Tapping public sentiments on Twitter for tourism insights: a study of famous Indian heritage sites

Shruti Gulati
Department of Commerce, Delhi School of Economics, University of Delhi, New Delhi, India

Abstract
Purpose – Twitter is the most widely used platform with an open network; hence, tourists often resort to Twitter to share their travel experiences, satisfaction/dissatisfaction and other opinions. This study is divided into two sections, first to provide a framework for understanding public sentiments through Twitter for tourism insights, second to provide real-time insights of three Indian heritage sites i.e., the Taj Mahal, Red Fort and Golden Temple by extracting 5,000 tweets each (n = 15,000) using Twitter API. Results are interpreted using NRC emotion lexicon and data visualisation using R.

Design/methodology/approach – This study attempts to understand the public sentiment on three globally acclaimed Indian heritage sites, i.e. the Taj Mahal, Red Fort and Golden temple using a step-by-step approach, hence proposing a framework using Twitter analytics. Extensive use of various packages of R programming from the libraries has been done for various purposes such as extraction, processing and analysing the data from Twitter. A total of 15,000 tweets from January 2015 to January 2021 were collected of the three sites using different key words. An exploratory design and data visualisation technique has been used to interpret results.

Findings – After data processing, 12,409 sentiments are extracted. Amongst the three tourists’ spots, the greatest number of positive sentiments is for the Taj Mahal and Golden temple with approximately 25% each. While the most negative sentiment can be seen for the Red Fort (17%). Amongst the positive emotions, the maximum joy sentiment (12%) can be seen in the Golden Temple and trust (21%) in the Red Fort. In terms of negative emotions, fear (13%) can be seen in the Red Fort. Overall, India’s heritage sites have a positive sentiment (20%), which surpasses the negative sentiment (13%). And can be said that the overall polarity is towards positive.

Originality/value – This study provides a framework on how to use Twitter for tourism insights through text mining public sentiments and provides real-time insights from famous Indian heritage sites.

Keywords Twitter analytics, Visitor, Sentiment analysis, Tourism, Social media, Lexicon, R

Paper type Research paper

Introduction
Social media has transformed the tourism and hospitality sector (Leung et al., 2013) by influencing the social life of its users (Zeng and Gerritsen, 2014), specifically of tourists with decision-making and information search (Power and Phillips-Wren, 2011) and sharing of experiences pre- and post-travel (Zeng and Gerritsen, 2014). This has attracted a lot of attention by the researchers as well (Nagle and Pope, 2013), but the full potential of this domain has not yet been attained and requires more work (Leung et al., 2013; Zeng and Gerritsen, 2014).

The social media platform is a hub of enormous data relating to tourist destinations, hotels, restaurants, leisure, services etc. Tourists’ love sharing their travel-related experiences on
social media, which often include their opinions and feedback. These real-time data help researchers to study tourism sector closely (Lu and Stepchenkova, 2015). Twitter is the world’s largest microblogging website with over 290.5 million users worldwide (Statista, 2019). It has also been considered the most “influential” microblogging website (Akehurst, 2009; Thelwall et al., 2011).

India’s tourism industry plays a vital role in its economy and has been growing enormously by the day. Even in the Asia Pacific region, it has been contributing at a growth rate of 10% (Runcle and Associates, Inc., 2015), and the Asian tourism sector has been anticipated to grow at the “quickest rate” of more than 6% per annum (Fuller, 2013). India is a home to various tourist spots, such as monuments, religious spots, parks and bird sanctuaries, etc. The Ministry of Tourism in their tourism statistics have noted that the Taj Mahal and Red Fort are the top two heritage sites for domestic and international tourists in India (India Tourism Statistics, 2019). The Golden Temple, on the other hand, has been awarded as the “most visited place of the world” by the World Book of Records (WBR, 2017).

Twitter allows its users to write the content in the form of alphabets, numeric or media with a maximum of 280 characters. This is also why Twitter can be said as “global pipeline for real-time information sharing and broadcasting” (Wang et al., 2013, p. 34). This makes Twitter a high utility platform for providing research data with its “people as sensor” network (Weng et al., 2010). Thus, Twitter data are often used by companies, government to understand the public sentiment. This study attempts to understand the public sentiment on three globally acclaimed Indian heritage sites, i.e. the Taj Mahal, Red Fort and Golden temple, using a step-by-step approach, hence proposing a framework using Twitter analytics. Extensive use of various packages of R programming from the libraries has been done for various purposes, such as extraction, processing and analysing the data from Twitter. A total of 15,000 tweets from January, 2015 to January, 2021 were collected of the three sites using different key words. An exploratory design and data visualisation technique has been used to interpret results.

The following are research questions for this study:

**RQ1.** How to gain public sentiment insights for tourism industry using Twitter analytics?

**RQ2.** What is the public sentiment on the three famous Indian heritage sites such as the Taj Mahal, Golden Temple and Red fort?

**Theoretical background**

**Twitter and social media in tourism**

Social media are interactive platforms that create, modify and share user-generated content (Kaplan and Haenlein, 2010), which is the content created and posted by web users. With the advent of smart devices and Internet, the utilisation of social media has been on a rise, specifically in tourism. Social media is used by hospitality players, such as airlines, hotels, restaurants etc. On the other hand, it is used by tourists for decision-making and for sharing travel experiences (Sreeja et al., 2019) and also for information collaboration (Leung et al., 2013).

Twitter is the source of million tweets (Velde et al., 2015) daily; from its very launch in 2006, it has been flooded with active users (Park et al., 2016). It is a popular microblogging website and a top pace player in social media space (Yang et al., 2015). Its unique way of allowing following other users, limited length (Bao et al., 2017), retweets (Philander and Zhong, 2016) has made it popular with official accounts of celebrities, governments and companies and also of general masses at large so as to express what they feel.

Twitter is often known for its popularity in terms of huge user base, high levels of engagement and thus a favourite amongst the tourism and hospitality industry. It allows the
industry and its players to promote, distribute, market and communicate (Leung et al., 2013). This has also been the reason why researchers have found its utility in decoding the public sentiment, and the fact that it is available at absolutely no cost makes it more attractive (Jiang and Erdem, 2017); thus, it serves as a great market research tool (Leung et al., 2013). Also, National Destination Marketing Organizations (DMOs) have started capitalising on the Twitter’s wide reach for destination promotions, as it is found to be more frequently used than Facebook (Hays et al., 2013). This could be accounted to Twitter’s liberal settings in comparison to Facebook as it has a fundamental “open network structure” in terms of following other people (Weng et al., 2010).

Sentiment analysis in tourism
Sentiment analysis, often known as opinion mining, uses information that is in textual form both subjective, such as feelings or opinions, and objective, such as facts and evidence, (Alaei et al., 2019; Kennedy, 2012) for understanding the underlying sentiment behind them. The basic premise of sentiment analysis is to analyse the polarity, i.e. positive or negative (Schuckert et al., 2015). Text mining facilitates the process of data extraction and conversion into useful data (Gupta et al., 2009). Text mining aims to analyse a textual document in order to extract data, transform it into information and make it useful for various types of decision-making (Gupta et al., 2009). This study makes use of a lexicon-based method that uses a dictionary-based approach of pre-coded words that are used to study “text semantic orientation” (Chiu et al., 2015; Thelwall et al., 2011).

For this study, we first propose a framework on how to use Twitter analytics using R programming language for tourism insights and then take an experiment of three globally acclaimed Indian heritage sites, i.e. the Taj Mahal, Red Fort and Golden temple for understanding real-time sentiments of these heritage sites.

Study area
Taj Mahal
Located in the city of Agra, Taj Mahal is the beautiful monument built by the Mughal emperor Shah Jahan for his wife Mumtaz Mahal using white marble the construction of which began in 1,632 and took more than 20,000 artisans for its completion. Spread across 42 acres, this monument includes a mosque and is situated between gardens. It was recognised as the UNESCO World Heritage Site in 1983 and also referred as the “Jewel of Muslim Art in India”. The Taj Mahal was recognised as a part of the “7 Wonders of the World” and later as the “New 7 Wonders of the World” (UNESCO http://whc.unesco.org/en/list/252) (see Plate 1).

Red Fort
The Red Fort is a historical monument situated in India’s capital, Delhi. It is a fort that was built by the emperor Shah Jahan in 1,639 while shifting his capital base from Agra to Delhi, the initiation of which started in 1526 AD. This fort served as the residence of Mughal emperors. It is named after its architectural style of red sandstone walls. It was also designated as a “UNESCO World Heritage Site” in 2007 (UNESCO https://whc.unesco.org/en/list/231rev). Every year the Red Fort witnesses a flag hoisting ceremony on India’s Independence Day 15th August (PTI, 2013) (see Plate 2).

Golden Temple
Also known as Harmandir Sahib, the Golden Temple is situated in the Amritsar city of Punjab, India meaning “abode of God” (Kerr, 2011). It is a spiritual site of Sikhism and its
foundation stone was laid in the year 1589. It is built around the man-made pool “sarovar”. It has more than 100,000 visitors daily who come to worship the holy shrine. It is known for its free meal community service in the form of a “langar”. It has been nominated as the UNESCO World Heritage Site with its application pending in the tentative list (UNESCO https://whc.unesco.org/en/tentativelists/1858/). It was conferred as the “most visited place of the world” by the World Book of Records (WBR, 2017) (see Plate 3).

Plate 1.
A glimpse of the Taj Mahal

Source(s): Author’s own

Plate 2.
A glimpse of the Red Fort

Source(s): Author’s own
Framework for sentiment analysis

Twitter authentication

Twitter provides an option to register for API which is an Application Programming Interface and allows to retrieve tweets using a developer account. Once the Twitter authentication is done through its server, the user is given a set of keys under the “keys and tokens tab”. These keys are consumer key, consumer secret, access token and access secret that are generated through the Twitter API and are used for mining data.

Access twitter data sets

After authentication with the Twitter authentication service and generation of token for the API, tweets are mined directly from Twitter using a search involving the names of the destinations chosen for the study. Only English tweets from January 2015 to January 2021 were collected. This is done using “searchTwitter()” function of R package as follows:

\[
\text{tweets} \leftarrow \text{searchTwitter(“heritage site name”, n = 5,000, lan = “en”, since = “2015-01-01”)}
\]

Collection of corpus

The corpus is collected using Twitter API authentication and “Twitter” package in R Studio. This method was repeated thrice for the three different heritage sites. A total of 15,000 tweets were randomly retrieved for this study. The language of the tweets chosen was English (en).

Text pre-processing and data cleaning

Before we perform analysis on data, text pre-processing and cleaning is a prerequisite due to the prevalent noise in the data. It becomes imperative to remove white spaces, punctuations and other signs (like “@”, “#” “;” “##”) and stop words (like “is”, “the,” etc.) included in the tweets as they provide hindrance in understanding the underlying sentiment of the tweet. For this purpose, “tm” and “gsub” functions are used to perform text mining and cleaning the data, respectively.

Technique used – sentiment analysis

To attach sentiments to the tweets, we perform sentiment analysis; this may be categorised as positive, negative or neutral. For this purpose, we make use of the “syuzhet,” which used the
NRC Emotion Lexicon for tweet classification. Under this library, there are eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two sentiments (negative and positive). We make use of the “get_nrc_sentiment” function to obtain scores of the same.

**Graphical representation using data visualisation**
For better understanding of the results of the sentiment analysis, we opt for a graphical representation using data visualisation. We create a sentiment histogram and word cloud. For this, we use the functions “word cloud” and “ggplots” and install “RcolorBrewer” for the same. For word cloud, we provide minimum frequency as 5 and maximum words as 200 (see Figure 1).

**Results**
Following are the results of the study.

**Taj Mahal**
The sentiment spread for the Taj Mahal is as follows: Anger (5.09%), anticipation (9.42%), disgust (2.9%), fear (7.24%), joy (11.9%), sadness (6.7%), surprise (6.8%), trust (13.51%), negative (11.1%) and positive (25.06%). The empirical results of sentient analysis performed on Twitter data extracted show that the users have a positive sentiment towards the famous Taj Mahal and much lessor negative sentiment (Please refer Figure 2). And can be said that the polarity is towards positive.

This can be accounted to the visitors experience at this 7 Wonders of the World and UNESCO Heritage site. In terms of emotion classification, it can be seen that maximum positive emotions, such as trust and joy, are visible indicating that visitors experience happiness while visiting the Taj Mahal and enjoy the experience with trust.
After the calculation of emotion scores shown in the graph above, there is a need to identify the topics within each emotion through content mining. This is shown in word cloud based on classification of word depicting emotions. These are in the form of TF-IDF (term frequency vis-à-vis documents) and showing words as per their frequencies as highlighted by different colours for the tweets of the heritage site Taj Mahal (Please refer Figure 3).
**Red Fort**

The sentiment spread for the Red Fort is as follows: Anger (10.08%), anticipation (5.25%), disgust (3.63%), fear (13.29%), joy (3.79%), sadness (6.74%), surprise (3.98%), trust (21.87%), negative (17.69%) and positive (13.62%). The empirical results of sentiment analysis performed on Twitter data extracted show that the users have a mixed sentiment towards the Red Fort. This tilt towards the negative sentiment can be accounted to various civil agitations and protests that are held in and around the area by civilians and also due to sarcasm prevalent in the tweets.

In terms of emotion classification, it can be seen that maximum positive emotions such as trust is visible indicating that visitors experience happiness while visiting this UNESCO World Heritage Site (Please refer Figure 4). While fear and anger represent the agitation sentiments that are prevalent in and around the site (old Delhi).

After the calculation of emotion scores shown in the graph above, there is a need to identify the topics within each emotion through content mining. This is shown in word cloud based on classification of word depicting emotions. These are in the form of TF-IDF and showing words as per their frequencies as highlighted by different colours for the tweets of the heritage site Red Fort (Please refer Figure 5).

**Golden Temple**

The sentiment spread for the Golden Temple is as follows: Anger (5.14%), anticipation (10.23%), disgust (2.93%), fear (6.81%), joy (12.08%), sadness (6.30%), surprise (6.56%), trust (13.63%), negative (10.94%) and positive (25.3%). The empirical results of sentiment analysis performed on Twitter data extracted shows that the users have a positive sentiment towards the famous Golden Temple and much lesser negative sentiment. And can be said that the polarity is towards positive.

This can be accounted to the visitors experience at this significant Sikh shrine. In terms of emotion classification, it can be seen that maximum positive emotions such as trust and joy are visible indicating that visitors experience happiness while visiting the Golden Temple and enjoy the experience with trust (Please refer Figure 6).

![Figure 4. Emotion classification using sentiment scores of Red Fort tweets](image-url)
After the calculation of emotion scores shown in the graph above, there is a need to identify the topics within each emotion through content mining. This is shown in word cloud based on classification of word depicting emotions. These are in the form of TF-IDF and showing words as per their frequencies as highlighted by different colours for the tweets of the heritage site Golden Temple (Please refer Figure 7).

Public sentiments on Twitter

Figure 5. Word cloud of Red Fort tweets

Figure 6. Emotion classification using sentiment scores of Golden Temple tweets
Conclusion
As the number of tweets after data processing is unequal, the average sentiment score is taken for comparison and understanding the overall sentiment about India’s heritage sites. Amongst the three tourists’ spots, the greatest number of positive sentiments is for the Taj Mahal and Golden temple with approximately 25% each. While the most negative sentiment can be seen for the Red Fort (17%). Amongst the positive emotions, the maximum joy sentiment (12%) can be seen in the Golden Temple and trust (21%) in the Red Fort. In terms of negative emotions, fear (13%) can be seen in the Red fort. A summary can be seen in Figure 8.

Overall, India’s heritage sites have a positive sentiment (20%), which surpasses the negative sentiment (13%). And can be said that the overall polarity is towards positive. It can also be seen that India’s heritage sites are considered trustworthy as the trust emotion prominently visible (16%). Table 1 shows the sentiment and emotion scores and their averages.
## Table 1.

Summary of sentiment scores of the three heritage sites.

<table>
<thead>
<tr>
<th>Sentiment emotion</th>
<th>Taj Mahal</th>
<th>Red Fort</th>
<th>Golden Temple</th>
<th>Overall sentiment</th>
<th>Avg. overall sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>187</td>
<td>483</td>
<td>203</td>
<td>873</td>
<td>291</td>
</tr>
<tr>
<td>Anticipation</td>
<td>346</td>
<td>252</td>
<td>404</td>
<td>1,002</td>
<td>334</td>
</tr>
<tr>
<td>Disgust</td>
<td>110</td>
<td>174</td>
<td>116</td>
<td>400</td>
<td>1,333333</td>
</tr>
<tr>
<td>Fear</td>
<td>266</td>
<td>637</td>
<td>269</td>
<td>1,172</td>
<td>3906667</td>
</tr>
<tr>
<td>Joy</td>
<td>437</td>
<td>182</td>
<td>477</td>
<td>1,096</td>
<td>3653333</td>
</tr>
<tr>
<td>Sadness</td>
<td>248</td>
<td>323</td>
<td>419</td>
<td>820</td>
<td>2733333</td>
</tr>
<tr>
<td>Surprise</td>
<td>253</td>
<td>191</td>
<td>259</td>
<td>703</td>
<td>2343333</td>
</tr>
<tr>
<td>Trust</td>
<td>496</td>
<td>1,048</td>
<td>538</td>
<td>2,082</td>
<td>694</td>
</tr>
<tr>
<td>Negative</td>
<td>408</td>
<td>848</td>
<td>432</td>
<td>1,688</td>
<td>5626667</td>
</tr>
<tr>
<td>Positive</td>
<td>920</td>
<td>653</td>
<td>1,000</td>
<td>2,573</td>
<td>8576667</td>
</tr>
<tr>
<td>Total</td>
<td>3,671</td>
<td>4,791</td>
<td>3,947</td>
<td>12,409</td>
<td>4136333</td>
</tr>
</tbody>
</table>

Public sentiments on Twitter.
Implications, limitation and scope for future research

Social media has become an “extensive repository” of the sentiments of visitors that are unforced (Thelwall et al., 2011). These aid the marketers to understand their real sentiments and experiences and thus help in providing the insights. Hence, big data in the form of data mining, web crawling and natural language processing is now being used (Taecharungroj and Mathayomchan, 2019; Xiang et al., 2017). Understanding visitor experiences in the form of sentiment analysis helps tourism players to understand the traveller’s satisfaction that in turn affects their purchase decisions (Ye et al., 2011). The newer age techniques serve as more beneficial over traditional methods of market research for its “real-time” nature that avoids “recall biases” (Rylander et al., 1995).

While some studies have attempted to analyse public sentiments using Twitter, such as Saini et al. (2019), Rathore and Ilavarasan (2020) and Sreeja et al. (2019), tourism industry still remains as an underexplored area. This study not just adds to the existing literature on tourism and big data and sentiment analysis but also provides real-time insights to DMOs for marketing tours and trips for India. It also provides empirical evidence of the visitor sentiment about these famous Indian heritage sites. It also encourages tourism players to understand the negative emotions depicted in the word cloud and attempt to improve the visitor experience by dealing with the concerns as pointed out by the sentiment scores and the word cloud. Specifically, security can be enhanced at the Red Fort to ensure the sense of fear, which is the maximum depicted emotion declines.

While this study provides an effective way of understanding the public sentiment about the three heritage sites, it is limited only to Twitter data. The study also suffers from the technical limitations of R programming language and Twitter, such as inability of understanding sarcasm and the existence of retweets and certain ambiguous characters in the tweets that are attempted to be removed in data processing, but since the whole process is on machine learning 100% accuracy cannot be guaranteed. This is also the reason why the actual sentiments retrieved are lessor (12,409) as best attempts were made to filter out noise during processing. The data are collected for the last six years and can be taken for prior to that as well. While this does not hamper the present results, generalisation can be better if the period can be extended. Also, the study area is limited to three globally acclaimed heritage sites of India, other heritage sites from other countries can also be taken in future. To monitor the results by the author at each step, only tweets in English language were taken, while several other languages are also used in tweets.

References
India Tourism Statistics (2019), Ministry of Tourism, Government of India.


Further reading


About the author
Shruti Gulati is a Ph.D Scholar at Department of Commerce, Delhi School of Economics University of Delhi. She is also working as an Assistant Professor at Shaheed Bhagat Singh College, University of Delhi. She was awarded Lady Shri Ram College Colors Awards in her graduation and Central Student’s Scholarship in her post-graduation. She has presented papers at various National and International Conferences and Seminars, some of which have also been recognised as “Best Paper”. She has published several papers in journals of repute. Her areas of interest include Consumer Behaviour, Hospitality and Tourism Management, Social Media and Marketing. Shruti Gulati can be contacted at: gulati_shruti@yahoo.co.in

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com