Data Science and Engineering (DSE) Record, Volume 4, issue 1.

Analysis and Visualization of Tourist Behavior of Hospitality Service in Chiang Mai Province Using Sentiment Analysis Method

Mallika Chali[1] and Arinya Pongwat [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

mallika_chali@cmu.ac.th

Abstract. This independent study aims to understand the behaviors of tourists affected by hostel accommodation services in a Mueang Chiang Mai district, Chiang Mai province. The data used in this research was collected from TripAdvisor.com, 5,108 messages were crawled and separated into 17,092 sentences. In-depth interviews with hostel entrepreneurs provide insights for the essential aspects considered when managing their businesses. These aspects together with aspects from related studies serve as classification criteria for sentiment analysis using Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) algorithms. The SVM model achieves 93% accuracy, outperforming MNB's 82%. Text-mining analysis explores hostel business development. The findings reveal that SVM is suitable for classifying customer review messages, and exhibiting satisfactory performance and accuracy. The aspects discovered in this studies include cleanliness, facility, location, quality of staff, security, social atmosphere, and value of money. The results of the current study contribute to the theoretical context for academic as well as practical guidelines for the hostel managers in general.

Keywords: Hostel, Hospitality Service, Sentiment Analysis, SVM, MNB.

1 Introduction

The tourism industry holds significant importance as a key driver of the country's economy, making it a crucial sector to consider. In Thailand, it is regarded as one of the key economic indicators. Prior to the outbreak of COVID-19, the tourism report from the Ministry of Tourism and Sports highlighted substantial growth in the accommodation branch and food services within the tourism industry. This sector experienced a growth rate of 6.8 percent, contributing significantly to the country's Gross Domestic Product (GDP). [1] Tourism is widely recognized as a vital source of income, as it contributes to job creation and income distribution across various regions of the country. One prominent sector within the tourism industry is the hotel business. In recent

times, there has been a noticeable increase in the popularity of solo travel or traveling in small groups. This trend is gaining traction and becoming increasingly popular among travelers. [2] Consequently, the accommodation business has shifted its focus to cater more specifically to this group of tourists. Recognizing the growing demand for solo and small group travelers, accommodation providers have redirected their attention to better accommodate the needs and preferences of these individuals.

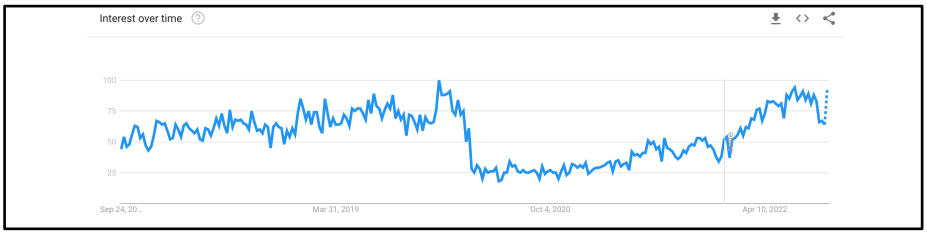

Fig. 1 Google search term: Solo Travel over past 5 years (https://trends.google.co.th/trends/)

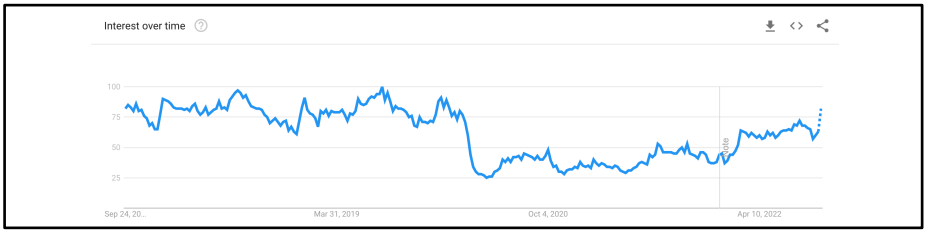

Fig. 2 Google search term: Hostel over past 5 years (https://trends.google.co.th/trends/)

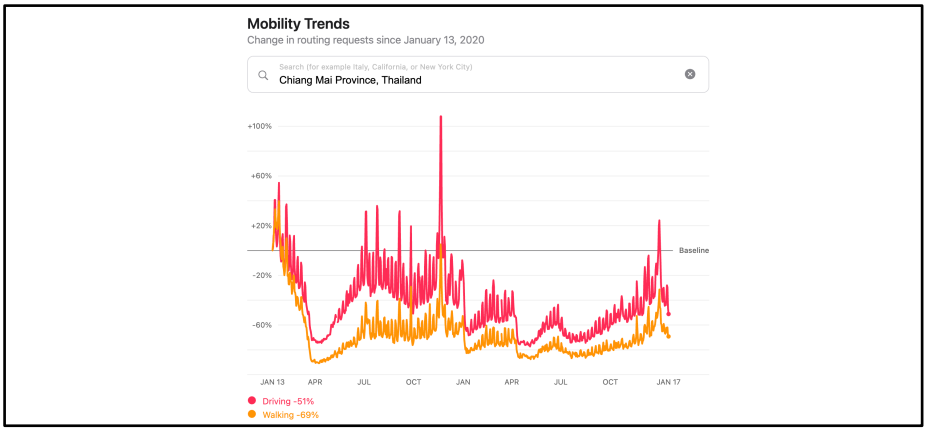

Fig. 3 Apple Mobility Trends: Chiang Mai province in 2020 – 2021
(Apple Mobility Trends Reports (As of Jan 17, 2022))

The hostel business has emerged as an intriguing industry because it aims to cater to the needs of various types of travelers such as solo travelers, digital nomads, and small groups of friends. These accommodations offer affordable lodging options that are

particularly suitable for individuals traveling for work purposes (also known as "workcation"). Chiang Mai, being one of the three provinces in Thailand popular among these types of travelers, further adds to the appeal of the hostel business. [3] To effectively serve their guests, hostel operators need to plan and adapt their services based on customer needs. Gathering information about customer preferences is crucial, and this can be done through various sources, including offline channels and online platforms such as social media and travel review websites. However, the sheer volume of online reviews can make it challenging for entrepreneurs to manually read and analyze all of them within a limited timeframe.

Previous research studies have utilized machine learning techniques to analyze and categorize services in the hotel industry. However, it is important to recognize that hostels have unique service characteristics that may differ from traditional hotels. Therefore, there is a need for further classification of service aspects specific to hostels. This classification will enable the development of systems that can accurately classify messages according to these specific service characteristics, aligning with the distinct context of hostels.

2 Literature Review

2.1 Sentiment Analysis Method

Bachtiar, Paulina, and Rusydi [4], explore the role of online travel agent (OTA) websites in the importance of authentic reviews in customer decision-making. They utilize text mining techniques, specifically sentiment analysis, to analyze reviews from popular websites such as Agoda.com, Expedia, Pegi-Pegi, Booking.com, and Tripadvisor. The study focuses on classifying reviews into different groups based on service as-pects like location, rooms, food, price, and service quality. The analysis highlights the need for improvement in the food sector, as indicated by satisfaction scores obtained from the Support Vector Machine (SVM) algorithm.

Abro et al. [5], analyzed user opinions and experiences shared online about products and services. They employed three approaches: Model A and B classified as-pects mentioned in restaurant reviews using SemEval, while Model C determined sentiment (positive/negative). The study compared feature engineering techniques and machine learning algorithms. Combining word2vec with SVM yielded better performance, achieving 76% accuracy for Model A, 72% for Model B, and 79% for Model C.

Yu [6] investigated sentiment analysis using supervised machine learning on textual comments from Tripadvisor. The study classified opinions based on aspects such as service, room quality, location, value, cleanliness, quality of sleep, and customer service. The developed model achieved 70% to 75% accuracy for star prediction and 85% to 90% accuracy for positive/negative classification using the support vector machine (SVM) algorithm.

From the relevant research studies, it was discovered that sentiment analysis utilizing the Support Vector Machine (SVM) method outperformed other supervised learning methods significantly. Therefore, this research involved creating two teaching models, Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB), for the learning system to classify service aspects. The performance of these models was then compared. Afterward, the content of each positive and negative text polarity was extracted to gain a deeper understanding of the behavior and needs of hostel guests. This analysis aimed to provide valuable insights that hostel operators can utilize to make informed decisions for further business development.

2.2 Hostel Business

Ana and Paulo [7], the quality of service in hostels encompasses four key dimensions: the quality of staff, social atmosphere, tangible aspects of hostel accommodation, and proximity to the city. These dimensions are interconnected and play a vital role in assessing guest satisfaction, making recommendations, and predicting guest loyalty. They also help identify variations among market segments, providing insights into guest preferences and needs. Overall, these dimensions contribute to effectively meeting the expectations of hostel guests and enhancing their experience.

Mukherjee et al [8] categorized the aspects of hostel accommodation into eight facets. These facets include: namely Safety and Security , Facilities , Col-lege Community , Quality and Bed Options , Food , Cost and Price and other miscellaneous aspects Other (Miscellaneous).

Mylocopos and Dickinger [9] utilized text mining and sentiment analysis techniques to gain insights into the expectations and emotions of backpackers. The re-sults of the study can be categorized into four main aspects: The staff (Staff), the location (Location) , the atmosphere (Atmosphere) and the facilities (Facilities).

The purpose of this study was to classify customer reviews based on service aspects derived from the relevant documents above. Furthermore, the researcher conducted additional interviews with hostel operators in Chiang Mai to identify the appropriate Aspects for analyzing the sentiment of hostel guests in the specific context of Chiang Mai. These identified variables played a significant role in shaping the customer's sentiment and facilitated the examination of how each service aspect influences the customer's sentiment.

3 Data and Methodology

3.1 Data

Data for this study was collected by crawling 5,108 hostel review text messages from Tripadvisor.com in Mueang Chiang Mai district, Chiang Mai province, from 2019 to

2021. The data collection process utilized the Selenium library in Python. The dataset includes information from 161 hostels and is stored in an Excel file. The selected variables for analysis pertain to hostel hospitality. with the following data fields: h_name, h_location, review_name, review_date, stay_date, review_title, and review. These data fields capture the essential information necessary for analyzing and understanding the sentiments, aspects, and overall hospitality experiences expressed in the hostel reviews.

3.2 Methodology

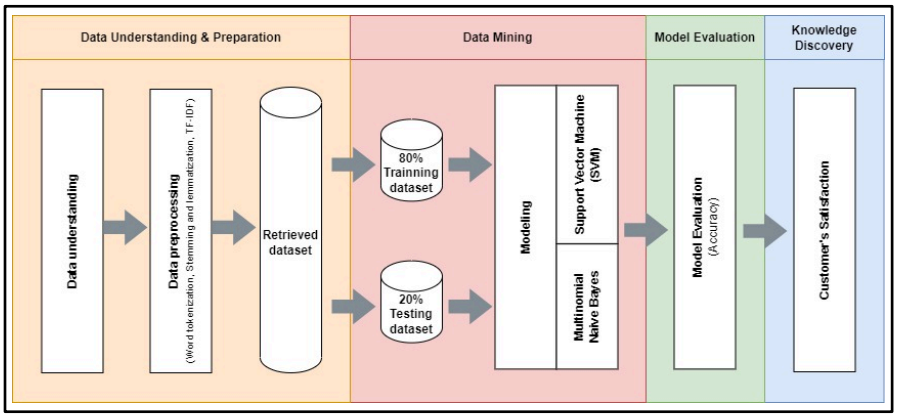

Fig. 4 Study Process Design

1) Business Understanding

In-depth interviews were conducted with 3 domain experts and successful entrepreneurs in the hostel accommodation industry. Entrepreneurs operating hostels in the Mueang Chiang Mai District of Chiang Mai Province were purposively selected based on their achievement of the prestigious Certificate of Excellence from TripAdvisor.com. This certificate recognizes businesses that consistently deliver exceptional service, with only around 10% of listed businesses receiving this distinction. By targeting these entrepreneurs, valuable insights on hostel service aspects were gathered from industry leaders.

2) Data Understanding

Data understanding is one of the important processes before developing the model. It is a process for data analysis, leading to the appropriate use of variables. This study applied descriptive statistics. The results reveal that the data in each of selected data there are no outliner (Customer votes) using data validation techniques. There are 9 Sub-district from our data using data extraction techniques. The data distribution is left-skewed and skewed toward high votes. It is possible that reviews are more positive than negative.

3) Data Preparation

This study uses the data preparation consisting of a data cleansing technique to clean data for modeling using Microsoft Excel, there are no missing value, duplicate data, and collect to suitable form. Also, a tokenization technique, word stemming and lemmatization technique, and label encoding to prepare data for modeling using Natural Language Toolkit (nltk) and Scikit-learn (sklearn) in Python.

4) Modeling

Support Vector Machine (SVM)

This study utilizes the Support Vector Machine (SVM) approach for review aspect classification due to its higher accuracy compared to the Multinomial Naïve Bayes (MNB) algorithm [4-6]. SVM uses a subset of the training set called support vectors and achieves a clear margin of separation. It determines the optimal decision boundary between vectors belonging to a category and those that do not, dividing the space into two subspaces. [10]

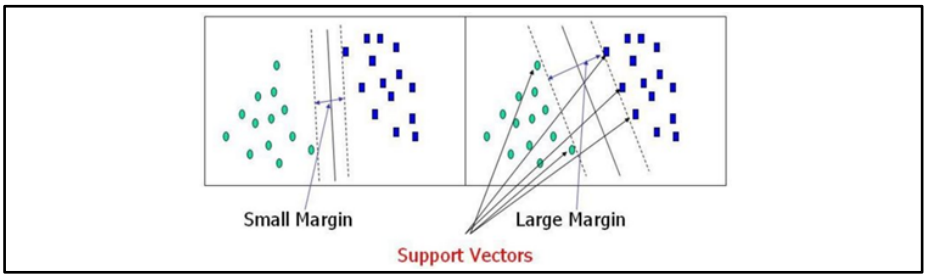

Fig. 5 Support Vector Machine (SVM) (From https://towardsdatascience.com)

Multinomial Naïve Bayes (MNB)

MNB (Multinomial Naïve Bayes) is a simple and efficient algorithm widely used in NLP problems such as sentiment analysis, spam detection, and topic classification. It calculates the probability of a class given an instance using Bayes theorem and considers certain features. Naive Bayes is commonly employed in NLP tasks, predicting the tag of a text by calculating the probability of each tag for the given text and selecting the one with the highest probability. [11]

To develop a model for classifying hostel guest reviews, both Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) models will be constructed using Scikit-Learn in Python. The performance of these models will be compared to determine the most suitable and efficient one. The dataset, covering the period from 2019 to 2021, will be divided into two sets: 80% for training the models and the remaining data for testing. Each model will be individually established, and the parameters will be adjusted for training. This study will consider the service aspects of hostel accommodation based on seven types; Quality of Staff, Facilities, Atmosphere,

Location, Cleanliness, Security, and Value of money, using the model for classifying reviews content.

5) Evaluation

The model performance is evaluated using the confusion matrix and running time to measure the maximized accuracy in the testing dataset.

4 Results

The experiment yielded results in two main areas. Firstly, the performance of two models, Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB), was compared. Various performance metrics, such as accuracy, precision, recall, etc., were used to assess the models' classification or analysis of service aspects in hostel guest reviews. The objective was to identify the superior model for this task. Secondly, the data analysis and visualization focused on extracting insights and patterns from hostel guest reviews, with a specific emphasis on different service aspects such as staff, location, cleanliness, facilities, etc. This analysis involved examining the reviews and applying visualization techniques such as charts, graphs, and word clouds to present the findings in a visually appealing and easily understandable format.

4.1 Model Performance Comparison

The experiment involved measuring the accuracy of classifying hostel guest reviews data based on predicting service aspects using two different models: Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB). To evaluate the efficiency of these models, a testing dataset was used. The accuracy of the classification was measured using the Accuracy_Score function, which quantifies how well the system learned to predict the service aspects of the hostel guest reviews. The results of this evaluation are shown in Table 1.

Table 1. Accuracy scores

Approach	Running Time	Accuracy
Support Vector Machine: SVM	3 Seconds	93 %
Multinomial Naïve Bayes: MNB	10 Seconds	82 %

Table 1 shows that the SVM model achieved an accuracy of 93%, while the MNB model achieved 82%. The SVM model also had a faster running time of 3 seconds compared to the MNB model's 10 seconds. These results indicate that the SVM model outperformed the MNB model in terms of accuracy and efficiency. The mention of the MNB model's performance with unseen data suggests its limitations with unfamiliar data points.

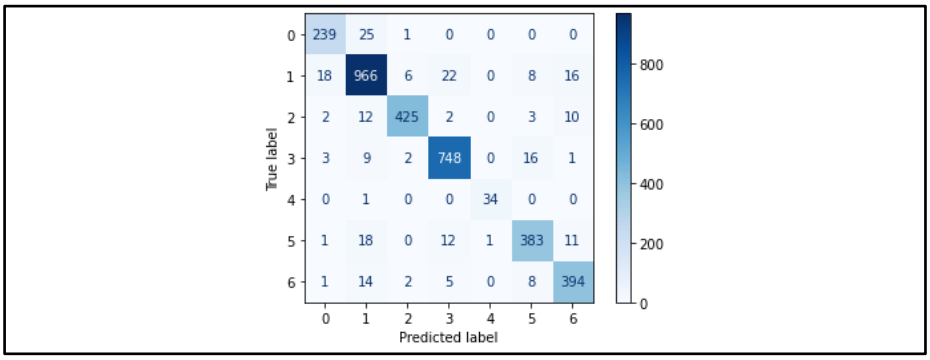

Fig. 6 Confusion Matrix of Support Vector Machine: SVM

Table 2. Confusion Matrix of Support Vector Machine: SVM

	Precision	**Recall**	**F1-Score**	**Support**
0 (Cleanliness)	0.91	0.90	0.90	265
1 (Facility)	0.92	0.93	0.93	1036
2 (Location)	0.97	0.94	0.96	454
3 (Quality of Staff)	0.95	0.96	0.95	779
4 (Security)	0.97	0.97	0.97	35
5 (Social Atmosphere)	0.92	0.90	0.91	426
6 (Value of Money)	0.91	0.93	0.92	424
Accuracy			0.93	3419
Macro avg	0.94	0.93	0.93	3419
Weighted avg	0.93	0.93	0.93	3419

Table 2 shows that the results demonstrate that the SVM model was found overall ac-curacy from F1-score is 93%. Fig. 6 illustrates the comparison graph of prediction label obtained from the SVM model with the true label.

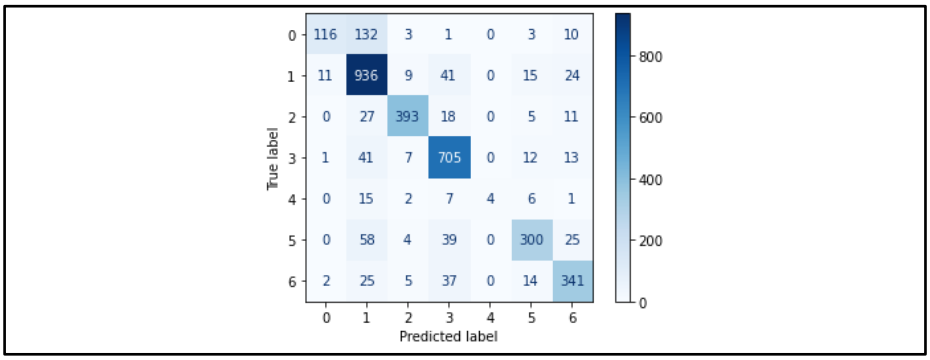

Fig. 7 Confusion Matrix of Multinomial Naïve Bayes: MNB

Table 3. Confusion Matrix of Multinomial Naïve Bayes: MNB

	Precision	**Recall**	**F1-Score**	**Support**
0 (Cleanliness)	0.89	0.77	0.59	265
1 (Facility)	0.76	0.90	0.82	1036
2 (Location)	0.93	0.87	0.90	454
3 (Quality of Staff)	0.83	0.91	0.87	779
4 (Security)	1.00	0.11	0.21	35
5 (Social Atmosphere)	0.85	0.70	0.77	426
6 (Value of Money)	0.80	0.80	0.80	424
Accuracy			0.82	3419
Macro avg	0.87	0.68	0.71	3419
Weighted avg	0.83	0.82	0.81	3419

Table 3 shows that the results demonstrate that the MNB model was found overall accuracy from F1-score is 82%. Fig. 7 illustrates the comparison graph of prediction label obtained from the MNB model with the true label.

4.2 Data Analysis and Visualization

Based on the in-depth interview, it can be concluded that hostel accommodation entrepreneurs start their businesses by researching the needs of tourists and target groups. They gather information from studying other hostels, interacting with operators, and reading online reviews. This helps them make informed decisions regarding their business, such as bed capacity and facility layout. Communication with customers is important, as hostels often cater to specific groups who value social interaction and shared accommodations. Entrepreneurs focus on online channels and word-of-mouth referrals through social media for customer communication and public relations. They actively respond to customer comments and prioritize engagement through activities and communal experiences. Continuous improvement is emphasized, and customer feedback plays a vital role in enhancing service quality.

In the study of data analysis and data visualization, the classification of aspects using a Support Vector Machine (SVM) allows for an examination of how guests value each aspect through positive and negative comments. To present this information effectively, a dashboard was created, which includes filters to read messages based on the service aspect and sentiment. The dashboard provides a comprehensive view of the data, allowing users to select filters to explore specific aspects and sentiments. Additionally, a word cloud is included, which can be filtered by specific words of interest. This allows for a quick visual representation of the most frequently mentioned words related to the hostel shown as Fig. 8

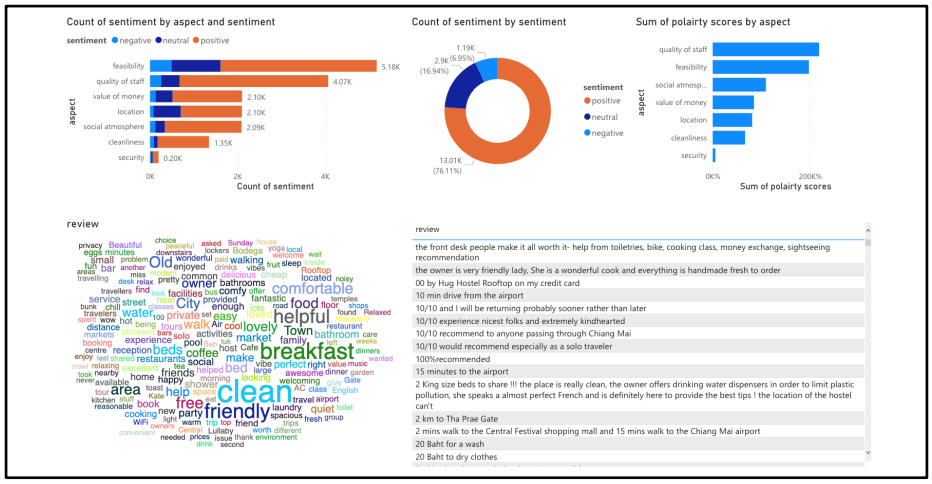

Fig. 8 Example hostel review dashboard

From fig. 8, It provides an overview of the different components included in the dashboard. These components are:

1) Bar Chart and Stack Bar Chart: This component displays the number of messages categorized as positive and negative reviews for each service aspect. It helps visualize the distribution of sentiments across different aspects.
2) Service Aspect Pie Chart: This pie chart provides an overview of the sentiment distribution for all the reviews. It shows the proportion of positive, negative, and possibly neutral sentiments for the entire dataset.
3) Bar Chart for Polarity Scores: This bar chart represents the polarity scores for each aspect. Higher polarity scores indicate a more positive mention of that aspect in the reviews. It allows users to assess the overall sentiment associated with each aspect.
4) Word Cloud: The word cloud visually represents the frequency of words mentioned in the text of all the reviews. The size of each word in the cloud corresponds to its frequency of occurrence. Users can filter the word cloud based on specific words of interest.
5) Raw Data Table: The lower right table presents the raw data, allowing users to access the detailed information and text of the reviews.

These charts, diagrams, and the raw data table provide a comprehensive and interactive way to explore and analyze the sentiment and content of the hostel reviews. Users can select and filter these components to gain insights and understand the guests' opinions and experiences more effectively.

5 Discussion and Conclusion

Based on the analysis conducted in this study, it can be concluded that the SVM (Support Vector Machine) technique is highly effective in analyzing the aspect-based sentiment of custom-ers staying in hostel accommodations in Mueang Chiang Mai

district, Chiang Mai province. The SVM model exhibited superior accuracy in classifying customer reviews compared to the MNB (Multinomial Naïve Bayes) models. The results demonstrated that the proposed SVM model achieved a high level of efficiency and accuracy in classifying text reviews, with an accuracy rate of 93%. On the other hand, the MNB models showed lower accuracy, with a rate of 82%. This study provides valuable knowledge and understanding regarding the aspect-based sentiment of customers staying in hostel accommodations in Mueang Chiang Mai district, Chiang Mai prov-ince. The findings can be applied to further enhance the development and improvement of hostel businesses in the area.

From the generated dashboard, the following conclusions can be drawn from the feedback da-ta:

1) Facilities (Facilities) received the highest number of mentions with 5,180 messages, ac-counting for 30.31% of the total 17,092 messages. This indicates that guests frequently mention the facilities provided by the hostels.

2) Quality of Staff received 4,072 messages, representing 28.82% of the total. This aspect is highly praised by guests, indicating that the quality of staff service is a standout feature in the hostels of Mueang Chiang Mai, Chiang Mai.

3) Value of Money received 2,104 messages, representing 12.31% of the total. Guests of-ten discuss the affordability and value they perceive in relation to the price they pay for their stay.

4) Location received 2,100 messages, representing 12.29% of the total. Guests frequently mention the location of the hostels, indicating its significance in their reviews.

5) Social Atmosphere received 2,092 messages, representing 12.24% of the total. This as-pect highlights the importance of a friendly and sociable atmosphere in the hostels.

6) Cleanliness received 1,348 messages, representing 7.89% of the total. Guests prioritize cleanliness, particularly in shared spaces like rooms, bedsheets, and bathrooms.

7) Security received 196 messages, representing 1.15% of the total. While it had the low-est number of mentions, it still holds importance to guests, highlighting the significance of providing a secure environment.

Regarding the sentiment analysis:

1) Positive Sentiment accounted for the highest proportion of mentions with 13,008 mes-sages, representing 76.11% of the total. Guests mostly expressed positive sentiments in their reviews.

2) Neutral Sentiment accounted for 2,896 messages, representing 16.94% of the total. These messages had a more balanced or neutral tone.

3) Negative Sentiment accounted for 1,188 messages, representing 6.95% of the total. These messages expressed negative sentiments in the reviews.

When considering the Polarity Scores, it can be observed that the Quality of Staff received the highest positive comments, indicating that guests highly appreciate the staff service.

The Word Cloud analysis highlights that guests often mention cleanliness, friendliness, and breakfast. Cleanliness is a crucial aspect for guests, and positive comments indicate their satisfac-tion with the cleanliness of rooms and shared facilities. The friendliness and service-mindedness of the staff are highly valued by guests. Additionally, guests appreciate hostels that offer break-fast or have hosts preparing breakfast together with guests, which adds to their overall satisfac-tion.

Overall, these insights help hostel entrepreneurs in Mueang Chiang Mai, Chiang Mai, to un-derstand the needs and preferences of their guests, and focus on areas such as staff quality, clean-liness, and creating a friendly and welcoming atmosphere to enhance guest satisfaction.

6 Limitation and future research.

1) Due to the limited time spent in this research, the developed system was able to process only English text. Unable to process text in other languages.

2) The conclusions obtained from this independent study can explain the feelings and behaviors of travelers who stay hostel in the Mueang Chiang Mai district, Chiang Mai province only. It cannot describe the feelings and behavior of travelers staying in other types of accommodations or in other locations.

3) The findings of this independent study explain the feelings and behaviors of travelers who write reviews on TripAdvisor only. Comments cannot be included on other websites.

4) The conclusion for the hostel accommodation business may be a guideline for operators at any given moment. It may not cover all the current traveler behavior.

5) Due to COVID-19 in the past several years, the number of travelers has decreased, and some hotels have closed, resulting in reviews are less than expected

References

1. สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, “GDP ไตรมาสที่สี่ทั้งปี 2562 และแนวโน้มปี 2563,” 17 ก.พ. 2563.
2. SolotravelerworldSolo, “Travel Statistics and Data”, 2022 [Online]. Available: https://solotravelerworld.com/about/solo-travel-statistics-data/ [Access April 10, 2022]
3. Nationthailand. “A new generation of workers are living their lives as digital nomads, travelling around the world while working online to earn their living. Thailand is one of the most popular destinations for these workers.”, October 25, 2022 [Online]. Available: https://www.nationthailand.com/thailand/tourism/40021375 [Access November 1, 2022]

4. Fitra A. Bachtiar el al, “Text Mining for Aspect Based Sentiment Analysis on Customer Review – A Case Study in The hotel Industry,” Faculty of Computer Science, Brawijaya University.
5. Sindhu Abro et al, “Aspect Based Sentimental Analysis of Hotel Reviews A Comparative Study,” Sukkur IRA Journal of Computing and Mathematical Science, Vol. 4, 2020.
6. Yang Yu, “Aspect-based Sentiment Analysis on Hotel Reviews,” Stanford University.
7. Ana B. and Paulo R, “Exploring heterogeneity among backpackers in hostels,” 2018.
8. Aniket Mukherjee et al, “Aspect Based Sentiment Analysis of Student Housing Reviews, ” 2020 sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 465-470, 2020.
9. Samantha Mylocopos and Astrid Dickinger, “Backpackers’ Expectations of Hybrid Hotels : A Text Mining Approach,” The 50th Annual EMAC Conference, Madrid, May 25-28, 2021. European Marketing Academy, 2021.
10. Rohith Gandhi, “Support Vector Machine – Introduction to Machine Learning Algoruthms,” June 7, 2018 [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [Access March 2, 2023]
11. Kong Ruksiam, “สรุป Machine Learning (EP.5) - การจัดหมวดหมู่ด้วย Naïve Bayes,” 27 มีนาคม 2565 [ออนไลน์]. จาก: https://kongruksiam.medium.com/ [เข้าถึง 2 มีนาคม 2566]