UNDERSTANDING THE TRAITS CAUSING DISSATISFACTION AMONG FOREIGN TOURISTS: SENTIMENTAL ANALYSIS AND TEXT MINING OF ONLINE REVIEWS

Abhishek Anjana - 2014IPG-001
Akshit Agarwal - 2014IPG-008
Ayush Gupta - 2014IPG-026

October 30, 2017

ABV-IIITM Gwalior
Gwalior-474 010, MP, India

INTRODUCTION

[fragile] black)

- Tourism contributes as much as 6.77% to India's total Gross Domestic Product through direct and indirect impact (National Council for Applied Economic Research).

- According to the ministry of tourism, India hosted 8.89 million tourists last year compared to only 2.65 million tourists in 2000.

- Although foreign tourist arrivals in India have continued to grow for the last few years, the relative growth has dipped sharply. From 26% in 2004 to just 4.5% in 2015. Such a low growth of foreign tourist arrivals is a serious cause of concern (Ministry of Tourism).

INTRODUCTION

- Tourist satisfaction is one of the most important factors in luring tourist as it has a direct effect on the choice of a destination, consumption of products and services, decision to return and improved destination reputation (Kozak & Rimmington, 2000; Yoon & Uysal, 2005; Zabkar, Brenčič, & Dmitrović, 2010).

PROBLEM STATEMENT AND OBJECTIVES

PROBLEM STATEMENT

The project aims to assess the international tourists' experience at the most iconic and leading sites of cultural importance in the country. In this regard, text mining and sentimental analysis of reviews given by foreign tourists' based on their experiences at the leading cultural and heritage associated tourist places of India has been conducted.

OBJECTIVES

To refine the broad project aim, several sub-aims are set. These sub-aims include:

- Assimilation of unstructured tourist reviews from tripadvisor.com via web scraping technique.
- Sentiment Analysis and Text Mining of the so collected reviews in order to extract dissatisfaction traits.
- Interpret the findings and narrowing down those into themes and categories with the wider ambit of understanding the full trajectory of tourists' interactions with the sites.

MOTIVATION

MOTIVATION

India possesses a plethora of heritage sites (36 of which have been termed as world heritage sites), and an overall culture which is unique. Even after this rich cultural heritage, India is still not as lucrative a tourist destination. This problem has thus lend us a motivation behind studying the topic.

SOCIAL MOTIVATION:

- Encourages infrastructure preservation and development of health care facilities, recreation areas, hotels and restaurants. etc.

ECONOMIC MOTIVATION:

- Generate income, resulted in poverty alleviation and a great number of jobs.

LITERATURE REVIEW

CONSUMER SATISFACTION AND IMAGE OF A LOCATION

- Consumer satisfaction is defined as the measure of how products and/or services supplied by a company meet or surpass a customer's expectations (Jamejami et al).
- A satisfied customer not only spends more time, plans more frequent visits but also brings with him loyalty and a good word of the mouth publicity, potentially providing a new customer and thereby a new source of income (Bowen & Shoemaker,1998; Shoemaker & Lewis, 1999).

LITERATURE REVIEW

- The reviews and comments given by a tourist thus are a very important link to define the perception of a location in the minds of the other potential tourists.
- Working closely upon and critically analyzing the reviews and recommendations of the tourists, we can come at the conclusion upon what are some of the most common issues faced by the tourists
- The above found issues can be worked upon and improved thereby increasing tourist satisfaction and thus in-turn bolstering the number of tourist arrivals in the country.

E-WOM AND ONLINE REVIEWS

- The advent of Web 2.0 engendered an influx of opinions and user generated content.
- A person who could earlier share his/her opinion in a closed group containing a few people (through word of the mouth) can now effectively tell everyone in the world because of the online websites (Electronic Word Of Mouth) (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004).

LITERATURE REVIEW

- Such an online presence of any reviews has brought about transparency in the business for the consumers, as, the travellers can now bank upon a rather reliable source of information.
- The use of review systems in travel planning is even more prominent as research indicates that searching for travel related information is considered one of the most popular online activities.
- Opinion Research Corporation (2008) found that more than 60% of the responded confirmed to have consulted online reviews and blogs before purchasing a service and grossly 80% of such people, accepted that the reviews had some impact on their ensuing choices.

TRIP ADVISOR

- Sites like TripAdvisor have come into prominence in the past decade because of the development of Web 2.0, and have been able to create a huge market for themselves.
- TripAdvisor, today is the world's largest travel site (comScore Media Metrix, 2016). The effectiveness and popularity of TripAdvisor can be judged from the fact that there are 290 new user contributions to TripAdvisor every minute (TripAdvisor, 2017).
- Of a survey conducted on TripAdvisor users in 2007, 96.4% of the respondents accepted to use the internet as their primary source of information for travel planning decisions and out of those 82.5% always use the internet for planning their trips (54 Gretzel, Yoo & Purifoy).

LITERATURE REVIEW

- Opinions and reviews given by fellow tourists are seen as being more helpful and authentic than those endorsed by travel and tourism agencies and/or advertisement.
- Travellers can go on the site and post their reviews about a location, rate the location on a 5-star scale, where 1 star denotes terrible, 2 stars denote poor, 3 stars denote average, 4 stars denote very good and 5 stars denote excellent.

LITERATURE REVIEW

- One issue that any such website faces is the issue of fake reviews.
- While, TripAdvisor says ”it would be impractical for us to fact check the details of reviews” it does also mention that such an issue is sometimes taken care of by itself, because the sheer number of genuine reviews ”allows travellers to get the facts, spot trends among reviews and determine whether a property is right for them”.
- The power of the masses is an aspect that is helpful to TripAdvisor. Considering the large number of authentic and genuine user base the company caters to, in the long run, the effect of the fake reviews is bound to get undermined and diminished by the large number of the actual reviews that are posted (TripAdvisor Review Moderation and Fraud Detection FAQ).

RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

DATA COLLECTION

- Web Scrapping
- Scrapping Aggression And Politeness Policy
- What Was Collected?
- Storing the data

DATA ANALYSIS

- Sentiment Analysis
- Sentiment Classification
- Binary Versus Multi-polarity Analysis
- Review Classification
- Internal Validity
- Part Of Speech Tagging
- Stop Word Removal
- Tokenization
- Text Link Analysis
- Clustering

WEB SCRAPPING

- Web scrapping technique used for extraction of data and desired contents from a website.
- Python was chosen as the preferred language for writing this code because of its open source availability and ease of use.
- Beautiful Soup library was utilized along with multi-threading for the process.

SCRAPPING AGGRESSION

- An overly aggressive web scrapper can potentially result in downtime for the target site. A short sleep time of 1 second before each request was implemented.

WHAT WAS COLLECTED?

- Data was collected for top 5 locations chosen from the list of top 10 cultural tourist attractions across India. (Ministry of Tourism, 2016).
  - Taj Mahal, Agra
  - Agra Fort, Agra
  - Qutub Minar, Delhi
  - Humayun Tomb, Delhi
  - Red Fort, Delhi

STORING THE DATA

- Microsoft Excel was chosen to be the preferred software for the same.
- The spread sheet was split into categories corresponding to the data fields collected and the data was stored for in separate files for each tourist site further segregated by the language the data was collected in.

SENTIMENT ANALYSIS

- Sentiment analysis is referred to the ability to classify the opinions in text into different categories, typically ”positive”, ”negative” and ”neutral” (Ding, Liu & Yu 2008).

SENTIMENT CLASSIFICATION

- Given a set of documents, a sentiment classifier categorizes each document d into different classes, typically positive, neutral and negative (Liu 2007).
- The polarity score of the text between [-1 and 1], where -1 defines extremely negative and 1 defines extremely positive mood of the text with the central figure of 0 defining a neutral mood of the speaker are used.

BINARY VERSUS MULTI-POLARITY ANALYSIS

- Binary, or standard polarity analysis, means purely positive or negative sentiment. Multi-polarity analysis, results in degrees of sentiment, for instance slightly happy, a little sad and so on. Numeric values between a given range (say -1 and 1) are often used to separate such polarities.

REVIEW CLASSIFICATION

- Reviews with a rating of 4 stars and above were automatically considered to be positive overall. Reviews with a rating of 2 stars or below were classified as having a negative sentiment. For the reviews which had a rating of 3 stars, the star rating did not clarify anything in much detail as, a star rating of 3 stars classifies user rating as being average.

PART OF SPEECH TAGGING

- Part of speech (also referred to as POS) tagging, is the process of assigning to every word in a text a tag, corresponding to a particular part of the text.
- POS tagging can be seen as a way to disambiguate part of speech, where the tagger algorithm returns the most likely tag for a word based on the surrounding context.

STOP WORD REMOVAL

- Stop words are a list of the most commonly used words in the English language. Words such as and, are, as, it, of, the, is, which, at, on, among are examples.
- Such words add meaning and structure to a sentence, but, for the purpose of this project, such words do not add any viable functionality to the corpus of collected words.
- Such specific words were thus removed after POS tagging.

TOKENIZATION

- Word tokenization is the process of conversion of text in sentences to tokens of words. In our project, tokens of a single word were created.
- A simple Linux shell command was written to count the occurrence of a word and show an output representing words and their count throughout a document.

TEXT LINK ANALYSIS

- Text analysis is a technique used to recognize patterns in the text. This technique was employed to recognize words which appear with ‘no’ and ‘not’ in the review files to recognize the negative sentiment.

CLUSTERING

- Clustering refers to bringing together a group of words (similar meanings and possible synonyms) that may define a similar sentiment and share a certain degree of correlation among themselves under one umbrella category.

RESULTS AND CONCLUSIONS

RESULTS

The results revealed 5 major categories which the tourists show dissatisfaction about and also revealed some location specific and minor categories of problems.

MAJOR FACTORS

- Cleanliness, Sanitation And Hygiene
- Maintenance
- Crowd
- Value For Money And Expense
- Security And Harassment

OTHER FACTORS

- Beggars And Poverty
- Climate
- Animals
- Facilities

The data in the chart below has been arranged in the order of Humayun Tomb followed by Qutub Minar, Agra Fort, Taj Mahal and Red Fort.

Figure: Graph representation of themes and their percentages at various sites.

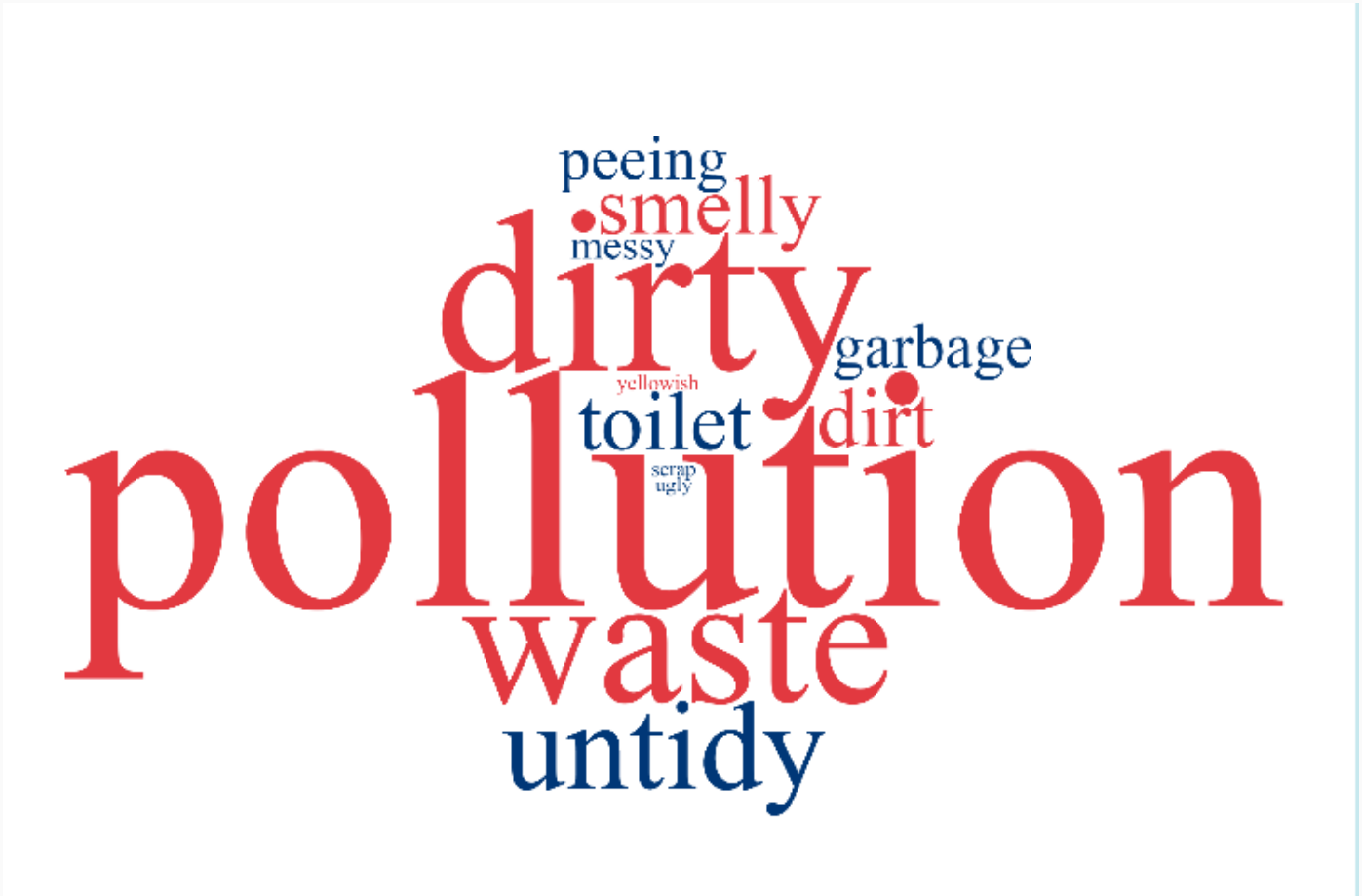

Figure: Word Cloud representation of terms used for showcasing dissatisfaction with cleanliness and hygiene.

CONCLUSIONS

- It was found out that the factors relating to tourists satisfaction and dissatisfaction belonged to both controllable and uncontrollable factors.
- Factors like weather are uncontrollable and even if they were to create the feeling of dissatisfaction, there is no way to actually control the factors.
- Other factors like grade of service and maintenance and cleanliness are factors that are controllable factors and work can be done upon these factors so in order to improve the underlying issues in these areas and increase the satisfaction level of tourists.

CONCLUSIONS

- Factors like cost, behaviour and management were pointed put in great detail in the negative reviews.
- Management and the staff that works at a tourist locations including the guides and photographers play a key role in customer satisfaction.
- The unavailability and/or worse condition of public toilets for the tourists, was seen as being a major concerning factor and is a basic need that must to be addressed at the earliest.

CONCLUSIONS

- The results of our analysis are in some agreement to other studies.
- Marin and Taberner in their research paper pointed out that cleanliness and hygiene, expense and crowd are some underlying factors that lead to a dissatisfaction response among the tourist. Also they pointed out that climate and safety were among the other issues that are taken into confederation by the tourists.
- Kozak and Remmington in their study found out factors like Value for money, hygiene, cleanliness and sanitation and feeling of safety and security.
- The findings from our project are not only in line with the above mentioned studies but also point out other factors which are somewhat unique to the Indian sphere.

THANK YOU