International Journal of Hospitality Management 92 (2021) 102658

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhm

Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach

Tianxiang Zheng [a], Feiran Wu [a], Rob Law [b], Qihang Qiu [c], Rong Wu [d,*]

[a] Shenzhen Tourism College/JNU-UF International Joint Laboratory on Information Technology & Tourism, Jinan University, No.6, Qiaocheng East Avenue, Overseas Chinese Town, Nanshan District, Shenzhen, Guangdong, 518053, PR China
[b] School of Hotel and Tourism Management, The Hong Kong Polytechnic University, 17 Science Museum Road, TST East, Kowloon, Hong Kong SAR, 999077, China
[c] Faculty of Human Geography and Planning, Adam Mickiewicz University, Krygowskiego 10, 61-680 Poznan, Poland
[d] School of Architecture and Urban Planning, Guangdong University of Technology, No.729, Dongfeng East Road, Yuexiu District, Guangzhou, Guangdong, 510090, PR China

ARTICLE INFO

Keywords:
Online customer review
Review reliability
Review rating prediction
Deep learning
Information quality

ABSTRACT

This study considers the review reliability problem by identifying biased user-given ratings through rating prediction on the basis of the textual content. Deep learning approaches were introduced to investigate the textual review and validate the effect of rating prediction using a dataset collected from Yelp. The definition of "biased rating" was clarified and influenced the matching rules. The approach obtains high performance on a total of 1,000,000 reviews for prediction, with user-given ratings as the benchmark. Using the revealed biased ratings, unreliable reviews were detected by combining the results of several deep learning kernels. Findings shed light on understanding review quality by distinguishing biased ratings and unreliable reviews that may cause inconsistency and ambiguity to readers. Hence, theoretical and managerial areas for social media analytics are enriched on the basis of online review meta-data in hospitality and tourism.

1. Introduction

Review data, posted on tourism and hospitality websites, such as TripAdvisor, Yelp, and Expedia, have received considerable academic attention from different theoretical or practical perspectives (Ma et al., 2018). Furthermore, review data represent a promising research direction (Xiang et al., 2017) possibly due to their low cost and accessibility. Analytical approaches are increasingly applied to collect, analyse, summarise and interpret online review data to extract useful patterns and insights pertaining to managerial problems.

Previous research has shown that the review quality problem cannot be ignored (Li et al., 2018; Ma et al., 2018; Xiang et al., 2017). To illustrate, the review rating and verbal review given by a user on a social media platform have been found to be the second most-trusted source of brand information (after recommendations from friends and family) (Gavilan et al., 2018). Compared with review text, review rating is faster, more instantaneous and more easily accessible because important information can be straightforwardly obscured unless users are willing to spend considerable time and effort on thoroughly reading the textual reviews. Hence, as arguably the first influence on the judgement of a consumer, review rating is an important heuristic element and information cue to simplify customer search. Indeed, consumers have faith in these ratings (Gavilan et al., 2018; Gössling et al., 2019) that could reduce their cognitive effort in making decisions (Gursoy, 2019; Schuckert et al., 2016). Nevertheless, user-given rating can be subjective and tends to be slanted based on the thinking of the writer of the ideal rating for the review on his/her purchase, including personal knowledge and past experiences. A biased rating may mislead potential customers in making purchase errors, especially those who believe that reading reviews word-by-word is cumbersome. Moreover, business ratings on various platforms are sometimes computed by the average of review ratings for the business (see Fig. 1 for details). Thus, these biased ratings may induce unfairness to the business if the business and the review ratings are dependent because the former contributes to which business would 'survive' after the shortlist of filtering of users.

Inconsistencies were likewise found (Antonio et al., 2018; Schuckert et al., 2016; Xiang et al., 2017) between a textual review and the satisfaction score assigned to the product. Review ratings were likewise

* Corresponding author.
E-mail addresses: zheng_tx@jnu.edu.cn, tianxiang_z@sina.com (T. Zheng), wufeiran@stu2018.jnu.edu.cn (F. Wu), rob.law@polyu.edu.hk (R. Law), qihang.qiu@amu.edu.pl (Q. Qiu), wurong5@mail2.sysu.edu.cn (R. Wu).

https://doi.org/10.1016/j.ijhm.2020.102658
Received 15 August 2019; Received in revised form 12 February 2020; Accepted 16 August 2020
Available online 10 October 2020

Fig. 1. Typical review on a hospitality platform (text with rating).

predicted by taking advantage of machine learning algorithms, regression models and recommendation systems (Antonio et al., 2018; Ganu et al., 2013; Lei and Qian, 2015). Nevertheless, previous studies have contained a few limitations as follows: (1) No test dataset was utilised (Antonio et al., 2018), thereby rendering difficulties in a good estimation on the generalisation capability of the machine. (2) Performance evaluation used mean absolute error (MAE) and/or root mean square error (RMSE), which measure accumulated errors and thus are commonly used in regression but less accurate in classification. Therefore, a low MAE or RMSE value does not necessarily indicate good performance. For example, if we have four samples, then the actual value for their labels is [1,1,1,1], if the predicted value of the method is [2,2,2,2], then the MAE is 1 and if the predicted value of another method is [5,2,1,1], then the MAE is 1.25. However, the latter method performs better because the accuracy score is 2/4, whereas the former is 0. Individually importing performance evaluation to each review would be intriguing. (3) Traditional models were used and thus were inherently limited in their power by the failure to handle ever changing, complex problems. For instance, regression models (Ganu et al., 2013) were not highly effective in processing nonlinear data, whilst recommendation systems (Ganu et al., 2013; Lei and Qian, 2015) represent a matrix factorisation, so they cannot achieve desirable results when an item (business here) or a user (reviewer here) is newly added to the system. Deep learning (Hochreiter and Schmidhuber, 1997), an advanced machine learning technique, has been found to outperform conventional machine learning models in predicting review reliability (Ma et al., 2018) and forecasting tourist arrival volumes (Law et al., 2019). Although considerable efforts have been invested in nearly every field of business domain, deep learning remains in its incipient stage for hospitality and tourism research. Hence, the practical relevance of recognising whether or not a review rating is slanted by using this novel approach would be interesting. (4) Homogeneous users, according to their similarities of reviewing behaviour (Ganu et al., 2013) or the subjective sentiments of users toward an item to reflect their interest preferences (Lei and Qian, 2015), were grouped before making recommendations or predictions. Although the goal was to capture the underlying inter-dependencies between user ratings, the prediction accuracy was highly dependent on the clustering outcome. As we aim to eliminate the biased effect from different users and uncover how the review text contributes to the corresponding rating, user related information was excluded in this study, regardless of individual preferences. (5) Previous studies have been based on hardcoded features, such as the reputation similarity of items (Lei and Qian, 2015), manual annotation on six topics (Ganu et al., 2013) and sentiment strength (Antonio et al., 2018). Consequently, important information in the data may be diluted. For instance, informative value containing various sentiments in a review was compressed into a numeric score named "average sentiment strength" (Antonio et al., 2018), which included most but not all of the necessary information into the related feature on review sentiment. Thus, these manually designed indexes were limited to understanding the in-depth meanings embedded in reviews that might aid the interpretation of textual data in a holistic way. (6) The main purpose of the studies of Lei and Ganu is to enhance the quality of prediction rather than to reduce the biased effect. Antonio partially focused on detecting fraudulent reviews. However, several drawbacks mentioned above would hinder the practical application of the study.

Given this context, determining or detecting suspicious online reviews (e.g. fake, manipulated, biased, irresponsible or perfunctory) can be profoundly important and remains an interesting topic due to the tremendous advancements of machine learning. As such, the search continues for a holistic feature, a fine-grained evaluation measurement and a complete verification to estimate the generalisation capability of the machine. Hence, the current study aims to fill this gap by considering the problem as classification prediction and attempts to forecast ratings solely from reviews by taking the novel approach of combining emotional word searches, word embeddings and deep learning models to harness the wealth of review reliability.

2. Research background

In recent years, several aspects within the big-data domain have been addressed and used to contribute to the research agenda in tourism. Review ratings, sentiment analysis, perceived value and review helpfulness have also been addressed using a big-data approach. A comprehensive literature review has likewise conducted on online review studies (Gursoy, 2019; Kwok et al., 2017).

2.1. Information quality of online reviews

Information quality acts as the antecedents of online word-of-mouth (eWOM) for hotels (Sijoria et al., 2019). Hence, suspicious online reviews (or ratings) have attracted growing criticisms and concerns when using user-generated content (UGC) as research data in recent years. In the early stage, statistical methods were proposed to detect online review manipulation (Hu et al., 2012) and demographic bias of reviewers (Wijnhoven and Bloemen, 2014). Xiang et al. (2017) comparatively investigated three major online review platforms in terms of information quality about hotels. Later, the reliability of social media data was assessed, and a considerable amount of data noise was found to possibly lead to misclassification (Xiang et al., 2018). Recently, Antonio et al. (2018) developed a review rating prediction to detect fraudulent reviews and create proprietary rating indexes. Costa et al. (2019) used a data mining approach to predict whether or not a new review published was incentivised. Mellinas et al. (2019) found that ratings of hotel location can be spurious and their values may be biased due to the influence of other attributes.

The external causes that may affect review quality were also considered. For instance, Liu et al. (2018) found that low-status (in terms of membership) reviewers were utilitarian and used few words in their reviews. Hong et al. (2019) conducted research to determine the effects of price promotion on online restaurant reviews and showed that the review rating is high from consumers who received a discount. Lastly, Gössling et al. (2019) presented the results from a qualitative study involving 20 hotel managers in southern Sweden about their perspectives on manipulation.

2.2. Inconsistency amongst textual review and ratings

Prior studies have addressed the inconsistent tendency either between overall rating and specific ratings or between textual review and its rating (Antonio et al., 2018; Schuckert et al., 2016; Xiang et al., 2017). Schuckert et al. (2016) obtained evidence of suspicious online ratings on TripAdvisor to investigate the self-contradiction of overall and specific ratings. They found that there was a significant difference between the two kinds of ratings. Xiang et al. (2017) used several key metrics to assess online reviews on three major platforms. They found that the inconsistency between review content and the reviewer's satisfaction rating could be captured by these metrics. Antonio et al. (2018) created a machine learning model to predict online review rating by using several features, and found that the difference between the predicted review rating and the actual rating could be used to detect possible fraudulent reviews.

2.3. Deep learning

The review rating prediction problem in hospitality research can be considered as a classification, which means that the prediction outcome was a class/category/discrete value (Antonio et al., 2018; Ganu et al., 2013). Machine learning provides solutions to such type of problems. With machine learning, data and answers are inputs, and rules are produced. Typically, machine learning is about mapping inputs to targets (i.e., the answers), which is done by observing numerous examples of inputs and targets. Machine learning must be trained rather than explicitly programmed. Conventional machine learning techniques, such as regression, support vector machines and early neural network, mainly rely on manually designed features that require considerable specific domain knowledge (Ma et al., 2018). Deep learning is a specific subfield of machine learning. The prevalence of deep learning lies in its success in a wide range of fields by providing effective solutions for a number of problems. This technique benefits from the ability of scaling to large datasets and simplified workloads of obtaining refined representations within a predefined hypothesis space (network topology) in feature engineering. Despite its great power, deep learning in the tourism and hospitality literature is scarce. Only a handful of relevant studies have been made to date (Law et al., 2019; Ma et al., 2018; Starosta et al., 2019; Zhang et al., 2020, 2019).

The network topology determines which kernel is used to map the input data (Chollet, 2017). For instance, densely connected kernel, also called fully connected (FC) or dense layers, can be used to store vector data. 2D convolution neural network (CNN) can be used to process image data. Long short-term memory (LSTM) can be used to map sequence data. The fundamental difference between FC and CNN is that the former learns global patterns in hypothesis space, whereas the latter learns local patterns. LSTM can save information for later use, thereby preventing older signals from gradually vanishing during the process.

2.4. Research motivations and questions

Our goal is to investigate the reliability problem of online reviews by its rating. Hence, before going any further, we consider the star rating on a hospitality platform. A business/review may contain several sub-ratings on different product attributes instead of one overall rating. Overall rating and sub-ratings are arguably similar in hypothesis development (Xie et al., 2014). Hence, if we ignore the self-contradiction between the overall rating and sub-ratings (Schuckert et al., 2016), then we retain brevity by regarding these two types of ratings as the same. Fig. 1 shows a review on one particular business.

Notably, the rating may not be highly trustworthy because each reviewer has different standard for rating. Alternatively, the customer may merely regard the rating assignment as a routine and thus posts them casually without careful consideration (Schuckert et al., 2016). On the premise that the review content is a truthful reflection of the evaluation of a customer regarding a business, immediate questions that arise in terms of review reliability are as follows. Can we describe this inconsistency of a particular review by utilising its rating? In addition, how do we detect suspicious reviews with biased ratings that would induce ambiguity to readers? Could it be possible to use only the review text in predicting its rating to lessen the biased effect? All these questions serve as guides to this study.

3. Research design, dataset and methods

3.1. Research design

We formally define the analysis of the rating impact on review reliability as a multinomial classification (with five classes in total) prediction problem in machine learning. The difficulty lies in effectively extracting useful features from the textual reviews and then quantifying their relative importance with respect to rating. Review rating was relative to the emotional view, emotion and attitude of the writer (Xie et al., 2019). Sentiment analysis has frequently been introduced to understand these opinions, sentiments and emotions in review text (Alaei et al., 2019; Calheiros et al., 2017; Duan et al., 2016; Kirilenko et al., 2018; Liu et al., 2019). Sentiment polarity was found to have a positive effect on review rating (Antonio et al., 2018; Lei and Qian, 2015; Zhao et al., 2019). Sentiment-related approaches were realised in dozens of computer programs, including those available online (e.g. TextBlob, SnowNLP and NLTK). Moreover, off-the-shelf software, such as the Stanford CoreNLP (http://nlp.stanford.edu:8080/sentiment/rntnDemo.html/), may be used to compute sentiment polarity. At first glance, sentiment is a good indicator of review rating, and thus an automated sentiment analysis may be attributed to the review, and then machine learning is applied to predict the rating from that polarity. However, from an analytical perspective, these methods have provided results that are far from practical applications, and the solution to sentiment analysis remains underexplored (Alaei et al., 2019; Kirilenko et al., 2018; Xie et al., 2019). A trade-off is undertaken by using keywords derived from the review text and then mapping these words to dense vectors by word embeddings.

We develop a predictor to create the model after feature extraction. The input denotes the review words, and the output indicates the predicted rating score. Given a review with words, we calculate its probability of belonging to a certain class. Thereafter, the predicted rating is compared with its user-given rating to further determine whether it is slanted. Thus, the inconsistency between textual reviews and user-given ratings can be characterised. Fig. 2 presents the framework with an accompanying example to demonstrate how a textual review is transformed into features (word vectors) that are fed into the classifier model to make a prediction on ratings (an integer from 1 to 5) and comparison with the user-given one.

3.2. Dataset

This study uses the publicly available Yelp dataset that has been frequently used as a primary data source in the academic literature within the hospitality and tourism field (Ma et al., 2018; Xiang et al., 2017). This dataset is chosen for the following reasons: (1) its simplicity for one overall rating instead of several additional ratings (e.g. in Expedia), (2) its relative balanced review length (Xiang et al., 2017), which is vital because we extract as many emotional words as features, (3) its relative 'saddle' sentiment score, with which the sample size in a different class (with respect to rating score) can be easily balanced, (4) its strong consistency between rating and review text (Xiang et al., 2017), which in turn leads to justified user-given ratings in the training process and (5) its strongest performance in attracting consumers to voice their dissatisfaction or complaints into the review content (Xiang et al., 2017), which helps extract additional meaningful words.

Data were downloaded in May 2019 from the official website (see

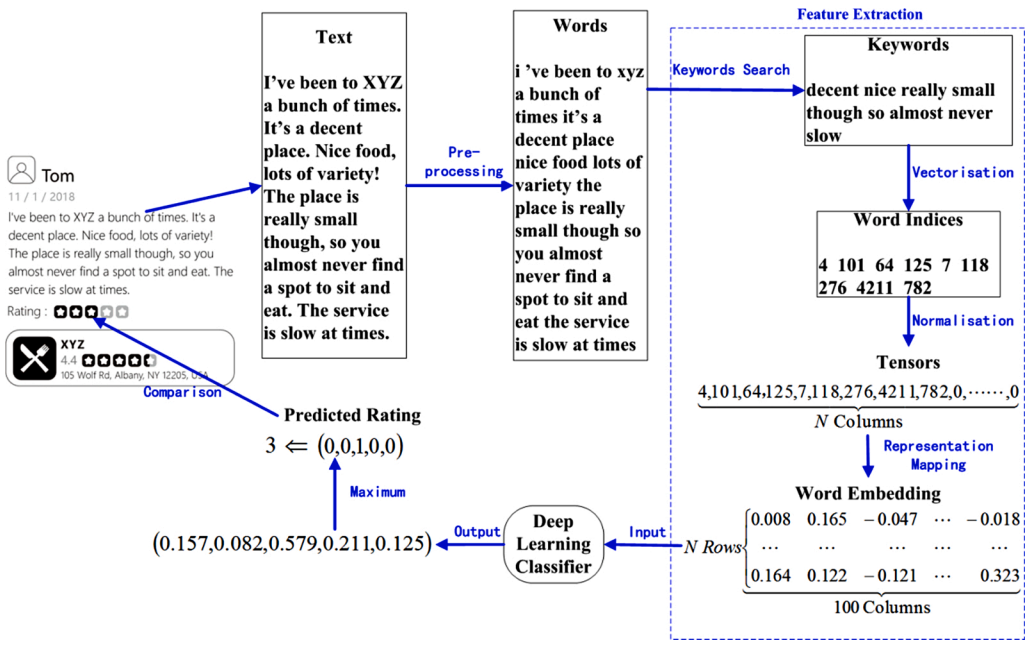

Fig. 2. Analytical framework to distinguish biased user-given ratings.

https://www.yelp.com/dataset). The raw Yelp data consist of six files, one for each object type, namely, review, tip, user, photo, check-in and business. The necessary data for this study are contained in the business.json and review.json files, and those for social information, photos and other data were excluded. In particular, the star rating in review.json is an integer in the range of 1–5.

The entire data set remains extremely rich, and its usage for analysis may exhaust the resources of our machine. Therefore, we are forced to investigate only a portion of the entire data (see the following section for details).

3.3. Methods

On the basis of the research design, the following five steps constitute the proposed method.

3.3.1. Data preprocessing: cleaning and tokenisation

The businesses described in the Yelp dataset belong to different categories, such as restaurants, shopping, hotels and travel. However, the textual reviews for different business categories may vary. Therefore, separately performing a review rating prediction for each business category is rational. Restaurants and food comprise 39% (74,599) of the 192,127 businesses, and 67% (4,502,556) of the 6,685,900 text reviews discuss eating or drinking. Hence, this study narrowed down the investigation to those reviews for eating or drinking only. Reviews with a category of 'restaurant' or 'food' were selected as our clean dataset.

Reviews written in English serve as the basis for the subsequent analysis. Review records from the clean dataset with an equal number in each class were randomly selected to construct the final dataset. Lastly, the review data were tokenised in the final dataset using the following steps.

- All spaces were removed from the review data.
- All punctuation marks and other non-letter symbols (e.g. '&, /') were removed.
- Elements with consecutive duplicates were eliminated.
- Capital letters were converted to lower case to reduce redundancy.

In total, the database (final dataset) contains about M reviews (for model training) and K reviews (for model testing). Each review was split into a collection of single words. The study set M to 900,000 and K to 100,000.

3.3.2. Feature extraction

The accurate calculation of the sentiment polarity remains highly challenging. This practical constraint led us to identify emotional seeds (Xie et al., 2019) directly from the raw data. The featured engineering consisted of three steps:

(1) Keywords search: Emotional words with the following parts of speech (POS) were considered to be feature candidates: adjective, adverb, determiner, interjection, conjunction and negative verb. In mining these emotional seed words, we used the natural language processing package (i.e. TextBlob) in Python (see https://textblob.readthedocs.io/en/dev/) to detect the POS tagging. Hence, a list of emotional words was obtained for each review text.

(2) Turning into Tensors: Classifiers used in the models only accept the input features of fixed length. Hence, the keyword lists were first converted into integer indices by a function called 'text_to_sequence' in Python's 'keras.preprocessing.text' package. Thereafter, the index lists were padded with zeros after the last word of each sequence to provide them with the same length of N, which is specified as the average number of keywords in the lists. Thus, the lists were turned into a 2D integer tensor of shape (M, N).

(3) Word embeddings: When dealing with extremely large vocabularies, word embeddings pack information into far few dimensions that allow us to efficiently train the neural network. The vector dimension was empirically set as 100, and word embeddings were learnt together with the main task (i.e. rating prediction) starting with random word vectors. Thus, each distinct word in the vocabulary was mapped into a 100-dimension vector to obtain a word embedding. Overall, we had a 3D floating-point tensor of shape (M, N, 100) as the input data of the classifier.

3.3.3. Classifier building

Considering that deep learning models outperformed various baseline models (Law et al., 2019; Ma et al., 2018), the current study used

deep learning models as classifiers for rating prediction. We applied three different deep learning kernels, namely, FC, CNN and LSTM, to effectively pinpoint biased reviews.

3.3.4. Model training and predicting

The final dataset was simply split essentially into an 80-10-10 partition (Chollet, 2017; Wen et al., 2019) on a total of 800,000 training data, 100,000 validation data and 100,000 test data, respectively. Multiple shuffling of the dataset and a 10-fold cross-validation technique (Bengio and Grandvalet, 2004; Refaeilzadeh et al., 2009) were performed to prevent over fitting, thereby assuring the independence of the split between training and validation data. The final test result of our experiments was the average performance of the 10-fold cross validation.

3.3.5. Model evaluation

Results were reported using Precision, Recall, Accuracy, F-measure (Powers, 2011) and the Area Under the Receiver Operating Characteristic Curve (AUC) (Fawcett, 2006) for the proposed model. Given that our purpose was to determine biased ratings, this study did not entirely seek performance on the basis of the above indices. Accordingly, several reviews were rooted out along with their predicted and user-given ratings to intuitively see which of these two ratings was more reliable.

4. Findings

4.1. Descriptive analysis

Table 1 shows the distribution of review properties across the data. Compared with the clean dataset that contains a total of 4,502,556 reviews, only one-fourth of the total data were left. However, either the number of distribution of reviews (see Fig. 3 for further details) in the test data or the average review length (excluding all possible punctuations) in each class in the final dataset is similar to that in the clean dataset. Table 1 shows that customers tend to post extra words and sentences with further detailed descriptions of the negative aspects of their experience, which is in accordance with existing findings (Zhao et al., 2019).

4.2. Predicting review rating with deep learning models

We applied three modules to test their capabilities in predicting review ratings. The first goal was to see if any differences existed between these models owing to the sequential order of the keywords. Table 2 shows the results with three predictors demonstrating the performance of the experiment on the basis of reviews on training, validation and test dataset. The results of the test dataset show that FC generates test scores (Precision, Recall, Accuracy and F-measure) below 0.59, whereas CNN and LSTM have these four test scores above 0.59. This finding implies that using 1D CNN and LSTM is beneficial for sequence data generation. Using these modules renders LSTM and CNN efficient when processing keywords to learn representations with generalisation power. The three kernels achieve an AUC value of 0.70 for training and test data, representing a fair model for each.

Table 2 shows that at first glance, none of the deep learning models demonstrates a very good predictor. On the basis of our observation, the difficulty lies in the fact that distinguishing the reviews from one another with star ratings of two, three and four is difficult. Another plausible explanation may be that no rigorous standard exists to classify a review text into five mutually exclusive classes. After all, text classification performance may differ significantly depending on many factors. The examples of such factors are the application domain and the language (Ma et al., 2018). Other factors like the feature engineering and the raw data may also affect the classification. Despite the imperfect performance, the accuracy reached by a purely random classifier in this five-class case is notably closer to 27% (after randomly shuffling the labels of test data). This situation is different from the balanced binary classification problem where an accuracy by a random classifier would be as high as 50%. Given all these aspects, the results here seem acceptable, at least when compared with a random baseline.

Table 1
Distribution of the review properties.

	Class	Final dataset		Clean dataset
		Training and validation data	Test data	
No. of reviews		900,000	100,000	4,502,556
No. of reviews in each class	1	180,000	11,747	544,186
	2		8,859	420,640
	3		13,068	604,373
	4		25,620	1,191,120
	5		40,706	1,820,145
Average word count (review length) in each class	1	127.3	131.7	126.7
	2	138.3	141.8	137.8
	3	131.6	135.1	131.1
	4	116.1	119	114.3
	5	88.6	91.1	86.6
Average rating		3	3.75	3.73

4.3. Distinguishing unreliable reviews based on biased ratings

Accuracy or precision is not the only purpose. We further validate the efficacy of the deep learning algorithm via an illustration of how we can use the model to uncover the unreliable reviews by distinguishing biased user-given ratings. As mentioned above, the agreement between two people classifying the rating of the same review text will never be perfect (Kirilenko et al., 2018). Accordingly, we use the intersection amongst the results of three predictors to determine the unreliable reviews and thus obtain a robust identification. Such an objective further explains our application of different deep learning kernels in the present study. Consequently, an unreliable review would be justified further using the following less rigorous definition of the matching rules for a biased rating.

Definition (biased rating and unreliable review): If the difference between the predicted and user-given ratings is greater than or equal to two, then the user-given rating is regarded as biased. If a particular review has been recognised as having a biased rating by all the deep learning modules, then such a review is deemed unreliable (with regard to its incongruency between review content and review rating).

Considering that the matching rule between the predicted and user-given ratings changes, we re-investigated our obtained results. The gains in this compromise are significant with an average of approximately 50% increase for any of the three cases (see Table 3 for details, where the corresponding indices are prefixed by 'adjusted'). Table 3 shows that the percentages of biased ratings (refer to quantity) on test data are approximately 10%, 9% and 6% for FC, CNN and LSTM models, respectively.

Under the new definition, we successfully identified the unreliable reviews by the overlaps of the resultant reviews with biased ratings. A total of 2839 unreliable reviews were determined out of the 100,000 test data. We rooted out certain examples to visually validate the effect. Table 4 shows the examples marked as unreliable.

Table 4 shows that our proposed method clearly achieved reasonably effective predictions in tasks related to inconsistency problems. Generally, the predicted model illustrates that the first half of the records we selected (i.e. from Review Nos. 1 to 5) are overrated, whereas the second half (i.e. from Review Nos. 6–10) are underrated. To analyse further, in Review No. 1, the reviewer claimed that the experience was worth 4.5 stars. However, from the review content below, the reviewer nearly had the same length of descriptions for either positive or negative perception of the experience. Thus, three modules unanimously anticipated this mingled review with a three-star rating. In Review No. 2, the reviewer

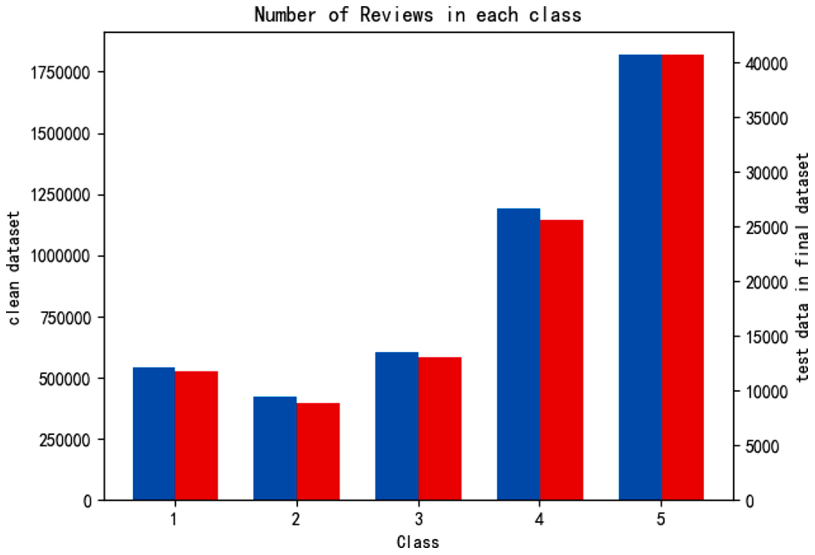

Fig. 3. Distribution of review number in each class on clean and final datasets.

Table 2
Results of review rating prediction on the final dataset.

	Training data			Validation data			Test data		
	FC	CNN	LSTM	FC	CNN	LSTM	FC	CNN	LSTM
No. of reviews	800,000			100,000			100,000		
Precision	0.62	0.62	0.64	0.58	0.59	0.59	0.59	0.61	0.63
Recall	0.62	0.62	0.64	0.57	0.59	0.59	0.58	0.59	0.61
Accuracy	0.62	0.63	0.64	0.57	0.59	0.59	0.57	0.60	0.62
F-measure	0.62	0.62	0.64	0.57	0.59	0.59	0.57	0.59	0.62
AUC	0.70	0.71	0.72	0.67	0.68	0.69	0.70	0.71	0.73

Table 3
Adjusted results of the review rating prediction on the final dataset.

	Training data			Validation data			Test data		
	FC	CNN	LSTM	FC	CNN	LSTM	FC	CNN	LSTM
No. of reviews	800,000			100,000			100,000		
Adjusted precision	0.94	0.95	0.95	0.94	0.95	0.95	0.91	0.92	0.94
Adjusted recall	0.94	0.95	0.95	0.94	0.95	0.95	0.91	0.91	0.94
Adjusted accuracy	0.94	0.95	0.95	0.94	0.95	0.95	0.90	0.91	0.94
Adjusted F-measure	0.94	0.95	0.95	0.94	0.95	0.95	0.91	0.91	0.94
Adjusted AUC	0.85	0.86	0.85	0.85	0.86	0.85	0.85	0.87	0.87

apparently made a very negative statement about the service by using words or phrases, such as ‘never’, ‘didn’t work’, ‘finally’ and ‘but no’. Correspondingly, a slight praise of the food could not compensate for this negative deficit. The LSTM model even generated the lowest star rating to such a textual review. Review No. 3 is similar to Review No. 2, but the tone is less negative, resulting in an ordinary negative score of two stars. Review Nos. 4 and 5 nearly have the same issue as that of Review No. 2, except for the one-star discrepancy between LSTM and the other two models in Review No. 5. Review Nos. 6 and 10 reflected a neutral rating to the experiences of the reviewers. However, the sentences clearly have a very positive tone, thereby explaining the five-star produced by our models for either case. In Review No. 7, the reviewer was extremely upset for not having lobsters and thereafter gave a bad rating of two stars. Another user reading this review would probably comprehend the reviewer’s frustration and pay extra attention to the positive parts, such as ‘good’, ‘enjoy’, ‘superb’ and ‘come back’. The models forecasted that the rating should be five or at least four stars, which appear reasonable with intuitive inference. Review Nos. 8 and 9 reflected the lowest stars to the texts of reviewers. Nevertheless, the sentences are positive about the food and only slightly negative about the service (tips) or size (sandwiches). According to our predictive model, a negative sentiment about such a small aspect cannot defeat the positive perception of the entire review owing to the predominantly large numbers of compliments.

These ten predicted ratings are more useful in undermining informative values than the user-given one, at least in the range of the examples we have chosen. Thus, the proposed method can potentially identify the unreliable reviews by ‘learning’ from the unbiased samples, resulting in the elimination of the biased effect from different users. The increase in the prediction performance strengthens its accuracy to unveil those ‘unreliable’ reviews.

5. Discussion

Knowledge of rating quality in social media-related studies in the present work is limited. Hence, we focused on the publicly available dataset in Yelp whose enormous amounts of the reviews of customers provide direct evidence of suspicious review ratings. Methodologically,

Table 4
Examples of unreliable reviews depicted by our models.

No.	Review content (original text)	User	FC	CNN	LSTM
1	(Attached in Supplementary Material)	5	3	3	3
2	(Attached in Supplementary Material)	4	2	2	1
3	*The food was wonderful!! Unfortunately, the service was less than attentive. Seems that no server wanted a table of 5. The one who did get stuck with us paid as little attention as possible until she eased the check on the table without saying anything and took off again. Maybe we'll try some other time, but 5 people who feel ignored translated into lackluster tipping.*	4	2	2	2
4	*Store is great! But Laureen P. Would have to be the rudest cashier I've ever come across at any store! Had no interest in wanting to help and just seemed irritated to be at work. Just stay home! Otherwise, the store is my go to for my mid-day Tea!*	4	1	1	1
5	*(Attached in* Supplementary Material*)* *For one thing we got a free buffet voucher. Then we saw the stars given for this place and didn't care. The staff were very kind and gave us mimosas for $2 each. Decor is beautiful done it feels like we step on to a movie set.*	4	1	1	2
6	*Just wanted to say thank you Luxor for an amazing treat.* *P.S. my husband is a picky eat so you made him happy plus he tried some thing new.*	3	5	5	5
7	*Good food. But no lobsters. enjoyed the place and the dining experience was superb. would be coming back to this place when we visit Vegas again* *Very disorganized. The food is awesome as I've been here before but the wait and the disorganization does not make it worth it.*	2	5	5	4
8	*Better to make reservations at a delicious steak and seafood restaurant as Vegas has an abundance than stand in line for HOURS on a Thursday night!*	1	3	3	3
9	*i've been here many times. very good food. esp the pork bone soup and kimchi pancake! what i don't like is they actually ask for tips.*	1	4	4	4
10	*Their sandwiches are very delicious for $5.99. I got the Italian (we also got a* 10% *employee discount). It measure about* 8″ *long - it is more filling than it looks. It doesn't quite fill you up like a Subways Sandwich, but for $5.99, it's an amazing deal.* *The Italian is absolutely delicious. The juices are strong and the meat is fresh. It was served quickly and devoured even quicker.* *This place is an amazing place to eat dinner. Why spend $500 on a fancy dinner for two when the food is just as good here?*	3	5	5	5

User, rating given by user; FC, rating predicted by FC model; CNN, rating predicted by CNN model; LSTM, rating predicted by LSTM model.

the uniqueness of the present study lies in the use of a combination of three techniques in data mining and neural network domain in review rating prediction. Such techniques are POS word search, word vector representation and deep learning models. The objective is to identify biased user-given ratings on a scale unavailable in traditional online review-related studies. Firstly, a sequence of words with certain emotions was utilised to maintain meaningful interpretation and complete the comprehension of the review text in the semantic space. Secondly, word embedding provides a position reorganisation. Such reorganisation results in the aggregation of similar words and thus provides adequate 'features' to the classifier to distinguish those semantically different words. Finally, on the basis of layered and hierarchical structures, deep learning models transform the inputs into representations that are increasingly informative about the final result, similar to a multistage information-distillation operation. Features (feature maps) are learned from the word embeddings of keywords. Likewise, word embeddings are learned from keywords that are fed into the classifier. Unlike hardcoded features typically used in prior studies, features based on keywords and word embeddings provide a solution to make full use of the raw data (review text here). The combination of the holistic feature, the large quantities of data and processing power of deep learning techniques makes the rating prediction task becomes successful by reaching their full capabilities. Without which, classical methods could not have hypothesis spaces rich enough to learn discriminative features by themselves and the power to differentiate each class from the other four classes by these features. Through the above findings, the present study offers several important implications for research and practice.

5.1. Theoretical insights

Theoretically, the present study bridges previous studies focusing on the content analysis of online customer reviews (Berezina et al., 2016; Levy et al., 2013; Shin et al., 2019; Wang et al., 2019), studies focusing on customer perceptions from online reviews (Cheng and Jin, 2019; Hu et al., 2019; Kuhzady and Ghasemi, 2019; Ma et al., 2018; Xiang et al., 2015; Zhang and Cole, 2016) and information quality of online reviews (Antonio et al., 2018; Schuckert et al., 2016; Xiang et al., 2017). The contribution develops novel and meaningful insights into social media analytics in hospitality and tourism using big data analytics.

Firstly, whereas many studies of online reviews focus on valence, the present study concentrates on the less-studied characteristic: arousal level (Ren and Nickerson, 2019). Such an arousal level is the intensity of the emotions embedded in the review through its rating. Secondly, the study proposes an analytical framework for review rating prediction on a hospitality platform: Yelp. Although the accuracy achieved in this study is arguably data specific, rating prediction solely using textual review can potentially enrich our existing knowledge about the review quality of online data. Thirdly, the present study adds to the growing body of tourism research on online customer review by demonstrating how the impact of textual content contributes to the review rating. We cannot specify which words explicitly and exactly lead to a specific rating. However, certain existing cues or principles embedded in the text content actually carry a core message. The neural network can be drawn from such a message to help in the process of rating prediction. This observation is consistent with the findings of Xiang et al. (2015) in which topical factors generated from review texts were found highly correlated with a rating. Fourthly, the present study extends the prior studies with its application of ensemble learning (Antonio et al., 2018; Ma et al., 2018; Singh et al., 2017). Such an application is for verifying the effects of different characteristics of review (rating here) on review quality (review reliability here). Fifthly, the present study provides new insights into the current understanding of the incongruence between the review and user-given rating (Antonio et al., 2018; Schuckert et al., 2016; Xiang et al., 2017, 2015). The star rating briefly outlines the overall rate level of the review, and the review content is supposed to be a truthful reflection of one's evaluation of the business. The present study reveals that the satisfaction rating given by the majority of reviewers actually reflects the review content containing similar words. The rationale stems from the old adage, 'the minority is subordinate to the majority'. The deep learning model has the capability of learning the rules from most of the unbiased pairs (textual content associated with its corresponding rating) in training. Owing to this model, the consistency can therefore be maintained. Sixthly, the present study provides a way to quantify an automated assessment on how a biased rating is proven: the difference between the predicted and user-given rating above a pre-defined threshold (two here) would be regarded as suspicious. This method extends the results of prior research (Antonio et al., 2018) where such an assessment would be manually verified and validated. Lastly, we suggest an objective method for ascertaining whether an online review posted on a particular hospitality platform (Yelp) is unauthentic. In other words, the present study provides an effective approach of

uncovering the reviews with biased or perfunctory ratings that may induce misleading ideas to readers. The biased ratings (i.e. the minority) become evidence from those unbiased ones (i.e. the majority) by using the proposed model. As such, the present study is beyond focusing on the accuracy of the entire model. It attempts to draw additional value from those reviews erroneously detected by the model to lessen the biased effect.

5.2. Managerial implications

Discovering an accurate model to predict ratings based on Yelp reviews would be beneficial to consumers, business owners and platform managers alike. Firstly, consumers may encounter increasing difficulty in assessing which review ratings are trustworthy or whether suspicious rating occurs due to the limits on the human capacity to process information. Hence, the proposed model in the present study serves as a tool for quality inspection. Such a model, in turn, helps consumers identify truly helpful ratings, and it reduces the probability of consumers making any purchase errors. Secondly, these unbiased review ratings predicted by our model have significant business value. The current proliferation or widespread use of online reviews is not only an opportunity but also a threat to the business. Hence, business owners will benefit from this consistency between review content and its star rating. Furthermore, they will no longer worry about the contradictory behaviours of customers where, for example, a positive review content with a negative star rating would occur. More importantly, our proposed model can reduce the difficulty faced by business owners in a diagnostic assessment where any malicious review may materially affect sales. In fact, such an automatic prediction of review rating would assist in rooting out malicious reviews that are harmful to the business. Knowing when the feedback of a customer is significantly damaging might be useful to a business owner. The proposed model would provide the opportunity to intervene before a negative review (refers to rating scores 1 and 2) is to be posted. The business can leverage appropriate tactics in advance to manage the influence of these malicious reviews through a prompt managerial response. For instance, the reviewer could be directed to another page that contains a courtesy warning when a lower rating is detected. Such a page will show that a business might probably be ruined by malicious rating and appealing to the reviewer for a second thought. Thirdly, some social media platforms, such as Twitter and Facebook, allow users to write freeform textual reviews without specifying a star rating. In these cases, a review rating prediction provides a solution to quantify the review easily and a criterion for consumers to choose between similar products in handy. Fourthly, platform managers can partially prevent review manipulation in writing by using insights from the underlying framework here. For instance, one approach to maintain a fair and just environment is not to disclose exactly how the rating is calculated. If reviewers cannot take advantage of such disclosures, then weighing various indicators to derive an overall rating remains unknown to them. Thus, they cannot identify which word in the textual review exactly leads to a high score (in terms of predicted rating) upon review publishing. Platform managers or business owners can implicitly use a preferred evaluation system. Such a system highlights the aspect in which customer reviews would help form an accurate perception of the platforms or businesses, and reflect the strategy to maximise the effects on its targeted audience. Lastly, the present study opens the direction to the automatic verification for the business to recognise whether a user-given rating might have been posted fraudulently. Instead of requiring the business owners to go through each review and pinpoint possible suspicious reviews, the business could apply this model to automatically select which reviews could be problematic.

6. Conclusions, limitations and future directions

More than ever before, the decisions of people concerning where to visit or what to eat are subject to the opinions of other people. The aggregate online reviews on hospitality platforms play an important role in information dissemination. These reviews also explain the type of experience a future consumer could anticipate when purchasing a product from the business. However, owing to an overwhelming number of reviews, a user can only read a limited number of reviews before reaching a decision. In facilitating this process, the star rating is a common solution for the business to briefly outline the overall rate of the review. However, the rating given by the users may not be the ideal indicators of their perceived service quality and satisfaction. They may suffer from subjectivity and be slanted towards the personalities of the users or manipulated either by the business owners or competitors. The present study is far beyond the first attempt to conduct a quantitative study of suspicious online reviews in the hospitality industry based on five-class rating prediction. Nonetheless, this study represents a pioneering endeavour to tackle the rating reliability problem. The problem is tackled by using massive-scale data and describing a series of experiments with deep learning neural networks built on top of word vector embeddings. In addition, it is addressed by importing a complete verification (i.e. test dataset), elaborated assessments with fine-grained performance evaluation and justified matching rules based on classification. Experimentation exhibits good predictive power of our proposed model. Performance by CNN and LSTM also indicates that the sequence of the review words (keywords) contributes to the rating prediction. Keeping in mind that accuracy is not the only concern of this study, we purport a state-of-the-art definition of biased rating and unreliable review. We gain substantially high forecasting performance to distinguish those pairs that may cause inconsistency and may be misleading to readers.

Our study unveils certain intriguing findings on review reliability within user-given rating and initial experiments using this approach proved promising. However, certain existing limitations require consideration. Moreover, a room for growth exists in extrapolating even deep patterns from our data in the following ways. Firstly, the review text and its rating are remarkably not a perfect match than the traditional classification problem in which a clear boundary exists amongst different classes. In other words, the training samples are mixed with selected uncertainties towards informational value. The major challenge is the huge sample size here (million orders of magnitude). Ascertaining whether or not the rating is a genuine reflection of the review content via manual checking is difficult. Thus, the performance is limited in its power to reach a high level above 65% before being adjusted. How to purify the training samples to lessen the negative effect on the blend of raw data deserves further investigation. Secondly, the present study only uses emotional words in feature extraction. We also have topics, concerns and other relevant information in the raw data that affect the sentiment (Antonio et al., 2018). This information suggests the possibility of extracting node-level meta-data to incorporate into our model as an aid for learning programmes to fully understand the inputs. Whether employing other features to fuse into the model yields better results would be interesting. Thirdly, punctuation marks were removed, and words were lower-cased in our study. However, the exclamation marks, question marks and all-caps words may count as a rough measure of extreme emotion in the reviews. For instance, reviews with high volumes of exclamation sentences or all-caps tend to be overly emphatic and may lead to a negative rating. Similarly, reviews with too many question marks may not be positive to the star rating because these reviewers question the quality of the business. In addition, abbreviations should be transferred to its original forms (e.g. ‘isn’t’ to ‘is not’, ‘I’ve’ to ‘I have’) to further improve the accuracy of prediction. Some unintentional misspellings in writing also require careful checking (e.g. ‘writting’ instead of ‘writing’, ‘didnt’ instead of ‘didn’t’). Fourthly, our overall research design is based on a specific case (i.e. Yelp) using only a sample of reviews relevant to eating or drinking. Hence, the efficacy of the deep learning models requires testing in future research across different categories and review platforms to demonstrate its robustness and generalise the findings to other datasets. Lastly, our approach is

within the general assumption that the review content is a truthful reflection of the evaluation of a customer regarding the business. Hence, using deep learning models would be extremely important to possibly determine how we can prevent customers from posting fake or manipulated reviews. Nonetheless, the potential limitations do not reduce the internal validity of the proposed model to demonstrate the power of big data analytics using deep learning models in hospitality.

Acknowledgement

The authors express their sincere appreciations to Weimin Zheng of the School of Management, Xiamen University, for his helpful discussions and valuable suggestions, to Weiheng (Vincent) Liang of Linxoon (Guangzhou) Information Technology & Service Co., Ltd, for his technical assistance, and to Rui Li of the Department of Geography and Planning, SUNY at Albany, for his encouragement and critical comments without which this research could not have proceeded to its present form.

This work was partially supported by the Special Funds of High-level University Construction Program of Guangdong Province under Grant No. 88018052.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.ijhm.2020.102658.

References

Alaei, A.R., Becken, S., Stantic, B., 2019. Sentiment analysis in tourism: capitalizing on big data. J. Travel. Res. 58 (2), 175–191.
Antonio, N., de Almeida, A.M., Nunes, L., Batista, F., Ribeiro, R., 2018. Hotel online reviews: creating a multi-source aggregated index. Int. J. Contemp. Hosp. Manage. 30 (12), 3574–3591.
Bengio, Y., Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn. Res. 5, 1089–1105.
Berezina, K., Bilgihan, A., Cobanoglu, C., Okumus, F., 2016. Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. J. Hosp. Mark. Manage. 25 (1), 1–24.
Calheiros, A.C., Moro, S., Rita, P., 2017. Sentiment classification of consumer-generated online reviews using topic modeling. J. Hosp. Mark. Manage. 26 (7), 675–693.
Cheng, M., Jin, X., 2019. What do Airbnb users care about? An analysis of online review comments. Int. J. Hosp. Manag. 76, 58–70.
Chollet, F., 2017. Deep Learning With Python. Manning Publications, Shelter Island, NY.
Costa, A., Guerreiro, J., Moro, S., Henriques, R., 2019. Unfolding the characteristics of incentivized online reviews. J. Retail. Consum. Serv. 47, 272–281.
Duan, W., Yu, Y., Cao, Q., Levy, S., 2016. Exploring the impact of social media on hotel service performance: a sentimental analysis approach. Cornell Hosp. Q. 57 (3), 282–296.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.
Ganu, G., Kakodkar, Y., Marian, A., 2013. Improving the quality of predictions using textual information in online user reviews. Inf. Syst. 38 (1), 1–15.
Gavilan, D., Avello, M., Martinez-Navarro, G., 2018. The influence of online ratings and reviews on hotel booking consideration. Tour. Manag. 66, 53–61.
Gössling, S., Zeiss, H., Hall, C.M., Martinrios, C., Ram, Y., Grøtte, I.P., 2019. A cross-country comparison of accommodation manager perspectives on online review manipulation. Curr. Issues Tour. 22 (14), 1744–1763.
Gursoy, D., 2019. A critical review of determinants of information search behavior and utilization of online reviews in decision making process. Int. J. Hosp. Manag. 76, 53–60.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. Appl. 9 (8), 1735–1780.
Hong, Z.D., Jie, Z.Z., Ping, C.Y., Shichang, L., 2019. Good discounts earn good reviews in return? Effects of price promotion on online restaurant reviews. Int. J. Hosp. Manag. 77, 178–186.
Hu, N., Bose, I., Koh, N.S., Liu, L., 2012. Manipulation of online reviews: an analysis of ratings, readability, and sentiments. Decis. Support Syst. 52, 674–684.
Hu, N., Zhang, T., Gao, B.J., Bose, I., 2019. What do hotel customers complain about? Text analysis using structural topic model. Tour. Manag. 72, 417–426.
Kirilenko, A.P., Stepchenkova, S.O., Kim, H., Li, X., 2018. Automated sentiment analysis in tourism: comparison of approaches. J. Travel. Res. 57 (8), 1012–1025.
Kuhzady, S., Ghasemi, V., 2019. Factors influencing customers' satisfaction and dissatisfaction with hotels: a text-mining approach. Tour. Anal. 24 (1), 69–79.
Kwok, L., Xie, K.L., Richards, T., 2017. Thematic framework of online review research: a systematic analysis of contemporary literature on seven major hospitality and tourism journals. Int. J. Contemp. Hosp. Manage. 29 (1), 307–354.
Law, R., Li, G., Fong, D.K.C., Han, X., 2019. Tourism demand forecasting: a deep learning approach. Ann. Tour. Res. 75, 410–423.
Lei, X., Qian, X., 2015. Rating prediction via exploring service reputation. In: 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP). IEEE, Xiamen, China, pp. 1–6.
Levy, S.E., Duan, W., Boo, S., 2013. An analysis of one-star online reviews and responses in the Washington, D.C., lodging market. Cornell Hosp. Q. 54 (1), 49–63.
Li, J., Xu, L., Tang, L., Wang, S., Li, L., 2018. Big data in tourism research: a literature review. Tour. Manag. 68, 301–323.
Liu, X., Schuckert, M., Law, R., 2018. Utilitarianism and knowledge growth during status seeking: evidence from text mining of online reviews. Tour. Manag. 66, 38–46.
Liu, Y., Huang, K., Bao, J., Chen, K., 2019. Listen to the voices from home: an analysis of Chinese tourists' sentiments regarding Australian destinations. Tour. Manag. 71, 337–347.
Ma, Y., Xiang, Z., Du, Q., Fan, W., 2018. Effects of user-provided photos on hotel review helpfulness: an analytical approach with deep leaning. Int. J. Hosp. Manag. 71, 120–131.
Mellinas, J.P., Nicolau, J.L., Park, S., 2019. Inconsistent behavior in online consumer reviews: the effects of hotel attribute ratings on location. Tour. Manag. 71, 421–427.
Powers, D.M.W., 2011. Evaluation: from precision, recall and fmeasure to Roc, Informedness, markedness and correlation. J. Mach. Learn. Technol. 2 (1), 37–63.
Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. In: Liu, L., ÖZsu, M.T. (Eds.), Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 532–538.
Ren, J., Nickerson, J.V., 2019. Arousal, valence, and volume: how the influence of online review characteristics differs with respect to utilitarian and hedonic products. Eur. J. Inf. Syst. 28 (3), 272–290.
Schuckert, M., Liu, X.W., Law, R., 2016. Insights into suspicious online ratings: direct evidence from TripAdvisor. Asia Pacific J. Tour. Res. 21 (3), 259–272.
Shin, S., Chung, N., Xiang, Z., Koo, C., 2019. Assessing the impact of textual content concreteness on helpfulness in online travel reviews. J. Travel. Res. 58 (4), 579–593.
Sijoria, C., Mukherjee, S., Datta, B., 2019. Impact of the antecedents of electronic word of mouth on consumer based brand equity: a study on the hotel industry. J. Hosp. Mark. Manage. 28 (1), 1–27.
Singh, J.P., Irani, S., Rana, N.P., Dwivedi, Y.K., Saumya, S., Roy, P.K., 2017. Predicting the "helpfulness" of online consumer reviews. J. Bus. Res. 70, 346–355.
Starosta, K., Onete, C.B., Budz, S., Krutwig, M., 2019. Differences in travelers' perceptions of popular tourist destinations estimated by a LSTM neural network: a comparison between the UK and Germany. Tourism. 67 (4), 405–422.
Wang, X., Tang, L.R., Kim, E., 2019. More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness? Int. J. Hosp. Manag. 77, 438–447.
Wen, L., Liu, C., Song, H., 2019. Forecasting tourism demand using search query data: a hybrid modelling approach. Tour. Econ. 25 (3), 309–329.
Wijnhoven, F., Bloemen, O., 2014. External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews? Decis. Support Syst. 59, 262–273.
Xiang, Z., Schwartz Jr., Z., J.H.G, Uysal, M., 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? Int. J. Hosp. Manag. 44, 120–130.
Xiang, Z., Du, Q., Ma, Y., Fan, W., 2017. A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism. Tour. Manag. 58, 51–65.
Xiang, Z., Du, Q., Ma, Y., Fan, W., 2018. Assessing reliability of social media data: lessons from mining TripAdvisor hotel reviews. Inf. Technol. Tour. 18 (1-4), 43–59.
Xie, K.L., Zhang, Z., Zhang, Z., 2014. The business value of online consumer reviews and management response to hotel performance. Int. J. Hosp. Manag. 43, 1–12.
Xie, X., Ge, S., Hu, F., Xie, M., Jiang, N., 2019. An improved algorithm for sentiment analysis based on maximum entropy. Soft comput. 23 (2), 599–611.
Zhang, Y., Cole, S.T., 2016. Dimensions of lodging guest satisfaction among guests with mobility challenges: a mixed-method analysis of web-based texts. Tour. Manag. 53, 13–27.
Zhang, K., Chen, Y., Li, C., 2019. Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: the case of Beijing. Tour. Manag. 75, 595–608.
Zhang, B., Li, N., Shi, F., Law, R., 2020. A deep learning approach for daily tourist flow forecasting with consumer search data. Asia Pacific J. Tour. Res. 25 (3), 323–339.
Zhao, Y., Xu, X., Wang, M., 2019. Predicting overall customer satisfaction: big data evidence from hotel online textual reviews. Int. J. Hosp. Manag. 76, 111–121.