Examining the Relationship of Keyword Analysis using Online Traveller Hotel Reviews

Zuraini Zainol[1], Angela SH Lee[2], Puteri NE Nohuddin[3], Noor Farizah Ibrahim[4], Mohd Hanafi Ahmad Hijazi[5]

[1]Department of Computer Science, Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, Malaysia
zuraini@upnm.edu.my
[2]Department of Computing and Information Systems, Sunway University, Selangor, Malaysia
[3]Institute of IR4.0, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
[4]School of Computer Science, Universiti Sains Malaysia, Malaysia
[5]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Sabah, Malaysia

Abstract—With a large number of reviews being posted daily, online travel forums have become a very common platform for sharing travel information. Travel websites including TripAdvisor.com and Booking.com have become invaluable tools for hoteliers and travellers alike, enabling them to advertise hotel rooms, select hotels, and solicit and exchange reviews. This research paper uses data collected from TripAdvisor websites and 5501 reviews were collected from 4 different four-star hotels in Kuala Lumpur from 2012-2017. The work aims to extract the keyword (factors) that influence travellers in their choice of hotels using the framework of Visualizing Keyword Relationship Analysis (VKRA). The relationships of these keywords are illustrated using text analytics visualization techniques which are term frequency, keyword association and keyword network graph. The incorporation of these text analytics techniques offers insights for (i) travellers to view the opinions of other travellers on these hotels and (ii) hoteliers to understand travellers' preferences (location, room, services) and experiences, and improve their services in the future.

Keywords— relationship, keyword analysis, network graph

I. INTRODUCTION

The tourism industry, in a broad sense, is recognized as one of the major industries in terms of foreign exchange earnings, and it encompasses all businesses that directly provide products or services to promote business, pleasure, or leisure activities away from the home setting. According to [1], the tourism industry's foreign exchange earnings rose from RM17.3 billion in 2000 to RM46.1 billion in 2007, reflecting a 166 per cent rise in revenue in just seven years. Statistica reported that there are approximately 4830 hotels in Malaysia in 2019, with various ratings and sizes [2]. Over the last decade, information technology (IT) has played a major role in the tourism and hospitality industry, lowering prices, rising operational performance, and enhancing services and customer experience. Today, visitors and customers use IT to make decisions, identify hotels, receive hotel reviews, and book hotel accommodations via booking applications. Word-of-mouth from close friends and relatives, online reviews, ratings, feedback, price comparisons, and other sources of information play a major role in travel and hospitality services.

The hospitality industry is becoming more aware of and embracing Big Data methods for retrieving, storing, analysing, monitoring, and visualizing data to enhance overall customer service, customer loyalty, and business operations [3]. To make the most of the data, hoteliers must go beyond the what and why by advising hotels about how to identify measurement platforms and implement a customer-centric business strategy [4]. Since the hotel industry deals with millions of travellers from all over the world, each with their own set of needs, wishes, hopes, preferences, it is difficult to listen to and satisfy their needs on a large scale. As a result, the hotel industry is progressing in the use of predictive exploitation of data, moving beyond conventional loyalty schemes and deepening their awareness of current customers to find new ways to increase revenue, satisfy customers by gaining a thorough understanding of customer segment actions, needs, and expectations while also providing ways to draw new customers [5]. With the capability and process of leveraging information regarding the usage and combination of internal data, such as occupancy rates and current bookings, with external such as information about peak demand seasons, weather forecast, local events, or school holidays, big data analytics can provide hoteliers with the ability to analyse this information and create promotional campaigns and adjust room prices correlating with the customers' information. It makes hotels better, both from the guests' perspective and as a company, by identifying and developing new ways to connect with customers and find out their needs and desires. According to research, gaining customers' confidence and enabling them to express and share their opinions help in growing the business. Online

reviews and feedback provide an opportunity to share your thoughts and satisfaction levels.

The aim of this paper is to extract the keyword (factors) that influence travellers in their choices of hotels using the framework of Visualizing Keyword Relationship Analysis (VKRA). The relationships of these keywords are presented using text analytics visualization techniques such as term frequency, keyword association and keyword network graph. The incorporation of these text analytics techniques offers insights for (i) travellers to view the opinions of other travellers on these hotels and (ii) hoteliers to understand travellers' preferences (location, room, services) and experiences, and improve their services in the future.

II. Related Work

In a consumer-centric industry, several forms of research have been developed and publicized to understand customers' needs in order to expand and develop the company [6]. Hoteliers can improve their services by analysing text data – in this case, customer reviews where the analysis result can be used to improve business decision-making. Allowing consumers to write reviews serves as a way for them to share their thoughts or experiences about a product or service [7]. In the field of opinion mining, it also reflects the customer's sentiment or voice, which has a major effect on the market. It is known as sentiment analysis where it analyses customer sentiment, examine the mood, experiences, and emotions of the general public [8]. The hotel and hospitality industry serves millions of travellers from all over the world, and each of these travellers come from different countries and culture. Each customer has their own set of expectations and preferences, making it difficult for hotels to consistently fulfil those expectations. As a result, data analytics plays a vital role in this analysis with the hotel industry now turning to sophisticated analytical solutions for clues about how to please customers and keep them satisfied [9, 10].

In scholarly journals, keywords are natural language terms that convey the document's thematic concepts. The study of keyword co-occurrence is based on the statistics of the number of times a pair of keywords is listed in the same text, as well as to perform network and cluster research for these. Thus, the information structure and science frontier are shown on a specific subject. Mapping the knowledge domain is a type of image that depicts the scientific knowledge creation process and structural relationships. As a result of comprehending these complex knowledge relationships, new knowledge can be created. In this situation, big data analytics would be able to distinguish between the customers and make a comparison [9, 11].

Many have developed and publicised the value of listening to customers' voices [10-12]. Many businesses rely on customer surveys or customer feedback forms to collect data in organised formats. This type of data is normally analysed using business intelligence software to help pinpoint the best course of action for improving customer support, product development, and identifying the weakness of competitors. In several studies, text data from customer feedback, emails, social media contents are found important for businesses to uncover hidden insights by understanding the languages, their contexts, and how they vary and are used in everyday conversations [11, 13]. Previous works showed that the unstructured text data was pre-processed thoroughly to convert customers opinions into a more structured format and setting[13, 14]. Using the text analytics approach, researchers have been able to understand and explore text data in different fields such as detecting cyberbullying patterns using Twitter data [15, 16], finding the relationship of keywords in paediatric cancer patients' experience [17], social media analysis [18, 19].

III. Framework of Visualising Keyword Relationship Analysis

Fig. 1 shows the proposed framework of Visualizing Keyword Relationship Analysis (VKRA) over online traveller's reviews. This framework contains 3 main modules including (i) data collection, (ii) text pre-processing and term-document matrix, and (iii) visualizing text analysis.

A. Data Collection

The data was collected from the TripAdvisor website from 2012-2017. The website offers travellers/tourists information about hotels, vacations, rental, and reviews/feedback on hotels. About 5501 reviews from 4 four-star hotels were collected in this study. The sample of collected data is transformed into CSV format for data analysis.

B. Text Pre-Processing and Term Document Matrix

Data collected from any website often contain highly noisy and unstructured text with a mixture of languages and typos. Noisy text is unstructured text data that can be found in SMS, online chatting, email [20], online reviews, blogs, surveys, etc. This text data needs to be pre-processed into a structured form. Therefore, a pre-processing step is carried out to generate clean text data that will be applied for the next process. Some common text-processing tasks are listed below:

- Removing unnecessary contents such as punctuations marks, symbols, usernames, quotes, expressions, etc.
- Transforming text data to lower case letters
- Removing numbers, and stop words (e.g., 'he', 'i', 'am', 'is', 'she', 'he', 'has', 'for', 'have', 'with', 'for', etc.)
- Striping extra whitespaces
- Applying stemming algorithms to reduce common words to their roots such as 'locat' (location), 'servic' (services/service), 'walk'(walk/walks/walked), etc.

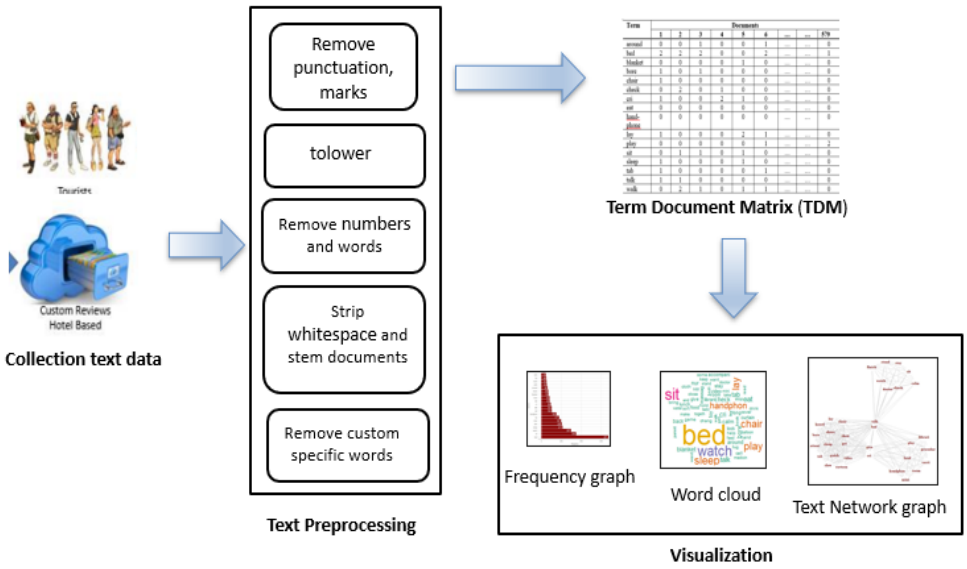

Fig. 1 Framework of Visualizing Keyword Relationship Analysis (VKRA)

From the cleaned corpus, the occurrence of each keyword was counted using a Term Document Matrix (TDM). In this study, the function *TermDocumentMatrix()* from tm package is applied to perform the analysis. TDM is defined as a m x n matrix that consists of the frequency of keywords in a collection of documents. TDM is developed based on the concept of TF-IDF, counting and ranking the keywords in the given content, followed by selecting keywords that occur more than the threshold. In a TDM, rows correspond to documents in the collection and columns correspond to terms. The result of keyword extraction from the 5501 documents is represented in the form of the 2D matrix.

C. *Visualising Text Analysis*

One of the popular ways to demonstrate the most frequent keywords in a corpus is through a visualisation of word cloud. Word cloud is a visualization technique used for representing a set of words in different sizes and colours. As shown in Fig. 2, the size of keywords corresponds to the frequency of the terms is observed. The keywords "room", "hotel", "stay", "good", "locat", "staff", "breakfast", and "play" are the top eight (8) most important keywords in 5501 text documents. Based on this initial visualization, it is observed that keywords such as "hotel", "room", "stay", "breakfast" "good" and "locat" (location) are the most frequently mentioned by travellers during their stay as their main concerns. These keywords help us to increase better understanding of the hotel review dataset.

Fig. 2 Sample of Word Cloud Adopted from [21]

Fig. 3 shows 20 keywords with the highest frequency terms in the dataset indicates that these keywords were frequently talked by the travellers. Based on the initial visualisation, a set of specific keywords that are found common in the dataset such as 'hotel', 'shop', 'mall', 'also', were removed. From the graph, the keywords such as "room" (8638), "stay" (5161), "locat" (3972), "good" (3852), "staff" (3692), "great" (2737), "breakfast" (2571), "servic" (2251), "time" (2078), "pool" (1910), "night" (1777), "check"

(1707), "friend"(1706), "walk" (1682), "clean" (1661), "bed" (1623), "nice" (1621), "station" (1550) , "food" (1544) and "well" (1440) demonstrates the highest frequencies of what customers shared and talked in their reviews. We can see that customers are more concerned about their room, stay and location compared to other facilities such as food and bed that were mentioned less from the analysis.

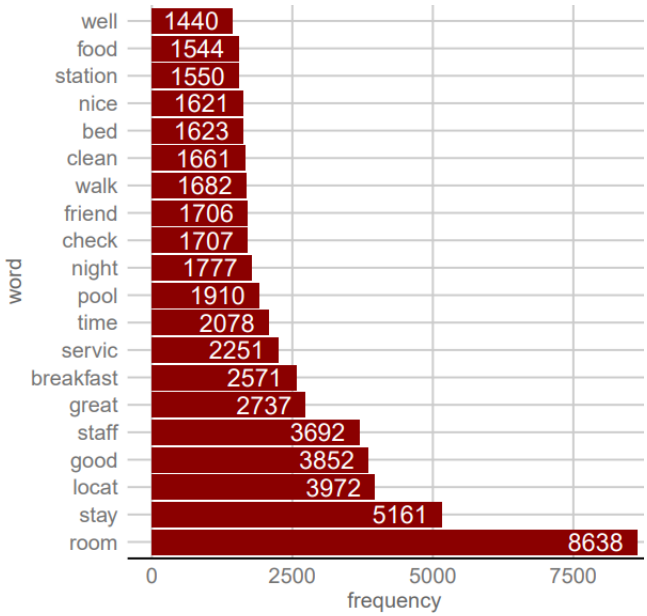

Fig. 3 Keyword Frequencies for Top 20 Words

TABLE 1
SAMPLE OF TERM DOCUMENT MATRIX (TDM)

	1	2	3	4	5	6	7	8	9	10	11
breakfast	0	1	1	0	0	0	1	0	1	0	0
clean	0	0	1	0	0	0	0	0	0	0	0
friend	2	0	0	0	1	1	0	0	0	0	0
good	0	0	1	0	0	0	0	1	0	0	2
great	0	0	0	0	0	0	0	1	0	1	0
locat	0	2	1	1	0	1	0	0	1	0	0
night	0	0	0	2	0	0	0	0	0	0	0
pool	0	0	0	0	0	0	0	0	0	0	0
room	0	3	1	4	1	0	1	0	0	0	1
sentral	0	0	0	0	0	0	0	0	0	0	0
servic	2	0	1	0	0	1	0	1	0	0	0
staff	3	2	0	0	2	1	0	0	3	1	0
stay	1	1	0	1	0	1	0	0	4	3	3
time	0	0	1	2	1	0	0	0	0	0	1

This experiment consists of 12110 keywords as rows and 5501 documents as columns with 100% sparsity. This size of generated TDM becomes huge where 100% of the rows are recorded with zero. This situation is common because there is a large number of unusual keywords, misspellings, etc. appear very infrequently in the corpus. Therefore, sparse terms need to be removed. With that, 75% of empty space is set in this experiment. Table 1 shows the results of removing sparse in TDM with 14 keywords with 61% of sparsity. This TDM table will be used as an input for visualizing association graphs and keyword network graphs in the later experiment. As shown in Table 1, the keyword "breakfast" appears in almost every document, except in documents 1, 4, 5, 6, 8, 10, 11 followed by keyword "servic" that exists 2 times in document 1, 1 time in document 3.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Various operations can be performed on TDM such as keyword association or is also known as text correlation in text analysis. Correlation is a statistical technique used to measure the co-occurrence of words in many documents [22]. For example, when keyword x appears, the other keyword y is associated with it [23] which is known as keyword pairings. In this experiment, we further explore the keywords associated with the keyword "locat" (location) based on the review. The keyword "locat" is randomly selected from the top 8 common keywords that influence travellers in the choice of hotels. This keyword is also in the top 3 keywords mentioned by travellers in their reviews indicates its importance and the main focus of the travellers. The created TDM table is used as an input to visualise this association graph. As such, findAssocs() function is applied to compute the correlation with every other keyword in a TDM.

```
#find associations
associations = findAssocs( tdm, 'locat', 0.15)
associations = as.data.frame( associations)
associations $ terms = row.names( associations)
associations $ terms <- factor( associations $ terms, levels = associations $ terms)

#plot the associations
ggplot( associations, aes( y = terms)) +
  geom_point( aes( x = locat), data = associations, size = 5) +
  theme_gdocs() +
  geom_text( aes( x = locat, label = locat), colour ="darkred", hjust = -0.25, size = 6) +
  theme( text = element_text( size = 20), axis.title.y = element_blank())
```

Fig. 4 Sample of R Script for Keyword Association Analysis for Keyword 'locat'

Based on the R script in Fig. 4, the search keyword is set as 'locat' and the minimum threshold for correlation is set as 0.15. The findAssocs() function produces a list of keywords that are most highly correlated with keyword 'locat' that meet with the minimum threshold. The result of codes in Fig. 4 is then visualised in the form of a graph using ggplot() function. The most common associated keywords from the hotel reviews dataset with keyword 'locat' such as "walk" (0.17), "area" (0.16), "bukitbintang" (0.16), "conveni" (0.16), "citi" and "station" were plotted in Figure 5. Based on these keywords, it can be confirmed that most travellers would prefer the location of hotels as their main reason for choosing hotels in Kuala Lumpur. This finding is aligned with some of the customers' feedback related to a location such

as: (1) "Located right in the heart of KL city. This hotel is a superb location for tourists and business stay approximately 500m from the Petronas Novotel that connects Petronas with a walkway making it very convenient to roam around this part of the city", (2) "The hotel is in a great central location in KL city", (3) "This stay was my 3rd times at this hotel because the good location", (4) "Location is great right in the middle of everything across the road from the Pavilion shopping mall and about 510 walk to the towers", (5) "Stayed here for a trade show at the nearby KLCC Convention Centre very convenient location with Pavilion Mall at the doorstep and the Convention Centre".

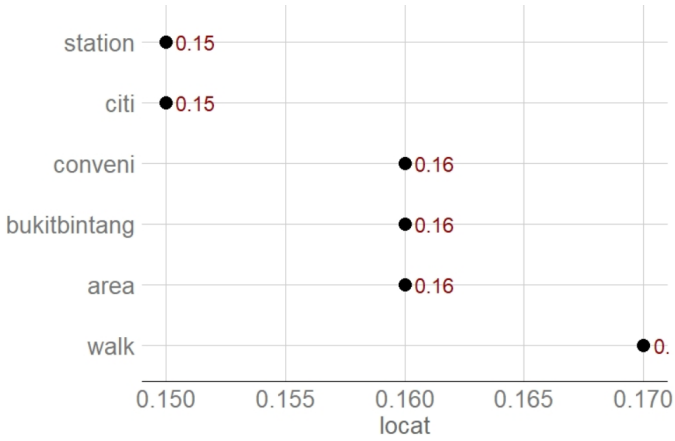

Fig. 5 An Associate Graph for Keyword 'locat'

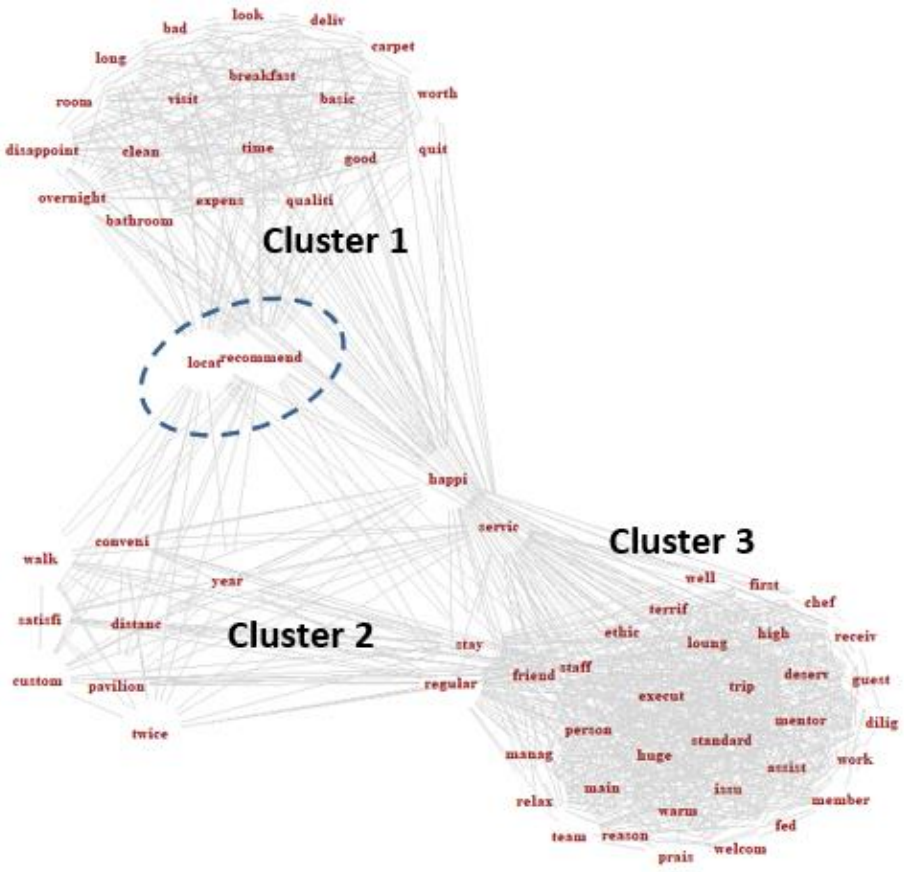

Fig. 6 A Network Graph for Keyword 'locat' for Three Clusters

Fig. 6 illustrates a graph of keyword 'locat' with 3 different keyword network clusters. The graph demonstrates that the keywords 'locat' and 'recommend' are linked to a group of keywords: Cluster 1 – 'time', 'bathroom', 'breakfast', etc.; and cluster 2 – 'walk', 'pavilion', 'conveni', 'satisfi' ,'twice', etc. The result connecting the two clusters shows a strong connection between the two keywords of "locat" and "recommend". This connection is shown in the dotted blue circle. This strong connection of keywords confirms the associative relationship between the keywords. It indicates that both clusters that shared the same keywords are interconnected and have relationships

that are affected by one another. These connections represent the node (keyword) "locat" is one of the most frequent factors that influenced the travellers in choosing their hotels. This relates to the observation that most of the reviews are positive (recommend) towards the hotel services.

V. CONCLUSIONS

In conclusion, tourists' opinions and reviews are important to all hoteliers to improvise their services and hospitality. These reviews and opinions can be retrieved from Travel websites such as TripAdvisor and Booking.com. In this paper, text (opinion) mining technique is applied on a dataset extracted from TripAdvisor containing reviews submitted by travellers who stayed at 4 four-star hotels in Kuala Lumpur. This experiment is conducted to analyse and deduce factors that influence travellers in their choice of hotels. The top frequent keywords mentioned in the hotel reviews are "room", "stay", "locat", "good", "staff", "great", "breakfast", "servic" indicates that the customers focus on these elements when they stay at the hotels. The analysis was extended by grouping and categorizing factors using correlation analysis to find clusters and relationships between the emerging keywords. The findings demonstrate that the keywords "locat", is one of the most frequent factors that influenced travellers in choosing their hotels. Furthermore, this relates to the impression that most reviews are positive (recommend) towards the hotel services. The study also enlightens some future works to improve the text mining process in the areas of the text mining tool, the representation of the results, and most importantly the final analysis.

REFERENCES

[1] B. N. Malaysia, "SME Annual Report 2007," National SME Development Council, 2008.

[2] R. Hirschmann. "Malaysia: number of hotels from 2009 to 2019 " https://www.statista.com/statistics/1004729/number-of-hotels-malaysia/#:~:text=Number%20of%20hotels%20in%20Malaysia%202009%2D2019&text=In%202019%2C%20there%20were%204.83%20thousand%20hotels%20in%20Malaysia (accessed Mac 20, 2021).

[3] M. Mariani, "Big data and analytics in tourism and hospitality: a perspective article," *Tourism Review*, 2019.

[4] D. Williams, "Connected CRM: Implementing a Data-Driven," *Customer-Centric Business Strategy, New York*, 2014.

[5] G. Dragosavac. "Big Data Analytics in Hotel Industry." http://goranxview.blogspot.com/2015/08/big-data-analytics-in-hotel-industry.html (accessed Mac 19, 2021).

[6] R. P. Ramsey and R. S. Sohi, "Listening to your customers: The impact of perceived salesperson listening behavior on relationship outcomes," *Journal of the Academy of marketing Science*, vol. 25, no. 2, p. 127, 1997.

[7] H. Öğüta and A. Cezara, "The factors affecting writing reviews in hotel websites," *Procedia-Social and Behavioral Sciences*, vol. 58, pp. 980-986, 2012.

[8] J. Liu and Y. Zhang, "Attention modeling for targeted sentiment," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 572-577.

[9] M. Bernard. "How Big Data and Analytics are Changing Hotels and the Hospitality Industry." https://www.forbes.com/sites/bernardmarr/2016/01/26/how-big-data-and-analytics-changing-hotels-and-the-hospitality-industry/?sh=621b0db51c22 (accessed April 1, 2021).

[10] A. S. H. Lee, T. M. Lim, S. K. Leow, and J. L. R. Aun, "A study on the use of "Yams" for enterprise knowledge sharing," in *2012 Second International Conference on Digital Information and Communication Technology and it's Applications (DICTAP)*, 2012: IEEE, pp. 183-188.

[11] H. J. Han, S. Mankad, N. Gavirneni, and R. Verma, "What guests really think of your hotel: Text analytics of online customer reviews," 2016.

[12] A. S. H. Lee, K. L. D. Chong, and K. H. Chan, "OpinionSeer: Text visualization on hotel customer reviews of services and physical environment," in *International Conference on Information Science and Applications*, 2018: Springer, pp. 337-349.

[13] K. H. Chan and A. S. H. Lee, "Voice of customers: Text analysis of hotel customer reviews (cleanliness, overall environment & value for money)," in *Proceedings of the 2017 International Conference on Big Data Research*, 2017: ACM, pp. 104-111.

[14] Z. Zainol, M. T. H. Jaymes, and P. N. E. Nohuddin, "VisualUrText: A Text Analytics Tool for Unstructured Textual Data," *Journal of Physics: Conference Series*, vol. 1018, no. 1, p. 012011, 2018, doi: 10.1088/1742-6596/1018/1/012011.

[15] Z. Zainol, S. Wani, P. N. E. Nohuddin, W. M. U. Noormanshah, and S. Marzukhi, "Association Analysis of Cyberbullying on Social Media using Apriori Algorithm," *International Journal of Engineering & Technology*, vol. 7, no. 4.29, pp. 72-75, 2018, doi: 10.14419/ijet.v7i4.29.21847

[16] M. O. Raza, M. Memon, S. Bhatti, and R. Bux, "Detecting cyberbullying in social commentary using supervised machine learning," in *Future of Information and Communication Conference*, 2020: Springer, pp. 621-630.

[17] Z. Zainol, P.N.E. Nohuddin, N. Rasid, H. Alias, and A. I. Nordin, "Relationship Analysis on the Experience of Hospitalised Paediatric Cancer Patient in Malaysia using Text Analytics Approach," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, pp. 72-79, 2019.

[18] D. Li, R. Rzepka, M. Ptaszynski, and K. Araki, "HEMOS: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media," *Information Processing & Management*, vol. 57, no. 6, p. 102290, 2020.

[19] K. Dashtipour, C. Ieracitano, F. C. Morabito, A. Raza, and A. Hussain, "An ensemble based classification approach for persian sentiment analysis," in *Progresses in Artificial Intelligence and Neural Systems*: Springer, 2021, pp. 207-215

[20] D. Contractor, T. A. Faruquie, and L. V. Subramaniam, "Unsupervised cleansing of noisy text," in *Coling 2010: Posters*, 2010, pp. 189-196.

[21] A. S. H. Lee, Z. Yusoff, Z. Zainol, and V. Pillai, "Know your hotels well! -- An Online Review Analysis using Text Analytics," *International Journal of Engineering & Technology*, vol. 7, no. 4.31, pp. 341-347, 2018.

[22] A. Kumar and A. Paul, *Mastering text mining with R*. Packt Publishing Ltd, 2016.

[23] T. Kwartler, *Text mining in practice with R*. John Wiley & Sons, 2017