The Analysis of Tourism Attitudes using Natural Language Processing Techniques: A Case of Malaysian Tourists

Md Tareq Bin Hossain[1] and Ruchdee Binmad[2]*

[1]Thammasat Business School, Tha Prachan, Bangkok, 10200, Thailand

[2]Research Center for Business Intelligence and Analytics, Faculty of Management Sciences, Prince of Songkla University, 90110, Thailand

* Corresponding author. E-mail address: ruchdee.bi@psu.ac.th

Received: 10 November 2023; Revised: 8 July 2024; Accepted: 6 September 2024; Available online: 18 September 2024

Abstract

The spread of COVID-19 has had a significant impact on all facets of the global tourism sector, particularly in Thailand, one of the world's leading travel destinations. At the height of the epidemic, many countries imposed a nationwide lockdown, prohibiting all citizens from leaving the country and all foreign tourists from entering. This led to a global shutdown that significantly affected the daily lives of billions of people and seriously impacted the travel and tourism industry. After a two-year hiatus due to the epidemic, the situation eased and the lockdown restrictions were lifted. An interesting question is how visitors' attitudes and preferences changed when compared to the time before the outbreak. This study attempts to answer this question by focusing on Malaysian visitors' attitudes and perceptions toward destinations in southern Thailand. The study examines the perceptions of Malaysian Twitter (now **X**) users from three areas in Malaysia; Kedah, Perlis, and Kuala Lumpur, by employing Natural Language Processing (NLP) techniques such as sentiment analysis and topic modeling. Then, tweet data before and after the lockdowns were gathered, analyzed, and compared. For sentiment analysis, it was found that, when neutral tweets were disregarded, results both before and after the COVID-19 lockdowns revealed that the attitudes conveyed by Malaysian tourists were overall positive especially a territory and a state that are more far away from Thailand, i.e., Kuala Lumpur and Kedah. The results from the topic modeling analysis showed a meaningful distinction between before and after the COVID-19 lockdowns. Practical suggestions are offered for tourism policymakers to identify and address both the strengths and weaknesses of tourism development in Southern Thailand.

Keywords: Sentiment analysis, Topic modeling, COVID-19, Tourist's attitudes, Malaysian Tourists

Introduction

Globally, Thailand ranks among the world's leading travel destinations. The variety of cultural uniqueness: from exotic food to beautiful landscapes and rich cultural heritage has driven the tourism sector to play a significant role in driving the Thai economy. There were almost 40 million international visitors to Thailand in 2019, contributing to an estimated 2 trillion Baht (11% of GDP) and providing employment for more than 7 million people (20% of total employment). Among all foreign tourists, China and Malaysia were the largest sources of Thailand's tourism market, with Chinese visitors accounting for 29% of arrivals and Malaysians being 8% of total arrivals (Tepanon, et al., 2021). In the south of Thailand Malaysian tourists dominated, especially in the areas near the Thai-Malaysian border such as Hat Yai City in Songkhla province (Bunnoon et al., 2021; Praprom & Laipaporn, 2021).

Unfortunately, in late 2019, the COVID-19 pandemic struck (World Health Organization, 2020; Centers for Disease Control and Prevention, 2023). Due to its magnitude, the outbreak of COVID-19 created an unprecedented public health, social, and economic emergency, causing immeasurable losses and affecting the normal lives of billions of people around the world. The worldwide shutdown instituted to prevent the spread of

COVID-19 had an especially negative effect on the travel and tourism industries. At the peak of the pandemic, both Thailand and Malaysia imposed a nationwide lockdown prohibiting all citizens from leaving the country and all foreign tourists from entering the country (Wongmonta, 2021; Wun'Gaeo & Wun'Gaeo, 2021). Consequently, the total number of foreign tourists entering Thailand has tremendously dropped by 83% in 2020 with a further disastrous nosedive in arrivals of 93% by the end of 2021 (Wongmonta, 2021).

A two-year break due to the COVID-19 pandemic could greatly influence all aspects of the Thai tourism domain including the changing of tourists' behavior and preferences when the pandemic situation improved. Therefore, identifying such changes is relevant to not only Thailand but also to the global tourism industry. The important research questions associated with this challenge include: How would Malaysian tourists' attitudes and opinions towards destinations in Southern Thailand look like after the pandemic? And how different were they when compared with attitudes and opinions before the pandemic?

The purpose of this study was to find answers to the above questions by analysing a large volume of user-generated data from the Internet, especially from a popular social media platform such as Twitter (now known as **X**). As of April 2022, the Internet reached 63% of the world's population and of this total, over 93% were social media users (Kemp, 2023). Specifically, Twitter's audience accounted for over 368 million monthly active users worldwide and generated 3.47 hundred thousand tweets per minute (Kemp, 2023; Statista, 2022).

Social Network Analysis

Social network analysis offers an extensive set of widely adapted techniques for quantitatively analyzing such networks (Tabassum, et al., 2018). Social network analysis diverges from traditional sociology by focusing on the relationships between agents rather than individual agent attributes. Social network analysis has proven useful for explaining many social phenomena. In a popular publication on social network analysis, Christiakis & Fowler (2009) showed how characteristics and happiness spread through social networks. This strong dependence on people's behavior and their friends (and their friends' friends and their friend's friends' friends...) makes social network analysis an important model for the study of social norms (Elsenbroich & Gilbert, 2014).

Common SNA tasks include identifying the most influential, prestigious, or central actors through the use of statistical measures; identifying nodes and authorities through the use of link analysis algorithms; discovering communities through the use of community detection techniques; and analyzing how information spreads through the network through the use of diffusion algorithms. By extracting knowledge from networks and then using it to solve problems, these activities are of great benefit (Tabassum et al., 2018). Due to the attractiveness of such tasks and the enormous potential offered by this type of analysis, SNA has gained popularity in several fields including biology, economics, and psychology (Tabassum et al., 2018; Binmad & Li, 2018).

Facebook has been utilized by the Korean government for tourist marketing and expansion (Park et al., 2016). The findings suggest that many Korean local governments make good use of social media platforms such as Facebook to disseminate information and engage with citizens to boost the country's tourism industry. That study also used network analysis to demonstrate how Korean local governments' Facebook pages are interconnected within the Facebook system as a component of a smart tourism ecosystem, and how the smart tourism ecosystem enabled by Facebook's offerings is linked to the utilization of Facebook to activate local tourism.

People use social networking services such as Twitter, now **X**, to express their thoughts, report real-life occurrences, and provide a viewpoint on what is going on in the world (Trajkova et al., 2020). As for social media and the big data phenomenon, it is estimated that **X** users generate about 15 GB of data per day. It is extensively utilized by the general public, who use it to voice their thoughts on a wide range of public issues and to express their concerns or complaints to businesses and government authorities (Srivastav et al., 2020). Tourism is considered one of the most important sectors of a country's economy as it generates not only a large amount of revenue but also a massive content on various social media websites.

In the domain of tourism management, SNA was used to provide an overview of how quantitative analysis approaches might be applied to improve tourism destination competitiveness (Valeri & Baggio, 2021). This study uncovered the network of relationships that can provide powerful leverage for tourism organization managers to increase information flow and take advantage of opportunities where that flow can have the greatest impact on regulatory or business operations.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is a rapidly evolving field that enables machines to understand, generate, and translate human language. Key NLP techniques include syntactic, semantic, and pragmatic analysis, which focus on sentence structure, word meaning, and context respectively(Bayer et al., 2021). Recent research explores NLP's applications in sociology, interactive language models, and machine learning tasks. Németh and Koltai (2023) highlight NLP's potential and challenges in quantitative text analysis within sociological research. Wang et al. (2023) propose a framework for Interactive NLP (iNLP), emphasizing its applications and ethical considerations. Machine learning plays a crucial role in NLP tasks like sentiment analysis and text classification, requiring large datasets (Sharma, 2023). Khan (2021) evaluates Voice note-taking technology, highlighting its transcription accuracy and the need for improvements in noisy environments.

NLP techniques have significantly improved user experience in the tourism industry by providing personalized recommendations and enhancing service quality. Camilleri and Troise (2023) highlight AI-powered chatbots' benefits and limitations, noting their efficiency in handling customer inquiries while acknowledging challenges in complex scenarios. Sentiment analysis of homestay reviews helps stakeholders understand customer satisfaction and identify areas for improvement (Vajpai & Pattanaik, 2022). Binabdullah and Tongtep (2021) investigate NLP techniques used in tourism suggestion systems, highlighting their role in processing tourist attitudes and providing tailored recommendations. Liang et al. (2023) propose an advanced recommendation model using a dual-channel DQN to further enhance accuracy and personalization. These studies collectively demonstrate NLP's transformative impact on making tourism services more interactive, personalized, and user-friendly.

Sentiment Analysis

Sentiment analysis is a common text classification method that analyzes text data to determine its underlying sentiment (positive, negative, or neutral) (Birjali et al., 2021). This technique has broad applications, particularly in marketing. Marketing teams can use sentiment analysis to gauge public opinion on new products or determine the popularity of existing ones by analyzing social media reviews. There are three main streams of sentiment definition in research: opinion-based, feeling-based, and a combination of both (Ge et al., 2018).

Given its potential usefulness in a wide range of fields, sentiment analysis has attracted several academics (Ge et al., 2018; Bonta et al., 2019). Existing research has focused extensively on online consumer reviews for products, hotels, movies, etc. Individual customers desire to know the opinions of existing product users before making a purchase, as well as the opinions of others on a political candidate before making a voting decision. Sentiment analysis has demonstrated its influence on social media platforms like Twitter and Facebook to comprehend user tweets and behavior (Zimbra et al., 2018).

Traditionally, gathering public opinion involved surveys, polls, and focus groups. With the rise of social media, businesses can now easily access and analyze public sentiment online. This automation reduces costs and saves time. Sentiment analysis tools provide real-time data on customer opinions, helping businesses understand their needs, expand their customer base, and exceed expectations.

Lwin et al. (2020) analyzed over 20 million Twitter tweets from January 28 to April 9, 2020, focusing on COVID-19-related emotions. Using a lexical approach and the CrystalFeel algorithm, the study classified emotions (fear, anger, sadness, joy) and their associated reasons. Key findings include growing concerns about testing and medical supplies, a shift in anger topics, and the impact of personal loss on sadness. Stella et al. (2020) also explored emotional and social consequences using Twitter data, developing the MERCURIAL framework for emotional profiling.

In Mishra et al. (2021), researchers examined Twitter data for the tourism industry, focusing on the hospitality and healthcare sub-domains. By using topic modeling techniques and the Valence Aware Dictionary for Sentiment Reasoning (VADER), the sentiment of around 20,000 tweets has been determined. To predict and categorize people's emotions, a state-of-the-art deep learning classification model with varying epoch sizes of the dataset was utilized. Sentiment analysis based on different deep-learning techniques and architectures was used in the study by (Martín et al., 2018) to evaluate sentiments about the reviews that tourists published online and studied how new tourists planned their trips using data from the TripAdvisor and Booking.com websites. In a similar study by Yu et al. (2019), a sentiment analysis was conducted on the Japanese tourism website, 4Travel, to determine how users perceive Chinese attractions. The findings help to elucidate the intrinsic characteristics of the Japanese language and also provide practical usage guidelines within the field of sentiment analysis of Japanese online tourism reviews.

Using Twitter as a source of data has also been performed by Ainin et al. (2020). This study aimed to examine halal tourism trends globally and the findings showed that Japan is the most-tweeted-about halal tourist destination, followed by Malaysia and Indonesia. In Pano and Kashef (2020), researchers used VADER as a sentiment analysis method to analyze bitcoin-related tweets during the outbreak of COVID-19 by comparing 13 different text preparation algorithms to correlate sentiment scores of tweets with bitcoin prices.

Topic Modeling Analysis

Other than sentiment analysis, topic modeling is also employed as a text analytics technique, to derive information from massive amounts of textual content. Topic modeling reflects the information in the set using an algorithm of machine learning and natural language processing (George et al., 2017). As the use of social media increases, so do the number of studies that attempt to mine the platforms for useful insights. The primary goals of topic modeling are (1) to find hidden topics in text data by clustering related word groupings and (2) to categorize the document according to the detected topic. Topic modeling has a positive impact on categorization by grouping similar phrases into topics and detecting patterns in social media (Kherwa & Bansal, 2019). It is often used in recommender systems with similarity measures (Jeong et al., 2019; Anupama & Elayidom, 2022).

There are two basic topic models: Latent Semantic Analysis (LSA, or Latent Semantic Index (LSI)) and Latent Dirichlet Allocation (LDA). LSA can illustrate the strong connection between documents and expressions. LSA performs well in categorizing short sentences like tweet data according to several previous studies (George et al., 2017; Neogi et al., 2020; Mujahid et al., 2021). LDA is used to look for structures, themes, and patterns in the texts and to determine how these themes are connected. Large amounts of data are efficiently categorized into groups based on patterns and attributes (Casillano, 2022). The study by Kherwa & Bansal (2019) showed that LDA is an effective topic modeling technique that can be used for classification, feature selection, and information retrieval. Table 1 shows the comparison of these two topic modeling techniques.

Table 1 Comparison of Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA)

Aspect	Latent Semantic Analysis (LSA)	Latent Dirichlet Allocation (LDA)
Methodology	Utilizes Singular Value Decomposition (SVD) to reduce dimensionality and uncover latent structure.	Uses a generative probabilistic model to represent documents as mixtures of topics, each defined by word distributions.
Basic	Linear algebra-based technique.	Probabilistic model based on Bayesian inference.
Core Concept	Identifies patterns in the relationships between terms and documents by transforming term-document matrices into lower-dimensional space.	Models documents as a mixture of topics where each topic is a distribution over words.
Advantages	Simple, fast, and effective for capturing semantic similarity. Improves information retrieval and semantic understanding.	Provides interpretable and nuanced topic distributions. Captures the thematic structure of large text corpora.
Recent Enhancements	Integration with deep learning and neural networks for improved scalability and semantic understanding (Gadamshetti et al., 2022).	Hybrid models with neural networks for better coherence (Wolpe & Waal, 2019).

Numerous studies on topic modeling have been conducted. In the study by Jeong et al. (2019) on the topic modeling and sentiment analysis of social media data, the authors of this research introduced a strategy for mining opportunities to discover potential new products. A customer-centric opportunity algorithm finds the potential value and improvement direction of each product subject based on the significance and satisfaction of product subjects. In the area of policy research, topic modeling for textual analysis has been studied by Isoaho et al. (2021). The results showed that different mixed-method research designs are appropriate when combining topic modeling with the two groups of methods enabling researchers to apply policy theories and concepts to much larger sets of data.

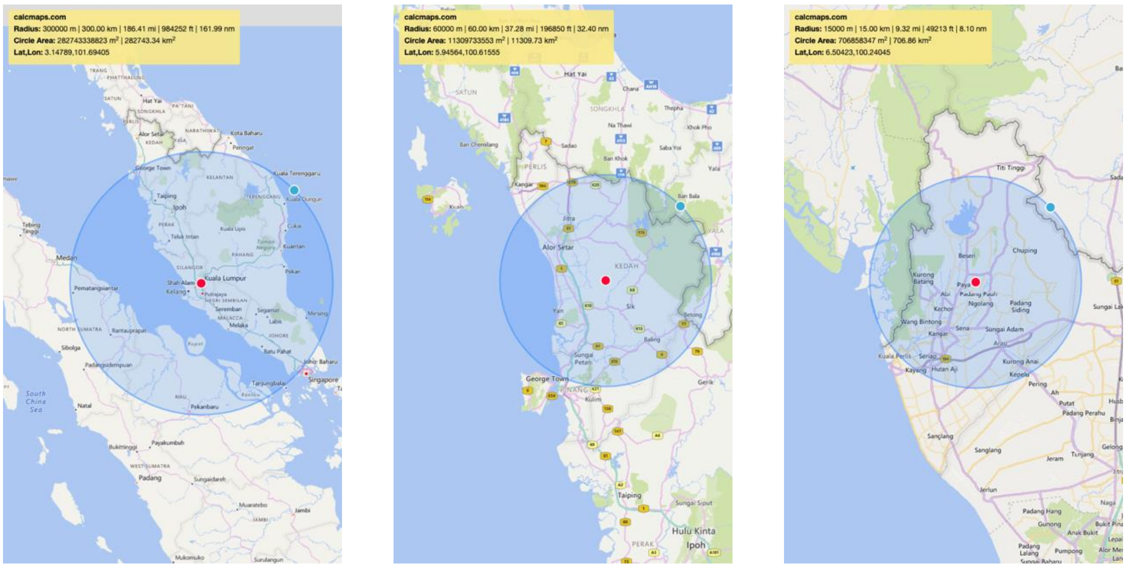
In Abd-Alrazaq et al. (2020), the authors identified the main topics posted by Twitter users related to the COVID-19 period from February 2, 2020, to March 15, 2020, using Latent Dirichlet Allocation (LDA). Twelve topics were identified, ten of which were rated as positive and two as negative; deaths from COVID-19 and increasing racism. The study of the transformation from conventional to online education based on sentiment and topic modeling analyses was conducted by Mujahid et al. (2021). The LSA model was used in this study to identify issues with e-learning, and it was found that the three significant challenges were related to the uncertainty of a campus' planned opening date, children's inability to understand online instruction and inadequate infrastructure for online learning.

This study examined the perceptions of Malaysian Twitter users concerning Southern Thailand tourism, particularly in the areas of Hat Yai City in Songkhla province. The research was carried out using Natural Language Processing (NLP) techniques such as sentiment analysis and topic modeling. Sentiment Analysis is used for the analysis of textual feelings and is sensitive to both polarity (positive/negative) and emotional intensity (strongness) of the scraped tweet data using VADER. Topic modeling is utilized for automatically identifying themes contained in tweets and deriving hidden patterns using Latent Semantic Analysis (LSA). Results were analyzed and compared on real-world tweet data between a year before the lockdown was introduced and 8 months after the lockdown was lifted. The study is structured in a systematic flow, with the literature review described in Section 2, followed by the methodology in Section 3, then the results and discussion in Section 4, and finally the conclusion and potential future applications discussed in Section 5.

Materials and Methods

Dataset Collection

The dataset for this study was collected from the famous Social Network platform, Twitter. Two Malaysian states and one federal territory were chosen as the sources for scraping tweet data, i.e., Kuala Lumpur, Kedah, and Perlis with a radius to bound the search area of 300 Km., 60 Km., and 15 Km. Respectively, as shown in the maps in Fig.1. To obtain the desired tweet data, the search results needed to contain one of these two keywords i.e., “hat yai” or “songkhla”, the city and province located near the Thai-Malaysian border and the travel destinations of choice for many tourists, especially from Malaysia.


calcmaps.com
Radius: 300000 m | 300.00 km | 186.41 mi | 984252 ft | 161.99 nm
Circle Area: 282743338823 m² | 282743.34 km²
Lat,Lon: 3.14789,101.69405
Hat Yai
Alor Setar
Kota Baharu
George Town
Kuala Terengganu
Ipoh
Medan
Kuala Lumpur
Shah Alam
Seremban
Melaka
Pekanbaru
Singapore
Padang
calcmaps.com
Radius: 60000 m | 60.00 km | 37.28 mi | 196850 ft | 32.40 nm
Circle Area: 11309733553 m² | 11309.73 km²
Lat,Lon: 5.94564,100.61555
Hat Yai
SATUN
SONGKHLA
PERLIS
Alor Setar
KEDAH
George Town
Sungai Petani
Taiping
Hulu Kinta
Ipoh
PERAK
calcmaps.com
Radius: 15000 m | 15.00 km | 9.32 mi | 49213 ft | 8.10 nm
Circle Area: 706858347 m² | 706.86 km²
Lat,Lon: 6.50423,100.24045
Titi Tinggi
Beseri
Chuping
Kangar
Arau
Kuala Perlis
Sanglang

[a] [b] [c]

Figure 1 Maps of scraping tweet data by locations [a] Kuala Lumpur, [b] Kedah, and [c] Perlis

Additionally, the one-year tweet data before the COVID-19 lockdown was scraped between December 31, 2018 and December 30, 2019. The period after lockdown has been set to scrap tweet data from the first day of the re-opening of the Thailand-Malaysia border which is April 1, 2022, to November 30, 2022. Table 2 shows the summary of the search query setting for this study.

Twitter has provided its official API (https://developer.x.com/en) for developers who want to scrape public conversations, users, and hashtags available on tweet data. However, Twitter API limits scraping only for the last seven days. Another limitation is that the removal of its free scrape APIs in April 2023 made access to the current version of the Twitter API no longer free.

Table 2 Summary of tweet data search query setting

Parameter	Query
Keyword	“Hat Yai” or “songkhla”
Location	Kuala Lumpur, Kedah, Perlis
Radius	300 KM, 60 KM, 15 KM
Time period before lockdown	Between Dec 31, 2018, and Dec 30, 2019
Time period after lockdown	Between Apr 1, 2022, and Nov 30, 2022

Since this study needed to access historic tweets, the Python wrapper library named “snscrape” was used to obtain raw data from Twitter (JustAnotherArchivist, 2022). Although there are several ways to scrape tweet data, the ability to retrieve tweets using snscape without requiring personal API keys, return thousands of tweets in seconds without any restrictions or request limits, and provide powerful search tools that allow for highly customizable searches shows that the key features of snscrape make the library superior over other methods.

Subsequently, the tweet data for the period before the COVID-19 lockdowns contained 1568 tweets, and after the lockdown contained 494 tweets. Most of the tweets were in Bahasa Malaysia (Malay) language; 69.77% before the lockdown and 80.77% after the lockdown (Table 3).

Table 3 Numbers of scraped tweet data

Location	Radius	Before Lockdown (rows)	After Lockdown (rows)
Kuala Lumpur	300 KM.	1001	367
Kedah	60 KM.	184	88
Perlis	15 KM.	383	39

Description of Scraped Data

The scraped tweet data contains 9 columns that represent different attributes for each tweet. Each column is explained in Table 4.

Table 4 Description of data

Attribute	Description
Unnamed	Row index
User	Twitter’s user name posting the tweet
Location	User’s location in profile
Date Created	Date tweet was created
Number of Likes	Count of likes
Source of Tweet	Where tweet was posted from
Hashtags	User’s hashtags appeared in tweet content
Language	Machine-generated, assume the language of the tweet
Tweet	The text content of the tweet

Text Preprocessing

Obtaining data that is machine-readable is one of the most crucial aspects of data analysis. This can be achieved by removing superfluous data to increase the learning process of classification models. At this stage, raw scraped data collected from Twitter can be preprocessed using Python's NLP toolkit (Bird et al., 2009) including the following steps:

(1) Removing duplicate tweet data: The radiuses of the selected locations might be overlapping suggesting that the same tweet data has appeared more than once.

(2) Translating into English: 67% of Twitter users in Malaysia tweet in Bahasa Malaysia (Malay) language (Kong et al., 2023). The sentiment analysis model used in this study supports text sentences only in English (Hutto & Gilbert, 2014).

(3) Converting to lowercase: Models classify uppercase and lowercase words as different words. Therefore, converting text to lowercase can reduce the effect of classification performance.

(4) Removing URL links and punctuation: URL links and punctuation provide no meaning for the classification performance.

(5) Replacing emojis and emoticons: To maintain consistency and integrity, emojis and emoticons are converted to standardized text representations defined by the Unicode Consortium (https://home.unicode.org/emoji/). For instance, the emoji "😊" is transformed to "smile", and "😭" becomes "cry". Python libraries, such as "emoji" for emojis and "emot" for emoticons, facilitate this transformation.

(6) Tokenizing tweets: Tweets are broken down into smaller tokens so that they can be analyzed by the model.

(7) Stemming and Lemmatization: Words are converted from inflectional forms into their common base form.

(8) Removing stopwords: Discarding words that give no useful information or significance towards the emotion of the text (tweet).

Table 5 Sample tweets before and after preprocessing

Tweet before Preprocessing	Tweet after Preprocessing
It is best to get a massage near Hat Yai yesterday even though it hurts	[‘best’, ‘massage’, ‘near’, ‘Hat Yai’, ‘yesterday’, ‘hurt’]
I'm at Kak Su Orange Seafood in Tambon Samnak Kham, Chang Wat Songkhla	[‘kak’, ‘orang’, ‘seafood’, ‘tambon’, ‘samnak’, ‘kham’, ‘chang’, ‘wat’, ‘songkhla’]
If you want to know the place we have to go if you go to Hat Yai this is it.. floating market. What's up here? There are original Thailand foods and all cheap.. halal don't worry	[‘want’, ‘know’, ‘place’, ‘Hat Yai’, ‘float’, ‘market’, ‘what’, ‘here’, ‘there’, ‘origin’, ‘thailand’, ‘food’, ‘cheap’, ‘halal’, ‘don’, ‘worri’]

VADER Sentiment Analysis

VADER (Valence Aware Dictionary and sentiment Reasoner, https://github.com/cjhutto/vaderSentiment) is a lexicon and rule-based sentiment analysis tool. It can be applied directly to unlabeled text data and is specifically attuned to sentiments expressed in social media. It is fully open-sourced under the MIT License (Hutto & Gilbert, 2014). VADER sentiment analysis is a model used for the analysis of textual feelings, sensitive to both polarity (positive/negative) and emotional intensity (strongness) of a corpus or set of documents. It relies on a lexicon that links lexicological characteristics with so-called emotional values.

VADER uses a dictionary to associate words with emotion intensities, called sentiment scores, to measure the emotion of a word. The score of a text is calculated by adding the intensity scores of each word in a text corpus. The scale used to measure sentiment intensity ranges from -4 to +4, with -4 being the most negative value (very negative) and +4 being the most positive. Meanwhile, 0 (the midpoint) is seen as neutral. The sum of positive, negative, and neutral scores which is then normalized to the overall emotion intensity to map the score between -1 and +1 called a compound score (Mishra et al., 2021; Pano & Kashef, 2020; Hutto & Gilbert, 2014). In a study conducted at the Georgia Institute of Technology, VADER and other widely used and well-regarded sentiment analysis tools were compared and assessed based on their ability to categorize emotions. VADER was found to perform well and generally outperform the others (Bonta et al., 2019; Hutto & Gilbert, 2014). This study utilizes a compound score primarily to understand the sentiments of the users' tweets.

LSA Topic Modeling

One of the machine learning techniques for mining texts is the so-called topic modeling (George et al., 2017; Kherwa & Bansal, 2019). It is a method for extracting hidden patterns in a text corpus and determining the underlying themes of a given text item automatically. Topic modeling differs from rule-based text mining algorithms that utilize regular expressions or dictionary-based keyword searches. It is an unsupervised approach for identifying and tracking a set of words in large amounts of text. Topic modeling uses a probabilistic approach to identify recurring abstract themes within a document collection (Mishra et al., 2021). This study used the Latent Semantic Analysis (LSA) which is also known as the Latent Semantic Index (LSI) model to perform topic discoveries from the existing tweets. LSA can illustrate the strong connection between documents and expressions. Several studies show that LSA performs well in categorizing short sentences (George et al., 2017; Neogi et al., 2020; Mujahid et al., 2021). Compared to other automatic indexing and retrieval techniques, LSA provides equivalent meaning with fewer dimensions while consuming less energy.

LSA is a statistical model of word usage that enables semantic similarity analyses across textual data. The fundamental concept of latent semantic analysis (LSA) is that statistical methods can be used to estimate an underlying or "latent" structure in the distribution of words in texts. In this scenario, documents as well as smaller text pieces such as individual paragraphs or phrases, can be viewed as contexts in which words appear (Salloum et al., 2020). Based on massive corpus studies, LSA creates a high-dimensional vector representation. However, LSA examines co-occurrence across the corpus by employing a predetermined context frame (e.g., the paragraph level). The co-occurrence matrix is then reduced to a smaller number of dimensions using a factor analytic technique (singular value decomposition). The dimension reduction produces similar vectors for words that are used in comparable contexts, even if they are not used in the same context. Their vector representations in LSA, however, would be equivalent (Anandarajan et al., 2019). This enables the comparison of bigger portions of text, such as the meaning of phrases, paragraphs, or whole texts. As a theoretical model and a technique, semantic relationships between linguistic units have been characterized using LSA. The outcomes of its performance on conventional vocabulary and topic matter are comparable to those of humans, it replicates human categorization and word-sorting processes, it simulates lexical priming data for words and passages, and it adequately evaluates textual coherence and learnability of texts.

The LSA model used in this study involves the following steps:

(1) Convert the text corpus into a document-term matrix

The tweet text is converted into a document-term matrix. This is achieved with the bag of words algorithm.

(2) Implement truncated singular value decomposition

The LSA model is predicated on truncated singular value decomposition (SVD). The operation is essential for extracting themes from the provided document collection. It can be expressed mathematically using the following formula:

$$A_{nxm} = U_{nxr}S_{rxr}V_{mxr}^{T}$$

Where A represents the document- term matrix with a count- based value assigned to each document- term pairing. The matrix has $n\,x\,m$ dimensions, with n representing the number of documents and m representing the number of words. U represents the document- topic matrix. In essence, its values indicate the strength of the relationship between each document and its derived topics. The matrix has $n\,x\,r$ dimensions, with n representing the number of documents and r representing the number of topics. S represents a diagonal matrix that evaluates the "strength" of each topic in the document collection. The matrix has $r\,x\,r$ dimensions. And V represents the word- topic matrix. Its values show the strength of the relationship between each word and the derived topics. The matrix has $m\,x\,r$ dimensions.

(3) Encode the words/documents with the extracted topics.

The SVD operation transforms a document-term matrix into a document-topic matrix (U) and a word-topic matrix (V). These matrices are used to identify the terms with the strongest relationship with each topic. This information can be used to decide what each derived topic represents and also to determine which documents belong to which topic.

This study primarily utilizes the Gensim library (https://pypi.org/project/gensim/), an open-source Python library for unsupervised topic modeling, document indexing, retrieval by similarity, and other natural language processing functionalities, using modern statistical machine learning (Rehurek & Sojka, 2010).

Ethical Considerations

All the collected tweets belong to the public domain and are accessible to the public, so no ethical assessment was required. Despite this, the authors adhered to the utmost ethical standards when handling the scraped data; no individual tweets were evaluated or shown in this work. Even though several tweets were obtained, after calculating average sentiment by location, all identifying information and the content of each tweet was removed. The main objective of this research is to identify the underlying sentiments and themes among Malaysian Twitter users about Thailand as a travel destination. If required, the data can be easily accessed again using the procedures mentioned in this study.

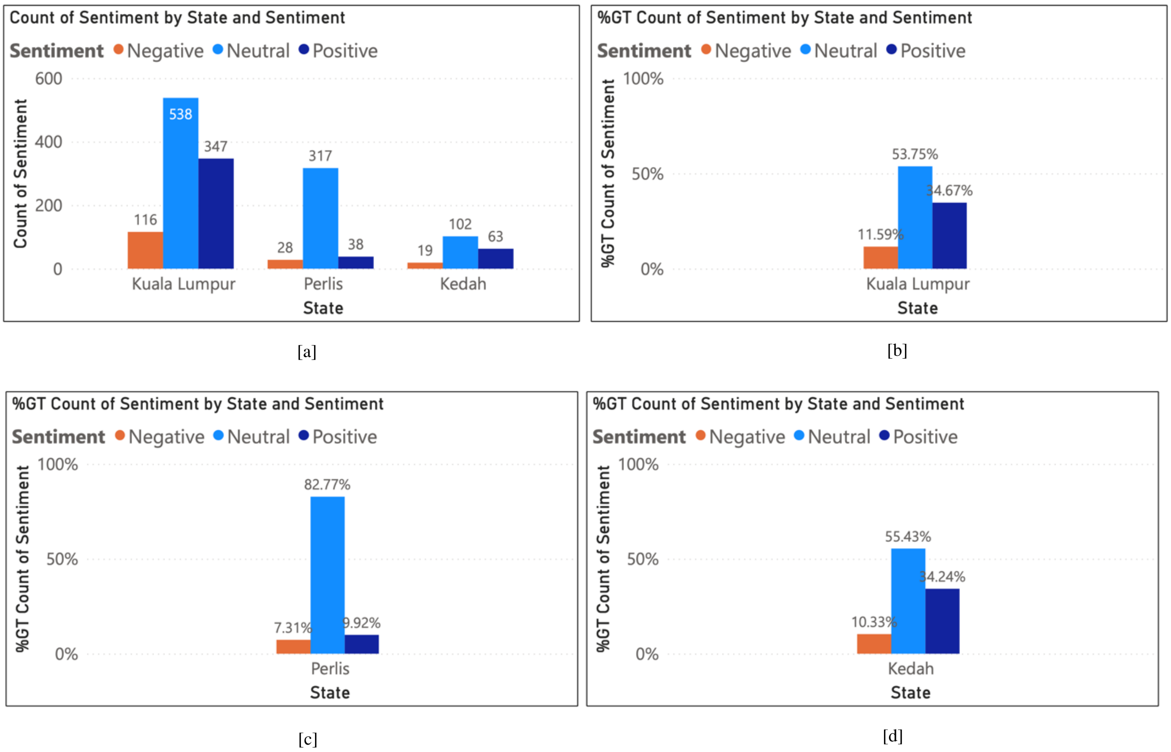
Results

Sentiment Analysis Results

Before the COVID-19 Lockdown

Fig. 2. visualizes the sentiment analysis of each tweet, namely positive, negative, and neutral in terms of the number and percentage of the sentiment polarities. This can be seen that even the majority of tweets are neutral, however, the number of positive tweets is greater than that of negative tweets in both states and a territory. Specifically, positive tweets from users in Kuala Lumpur and Kedah are almost highly the same. The number of users in Perlis who tweeted positively is significantly less than those from Kuala Lumpur territory and Kedah state.

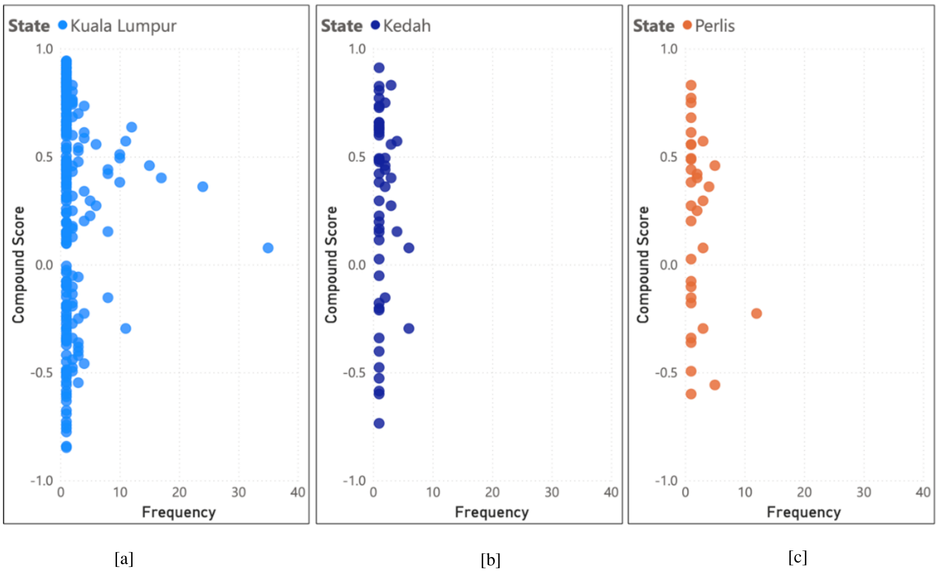
The frequencies of compound scores of positive and negative tweets are shown in Fig. 3. The compound scores of tweets from Kuala Lumpur, Kedah, and Perlis lie in the range [-0.84, 0.94], [-0.74, 0.91], and [-0.60, 0.83] respectively. Additionally, the highest frequencies of the positive tweets are between [0.08, 0.64], [0.08, 0.57], and [0.08, 0.57] from Kuala Lumpur, Kedah, and Perlis respectively. While, the highest frequencies of the negative tweets are between [-0.46, -0.15] from Kuala Lumpur, [-0.30, -0.15] from Kedah, and [-0.56, -0.23] from Perlis.



[a] [b]

[c] [d]

Figure 2 [a] Total number [b]-[d] Percentage of sentiment distribution of tweets before the COVID-19 lockdown across the 2 states and a territory of Malaysia

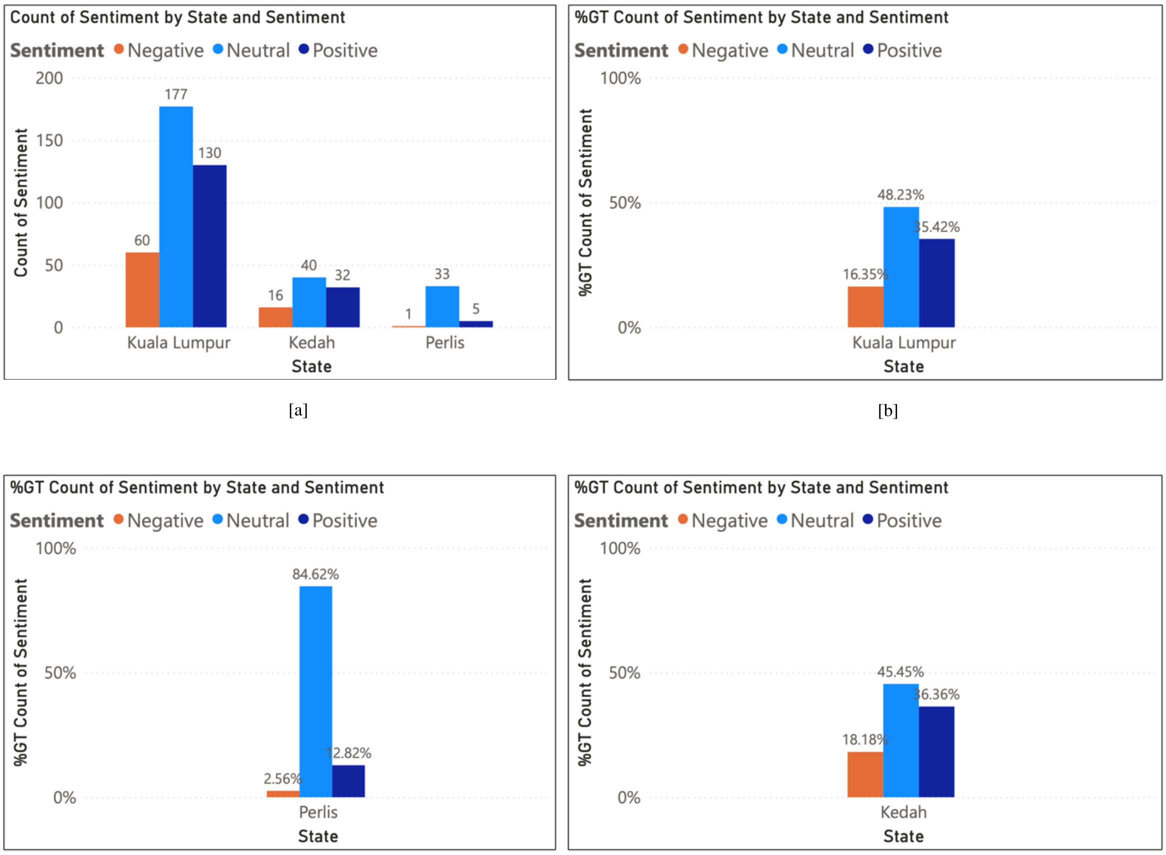


[a] [b] [c]

Figure 3 Frequency of compound scores of tweets before the COVID-19 lockdown from [a] Kuala Lumpur territory [b] Kedah state and [c] Perlis state

After the COVID-19 Lockdown

Similar to the results before the COVID-19 lockdown as visualized in Fig. 4, most of the tweets after the lockdown from both states and the territory were neutral and the number of positive tweets was also higher than the number of negative ones. Furthermore, the number of positive tweets after the lockdown was significantly more than before the lockdown in both states and the territory.



[a] [b]

[c] [d]

Figure 4 [a] Total number [b]-[d] Percentage of sentiment distribution of tweets after the COVID-19 lockdown across the 2 states and a territory of Malaysia

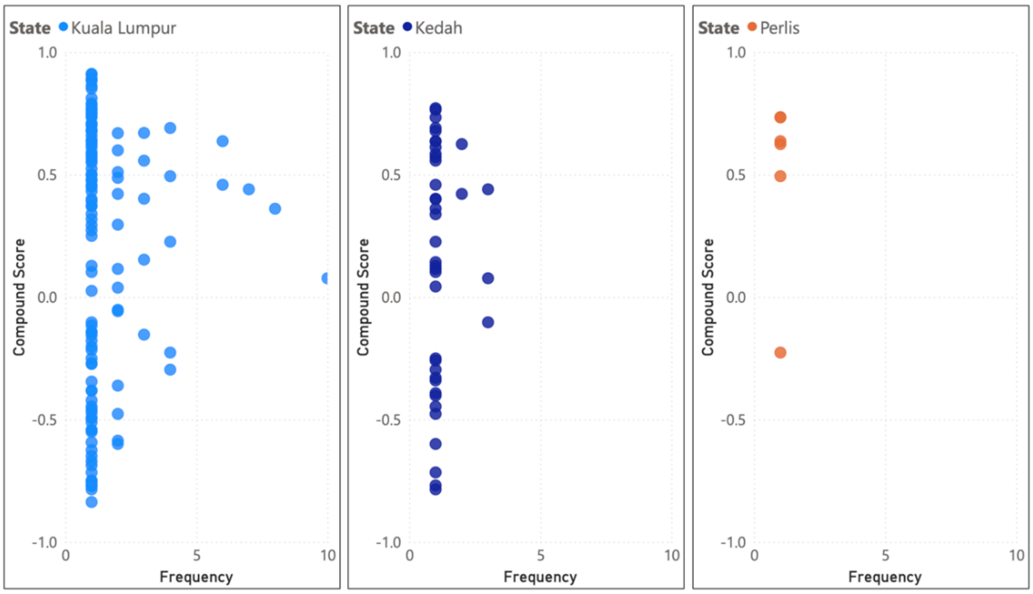
As shown in Fig. 5, tweets from Kuala Lumpur, Kedah, and Perlis have compound scores that lie in the range [-0.84, 0.91], [-0.78, 0.77], and [-0.23, 0.74] respectively. The highest frequencies of positive and negative tweets from Kuala Lumpur are between [0.23, 0.69] and [-0.30, -0.15] respectively. Tweets from Kedah have the highest frequencies of positive compound scores lying between [0.08, 0.44] and of negative compound scores only of -0.10. There are no the highest frequencies of positive and negative tweets from Perlis as all compound scores are equally distributed with one score.

Topic Modeling Results

Before the COVID-19 Lockdown

To optimize the results of the topic modeling analysis, the coherence scores are calculated to determine the best number of topics that should be extracted from each tweet dataset. In terms of semantic value, the coherence value indicates how similar the words of each topic are, with a higher value indicating greater similarity.

As mentioned in Methods and Materials Section 3, this study utilizes LSA to assess Malaysian tourists' general public attitudes concerning travelling to Southern Thailand. The top five most contributing keywords in tweets from Kuala Lumpur territory, Kedah state, and Perlis state are listed in Tables 6, 7, and 8 respectively.



[a] [b] [c]

Figure 5 Frequency of compound scores of tweets after the COVID-19 lockdown from [a] Kuala Lumpur territory [b] Kedah state and [c] Perlis state

Table 6 Words in tweets from Kuala Lumpur territory before the COVID-19 lockdown with the strongest association to the derived topics

Topic/Word	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	hat yai	songkhla	trip	want	thailand
Topic 1*	best	hat yai	songkhla	thailand	want
Topic 2	trip	want	2019	seat	bus

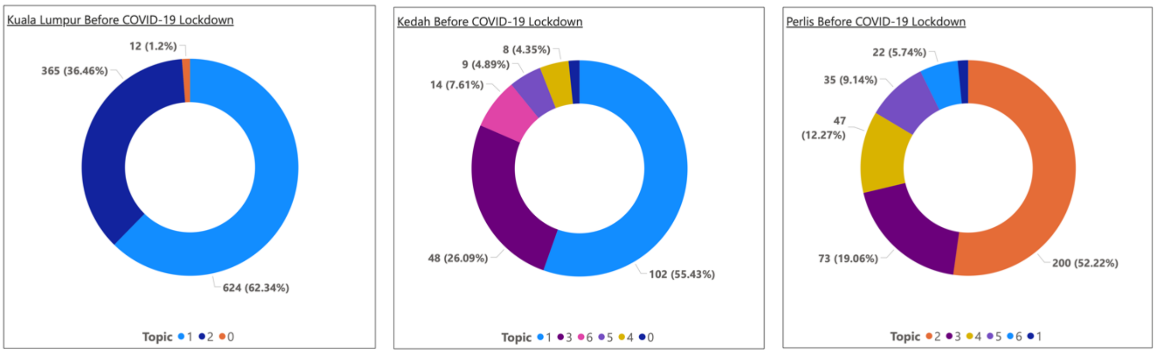
Table 7 Words in tweets from Kedah state before the COVID-19 lockdown with the strongest association to the derived topics

Topic/Word	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	hat yai	songkhla	want	thailand	order
Topic 1*	best	hat yai	thailand	songkhla	order
Topic 2	hat yai	van	tut	radio	damag
Topic 3	want	don	dock	crazi	like
Topic 4	time	what	ask	junction	har
Topic 5	har	boyfriend	tomorrow	month	son
Topic 6	like	order	size	hat yai	don

Table 8 Words in tweets from Perlis state before the COVID-19 lockdown with the strongest association to the derived topics

Topic/Word	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	sadao	immigration	border	hat yai	songkhla
Topic 1	hat yai	songkhla	sadao	dan nok	immigr
Topic 2*	sadao	songkhla	border	immigration	and
Topic 3	hat yai	sadao	dan nok	zainab	bangdon
Topic 4	hat yai	sadao	thailand	dan nok	border
Topic 5	thailand	songkhla	trang	patthalung	roadtrip
Topic 6	songkhla	boyfriend	har	sadao	hat yai
Topic 7	har	boyfriend	sadao	seafood	tomorrow

It can be observed in Fig. 6 that 62.34% of tweets from Kuala Lumpur were mainly talking about topic 1. Example tweets are "Have fun in Hat Yai! Asean Night Bazaar is the best", "Where to buy a very delicious frozen dumpling. I want to dumpling in Hat Yai". The most prevalent topic of tweets from Kedah is topic 1 with 55.43% and example tweets are "This coffee shop will be added to my-visit-list when in Hat Yai", "If you worry about the rainy season, Hat Yai is the best. Food Heaven". For Perlis state, topic 2 is the most dominant topic with 52.22% and example tweets such as "Till Next Time @Dannok Immigration Check Point in Sadao, Songkhla", "I'm at Dan Nok in Sadao, Songkhla".



[a] [b] [c]

Figure 6 The most discussed topics in tweets before the COVID-19 lockdown from [a] Kuala Lumpur territory [b] Kedah state and [c] Perlis state

After the COVID-19 Lockdown

To find out topics most Twitter users in Kuala Lumpur territory, Kedah state, and Perlis state used when talking about tourism in Southern Thailand after the COVID-19 lockdown, Tables 9, 10, and 11, present the lists of top five prevalent keywords that are strongly associated to the derived topics generated from users' tweets in one territory and two mentioned states.

Table 9 Words in tweets from Kuala Lumpur territory after the COVID-19 lockdown with the strongest association to the derived topics

Topic/Word	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	hat yai	songkhla	want	eat	thailand
Topic 1*	price	hat yai	hotel	best	songkhla
Topic 2	ticket	want	eat	don	price
Topic 3	ticket	want	price	wai	hat yai

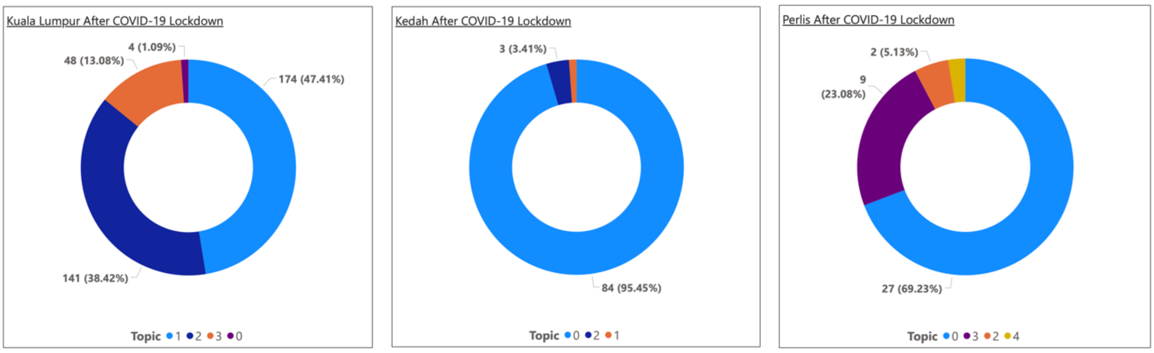
Table 10 Words in tweets from Kedah state after the COVID-19 lockdown with the strongest association to the derived topics

Topic/Word	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0*	hat yai	cheap	nei	hotel	week
Topic 1	holidai	train	malaysia	hat yai	songkhla
Topic 2	hotel	hat yai	holidai	mom	week

Table 11 Words in tweets from Perlis state after the COVID-19 lockdown with the strongest association to the derived topics

Topic/Word	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0*	hello	songkhla	sadao	hat yai	dan nok
Topic 1	hat yai	minut	mom	thailand	sadao
Topic 2	songkhla	minut	mom	seafood	train
Topic 3	hat yai	place	dan nok	debit	dimsum
Topic 4	sadao	seafood	nurlaila	restaur	tomyam
Topic 5	train	place	trend	ride	bring

Fig. 7. depicts the dominant topics of each state, i.e., topic 1 belongs to tweets from Kuala Lumpur sitting on top with 47.41% and topic 0 belongs to both from Kedah and Perlis states dominating 95.45 and 69.23% respectively. More specifically, after the lockdown was lifted, Twitter users in Kuala Lumpur were most concerned about hotels in Hat Yai City and Songkhla Province. Example tweets e.g., "After that when people prefer to go on vacation to Hat Yai much cheaper, Malaysians say Malaysians do not support domestic tourism", "The hotel in Hat Yai is full this weekend". This is similar to the prevalent topic from Kedah which most Twitter users also talked about accommodations in Hat Yai. Example tweets such as "Hotel train Hat Yai RM76. It's expensive because it's on the train station, hahahahahaha", "Hat Yai two years no tourist dies all, now still cheap all hotels". While tweets from Perlis mainly contributed to places like Sadao city and Dan Nok town. Example tweets are "I'm at Nurlaila Tomyam Restaurant - Seafood in Sadao, Songkhla", and "After difficult years, hello Dan Nok!, Sadao".



[a] [b] [c]

Figure 7 The most discussed topics in tweets after the COVID-19 lockdown from [a] Kuala Lumpur territory [b] Kedah state and [c] Perlis state

Discussion

Sentiment Analysis

In this study, sentiment analysis is performed with a very distinctive approach called VADER, which has the largest benefit of requiring very little data cleansing since it recognizes the magnitude of the words. After the analysis, it is found that in case neutral tweets are disregarded, results of both before and after the COVID-19 lockdowns revealed that the attitudes conveyed by Malaysian tourists were positive in both states and a territory especially a territory and state that are further away from Thailand, i.e., Kuala Lumpur and Kedah. More specifically, the results of positive tweets after the lockdown are greater than those before the lockdown by 0.75%, 2.12%, and 2.9% for Kuala Lumpur, Kedah, and Perlis respectively. Negative tweets, on the other hand, only tweets from Perlis provided the result after lockdown less than that before the lockdown by 4.75%. Interestingly, tourists from Perlis, the northern Malaysian state bordering Thailand, exhibited predominantly neutral sentiment. This is likely due to the ease of cross-border travel. However, compared to Kuala Lumpur and Kedah, Perlis also showed a greater upward trend in positive sentiment towards Southern Thai destinations.

Our sentiment analysis results, showing a dominance of positive tweets regarding travel both before and after lockdowns, align with Asan (2021) who suggested the pandemic's impact on travel attitudes and behavior, particularly for youth tourism, was relatively low. Furthermore, Mishra et al. (2021) posited that the prevalence of positive sentiment holds promise for a substantial and beneficial influence on worldwide tourism in the coming years.

Topic Modeling

The results of topic modeling analysis are meaningfully distinct between before and after the COVID-19 lockdowns. It was found that before the lockdown was announced, the prevalent topics of tweets from Kuala Lumpur and Kedah were similar. They, in many instances, talked about desires, for example, the desire to eat Thai foods, the desire to go to places in Hat Yai or Songkhla, etc. The prevalent topic of tweets from Perlis was about places that are near Thailand and Malaysia borders such as Sadao city and Dan Nok town. Interestingly, results after lockdown found that the dominant topics of Twitter users both in Kuala Lumpur and Kedah were concerned about hotels in Hat Yai or Songkhla. In the same way as before the lockdown, Sadao city and Dan Nok town were also the prevalent topics provided by Twitter users in Perlis.

Our topic modeling results corroborate the findings of Balasubramanian et al. (2021) who identified hygiene, health & safety, and access to suitable accommodation and restaurants as primary concerns for tourists in the post-COVID-19. Furthermore, considering Malaysia's Muslim-majority population, our findings align with Feizollah et al. (2021) who suggested that Malaysian tourists prioritize factors related to halal tourism, including food and accommodation options.

Conclusion and Suggestions

Social media are defined as platforms that allow their users to share opinions, and thoughts and participate in social networking activities. Since the occurrence of the COVID-19 pandemic, travel and tourism have been among the most affected industries. Analysis of these historical and unstructured contents may help not only to understand the attitudes of tourists in the past but also to prepare both private and public authorities for upcoming tourists once the COVID-19 situation normalises.

After the recovery of the COVID-19 situation and the re-opening of the Thailand-Malaysia border on April 1, 2022, the influx of Malaysian tourists coming to Thailand has increased especially in cities in the South such as Hat Yai City, Songkhla Province. A two-year break due to the pandemic could greatly influence all aspects of the tourism domain including the tourists' attitudes and activities. This study aimed to scrape and analyze user-generated posts from the renowned social network platform, Twitter, about the sentiment and the underlying themes of Malaysian tourists travelling in Southern Thailand. The study began by performing Twitter data extraction with the use of The Python library. As a result, two datasets were collected from Malaysia Twitter users in one territory and two states which are Kuala Lumpur, Kedah, and Perlis with the periods before and after the COVID-19 lockdowns of the two neighboring countries. The subsequent steps were to use Natural Language Processing (NLP) techniques such as sentiment analysis and topic modeling to identify attitudes and themes contained in scraped tweets. The results have shown that Malaysian tourists in Kuala Lumpur territory and Kedah state have different attitudes and themes from those in Perlis state regarding tourism in the South of Thailand. In other words, despite limitations on travel due to the COVID-19 pandemic, after the restrictions Malaysian tourists continued to view Southern Thailand positively and still considered it an attractive destination as evidenced by Twitter content.

This study's findings offer practical implications for tourism policymakers. The results can inform the development of targeted promotional and campaigning strategies. Additionally, these insights can be utilized by tourism stakeholders, including the Tourism Authority of Thailand, Hat Yai City Municipality, and Songkhla Provincial Administrative Organization, to identify and address both the strengths and weaknesses of tourism development in Southern Thailand. For instance, the prevalence of positive sentiment can suggest an opportunity for hospitality stakeholders like Hat Yai Songkhla Hotels Association to strategically reallocate resources towards recovery efforts including renovations or service improvements targeted at meeting the evolving needs of Malaysian tourists in the forthcoming years.

One of this study's future works is to see whether the analyses of sentiment and topic modeling can suggest or can be expanded to other sectors like business or healthcare to see what kinds of comparable words are impacting and influencing the tourism industries between Thailand and Malaysia. We were limited in the quantity of data we could collect because of the period and location restrictions, which was one of the study's constraints.

Acknowledgements

We wish to acknowledge the constructive comments and suggestions provided by anonymous reviewers and editorial team. These comments and suggestions have significantly improved the article. We also wish to thank Mr Roy I. Morien of the Naresuan University Graduate School for his editing of the grammar, syntax and general English expression in this manuscript.

Author Contributions

Author 1 (Md Tareq Bin Hossain): Conceptualization, Design of methodology, Manuscript writing, and Manuscript review.

Author 2 (Ruchdee Binmad): Conceptualization, Development of methodology, Collection of data, Data analysis, Manuscript writing, Manuscript review, and Manuscript editing.

Conflict of Interests

All authors declare that they have no conflicts of interest.

Funding

The authors received no financial support for the research of this article.

References

Abd- Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *Journal of Medical Internet Research*, *22*(4), 1-9. http://dx.doi.org/10.2196/19016

Ainin, S., Feizollah, A., Anuar, N. B. & Abdullah, N. A. (2020). Sentiment Analyses of Multilingual Tweets on Halal Tourism. *Tourism Management Perspectives, 34*, 1-8. https://doi.org/10.1016/j.tmp.2020.100658

Anandarajan, M., Hill, C., & Nolan, T. (2019). Latent Semantic Analysis (LSA) in Python. In *Practical Text Analytics, Maximizing the Value of Text Data: Advances in Analytics and Data Science*, 2, (pp. 221-242). Springer, Cham. https://doi.org/10.1007/978-3-319-95663-3_14

Anupama, V., & Elayidom, M. S. (2022). *Course Recommendation System: Collaborative Filtering, Machine Learning and Topic Modelling* [Conference session]. 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India. https://doi.org/10.1109/ICACCS54159.2022.9785353

Asan, K. (2021). Covid-19 Pandemic on Youth Tourism. *Journal of Mediterranean Tourism Research*, *1*(1), 12-21.

Balasubramanian, S., Kaitheri, S., Nanath, K., Sreejith, S., & Paris, C. M. (2021). Examining Post COVID-19 Tourist Concerns Using Sentiment Analysis and Topic Modeling. In W. Wörndl, C. Koo, & J. L. Stienmetz (Eds.), *Information and Communication Technologies in Tourism 2021*. Springer, Cham. https://doi.org/10.1007/978-3-030-65785-7_54

Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., & Reuter, C. (2021). Data Augmentation in Natural Language Processing: A Novel Text Generation Approach for Long and Short Text Classifiers. *International Journal of Machine Learning and Cybernetics, 14*(1), 135-150. https://doi.org/10.1007/s13042-022-01553-3

Binabdullah, K., & Tongtep, N. (2021). *Comparative Study on Natural Language Processing for Tourism Suggestion System* [Conference session]. 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Korea (South). IEEE. https://doi.org/10.1109/ITC-CSCC52171.2021.9501422

Binmad, R., & Li, M. (2018). Psychology-Inspired Trust Restoration Framework in Distributed Multi-Agent Systems. *Scientific Programming*, *2018*, 1-15. https://doi.org/10.1155/2018/7515860

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A Comprehensive Survey on Sentiment Analysis: Approaches, Challenges and Trends. *Knowledge-Based Systems*, *226*, 107-134. https://doi.org/10.1016/j.knosys.2021.107134

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian Journal of Computer Science and Technology, 8*(S2), 1-6. https://doi.org/10.51983/ajcst-2019.8.s2.2037

Bunnoon, P., Thongtang, L., Madsa, T., & Suntiniyompakdee, A. (2021). Satisfaction and Behavior of Foreign Tourists during the Vegetarian Festival in Food Routes of Chue-Chang Community, Tourist Attractions at Hat Yai District in Songkhla Province. *Parichart Journal*, *34*(1), 42-58.

Camilleri, M. A., & Troise, C. (2023). *Chatbot Recommender Systems in Tourism: A Systematic Review and A Benefit-Cost Analysis* [Conference session]. 8th International Conference on Machine Learning Technologies, Stockholm, Sweden. https://doi.org/10.1145/3589883.3589906

Casillano, N. F. B. (2022). Discovering Sentiments and Latent Themes in the Views of Faculty Members towards the Shift from Conventional to Online Teaching Using VADER and Latent Dirichlet Allocation. *International Journal of Information and Education Technology*, *12*(4), 290-298.

Centers for Disease Control and Prevention. (2023). *CDC Museum COVID-19 Timeline.* https://www.cdc.gov/museum/timeline/covid19.html

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown Spark.

Elsenbroich, C., & Gilbert, N. (2014). *Modeling Norms*. Springer.

Feizollah, A., Mostafa, M. M., Sulaiman, A., Zakaria, Z., & Firdaus, A. (2021) Exploring Halal Tourism Tweets on Social Media. *Journal of Big Data*, *8*, 72. https://doi.org/10.1186/s40537-021-00463-5

Gadamshetti, S., Deepak, G., Santhanavijayan, A., & Venugopal, K. R. (2022). RDRLLJ: Integrating Deep Learning Approach with Latent Semantic Analysis for Document Retrieval. In N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath, & N. Nalini, (Eds.), *Emerging Research in Computing, Information, Communication and Applications*. Lecture Notes in Electrical Engineering, vol 790. Springer, Singapore. https://doi.org/10.1007/978-981-16-1342-5_79

Ge, J., Vazquez, M. A., & Gretzel, U. (2018). Sentiment Analysis: A Review. In M. Sigala, & U. Gretzel (Eds.), *Advances in Social Media for Travel, Tourism and Hospitality*. Routledge.

George, M. I. N. O., Soundarabai, P. B. & Krishnamurthi, K. (2017). Impact of Topic Modeling Methods and Text Classification Techniques in Text Mining: A Survey. *International Journal of Advances in Electronics and Computer Science*, *4*(3), 72-77.

Hutto, C. J., & Gilbert, E. E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text* [Conference session]. 8th International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*, *49*(1), 300-324. https://doi.org/10.1111/psj.12343

Jeong, B., Yoon, J., & Lee, J. M. (2019). Social Media Mining for Product Planning: A Product Opportunity Mining Approach based on Topic Modeling and Sentiment Analysis. *International Journal of Information Management*, *48*, 280-290. https://doi.org/j.ijinfomgt.2017.09.009

JustAnotherArchivist. (2022). *Snscrape: A social networking service scraper in Python*. https://github.com/JustAnotherArchivist/snscrape

Kemp, S. (2023). *Digital 2023: Global Overview Report*. Datareportal. https://datareportal.com/reports/digital-2023-global-overview-report/

Khan, A. A., Newn, J., Kelly, R. M., Srivastava, N., Bailey, J., & Velloso, E. (2021). GAVIN: Gaze-Assisted Voice-based Implicit Note-Taking. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *28*(4), 1-32. https://doi.org/10.1145/3453988

Kherwa, P., & Bansal, P. (2019). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), 1-16. http://dx.doi.org/10.4108/eai.13-7-2018.159623

Liang, S., Jin, J., Ren, J., Du, W., & Qu, S. (2023). An Improved Dual-Channel Deep Q-Network Model for Tourism Recommendation. *Big Data*, *11*(4), 268-281. https://doi.org/10.1089/big.2021.0353

Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., & Yang, Y. (2020). Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends. *JMIR Public Health and Surveillance*, *6*(2), 1-4.

Martín, C. A., Torres, J. M., Aguilar, R. M., & Diaz, S. (2018). Using Deep Learning to Predict Sentiments: Case Study in Tourism. *Complexity*. https://doi.org/10.1155/2018/7408431

Mishra, R. K., Urolagin, S., Jothi, J. A. A., Neogi, A. S., & Nawaz, N. (2021). Deep Learning-based Sentiment Analysis and Topic Modeling on Tourism During Covid-19 Pandemic. *Frontiers in Computer Science*, *3*, 1-14. https://doi.org/10.3389/fcomp.2021.775368

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Applied Sciences (Switzerland)*, *11*(18), 1-25. https://doi.org/10.3390/app11188438

Németh, R., & Koltai, J. (2023). Natural Language Processing: The Integration of A New Methodological Paradigm into Sociology. Intersections. *East European Journal of Society and Politics*, *9*(1), 5-22. https://doi.org/10.17356/ieejsp.v9i1.871

Neogi, P. P. G., Das, A. K., Goswami, S., & Mustafi, J. (2020). Topic Modeling for Text Classification. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics*. Springer.

Pano, T., & Kashef, R. (2020). A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing*, *4*(33). https://doi.org/10.3390/bdcc4040033

Park, J. H., Lee, C., Yoo, C., & Nam, Y. (2016). An Analysis of the Utilization of Facebook by Local Korean Governments for Tourism Development and the Network of Smart Tourism Ecosystem. *International Journal of Information Management*, *36*(6), 1320-1327. https://doi.org/10.1016/J.IJINFOMGT.2016.05.027

Kong, J. T. H., Juwono, F. H., Ngu, I. Y., Nugraha, I. G. D., Maraden, Y., & Wong, W. K. (2023). A Mixed Malay-English Language COVID-19 Twitter Dataset: A Sentiment Analysis. *Big Data and Cognitive Computing*, *7*(2), 61. https://doi.org/10.3390/bdcc7020061

Praprom, C., & Laipaporn, J. (2021). The Intervention Analysis of the Interrupted Incidents' Impacts on Malaysian Tourist Arrivals to Songkhla Province in Thailand. *Journal of Environmental Management and Tourism*, *12*(6), 1513-1522. https://doi.org/10.14505//jemt.v12.6(54).08

Rehurek, R., & Sojka, P. (2010). *Software Framework for Topic Modeling with Large Corpora* [Conference session]. The LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta. https://radimrehurek.com/lrec2010_final.pdf

Salloum, S. A., Khan, R., & Shaalan, K. (2020). *A Survey of Semantic Analysis Approaches*. In A. E. Hassanien, A. Azar, T. Gaber, D. Oliva, & F. Tolba (Eds.), *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). AICV 2020. Advances in Intelligent Systems and Computing, vol 1153*. Springer, Cham. https://doi.org/10.1007/978-3-030-44289-7_6

Sharma, H., Jindal, H., & Devi, B. (2023). *Advancements in Natural Language Processing: Techniques and Applications* [Conference session]. International Conference on Advanced Computing and Communication Technologies, ICACCTech 2023. https://doi.org/10.1109/ICACCTech61146.2023.00019

Srivastav, A., Khan, H., & Mishra, A. K. (2020). Advances in Computational Linguistics and Text Processing Frameworks. In G. Loveleen, A. Solanki, V. Jain, & K. Deepak (Eds.), *Handbook of Research on Engineering Innovations and Technology Management in Organizations*. IGI Global.

Statista. (2022). *Number of Twitter Users Worldwide From 2019 To 2024*. https://www.statista.com/statistics/303681/twitter-users-worldwide/

Stella, M., Restocchi, V., & De Deyne, S. (2020). #lockdown: Network-Enhanced Emotional Profiling in the Time of COVID-19. *Big Data and Cognitive Computing*, *4*(2), 14. https://doi.org/10.3390/bdcc4020014

Tabassum, S., Pereira, F. S. F., Fernandes, S., & Gama, J. (2018). Social Network Analysis: An Overview. *WIREs Data Mining Knowledge Discovery*, *8*(5), 1-30. https://doi.org/10.1002/widm.1256

Tepanon, Y., Saiprasert, W., & Tavitiyaman, P. (2021). Destination Images of Thailand: Current and Future Development. In J. Zhao, L. Ron & X. Li (Eds.), *The Hospitality and Tourism Industry in ASEAN and East Asian Destinations: New Growth, Trends, and Developments*. Apple Academic Press. https://doi.org/10.1201/9781003082200

Trajkova, M., lhakamy, A., Cafaro, F., Vedak, S., Mallappa, R., & Kankara, S. R. (2020). Exploring Casual COVID-19 Data Visualizations on Twitter: Topics and Challenges. *Informatics*, *7*(3), 1-22.

Vajpai, G. N., & Pattanaik, D. (2022). Analyzing Visitors' Review of Homestays Located in Nature-Based Settings: An NLP Based Approach. *NMIMS Management Review*, *30*(2), 8-17. https://doi.org/10.53908/NMMR.300201

Valeri, M., & Baggio, R. (2021). Social Network Analysis: Organizational Implications in Tourism Management. *International Journal of Organizational Analysis*, *29*(2), 342-353. https://doi.org/10.1108/IJOA-12-2019-1971

Wang, Z., Zhang, G., Yang, K., Shi, N., Zhou, W., Hao, S., Xiong, G., Li, Y., Sim, M., Chen, X., Zhu, Q., Yang, Z., Nik, A., Liu, Q., Lin, C., Wang, S., Liu, R., Chen, W., Xu, K., Liu, D., Guo, Y., & Fu, J. (2023). *Interactive Natural Language Processing*. *ArXiv*. https://doi.org/10.48550/arXiv.2305.13246

Wolpe, Z., & Waal, A. D. (2019). *Autoencoding variational Bayes for Latent Dirichlet Allocation* [Conference session]. South African Forum for Artificial Intelligence Research (FAIR 2019), Cape Town, South Africa.

Wongmonta, S. (2021). Post-COVID 19 Tourism Recovery and Resilience: Thailand Context. *International Journal of Multidisciplinary in Management and Tourism*, *5*(2), 137-148. https://doi.org/10.14456/ijmmt.2021.12

World Health Organization. (2020). *Archived: WHO Timeline-COVID-19*. https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

Wun'Gaeo, C., & Wun'Gaeo, S. (2021). Thailand and Covid-19: Institutions and Social Dynamics from Below. In J. Nederveen Pieterse, H. Lim, & H. Khondker (Eds.), *Covid-19 and Governance: Crisis Reveals*. Routledge. https://doi.org/10.4324/9781003154037

Yu, C., Zhu, X., Feng, B., Cai, L., & An, L. (2019). Sentiment Analysis of Japanese Tourism Online Reviews. *Journal of Data and Information Science*, *4*(1), 89-113. https://doi.org/10.2478/jdis-2019-0005

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Transactions on Management Information Systems*, *9*(2), 1-29. https://doi.org/10.1145/3185045