A Latent Dirichlet Allocation Technique for Opinion Mining of Online Reviews of Global Chain Hotels

Amity International Business School
Amity University
Noida, India
anam24afaq@gmail.com

Loveleen Gaur
Amity International Business School
Amity University
Noida, India
lgaur@amity.edu

Gurmeet Singh
School of Business and Management
The University of the South Pacific
Suva, Fiji
singh_g@usp.ac.fj

Abstract—The hospitality industry has faced unprecedented challenges with the outbreak of Covid-19, which has changed customers' expectations. Therefore, it is essential to identify customers' new perceptions and expectations that lead to positive and negative opinions towards the service providers. Accordingly, this study aims to perform topic modeling and sentiment analysis on 94,200 online reviews of five global chain hotels in South Asia. Topic modeling as a text mining, unsupervised machine learning technique can decipher topics from a corpus such as online reviews, online reports, news covers, etc. In this study, the data is extracted from Trip Advisor through web scraping. Topic modeling is performed using the Latent Dirichlet Allocation (LDA) on the extracted data set to analyze the key topics mentioned by the customers in the online reviews. The analysis depicted that cleanliness, food, staff, and service were the main concerns of the hotel guests. Furthermore, the findings represented that the main issues impacting the hotel guests were service delays. However, food and services were the keywords with the maximum word count as depicted by topic modeling.

Keywords—Topic Modeling, Text Mining, Sentiment Analysis, Online Reviews, Hotels

I. INTRODUCTION

In today's era of technological advancement and the increasing trend of social media usage, the hospitality industry faces many challenges in maintaining its online image particularly after the outbreak of Covid-19 [1]. Moreover, the growing competition and rising hotel guests' expectations affect the hotels, as they have to retain their existing customers, attract new customers, and maintain a balance between services and pricing [2]. The study used Python for text mining and extracted 94,200 reviews from Trip Advisor, an online review site, to perform topic modeling and sentiment analysis. The reviews shared by the customers on online review sites also assist future customers in their decision-making process by formulating effective marketing strategies [3].

Topic modeling is rapidly gaining a lot of attention from researchers. This is an effective research tool that understands customers' in-depth ideas. These insights are helpful to make strategies as per the expectations of customers. Topic modeling as a technique is adequate for research trend analysis. The LDA-based topic modelling is most widely used as it clusters the topics and calculates the probability distribution on a set of words [4]. It shows the themes from the text data and identifies the hidden views that can comprehend the trend. The natural language processing (NLP) approach assists in sentiment analysis (SA) and depicts the customers' positive, negative, and neutral sentiments [5]. SA is a part of text mining and is also known as opinion mining [6]. The hotel-related research on customer reviews highlights big data analysis's relevance apart from handling the imbalanced data concerns [5].

Furthermore, although SA of online customer reviews is advancing, it is still in a preliminary stage. Therefore, this study adopted topic modeling and sentiment analysis to analyze the importance of hotel guests' positive, negative, and neutral reviews. The study aims to draw adequate conclusions from the data set about the factors affecting the customers' positive and negative experiences. Based on the reviews and ratings given by customers, the study depicts the sentiments of the customers towards the services provided by the hotels. The study findings will assist hospitality practitioners in their decision-making process by formulating effective marketing strategies as per the expectations and experiences of the customers.

II. LITERATURE REVIEW

Electronic word of mouth plays a vital role in creating the hotel's image and affecting its sales and profitability [8]. The customer reviews posted on social media platforms hold a lot of credibility for prospective customers, and businesses' engagement on social media platforms further impacts customers' purchase decisions [9]. The first-hand experiences shared by guests are considered impartial compared to the company's information [10]. Being a competitive sector, the hotels must pay attention to the online reviews of their hotels and devise suitable measures to upgrade their services [11]. Therefore the hospitality industry must undertake techniques like topic modeling and sentiment analysis to evaluate the factors affecting the image of their hotels [12]. Furthermore, Covid-19 has also created a lot of fear among individuals to trust the service providers; therefore, it is essential to use technologies to get insights into the changing expectations of the customers and formulate strategies to meet expectations effectively [13]. Furthermore, researchers are giving much attention to deep learning and topic modelling techniques in different domains like hospitality, healthcare, retail, etc [14]. Deep learning techniques are recently gaining a lot of
popularity during Covid-19 crisis [15]. Furthermore, topic modeling techniques are very effective for reducing employee churn as the assessment of employees positive and negative sentiments can help mitigate negative sentiments of employees [16]. The effectiveness of various machine learning techniques can bring a positive transformation for different service sectors [17].

Topic modeling is a part of machine learning techniques. It ascertains topics from predefined documents [18]. The ascertainment of topics provides the narratives of the dataset. [19] Adopted the topic modeling technique to analyze customer satisfaction in all continents during the initial outbreak of Covid-19. They formulated a new scale of metrics to analyze customer satisfaction. [20] Proposed a framework wherein the latent Dirichlet model (LDA) was applied to depict the satisfaction criteria of airline travelers using reviews from an online review site. [21] Recommended an inductive technique, known as the World2vec-based unique subject modeling technique, to perform the blockchain research analysis on the articles published on the blockchain during a timeframe of five years. Another research [22] suggested an LDA model that analyses essential subjects for the transport research domain. The themes depicted from the subjects are representative of crucial topics in the transport sector. [23] Utilized topic modeling on online reviews to show how unstructured data can effectively illustrate customer satisfaction and dissatisfaction towards the services offered by the service providers.

SA makes a sentiment lexicon with polarities representing positive, negative, and neutral sentiments. SA is also a machine learning technique that uses Natural Language Processing (NLP) to ascertain individuals' sentiments through their comments, tweets, and reviews [24]. SA technique is highly relevant for businesses for getting the customer's insights [25]. The sentiments expressed on social media platforms can assist in predicting major and minor fluctuations in businesses. [26] Conducted sentiment analysis on hotel reviews extracted from online review sites and found that the review star ratings on online review sites coincide with the title of online reviews. Deep learning techniques also play a substantial role in sentiment analysis. [27] Used deep learning models incorporating DNN, CNN, and RNN to analyze issues concerning sentiment analysis (aspect-based sentiment). The convolution layers scrape data from their inputs by merging many filters outputs. Many researchers have utilized deep learning models accompanying the TF-IDF for interpreting Twitter data. [28] Proposed a hotel recommendation system by analyzing the online customer reviews. The reviews were categorized using fuzzy logic, and the Random forest model and BERT were used for sentiment analysis. The model achieved more than 90% test accuracy.

Furthermore, [29] compared machine learning algorithms Naive Bayes and K-Nearest Neighbour (K-NN) for sentiment analysis for movie reviews and hotel reviews and found that Nave Baise gave more accurate results for movie reviews. However, both algorithms gave a nearly identical accuracy for hotel reviews. Another researcher [30] performed SA on customer reviews on online review sites using binary sentiment analysis and the weighted approach. This approach gave two lexicons (weighted lexicon and manually identified lexicon). Different methodologies validated these lexicon-based sentiment analyses to compare the accuracy metrics of the two lexicons. A study by [31] identified service determinants of robot hotels using sentiment analysis of customer reviews and found that robotic hotels' services determinants positively correlate with hotel guests' satisfaction.

To analyze the factors contributing to hotel guest satisfaction, the authors analyzed the reviews of customers posted on Tripadvisor, including frequency analysis, topic modeling, sentiment analysis, and word cloud.

### III. METHODOLOGY

Previous studies have stated Tripadvisor to be a valuable platform for analyzing customer reviews and sentiments, mainly for hospitality and tourism studies [32]. Therefore, we used Tripadvisor as a web portal to extract customer reviews. The authors used the extracted reviews for sentiment analysis and topic modeling. This study identifies the underlying meaning by looking into various text mining algorithms' findings, focusing on the online reviews of customers on Tripadvisor. The entire data was analyzed using the programming language Python. The technique adopted for topic modeling is segregated into three parts. First, the collection of data on which topic modeling is applied. The second is the preprocessing stage, wherein the unstructured data gets converted into appropriate data for topic modeling. The third stage involves the analysis of data. “Fig. 1” depicts the workflow of the methodology adopted for this study.

#### A. Data Collection and Preprocessing

Web scraping extracted 94,200 reviews of five global chain hotels in South Asia. Selenium library in Python was used to extract the review pages (unstructured data), and then beautiful soup library was used to gather the reviews from the unstructured HTML data. The reviews extracted are from September 2021 to December 2021. The entire dataset was
then saved in a .csv file. The next step was data preprocessing, which involves data cleaning, where the unnecessary clutter gets removed from the data set. For example, all the stop words and special characters are removed. Tokenization, stemming, and lemmatization are a part of the preprocessing stage. Tokenization involves tokenizing data into terms. The word variations were reduced by using stemming and lemmatization that converted the inflected words to a common base.

B. Data Analysis

The top 48 words derived from the frequency analysis were first identified in this stage. Next, a word cloud was made from the dataset to depict the complete visualization. The topic modeling analysis using the data mining technique (Latent Dirichlet Allocation) that illustrated the significant topics that emerged after reviewing the dataset. Lastly, the python package Text blob is applied to perform sentiment analysis to check the polarity. The sentiment analysis results depicted the hotel guests' positive, negative, and neutral sentiments towards the hotels.

IV. RESULTS

This section represents the results of the data analysis conducted for the study.

A. Frequency Analysis

The frequency analysis was utilized to identify the rating-review associations. Frequency analysis distinguished the positive and negative themes. Notably, the weight of the topic was ascertained by the following equation- as in:

\[ TW_{n,x} = \frac{w_{n,x}}{\sum_m w_{m,x}} \]

Where in “Eq. (1)”, \( TW_{n,x} \) is nth Topic Weight in document x, and \( w_{n,x} \) depicts the number of times that word in the document x is assigned to nth Topic.

TABLE I: Frequency Analysis Results

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotel</td>
<td>71101</td>
<td>stay</td>
<td>67989</td>
<td>staff</td>
<td>61422</td>
</tr>
<tr>
<td>rooms</td>
<td>55980</td>
<td>services</td>
<td>54018</td>
<td>good</td>
<td>53365</td>
</tr>
<tr>
<td>food</td>
<td>47495</td>
<td>thank</td>
<td>37080</td>
<td>great</td>
<td>35432</td>
</tr>
<tr>
<td>experience</td>
<td>25534</td>
<td>excellent</td>
<td>21244</td>
<td>Like</td>
<td>21244</td>
</tr>
<tr>
<td>nice</td>
<td>20361</td>
<td>amazing</td>
<td>20274</td>
<td>best</td>
<td>19748</td>
</tr>
<tr>
<td>restaurants</td>
<td>17262</td>
<td>special</td>
<td>17029</td>
<td>breakfast</td>
<td>16978</td>
</tr>
<tr>
<td>helpful</td>
<td>16498</td>
<td>hospitality</td>
<td>15868</td>
<td>cleanliness</td>
<td>13108</td>
</tr>
<tr>
<td>chef</td>
<td>13068</td>
<td>comfortable</td>
<td>12555</td>
<td>care</td>
<td>11928</td>
</tr>
<tr>
<td>friendly</td>
<td>11308</td>
<td>enjoy</td>
<td>10465</td>
<td>wonderful</td>
<td>10279</td>
</tr>
<tr>
<td>ambiente</td>
<td>10017</td>
<td>check-in</td>
<td>8826</td>
<td>beautiful</td>
<td>8809</td>
</tr>
<tr>
<td>view</td>
<td>8264</td>
<td>awesome</td>
<td>7787</td>
<td>feel</td>
<td>7414</td>
</tr>
<tr>
<td>housekeeping</td>
<td>7414</td>
<td>buffet</td>
<td>7289</td>
<td>Location</td>
<td>7083</td>
</tr>
<tr>
<td>Pool</td>
<td>7041</td>
<td>delicious</td>
<td>6795</td>
<td>dinner</td>
<td>6781</td>
</tr>
<tr>
<td>manager</td>
<td>6608</td>
<td>bar</td>
<td>6371</td>
<td>guest</td>
<td>6370</td>
</tr>
<tr>
<td>love</td>
<td>6284</td>
<td>reception</td>
<td>5917</td>
<td>help</td>
<td>5789</td>
</tr>
<tr>
<td>covid</td>
<td>5509</td>
<td>memorable</td>
<td>5474</td>
<td>happy</td>
<td>5224</td>
</tr>
</tbody>
</table>

B. Word Cloud

The word cloud visually depicts the most frequent words expressed in the dataset. Word cloud provides visual insights that can offer more in-depth insights from the dataset by denoting the corpus in association with its frequency. The word cloud effectively helps spot the top words at first glance compared to spotting the keywords from the table. The findings of the word cloud are depicted in “Fig. 2”. This word cloud illustrates that the global chain hotels in South Asia have mostly received a positive review by the hotel guests in the form of visible words like 'good,' 'happy,' 'food,' 'thank,' 'service,' 'clean' and, 'comfortable.' On the other hand, the negative keywords mainly focus on 'delayed,' 'poor,' and 'overcrowding.' However, as only top 200 words are utilized to make the word cloud, therefore negative keywords are not visible due to low frequency. The neutral expressions depicted by the word cloud are 'hotel,' 'stay,' and 'staff.' The word cloud reflects that most hotel guests are satisfied with the food, cleanliness, and comfort provided by the global chain hotels in South Asia.
C. Topic Modeling

Topic modeling is a part of machine learning techniques. It ascertains topics out of predefined documents. The ascertaining of topics provides the narratives of the dataset. We adopted the Latent Dirichlet Allocation technique (LDA), which classifies the given text into relevant topics to provide insights into the frequency of words used by different individuals to offer their reviews. LDA as an approach brings up the latent semantic structures from a cluster of text. LDA uses Bayesian learning for extracting unnoticed constructs. This executes in two steps. First, the relevant topics are allocated to the documents, and second, a probability distribution is assigned to the words of every topic.

Developing Topic Modeling using LDA: It is a statistical model in NLP for topic modeling. It adopts matrix factorization approach to depict a particular corpus specified as a “document-term matrix”. The matrices are processed using sampling techniques till they achieve a steady point. The combined distribution of unseen and calculated variables of LDA are presented in “Eq. (2)”: \[
p(B_{1:k}, \theta_{1:d}, z_{d:m}, w_{1:d}) = \prod_{k=1}^{K} p(B_k) \prod_{d=1}^{D} p(\theta_d) \prod_{m=1}^{M} p(z_{d:m}) p(w_{d,m} | B_k, Z_{d:m}) \]  

(2)

Where \( B_{1:k} \) are topics, where \( B_k \) is a distribution over the words, \( \theta \) is Document-topic distribution parameter, \( \theta_{1:d} \) is the topic proportion for topic \( k \) in document \( d \), \( z \) is the word-topic assignment for the \( m \)-th word in document \( d \), \( w \) is the observed word \( w_{d,m} \) is the \( m \)-th word in document \( d \)

This distribution depicts interrelationships between the topics assigned and the topic distribution, terms in all the documents along with their dependence on the topics and topics assigned.

The LDA model is enacted on the TripAdvisor dataset of 94, 200 customer reviews. For calculations, we passed 100 samples with 12 topics, A 0.01, B, 0.1 through 550 iterations. The model generates the topic distribution. “TABLE II” presents the sample topics and word’s distribution with regards to each topic. The authors identified the keywords from the prevalent topics. The topics are labeled to make them more transparent, as depicted in “TABLE II”. These topics have sample themes of ten words. The topics are—conversation, responses during Covid-19, hotel stay experience, Rooms conditions, aesthetics, booking facility, welcome drinks, view, festivities experience, ambience, hotel architecture, and daily activities. All twelve dominant topics are identified by adopting the topic modeling technique. LDA provides relevant topics having the maximum weights. The topics are presented by grouping the keywords with the maximum possible occurrence, distinguishing all topics from one another.

<table>
<thead>
<tr>
<th>ID</th>
<th>Probability distribution</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.047&quot;good&quot; + 0.026&quot;hotel&quot; + 0.014&quot;food&quot; + 0.026&quot;hotel&quot; + 0.014&quot;amazing&quot; + '0.012&quot;amazing&quot; + '0.012&quot;staff&quot; + '0.011&quot;time&quot; + '0.010&quot;satisfactory&quot; + '0.009&quot;breakfast&quot;</td>
<td>hotel stay experience,</td>
</tr>
<tr>
<td>1</td>
<td>0.029&quot;hotel&quot; + 0.026&quot;staff&quot; + 0.025&quot;excellent&quot; + 0.017&quot;call&quot; + '0.016&quot;mice&quot; + '0.014&quot;great&quot; + '0.014&quot;reservation&quot; + '0.012&quot;hospitality&quot; + '0.011&quot;process&quot; + '0.011&quot;listen&quot;</td>
<td>Conversation</td>
</tr>
<tr>
<td>2</td>
<td>0.019&quot;good&quot; + 0.016&quot;best&quot; + 0.015&quot;traditional&quot; + 0.014&quot;stay&quot; + 0.013&quot;hotel&quot; + '0.013&quot;stylish&quot; + '0.011&quot;street&quot; + '0.010&quot;modern&quot; + '0.010&quot;great&quot; + '0.008&quot;place&quot;</td>
<td>aesthetics</td>
</tr>
<tr>
<td>3</td>
<td>0.018&quot;staff&quot; + 0.015&quot;fast&quot; + 0.014&quot;good&quot; + 0.014&quot;hotel&quot; + 0.014&quot;rush&quot; + '0.013&quot;waiting&quot; + '0.011&quot;helpful&quot; + '0.011&quot;room&quot; + '0.011&quot;reception&quot; + '0.010&quot;also&quot;</td>
<td>booking facility</td>
</tr>
<tr>
<td>4</td>
<td>0.015&quot;staff&quot; + 0.014&quot;beer&quot; + 0.014&quot;drinks&quot; + 0.012&quot;arrival&quot; + '0.010&quot;stay&quot; + '0.010&quot;friendly&quot; + '0.008&quot;water&quot; + '0.008&quot;hotel&quot; + '0.008&quot;hot&quot; + '0.008&quot;great&quot;</td>
<td>Welcome drinks</td>
</tr>
<tr>
<td>5</td>
<td>0.012&quot;hotel&quot; + 0.011&quot;good&quot; + 0.011&quot;swimming&quot; + 0.009&quot;relaxing&quot; + 0.009&quot;shopping&quot; + '0.008&quot;special&quot; + '0.008&quot;overall&quot; + '0.007&quot;natur&quot; + '0.007&quot;view&quot; + '0.007&quot;noisy&quot;</td>
<td>view</td>
</tr>
<tr>
<td>6</td>
<td>0.022&quot;hotel&quot; + 0.017&quot;club&quot; + 0.019&quot;stay&quot; + 0.007&quot;great&quot; + '0.007&quot;activities&quot; + '0.007&quot;rooms&quot; + '0.007&quot;good&quot; + '0.007&quot;beautiful&quot; + '0.006&quot;memorable&quot; + '0.006&quot;food&quot;</td>
<td>festivities experience</td>
</tr>
<tr>
<td>7</td>
<td>0.013&quot;staff&quot; + 0.013&quot;classy&quot; + '0.012&quot;attractive&quot; + '0.012&quot;rooms&quot; + '0.011&quot;soothing&quot; + '0.009&quot;beautiful&quot; + '0.009&quot;room&quot; + '0.009&quot;good&quot; + '0.009&quot;hotel&quot; + '0.009&quot;music&quot;</td>
<td>Ambience</td>
</tr>
<tr>
<td>8</td>
<td>0.016&quot;hotel&quot; + 0.013&quot;good&quot; + 0.013&quot;staff&quot; + 0.009&quot;guard&quot; + '0.007&quot;lift&quot;</td>
<td>hotel architecture</td>
</tr>
</tbody>
</table>
Sentiment Analysis (SA)

SA is also a machine learning technique that uses Natural language Processing (NLP) to ascertain individuals' sentiments through their comments, tweets, and reviews. SA depicts the sentiments as positive, neutral, and negative. This technique is highly relevant for businesses for getting the customer's insights. The sentiments expressed on social media platforms can assist in predicting major and minor fluctuations in businesses. The two prime sentiment analysis methods are supervised machine learning and lexicon-based sentiment analysis. We adopted the lexicon-based sentiment analysis for this study using the python package TextBlob.

The visualization of LDA results are illustrated in “Fig. 3” in the form of an intertopic distance map that aids in visualizing the topics and keywords. The various bubbles on the map are the depiction of the topics. LDA provides relevant topics having the maximum weights. The topics are presented by grouping the keywords with the maximum possible occurrence, distinguishing all topics from one another. The prominence of the topics depends upon the size of the bubbles. Further, the overlap of bubbles depicts the presence of common keywords in the overlapping topics. The bar chart attached with the map in LDA visualization illustrates the local frequency of keywords found in topic 1.

The growing use of social media for expressing views after availing services is gaining a lot of attention from the service providers. As one of the dominant service sectors, the Hospitality Industry has understood the significance of big data available on social media platforms that can assist them in formulating strategies and meeting customers' growing expectations. This study adopted topic modeling and sentiment analysis as powerful techniques to analyze the concerns and sentiments of the hotel customers towards the global hotel chains of South Asia. The study also provides certain limitations. The first limitation of this study is using only one online review platform TripAdvisor for the study. However, future research can consider multiple online review platforms for validating the results of this study. Second, this study is undertaken for only a limited period. Future studies can include a more considerable period to get more detailed insights of the customers, including pre and post-pandemic reviews. This research can also be extended to various other sectors that can use the advantage of big data to understand their customers by analyzing their views and providing more customer-centric services.

Overall reviews depicting a positive sentiment towards the hotels were 46.73%, negative sentiments accounted for 15.09%, and neutral sentiments were 38.18%. The study's findings show that global chain hotels can deliver customer satisfaction significantly despite the uncertainties and challenges that occurred due to the pandemic. The findings also reflect that mostly the negative reviews are due to some delayed services given by the hotels. The keywords like happy, good, clean, comfortable, and friendly reflect that the hotel guests are satisfied with the services.

![Figure 3. Top 12 Topics LDA Visualization](image)

![Figure 4. Sentiment Analysis](image)

V. CONCLUSION AND LIMITATIONS

REFERENCES


