Social Media Analytics: Extracting and Visualising TripAdvisor Reviews, a Case of Taj Hotels in India

[1]Alfred George, [2]R. Sreejith, [3]Smitha Siji

[1]Student, Rajagiri Business School Kochi, India
[2]Asst.Professor, Rajagiri College of Social Sciences, Kochi, India
[3]Phd, Asso.Professor, Rajagiri College of Social Sciences, Kochi, India
[1]alfredgeorgemampilly@gmail.com, [2]sreejithr@rajagiri.edu, [3]smitha@rajagiri.edu

Article Info
Volume 83
Page Number: 3881-3888
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

This research is conducted with an objective to develop a methodology that can use machine learning techniques to analyze online reviews. This will be of help to tourism and destination management professionals to understand and apply these techniques to enhance their attractions. Online reviews of Taj hotels in India were collected and TF-IDF along with LDA method was used to model topics in the reviews. VADER was applied to the reviews to analyse the sentiment of customers towards the hotel brand.

Keywords: *Hospitality, Natural language processing,Social media analytics, Visual analytics, Google trends,TripAdvisor, LDA, VADER)*

1. Introduction

Before purchasing a product or service, consumers look for information about it online. User ratings and reviews of the product/service give them important cues about the product/service. Organizations have ensured their presence online and also given consumers an opportunity to post a review about the services they received. Apart from the organisation's own website, there are third-party websites that also compares the different service organisations and post a rating as well as a review about these organizations. Hence, from being a company generated information on quality (in the form of advertisements, price, brand name etc.) perception of quality has now become user-controlled (in the form of ratings and reviews). These user-generated reviews have become a significant driver of sales across products and industries. This study focuses on the actual and perceived utility of online user ratings and reviews.

In the case of the service industry, most often, it's not possible for customers to personally visit and review the service facility. This necessitates the customer to rely upon the reviews posted by other customers or experts in that area.

Consumer-generated content has created an important new information medium for tourists, changing the way visitors assess, choose and share tourism experiences throughout the buying lifecycle.Numerous studies have suggested that customer satisfaction plays an important role in promoting behavioural loyalty of consumers, such as positive reviews, returns or recommendations..

Extant work has shown that online user feedback can be used as a significant source of information for analysts and analysts who can seek to interpret consu mer preferences and demand correctly: for example, f orecasting financial performance or trying to increase revenue.

When online ratings, that are numerical and easily understood are compared to online reviews which are text based and often comprise large repositories of information, it is seen that online reviews are beyond the analytical capabilities of traditional econo metric and statistical methods. To date, large scale online reviews provide insufficient scientific evidence to help explain consumer satisfaction and its antecedents.

User-generated content (UGCs), like hotel reviews, are important for certain prospective customers, distributors and product manufacturers and also for owners of companies as they reveal consumers ' opinions on their services and products. [8] UGC can be viewed as random, informative and enthusiastic consumer feedback that is easily available, costs nothing or is low-cost, and that can be accessed easily everywhere, anywhere.

Social status and other topics like this, on the other hand, are latent dimensions which may not be explicitly mentioned by customers but which absorb or represent a large range of characteristics, many a times indirectly from other indices (e.g., income and employment). Analysis on online reviews has several benefits, including data access, data collection speed and usability and non-intrusiveness of human subjects. [1]This paper attempts to illustrate the utility of text mining techniques in the study of the textual data contained in the narratives of the Taj hotels customer reviews on TripAdvisor. There was no published paper examining the Taj Hotels TripAdvisor reviews by text mining methods to the best of the authors ' knowledge. However, previous research has been carried out to explore the use of text mining techniques to decode consumer reviews relating to other hotels such as Hilton.[2]

The sequence of the research is as follows; First data was collected from TripAdvisor using R studio using a library called 'rvest' and the review for each Taj hotel was collected and converted into a .csv file. After collecting all the hotel reviews it was joined to one .xlsx file and imported to python environment using Jupyter notebook.

The necessary libraries were added which included nltk, NumPy, string, sklearn, TfidfVectorizer,

LatentDirichletAllocation, stopwords from nltk, matplotlib, pos_tag and word_tokenize from nltk.

In our approach, the reviews were pre-processed with tokenization, filtering stop words, stemming and non-removal of special characters like "!","?" and emoticons like ":)",":(",":D". Case transformation and removal of special characters were purposely ignored to preserve the raw nature of reviews as in the study Vader sentiment analysis method was used which considers special characters, emoticons and case sensitive words. This will be explained further in the coming chapters. Further, the Term Frequency / Inverted Document Frequency (TF/IDF) has been adopted. TfidfVectorizer was used to create a term-document matrix specifying the ngram. The study used a unigram method to understand the latent topics in the corpus. The number of occurrences of each word for each document, and imposes appropriate standardization on the terms ' frequencies. Tf-idf's final result is a term-document matrix, whose columns include the normalized frequency of words for each of the corpus documents.

Then LDA model was used to divide the documents into 5 topics. The number of topics was found out by trial and error. The dominant topics in each document were identified and summed. Top 10 tri-grams in each topic in descending order of weights were identified and used to name the topics or aspects. Further, Vader sentiment analysis method was applied to find the polarity scores of the reviews in each topic. The sentiment was classified into five categories namely high positive, positive, neutral, negative and highly negative. Various visualization techniques were applied using matplotlib library which will be discussed in the visualisation and findings section.

In this study only Taj Hotels were considered, a premium hotel chain from India having presence internationally. Taj is a symbol of Indian hospitality. Taj brand of the hotel was selected primarily because; Taj hotels have an ample number of reviews on TripAdvisor. Considering google trends data for the past 16 years it was identified that Taj has been searched more frequently than other Indian hotel brands like Lalit, ITC,Leela Group and Oberoi group.TripAdvisor was chosen because of its rich WOM information which may affect other consumers' behaviours.

The remaining of the article is articulated as follows. First, a literature review of contemporary research of ratings and reviews, social media analytics, outcomes of internet search, hospitality and tourism, and sentiment as well as aspect-based analysis of sentiments. Next, a scraping program was developed to collect Taj hotel reviews and ratings from TripAdvisor using R. Then the methodology will be discussed. Visualization tools are used to interpret details from TripAdvisor and Google Trends to discover hidden information and connections that can be used for the purposes of marketing or decision making. Finally, conclusions were drawn and future lines of the study were suggested.The scope of this research is limited to the luxury hotels of Taj in India. The research questions the paper aims to answer are as follows:

1. What are the dimensions/ aspects of customer reviews?
2. How many reviews are there in each aspect?
3. How positive or negative are each aspect of the reviews?
4. What are the managerial and marketing implications from the reviews?

2. Literaturereview

Ratings and Reviews

Online reviews generally have qualitative and quantitative features, quantitative in the form of ratings and stars and qualitative is in the form of a description. One of the important partsof the review is content. The reviews can be positive or negative. The quality of the review content is an important determinant of adoption of the product andthe intention to purchase. Previous researches have indicated that consumers trust reviews posted online. Their perceptions about the firm as well as purchase decisions are influenced by such reviews (. But the problem faced by firms is the posting of positive as well as negative reviews. These reviews remain on the sites for a long term affecting the reputation of the firms for an extended period of time. Thus, it is important for firms to know how do customers evaluate these ratings and reviews posted on different online platforms. Though there are studies that have investigated the influence of online reviews on firm performance indicators such as hotel room bookingsor popularity of the firm, trust in the hotel and intentions to book, there are very limited studies on the impact of the online reviews based on its presentation and content as well as the utility of such reviews.

Social Media Analytics

Social media analytics involves techniques to collect, extract, analyze, and present user-generated data to support decision making, insight discovery, or other business-related operations. [3]
Websites on social media allow users to create and distributeinformation and interactions through immersive Web 2.0 technology and apps, such as Facebook, Twitter, TripAdvisor, etc. This has led to an explosion of online UGCs, e.g., tweets, opinions, and reviews. [2]

Internet Search results

Google Trends provides access to user trends on the Web byevaluating a percentage of all online searches on the Google Search platform and other Google-related pages.

Search engines allow users to access relevant data by entering keywords. Such search queries not only lead to a huge amount of UGC but also provide valuable information about users' interests and intentions.

Google Trends is a search trend feature that indicates how often a given search term is used in Google's search engine compared to the overall search volume of the internet over a given time period. Google Trends offers keyword related data including index search volume and search engine usage geographic detail.

Researchers adopted data on search volumes to predict manysocial and economic activities. The forecasted factors contain levels of unemployment.consumer priceshousing prices [4], and stock prices[5].

Researchers have the relationship between consumers' hotel searching behaviour and hotel booking data on Expedia.com. They found that the higher volume of search queries for a hotel was associated with a higher room booking rate.

Fig. 1 shows a sample of query results on Google Trends over a 16-year period, 2004–2020, across five popular Indian hotel brands, Taj, Leela, Lalit, ITC and Oberoi. During this tenure,the Taj brand of hotels has had the most worldwide search queries during the last 16 years.

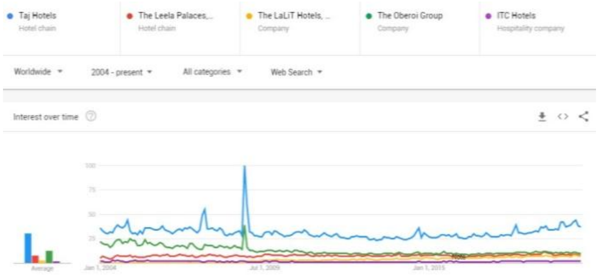


Figure 1 : The trend between leading hotel brands in India

Hospitality, tourism and hotel reviews

Online reviews are important sources of research and practice in the area of tourism. They depict the joy of visitors and their disaffection with their experiences. [6]

Online reviews have a number of functions: communicating the feelings of the readers, explaining real experiences, providing feedback and supplying important information. [7]

Online reviews are an essential tool for a destination since they can boost their competitive advantage.
[8]Researchers found that the ratings would assess service providers ' efficiency and affect perceived value , image and reputation.

These often influence customers ' choices, such as their intention to travel, probability of booking, conversion rate [9], commitment[10], and intention to suggest; hence, they affect popularity. Though there have been several recent studies on user reviews such as Expedia, Yelp [11], Booking.com, AirBnb and Google reviews, TripAdvisor is more preferred by researchers.

According to its official website, TripAdvisor had an average of over 490 million monthly users, and a record of over 73 0 million ratings and views by the end of 2018.Not only do the reviews on TripAdvisor typically contain rich information [12]but the platform has also maintained its reputation by policing the system to avoid false reviews despite some past controversies

Sentiment analysis

Furthermore, sentiment analysis can be used in various applications such as financial scenarios and political prediction, e-health and e-tourism, user profiles and user influence, community detectionand dialogue systems. Also important at the level of individual reviews is the valence of reviews. Reviews that give a high rating, or positive reviews, posted when visitors are pleased with an event are deemed a powerful tool for marketing products efficiently and effectively and enhancing brand recall.[13]Positive reviews are very relevant as consumers evaluate hedonic usage products and services since they contribute to favourable product and service perceptions [14].

Research has found that people find negative reviews to be more reliable, beneficial, predictive, insightful, convincing, trustworthy and meaningful than positive reviews [15]. The reason is that people are more inclined to avoid failure than to strive for benefits. [16]. It is therefore important that the destinations pay close attention to the negative.

Aspect based sentiment analysis

Aspect based sentiment analysis (ABSA) is part of an analysis of fine-grained sentiments. It is intended to define in a paragraph the polarity of a portion of an entity or element. ABSA can be classified into two further forms according to the subject of study: the aspect term (or target) analysis of feeling and the aspect type/dimension analysis of sentiment As the name suggests, the aspect term sentiment analysis detects the polarity of a specific aspect which occurs in the text.

Dimension type sentiment analysis attempts to capturing the polarity of the group of dimension/aspect to which a focus belongs. For example,A review on a hotel "The rooms was clean and the staff were friendly", aspect terms are "rooms"," clean", "staff" and "friendly" with sentiment towards these terms being positive. The aspect terms "room" and "clean" belong to the aspect category of "ROOM" signifying they had a positive experience regarding the rooms. The aspect terms "staff" and "friendly" belong to the aspect category of "SERVICE" and the sentiment signifies that they had a positive experience with the staff of the hotel. In this study, the focus will be on aspect category/dimension detection.

Aspects extraction in sentiment analysis is now becoming an active area of research as it is the most vital task in the aspect basedrecognition.. Aspects of reviews are domain-based and differ from context to context. Assessment of feelings based on dimension has been widely used in various application domains such as product reviews, social media, hotel reviews and restaurant reviews.

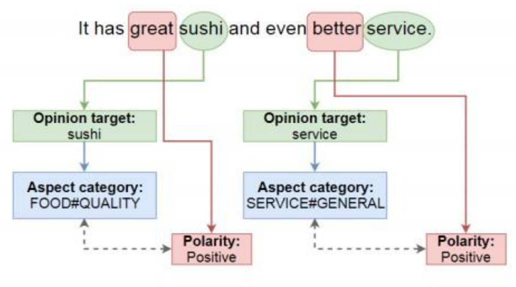

It has great sushi and even better service.
Opinion target: sushi
Aspect category: FOOD#QUALITY
Polarity: Positive
Opinion target: service
Aspect category: SERVICE#GENERAL
Polarity: Positive

Figure 2 : Aspect based review breakdown

Another example of aspect detection is shown in fig.2." Sushi" is one of the aspect term or opinion target and "great sushi" signifies a positive sentiment on the aspect category of food quality. Similarly, "service" is an aspect term or opinion target and "better service" indicates a positive sentiment towards the aspect category of service.

3. Methodology

Data collection and preprocessing

A robotic technique was used to collect and analyze the results. Reviews are gathered at a disaggregated level within Trip Advisor (i.e., user reviews) and thus it was necessary to extract individual website reviews and compile them for our study. Text pre-processing used steps which were very similar to those adopted in earlierr studies and that includeddiscarding non-English characters and words, word text tokenization, word stemming, and removing stopwords. For example, an original review appeared as:

"Located near one of the largest malls in Kochi, this hotel has an excellent buffet and service. A quick and complimentary buggy ride to the Lulu mall is convenient. Hotel staff was courteous and prompt. The servers in the restaurant from Sikkim, Nepal and Assam were very polite."

After pre-processing of the text, the review became. "located near one largest Kochi, hotel excellent buffet service a quick complimentary buggy ride convenient hotel staff courteous prompt the servers restaurant Sikkim, NepalAssam polite"

The study also implemented a text pre-processing by using modules of the Natural Language Toolkit (www.nltk.org) in the Python programming environment.

TIFD Vectorizer (ngram)

TF-IDF is a numerical statistic that shows the relevance of keywords to some specific documents. In other words, it can be said that it provides those keywords which helps in identifying or categorizing some specific documents. Previous researches have used various vector space models to automatically measure semantic text similarity. These include TFIDF, the models of topics and neural models.Term Frequency –

Inverse Document Frequency (TFIDF) is one of the most common vectorizing techniques with many possi ble variations for textual data .The theory behind TFI DF is to scale down the importance ofterms that are w idespread in many documents, considering that all tho se terms that hold less specific information to a focal t ext..

Term Frequency

TF is used to measure how many times a term is present in a document. Assume there is a text "Doc 1" comprising 10000 characters, and precisely 20 times the word "Beta" is included in the document. It is well known that the overall length of documents may vary from very small to large, so it is possible that any term may occur more frequently in large documents than in small documents. So to rectify this issue, to find the word frequency, the incidence of some term

in a document is separated by the total words present in that document. The term frequency of the word "Beta" in the document "Doc 1"
will be therefore in this case

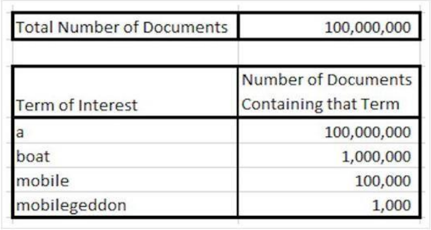
TF= 20/10000 = 0.002

Inverse Document Frequency

Simply put it's a test of a term's rareness. The inverse document frequency assigns a lower weight to frequent words and assigns a greater weight for the words that are infrequent.

$$idf_j = log\left[\frac{n}{df_j}\right]$$

Let's consider the following example.



Total Number of Documents	100,000,000

Term of Interest	Number of Documents Containing that Term
a	100,000,000
boat	1,000,000
mobile	100,000
mobilegeddon	1,000

Figure 3 : Example of Inverse Document Frequency

In fig 3, it can be inferred that every document in the document collection displays the term "a." What this teaches us is that saying the papers apart doesn't provide any benefit. It's all in it. Now, look at the word "mobilegeddon." It appears in 1,000 or a thousandth of one per cent of the documents. This term explicitly offers much more distinction for the documents which contain them.
Here IDF of mobilegeddon will be =
log_e(100000000/1000)= 5

Term Frequency - Inverse Document Frequency (TFIDF)

TFIDF means that higher term frequency will be given to words that have higher occurrence and the less occurrence of the word in documents will generate higher importance (IDF) for that keyword when searched in a particular document.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the most common method for topic modelling and is a generalization of probabilistic latent semantic indexing (PLSI).
Fundamental concepts are explored from massive volumes of unstructured text data and big data through Topic modelling that utilizes LDA. Using this method, one can quickly discover a combination of topics (such as aspects influencing hotel customers ' satisfaction) from large numbers of documents (i.e., reviews). It also helps in achieving a series of specific purposes, including identifying an optimum number of dimensions, labelling the dimensions, and evaluating the heterogeneity and relative importance of dimensions based on different reviewer characteristics.

LDA is understood as a corpus generative probabilistic model. Reviews are explained as random mixtures over K latent dimensions, in which each dimension is described by word distribution. In other terms, customer feedback reflects with percentages the different aspects of K. For example, a consumer can give their textual review based on personal opinions that mean 20 per cent for service, 20 per cent for cleanliness, 25 per cent for cost and 35 per cent for the friendliness of the staff.

Instead of other text analysis methods contained in the literature, our study implemented the LDA approach, based on the following criteria. First, the LDA model excels in effectively processing extremely granular large-scale data and thus helps us to investigate the variation of measurements across different customer categories (e.g., male vs. female). Furthermore, LDA allows us to measure the realistic level of incidence in hotel reviews for each derived factor dependent on its severity. Tourists, for example, choose terms from their own vocabulary to convey personal opinions regarding various aspects of hotels, such as location, services, size, etc.

Based on the assumption of the bag of words LDA describes a text as a mixture of latent subjects in which a topic is a multinomial distribution of phrases. The paper will have its own combining ratio of topics and each topic will have its own term distribution. LDA helps to uncover latent subjects

from our vast unstructured analysis data

For working with collections of structured, unstructured, and semistructured text documents, LDA is considered as a powerful tool, from the plenty of tools available in software repositories.

Vader Sentiment Analysis

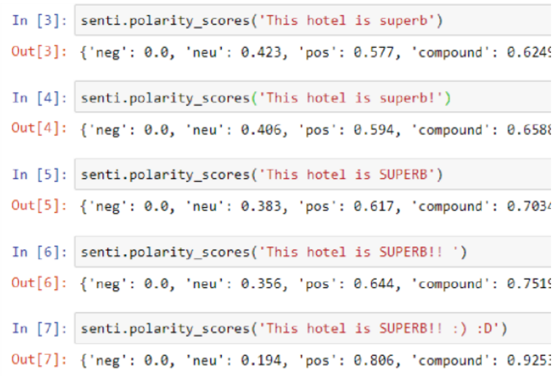
VADER stands for Valence Aware Dictionary for sEntiment Reasoning. Researchers contrasted the VADER opinion lexicon to seven other wellestabl ished sentiment analysis lexicons in the research cond ucted by C.J. Hutto, 2015): Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, and the Hu-Liu04 opinion lexicon and found out that the VADER lexicon performs exceptionally well in the social media domain, and generalizes favourably.

In the study, expresses the goals on which they based the creation of VADER as the following:
“1) operates well on social media type language, but g eneralizes effortlessly to multiple domains, 2) needs n

o training data, but is developed from a generalizable, valence-based, human cured gold standard emotion lexicon 3) is simple enough to be used online with streaming data, and 4) does not suffer significantly from a speed-performance tradeoff.”

VADER was constructed by examining and selecting features from three previously constructed and validated lexicons as a candidate list [7]; Linguistic Inquiry and Word Count (LIWC) [23], Affective Norms for English Words (ANEW) [24], and General Inquirer (GI) [25]. The authors also added common social media abbreviations, slang, and emoticons.

VADER sentiment analysis relies primarily on a dictionary that compares lexical characteristics to the intensities of the emotions called sentiment ratings. You can get the sentiment score of a document by summing up the intensity of each term in the sentence.We will include not only words but also emoticons like":-)" in a standard post, acronyms like "LOL," and phrases like "meh ». The interesting thing about a study of VADER's opinion is that these colloquialisms are also translated to strength levels. As far as the authors are concerned there has not been a previous study on TripAdvisor reviews for Taj hotels in India.



```
In [3]:  senti.polarity_scores('This hotel is superb')

Out[3]:  {'neg': 0.0, 'neu': 0.423, 'pos': 0.577, 'compound': 0.624

In [4]:  senti.polarity_scores('This hotel is superb!')

Out[4]:  {'neg': 0.0, 'neu': 0.406, 'pos': 0.594, 'compound': 0.658

In [5]:  senti.polarity_scores('This hotel is SUPERB')

Out[5]:  {'neg': 0.0, 'neu': 0.383, 'pos': 0.617, 'compound': 0.703

In [6]:  senti.polarity_scores('This hotel is SUPERB!! ')

Out[6]:  {'neg': 0.0, 'neu': 0.356, 'pos': 0.644, 'compound': 0.751

In [7]:  senti.polarity_scores('This hotel is SUPERB!! :) :D')

Out[7]:  {'neg': 0.0, 'neu': 0.194, 'pos': 0.806, 'compound': 0.925
```

Figure 4 : Example using VADER

From this example above of a simple text “This hotel is superb” we can see that the polarity scores change significantly each time a small change is made to the word “superb”.

In the first case, the text is a simple text and it receives a compound score of 0.6249. In the second instance,an exclamation mark was added after the word “superb” and the compound score increased to 0.6588 indicating that more excitement was shown by the customer as he wrote an exclamation mark after the word “superb”. Similarly,it can be seen that compound score again rises significantly to 0.7034 and 0.7519 as the word “superb” was changed to uppercase and when an exclamation mark was added after the word “superb” in uppercase. The most significant change in this example can be seen when the word “superb” was both changed to uppercase and added an exclamation mark and also added emoticons at the end “:)” and “:D”. I can be seen that the compound score increased from 0.7519 to 0.9253 indicating that the reviewer was very much happy about the hotel experience compared to the previous reviewers

Thus, it can be understood that Vader is a good method of sentimental analysis as it takes into consideration aspects like uppercase, exclamation marks and emoticons which are very common in social media reviews

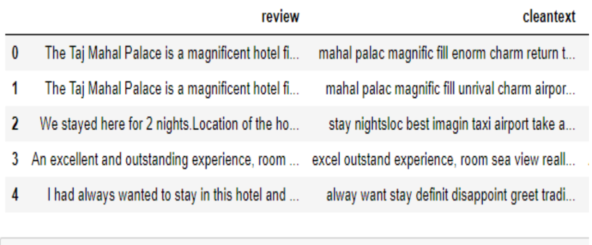
Visual analytics

Visual analytics provide an interactive visual interface to present hidden structures and details. Machine and human strength are combined to process and explore large data and provides decision-support information. The study uses Tableau to visualize google trends data and aspect sentiment of the reviews from TripAdvisor.

Timeline analysis, location-based analysis, and dashboard analysis are the visual analyses provided. Apart from that, it also provides trendline, forecast, and cluster models for data analytics.

4. Results and Discussions

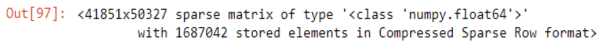
In fig.4 the reviews were first pre-processed by removing special characters, converting to lower-case, stemming and removing stopwords. The stopwords were imported from the nltk package and certain words were added to the existing words using extend function. str.replace function was used to replace special characters except !,?,:,;,(,) which are used as emoticons and expressions in reviews.



	review	cleantext
0	The Taj Mahal Palace is a magnificent hotel fi...	mahal palac magnific fill enorm charm return t...
1	The Taj Mahal Palace is a magnificent hotel fi...	mahal palac magnific fill unrival charm airpor...
2	We stayed here for 2 nights.Location of the ho...	stay nightsloc best imagin taxi airport take a...
3	An excellent and outstanding experience, room ...	excel outstand experience, room sea view reall...
4	I had always wanted to stay in this hotel and ...	alway want stay definit disappoint greet tradi...

Figure 4: Raw review and pre-processed review

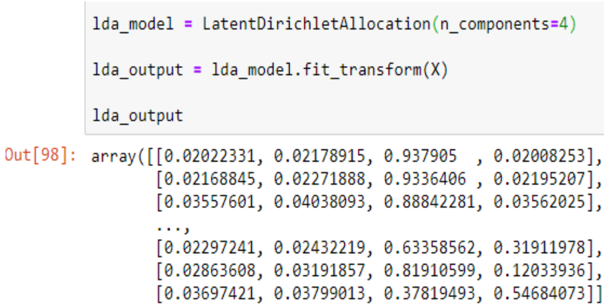
Using the TfidfVectorizer a sparse matrix of 41851 documents and 50327 terms were created which contained a total of 1687042 elements. (fig 5)



```
Out[97]: <41851x50327 sparse matrix of type '<class 'numpy.float64'>'
                with 1687042 stored elements in Compressed Sparse Row format>
```

Figure 5: Sparse matrix creation

Using the LDA model the reviews were modelled into 4 topics in fig 6.

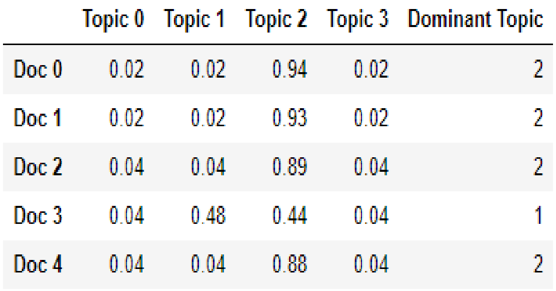


```
lda_model = LatentDirichletAllocation(n_components=4)

lda_output = lda_model.fit_transform(X)

lda_output
```

```
Out[98]: array([[0.02022331, 0.02178915, 0.937905  , 0.02008253],
                [0.02168845, 0.02271888, 0.9336406 , 0.02195207],
                [0.03557601, 0.04038093, 0.88842281, 0.03562025],
                ...,
                [0.02297241, 0.02432219, 0.63358562, 0.31911978],
                [0.02863608, 0.03191857, 0.81910599, 0.12033936],
                [0.03697421, 0.03799013, 0.37819493, 0.54684073]]
```

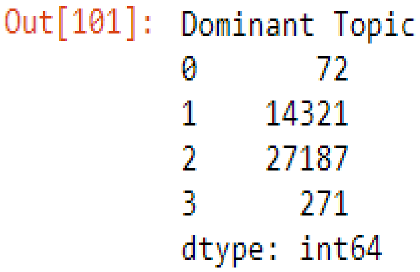
Figure 6: Topic modelling using LDA

In fig 7, a new column named “dominant topic” was added to the dataframe to indicate in each document which topic was more dominant. When a sum of dominant topics was done it could be seen that 72, 14321, 27187, 271 were document sum for topics 0,1,2 and 3. (fig 8)



	Topic 0	Topic 1	Topic 2	Topic 3	Dominant Topic
Doc 0	0.02	0.02	0.94	0.02	2
Doc 1	0.02	0.02	0.93	0.02	2
Doc 2	0.04	0.04	0.89	0.04	2
Doc 3	0.04	0.48	0.44	0.04	1
Doc 4	0.04	0.04	0.88	0.04	2

Figure 7: Identifying the dominant topics in each document



```
Out[101]: Dominant Topic
          0       72
          1    14321
          2    27187
          3      271
          dtype: int64
```

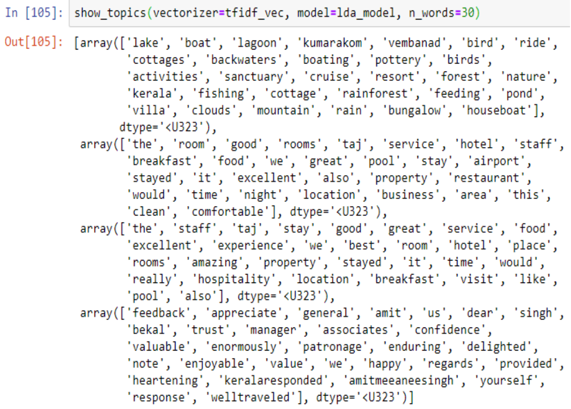
Figure 8: Sum of documents in each topic

In fig 9, the 4 topic terms are categorized. The first array contains words like ‘lake’, ‘ride’, ‘backwaters’, ‘boating’, ‘cruise’, ‘forest’, ‘clouds’, ‘mountains’ signifying the ambience of the stay at Taj hotels around India.

The second array has words like service, staff, location, clean, rooms signifying the amenities of the Taj hotels.

The third array contains words like stay, great, property, hospitality signifying the quality of stay.

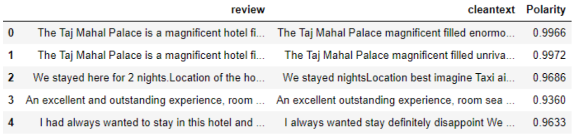
The fourth array has words like manager, associates, confidence, value, patronage, trust signifying the values and behaviour of employees and managers of Taj that the customers could identify.



```
In [105]: show_topics(vectorizer=tfidf_vec, model=lda_model, n_words=30)

Out[105]: [array(['lake', 'boat', 'lagoon', 'kumarakom', 'vembanad', 'bird', 'ride',
                  'cottages', 'backwaters', 'boating', 'pottery', 'birds',
                  'activities', 'sanctuary', 'cruise', 'resort', 'forest', 'nature',
                  'kerala', 'fishing', 'cottage', 'rainforest', 'feeding', 'pond',
                  'villa', 'clouds', 'mountain', 'rain', 'bungalow', 'houseboat'],
                 dtype='<U323'),
           array(['the', 'room', 'good', 'rooms', 'taj', 'service', 'hotel', 'staff',
                  'breakfast', 'food', 'we', 'great', 'pool', 'stay', 'airport',
                  'stayed', 'it', 'excellent', 'also', 'property', 'restaurant',
                  'would', 'time', 'night', 'location', 'business', 'area', 'this',
                  'clean', 'comfortable'], dtype='<U323'),
           array(['the', 'staff', 'taj', 'stay', 'good', 'great', 'service', 'food',
                  'excellent', 'experience', 'we', 'best', 'room', 'hotel', 'place',
                  'rooms', 'amazing', 'property', 'stayed', 'it', 'time', 'would',
                  'really', 'hospitality', 'location', 'breakfast', 'visit', 'like',
                  'pool', 'also'], dtype='<U323'),
           array(['feedback', 'appreciate', 'general', 'amit', 'us', 'dear', 'singh',
                  'bekal', 'trust', 'manager', 'associates', 'confidence',
                  'valuable', 'enormously', 'patronage', 'enduring', 'delighted',
                  'note', 'enjoyable', 'value', 'we', 'happy', 'regards', 'provided',
                  'heartening', 'keralaresponded', 'amitmeeaneesingh', 'yourself',
                  'response', 'welltraveled'], dtype='<U323')]
```

Figure 9: List of keywords associated with a topic

In fig 10, the polarity of each document was analysed. A new column of polarity was created and the polarity scores were calculated using VADER



	review	cleantext	Polarity
0	The Taj Mahal Palace is a magnificent hotel fi...	The Taj Mahal Palace magnificent filled enormo...	0.9966
1	The Taj Mahal Palace is a magnificent hotel fi...	The Taj Mahal Palace magnificent filled unriva...	0.9972
2	We stayed here for 2 nights.Location of the ho...	We stayed nightsLocation best imagine Taxi ai...	0.9686
3	An excellent and outstanding experience, room ...	An excellent outstanding experience, room sea ...	0.9360
4	I had always wanted to stay in this hotel and ...	I always wanted stay definitely disappoint We ...	0.9633

Figure 10: Polarity scores of each review

In fig 11, a visualisation of reviews with polarity is depicted. It is clearly visible that the majority of reviews tend to have a polarity score between .5 and 1 indicating that the majority of reviews were positive.

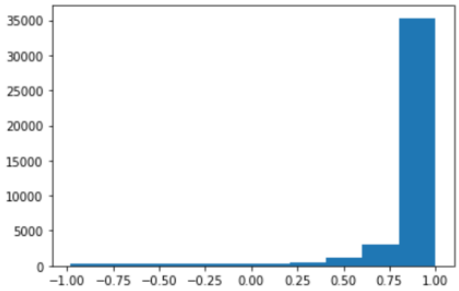


Figure 11: Graph showing polarity scores

In fig 12, a deep analysis was done to segregate polarity scores as high positive, positive, neutral, negative and highly negative. A threshold value was given to segregate the sentiment based on polarity scores.

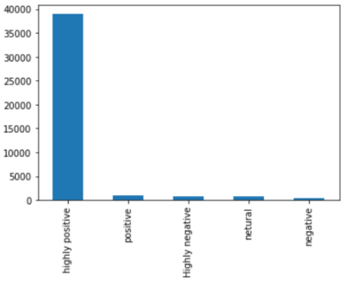


Figure 12: Graph showing the total number of reviews in each sentiment category

5. Conclusion

The study aimed to identify the latent topics inside reviews of Taj hotels in India thus allowing us to dive deeper into the meaning of customer feedback. The study also aimed at identifying the overall sentiment towards the brand and classify the customers based on high positive sentiment customers to high negative sentiment customers. In the study, it was identified that the brand Taj highly liked by its customers due to various reasons like ambience, amenities, quality of stay and values and behaviours of employees towards the customers.

6. Limitation and Future research

Although the elbow approach is an appropriate way of identifying the number of dimensions no procedure is universally accepted and reliable. This requires interpretation of the researcher in a system of study which is mostly unsupervised. This work only collected data from TripAdvisor, which could likely have a platform bias. It is therefore proposed that this approach be used in future research with input from other platforms or multiple sites together. The use of mobile devices to leave feedback may be integrated into studies to determine if the device type influences the reviews. Therefore, growing concerns about false and paying online reviews have arisen, and scores and comments may have bee n distorted. Future research can look into these matters.

References

[1] S. S. Weilin Lu, "User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software," 2015.

[2] C.-H. K. C.-H. C. Yung-Chun Chang, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," International Journal of Information Management, p. 17, 2017.

[3] S.-H. H. R. P. Clyde W. Holsapple, "Business Social Media Analytics: Definition, Benefits and Challenges," Americas Conference on Information Systems (AMCIS) 2014, At Savannah, GA, USA, 2014.

[4] E. B. Lynn Wu, "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales," National Bureau of Economic Research, 2015.

[5] O. S. N. M. A. S. K. Alqaryouti, " Aspect-Based Sentiment Analysis Using Smart Government Review Data," Applied Computing and Informatics, 2019.

[6] S. B. Alton Yeow-Kuan Chua, "Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality," Computers in Human Behavior , pp. 547-554, 2016.

[7] H. L. Z. W. R. L. Qiang Ye, "The influence of hotel price on perceived service quality and value in e-tourism: an empirical investigation based on online traveler reviews," Journal of Hospitality & Tourism Research, 2014.

[8] P. B. F. C. A. M. Maria Immacolata Simeon, "Exploring tourists' cultural experiences in Naples through online reviews," Journal of Hospitality and Tourism Technology, 2017.

[9] H. Ö. Asunur Cezar, "Analyzing conversion rates in online hotel booking," International Journal of Contemporary Hospitality Management, pp. 286-304, 2016.

[10] D. G. M. A. G. Martinez, "The influence of online ratings and reviews on hotel booking consideration," Tourism Management, pp. 53-61 , 2017.

[11] Y. W. Makoto Nakayama, "The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews," Information & Management , 2018.

[12] Q. D. Y. M. W. F. Zheng Xiang, "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism," Tourism Management, 2017.

[13] Q. Y. R. L. Huiying Li, "Determinants of Customer Satisfaction in the Hotel Industry: An Application of Online Review Analysis," Asia Pacific Journal of Tourism Research, 2013.

[14] X. J. Mingming Cheng, "What do Airbnb users care about? An analysis of online review comments," International Journal of Hospitality Management, 2018.

[15] L. X. R. L. Markus Schuckert, "Hospitality and Tourism Online Reviews: Recent Trends and Future Directions," Journal of Travel & Tourism Marketing, pp. 608-621 , 2015.