation period of lockdown and preventive measures, both domestic and international travel was banned, significantly affecting revenue from the tourism-related sector. This revenue accounted for approximately 12% of Thailand's gross domestic product (GDP) and approximately 50% of Thailand's export GDP, the latter being further affected by the COVID-19-induced global recession [3]. The social distancing measures also affected domestic spending, especially face-to-face activities such as tourism, recreation, hospitality, and the purchase of luxury products [4]. Although the COVID-19 situation in Thailand is uncertain and may continue as new waves of infection occur, the government has eased preventive measures to enable economic recovery and has initiated campaigns to support visits by foreigners.

Tourism is a rapidly evolving industry that constantly faces new challenges, and a major challenge at present is to develop sustainable tourism models that will ensure the longevity of the industry. Information and communication technology (ICT) has become essential in tourism destination management and is used by both business owners and stakeholders to ensure the survivability and competitiveness of the tourism industry. Therefore, ICT has been embraced as an integral part of sustainable tourism models [5].

As tourists have become more experienced and digitally literate, they have gained access to information and increased their negotiating power by utilizing the latest ICT [6]. Big data, the Internet, and social media have changed the way people travel. For example, when choosing a travel destination, people often seek to optimize their travel experiences through online research. This pre-travel phase is followed by their decision to travel, undertaking of the travel, and post-travel phase, during which they often share their experiences and journeys online through

* Corresponding author.
E-mail address: jing.t@chula.ac.th (J. Tang).

https://doi.org/10.1016/j.heliyon.2022.e10894
Received 20 March 2022; Received in revised form 17 June 2022; Accepted 27 September 2022

social media [7]. Accordingly, social media has become the most popular source of inspiration and opinions for Internet users searching for leisure and entertainment opportunities, new destinations, and activities at tourist destinations.

The tourism industry is one of the highest income industries for Thailand [8]. Not only do foreign travelers enjoy traveling to Thailand, but Thai people also enjoy traveling in their own country [9]. Bangkok, the capital of Thailand, is one of the most popular travel destinations in Thailand due to being a major hub for international travel in Southeast Asia [10]. In addition to Bangkok, Chiang Mai and Phuket, which are located in the northern and southern parts of Thailand, respectively, are also popular tourist destinations in Thailand [8]. However, due to COVID-19, the number of both foreign and domestic travelers has significantly decreased since the beginning of 2020 due to COVID-19 restriction policies, such as the travel restriction and lockdown policy, which were implemented several times since COVID-19 first appeared in Thailand. Figure 1 and Figure 2 illustrate the effect of the COVID-19 pandemic based on the decrease in the number of tourists in Bangkok, Chiang Mai, and Phuket.

Social media platforms, such as Facebook, Twitter, and Instagram, have grown exponentially in recent years, providing valuable sources of information for analyzing social trends and opinions. Social media platforms also benefit travelers who wish to better understand the travel industry. Twitter has over 330 million monthly active users and 145 million daily active users worldwide [8]. Using specific hashtags on Twitter, it is possible to find tweets related to a particular topic and thereby analyze people's opinions on tourism in Thailand. In a previous study, Mehraliyev et al. [11] recommended that tourism researchers utilize Twitter data. In addition, Leelawat et al. [10] extracted trends in topics related to COVID-19 since the beginning of the pandemic. Similar research during the COVID-19 pandemic was conducted by Lu and Zheng [12], who reported the time series dynamics of public sentiment on Twitter toward cruise tourism and its driving factors during the COVID-19 pandemic.

Mehraliyev et al. [11] performed an extensive literature review of studies on sentiment analysis in the field of hospitality and tourism and found that when first developed, sentiment analysis was often used as an alternative to the traditional data collection and analysis techniques of surveys and interviews. As a result, sentiment analysis has been increasingly used to provide novel perspectives on existing research questions and relationships in hospitality and tourism research. However, to the best of our knowledge, the effects of personalization on customer experience and sentiment and the effects of review sentiment on other customers' perceptions of tourism products and experiences have rarely been investigated. With changes in travel behavior and experiences during the COVID-19 pandemic, it is necessary to investigate the shift in traveler perceptions of tourism products to identify appropriate strategies to recover from the loss caused by the pandemic.

Currently, sentiment analysis is one of the most popular research techniques to understand the opinions and feelings of the people based on the text data [10]. Machine learning is also one of the popular tools for analyzing the data based on the large amount of data. Nowadays, many researchers try to use the good points of the machine learning to perform the sentiment and opinion mining. Hasan et al. [13] analyzed the political sentiment of the people using the Naïve Bayes and Support Vector Machines (SVM). Ahmad et al. [14] explained the popularity of SVM for using in many research analyses and especially for the sentiment analysis which also been used in the research. Furthermore, Ahuja et al. [15] also extracted the features from the text data which is considered as the unstructured data using the TF-IDF and performed the analysis using the six machine learning classification techniques such as Random forest, Decision tree, etc.

Machine learning is a type of artificial intelligence that can automatically improve its predictive power through experience [13]. Experience is gained by training on a dataset with a given number of samples. SVM is a supervised learning model with associated learning algorithms that analyze data used for regression analysis. It performs classification tasks by maximizing the margin separating classes while minimizing the classification error. Specifically, it classifies samples by finding the optimal hyperplane (n-dimensional region) in the hyperspace (n-dimensional space) to divide the samples into classes [14]. SVM can only take numerical attributes as input because each sample is plotted in a virtual hyperspace, which is then divided by a hyperplane. Another machine learning algorithm is Classification and Regression Tree (CART), which is used to generate a decision tree for classifying samples based on sample attributes [15]. A decision tree is a decision support tool that uses a tree-like model of decisions and possible outcomes. This algorithm recursively separates observations to construct a tree to improve the prediction accuracy. The decision tree model can reveal how each attribute relates to the target. For example, decision trees have been employed in a machine learning system for automated cataloging of large-scale sky surveys [16]. In addition, decision trees have been compared with logistic regression for credit risk analysis [17], and it was concluded that the decision tree provide higher performance than logistic regression. Another machine learning algorithm is the random forest algorithm, which uses multiple decision trees to classify samples based on the sample attributes [18]. Each decision tree has a different order of nodes and a different number of branches. The result of each sample is determined by the majority rule of the decision trees.

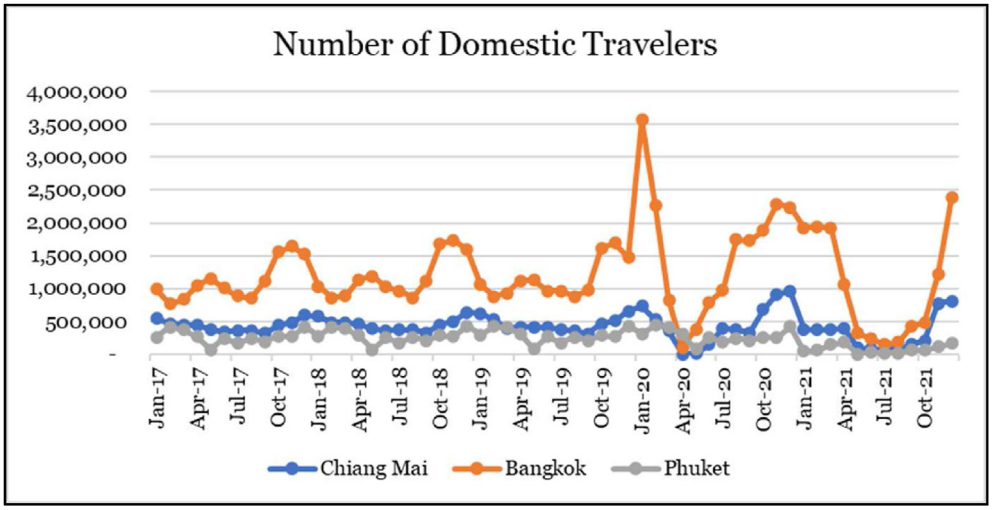

Figure 1. Number of domestic travelers for Chaing mai, Bangkok, and Phuket (2017–2021).

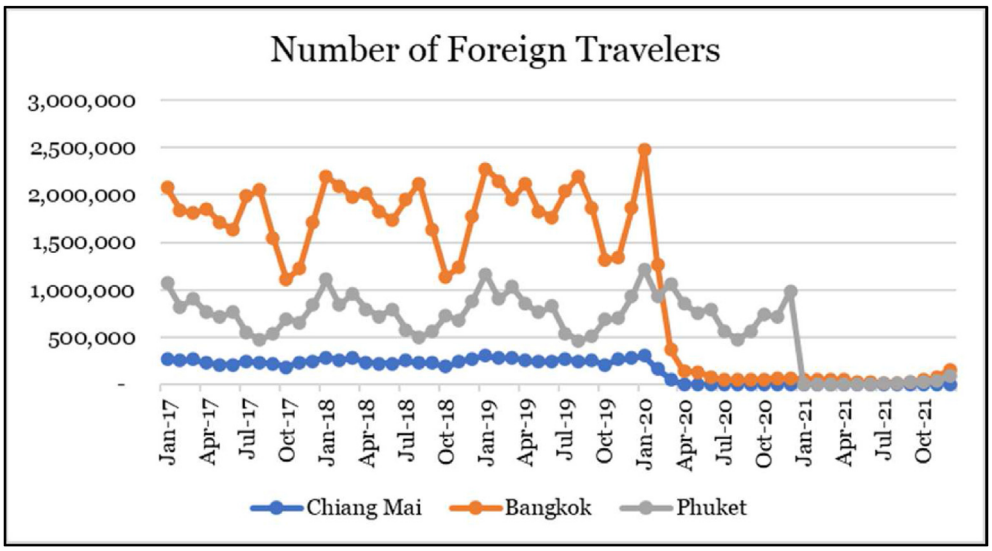

Figure 2. Number of foreign travelers for Chaing mai, Bangkok, and Phuket (2017–2021).

The present study analyzes the sentiments of tourists in Thailand using machine learning algorithms, which learn more quickly and with less complexity than deep learning algorithms. In the present study, machine learning algorithms are optimized with English-language tweets expressing sentiments about tourist attractions, events, festivals, and experiences from July to December 2020. After gaining insights from the obtained sentiments, this study provides informed suggestions on restoring Thai tourism, particularly regarding changes that can help Thai tourism adapt to the changing COVID-19 and political situation. The results can help our understanding of how COVID-19 has affected the tourism industry in Thailand and altered tourists' sentiments about attractions, events, and festivals in Thailand. The results also provide additional insights into the tourism industry in Thailand as well as suggestions for improvement.

2. Materials and methods

Figure 3 presents a flowchart of the present study. First, the necessary datasets were obtained and divided into training and testing datasets. After cleaning and preprocessing, the training datasets were input into three machine learning algorithms: SVM, CART, and random forest. Among these algorithms, only SVM required vectors weighted by the term frequency–inverse document frequency (TF–IDF). The best parameters were determined using randomized search cross-validation (RandomizedSearchCV). Then, the testing datasets were input to the trained machine learning models. RandomizedSearchCV was iterated until the prediction accuracy reached an acceptable level (i.e., >50%). When satisfactory results were obtained, they were compared across multiple models.

2.1. Data collection

Data collection was the first step of the research process. English-based tweets related to Thailand tourism from July 1 to December 31, 2020, were retrieved from the Twitter Application Programming Interface (API) accessed through the Tweepy library. Valid tweets needed to mention Bangkok, Chiang Mai, or Phuket. We selected the aforementioned study period because it was the period in which Thailand first began to relax its lockdown policy and preventive measures due to relief from the first wave of COVID-19 [16]. At that time, tourism demand in Thailand also increased due to several famous festivals, such as the Loy Krathong Festival and New Year Festival. Keywords for data collection included the top cities in Thailand ("*Bangkok*," "*Chiang Mai*," and "*Phuket*"). Next, we combined the top cities with keywords related to the most popular places and activities among travelers in Thailand by using "AND" as the logic term. These keywords were "*travel*," "*tourist*," "*trip*," "*tour*," "*vacation*," "*adventure*," "*landmark*," "*photo*," "*temple*," "*journey*," "*wat*" (i.e., "temple" in Thai), and "*river*." For each keyword, we used "OR" as the logic term to collect all related content. In addition, keywords that were used together with the three target provinces of focus were related to the tourism industry. We selected these logic terms based on the most popular activities among tourists in Thailand, which helped us to filter out unrelated tweets. Moreover, because this study focuses on foreign visitors to Thailand, the language of the Twitter data was restricted to English. Table 1 presents the sizes of the datasets for the three main cities—Bangkok, Phuket, and Chiang Mai—used in this study.

Furthermore, Figures 4 and 5 illustrate the number of tweets based on the sentiment (positive, neutral, or negative) and the intention of users (to visit or not visit).

2.2. Data cleaning

Data cleaning removes noise and unrelated content from Twitter data, such as advertisement tweets, news-related tweets, spam tweets, and tweets unrelated to tourism. This process was performed by visual inspection. The tweets remaining after the data cleaning step were reserved for further analysis. Twitter accounts identified as spam or reposts of the same tweets were also blocked and filtered out. During this process, the contents of the actual tweets were neither modified nor removed. Before processing by the machine learning models, the tweets were subjected to the following natural language processing steps:

1. Removal of all punctuation from the tweet.
2. Removal of stop words, such as "*I*," "*me*, "*my*," "*myself*," "*you*," "*you're*," and "*you've*," from the tweets using the Natural Language Toolkit (NLTK) libraries.
3. Lemmatization of each word into its root using the WordNet Lemmatizer of the NLTK. For example, this process removes the end of "*looked*" to form "*look*," and inflects "*ate*" to its root, "*eat*."
4. Removal of URLs and HTML tags that were retained in the tweets by previous collection processes.
5. Tokenization of the tweets by splitting them into individual words (unigrams) or two consecutive words (bigrams) in the feature sets.

2.3. Data labeling

This step classified the labeled data into three main categories (positive, negative, and neutral) and two intention values ("to visit" and "not to visit"). Tables 2 and 3 provide the specific criteria for labeling the

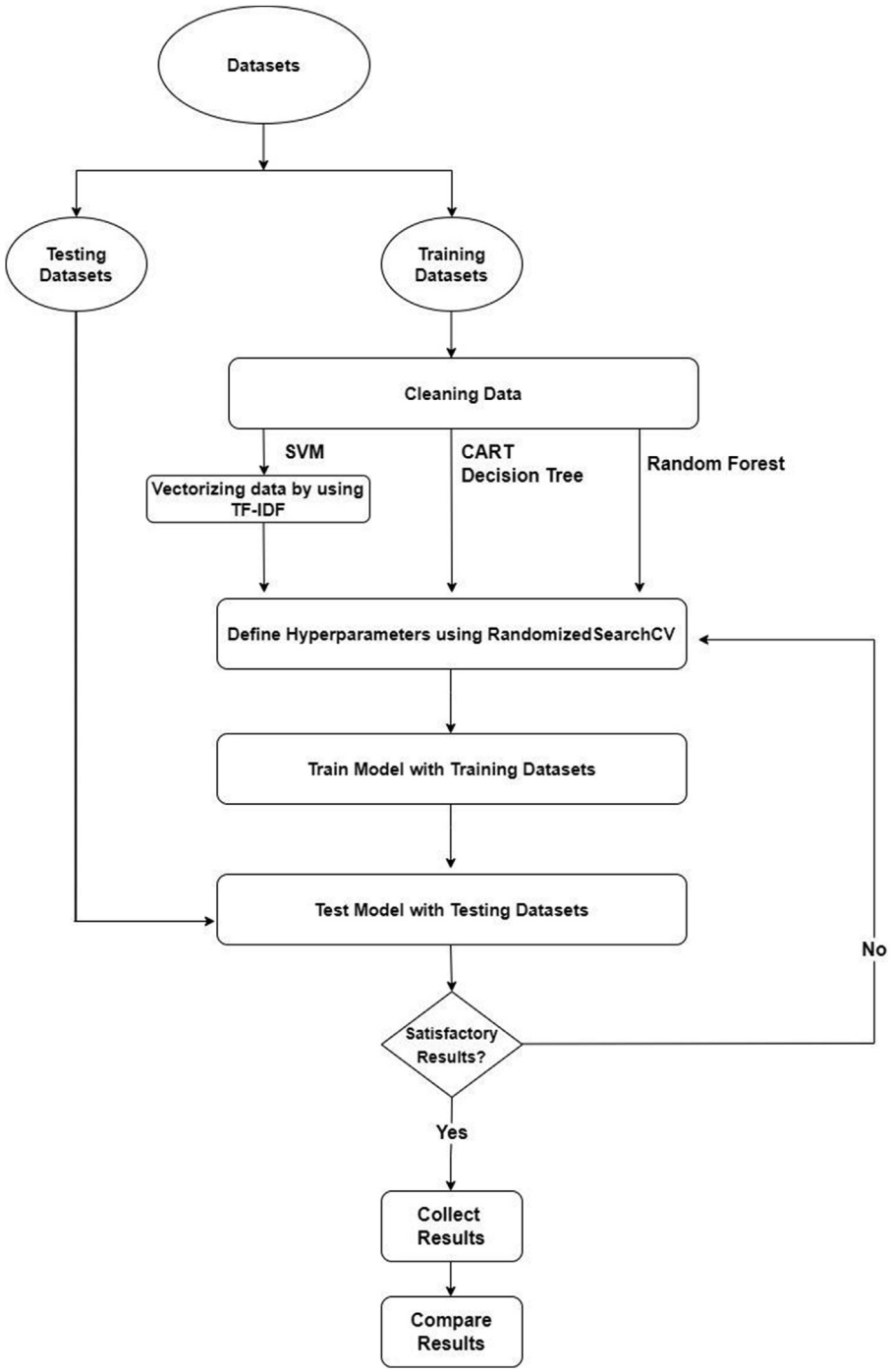

Figure 3. Flowchart of the present study.

Table 1. Number of foreign tweets collected for three main cities in Thailand.

	Jul	Aug	Sep	Oct	Nov	Dec	Total
Bangkok	10,746	9,913	9,778	11,236	9,698	10,614	61,985
Phuket	2,980	10,519	13,345	14,965	16,547	25,546	83,902
Chiang Mai	831	807	759	899	794	603	4,693

categories and intentions of the Twitter data, respectively. Tweets expressing positive opinions or perceptions of Thailand were labeled as positive, whereas tweets expressing negative opinions or criticism were labeled as negative. Tweets expressing neither positive nor negative attitudes were labeled as neutral. In this process, the tweets were labeled by the consensus of three researchers. If the sentiments of the tweets were not judged identically by all researchers, they were reassessed until a consistent judgment was made. From the total of 150,580 tweets,

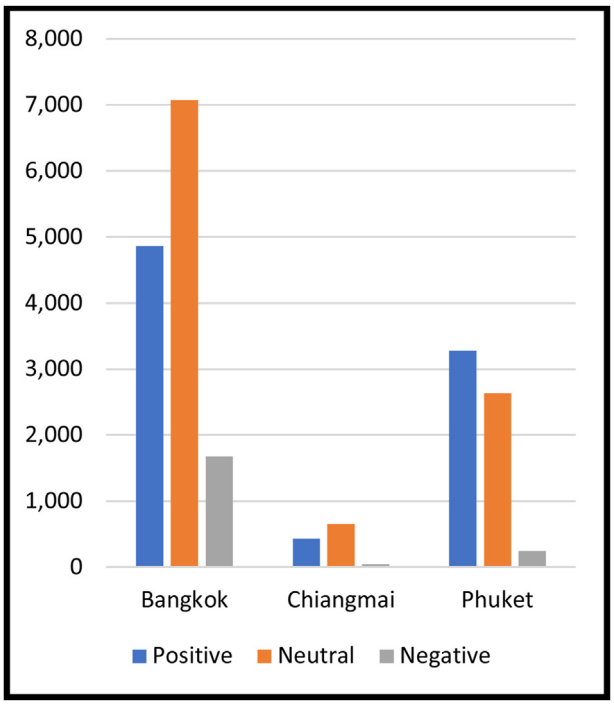

Figure 4. Number of tweets separated by sentiment.

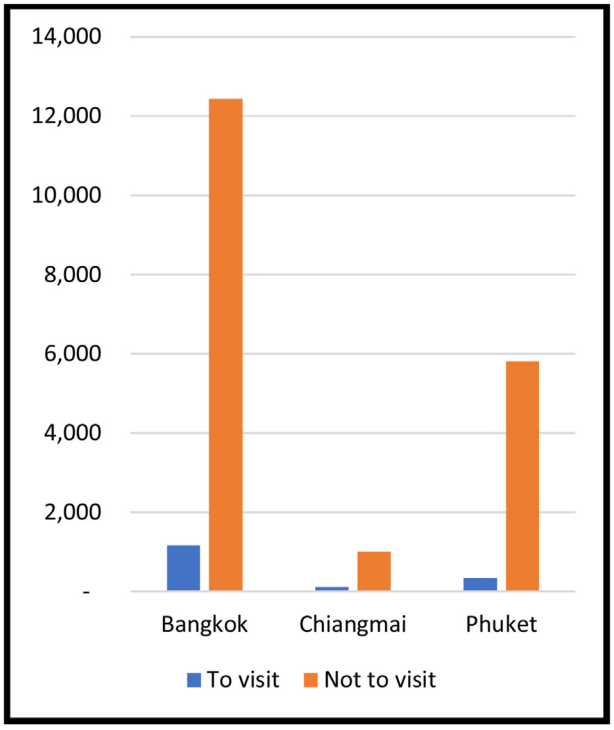

Figure 5. Numbers of tweets separated by intention to visit or not visit.

approximately 18,000 tweets were labeled for each sentiment category. Therefore, for the total number of tweets per day, at least 100 tweets were randomly assigned a label.

2.4. Data preprocessing

As SVM requires numerical values, the words were tokenized using the TF–IDF to form a set of features for each tweet. The TF–IDF provides numerical values associated with the feature sets of unigrams and bigrams. The TF–IDF is a quantitative measure (importance score) of a word in the entire set of collected samples. Its value is the product of the TF (i.e., number of times a word appears in a sample document) and IDF (i.e., inverse of the number of times a word appears in the entire set of samples). According to [19], the TF is computed as shown in (1):

$$tf(term, document) = \frac{f(term, document)}{\sum_{term' \in document} f(term', document)}, \quad (1)$$

and the IDF is given by (2):

$$idf(term,\ allDocuments) = log\frac{N}{df(t)}. \quad (2)$$

After computing the TF–IDF, the data for each city were split into training and testing datasets at a 90:10 ratio. The training set was further split into training and validation datasets at a ratio of 4.5:1 based on the use of RandomizedSearchCV. The training set was used for training the model and determining the optimal set of hyperparameters in different situations, such as selecting the type of machine learning algorithm, analyzing the sentiment or intension to visit, and distinguishing unigrams from bigrams. The testing set was used to compare the accuracy of all machine learning models in the above-mentioned situations. Each dataset had the same ratio of sentiment classes and intension-to-visit classes.

The number of tweets in the sentiment and intention classes within the training set was unbalanced, which biased the model prediction toward the class with the highest number of tweets. These biases were resolved by oversampling the sentiment data and undersampling the intention-to-visit data. At the end of preprocessing, the same number of sample tweets were assigned to each class.

Table 2. Criteria for sentiment classification.

Positive	Neutral	Negative
• Expresses an eagerness to visit • Describes positive experiences of Thailand • Uses positive adjectives • Suggests why others should visit a particular tourist destination	• Displays the current location at the time of creating a tweet • Describes a tourist destination without expressing positive or negative feelings • Tweets a question related to Thailand tourism	• Describes negative experiences of Thailand • Uses profanity • Displays a lack of willingness to visit • Exhibits a willingness to avoid Thailand

Table 3. Criteria for intention labeling.

To Visit	Not to Visit
• Displays an eagerness to visit	• Displays a lack of willingness to visit • Displays a willingness to avoid Thailand

2.5. Model training

The model training and optimization process was assisted by RandomizedSearchCV, which automatically split the training dataset into training and validation datasets at a ratio of 4.5:1 and obtained the optimal set of hyperparameters by comparing the accuracies of the model after testing on the validation dataset. To select the hyperparameters, randomized search cross-validation was performed 10 times (or more if the accuracy was below 60.0%). On one occasion, the random forest algorithm required 100 runs of randomized search cross-validation to achieve the required accuracy. As a result, the hyperparameters that had been used for tuning before performing the analysis are illustrated in Table 4.

2.6. Model testing and comparison

The accuracy, weighted precision, weighted recall, and weighted F1 score of the trained models with the optimal hyperparameters were

Table 4. Tuning hyperparameter for each type of selected machine learning algorithm.

CART	Random forest	SVM
Decision criterion (Criterion)	Number of trees	Kernel type
Maximum tree depth (Max depth)	Decision criterion (Criterion),	Regularization parameter (C)
Minimum number of samples for splitting nodes (Min sample splits)	Maximum tree depth (Max depth)	Gamma parameter (Gamma)
Minimum number of samples in a leaf (Min sample leaf)	Minimum number of samples for splitting nodes (Min sample splits)	
	Minimum number of samples in a leaf (Min sample leaf).	

determined on the testing set. Prior to training, the testing set was split to ensure a fair comparison between all models.

2.7. Content analysis

Preliminary qualitative content analysis was performed on the top 10 most common words in each class to understand more about the characteristic of data, which were found using FreqDist from the NLTK library, omitting common and uninformative words for each province. For example, common words, such as "Bangkok," "travel," "tour," and "tourist," along with uninformative words, such as "riv," "\u200b," and "19," were removed from the Bangkok dataset.

3. Results

3.1. Sentiment analysis

Sentiment analysis was performed on the separate Bangkok, Chiang Mai, and Phuket datasets and on a combination of these datasets. The results of data preprocessing method that provide the highest accuracy and F1-score can are presented in Tables 5, 6, and 7. SVM yielded the highest accuracy (up to 77.4%) with F1 score of 0.771 in sentiment analysis, which we considered acceptable. The lower accuracy of SVM on the Chiang Mai dataset (66.7%) with F1-score of 0.662 than on the other datasets can be attributed to the inadequate number of tweets related to Chiang Mai. The results of SVM on the two larger datasets indicated that the lower accuracy was not due to SVM.

3.2. Intention analysis

In the intention analysis (*see* Tables 8, 9, and 10), the random forest algorithm achieved the highest accuracy on the Bangkok dataset (95.4%) with the F1-score of 0.950, whereas SVM achieved the highest accuracy on the Chiang Mai and Phuket datasets (91.2% and 93.8%, respectively) with the F1-score of 0.893 and 0.926 respectively.

3.3. Content analysis

3.3.1. Sentiments

After performing sentiment analysis, we continued data analysis by using the text data pertaining to each sentiment, and extracted the frequency of the top keywords. As illustrated in Table 11, common positive and negative themes emerged for the three cities. Two common positive themes were tourist destinations (e.g., city, beach, and island) and hospitality (e.g., luxury, bedroom, and private). Common negative themes were related to the political situation and COVID-19 pandemic. The neutral tweets differed among the three cities, and no clear themes emerged. The neutral tweets included statements such as "*I'm at …*" followed by the tweeter's current location at the time of creating the tweet but expressing neither positive nor negative feelings about the location. Such statements constituted the largest class for all three cities.

Table 12 lists the focal tourism-related words in the three datasets which are Bangkok (BKK), Chiang Mai (CM), and Phuket (PK) after a more specific analysis of the datasets. This table displays the positive sentiments specific to each city. In Bangkok, most of the positive sentiments referred to tourist destinations ("*Temple*," "*Wat*," and "*Market*"), nightlife ("*Night*," "*Bar*," and "*Friend*"), and food ("*Bar*" and "*Restaurant*"). Visitors to Chiang Mai focused on natural attractions ("*Doi*" ["mountain" in Thai], "*Waterfall*," and "*Mountain*") along with "*Restaurant*" and "*Coffee*." Impressions of Phuket were also dominated by natural attractions ("*Beach*," "*Island*," and "*Sea*") and hospitality ("*Luxury*," "*Bedroom*," and "*Bathroom*"). Apart from the political situation and COVID-19 pandemic, common negative sentiments were "*Sexism*," "*Airport*," "*Refund*," and, surprisingly, "*Elephant*."

3.3.2. Intentions

As indicated in Table 13, the "to visit" class was associated with words such as "*go*," "*want*," "*visit*," and "*take*." Many of these terms were also

Table 5. Results from sentiment analysis for Bangkok dataset.

Algorithm	Data preprocessing	Accuracy	F1-score
CART	Unigram, Over-sampling	0.637	0.601
Random Forest	Unigram, Under-sampling	0.681	0.682
Support Vector Machine	Unigram, Under-sampling	0.721	0.701

Table 6. Results from sentiment analysis for Chiang Mai dataset

Algorithm	Data preprocessing	Accuracy	F1-score
CART	Unigram, Over-sampling	0.637	0.578
Random Forest	Unigram, Over-sampling	0.637	0.634
Support Vector Machine	Unigram, Over-sampling	0.667	0.662

Table 7. Results from sentiment analysis for Phuket dataset

Algorithm	Data preprocessing	Accuracy	F1-score
CART	Unigram, Under-sampling	0.639	0.632
Random Forest	Unigram, Over-sampling	0.708	0.713
Support Vector Machine	Unigram, Over-sampling	0.774	0.771

Table 8. Results from Intention Analysis for Bangkok dataset.

Algorithm	Data preprocessing	Accuracy	F1-score
CART	Bigram, Over-sampling	0.943	0.933
Random Forest	Bigram, Over-sampling	0.954	0.950
Support Vector Machine	Bigram, Over-sampling	0.920	0.886

Table 9. Results from Intention Analysis for Chiang Mai dataset

Algorithm	Data preprocessing	Accuracy	F1-score
CART	Bigram, Over-sampling	0.911	0.879
Random Forest	Unigram, Over-sampling	0.775	0.804
Support Vector Machine	Unigram, Over-sampling	0.912	0.893

Table 10. Results from Intention Analysis for Phuket dataset

Algorithm	Data preprocessing	Accuracy	F1-score
CART	Bigram, Over-sampling	0.932	0.914
Random Forest	Unigram, Over sampling	0.797	0.797
Support Vector Machine	Unigram, Over-sampling	0.938	0.926

Table 11. Top 10 words in the positive, neutral, and negative classes in sentiment analysis.

	Positive			Neutral			Negative		
	BKK	CM	PK	BKK	CM	PK	BKK	CM	PK
1	Day	Good	Beach	Post	Coffee	Beach	Protest	Go	Island
2	One	Day	Pool	Wat	Post	Post	Police	Still	Covid
3	Go	One	View	2020	Day	View	People	Covid	Go
4	Love	Place	Luxury	New	Morning	Island	Govnment[a]	People	Beach
5	Best	Go	Beautiful	Temple	Go	Pool	Wat	New	Time
6	Time	Beautiful	Bedroom	Day	Time	Chalong	Covid	Year	Town
7	Food	Love	One	Go	Good	Resort	Go	Night	Hotel
8	Good	Visit	Sea	One	Cafe	Locate	Traffic	Last	New
9	City	Time	Island	Hotel	Wat	House	Pro-democracy	Use	Year
10	See	Best	private	Protest	One	Apartment	Rally	Province	Pandemic

[a] Many tweets intentionally referred "government" as "govnment."

Table 12. Top 10 tourism-related words in the positive, neutral, and negative classes in sentiment analysis.

	Positive			Neutral			Negative		
	BKK	CM	PK	BKK	CM	PK	BKK	CM	PK
1	Night	Temple	Beach	Temple	Coffee	Beach	Protest	Night	Island
2	Temple	Festival	Pool	Wat	Morning	View	Police	COVID	COVID
3	Street	Life	Luxury	Night	Life	Island	Govnment[a]	Province	Beach
4	Market	Night	Beautiful	Protest	Garden	Pool	Traffic	Buddhist	Town
5	Wat	Doi	Bedroom	Art	Krathong	Chalong	Monarchy	Elephant	Pandemic
6	View	Waterfall	Sea	Street	Cafe	Resort	Reform	Festival	Ghost
7	Restaurant	Mountain	Island	Market	Date	Sea	Road	Airport	Industry
8	Road	Coffee	Sale	Park	Park	Patong	Pro-democracy	Protest	Patong
9	Bar	Wat	Apartment	Bar	Wat	Bedroom	Chaos	Sexism	Refund
10	Friend	Restaurant	Bathroom	Station	Airport	Kata	COVID	Crowd	Hurt

[a] Many tweets intentionally referred "government" as "govnment."

Table 13. Top 10 words in the "to visit" and "not to visit" classes obtained in the intention analysis.

	To Visit			Not to Visit		
	BKK	CM	PK	BKK	CM	PK
1	Go	Go	Go	Post	Day	Beach
2	Next	Back	Island	Protest	Good	Pool
3	Want	Visit	Beach	Day	One	View
4	Year	Day	Take	Wat	City	Island
5	Miss	Place	Day	One	Year	Bedroom
6	Back	One	Visit	2020	Time	Luxury
7	Visit	Time	Want	New	Place	One
8	Time	Want	Time	Go	Go	Sea
9	See	Take	Year	Time	Coffee	Beautiful
10	Day	See	Soon	City	Morning	Locate

Table 14. Top 10 words in the "to visit" and "not to visit" classes obtained in the intention analysis of tourism-related words.

	To Visit			Not to Visit		
	BKK	CM	PK	BKK	CM	PK
1	Flight	Home	Island	Protest	Coffee	Beach
2	Temple	Temple	Beach	Wat	Life	Pool
3	Home	Plan	Visit	Temple	Festival	View
4	Friend	Mountain	Want	Night	Temple	Bedroom
5	Wait	Car	Soon	Street	Event	Sea
6	Night	Work	Need	Market	Wat	Beautiful
7	COVID	Family	Miss	Police	Doi	Apartment
8	Visit	Everything	Back	Bar	Krathong	Resort
9	Life	Street	Hope	Restaurant	Garden	House
10	Plan	Doi	Like	Park	Airport	Patong

found in the "not to visit" class. Some tweets about visits to tourist destinations indicated no intention to visit again.

Table 14 presents the results of a detailed analysis of tourism-related words in the intention analysis. The themes of the "to visit" class (i.e., tourist and natural destinations) were similar to those of positive-sentiment class in the sentiment analysis. However, in the "to visit" class, the term "COVID" could be interpreted as being the only obstacle stopping tourists from visiting Thailand.

Furthermore, we also perform further analysis by illustrating the term in each type of sentiment and intention of the traveler who want to visit and not to visit the focused research areas by using WordCloud [20]. The result of WordCloud for each Bangkok, Chiang Mai, and Phuket are shown in Figures 6, 7, and 8, respectively.

4. Discussion

4.1. Sentiment analysis

Although SVM achieved the highest accuracy on all three datasets, the accuracy on the Chiang Mai dataset was below 70.0% because fewer tweets for training were available than in the Bangkok and Phuket datasets. If the number of tweets in the Chiang Mai dataset were equal to or

Figure 6. WordCloud result for Bangkok. Positive sentiment (a); Negative sentiment (b); Neutral sentiment (c); Intension not to visit (d); and Intension to visit (e).

greater than the number of tweets in the Bangkok and Phuket datasets, the accuracy of SVM would improve. The result of this analysis was also inline with the studies of Ahmad et al. [14] and Sontayasara et al. [21] based on the positive view for the used of SVM classification in Sentiment analysis. Furthermore, Chen et al. [22] also purposed the uses of the combination between Convolutional Neural Networks (CNN) and SVM to conduct sentiment analysis which can also yield the effective sentiment analysis.

Our analysis aligned with that of Kuhamanee et al. [18], who found that among the decision tree, naïve Bayes, SVM, and artificial neural network (ANN), ANN provided the highest accuracy at 80.33%, followed by SVM at 80.11%. Omitting deep learning algorithms, such as ANN, which were not considered in the present study, SVM was the best machine learning method in both our study and the aforementioned study [18].

The CART and random forest algorithms obtained the same accuracy on the Chiang Mai dataset because the random forest algorithm determines the votes of multiple decision trees as an ensemble. The smaller dataset for Chiang Mai than that for Bangkok and Phuket reduced the complexity of the trees and thus the accuracy of the algorithm.

4.2. Intention analysis

Although SVM achieved the highest intention accuracy on two out of the three datasets, the Bangkok dataset was larger than the Phuket dataset. Providing larger datasets is expected to increase the performance of both the CART and random forest algorithms beyond that of SVM. To support this expectation, CART and random forest achieved accuracies of 94.3% and 95.4%, with the F1-score of 0.933 and 0.950 respectively, in the intention analysis of the Bangkok dataset, whereas SVM achieved an accuracy of only 92.0% with the F1-score of 0.886.

In the intention analysis of a large dataset, the CART and random forest algorithms can outperform SVM because words related to intention are more direct than sentiments ("*Want*," "*Visit*," and "*Miss*"). The

Figure 7. WordCloud result for Chiang Mai. Positive sentiment (a); Negative sentiment (b); Neutral sentiment (c); Intension not to visit (d); and Intension to visit (e).

random forest algorithm simply determines the ensemble votes of multiple decision trees with different configurations [23].

4.3. Word used for sentiments

As displayed in Table 11, the positive-sentiment analysis mainly yielded positive adjectives, such as "*Best*," "*Good*," and "*Beautiful*." The use of positive adjectives is one of the sentiment conditions listed in Table 2. Words such as "*Food*," "*Beach*," "*Pool*," "*Bedroom*," and "*Luxury*" reveal that Twitter users enjoy Thai food, tourist destinations, and hospitality. In contrast, many negative sentiments referred to the political situation in Thailand (e.g., "*Protest*," "*Police*," and "*Pro-democracy*") and the COVID-19 pandemic (e.g., "*COVID*" and "*Pandemic*").

Kuhamanee et al. [18] reported that positive impressions of Bangkok included *"nightlife activity, temple, and historical sites, Thai cuisine, and nature."* Nightlife, tourist destinations, food, and natural attractions were also found in the positive-sentiment class across the three locations in the present study (see Table 12). However, the negative sentiments identified by Kuhamanee et al. [18] (Thai culture) differed from those in our study (COVID-19 and political tensions in Thailand). These differences can be explained by the time period of our datasets (July–December 2020), which corresponds to the first wave of the COVID-19 pandemic.

One surprising observation was the frequent reference to coffee-related words ("*Coffee*," "*Morning*," and "*Café*") in the positive and neutral sentiments about Chiang Mai. This result supports the reputation of Chiang Mai as a province with a high degree of coffee culture [24].

4.4. Word used for intentions

Intentions to visit Thailand were commonly expressed as "*Go*," "*Want*," "*Visit*," and "*Miss*" (*see* Table 13). Among the top 10 intentional words were "*Next*," "*Year*," "*Time*," and "*Soon*," which were possibly

Figure 8. WordCloud result for Phuket. Positive sentiment (a); Negative sentiment (b); Neutral sentiment (c); Intension not to visit (d); and Intension to visit (e).

derived from "*next year*," "*next time*," and "*soon*," indicating a strong eagerness to visit Thailand as soon as possible. However, visiting Thailand was prohibited during the second and third waves of COVID-19.

Based on the process for labeling the intentions in Table 3, the "not to visit" class included many of the words found in the "to visit" class and positive-sentiment class (*see* Table 11). Thus, there was no distinction between "intention never to visit Thailand" and "lack of willingness to visit Thailand." A more detailed analysis would expand the intention analysis from two to three classes, as performed in the sentiment analysis.

5. Conclusions

In this study, English-language tweets related to Thai tourism were manually labeled and assigned to sentiment and intention classes. The accuracies of three machine learning algorithms (CART, random forest, and SVM) in predicting the sentiments and intentions of tweets were then compared. Then, the top 10 words found in each sentiment and intention class were analyzed to gain insights and provide informed suggestions.

Among the three machine learning algorithms, SVM most accurately detected the sentiments and intentions of the English-based tweets mentioning Bangkok, Phuket, and Chiang Mai. Oversampling tends to provide better results than undersampling. However, for the number of analyzed terms unigrams can even provide higher accuracy than bigrams. The top 10 words implied that Twitter users have positive sentiments toward food, tourist destinations, and hospitality in Thailand, but negative sentiments toward the political situation in Thailand and the COVID-19 pandemic. It is likely that these negative impressions have greatly affected the Thai tourism industry.

This study demonstrates that machine learning algorithms can effectively predict the sentiments and intentions of source material. After separating the text into sentiment and intention classes, the word usage in each class can help to identify tourists' sentiments and intentions to visit.

However, this study has some limitations. Although Twitter data are useful for extracting general sentiments, people share their opinions on platforms other than Twitter. Other online platforms include (but are not limited to) Facebook, Instagram, message boards, and personal blogs. Moreover, offline opinion sharing may never reach publicly accessible sources. In addition, because English is considered the *lingua franca* of the world, opinions on Thai tourism from non-English Twitter users were excluded from this study, which focused on English-language tweets. Therefore, the research results are limited to Twitter-accessible English-based tourists and are not generalizable to the sentiments of all tourists. Finally, Thailand has numerous tourist destinations outside Bangkok, Chiang Mai, and Phuket. Destinations such as Hua Hin, Nakhon Ratchasima (Khorat), Rayong, and Kanchanaburi were excluded from the present study.

5.1. Suggestions for restoring Thai tourism

To restore tourism, the effects of the two main concerns of tourists, namely, COVID-19 and the political situation, should be reduced to avoid mentioning the government. After the COVID-19 pandemic recedes, the following four aspects of tourism should be promoted: destinations, nature, food, and nightlife.

Declarations

Author contribution statement

Natt Leelawat: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Sirawit Jariyapongpaiboon; Arnon Promjun; Samit Boonyarak: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Kumpol Saengtabtim; Ampan Laosunthara; Alfan Kurnia Yudha: Analyzed and interpreted the data; Wrote the paper.

Jing Tang: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by the Thailand Science Research and Innovation Fund Chulalongkorn University (CU-FRB65-dis (22)-147-21-13).

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] A.J. Rodriguez-Morales, V. Gallego, J.P. Escalera-Antezana, C.A. Méndez, L.I. Zambrano, C. Franco-Paredes, S. Cimerman, COVID-19 in Latin America : the Implications of the First Confirmed Case in Brazil, Trav. Med. Infect. Dis. (2020).
[2] O.B. Group, Oxford Business Group, 2016.
[3] SCBTV, "SCB," ed: SCBTV.
[4] l. lAEVEN, "Vox," ed: CEPR Policy Portal.
[5] M. Roman, A. Niedziółka, A. Krasnodĺebski, Respondents' involvement in tourist activities at the time of the COVID-19, Sustainability 12 (2020) 9610.
[6] P.A. Valdivia, P.L. Arteaga, M.E. Escortel, C.S. Monge, R.J. Villares, Analysis of complaints in primary care using statistical, Rev. Calid. Assist. 24 (2009) 155–161.
[7] D. Flores-Ruiz, A. Elizondo-Salto, M.d. l.O. Barroso-González, Using social media in tourist sentiment analysis: a case study of andalusia during the Covid-19 pandemic, Sustainability 13 (2021) 3836.
[8] M.A. Sharafuddin, Types of tourism in Thailand, E-review of Tourism Research 12 (2015).
[9] R. Henkel, P. Henkel, W. Agrusa, J. Agrusa, J. Tanner, Thailand as a tourist destination: perceptions of international visitors and Thai residents, Asia Pac. J. Tourism Res. 11 (3) (2006) 269–287.
[10] C. Suttikun, et al., Sociodemographic and travel characteristics affecting the purpose of selecting Bangkok as a tourist destination, Tourism Hospit. Res. 18 (2) (2018) 152–162.
[11] F. Mehraliyev, I.C.C. Chan, A.P. Kirilenko, Sentiment Analysis in Hospitality and Tourism: a Thematic and Methodological Review, International Journal of Contemporary Hospitality Management, 2021.
[12] Y. Lu, Q. Zheng, Twitter public sentiment dynamics on cruise tourism during the COVID-19 pandemic, Curr. Issues Tourism 24 (7) (2021) 892–898.
[13] A. Hasan, S. Moin, A. Karim, S. Shamshirband, Machine learning-based sentiment analysis for twitter accounts, Math. Comput. Appl. 23 (1) (2018) 11.
[14] M. Ahmad, S. Aftab, I. Ali, Sentiment analysis of tweets using svm, Int. J. Comput. Appl. 177 (5) (2017) 25–29.
[15] R. Ahuja, A. Chug, S. Kohli, S. Gupta, P. Ahuja, The impact of features extraction on the sentiment analysis, Procedia Comput. Sci. 152 (2019) 341–348.
[16] S. Triukose, et al., Effects of public health interventions on the epidemiological spread during the first wave of the COVID-19 outbreak in Thailand, PLoS One 16 (2) (2021) e0246274.
[17] S.S. Satchidananda, J.B. Simha, Comparing Decision Trees with Logistic Regression for Credit Risk Analysis, International Institute of Information Technology, Bangalore, India, 2006.
[18] T. Kuhamanee, N. Talmongkol, K. Chaisuriyakul, W. San-Um, N. Pongpisuttinun, S. Pongyupinpanich, Sentiment analysis of foreign tourists to Bangkok using data mining through online social network, in: IEEE 15th International Conference on Industrial Informatics, IEEE, 2017, pp. 1068–1073.
[19] A. Aizawa, An information-theoretic perspective of tf–idf measures, Inf. Process. Manag. 39 (1) (2003) 45–65.
[20] L. Oesper, D. Merico, R. Isserlin, G.D. Bader, WordCloud: a Cytoscape plugin to create a visual semantic summary of networks, Source Code Biol. Med. 6 (1) (2011) 1–4.
[21] T. Sontayasara, et al., Twitter sentiment analysis of Bangkok tourism during COVID-19 pandemic using support vector machine algorithm, J. Disaster Res. 16 (1) (2021) 24–30.
[22] Y. Chen, Z. Zhang, Research on Text Sentiment Analysis Based on CNNs and SVM, IEEE, 2018, pp. 2731–2734.
[23] A. Paul, et al., Improved random forest for classification, IEEE Trans. Image Process. 27 (8) (2018) 4012–4024.
[24] P. Tanitnon, BK Magazine Online, 2015.