Tourism analytics: social media analytics framework for promoting Asian tourist destinations using big data approach

Sahil Raj* and Tanveer Kajla

School of Management Studies,
Punjabi University,
Patiala, 147002, India
Email: dr.sahilraj47@gmail.com
Email: er.tanveer47@gmail.com
*Corresponding author

Abstract: This study is committed to social media analytics (SMA) which is the one of the most talked about research areas. The research contributes to building up a big data analytics (BDA) framework for carrying out SMA on Twitter data by using R and Hadoop. The main objective of this research is to check the popularity of tourism places on social media, and performing various text analytics methods to gain some insights out of the data. The results of the study show the significance of social web in getting an in-depth understanding of opinions of Twitter handlers. Visualisations in the research should assist in getting a quick analysis of the situation and aids in destination promotion. The research paper also helps in disaster management and building up the strategies to broadcast the tweets to the wider audience by discovering the influential persons on Twitter.

Keywords: tourism; analytics; social media; big data; Hive; word cloud; flume; sentiment analysis; Twitter.


Biographical notes: Sahil Raj is a faculty member at School of Management Studies, Punjabi University, Patiala, India. His main area of interest is the application of information systems in business organisations and big data analytics. He has authored four books which include Management Information Systems and Business Analytics with leading global publishers like Pearson Publication and Cengage Publication.

Tanveer Kajla is a Research Scholar in School of Management Studies, Punjabi University, Patiala, India. He has done his BTech in Computer Sciences and MBA in Information Technology. His current area of research is big data analytics, text analytics, unstructured data analytics in the tourism industry.

This paper is a revised and expanded version of a paper entitled ‘User generated content analysis of tourism industry of Indonesia’ presented at AGBA 13th Annual World Congress, Indonesia, 26–27 November, 2016.
1 Introduction

Tourism has evolved as the foremost measure in terms of economic growth. After fuels, chemicals and food, tourism has become the largest export industry (Law et al., 2010) holding 6% of the world’s total goods and services.

The definition given by UNWTO describes tourism as a phenomenon restricting the people to move to the places outside their location for professional or personal purposes (UNWTO, 2017). Tourism is considered as the transfer of individuals to some other place for income and technology. The visitors are the people who visit the recreational places. When the visitor is visiting the place for some business purpose, then that visitor is known as a business visitor. With the coming of tourists to a particular place, there is a lot of awareness among the tourists about the culture of the residents of that place. It improves the quality of living of the public as it generates a lot of revenue and through this revenue, the infrastructure of a place is also improved. Even the facilities are developed which can be utilised by local residents as well.

Asia-Pacific region contributed to 23% of global arrival and 30% of the total revenues where growth rates posted by India and Indonesia are 11% and 7% (Runclle & Associates, Inc., 2015). According to the article, “Asia-global tourism’s driving force” published in Forbes, the prediction given by WTTC on Asia’s tourism sector’s growth is more than 6% every year, which is among the quickest in the world (Fuller, 2013). Linking with economic growth, this positive change in statistics in Asia-Pacific region also increased per capita income of Indonesia (Kim and Park, 2015). The rising number in terms of tourism arrivals reflects a great success of tourism industry in Indonesia. But the tourist arrivals are higher in neighbouring countries like Malaysia, Singapore and Thailand. The study conducted by Law et al. (2010), evaluated the low performance of tourism destination websites, and the majority of websites need improvement in terms of information and technical support.

Bali is the most popular tourist destination among the destinations of Indonesia (Tajeddini et al., 2017). The population of Bali is around 30 lakhs, having Hinduism as its major religion. It is famous for its temples and beaches among tourists. Goa is a popular tourist destination in India loaded with beaches. Goa contributes a lot to the tourism industry of India and is the major service industry in the country.

Social media has effectively reformed the tourism and hospitality industry (Leung et al., 2013). It influences the social and economic lives of users (Zeng and Gerritsen, 2014), specifically in decision making and searching for content on the web (Power and Phillips-Wren, 2011). The influence of social media on tourism industry has been recognised as an upcoming area of research (Nagle and Pope, 2013). Social media allows users to share their pre-travel and post-travel experiences (Zeng and Gerritsen, 2014). Since the evolution of social media from broadcasting medium to a participatory platform, internet users became the means of sharing and collaborating information on the internet (Leung et al., 2013). The wide adoption of social media by travellers is primarily to search, share and elucidate their journey experiences by using blogs and microblogs like Twitter and online communities like Facebook in a collective manner (Leung et al., 2013).

Big data is considered as very gigantic and complicated data sets that are impossible to process and store in the traditional data warehouse systems (Jukic et al., 2015). There are very few and recent publications on social networks in the tourism sector, which shows that very little work has been done in this area (Leung et al., 2013; Zeng and...
Gerritsen, 2014). This opens new avenues of research in the field of social media analytics (SMA). SMA is the development and evaluation of data science tools. It is a framework to extract the data, process and analyse data to discover some hidden pattern and gain some intelligence by visualising those hidden patterns (Thelwall, 2017).

To contribute more to the tourism sector, the research needs to be done on analysing the data generated from the very popular microblogging site, Twitter. Twitter users posted millions of tweets every day (Velde et al., 2015). Since its inception in 2006 into the world of social media, Twitter has added millions of active users (Park et al., 2016). It is one of the popular microblogs available on social media (Yang et al., 2015). Twitter users can post messages in the form of tweets. Length of tweets is less than 140 characters (Bao et al., 2017), which is close to 11 words per tweet. Users have the privilege to follow other users on Twitter. A retweet is a tweet, resent by some other user (Philander and Zhong, 2016). Hashtags (#, suffix with word), are used to start a discussion on any topic or event, and it tends to group the tweets by topic. @(suffix with name) mentions the name of the user in the tweets. The popularity of Twitter among researchers, developers and practitioners is due to its rich content available at zero cost (Jiang and Erdem, 2017). Twitter provides two different API’s to collect the data. REST API is used to collect the historical data, whereas the streaming API plays a significant role in getting real-time data (Durahim and Coskun, 2015).

This paper prepares a research model to capture, analyse, store and visualise Twitter data to find the hidden patterns. This framework will also find the popular tourist places on the social network by calculating the web traffic generated. Various opinion mining techniques based on machine learning algorithms will be used for classification of social media data. Word cloud is created for the topic analysis of social data. This study also contributes towards research by building a big data framework to collect the Geo-tagged Twitter data and analyse the data from different perspectives.

2 Review of literature

2.1 Tourism vis-à-vis ICT

ICT has transformed the organisational structures, practices, and strategies in the tourism sector (Makkonen and Hokkanen, 2013). Tourism industry changed rapidly with the emergence of ICT, and particularly the internet (Pena et al., 2013). Embedment of ICT into tourism is also known as e-tourism. E-tourism leads to process digitisation in tourism and hospitality sector (Zhong et al., 2013).

The rapid upsurge in the number of internet users and online transactions signifies the popularity of technology (Law et al., 2010). The number of internet users using desktops has declined, whereas tablet computer users have risen significantly (Xiang et al., 2015), which shows users having portable devices are increasing in number. According to the stats given by Santos (2011) on social media,

- 600 million search queries were submitted to the search engine of Twitter
- 40% of the tourism sector consider social media as a big thing in the coming five years
Tourism analytics: social media analytics framework

- recommendations given on social media were trusted by 70% of the consumers in the tourism industry
- in the next five years, new technologies will be used by the consumers of tourism and hospitality industry to book the holiday packages.

25% of the people got access to the internet. There was an 18% rise in users using social media. The quarter percentage of the total population of the globe is using social media, where Asia showed a significant 76% increase in internet surfing in 2013 (Kiralova, 2014).

This shows the importance of social media and information technology in the tourism sector. The internet can be used as an effective tool for doing promotions (Law et al., 2010). ICT played an important role in mitigating the risks by providing information from all the corners to the travellers by connecting them to the internet (Xiang et al., 2015). The traditional sources of searching information declined, whereas there is a rise in non-conventional search techniques used by travellers like online tourism organisations and search engines for planning their trip (Xiang et al., 2015). Social media is used expansively for sharing post-trip experiences. A lot of changes coming into technology and less budgets are forcing tourism organisations to adopt social media as a communication strategy (Huang, 2012).

Many researchers wanted to embed benefits of the web into the tourism industry (Law et al., 2010). Successful ICT implementation gives various benefits to an organisation (Lee and Hong, 2017). With the implementation of ICT, organisations can connect with their customers 24x7 and the speed of communication increases by embedding ICT in their organisation (Limayem and Cheung, 2011). The use of IT also gives lots of benefits to the internet users. The consumers in tourism industry get the luxury of sitting at home while shopping for their products or services at any time. Even the consumers can exchange their trip experiences with others (Kim et al., 2015).

Compared to conventional methods, ICT made the booking system convenient and efficient, which in return increased the convenience of the travellers and made them satisfied with their services (Firchow, 2013). This shows that the tourism industry has experienced a paradigm shift by using ICT and internet (Makkonen and Hokkanen, 2013).

2.2 Big data and Hadoop

There are 294 billion emails sent every day. People on the internet do more than one billion Google searches every day. 30+ petabytes content generated by web users is stored, retrieved and analysed. Twitter users posted more than 500 million tweets each day. According to Forbes, large volumes of data have been created in the last two years as compared to all the historical data. There are more than one billion users of Facebook in a single day viewing 2.77 million videos in a minute. YouTube got 300 hours of videos uploaded every minute (Marr, 2015).

The above facts describe the continuous growth of data. Social media are generating a variety of data, such as photos, posts, videos (Miah et al., 2017). This continuous generation of data by multiple users brought up a new concept known as big data. (Storey and Song, 2017) describes five V’s to define Big Data. These are Volume (a large amount of data), Variety (structured and unstructured data), Velocity (the rate at which data has been generated), Veracity (quality of data), Value (usefulness of data) (Jukic et al., 2015).
Big data will only provide the value when it is unlocked for decision making (Wang et al., 2016a). The biggest challenge of applications dealing with it is to get knowledge out of this tremendous amount of data (Silva et al., 2015). The techniques to get value out of these big data sets are known as big data analytics (BDA). BDA involves various activities such as capturing data, storing the captured data, and then analyse the data for decision-making (Miah et al., 2017). The challenges in BDA are to store and process these massive data sets. The traditional systems are not capable enough to handle big data. Apache Hadoop is the solution to all the hindrances created by big data. Hadoop is used for collection of structured data and unstructured data. Doug Cutting and Mike Cafarella created Hadoop in 2005. Then this project came under Apache Foundation and it became open sourced. The big companies like Google, Yahoo, Facebook and Amazon are extensively using Hadoop for performing data-intensive tasks (White, 2012). These companies generate and process terabytes and petabytes of data every day to get some insights out of these massive datasets.

Hadoop is a programming framework based on Java which runs applications for processing big datasets on commodity machines (White, 2012). Hadoop works on two principles, distributed storage and parallel processing. Distributed storage is done with Hadoop Distributed File System (HDFS) (Ghemawat et al., 2003), and parallel execution of data is with MapReduce (Dean and Ghemawat, 2008). HDFS and MapReduce are the two components of Hadoop. MapReduce is responsible for the parallel execution of data, whereas HDFS plays a big role in storing the data along various nodes/computers. HDFS is derived from the Google file system (Ghemawat et al., 2003). Google file system is used to support the fast-growing data at Google. The traditional distributed file systems are performing the same task of storing the data on different nodes. Google file system is handling the very large files up to terabytes in an efficient manner. There can be lots of errors while handling the large files. Therefore, monitoring the datasets for errors and faults and recovering from these fatal accidents is embedded into the systems (Ghemawat et al., 2003). The component used in this research to store the data is HDFS. It is used to store the data efficiently on commodity hardware with highly fault tolerant behaviour. HDFS accepts all formats of data like images, unstructured text, audio files, videos, etc.

Hive accumulate the data over HDFS. It is a data warehouse layer supporting the HiveQL for data stored on HDFS. It is developed by Facebook in 2007 to store and analyse the Big Data with Hive Query Language. It got open sourced in 2008. HiveQL compiles the SQL like queries into MapReduce programs, so that non-programmers can also run the MapReduce programs. Facebook used Apache Hadoop and Hive platform to implement its application, Facebook messages.

Flume is a component of Hadoop developed by Cloudera. It is used to store streaming data in HDFS. Hortonworks defined flume as a reliable service to collect and store large datasets in HDFS (Apache Flume, n.d.). Flume is very popular in social media and academics to collect the data from Twitter and save it to HDFS.

### 2.3 Social media and Twitter

In the era of Web 1.0, the unidirectional approach was followed to put the content on the web (Kaplan and Haenlein, 2010). There were static web pages where no interactive content was found. There were very strict norms to load the content on the web (Power and Phillips-Wren, 2011). Moreover, only the patented content was loaded
on a web page. At the time of publishing the material on the web, there was no direct interaction between the user who is viewing the content and the publisher of the content.

Unidirectional uploading of content on the web in a restricted way belongs to Web 1.0 (Kaplan and Haenlein, 2010). There is no hesitation in describing Web 2.0 as a platform for the evolution of social media. Since 2004, the social web is described with the term Web 2.0. Social networking sites embraced important place in the online activities (Power and Phillips-Wren, 2011).

There is a lot of confusion among researchers and academicians about social media. There must be some clarification on how social media is dissimilar from Web 2.0 and UGC (Kaplan and Haenlein, 2010). The survey did by Heidi Cohen found different definitions of social media. One respondent even specified that he is not sure what is social media (Cohen, 2011). (Kaplan and Haenlein, 2010) identified social media as an assemblage of applications grounded on internet and Web 2.0, that permits the foundation and interchange of UGC.

Social media do not mean only social networking sites. Apart from social networking sites, it also includes various content sharing media like YouTube (video sharing sites), LinkedIn (professional networking sites), Wikipedia (collaborative content sharing sites). Social media delivers interactive platforms to create, modify, share the user-generated content (UGC) (Kaplan and Haenlein, 2010).

UGC is the content posted by web users. The growth in internet and mobiles, enables users to instantly add UGC like giving reviews on services like airlines, hotels, restaurant, sharing travel experiences. Various social media platforms contain enormous data, appropriate for researchers in tourism and hospitality sector (Lu and Stepchenkova, 2015). UGC is sometimes called as consumer-generated media by some businesses and academicians.

The UGC available on social is additional information used by travellers for the information about the destination. Social media has a powerful presence in travel-related search (Xiang et al., 2015). With social media, customer retention can be done relationship building in virtual communities. The customer got attracted and engaged with user-generated content generated by members of these travel-related online communities (Daugherty et al., 2008). It is also noticed that related review of literature is also published recently which clearly indicates Social media in tourism is one of the emerging topics for researchers.

The beauty of social media is that the users having no IT background can also publish any content on social media. The internet users with no technical skills can put any content on social media (Kang and Schuett, 2013). The growth of social media is due to the presence of Web 2.0. The travellers can share their trip experiences easily on social media. Various business houses used social media for communication purpose (Hays et al., 2013). User-generated content is the priceless source of information for them. People show trust in information given by tourism websites maintained by the state government. Travellers consider the user generated content as less biased, credible, and trustworthy as compared to any communication channel of an organisation (Munar and Jacobsen, 2013). People can’t try services of any hotel or restaurant so they rely on information coming from user-generated content created on travel blogs or websites. Social media provides a platform where users post UGC in the form of reviews, advice, and personal recommendation (Tham et al., 2013). These type of user-generated content are in demand for reliable and unbiased information (Kang and Schuett, 2013), they influence travellers in making pre-trip and during trip decisions (Yeongbae et al., 2017).
Twitter and Facebook were initially used for social interactions. But with time these social networking tools were used for monitoring earthquakes, epidemics and elections (Bao et al., 2017). Among the various social networking tools, Twitter is a very popular microblog used by organisations and users (Philander and Zhong, 2016). It became the influential platform on the web in terms of microblogs (Thelwall, 2017). For providing authentic information, blogs were considered as a popular social media platform (Huang, 2012). Researchers and academicians used Twitter in various ways for their researches. Twitter is used for its Geo-location data and also for doing sentiment analysis on the text gathered from the tweets. Twitter provides an API that gives access to Twitter data according to the search criteria. API also gives the location of the tweets is the Twitter user wants to reveal this information (Durahim and Coskun, 2015).

2.4 Social media analytics

As already discussed Web 2.0 enabled internet users to post their opinions on various social media platforms (Power and Phillips-Wren, 2011). The web has become the repository which is holding the large datasets of user-generated content where new information gets added to this data each day (Power and Phillips-Wren, 2011). They generate lots of textual data which is understandable to humans but machines are not capable enough to get some insights out of that data (Hu and Liu, 2012). Automatically getting the opinions/content from the multiple sources on the web is getting popularity as it is impossible to manually extract the massive content on the web (Johnson et al., 2012). The task of opinion mining and sentiment analysis is to use various algorithms for extracting opinions from the user-generated content by using data mining techniques (Hu and Liu, 2012). Once the algorithms for opinion mining and sentiment analysis are finalised, it takes very less endeavour to get the results from massive data (Pang and Lee, 2008).

SMA is used to develop a framework where data is collected, analysed and then presented in visual form (Lawrence et al., 2010). The rapid expansion of social media data gives computing challenge (Fan and Gordon, 2014). To overcome the computing challenge, SMA process is used to get the meaning out of the data (Batrinca and Treleaven, 2015).

The SMA methodology or process is divided into three stages. It involves capturing of data, understanding the data, and then presenting that data in the visual form (Fan and Gordon, 2014). The data from social media is captured from various platforms with API's or web crawlers. The data collected from different platforms and can be integrated into a single dataset if there is a requirement of inclusive datasets. Capturing the data sometimes also involves pre-processing of data. It involves POS tagging, stemming and feature extraction.

When the data has been captured, the next step in SMA process is the understanding stage. It is the central theme of the analytics process. This stage will remove the noisy data. The analysis technique to get some insights from the data is performed at this stage. Analysis can be done on historical data as well as on real-time data (Batrinca and Treleaven, 2015).

The present stage is the third stage in the process. This is the visualisation stage where the results from previous stages are presented in the visual form by using customised dashboards (Fan and Gordon, 2014).
The most fundamental technique while doing monitoring and analysis is opinion mining or sentiment analysis (Fan and Gordon, 2014). The term opinion mining is the method of extracting the sentiments and emotions from the text (Severyn et al., 2016). In technical terms, opinion mining is also known as sentiment analysis, which does the classification of text into different emotions or feelings (Pang and Lee, 2008). Opinion mining is an upcoming topic in the previous years as the internet users give shape to their opinions by looking at the opinions of the others (He et al., 2015). There are various applications of opinion mining in areas of finance, research and development, recommendation systems. Opinion mining is used to make the recommendation of websites, spam detection, brand management and the impact (assessment) of product launch (Severyn et al., 2016). Managers make sound decisions from predictions made with the assistance of opinion mining (Yang et al., 2015). This technique is also used for making stock market predictions, trends in the market, finding defects in products and crisis management (Fan and Gordon, 2014).

3 Big data analytics framework

The BDA framework will be created for doing the analysis of social media data in the tourism industry. Hadoop as the standalone mode is installed on the Ubuntu machine. R, flume and hive were also installed on this machine. Integration of R and HDFS (storage component of Hadoop) is done with ‘rhdfs’ package that provides connectivity to R and Hadoop.

This framework is used in the tourism industry to handle the big data which is mostly generated by social media. R, Hive, Flume and Hadoop are used to create a framework to store and analyse the unstructured data generated from Twitter. The framework extracts the data from Twitter, and then dump all the data in a single storage space. The next step in the framework is used for cleansing and pre-processing of data. This is done separately in Hive and R. After that in the final stage the data is analysed with Hive and R to get insights out of the data (Figure 1).

Figure 1  Big data analytics framework (see online version for colours)
3.1 Data extraction

In this research, the data is acquired from the Twitter in the form of tweets. Tweets will be collected with the assistance of Twitter API. Twitter provides assistance to the researchers and analyst by giving access to the data in the form of Tweets. Two types of API’s are provided by Twitter to get the historical tweets and real-time tweets. REST API provides historical tweets whereas Streaming API provides real-time tweets posted by the users in that time frame (Durahim and Coskun, 2015). When connected with Twitter, real-time tweets will be collected by using streaming API. They are collected from Twitter through the authentication done with the Twitter application (API) (Figure 2).

Figure 2 Data extraction with flume (see online version for colours)

3.2 Data storage

All the tweets will be stored in the HDFS. HDFS can store all types of structured and unstructured data. Data collected in the form of tweets is unstructured data and is in JSON format. This data cannot be stored in tabular form. There needs to be a storage space which is fault tolerant and can store unstructured as well as structured data.

3.3 Data processing

Data processing will be done with Hive and R. Collected data is stored in a tabular structure in the databases with the help of hive. To do the pre-processing in R, the text of the tweets is fetched with Hive and then loaded into R in object form. Data cleansing techniques like stemming, removing stop words, punctuations will be performed with R so that the relevant text is used in the analysis.

3.4 Data analysis

Network analysis of tweets will be done by firing SQL like queries in Hive. Further, sentiment analysis, emotion classification and word cloud will be done to get the perception of Twitter users towards tourism places.

4 Research methodology

4.1 Data acquisition

4.1.1 Sample

The tweets will be collected for two popular tourist destinations of Asia. The two destinations are Goa (India) and Bali (Indonesia).
4.1.2 API

Twitter API acts as a backend server that facilitates the collection of data by fetching the tweets for the public at no cost (Philander and Zhong, 2016). For acquiring the data, a Twitter developer account is created. In the developer account, API will be made through which keys and tokens get generated. These keys and tokens will be used to authenticate with the Twitter to extract the data for the research. Twitter APIs are classified into REST API and streaming API. REST API uses ‘pull’ strategy, whereas Streaming API uses ‘push’ technique, in which continuous data will be retrieved for a particular time frame. In the research, streaming API is used to get the real-time tweets from Twitter (Durahim and Coskun, 2015).

4.1.3 Flume

Apache Flume is an open source product used in the Hadoop framework for real-time processing of data. Apache Flume, by default uses Streaming API to collect the tweets in real time. Flume stores streaming data into HDFS. Hortonworks defined flume as a reliable service to collect and store large datasets in HDFS (Apache Flume, n.d.). Flume is very popular in social media and academics to collect the data from Twitter and save it to HDFS.

4.2 Data storage

All the data storage takes place at HDFS which is the central repository in this framework. Tweets fetched by Flume are totally unstructured data. This data needs to be in structured form for cleansing and analysis. Hive is used to make it in a structured form by storing that in tabular form on its data warehouse layer. For further analysis in R, the text will be extracted from this data and stored in HDFS. The text (only tweets) will be extracted by R with ‘rdfs’ package (Figure 3).

Figure 3  R and Hive connectivity with HDFS

4.3 Data analysis

The popularity of tourism places will be found from the web traffic generated for these sites on Twitter. The places that get a higher number of tweets will be considered as
popular tourism place as compared to other tourism places included in the study. Further, the research will check the popular time zones from where the web traffic has been generated for these tourism places by using the geo-located Twitter data.

Network analysis of collected data will be done by the hive. The influential users will be found from the data by getting the Twitter users having the most number of friends count, friends count and retweet count. Popular hashtags used in the tweets by Twitter users were found with Hive to get the trending topics on the social network (Twitter).

R language is used in this research to get the perceptions of Twitter users towards the selected tourism places. It is checked by performing the sentiment analysis and emotion classification on the text. Sentiment analysis is used to categorise the text into positive, negative and neutral text. Emotion classification technique classifies the tweets into various emotions like joy, anger, sadness, surprise, fear and trust. The emotions of the tweets not falling into any of the categories come under unknown emotion. Advanced visualisations are made from the tweets by forming a word cloud from the collected data. Word cloud gives the most prompting words in the datasets in the middle of the cloud, to get the clear picture of the popular words and their significance in the tweets.

The analysis of the social media data generated by Twitter is visualised with graphs plotted in R with ggplot2 package. ggplot2 package for R is used for visualisation of data.

5 Results

5.1 Popularity of tourism places

The total number of tweets posted for Bali was 10,961 whereas Goa got 6882 tweets on Twitter. This shows that Bali is more popular than Goa (Figure 4).

Figure 4 Popularity of tourism places (see online version for colours)
5.2 Goa

5.2.1 Word cloud

Figure 5 shows the word cloud for Goa. Word cloud shows the most occurring words in the middle of the cloud. The word cloud of Goa shows some interesting discussions about Goa on Twitter. The prompting words in the word cloud are ‘Konkani’, ‘music’, ‘app’, ‘beach’.

![Figure 5 Word cloud of Goa](see online version for colours)

5.2.2 Sentiment analysis

Sentiment analysis shows very positive results about Goa as a lot of tweets fall into the positive category. However, there are some negative tweets also, but they are less in number (Figure 6).

![Figure 6 Sentiment analysis of Goa](see online version for colours)
5.2.3 Emotion classification

Figure 7 gives the emotion classification of tweets. There are lots of tweets in the unknown category. But still, other emotions give the clarity about the moods of the Twitter users. The results show users are in a joyous mood and other emotions such as ‘sadness’, ‘surprise’, ‘fear’, ‘anger’ and ‘disgust’ are less in proportion as compared to ‘joy’ category.

Figure 7 Emotion classification of Goa (see online version for colours)

5.2.4 Popular hashtags

Hashtags define the discussion topics on Twitter. Figure 8 defines the popular hashtags in descending order. The popular hashtags posted in the tweets are #goa, #golmaalagain, #konkani, #music, #incredibleindia, #nowplaying, #radiogoa.

Figure 8 Popular hashtags (see online version for colours)
5.2.5 Friends count

Popular users according to the Friends Count are ‘hashmumbai’, ‘BARATSQ’, ‘SureshM46’, ‘laxman_xy’, ‘averageUSgirl’, ‘KristineMHouser’. These users can be used to broadcast tweets if some relief management operations need to be carried out (Figure 9).

Figure 9  Popular twitter users according to friend’s count (see online version for colours)

5.2.6 Followers count

Figure 10 given below shows the popular twitter users according to their follower’s count.

Figure 10  Popular Twitter users according to follower’s count (see online version for colours)

5.2.7 Time zones

In Figure 11, ‘New Delhi’ time zone is the most popular one, whereas least popular time zone is ‘Central Time (USA and Canada)’. The other popular time zones shown in Figure 11 are ‘Pacific Time (USA and Canada)’, ‘Chennai’, ‘Mumbai’, ‘Jakarta’ and ‘Amsterdam’.
5.2.8 Retweets

According to the tweets posted, there are users whose tweets are retweeted the most. The most popular Twitter user shown in Figure 12 is ‘GurmeetRamRahim’, and the least popular user is ‘AdityaRajKaul’. The other popular users are ‘insan_honey’, ‘alluarjun’, ‘bhak_sala’, and ‘digvijaya_28’.

5.3 Bali

5.3.1 Word cloud

The total number of tweets collected from Twitter was 10,961. The words prompting in the word cloud of Bali are ‘happiness’, ‘pushawardskathniels’ and ‘amazing’, making the posted tweets fall into joy category in the emotion classification. These tweets put a positive impact on Bali’s image.

The tweets having these words are giving a positive impact to Bali in sentiment analysis (Figure 13).
5.3.2 Sentiment analysis

Bali shows a positive image when sentiment analysis is performed. The scenic beauty of Bali puts it on the positive side. But due to some issues like rabies, droughts and drugs, some negative tweets are also posted on Twitter (Figure 14).

Figure 14 Sentiment analysis of Bali (see online version for colours)
5.3.3 Emotion classification

Although a lot of tweets are falling into ‘unknown’ category, still the graph clearly suggests the joyous mood of the Twitter users towards the Bali Islands. There are push awards by ABS-CBN which are putting the native people and tourist of Bali in a joyous mood (Figure 15).

Figure 15  Emotion classification of Bali (see online version for colours)

5.3.4 Popular time zones

To do the further in-depth analysis, Apache Hive will be used to find the users and accounts having a number of followers and friends linked to their accounts. Popular time zones will also be found to find where the Bali is much talked about. Popularity will be checked by calculating the total web traffic generated by each time zone. From Figure 16, the Pacific Time (USA and Canada) time zone is very popular as a lot of web traffic is coming from these zones. Jakarta is at second place. It is also observed that Bali is also a talked about the destination in Beijing also. But the Eastern Time zone is giving very less attention to Bali as very less number of tweets was coming from Bali (refer Figure 16).

Figure 16  Tweets captured from popular time zones (see online version for colours)
5.3.5 Popular hashtags

The diagram below gives the list of popular hashtags from the tweets of Bali. Hashtags were found in Hive. Figure 17 shows that ‘bali’ is a very popular hashtag.

**Figure 17** Popular hashtags for Bali (see online version for colours)

5.3.6 Retweets

Popular retweets counts were also found by using Hive. Retweets are those tweets which are reposted. Figure 18 shows the Twitter users who are very popular, because when they put any tweet, then that tweet is retweeted most of the times.

**Figure 18** Popular Twitter users – retweets count (see online version for colours)
5.3.7 Followers’ count

After that, the users with a most number of followers will be found. Figure 19 shows the top 6 users having highest followers’ count. This confirms that MaliqMusic is very popular, according to Friends Count as they were having close to 37,000 followers for their account. The account ‘Balitoday’ is also getting attention here because a lot of Twitterati are following their account.

Figure 19 Popular Twitter users – followers’ count (see online version for colours)

*1e+05 = 10,000, 2e+05 = 20,000, 3e+05 = 30,000.

5.3.8 Friends’ count

Figure 20 will show the list of users who are having the most number of friends in their account. The list is shown in decreasing order. It shows that ‘Jasonsentell’ is the most popular user/account because it has maximum friends.

Figure 20 Popular Twitter users – friends’ count (see online version for colours)
6 Discussions

6.1 Goa

Konkani is the language of Goa. This word is popular on Twitter due to the Konkani Radio playing Konkani music and other regional events. Goa people are passionate about their language because Konkani is often used for any music album or song at some events or on radio channels. Goan music is popular on Twitter and most of the music is in their regional language. The word ‘beach’ is also coming at the centre of the cloud which shows the popularity of the beaches of Goa on social media. Varca beach is the most talked about beach in the tweets. This beach is known for its cleanliness and is among the top destinations among tourists. Vagator and Palolem beaches are also popular among tourists. Tourists and travellers coming to Goa prefer Goa’s mobile applications to book online cabs or for some information about Goa.

Most of the tweets posted by radio channels for playing Goan songs fall into the positive category which shows the love for music by the residents of Goa. Varca beach is also giving positive tweets on the Twitter because people adore the cleanliness of the beaches. Even the scenic beauty of Goa is praised on Twitter. Some positive tweets are also posted regarding the dolphins. The tweets regarding mining are negative in nature, as the users are complaining about the dust clouds formed by mining in Goa. Some negative tweets show that beaches are kept in bad shape, which shows disappointment on Twitter. Monkey fever is also a major concern on social media as one woman died in Goa due to this fever.

The Twitter users feel disgusted about the taxi mafia prevailing over Goa. The rise in mercury in Goa also made people feel disgusted about Goa. There are some incidents of rave parties in Goa which made people angry on social media. This kind of incidents hampers Goa’s image as a family tourist destination. There are some cases of monkey fever in Goa, according to social media. One woman died due to this fever. This disease is also known as Kyasanur Forest disease (KFD). People got feared about the side-effects of the vaccine. There was also fear among tourists as Gaurakshaks thrashed foreign tourist. Tweets posted in joy category considered Goa as paradise and most preferred place for spending holidays. Goa’s trance music puts users on Twitter in a joyous mood. Everybody loves the sunset on Candolim beach. Some Twitter users are surprised to know about Goa as one of the richest states in terms of GDP as discussed in social media.

6.2 Bali

The tourists are attracted by the beauty of Bali. The tweets collected when analysed, showed that the drug issue is prevailing in rural parts of Bali. Bali image is hampered because of the drug menace in Bali island. The government of Bali should handle these circumstances by launching anti-drug campaigns. Pictures of fake egg were posted on Twitter. Twitter handlers were disturbed by looking at these kinds of images. The government needs to counter these types of situations by raising its food standard to tackle these food issues. There are some fatalities because of rabies. The government or tourism board can put some vaccination camps for these dogs or put them at some common place so that they can’t harm the tourist or locals. There are also issues of drought in Bali. The tweets related to drought are negative in nature, and Bali should counter these concerns by putting up some tweets for awareness. Anti-drought policies
should be made by the government and put on Twitter, to assure the tourist and locals, that there will not be any problem regarding water issues.

The Pacific Time zone is generating more web traffic than Eastern Time zone. Customised tourism marketing can be done according to the time zones. If the Eastern Time zone is not so popular, then Bali can be promoted in this time zone to attract more tourists. New advertisement strategies can also be made for various time zones. Influential users having large followers base and friends’ base can be used to broadcast the messages in the time of emergency or for the promotion of any product or service. In the same way, any Twitter handler with the most number of retweets can also be considered for the same. Trending topics can also be found with the popular Hashtags also.

7 Managerial implications

This study contributes a lot to the field of analytics in the tourism industry. The research gave a BDA framework for the tourism industry, in which the data can be captured and analysed easily to get some knowledge or insights. Twitter is a popular marketing tool for the destination marketing organisations (Philander and Zhong, 2016). The framework helps the tourism organisations and tourism service providers to collect humongous data and use that data to understand the social media in a better way for building up new destination marketing strategies. When discussing the ‘Big Data’ concept, it is considered that more the data more is the understanding of the scenario. This research not only assists organisations in storing the data, but it also gives faster processing of data. The faster processing means quicker analysis, which aids the managers or any travel organisations to make strategies more rapidly by analysing the results. The framework created for the research is created with open source software. No additional cost is given for implementing this SMA framework. Any new organisation who wants to embed analytics in their organisation can find this framework very effective and cost-efficient. Even the government tourism boards can implement this framework to test the analytics on social media data and acquire the knowledge from the voluminous data.

The text analytics (opinion mining) techniques used in this framework analysis the perception of user not with the structured data, but with the unstructured data. The major benefit of tapping the unstructured data is to check the trustworthy and unbiased opinions given by the Twitter handlers. The methodology used in the study aids any organisation to capture this unstructured data even performing the analytics application on this data. The insights gained from this data can be used by the managers as feedbacks for their services or products. If anybody is talking negative about the services of the hotels, then the manager of the hotel can find why that person is not happy with their services by using analytics, and even improves the services by fetching their tweets. In a similar manner, they can analyse the strengths and weaknesses of their own organisations and even their competitors. The tourism industry used social media for promoting discounted products, resolving inquiries, and handling customer complaints (Philander and Zhong, 2016).

The second managerial implication of this study is giving is to capture and analyse the web traffic for tourism places and even comparing this with the past data or with some other destination to check the popularity of tourism places. The techniques like word cloud and popular hashtag finder give the trending topics to the manager. These trending
Tourism analytics: social media analytics framework

85

topics of discussions, when used by managers, can give more understanding of the situation.

Twitter was a very efficient channel for communication during the earthquake in Japan. This analytics framework can be used by government agencies in the times of emergency for disaster management, or relief operation by targeting the influential users to broadcast the messages. These influential users can also promote the tourism places or tour packages as they have wider coverage on social media.

8 Conclusion

The introduction of ICT and particularly social media made the tourism industry very dynamic. Social media changed the way how tourism destinations (Tan et al., 2016) are promoted. There is a need to analyse this ever growing data by building some SMA framework and devising various analytics techniques to capture and understand this humongous data. Tapping this social media data can provide lots of benefits to an organisation if that data is properly managed. Various researchers in their study elucidated the Twitter structure and concluded that it is widely used for well-timed updates and publishing colossal information (Philander and Zhong, 2016). Text analytics is still in its nascent stage, so the analysis requires a very strong understanding of captured data. The analytic techniques used to gain insights from data require expert knowledge, as this data is very noisy, and a lot of different meanings came from this data. The BDA framework should be adaptable to the future technologies also, as a lot of developments are going on in this area.

References


UNWTO (2017) UNWTO Tourism Highlights, UNWTO.


**Bibliography**


