### Introduction

### Overview of tourism and hospitality in India

India has always been one of most sought-after travel locations owing to its rich heritage, variety in ecology, and places of natural beauty. The tourism industry in India is one of the most critical pillars of the service sector and a large employer as well. Tourism makes up one-tenth of India’s GDP. Domestic tourism, especially off late, has been aiding the sector’s growth in the country courtesy of its rising middle class.

The country is also well advanced in terms of the use of technology including easy to use digital means of planning a journey. Owing to the same, foreign tourist arrivals have soared in the recent past with a massive increase in the issue of e-Tourist Visas (Ministry of Tourism, 2019).

### Changes in customers’ perspective and expectations especially domestic travellers

The tourism and hospitality sector has been hit hard by the pandemic in India (Kumar 2020). Due to this, there is bound to be a change in the rationale with which travellers think about making trips. Our study aims to investigate these changes in travellers’ perceptions about general travel and hotel stays by inspecting user reviews in detail and thereby bring out some of the implications thus obtained for hotel managers and researchers.

### User Generated Content

With the quick spread of UGC (User Generated Content) research in tourism, TripAdvisor has emerged as one of the best platforms in terms of acting as a credible data source (Ayeh et al., 2013). User reviews left on platforms like TripAdvisor about hotels are a very rich source of information about travellers’ firsthand experiences (Guo et al., 2017). With this, users get a chance to convey their thoughts to others on the large web platform which makes the reviews a strong form of electronic word-of-mouth (Filieri et al., 2015).

Reviews have been found to determine the performance of service providers (Ganzaroli et al., 2017; Nieto et al., 2014) and affect their perceived value and reputation as well (Rodríguez-Díaz et al., 2016; Baka 2016). These effects translate to loyalty (Gavilan et al., 2018), intention to visit (Zhang et al., 2010) and intention to recommend (Berezina et al., 2016). Collie (2014) states that 65% of leisure travellers will search online before deciding on a travel destination, and 69% of their plans are determined by online travel reviews.

Collecting and analysing these reviews ensure good data availability, quick data collection, non-intrusiveness with human subjects and useful end implications for hotel managers (Lu et.al., 2015).

Taecharungroj et al. (2019) elaborate on the immense significance of online reviews in the tourism research context. They also discuss how TripAdvisor contains rich information and employs strict supervision to police fake/potentially manipulated content as a result of which it has become one of the most credible review platforms over the years (Filieri et al., 2015; Schuckert et al., 2015).

It is natural for potential travellers to read review contents carefully to be acquainted with the hotels. This means that the ease with which they can read and understand the contents of the review plays an important role. One widely accepted measure for the above is “readability” which is likely to influence a review’s perceived value. Several studies in the past have studied the effect of readability on the perceived value of reviews (Liu et al., 2015; Kusumasondjaja et al., 2012; Ghose et al., 2011).

TripAdvisor also provides various data other than reviews including but not limited to the precautions the hotel employs, room features, room types, amenities, places of attractions in close proximity to the hotel, prices, and location.

With data mining becoming more prevalent in the UGC research domain, it has become common practice to apply techniques such as machine learning and statistical analysis come into picture to analyse and interpret data (Xiang et al., 2017). For reviews, the use of text mining methods is most recommended. In this study, we utilise the Flesch Reading Ease Scale (FRES) (Kincaid et al., 1981), tf-idf scores, word2vec (Mikolov et al., 2013), and sentiment analysis using coreNLP (Manning et al., 2013).

The effect of the pandemic on hospitality and what it means for hotel managers’ strategies is an under-researched topic. This study aims to present comprehensive comparisons of various parameters corresponding to pre-pandemic and post-pandemic times in order to understand what exactly has changed in the hospitality space. Some of the most interesting observations are that reviews in general carry a lot more positive sentiment post the pandemic for all traveller types, a noticeable decline is observed in the share of reviews left by business travellers and potentially their occupancy share in hotels, and a sizeable increase in the ratings and positive sentiment of reviews for hotels that have listed their precautions on TripAdvisor as opposed to the ones that have not.

Leveraging the results of the above research, we give crucial managerial implications for hotel managers, hotel marketing professionals, and other stakeholders and theoretical implications for tourism and hospitality researchers.

The rest of the study is organised as such. Section 2 contains a review of the existing literature on text analytics techniques used in UGC analysis and COVID studies conducted thus far in the tourism space. Section 3 contains the comprehensive methodology used in our research. Section 4 is a presentation of the findings of the conducted study. Following which we discuss the practical and theoretical implications based on the observed results in Section 5. Lastly, we present the conclusions, limitations, and scope for future research in Section 6.

# Literature Review

The volume of online data relative to the tourism and hospitality space is increasing exponentially. In fact, the data are generated as such a velocity that they can be used to capture real time events and hence monitor service quality and recovery spontaneously. Availability of data and advances in technology render the traditional pen and paper surveys useless (Alaei et al., 2017). This calls for automation in the analysis of these data. For this a plethora of state-of-the-art techniques are now available that can be explored and executed with ease using computers. We explain all the techniques used in this paper in this section.

Fang et al. (2016) analyse the effect of readability and reviewer characteristics on the perceived value of online tourism reviews. The perceived value is gauged by the “helpfulness” parameter which is the number of helpful votes a review has received on TripAdvisor. TripAdvisor provides a feature called “Was this review helpful?” to which a “yes” or “no” button can be clicked by users to quickly identify helpful reviews. They find that the readability of a review text is correlated with perceived helpfulness of the reviews i.e., reviews written with an easy-to-understand writing style will receive more helpfulness votes and that reviews carrying extreme sentiment are more likely to considered as valuable.

Tsai et al. (2020) predict the helpfulness of reviews with the help of parameters gauging different characteristics of a review. Some parameters include readability scores (6 different metrics), rating, and number of characters, syllables, and words. For reviewing review sentiments, they employ Stanford’s CoreNLP toolkit due to its favourable generalizability across research domains and high acceptance in the research community.

The term frequency-inverse document frequency (tf-idf) score (Razzaghnoori et al., 2017; Chen et al., 2021; Martinez-Torres et al., 2019) is a widely used metric in text analytics. This is a word-wise statistical score that represents the importance of words based on its frequency for any one document from an entire corpus. The tf-idf enhances the analysis process as it brings out words that appear in rare documents (reviews) but might be of great substance to the analysis as a whole. For example, some words like “spa” might have a low overall term frequency as most normal hotels in India don’t have spas, but reviews with mentions of the word “spa” will add value to resorts that have operating spas.

The Word2Vec algorithm is an unsupervised word-embedding language model. It is a widely used text analysis method used for similarity analysis of words in a corpus. It was given based on Google’s deep learning framework (Mikolov et al., 2013). The core principle of Word2Vec is that semantically and syntactically similar words occur in the same context with high probability. It learns the vector representations of words in a high-dimensional vector space and calculates cosine distance between word vectors as a measure of similarity. Since it is unsupervised, it can take pieces of natural text as inputs and returns a vocabulary of words, each having a corresponding vector of continuous numerical values that encode the information about the words that they represent. The algorithm uses one of the two variants (Mikolov et al., 2013a, 2013b, 2013c): continuous bag-of-words (CBOW) model and continuous skipgram model, to learn the word vector representations. The CBOW variant predicts the word in focus based on the context and the skipgram predicts surrounding (context) words with the help of the word in focus.

Sentiment analysis helps determine the overall contextual polarity of a text document. This polarity can be either positive, negative, or neutral. In the UGC (tourism) context, this text document is a user review on TripAdvisor. With advances in technology and readily available UGC, researchers have used sentiment analysis as a tool to study travellers’ perceptions and level of satisfaction. Analysing these large volumes of subjective electronic word of mouth is of irrefutable value to tourism and hospitality organisations (Alaei et al., 2017). Coming to recent times, Alamoodi et al. (2020) say that sentiment analysis can help reduce response time and risks especially during pandemic-like situations. Borg et al. (2020) explain that sentiment analysis can be used prepare actions for customers and provide feedback according to the sentiment via customer support. This holds well in the hospitality context too, as hotel managers are tasked with something very similar to the above, while replying to TripAdvisor reviews of past customers.

To tackle the problem of overly frequent and overly infrequent words showing up in the analysis without adding much value to it, we place a lower and upper bound on the frequencies of words that we consider for analysis (Kirilenko et al., 2021).

COVID centred research in tourism and hospitality is just starting to pick up pace and will continue to be prevalent till a cure is found. Most of the existing literature that uses data from sources like TripAdvisor only focuses on reviews and the topics thus described (Escandon-Barbosa et al., 2021; Yu et al., 2021; Uğur et al., 2020). There has been no study till date drawing out a clear cut comparison between pre and post pandemic times. Furthermore, no literature exists on analysing the precautions parameter (which is readily available on TripAdvisor) and its interactions with other crucial parameters like the travel type and review characteristics. There continues to be a comparative dearth of literature about the current situation for tourism and hospitality in India. The current study strives to bridge all the mentioned research gaps.

1. Methodology

This study is built on a 4-stage framework: data collection, a rundown of missing data, data cleaning and preprocessing, and data analysis. The first stage involves web scraping of TripAdvisor India’s website/platform for all necessary data for the analysis. The second stage explores the case of missing data and is often under-researched. In the third stage, all data to be analysed is preprocessed and cleaned using a standard set of techniques. The last stage comprises of exploratory data analysis and text mining.

# Data collection

We collect all our data from TripAdvisor India’s webpage (<https://www.tripadvisor.in/>). We employ a top-down approach to data collection. First, we look at a TripAdvisor section in which the most famous tourist locations are listed. Each location’s hyperlink in this section further holds a comprehensive list of hotels in the respective location. From here, each hotel’s hyperlink holds all information about the hotel from which we extract the hotel’s name, location, property amenities, room features, room types, listed covid precautions, reviews (only English reviews from January 2019 to July 2021), date of stay (if mentioned by the reviewer), trip type (if mentioned by reviewer), rating (1-5 bubbles), and price range (if listed on TripAdvisor). We have an initial filter of 150 reviews as we were observing that hotels with fewer reviews were harder to extract data from and were also missing a hefty amount of data, thus contributing very little to the current study. After removing duplicate reviews, we finally ended up with 430,876 reviews from 2405 hotels across 387 locations.

We use selenium, a software package that can be implemented in python, for all data collection activities. Selenium uses HTML tags of elements on webpages to identify and extract content from tags across the webpage. Selenium throws an error when it cannot find desired data on the webpage. We handled these with try-except blocks.

After collecting all data, we partition it by date to yield two datasets: the pre-covid dataset (dataset 1) and the post-covid dataset (dataset 2). Dataset 1 is based on the timeframe: 01/01/2019 – 28/02/2020. Dataset 2 is based on the timeframe: 01/03/2020 – 31/07/2021.

# Missing data

Some data is bound to be missing in a pan-Indian study of relatively large scale involving low-end hotels across a two-year timeline. Some categories of data can be missing due to incomplete information filled by travellers themselves. For example, the trip type is an optional field to enter for a user leaving behind a review.

Missing data can often affect the quality of analysis, and a conscious effort by hotel managers to list all data about the hotel and encourage their guests to leave behind reviews with as much information as possible can go a long way in improving the hotel’s online profile and contribute to the theoretical aspect of further enriching data-oriented studies like ours.

# Analysis

* + 1. Ratings

Users on TripAdvisor can leave behind rating scores ranging from 1 (worst rating possible) to 5 (best rating possible). Ratings convey information about the overall experience at hotels in very little time and space. Hence their analysis can often yield crisp insights about the quality of hotels and travel experiences.

* + 1. Trip type

Another vital variable in our study is the trip type. TripAdvisor lets users associate their stay with either one of the five below listed types of trips: business, solo, couple, family, and friends. We use this to segment travellers in parts of our analysis and to explore the interactions of this parameter with several others in the study, such as the precautions taken by the hotels.

* + 1. Precautions

A new and unique parameter to this study is the precautions listed by hotels. To explore potential relations between other parameters and precautions, we first quantify the precautions into a simple count of the number of precautions taken by the hotel and then discretise this count variable into subcategories.

* + 1. Review preprocessing

We first clean all review text using standard text preprocessing techniques to perform text analysis. We expanded contractions in the text (didn’t -> did not), apply lemmatisation along with POS tagging, discarded punctuations, stop words, extra spaces, converted all non-ASCII characters to ASCII characters using python’s unicode package, and finally, removed all words that do not exist in the English Dictionary. All above steps are executed in python using the Natural Language Toolkit (NLTK) (<https://www.nltk.org/>).

* + 1. Review length and readability

Some of the characteristics of the reviews we scrutinised are the length of the review, i.e., the number of words in a review and the readability measure of the review. Literature shows that these metrics have been a great indicator of the review’s perceived value and helpfulness, thus directly affecting decision-making (Shin et al., 2021; Fang et al., 2016; Liu et al., 2015).

We use the renowned Flesch Reading Ease Scale (FRES) developed by Kincaid et al. (1948) to measure the complexity of reviews. We calculate the metric for each review using the standard formula and implement it in python using the textstat package (<https://pypi.org/project/textstat/>). We further study the variation of this metric with the traveller type types of hotels (segmented by level of precautions taken) and compare the general scores of reviews in datasets 1 and 2.

Text

Description automatically generated

* + 1. Sentiment Analysis

To maintain the main goal of achieving a holistic understanding of the exact changes in customer mindset reflected by online reviews, we delve into the sentiment of the review texts. Studying this not only sums up the overall mood of the review, but also comparisons of the sentiment parameter in datasets 1 and 2, studying sentiment for each class of traveller during different timeframes, and exploring its relations with the precautions parameter gives us a lot of insights about how travel and stay sentiments for customers have changed over time and what the possible role of the pandemic has been in bringing about these changes. These insightful results further shed light on invaluable implications for practice and theory.

This analysis is implemented with Stanford’s CoreNLP (Manning et al., 2014; Valdivia et al., 2017). CoreNLP returns one of the five sentiment buckets for each review that it encounters: Very negative, Negative, Neutral, Positive, Very Positive.

* + 1. Word analysis

A vital step in text analytics is deriving numerical features from the cleaned text. This helps introduce a mathematical way to analyse text and empowers the usage of mathematical, statistical, and machine learning models. One of the most effective and widespread ways of feature extraction for textual data is the Term Frequency-Inverse Document Frequency (TF-IDF) method. To shortlist words to consider for the TF-IDF, we place a lower bound of 1.5% and an upper bound of 70% on the total occurrence of a word to filter out the most frequent and infrequent words, a standard practice for computer-assisted text analysis (Kirilenko et al., 2021).

To examine further intricacies in our review text, we decided to apply a semantic direction of thought. We use word2vec implemented via gensim in python (<https://radimrehurek.com/gensim/models/word2vec.html>) to measure semantic similarity between words.

In our context, word2vec is instrumental in distinguishing between the context in which users use certain words pre and post the pandemic. We aim to bring out exactly what has “changed” for travellers due to the pandemic by scrutinising these changes.

First, we examine the common words in datasets 1 and 2 with high TF-IDF scores (among the top 10). After this baseline exercise, we delve into a manually curated shortlist of more relevant words to the present scenario.

1. Results

The data collection process yields 2 datasets: dataset 1 with 301637 entries and dataset 2 with 129239 entries. Upon examining missing data, we find that 36.48% of reviews do not have mentions of the trip type in dataset 1 and the same figure rose to 39.89% in dataset 2. In dataset 2, we find that 1146 out of 2182 hotels which is 52.52%, have not listed precautions on TripAdvisor.

* 1. Ratings and sentiment

Highlighted in Table 1 are the proportion of total rating corresponding to each rating in datasets 1 and 2.

**Table 1:** Proportion of each rating for datasets 1 and 2

|  |  |  |
| --- | --- | --- |
|  | Dataset 1 | Dataset 2 |
| 5 | 0.816491 | 0.86574 |
| 4 | 0.118066 | 0.089215 |
| 3 | 0.031044 | 0.017224 |
| 2 | 0.012992 | 0.008991 |
| 1 | 0.021407 | 0.018787 |

There is a 2% increase in 4 and 5 star reviews post the pandemic.

* 1. Ratings and precautions

To explore the spread of ratings with precautions listed, we adopt a binary categorisation approach for the precautions variable. We coin two buckets: “Listed” and “Not Listed”, signifying hotels that have listed one or more precautions on TripAdvisor and the ones that have not listed any precautions on TripAdvisor.

Chart, histogram

Description automatically generated

**Figure 1:** Comparison of percentage volume of ratings of the reviews in hotels that have and have not listed precautions

From Not listed to Listed, we observe a 2.83% decrease in 1 and 2 star ratings, 1.3% decrease in 3 star ratings, and a 4.4% increase in 4 and 5 star ratings.

Given the above, we investigate within the “Listed” hotels category by binning the count variable into four bins by employing quantile splitting. The bins and their frequencies are:

* [1, 9] - 29538
* (9, 12] - 36236
* (12, 13] - 18480
* (13, 14] - 25670

Chart

Description automatically generated with low confidence

**Figure 2:** Comparison of percentage volume of ratings of the reviews in hotels categorised by number of precautions listed

Within hotels taking precautions, we observe almost no change in the rating distribution.

* 1. Sentiment

Chart, bar chart

Description automatically generated

**Figure 3:** Percentage frequency of reviews vs sentiment of reviews for Dataset 1 (left) and Dataset 2 (right)

Figure 3 suggests that the net percent of review with positive sentiment has gone up in dataset 2 by 9% with a 3% increase in the positive and 6% increase in the very positive bracket.

A comparison of the rating of a review with the sentiment associated with it, helps understand the extent to which ratings are subjective and how much they can potentially explain the sentiment of the review. It further sheds light on the subclass of reviews that have become more positive in sentiment as proved previously.

Chart

Description automatically generated

Chart

Description automatically generated

**Figure 4:** Sentiment distribution by rating category for dataset 1 (top) and dataset 2 (bottom)

Figure 4 suggests that reviews in general have become far more positive post the pandemic. The above especially holds for the reviews that have a corresponding rating of greater than 2.

* 1. Trip type

**Table 2:** Proportion of each trip type for datasets 1 and 2

|  |  |  |
| --- | --- | --- |
|  | Dataset 1 | Dataset 2 |
| Business | 0.316563 | 0.224541 |
| Family | 0.352517 | 0.390480 |
| Friends | 0.130950 | 0.142395 |
| Solo | 0.047435 | 0.061950 |
| Couple | 0.152535 | 0.180634 |

On comparing the proportion of the volume of reviews with the trip type, we find that dataset 2 consists of roughly 4% more family, 3% more couple, 2% more solo, 1% more friends and 9% lesser business reviews.

Chart

Description automatically generated

A picture containing chart

Description automatically generated

**Figure 5:** Sentiment distribution by trip type for dataset 1 (top) and dataset 2 (bottom)

Upon weighing the trip type against sentiment for datasets 1 and 2, we observe a uniform uptick in positive sentiment amongst all trip types.

To investigate if the traveller distribution changes with varying hotel precautions, we draw up a comparison between the precautions count parameter and the trip type. Similar to the rating analysis, we consider a binary approach first and then dive deeper into the subcategories within hotels taking precautions. We also aggregate the trip type parameter and present a binary categorisation. The two groups we coin are “Group” and “Individual” with the former subsuming the travel types friends, family and couple and the latter comprising of business and solo travellers. This approach helps succinctly bring out the interaction of the parameters in focus.

Chart, treemap chart

Description automatically generated

**Figure 6:** Percentage of group and individual traveller reviews in hotels that have and have not listed precautions

Data suggests that individual travellers make up 5% more of the reviews and potentially more occupancy as well in hotels that list precautions. To scrutinize further, we use the binned precautions count parameter.

Graphical user interface, funnel chart

Description automatically generated

**Figure 7:** Percentage of group and individual traveller reviews in hotels categorised by number of precautions listed

A more detailed comparison finds that the volume of reviews comprises of noticeably more individuals travellers as the amount of precautions taken go up which supports the previous finding.

To test if the sentiment of the review changes with the precautions listed by hotels on TripAdvisor, we perform a percentage comparison of the reviews in different buckets of the precautions variable.

* 1. Sentiment vs precautions

Chart, bar chart

Description automatically generated

**Figure 7:** Sentiment distribution in hotels that have and have not listed precautions

A binary comparison indicates a 7% increase in the positive and very positive reviews for hotels that have listed their precautions. A deeper comparison with the binned precautions count parameter brings out minimal changes in the distribution of the sentiment across the precaution buckets.

Chart

Description automatically generated with low confidence

**Figure 8:** Sentiment distribution in hotels categorised by number of precautions listed

A comparison of the ratings and the review length variable can speak volumes about how a vital factor contributing to review helpfulness relates with the ratings. It helps answer questions about the amount of content people leave behind when they are seemingly satisfied and or discontented.

* 1. Review length

Chart, box and whisker chart

Description automatically generated

**Figure 9:** Box plot of review lengths of reviews belonging to all rating categories for datasets 1 (left) and 2 (right)

Figure 9 suggests that the length of the review has an inverse relation with the rating and also attains a wider distribution with its increasing magnitude.

**Table 3:** Descriptive statistics for review length by rating: Dataset 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean | SD | Min | 25% | 50% | 75% | Max |
| 1 | 6457 | 64.379123 | 60.051238 | 6 | 29 | 49 | 74 | 1356 |
| 2 | 3919 | 64.571064 | 54.022664 | 13 | 29 | 52 | 75 | 867 |
| 3 | 9364 | 56.922362 | 48.649762 | 10 | 26 | 43 | 70 | 739 |
| 4 | 35613 | 42.967287 | 35.234065 | 13 | 23 | 30 | 52 | 808 |
| 5 | 246284 | 36.432391 | 28.161065 | 5 | 22 | 27 | 40 | 1396 |

**Table 4:** Descriptive statistics for review length by rating: Dataset 2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean | SD | Min | 25% | 50% | 75% | Max |
| 1 | 2428 | 62.761120 | 63.501482 | 10 | 27 | 47 | 71 | 938 |
| 2 | 1162 | 63.104991 | 71.265352 | 7 | 27 | 48 | 72 | 1211 |
| 3 | 2226 | 56.196765 | 49.740747 | 5 | 24 | 42 | 67 | 568 |
| 4 | 11530 | 37.852559 | 33.763657 | 5 | 20 | 26 | 44 | 595 |
| 5 | 111893 | 31.888393 | 25.319224 | 3 | 19 | 23 | 36 | 1122 |

Upon closer inspection, we find that the differences in review length corresponding to the same rating in datasets 1 and 2 to be very minute with there being close to a noticeable drop in the 4 and 5 star rating category.

Chart, box and whisker chart

Description automatically generated

**Figure 10:** Box plot of review lengths of reviews by sentiment categories for datasets 1 (left) and 2 (right)

As expected, the weighing of sentiment against the review length pointed toward positive review being inherently shorter and the negative ones being longer. As seen in the figure, very similar lengths for all sentiment buckets exist for datasets 1 and 2 suggesting that the interaction between the two parameters has not changed due to the pandemic.

Chart, box and whisker chart

Description automatically generated

**Figure 11:** Box plot of review lengths of reviews by trip type for datasets 1 (left) and 2 (right)

In accordance with earlier results, the review length indicates a down move homogenous across trip types for dataset 2. This can potentially be attributed to the boom in positive sentiment and higher average ratings post the pandemic.

Chart, box and whisker chart

Description automatically generated

**Figure 12:** Box plot of review lengths of reviews by trip type for hotels that have (right) and have not listed precautions (left)

**Table 5:** Descriptive statistics for review length by trip type for hotels that have and have not listed precautions

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Not Listed** | Count | Mean | SD | Min | 25% | 50% | 75% | Max |
| Business | 1818 | 28.319032 | 20.641487 | 9 | 18 | 22 | 29 | 273 |
| Family | 4798 | 36.740517 | 33.720608 | 7 | 20 | 26 | 41 | 938 |
| Friends | 1935 | 31.831008 | 23.601984 | 5 | 19 | 23 | 35 | 267 |
| Solo | 659 | 32.790592 | 26.030176 | 10 | 19 | 24 | 37 | 253 |
| Couple | 2148 | 36.298883 | 32.271122 | 10 | 20 | 25 | 40 | 375 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Listed** | Count | Mean | SD | Min | 25% | 50% | 75% | Max |
| Business | 15616 | 28.160092 | 21.395478 | 5 | 18 | 22 | 29 | 434 |
| Family | 25520 | 33.099060 | 28.956024 | 3 | 19 | 23 | 36 | 601 |
| Friends | 9121 | 28.831159 | 25.275560 | 4 | 18 | 22 | 30 | 752 |
| Solo | 4151 | 30.554083 | 24.858596 | 9 | 19 | 23 | 32 | 367 |
| Couple | 11877 | 33.041846 | 29.077028 | 4 | 19 | 23 | 36 | 511 |

Not too noticeable a change is observed in review length among trip types between hotels that have and have not listed precautions. A marginal drop is the longer reviews is the only thing to note.

**Table 6:** Descriptive statistics for review length by trip type for hotels categorised by number of precautions listed

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean | SD | Min | 25% | 50% | 75% | Max |
| [1, 9] | 29538 | 34.775746 | 30.680025 | 5 | 19 | 24 | 39 | 865 |
| (9, 12] | 36236 | 33.306712 | 29.272274 | 3 | 19 | 24 | 37 | 1122 |
| (12, 13] | 18480 | 30.866613 | 26.560598 | 6 | 18 | 23 | 33 | 1211 |
| (13, 14] | 25670 | 32.647293 | 27.603885 | 4 | 19 | 23 | 36 | 601 |

We observe negligible changes in review length across hotels that have listed precautions.

* 1. Readability

Inspecting the readability of the reviews gives us an idea of how potentially helpful the reviews are prospective travellers. Readability has proven to be one of the key drivers of review helpfulness (Yang et al., 2017). A widely accepted and highly accurate metric to measure readability is the Flesch Reading Ease Scale (Kincaid et al., 1981) that returns a score between 0 (extremely confusing) and 100 (extremely easy to read and interpret). While the maximum score is 121.22, there is no limit on how low the score can be. A negative score is valid (<https://pypi.org/project/textstat/>).

The average Flesch Reading ease score of reviews in datasets 1 and 2 linger around the mid-sixties showing negligible contrast. Furthermore, we find similar results on comparing readability scores by trip type (in the union of datasets 1 and 2) and the level of precautions (in dataset 2).

**Table 7:** Descriptive statistics for number of words, syllables, and sentences in reviews in Datasets 1 and 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean (Dataset 1) | SD (Dataset 1) | Mean (Dataset 2) | SD (Dataset 2) |
| Words | 71.638115 | 64.024589 | 70.935198 | 64.157983 |
| Syllables | 101.103800 | 87.112108 | 101.282121 | 87.947081 |
| Sentences | 4.736491 | 3.613461 | 4.662733 | 3.589529 |

The mean of the readability scores is comparable though a touch lesser to the ones Tsai et al. (2020) furnish in their study involving reviews from the United States. The standard deviation on the other hand is almost double in the case of Indian reviews. Along the same lines, a drastic decrease is observed in the average and standard deviation of the number of words, syllables, and sentences in Datasets 1 and 2 as compared to the results of Tsai et al. (2020).

An intriguing finding emerges when the length and the readability of the reviews are compared. The correlation of the length and the readability scores of the reviews is -0.045, p<0.01 in dataset 1 and -0.022, p<0.01, which is similar to the findings of Shin et al. (2021) who argue that more readable texts paradoxically contain lesser information.

* 1. Results of word analysis

We compare the top 20 words (as per TF-IDF score) and their 20 nearest vector neighbours (returned by Word2Vec) to try and understand the contrast if any in what context these words are used in datasets 1 and 2.

**Table 8:** Top words, their vector neighbours, and observations regarding change of usage in datasets 1 and 2

Table named “word2vec top 10” in Marketing paper excel sheet

After this initial exercise, we move onto a custom word list with words handpicked by us; the analysis of which we felt would shed light on the exact differences in peoples’ thoughts about certain subject words before and during the pandemic.

**Table 8:** Custom words, their vector neighbours, and observations regarding change of usage in datasets 1 and 2

Table named “word2vec top 10” in Marketing paper excel sheet

1. Discussions and implications
   1. Missing Data

In general, we find that Indian hotels can do a lot better in listing their precautions as we observe that 52% of the hotels in India have not listed their precautions on TripAdvisor. 3% more of the in dataset 2 do not have mentions of the type of trip. Urging customers to leave behind more comprehensive reviews can go a long way in helping data analytics of the reviews to enhance business decisions (Black et al., 2009).

* 1. Ratings and sentiment

Review ratings have become more positive post the pandemic with a 2% increase in 4 and 5 star reviews in dataset 2 as compared to dataset 1. This shows that the average traveller tends to leave behind more positive reviews post the pandemic. We feel this can potentially be attributed to the reliefs of people all over the country after being able to travel post stringent lockdowns.

Within dataset 2, there is an increase in 4 and 5 star ratings (combined) and a proportional decrease in 1, 2 and 3 star ratings for hotels that have listed precautions as compared to the ones that have not. Within the subset of hotels that have listed precautions in dataset 2, we find that there is little to no change in the rating distribution which indicates similar level of satisfaction in all hotels that have listed precautions.

A sizeable increase in the frequency of reviews in the “Positive” and “Very positive” sentiment brackets in dataset 2 further shows that the average traveller leaves behind much more positive reviews post the beginning of the pandemic. To drive the above point home, a ratings vs sentiment comparison also shows that reviews having 3, 4, and 5 star ratings have a much more positive sentiment to them in dataset 2.

* 1. Trip type

A decrease in reviews belonging to the “Business” trip type category indicates either decrease in travel for business purposes or business travellers’ increased disinclination to write reviews following their stay. Either which way, hotels must try and market themselves in a way that is more appealing to business travellers in order to achieve pre-pandemic numbers.

Consistent with our previous finding, sentiment is on the up for all trip types in dataset 2. This reveals that all traveller types have been leaving behind much more positive sentiment oriented reviews since the pandemic’s dawn. Hence the above effect of the pandemic on travellers is not limited to the nature of the trip made by the traveller.

A comparison of the trip type with the precautions parameter suggests that individual travellers (business+solo) tend to stay in hotels that have listed their precautions on TripAdvisor. A joint 7% increase in the positive and very positive reviews for hotels that have listed precautions in dataset 2 shows that travellers exude more positivity in their reviews which leads to more beneficial e-WOM for the hotels. Hence, hotel marketers must note to display each of their COVID safety precautions in all mediums where the hotel is marketed.

* 1. Text analysis

An inverse correlation between the length and the readability suggests that lengthier text contain less readable information. Hence hotel managers must keep the length of the reviews in mind while choosing to highlight certain helpful reviews while marketing the hotel.

Most of the top words (judged by TF-IDF and Word2Vec) have no change in the context that they are used in pre and post the beginning of the pandemic. This shows the absence of a very drastic change in the general context in which people use and describe subjects involving high frequency words in their reviews.

One word where we could strike a distinction is “hotel”; where mentions of the hotel’s proximity to airports and railway station takes more precedence in dataset 2 than in dataset 1. We think this can be attributed to the fact that people are naturally more concerned with anything travel related during the pandemic and hence tend to mention ease of travel more regularly in their reviews. More research on this front is required to be able to draw concrete conclusions.

Coming to the custom word list made by us, we find quite a lot of interesting takeaways that help capture exactly what has changed in travellers’ minds.

The word “hygiene” and “risk” are associated more with COVID and COVID safety norms in dataset 2.

The word “safety” is associated with COVID safety norms in dataset 2 whereas it was associated theft, harassment and general security in dataset 1.

The context in which the word “travel” is used in dataset 2 shows that mentions of travel apprehensions have increased multi-fold post the pandemic.

The use of the word “government” in dataset 1 as with respect to the local police, politics and ministries. However the focus shifts to the COVID norms levied by local governments in dataset 2.

We observe an increased association of buses and bus stations with the word "Airport" in recent times as opposed to trains and railway stations before the pandemic.

The words “disease” and “fever” are used along with very strong negatively polarised words in dataset 2 as compared to dataset 1.

The word “symptom” was associated with COVID in dataset 1, even before the 1st wave began in India. This shows the vigilance of people noticing the effects and potential of the pandemic.

The contrast in the context in which the word “doctor” is used in the datasets indicates a shift towards therapy and other mental ailments. This comes as no surprise as mental health has become one of the most prominent issues to come to centre stage due of the effects of the pandemic.

It is also worth mentioning that the words “emergency”, “hospital”, “spread” and “news” saw little to no change in the context they were used in.

* 1. Theoretical implications

This study provides researchers with a solid framework to compare and contrast the travel perspectives of customers before and after the dawn of the pandemic. We put forth that any study in the near future involving studying traveller’s preferences using UGC must consider and analyse the irrefutable role of the pandemic.

Though the phenomenon of COVID is not under-researched in the tourism and hospitality space, our study provides a unique methodology to gauge its effects on widely used theoretical parameters such as trip type, ratings, readability, precautions, review length, and sentiment. Moreover, little research in tourism has considered the precautions taken by the hotels and we develop an approach to analyse not only the precautions but also its effects on other variables.

We provide an in-depth framework for automated text analysis using a word list made using TF-IDF scores to study the general differences in opinion about certain important topics, and a custom word list to single out and study a particular phenomenon (here COVID) which may not always show up while considering only words that have a high TF-IDF score. The same approach can be employed in any research that aims to provide either a general overview or study a particular phenomenon using text data.

* + 1. Review length

All analysis of the review length broadly suggests that the length of the average review has become shorter post the pandemic. This is a standard result in tourism as writers are naturally more inclined to describe their negative experiences in much more detail than the positive ones (Herr et al., 1991; Shin et al., 2021).

* + 1. Readability

The mean readability for datasets 1 and 2 is almost the same suggesting that there has been very little to no influence of the pandemic in the readability of customer reviews. The mean readability obtained in this study when compared against the same in a study involving reviews from the U.S.A (Tsai et al., 2020) comes out to be a little lesser. This may be a hint that Indians in general write lesser readable reviews (according to the metric adopted to measure the same) than residents of the U.S.A. The standard deviation of the readability score being nearly double as that of the ones observed in Tsai et al. (2020)’s study shows that the readability in the Indian reviews is more erratic and spread out. More research on this front going forward can potentially bring out more meaningful insights.

* + 1. Precautions

We explore the effects of the types of precautions taken by hotels on various review parameters to shed light on the overall nature of the effect that precautions have on different types of customers.

1. Conclusions, limitations and further research

Our study analyses UGC from a different perspective to pinpoint differences in customers’ experiences in and overall mood for travel pre and post the beginning of the COVID pandemic. Through sentiment analysis of user reviews, we prove that the average traveller leaves behind much more positive reviews that the hotel managers can look to capitalise on by highlighting such reviews on the hotel’s webpage. We point out a decrease in the share of reviews and potentially occupancy of business travellers that hotel managers must be wary of and henceforth market their hotels in a way to attract more travellers from the above demographic. Most importantly, by comparing the precautions parameter with the review rating and sentiment, we shed light on the advantages of listing a hotel’s precautions on its webpage.

From a theoretical standpoint, our framework serves as a template for all studies that want to investigate various and more detailed effects of pandemics/other calamities on travellers especially when the aim is to juxtapose various situations pre and post the pandemic using UGC data.

Though we make some valuable contributions, our study does not come without limitations. We have not calculated the exact review helpfulness for reviews even though we use certain parameters that can be used to determine it such as length and readability. We use only one metric for readability which can lead to some uncertainty in observations. TripAdvisor has been shown to have some cases of fake/paid reviews which compromises the validity of those reviews in the analysis (Filieri et al., 2015; Chang et al., 2019). We consider all our data only from TripAdvisor, which some previous research shows can lead to bias in results (Xiang et al., 2017). As an extension to our study, we urge researchers to explore the highlighted avenues above.

References

<https://tourism.gov.in/sites/default/files/2020-09/ITS%20at%20a%20glance_Book%20%282%29.pdf>

Kumar, V. (2020). Indian tourism industry and COVID-19: present scenario. Journal of Tourism and Hospitality Education, 10, 179-185.

Ayeh, J. K., Au, N., & Law, R. (2013). “Do we believe in TripAdvisor?” Examining credibility perceptions and online travelers’ attitude toward using user-generated content. Journal of Travel Research, 52(4), 437-452.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59, 467-483.

Filieri, R., Alguezaui, S., & McLeay, F. (2015). Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. Tourism management, 51, 174-185.

Ganzaroli, A., De Noni, I., & van Baalen, P. (2017). Vicious advice: Analyzing the impact of TripAdvisor on the quality of restaurants as part of the cultural heritage of Venice. Tourism Management, 61, 501-510.

Nieto, J., Hernández-Maestro, R. M., & Muñoz-Gallego, P. A. (2014). Marketing decisions, customer reviews, and business performance: The use of the Toprural website by Spanish rural lodging establishments. Tourism Management, 45, 115-123.

Rodríguez-Díaz, M., & Espino-Rodríguez, T. F. (2018). A methodology for a comparative analysis of the lodging offer of tourism destinations based on online customer reviews. Journal of Destination Marketing & Management, 8, 147-160.

Baka, V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. Tourism Management, 53, 148-162.

Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. Tourism Management, 66, 53-61.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. Journal of Hospitality Marketing & Management, 25(1), 1-24.

Collie, M. (2014). Internet leads travel. Geton. com. Available online: http://www. geton. com/blog/geton/internetleads-travel (accessed on 30 April 2020).

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. International Journal of Hospitality Management, 29(4), 694-700.

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. Journal of Hospitality Marketing & Management, 24(2), 119-154.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. Journal of Travel & Tourism Marketing, 32(5), 608-621.

Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. Tourism Management, 75, 550-568.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. Tourism management, 47, 140-151.

Kusumasondjaja, S., Shanka, T., & Marchegiani, C. (2012). Credibility of online reviews and initial trust: The roles of reviewer’s identity and review valence. Journal of Vacation Marketing, 18(3), 185-195.

Ghose, A., & Ipeirotis, P. G. (2010). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE transactions on knowledge and data engineering, 23(10), 1498-1512.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. Tourism Management, 52, 498-506.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Kincaid, J. P., Aagard, J. A., O'Hara, J. W., & Cottrell, L. K. (1981). Computer readability editing system. IEEE Transactions on Professional Communication, (1), 38-42.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 58(2), 175-191.

Tsai, C. F., Chen, K., Hu, Y. H., & Chen, W. K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. Tourism Management, 80, 104122.

Razzaghnoori, M., Sajedi, H., & Jazani, I. K. (2018). Question classification in Persian using word vectors and frequencies. Cognitive Systems Research, 47, 16-27.

Martinez-Torres, M. D. R., & Toral, S. L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. Tourism Management, 75, 393-403.

Chen, S. X., Wang, X. K., Zhang, H. Y., Wang, J. Q., & Peng, J. J. (2021). Customer purchase forecasting for online tourism: A data-driven method with multiplex behavior data. Tourism Management, 87, 104357.

"Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word

representations in vector space. In Proceedings of workshop at ICLR.

Mikolov, T., Sutskever, I., Chen, K., Greg, C., & Jeffrey, D. (2013b). Distributed

representations of words and phrases and their compositionality. In Proceedings

of neural information processing systems.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013c). Linguistic regularities in continuous

space word representations. In Proceedings of the 2013 conference of the North

American chapter of the association for computational linguistics: Human language

technologies. Association for Computational Linguistics."

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., ... & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert systems with applications, 167, 114155.

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. Expert Systems with Applications, 162, 113746.

Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021). Automated topic modeling of tourist reviews: does the Anna Karenina principle apply?. Tourism Management, 83, 104241.

Escandon-Barbosa, D., & Salas-Paramo, J. (2021). Tourism Amidst COVID-19: consumer experience in luxury hotels booked through digital platforms. Tourism Recreation Research, 1-6.

Yu, M., Li, Z., Yu, Z., He, J., & Zhou, J. (2021). Communication related health crisis on social media: a case of COVID-19 outbreak. Current issues in tourism, 24(19), 2699-2705.2

Uğur, N. G., & Akbıyık, A. (2020). Impacts of COVID-19 on global tourism industry: A cross-regional comparison. Tourism Management Perspectives, 36, 100744.

Shin, S., Du, Q., Ma, Y., Fan, W., & Xiang, Z. (2021). Moderating effects of rating on text and helpfulness in online hotel reviews: an analytical approach. Journal of Hospitality Marketing & Management, 30(2), 159-177.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. Tourism management, 47, 140-151.

Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3), 221.

Valdivia, A., Luzón, M. V., & Herrera, F. (2017, June). Sentiment analysis on tripadvisor: Are there inconsistencies in user reviews?. In International Conference on Hybrid Artificial Intelligence Systems (pp. 15-25). Springer, Cham.

Yang, S. B., Shin, S. H., Joun, Y., & Koo, C. (2017). Exploring the comparative importance of online hotel reviews’ heuristic attributes in review helpfulness: a conjoint analysis approach. Journal of Travel & Tourism Marketing, 34(7), 963-985.

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. Journal of consumer research, 17(4), 454-462.

Shin, S., Du, Q., Ma, Y., Fan, W., & Xiang, Z. (2021). Moderating effects of rating on text and helpfulness in online hotel reviews: an analytical approach. Journal of Hospitality Marketing & Management, 30(2), 159-177.

Black, H. G., & Kelley, S. W. (2009). A storytelling perspective on online customer reviews reporting service failure and recovery. Journal of Travel & Tourism Marketing, 26(2), 169-179.

Chang, Y. C., Ku, C. H., & Chen, C. H. (2019). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. International Journal of Information Management, 48, 263-279.