SENTIMENT ANALYSIS OF BIG DATA IN TOURISM BY BUSINESS INTELLIGENCE

Diptiben Ghelani
Department of Computer Engineering, Gujrat Technological University, Ahmedabad, Gujrat, India

Abstract: There are several unsuccessful IT initiatives in today's market among specialized small and medium-sized businesses due to a lack of control over the gap between the business and its goal. In other words, purchased products are not being sold, which is a regular occurrence in tourism retail businesses. These firms purchase a number of trip packages from large corporations, which then expire because of a lack of demand, resulting in a cost rather than an investment. To address this issue, we suggest detecting flaws that restrict a firm by re-engineering processes, allowing the creation of a business architecture based on emotional analysis, allowing small and medium-sized tourist companies (SMEs) to make better decisions and evaluate data.

For more than a decade, business intelligence has been an important study subject in tourism, and with the arrival of big data, it has gotten even more attention. Big data summarises themes such as integrating large volumes of data from external data sources (e.g., online content), extracting information from any type of data source, particularly unstructured data (e.g., customer evaluations), and integrating data in real-time, as needed. For the tourist industry, business intelligence and big data are only beginning to reveal their full potential. Because of the critical function and relevance of social media and online product evaluations in tourism, the aforementioned trends are becoming increasingly important for tourist businesses to maintain their competitiveness. More sophisticated IT systems, as well as new algorithms and methodologies, particularly in the areas of online content mining and text mining, open up new application domains for business intelligence approaches that have already attracted a lot of study.

Keywords: sentimental analysis, Business Intelligence, tourism

Introduction

Big Data has been a popular phrase in recent years (a Google search provides more than 55 million items containing the expression). The phrase refers to the huge amount of organised and unstructured data that appears to be available on the Internet but is difficult to handle using typical software approaches or statistical methodologies. It is a fast developing field of study that is frequently regarded as a key component in boosting economic success and better
understanding or addressing societal issues. Many individuals regard Big Data (BD) as a fantastic opportunity because of its purported ability to answer virtually any query about people's actions, thoughts, and feelings (Martin-Sanchez & Verspoor, 2014). In fact, it's surprising to see how a phenomenon once regarded as perplexing and perplexing, the so-called information overload, has been renamed Big Data and is now regarded as a kind of silver bullet capable of providing a wealth of valuable and unquestionable insights into many aspects of modern life for individuals, organisations, and markets. Many of these statements, on the other hand, appear to be more than credible, and the capacity to investigate complicated phenomena by integrating widely available sources of data can be a significant advantage for those who can fully use it. BD, on the other hand, presents a variety of obstacles and hazards that have been extensively examined in a number of studies (Senthilkumar et al., 2018). They mostly pertain to the technical and methodological challenges of dealing with such vast amounts of quickly changing data sets. Aside from that, a decent set of specialised talents and resources is required, and a new approach to data collecting and analysis is thought necessary in comparison to the one that has characterised data collection and analysis for millennia. Nonetheless, scholars and practitioners agree that having access to a vast amount of data that covers virtually every aspect of human life has significant advantages, primarily because the data is "spontaneously" generated and does not suffer from the selection biases that can be present in traditional investigation methods. BD can, in any event, be a helpful and crucial addition to more solid, rigorous research methodologies, even when methodological concerns are taken into account. Business Intelligence (BI) initiatives have begun to use Big Data as a source. Although BI analytics has a longer history, the area is extremely sensitive to any data or information source that might improve the return on investment. As a result, both disciplines are quite complimentary. Advanced analytics and better, more comprehensive sources can give a more comprehensive view of the data, which can benefit from more organised and rigorous experience. Business intelligence's interpretation layer can thus be critical in making sophisticated BD analytics actionable (De Mauro et al., 2015).
In recent years, tourism has increasingly recognised the need for a more customer-focused approach that prioritises tourists' needs, preferences, and requirements in order to improve the quality of their experience and achieve higher satisfaction, which has proven to be a key factor in all travel decisions. Given these premises, the question is: to what degree is tourism academics aware of these issues and working on them? This short note offers the findings of a review of contemporary literature on Big Data and business intelligence, as well as the application of related approaches to the fields of travel, tourism, hospitality, and leisure, to address, at least partially, this issue (De Mauro et al., 2016). Big Data is fast infiltrating the field of tourist research. With the growing need for real-time and tailored information, the four Vs of Big Data, namely volume (size), variety (various sorts of data), velocity (high speed, and real time), and veracity (uncertainty, and validity), are especially essential in consumer research (IBM n.d.). Because customer experience is so important to the tourist industry's development and reputation, it has mostly adapted to changing technologies and the availability of new data sources. The majority of tourism services are now available via internet booking platforms. In addition, one of the most popular subjects on social media, such as Facebook and Twitter, is travel. As a result, it's not surprise that tourism has been named the top industry for online interaction (Mack, Blose, and Pan 2008). Every action you do on the internet leaves a digital trail. It's time to look at how these new sorts of data are being used by tourism researchers, and if they're part of a new research paradigm that includes unique approaches and has the potential to expand our theoretical knowledge of tourism (Baggio, 2016). To far, internet data sources have primarily been employed in applied research, where enormous and frequently free quantities of data that give insights into the tourism/travel sector and its clients have been used. Previous study has focused on company strategy creation, innovation and product development, and marketing strategies, which is unsurprising. The notion of tourist happiness is crucial in the tourism business, which is a service-based sector that relies on good consumer feelings and feedback. Satisfaction has long been studied and debated as a theoretical notion, and many tools exist to operationalize and assess it. It does, however, rely heavily on survey data for data
collection. It is generally known that survey-based systems have various flaws, including high costs and logistics, as well as the possibility for multiple bias. As a result of confirmation bias, visitors' replies to survey questions may reflect an intrinsically favourable appraisal as a result of their strong commitment in their trip. Other acknowledged issues with survey-based procedures include interviewer bias and cultural impact in responding specific questions. Furthermore, surveys only address preset features of the destination, making them insufficiently thorough (Song & Liu, 2017).

**Literature review**

Marco Rosetti et., al. researched on , "Analyzing User Evaluations in Tourism using Topic Models," proposed an innovative technique for analysing customer reviews using topic modeling's unsupervised learning. The technique extracts user profiles, which include topic-wise user preferences, as well as item profiles, which include topic-wise customer satisfaction, based on the themes expressed by users in their evaluations. A recommendation service uses the fit between concrete user profiles and item profiles as input. Furthermore, the method allows for a multi-criteria characterisation of user preferences and item ratings, allowing for a clearer explanation of a particular suggestion, comparison of different things, and even the calculation of ratings for specific reviews (Kaisler et al., 2013). The paper's contribution—the novel topic-criteria model and extended topic-sentiment-criteria model—were evaluated using a Yelp and TripAdvisor dataset, demonstrating not only comparable or even superior accuracy to existing approaches, but also practical relevance in scenario review and interpretation, recommendation, and rating calculation. Matthew James Krawczyk and Phil Zheng Xiang's research, "Perceptual Mapping of Hotel Brands Using Online Reviews: A Text Analytics Approach," employs a text analysis approach to generate perceptual maps using the most frequently used phrases in a data set acquired from an online travel agency (Hariri et al., 2019). The maps reflect the hotel industry's structure, which is reflected by hotel class and service features. These maps demonstrate how text analytics may help us gain a better understanding of how hotel brands,
which are made up of both tangible and intangible assets, are created and distinguished in the minds of consumers. Customers link brands with phrases that describe the hotel's class, according to the findings. The authors suggest a new line of inquiry into the nature of online reviews, specifically how they might be utilised to better comprehend the hotel and tourist industry's market structure. Aitor GarciaPablos, Montse Cuadros, and Maria Teresa Linaza's work "Automatic Analysis of Textual Hotel Reviews" describes a natural language processing (NLP) platform that includes text processing methods such as sentiment analysis. In the hotel area, opinion mining and named entity recognition (NER) will be applied to textual information in the form of user reviews. The offered technique is built on a homogenous data representation format that is used across all text processing phases (Lakshman & Malik, 2010). This so-called knowledge annotation format (KAF) allows for the supply of meta data on many layers for various text processing and analysis activities, allowing for the flexible integration of various text processing modules. The proposed technique was tested using hotel reviews taken from the review sites Zoover and Holiday Check, resulting in acceptable precisions for the various text processing jobs and proving its applicability in flexibly integrating diverse tasks into an overall text processing pipeline. Christopher Patrick Holland, Julia Andrea Jacobs, and Stefan Klein's paper, "The Role and Impact of Comparison Websites on the Consumer Search Process in the US and German Airline Markets," demonstrates a novel use of online panel data to investigate detailed aspects of customer search behaviour, in particular the interaction effects between different types of websites, in this case comparison websites and airline websites. As an innovative technique, set theory is used to analyse audience duplicate reports. The travellers' search process is represented using the notion of contemplation and data from ComScore, the top commercial supplier of consumer analytics based on a global panel of two million internet users (Hair, 2007).

Tourism as a Digitally Supported Industry The tourist business has been transformed from a brick-and-mortar and person-to-person service sector to a strongly digitally enabled and
ubiquitous travel service network as a result of technological advances connected to the Internet, including smartphones and tablets. Individual travellers and groups now have far more influence over the planning, construction, and customization of their journeys. They not only connect with a variety of platforms and online intermediaries to broaden their knowledge of travel and tourism decision-making, but they also network with other travellers who share their experiences. Travelers can submit comments and make suggestions for other travellers via internet platforms. As a result, new Internet technologies have given people a voice who previously had none (Diefenbach et al., 2018). Expedia, VirtualTourist, TripAdvisor, and LonelyPlanet are the most successful professional platforms in the travel and tourism industry. TripAdvisor alone receives 350 million monthly unique visits and provides over 320 million reviews covering hotels, restaurants, and activities (TripAdvisor, 2016). When compared to company websites and professional reviews, the information offered by these independent platforms has been shown to be superior and more trustworthy. Online social media sites like Twitter, Instagram, Facebook, Foursquare, Sina Weibo, and Google Plus, in addition to professional systems, play a vital part in producing electronic word-of-mouth (e-WOM). Importantly, online social media, travel professional websites and platforms, and blogs provide low-cost ways to collect extensive, real, and unsolicited data about passengers' views. While personal recommendations are frequently the most significant source of pre-trip information (Xu et al., 2020), the overall trustworthiness of blogs and online social media is very strong when compared to traditional WOM. As a result, opinions obtained from families, friends, coworkers, and official sources are now supplemented by social media and blogs. Searching, manipulating, and aggregating data to extract relevant and useful insights about tourists' attitude, behaviour, and experience quality has become a tedious and time-consuming task for both travellers and industry users, as well as professional and academic researchers, as the volume of online information continues to grow at a breakneck pace. The demand for autonomous multiaspect algorithmic and machine-operated systems is growing in order to better evaluate massive data volumes. In the literature, the value of leveraging social media data and data mining techniques and methods in tourism has been
examined (Chen et al., 2015). The primary phases employed in most applications in connection to social media data analysis in tourism are data collecting, data cleaning, mining procedure, and finally assessment and interpretation of the results. Text summarization and text categorization, as well as natural language processing (NLP), were formerly utilised to help with data processing and analysis. Machines may also model sentiment for automation and integration across many applications. Sentiment analysis is the process of analysing text and identifying subjective information using computational linguistics and natural language processing (NLP). While sentiment analysis has been studied since the 1970s, it has only lately gained popularity among scholars and practitioners. The escalation of web and social media-based information, the emergence of new technology, particularly machine learning techniques for text analysis, and the creation of new business models and apps that exploit this information are all driving the interest. Despite its popularity, sentiment analysis is still in its early stages when compared to other technologies like data mining and text summarization.

What Is Sentiment Analysis?

Opinion mining based on sentiment orientation has been investigated in recent years to establish the reliability of material and reasons for posting reviews, as well as to discover perspectives and features of demographic or market groupings. Diverse sentiment analysis approaches have been developed in various sectors, resulting in a modest number of review publications on the subject. To yet, there has been no mention of tourism in any of the evaluations (Liu, 2012).
Figure 1: machine-learning-based sentiment analysis system (Neethu & Rajasree, 2013).

Sentiment analysis is based on the assumption that information presented through text (e.g., a review) is either subjective (i.e., opinionated) or objective (i.e., factual). Subjective evaluations of entities or occurrences are based on personal feelings, beliefs, and judgments. Facts, data, and measurable observations are used to create objective reviews. Happiness, frustration, disappointment, joy, and other emotions are frequently expressed in customer evaluations and social media posts. Tourism groups and enterprises that want to enhance customer management and profitability might benefit greatly from tapping into these vast quantities of subjective e-WOM. Sentiment analysis is a polarity classification challenge in terms of methodology. Sentiment polarity categorization can be thought of as binary, ternary, or ordinal depending on the number of classes involved. We presume that a particular customer review is subjective in a
binary categorization (Neethu & Rajasree, 2013). In other words, a binary categorization implies that a particular text is mostly positive or negative, and then assigns a polarity to the review as "positive" or "negative." The positive and negative poles of sentiment are defined differently depending on the application and domain. In the context of tourism, "positive" and "negative" may relate to "satisfied" and "unsatisfied," respectively, although more study is needed to link sentiment polarity to theoretical frameworks of satisfaction. Because reviews are not always subjective, the binary classification should be expanded to a ternary classification that includes a third, "objective" category. In the ternary classification problem, the classifier conducts an implicit classification to distinguish between objective and subjective statements, assigning a "positive," "negative," or "neutral" class label. Positive and negative polarity are sometimes confused with neutral polarity (Agarwal & Mittal, 2016). A cascaded technique, consisting of a binary classifier to distinguish between subjective and objective reviews and a binary polarity classifier to further categorise subjective reviews into two categories, positive or negative, may also be used to address sentiment analysis. The terms that are clearly defined as positive or negative in a dictionary are rarely seen in objective assessments. They may also comprise polarities that are blended without a distinct sense of direction. In addition to straightforward binary and ternary classification, ordinal classification may be done using a sentiment strength rating system (e.g., one to five stars). It's also crucial to know what a sentiment refers to while doing sentiment analysis (Ahmad et al., 2017). The determination of the topic of an emotion expression is linked to the detection of a target and aspect. Aspect-based review mining is aided by sentiment analysis at the sentence level. A sentiment component might relate to a physical or tangible thing or a more abstract issue, depending on the level of granularity of analysis. An implicit or explicit reference to a target or an aspect is possible. It's easier to examine reviews with stated aims or elements than those with implicit ones. A hotel review can be made up of several features of a hotel, such as "the size of the bed was little and there was a noisy refrigerator," which expressly specifies two aspects of a "hotel room" as "small bed" and "noisy." The word "expensive" is an implicit part of the review that refers to the "price" of the hotel,
whereas "hotel was costly!" is an explicit aspect of the review that refers to the "price" of the hotel. The accuracy of sentiment analysis findings improves when both implicit and explicit parts of evaluations are effectively extracted, according to Aurchana, Iyyappan, and Periyasamy (Boiy & Moens, 2009). In addition to data on who gave the information and when it was delivered, a thorough sentiment analysis includes information on who provided the information and when it was provided (Singh et al., 2017).

**Sentiment Analysis Methods**

Sentiment analysis is a multistep procedure that includes data retrieval, data extraction and selection, data preprocessing, feature extraction, topic detection, and data mining. The identification and specification of the data source, such as a commercial service provider site or a social media network, is required for data retrieval. To gather review data from these sources, a particular web crawling method is required to retrieve data and then record it in a database that takes into account the data format. Review data must be retrieved from a group of heterogeneous data fields once data has been collected in a database (Gautam & Yadav, 2014). In the case of TripAdvisor data, for example, a review is contained within a retrieved HTML document, which is made up of various parts including footers or headers, tags, and the review content itself. Using proper phrases, the review text must be extracted. Each extracted review includes one or more phrases expressing the reviewer's point of view. To prepare reviews for the next stage, the preprocessing step performs activities such as separating a review into sentences, splitting a sentence into words, tokenization, stop-word filtering, part-of-speech (POS) tagging, stemming, and the transformation to lower/upper cases (i.e., feature extraction) POS tagging is an essential preprocessing operation that is usually done as part of sentiment analysis. It involves giving a label to each word (for example, noun, adjective) etc (Aydoğan & Akcayol, 2016). The technique of generating a collection of discriminative, informative, and non-redundant values to numerically represent a review or text is known as feature extraction. Term frequency (TF) or term frequency–invers document frequency is an extensively utilized feature extraction approach
based on term occurrences. Reviews or phrases are turned into a "term document matrix" using the TF feature extraction approach. Subject detection is a multiclass classification issue in which a text is assigned to a topic class based on its content and intended use. This section builds on the technical description of sentiment analysis by looking at how sentiment analysis has been used in tourism. It's also interesting to see if tourism-related research are adopting cutting-edge approaches or if there are other ways to improve the sentimental analysis field (Mitra, 2020).

**Use of Identified Studies and Datasets**

Instead of conducting a specific search within the Scopus and Web of Science websites, key words such as "sentiment analysis of tourism," "tourism sentiment data," "sentiment analysis of hotel reviews," and "sentiment analysis of restaurant reviews" were used to search for and retrieve relevant articles published on the Internet using the Google search engine. We went on to look at some recent review articles on sentiment analysis to see which ones mentioned tourism (Jagdale et al., 2019). As a consequence, we feel we have found a critical mass of tourism-related sentiment research for this analysis. An overview of significant tourism-related research is offered, as well as their unique datasets. For sentiment analysis, tourism researchers have traditionally employed two forms of online content: tourism reviews from professional websites (e.g., TripAdvisor, Booking, and Ctrip) and social media posts (e.g., Twitter) (Jagdale et al., 2019). Short text is common in both sorts of sources. Twitter, for example, limits tweets to only 140 characters, allowing enabling sentiment analysis at the sentence level. To train and assess sentiment analysis systems, manual and machine annotation procedures were utilised to label the reviews. It should also be mentioned that the majority of the datasets utilised in the literature are related to hotel stays. Restaurants and aeroplanes are the subject of a few studies. Shimada et al. designed an unsupervised machine learning strategy based on the Nave Bayes classifier (2011) to do a sentence-level sentiment analysis of tourist data. Using automatically labelled data, the Nave Bayes sentiment classification technique was trained. Instead of using words like "great" and "bad," emojis were utilised to indicate positive and negative seeds to classify data for training.
As a result, evaluations with a happy face were classified as good, while those with an angry face were classified as negative. Tourism data was also subjected to K-mean clustering approaches and statistical models based on the probability distribution of reviews in sentiment space (Schwartz and Uysal 2015). A dictionary-based method has been used in several tourist studies. Misopoulos et al. (2014) assessed the polarity of Twitter postings related to airline service delivery using a lexical type technique. To do a sentence-level sentiment analysis of tourist data using automatically labelled data, the Nave Bayes sentiment classification technique was trained. Instead of using words like "great" and "bad," emojis were utilised to indicate positive and negative seeds to classify data for training. As a result, evaluations with a happy face were classified as good, while those with an angry face were classified as negative. Tourism data was also subjected to K-mean clustering approaches and statistical models based on the probability distribution of reviews in sentiment space. A dictionary-based method has been used in several tourist studies. Misopoulos et al. (2014) assessed the polarity of Twitter postings related to airline service delivery using a lexical type technique (Basarslan & Kayaalp, 2020).

Evaluation Metrics

As previously stated, most sentiment analysis methods classify items into two categories (positive and negative) or three categories (positive, neutral, and negative). It is critical to assess and quantify the results of various procedures. A confusion matrix also known as a contingency, is a simple and straightforward approach to illustrate a classifier's prediction results (Gupta et al., 2017).
Figure 2: The knowledge destination framework

The proposed architectural framework, like , distinguishes between a knowledge creation layer and a knowledge application layer, with the former including various sources of customer-based data (e.g., web-search, booking, and feedback data), technical components for data extraction, transformation, and loading (ETL processes), a centralized data warehouse, and data mining, and the latter including data mining. The knowledge application layer, DMIS-re'cockpit, is responsible for the decentralised presentation and ad-hoc visualisation of data mining models and underlying data. Figure 2 depicts the essential components of the knowledge destination framework architecture, which will be discussed further below.

Data sources

Customer-based data can take the form of explicit tourist feedback, such as guest surveys and e-reviews, or implicit tourist information traces, such as web-navigation data, online requests, booking and payment data, and GPS-based coverage of tourists' spatial movements, which are all provided unknowingly and unintentionally. Unstructured data, which includes free text (e.g. e-reviews) and rich content from web 2.0 apps, may be distinguished from structured data, e.g. transaction data, polls, ratings (Anicic et al., 2011).
Data extraction

Depending on the data format, different data sources require different strategies for extracting, transforming, and loading (ETL) essential information. Typically, heterogeneous data sources are integrated by extracting structured and semi-structured data (html texts) using semantic, linguistic, or constraint-based information integration strategies. Wrappers or text mining are used to extract data (i.e. statistical language models, natural language processing approaches) (Bansal, 2014).

Data warehousing

Data from many sources is mapped into a common data format and stored in a single Data Warehouse that contains all necessary data for target stakeholders. It is feasible to conduct a destination-wide and all-stakeholder analytical approach using a harmonisation procedure. Individual data sources are turned into a central data model and a dimensional structure using a tourism-ontology. Data mining for knowledge generation: Data mining methods may be used to find interesting patterns and correlations in the data (Deb Nath et al., 2015). However, data mining has only lately become relevant in tourism because of its potential to uncover new patterns in large data sets and, unlike other statistical approaches, its ability to examine non-linear correlations in the data. Furthermore, as compared to other statistical approaches, data mining has less assumptions about data quality because data might be incomplete, noisy, redundant, and dynamic. Although data mining's potential in tourism has yet to be completely realised, all main data mining techniques can be found in the literature. For example, descriptive/explorative analysis in the form of reports (OLAP) may be used to depict tourism arrivals based on time/season, travel type, or customer origin (Bansal & Kagemann, 2015).
A business intelligence-based destination management information system

Designing and engineering a Business Intelligence-based destination management information system (DMIS) requires a profound understanding of the nature of knowledge behind management processes and an appropriate interpretation of the management objectives behind decision making at the level of tourism destinations. According to the literature, knowledge relevant in a tourism destination context subsumes knowledge about market cultivation (e.g. how to attract valuable customers) and knowledge relevant for destination management, development, and planning (e.g. the development of new product–market combinations for valuable customer segments, training, private–public partnerships, etc (Nath et al., 2017). Especially, customer-based knowledge is gained through customer segmentation techniques and service performance evaluation. Thus, the knowledge sources considered in our presented BI application reflect tourists’ search behaviour (i.e. Web navigation/search), tourists’ booking behaviour, and tourists’ feedback (i.e. feedback from a variety of polls and online review sites). To put it another way, visitors’ demographic and psychographic features, purchase motives and brand perceptions, as well as customers’ information usage and product consumption habits, are among the data gathered, saved, analysed, and displayed in DMIS-Åre (Juneja & Das, 2019).

Concluding Remarks and Future Directions

In practise, the challenge of obtaining and processing increasingly high velocity and massive volumes of data has become extremely complicated, necessitating the development of automated machine-based systems. There are several approaches for extracting sentiment from online material, which have been explored in this article from both a general and a tourist viewpoint. Aspect-oriented sentiment analysis remains a tough subject due to the difficulty of recognising and locating implicit elements in reviews (Nath et al., 2017). Future research in aspect-oriented sentiment analysis will necessitate close collaborations between domain experts (i.e., tourism researchers), information technology, and NLP scientists to create and make publicly available
some specific dictionaries for topics/aspects, as well as annotated review databases related to tourism industries. This will initially aid in the development of a more complex aspect-oriented sentiment analysis model to address the issue of implicit aspect identification in reviews. Second, it will advance tourism research by generating new hypotheses, such as discovering/understanding the relationship between satisfaction and sentiment, and then assessing tourist satisfaction (Juneja & Das, 2019). Furthermore, tourism research can benefit from employing Big Data and deep learning methodologies to identify dynamics based on vast linked sets of data and gain additional understanding from many elements of Big Data. Tourism research may progress into a new domain where theory-driven techniques and data-driven practices can complement one another in order to better understand or explain phenomena and realize new dimensions in ideas. The conclusion of this review study is that tourist sentiment analysis is just the tip of the iceberg in terms of a new tourism research paradigm. Sentiment analysis is only the beginning of more complicated Big Data methods. Integration of many forms of data, in particular, provides a lot of promise for creating new insights at previously unheard-of scales (Baumgartner et al., 2005).

References


24. Kilanko, V. The Potential Effects of Biden’s Infrastructure Bill on the American Economy.


32. Khan, M. S. Control of Autonomous License Plate Recognition Drone in GPS Denied Parking Lot.