Analyzing Visitors' Review of Homestays Located in Nature-Based Settings: An NLP Based Approach

Received: 16th October 2021
Review: 4th December 2021
Accepted: 12th Feb 2022

Gopi Nath Vajpai[1], Debashis Pattanaik[2]

Abstract

Objective: Sentiment analysis techniques such as Natural Language Processing (NLP) provide a powerful tool to analyze textual data. Along with machine learning and other big data methods, these techniques are used in improving customer service quality in different sectors. This paper utilizes sentiment analysis techniques to identify key themes surrounding visitors' homestay experience in nature-based settings.

Methodology/Approach/Scope: Analysis of 2369 TripAdvisor reviews through Structural Topic Modeling (STM) reveals how high rated homestay experiences differ from those rated low on various parameters.

Findings/ Implications/Conclusion – The paper contributes to the knowledge on text mining & its application in improving customer service in the hospitality & tourism domain. The research has practical usage for homestay stakeholders and future direction for further research.

Keywords: *Sentiment Analysis, Text Analysis, TripAdvisor, Topic Modeling, Homestay, Customer Review*

1. Introduction

In the age of web 2.0 culture, customers constantly look for feedback left by fellow users to make better purchase decisions. It identifies customer-driven content generated on different social media platforms on the internet. With quick access to products and services through internet access on handheld devices, customers often look for opinions from fellow users of similar demographic profiles and lifestyles (Misner, 1994). As more user-generated content dominates the internet, businesses

SVKM'S NMIMS Deemed to be UNIVERSITY
NMIMS Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022

https://doi.org/10.53908/NMMR.300201

must analyze consumer reviews and feedback. Thus, businesses analyze customer reviews to improve product/service quality and offer personalized experiences. Similarly, hotels & resorts regularly try to assess and maintain their market image by analyzing customer feedback from travel websites, travel review pages (e.g. www.tripadvisor.in), blogs, social media sites and other similar platforms.

Homestay accommodation is a subset of community-based tourism which Goodwin and Santilli (2009) define as "tourism owned and/or managed by communities and intended to deliver wider community benefits." Thus many travelers willing to experience simple rural life choose a homestay to have a meaningful learning experience through cultural exchange with host family (Boonratana, 2010). Bose and Biju (2020) suggest that homestay experiences are considered unique and highly satisfying by visitors compared to visiting a resort or hotel.

Tourism research has highlighted that traveler consult online reviews portals before finalizing their travel itinerary as it helps them get new ideas (Gretzel et al., 2007).

Advancements in Natural Language Processing (NLP) and software capability to analyze qualitative data has enabled researchers to analyze textual data. This form of analysis is also called sentiment analysis or content analysis, and these terms will be used interchangeably throughout this paper. Applications of sentiment analysis techniques are numerous, spanning all fields such as e-commerce, retail, airlines, movie and TV show ratings, stock market, election campaigns, social media, etc. (Feldman, 2013).

2. Objectives

Customer reviews are rich source of information regarding business performance and brand image. This information can be used to understand and extract themes about product and services. A limited number of studies have tried to understand visitors's review of homestay experience in nature based locations. This paper thus aims to contribute to tourism research by analyzing online homestay reviews using text analysis techniques. It specifically aims to answer following research questions:

- What are the central themes surrounding visitors' homestay experience?
- What are the key differences in visitors' positive and negative opinion towards homestay services?

3. Literature Review

Several researchers have tried to develop a system for analyzing consumer opinions and online reviews using content analysis techniques. Kasper & Vela (2011) developed a web-based opinion mining technique to help hotels in Germany analyze

SVKM'S NMIMS Deemed to be UNIVERSITY
NMIMS Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022

customer comments. Barreda and Bilgihan (2013) analyzed customers' reviews of hotel experience using content analysis technique. A similar study involving lexicon-based sentiment analysis of customer reviews in Spanish calculated a global sentiment score based on the frequency of negative and positive keywords matching the lexicon database (García, Gaines & Linaza, 2012). Gräbner et al. (2012) involved creating a domain-specific lexicon in the field of tourism. The lexicon was subsequently tested on test data to validate the proposed approach. Calheros, Moro and Rita (2017) performed topic modeling with a refined approach to review analysis wherein they used a custom dictionary with Latent Dirichlet Allocation (LDA) algorithm. The study concludes that food quality inspires positive sentiments among travelers, while hospitableness generates strong positive feelings.

More recently, Thapa, Malini and Manavi (2018) attempted to identify the most talked about attributes of homestay from online tourist reviews in nine Indian states. Apart from classifying the reviews as positive and negative, they further tried to distinguish between the type of emotion viz. fear, joy, surprise etc. In yet another study, Park et al. (2018) utilized 105126 user reviews to predict customer revisit behavior and find parameters associated with revisit intention towards hotels. The study reveals a significant difference in the review style of one-time visitors compared to repeat visitors.

Ray, Garain and Sarkar (2021) have created a recommendation system by using TripAdvisor reviews by categorizing hotel reviews. The outcome is an example of real life application of sentiment analysis & artificial intelligence tools.

4. Research Design

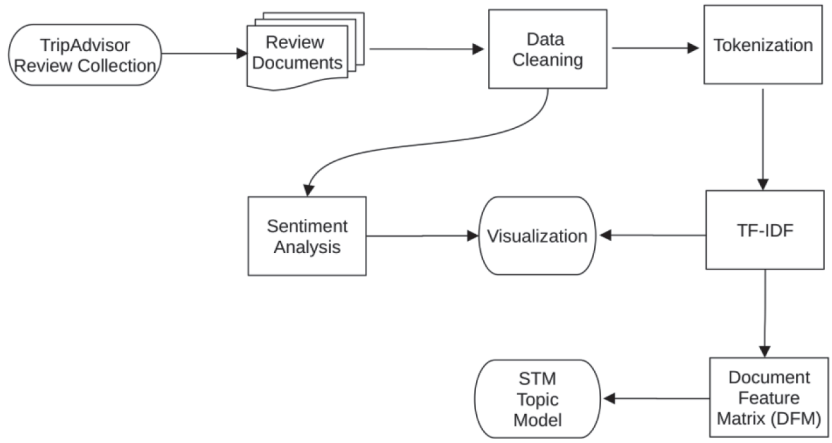
The method used in this paper is divided into following broad topics, steps for which are briefly explained below.

4.1 Data Collection:

In this part, we collected data from TripAdvisor website using python programming. The aim was to get reviews for all homestays in nature-based locations of Uttarakhand region. The website was chosen as it is widely used by travelers to post reviews about hotels & restaurant throughout the world. Moreover, we were unable to find any other website/source containing significant number of reviews for homestays in our study area. The purpose of choosing only nature-based homestays was to have homogeneity within samples as otherwise the analysis will not be robust.

The fields extracted from the webpage consisted of four parts: Homestay name, stay date, star rating (from 1 to 5 star) and the review text in a tabular form. We collected 2369 reviews which were then visually screened for any missing information or missing fields.

Figure 1: Research Design



4.2 Pre-Processing: In the next stage, we firstly calculated sentiment at review level using "sentimentr" algorithm. A positive sentiment score indicates a favorable visitor experience while score below zero indicate negative polarity. This sentiment value was then compared and plotted against the star rating to have a quick picture of the overall dataset. The next step involves cleaning up the review data. This is done by first changing all words to lowercase then removing punctuation marks, stopwords & numbers from it. By changing to lowercase, the text has uniformity without changing its meaning. This step is essential as computer programs are case sensitive. Stopwords are commonly used English words such as 'is', 'the', 'are' which are frequent in the text and they are of no use for our model. We then divided the reviews into "high" score and "low" score category based on the star rating. All reviews with four star and above were categorized as "high" while those scoring below four stars were labelled as "low". In the next stage, sentences were split into individual words (tokenization).

4.3. Topic Modeling: In the last step of our analysis, TF-IDF (term frequency, inverse document frequency) is calculated, which is based on frequency of a word in each document and the entire text. The top terms within the high and low category of review were highlighted as a preliminary analysis.

Document Feature Matrix (DFM) refers to documents in rows and features as columns. Features in our case are the individual words with their respective frequency in our text. DFM is created and is needed as an input to topic modeling. In the last step, Structural Topic Model (STM) algorithm is used to extract topics from our text data.

In topic modeling we try to analyze the topics occurring frequently in our documents. We can think of topics as group of words; while each document (review) can be thought of as composed of one or more topics. It essentially means uncovering latent

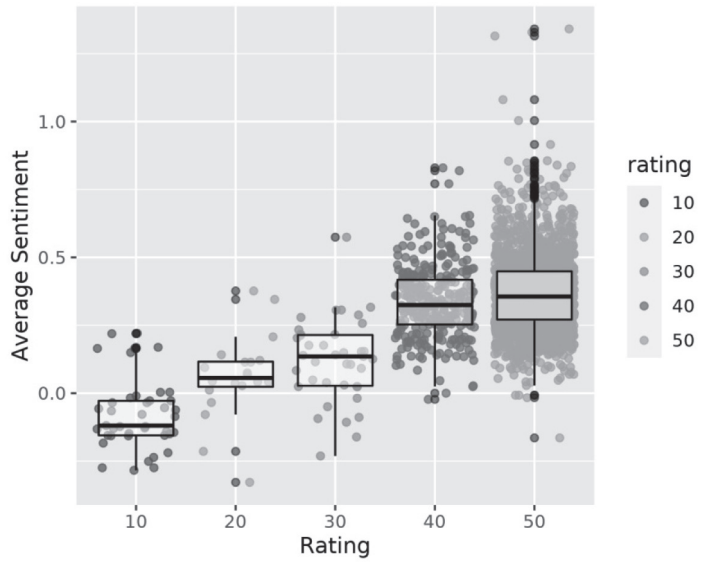
SVKM'S NMIMS Deemed to be UNIVERSITY
NMIMS Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022

topics within a corpus of documents. Through a process of comparing and fine tuning the algorithm, we decided to create three topic categories. These three topic categories summarize our review data as we will see in the analysis section.

5. Findings

The review data for our study has 2369 individual reviews consisting of 23671 sentences (317282words). Quick examination of the star rating given by visitors reveals that more than 95 % of the reviews were rated 4-star or above with only 100 reviews rated 3-star or below. A preliminary sentiment analysis of all reviews gives a basic idea of number of positive and negative customer sentiments. Majority of the reviews (2319) show positive sentiment (value above 0), suggesting that most homestay visitors had a positive experience at the homestay visited. The minimum sentiment value was -0.328 while1.341 was the maximum with a median sentiment of 0.3450. A boxplot of the sentiment scores was plotted against the star rating given by travelers (see Fig 2). The average sentiment increases with star rating meaning that the sentiment algorithm gives a fairly accurate picture of overall sentiment.

Figure 2: Boxplot of Star Rating against Sentiment Score

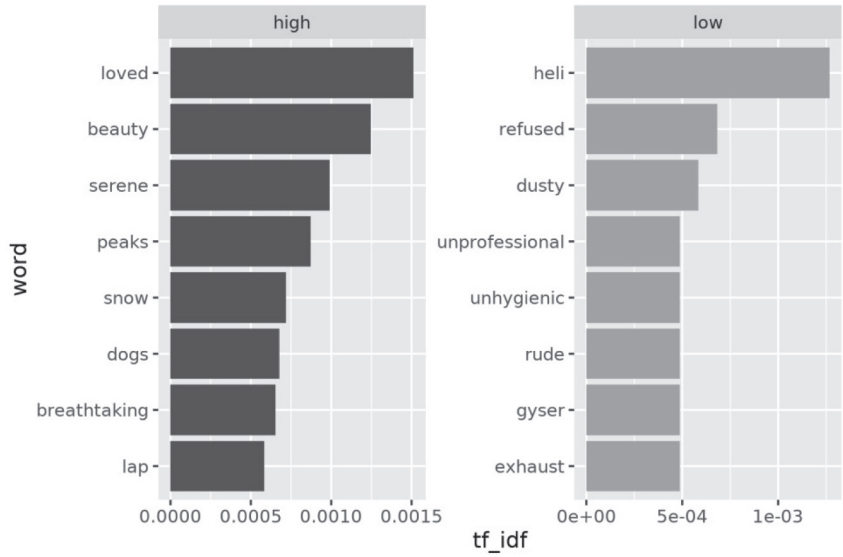


In the next step, reviews were categorized into “high” and “low” category. Where any review receiving 3-star or less has been categorized as “low” rated while 4-star & 5-star reviews labeled as “high”. This was useful for further analysis and comparison of the data. Afterwards, sentences were tokenized (broken down to individual words) and a frequency table created. The frequency were further used to calculate Term Frequency-Inverse Document Frequency (TF-IDF). The TF-IDF (see Silge & Robinson, 2017, p. 31 for a detailed explanation & calculation of TF-IDF) value

SVKM'S NMIMS Deemed to be UNIVERSITY
NMIMS Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022

shows the relative importance of a review document to the review category (high & low in our case).

Figure 3: TF-IDF



The highest scoring words are presented in Fig 3. A look at figure 3 reveals how word usage is different between high and low rated reviews. While the high rated reviews mostly used words like loved, beauty, serene, breathtaking etc. the low rated reviews mostly consisted of refused, dusty, unprofessional, rude, unhygienic etc. This preliminary analysis hints that there is significant difference in visitors' reviews & rating of homestays. That brings us to the next phase of our analysis i.e. topic modeling.

5.1 Topic Modeling

After processing the text data through the STM topic modeling algorithm, we derived three topics presented in table 1. These three topics summarize our review data in terms of topic keywords.

Table 1: Distribution of Keywords in Topics

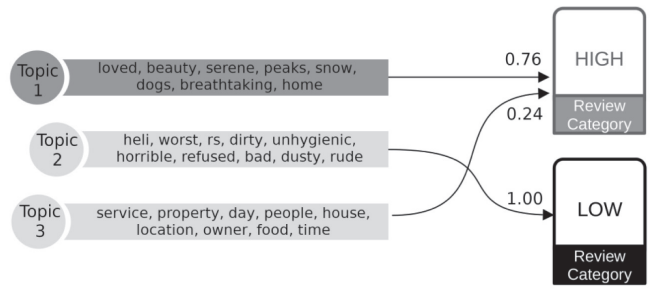
Topic Name	Keywords
Topic 1	loved, beauty, serene, peaks, snow, dogs, breathtaking, home
Topic 2	heli, worst, rs, dirty, horrible, refused, bad , dusty, exhaust, rude, unhygienic
Topic 3	service, property, people, house, location, food, owner, stay, time

We can see that the Topic 1 revolves around aesthetic or visual appeal of the surrounding of homestay as expressed in keywords beauty, loved, peaks, breathtaking. Topic 2 mostly talks about dissatisfaction expressed towards service quality, staff's

SVKM'S NMIMS Deemed to be UNIVERSITY
NMIMS Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022

unprofessional behavior, and lack of cleanliness. The theme of topic 3 is about hospitableness and adequate service levelsat the homestay.

Figure 4: Topic-wise Association with Review Categories.



The gamma values of these topics reveal the composition of different review category and their association with these topics (Fig 4). So, a gamma value of a topic reveals degree of presence of that topic in the specified text. Figure 4 reveals that while the "high" category or reviews were majorly composed of Topic 1 (76%) and Topic 3 (24%), the "low" category reviews were entirely from Topic 2. The analysis clearly maps correct keywords to the respective category. The reviews which received lower rating are exclusively composed of topic 2. Few sample reviews are given in figure 5. We can see that there is prevalence of topic 1 and topic 3 in high rated reviews while low rated ones are composed of topic 2.

Table 2: Sample Quotes from the Review Dataset

Rating	Review Text
High	Bliss..The Place is purely Heaven, the people are so good and so friendly, you just feel like home here, and the view is breathtaking, i wish to thank Mr. Mukesh to make finger-licking good food, the non-veg is highly recommended. Veg is just fine, i love the evening snack chai & pakoras. pics attached.
High	Food, ambiance, stay..everything is just amazing!!! Mohan Binsar retreat is one stop at Almora where one can fully relax and enjoy the beautiful mountain view..Also, Hospitality is an icying on the cake.
High	Place with a view. Went to this place for lunch and evening tea during our stay in Binsar recently, while this place was quite a drive from our hotel but the view here will keep encouraging you to take the drive. They serve you some really good pasta's, Coffee fresh juices. It's worth a visit atleast once when you visit the hills.
Low	Pangot as a hamlet is a bird watcher's paradise and is a quiet getaway from the hustle and polution of Nainital. Yet one has to think twice about choosing woodside retreat as the place to stay.. here is why 1. Woodside as minimal views of the grogeour valley as view is blocked by trees. 2. Woodeside has camping tents right next to it so be careful about the noise of trekkers returning 3. the place is infested with monkeys who attack and make having lunch and inner outside an impossibility 4. Pangot has a a few more places which have MUCH better views and comfort for cheaper rents like The NEST COttage which i sbuilt on organic living and another place called mountin quail. 5. Woodside is really a house with 6 rooms with only ONE large room on top whjich is worth living (if they fix the bathrooms) and the rest is an eyewash. The only saving grace at Pangot for us was the home cooked food by Bir SIngh the house keeper. Yet do try out pangot for its forest walks and a few easy treks nearby.
Low	Disappointed. I booked on line last month because of the good reviews, but our room was bad. Asked if there was another room available. We checked in and out. It was not very clean and what shocked us the most was the bed. It had not a mattress. It seemed like Wood board and a carpet The area didn't look very nice either. On the outside they were burning garbage. I must say that the owner was very nice with all the information he gave us to arrive correctly.

6. Conclusion

This study was primarily exploratory in nature wherein by analyzing homestay reviews in the study area many new themes and directions have emerged. The analysis of visitors' review in Uttarakhand reveals that homestays experiences are generally rated favorably by travelers despite the lack of luxurious facilities. The landscape plays an important role in enhancing the visitors' experience in nature-based locations. However, as evident from topic modeling exercise there are certain areas which homestay owners need to pay attention to reduce negative feedback from visitors. Specifically, these areas pertain to aspect of cleanliness & hygiene, staff behavior, service quality and value for money as evident from emerging themes in topic 2.

7. Contribution

Text data is mostly unstructured and qualitative in nature; this makes analysis of text data resource intensive. In most cases studies focusing on qualitative data tend to include limited samples. Present research was based on customer review data which was used for modeling and extracting sentiments & themes from unstructured text data. It builds upon the Natural Language Processing framework to apply text analysis techniques on large datasets. Once fine-tuned, the analysis can be used by homestay owners/managers & other stakeholders in the community-based tourism domain to leverage the power of text mining to improve visitor satisfaction.

8. Practical Implications

This research highlights the difference between high and low rated homestay experiences and explains the cause for difference in rating, the outcome of this research thus has immediate practical usage in Customer relationship management in hospitality domain.

Gopi Nath Vajpai is currently working at National Institute of Science Education and Research, Bhubaneswar, Odisha, India, Homi Bhabha National Institute, Training School Complex, Anushaktinagar, Mumbai, India and can be reached at gopinath.vajpai@niser.ac.in

Dr. Debashis Pattanaik is currently working at National Institute of Science Education and Research, Bhubaneswar, Odisha, India, Homi Bhabha National Institute, Training School Complex, Anushaktinagar, Mumbai, India.

SVKM'S NMIMS Deemed to be UNIVERSITY

NMIMS Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022

Reference

Barreda, A., & Bilgihan, A. (2013). An analysis of user-generated content for hotel experiences. *Journal of Hospitality and Tourism Technology*.

Boonratana, R. (2010). Community-based tourism in Thailand : The need and justification for an operational definition. *Kasetsart Journal - Social Sciences*, *31*(2), 280–289.

Bose, J., & Biju, M. K. (2020). Accommodation preferences, memorable tourism experience and its outcomes–a comparative study on homestays vs other accommodation among tourists. *Test Engineering & Management, 82*, 13022-13035.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management, 26*(7), 675-693.

Chakraborty, B. (2019, June). Homestay and women empowerment: A case study of women managed tourism product in kasar devi, uttarakhand, india. In *TISC-Tourism International Scientific Conference Vrnjačka Banja* , Vol. 4, No. 1, pp. 202-216.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

García, A., Gaines, S., & Linaza, M. T. (2012). A lexicon based sentiment analysis retrieval system for tourism domain. *Expert Syst Appl Int J*, *39*(10), 9166-9180.

Goodwin, H., & Santilli, R. (2009). Community-based tourism: A success. *ICRT Occasional paper*, *11*(1), 37.

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012, January). Classification of customer reviews based on sentiment analysis. In *ENTER,* pp. 460-470.

Gretzel, U., Yoo, K. H., & Purifoy, M. (2007). Online travel review study: Role and impact of online travel reviews.

Kasper, W., & Vela, M. (2011, October). Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference* , Vol. 231527, pp. 45-52.

Macek, I. C. (2013). *Homestays as Livelihood Strategies in Rural Economies: The case of Johar Valley, Uttarakhand, India* [Masters dissertation, University of Washington].

Misner, I. R. (1994). *The world's best-known marketing secret: building your business with word-of-mouth marketing*. Bard.

Oranratmanee, R. (2011). Re-utilizing space : accommodating tourists in homestay houses in northern thailand. *Jars*, *8*(1), 35–54.

Park, E., Kang, J., Choi, D., & Han, J. (2020). Understanding customers' hotel revisiting

behaviour: a sentiment analysis of online feedback reviews. *Current Issues in Tourism, 23*(5), 605-611.

Price, Martin F. (1992). Patterns of the development of tourism in mountain environments. *GeoJournal*, *Vol. 27*(1), Mountain Environments, 87-96.

Ray, B., Garain, A., & Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, 98, 106935.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

Thapa, B., Manavi, A. D. , & Malini, D. H. (2018). Unraveling tourists' preferred homestay attributes from online reviews: A sentiment analysis approach. *International Journal of Pure and Applied Mathematics*, Vol. *119*(15), 1567-1585.

SVKM'S NMIMS
Deemed to be UNIVERSITY

NMIMS
Management Review
ISSN: 0971-1023
Volume XXX
Issue-2 | April 2022