User Generated Big Data Analysis of Customer Ratings of Beaches in Andaman and Nicobar Islands of India

Anagha Prem, Sangeetha Gunasekar, Dileep G Menon

Abstract: In this study, we analyse user-generated content of beach tourist attractions across Andaman and Nicobar Islands in India to find the factors that influence tourist experience in terms of customer satisfaction in island beach destinations. The study is conducted to identify the specific attributes of beach tourism that has positive and negative influence over the high rating provided on online review portals. The study is conducted over the reviews obtained from online reviews over Trip Advisor. Text analytics and logistic regression is used to find the various attributing factors that affect the customer satisfaction in the form of rating. The online data is collected and the unique words are identified with respect to the frequency of occurrence of the unique words in the reviews and taken for analysis. The results indicate that there is more influence of factors such as cleanliness of the beach, the accommodation facilities available and the various beach activities in inducing better satisfaction among the tourists than the scenic beauty of the place. The study has policy implications as the tourism development activities undertaken by the government of Andaman & Nicobar Islands must focus on the positive influencing factors and focus should be given on the factors such as better restaurant facilities, crowd management and shopping facilities in order to enhance the tourist experience at Andaman & Nicobar Islands.

Index terms- Beach tourism, online review analysis

I. INTRODUCTION

Academic research has focused on island tourism since several years [1] [2]. While themes covering economic impact of foreign revenue earned [3][4], sustainability of island tourist destinations [5], coastal scenery assessments [6], cruise experience [7][8], eco-tourism and resulting conflicts between human settlements, increasing tourist footfalls and ecological sustainability of islands [9] have been the focus of several studies in the literature, not many have analysed the reactions of tourists visiting different tourist attractions across an island destination using user-generated content analysis [10]. Tourism industry with a revenue of US$733 billion and employment at 8% of global workforce, is expected to see over 1.6 billion international tourists by 2020 (UNWTO, 2008). Beaches are the most popular tourist attractions in islands. Many studies highlight the importance of beaches as a tourist destination in the literature [10]; [11]. The studies pertaining to Andaman & Nicobar Islands of India are scant in the literature.

One of the few studies that have focused on these islands [12] estimate the carrying capacity of beaches at Andaman & Nicobar Islands and find that while the islands tourist carrying capacity is well within the limits, accommodation and other infrastructure needs to be improved to at least match the effective carrying capacity of these islands. Others like [13] find these islands to have a high potential for ecotourism. The Andaman & Nicobar Islands of India are bestowed with beautiful coastline and sandy beaches and are a growing attraction for tourists across the world. The present study use the user generated content (online reviews from Trip Advisor) of tourist attractions across Andaman & Nicobar Islands to understand the factors that influence tourist experience in terms of customer satisfaction. These islands of India are endowed with some of most beautiful white-sand beaches, mangrove forests, a large number of rare species of flora and fauna [14]. It is a growing tourist destination of India comprising of over 572 islands, islets and rocks, of which only 38 are inhabited. Most of these islands have forest cover with substantial parts under reserve and protected forest area that are earmarked as primitive tribal reserves. Contact with local tribes is restricted. The islands hold the potential to grow into one of the most attractive tourist destinations of India. The inflow of tourists increased over the years as clearly seen in Figure 1 with domestic tourists showing stronger increasing growth trends. This increasing trend of domestic as well as foreign tourism can be capitalised for the growth of the tourism industry and there by the economy. The data was sourced from Basic Statistics 2016-17, Directorate of Economic and Statistics, Andaman and Nicobar Administration.

Fig1: Tourist arrival (foreign and domestic) at Andaman and Nicobar Island
LITERATURE REVIEW
The importance of social media is well established in the tourism industry. Tourists share their travel and leisure experiences at a tourist destination in the form of reviews, videos and photographs on various social media platforms like Facebook, twitter, Instagram and on travel websites like TripAdvisor, among others. While tourists are seen to contribute to this unbiased information repository on social media, they also benefit from the experience of others. Several studies in the literature have found tourists to rely on the reviews posted by other tourists on social media to make their decisions [15][16]. These decisions involve decisions regarding where to spend the next holiday, where to stay, what to eat and what things should be done while visiting a tourist destination. Several studies in the literature have focused on the online user-generated contents of tourists to understand what factors influence the tourists to book a hotel [17], their stay at the hotel [17], or eating at a restaurant [18]. While few studies [19] have focused on understanding the online written reviews of destination, specifically studies relating to factors impacting consumer satisfaction are not seen in the literature (to the best of our knowledge). Our study attempts to identify the factors that influence tourist satisfaction from a visit to a beach. This study is done in the context of island tourist destination, Andaman & Nicobar islands of India.

Island as tourist destination attract tourists in large numbers due to their remoteness, environmental beauty and serenity, sense of exclusion from other populated places [20], many a times leading to circumstances where tourists outnumber the local island population. These special features of tourism with unique cultural and natural elements make island destination an attractive academic research area. Thus this study explores the various attributes of beach tourism aspect of island tourism contributing towards customer satisfaction of tourists with respect to Andaman & Nicobar islands of India.

Research Gap
The study tries to focus on the various attributes that influence the customer satisfaction with reference to the customer rating provided for the beach tourism of Andaman & Nicobar Islands of India.

DATA
The data for the study was collected from TripAdvisor.in, a popular travel and destination website that hosts the reviews of tourists. It is one of the largest such websites where over 460 million user reviews are hosted for over 7 million hotels, restaurants, vacation rentals and tourist attractions. It provides the travellers with wisdom of crowd. A Web crawler program was written in R, an open-source programming language, to scrape the reviews. Each row of the data scraped had the Rating, Review, and Quote or the heading of the review. All reviews as on 31st March 2018, that was posted in TripAdvisor.in for Andaman & Nicobar ‘beaches’ has been scrapped. In all there were 5760 reviews that are collected for the proposed study.

Word clouds:

Word cloud was constructed for the 100 most frequently used words from the reviews. Word clouds are extensively used in academic research [21], [22]. They are majorly used to represent the most frequently occurring words in text analysis [23]. The word-size in the word cloud corresponds to its frequency in the total words.

Fig 2: Word Cloud for Beaches
This gives a glimpse of the major factors that influence beach tourism. As seen from the Figure 2, “cleanliness”, “beautiful”, “good”, “sunset” and best topped the list for most frequently used words in the review followed by visit and time.

To understand the text reviews, the reviews were taken from the TripAdvisor website using R programming. Most frequently used words in the reviews were mapped to the selected attributes identified from the literature. Further the details of cleaning the data and mapping to various attributes identified from the literature is given in Fig 3.

IV. EMPIRICAL METHODOLOGY
Given that the dependent variable takes only values of 1 or 0, logistic regression is used for the estimation. The dependent variable is customer satisfaction defined as:

\[ y = 1 , \text{ if customer rates the experience as “Excellent” (Rating } = 5 ); \]

\[ y = 0 , \text{ otherwise} \]
Fig 3: Research Methodology

Table 1: Words included under each attribute

<table>
<thead>
<tr>
<th>S No</th>
<th>Attribute</th>
<th>Words included</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clean</td>
<td>“clean,cleaned,cleaning,cleanly,cleaness,clear,crystal,crystals,neat,pristine”</td>
</tr>
<tr>
<td>2</td>
<td>Time</td>
<td>“sunset,’sunset’,sunsets,time,timely,times,timing,timings,eve ning,eve nings,morning,mornin gs,sunrise,sunrise,sunrises”</td>
</tr>
<tr>
<td>3</td>
<td>Scene</td>
<td>“forest,forests,scenic,coconut,’coconut,coconuts,greenery,tree,trees,view,viewed,viewing,views”</td>
</tr>
<tr>
<td>4</td>
<td>Crowd</td>
<td>“people,people,crowd,crowd,cleared,cleared,crowding,crowds,quite,quiet”</td>
</tr>
<tr>
<td>5</td>
<td>Stay</td>
<td>“room,rooms,resort,resorts,stay,stay,stayed,staying,stays”</td>
</tr>
<tr>
<td>6</td>
<td>Value for money</td>
<td>“worth,worthed”</td>
</tr>
<tr>
<td>7</td>
<td>Food</td>
<td>“food,foodings,foods,restaurant,restaurants,restaurant,restaurants”</td>
</tr>
<tr>
<td>8</td>
<td>Activity</td>
<td>“walk,walked,walking,walks,swim,swimmed,swimming,swims,ride,rides,riding,sport,sporting,sports”</td>
</tr>
<tr>
<td>9</td>
<td>Shop</td>
<td>“shop,shopped,shopping,shop pings,shops,’shops”</td>
</tr>
</tbody>
</table>

A. Logistic Regression

Logistic regression estimation is a special case of linear regression models where the Y variable is categorical. It helps compute the probability of occurrence of an event. Dependent variable is log of odds ratio with equation forms as:

\[ f(E(y)) = \alpha + \beta x_1 + \gamma x_2 + \epsilon \]

Here, \(f()\) is the link function, \(E(y)\) is the expectation of target variable and \(\alpha + \beta x_1 + \gamma x_2 + \epsilon\) is the linear predictor (\(\alpha,\beta,\gamma\) to be predicted) and \(\epsilon\) is the error term. The role of link function is to ‘link’ the expectation of \(y\) to linear predictor. In the study the dependent variable of customer satisfaction is dichotomous, as only the highest rating is taken into consideration. The unique words are grouped into constructs specified by the attributes and these are taken as the independent variables.
The following logistic regression model was used to test the model:

\[
\text{Customer Satisfaction} = \beta_1 + \beta_2 \text{clean} + \beta_3 \text{time} + \beta_4 \text{scene} + \beta_5 \text{crowd} + \beta_6 \text{stay} + \beta_7 \text{value for money} + \beta_8 \text{food} + \beta_9 \text{activity} + \beta_{10} \text{shop} + \epsilon
\]

Where \(\epsilon\) is standard error.

The above model was estimated using logistic regression and the results of the estimate are given in Table 2.

**Table 2: Results of logistic regression analysis**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Coefficient</th>
<th>Odds Ratio</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>.762</td>
<td>2.142</td>
<td>0.000*</td>
</tr>
<tr>
<td>Time</td>
<td>.026</td>
<td>1.026</td>
<td>0.643</td>
</tr>
<tr>
<td>Scene</td>
<td>.078</td>
<td>1.081</td>
<td>0.185</td>
</tr>
<tr>
<td>Crowd</td>
<td>-.282</td>
<td>.754</td>
<td>0.000*</td>
</tr>
<tr>
<td>Stay</td>
<td>.144</td>
<td>1.155</td>
<td>0.051**</td>
</tr>
<tr>
<td>Value for money</td>
<td>-.175</td>
<td>.839</td>
<td>0.144</td>
</tr>
<tr>
<td>Food</td>
<td>-.314</td>
<td>.730</td>
<td>0.001*</td>
</tr>
<tr>
<td>Activity</td>
<td>.183</td>
<td>1.200</td>
<td>0.006*</td>
</tr>
<tr>
<td>Shop</td>
<td>-.413</td>
<td>.661</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

Notations of *, ** imply significance at 10%, 5% and 1% level for \(p\) respectively.

V. RESULTS

According to the model, various variables seem to impact the customer rating given the beach fronts of Andaman. From the table we see that, people who find the beaches clean tend to rate higher than the scenic beauty offered or the time of the visit. The restaurant facilities and shopping facilities available to the customers tend to have significantly negative effect on the rating provided to the beaches. The behaviour of the people as well as the intensity of the crowd seems to have significantly negative influence as well. On the other hand the quality of accommodation and stay provided at the beach tends to prompt the customers to give a higher rating. The various beach activities such as swimming etc. tends to have significantly a positive effect on the higher rating of the beaches. Contrary to normal assumptions value for money does not seem to significantly contribute to the higher rating provided by customers.

VI. CONCLUSION

In the hospitality literature, studies have documented the use of websites and social media in providing useful information in terms of review of experiences shared by travellers for destination, property, facilities and restaurants [24]. The reviews and opinions which were recorded in traditional form is now digitalized and shared globally. Online reviews of a destination not only attract or detract a traveller, but also develop consumer expectations [25]. The present study used online reviews to analyse the various attributes that impact customer satisfaction for beach tourism. Our conclusion from the study indicates that the attributes contributing towards the higher rating given by customers are factors such as cleanliness, various beach activities and the resort facilities provided in the island which act as positive influence over the tourist satisfaction, whereas the poor state of restaurant facilities, the crowd and the lesser opportunities for shopping tends to have negative effect over the same. Therefore, government and private business dependent on beach tourism should focus on the above factors in order to enhance the tourist experience at Andaman & Nicobar Islands.

VII. LIMITATIONS

The scope of the study in this article has been limited to reviews taken from Trip Advisor. The study conducted poses the limitation of the focus being given to only 15 beaches of Andaman & Nicobar Islands. The study can also be broadened by doing a large-scale research with data collected from an extensive range of social media and other forums. Further study can be conducted with a broader sample and area to be focused upon.

REFERENCES


AUTHORS PROFILE

Anagha Prem, final year MBA student at Amrita School of Business Coimbatore, pursuing marketing and human resource management after graduating in computer science and engineering.

Dr. Sangeetha Gunasekar is currently working as the Associate Professor at Amrita School of Business, Coimbatore, India.

Dileep G Menon is a research scholar at Amrita School of Business, Coimbatore, India.