Extended use of online customer reviews for hospitality research: An application of imputation approach

Jewoo Kim*[1] and Jongho Im†[2]

[1]Department of Apparel, Events, and Hospitality Management, Iowa State University
[2]Department of Statistics, Iowa State University

October 31, 2017 (Submitted)

Abstract

Purpose - The purpose of this study is to introduce a new multiple imputation method that can effectively manage missing values in online review data, thereby allowing the online review analysis to yield valid results by using all available data.

Design/methodology/approach - This study adopts the multiple imputation chained equation (MICE) algorithm to generate imputed values for online reviews. Sentiment analysis is employed to incorporate customers textual opinions as the auxiliary information in the imputation procedures. To check the validity of this proposed imputation method, it is applied to missing values of sub-ratings on hotel attributes in both the simulated and real Honolulu hotel review datasets. Its estimation results are compared to those of different missing data techniquesnamely, listwise deletion and traditional multiple imputation which does not consider text reviews.

Findings - The findings from the simulation analysis show that our imputation method produces more efficient and less biased estimates compared to the other two missing data methods when text reviews are possibly associated with the rating scores and response mechanism. When applying our imputation method to the real hotel review data, the findings

*jjawoo@iastate.edu
†drive.jonghoim@iastate.edu

show that the text sentiment-based propensity score can effectively explain the missingness of sub-ratings on hotel attributes and our imputation method considering those propensity scores has better estimation results than the other methods as in the simulation analysis.

Originality/value - This study applies multiple imputation to online information considering its unstructured nature. This new method helps make the fuller use of the observed online data while avoiding potential missing problems.

Keywords - Online customer review, Missing data, Multiple imputation, Sentiment analysis, Unstructured data.

Paper type - Research paper.

1 Introduction

Recent technological advances enable consumers to share their experiences with goods or services through various online platforms, such as Yelp, TripAdvisor, and Expedia (Tirunillai and Tellis, 2014). For example, an average 390 million consumers visit TripAdvisor and submit more than 0.4 million reviews and opinions each month (TripAdvisor, 2017). As these user-generated contents (UGCs) are after-consumption customer opinions on products readily available to the public free of charge, they have been an important source of information in consumers pre-purchase information searches. Thus, UGCs have become in influential in formulating customer preferences and affect their purchase decisions (Baka, 2016; Hills and Cairncross, 2010; Kwok and Xie, 2016; Lee and Ro, 2016). According to Oxford Economics (2016), the contents on TripAdvisor in influenced 353 million trips, with $478 billion of travel spending, in 2014. Many researchers have also proved the effect of online reviews on firm performances such as room booking volume and restaurant sales (Kim et al., 2016; Viglia et al., 2016). In this regard, the in-depth analysis of online reviews is critical for understanding different customer preferences and overall business trends. The analysis results can be useful for hospitality firms that seek to improve customers' experiences and further have competitive advantages in their markets.

However, research on online reviews requires additional statistical techniques to address their spontaneous and unstructured nature for data analysis. One challenge in analyzing online reviews is to properly handle missing data. As online reviews are not created for

research, they are not structured in a pre-organized manner. Accordingly, missing values are likely to occur in most online review data, which leads to inconsistent or biased statistical results when applying regression-based analysis or classification (Ding and Simonoff, 2010; Horton and Kleinman, 2007; Saar-Tsechansky and Provost, 2007). For example, TripAdvisor allows users to evaluate not only their overall accommodation experience related to hotels, but also their satisfaction with specific attributes such as room, sleep quality, and service, along with a five-star scale. According to our investigation, 69% of Honolulu hotel reviewers have only partially (31%) or not at all (38%) evaluated these hotel attributes between January 2017 and the middle of October 2017. Such a considerable volume of missing values can invalidate analysis results and further severely limit the use of online information. Therefore, the adequate handling of missing values is essential for online information analysis.

Two approaches are primarily employed to handle incomplete data with missing values. A naïve approach is deletion under which observations with missing values are excluded from the data analysis. The other approach is to implement imputation that replaces missing values with plausible alternative values. In practice, imputation methods are preferred because deletion is inefficient and may cause large biases in the model parameter estimates (Buhi et al., 2008). Among imputation approaches, multiple imputation, initially proposed by Rubin (1978), is the most popular in both social science and biomedical science due to its statistical and practical advantages.

Under multiple imputation, multiple versions of complete datasets are generated replacing each missing value with a set of imputed values. The analysis results with all these multiply completed datasets are then simply combined to compute overall estimates and their standard errors. Multiple imputation yields precise estimates and accurate standard errors that can help obtain less biased results than when using single imputation (Schafer and Graham, 2002). These advantages make multiple imputation one of the best options for handling missing data (Schlomer et al., 2010). As many statistical packages have already been developed for implementing multiple imputation on the incomplete data, non-statisticians can easily handle missing data and then conduct statistical analyses on the imputed datasets. Given that online information is likely to have a considerable portion of missing values and be able to provide different forms of auxiliary data useful for creating alternative imputed

values, multiple imputation is of great importance for online review research. However, the unstructured characteristic of online data makes it difficult to use existing multiple imputation methods requiring additional procedures to create structured auxiliary data for multiple imputation. To the best of our knowledge, there have been no studies that focus on the application of multiple imputation to incomplete online review data.

To fill this research gap, we aim to introduce a new multiple imputation method for online review research that incorporates customer reviews including both structured opinions shared via sub-ratings on specific product attributes and unstructured opinions in the form of text. This imputation adopts sentiment analysis to convert each review into a latent propensity score, and in turn, the latent propensity score and sub-ratings are used to generate imputed values. To validate the proposed imputation method, we conduct two different analyses: Monte Carlo simulation analysis and real data analysis. First, the simulation analysis compares our imputation method to a deletion method and another imputation method in three different missing mechanisms. Second, we apply the proposed imputation method to actual online reviews on hotels in Honolulu, Hawaii in order to investigate the robustness of our method in practice.

This study can contribute to hospitality research by proposing a new imputation method that can be applied to unstructured online data. This imputation method can reduce inconsistency and bias in analysis results on online data compared to the deletion and other imputation methods. It also allows hospitality researchers to investigate diverse topics discussed and shared through UGC that are otherwise thrown away due to many missing values.

2 Literature Review

2.1 *Online Customer Reviews*

With the development of technological capability to digitize a large amount of data, various forms of online information about business products or services are created and shared by customers. Such online information primarily interests hospitality customers in influencing their preferences and decisions. The growing importance of online information forces hospitality managers to develop different and more effective approaches to interactive electronic

word of mouth (eWOM). The research on UGCs might support the development of new eWOM strategies. Following this trend, hospitality studies have focused on the impact of online reviews on customer and corporate behavior. For example, Levy et al. (2013) analyzed hotels' responses to poor reviews using content analysis. Casaló et al. (2016) investigated the relationship between online hotel ratings and customers attitude toward hotels and booking intentions adopting an experimental design. However, these studies might have limitations in conducting their investigation in a broader setting due to the unstructured characteristics of online reviews.

The recent development of text mining, sentiment mining, and web-scraping techniques has enabled researchers to transform unstructured online information into a structured form suitable for data analysis and further apply a big data analysis to online review studies. Geetha et al. (2017) conducted a sentiment analysis of online reviews for Indian hotels and investigated the relationship between customers' sentiments and their ratings of the hotels. The findings showed that customers' feelings about hotels were consistent with their ratings of the hotels. Using an automated web scrapper, Xiang et al. (2015) investigated 60,648 online reviews for 10,537 hotels in the U.S. and found that guest experiences related to deals, family friendliness, core product quality, and staff affected hotel customers satisfaction. These data transformation and collection techniques have enabled the analysis of a huge amount of diverse online information in both consumer and industry-level research.

Although the unstructured nature of online information can be addressed using data transformation techniques, online review studies need to deal with another challenge: missing values (Xiang et al., 2015). Online reviews are spontaneously produced by customers, and many reviews have incomplete information for data analysis. As previously mentioned, not all hotel reviewers on TripAdvisor provide their satisfaction scores for specific attributes. This missing data problem can be much more serious when investigating text reviews because the texts are likely to contain information about only several specific attributes that significantly satisfy or dissatisfy customers. This creates an online information paradox. Digitized online information allows researchers to access a large volume of valuable information in a complete form for analysis while they also have another large chunk of incomplete information that should be dealt with; otherwise, information loss results in the invalidation

of analysis results or makes it difficult to use the information itself. Thus, online review studies need to address the potential problems with missing values.

2.2 *Missing Data and Deletion*

We begin by introducing missing data mechanisms to discuss missing data techniques, including deletion and imputation. According to Little and Rubin (2002), missing data can be classified into three categories: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) not missing at random (NMAR). The MCAR holds when the missingness of variables is independent of the missing variables themselves and all other variables examined in a study. Under the MAR condition, the missing mechanism is related only to observed variables, not missing variables. Note that the MCAR is a special case of MAR. For the NMAR, missing data are non-random, and the missing mechanism depends on the missing variables themselves. Consequently, it is necessary to estimate the response model and the outcome model jointly (Ibrahim et al., 2005). However, as the joint modeling of the missing mechanism and the outcome model is often unfeasible or sometimes not tractable, the majority of research has focused on the cases where the missing data are MAR or MCAR (Graham, 2009).

One of the widely used missing data techniques is deletion, which excludes cases with missing values from the original dataset. There are two types of deletion: listwise deletion and pairwise deletion. Under listwise deletion, entire observations containing one or more missing values are removed from the sample. This method, called complete case analysis, is relatively simple to implement; as a result, it is frequently used as the default option for many statistical software packages (e.g., SPSS) (Buhi et al., 2008). In hospitality academia, listwise deletion is a popular method for handling missing values. Among survey-based studies published in the *International Journal of Contemporary Hospitality Management* between 2011 and 2016, 44.6% used listwise or pairwise deletion (see Table 1). However, because listwise deletion only considers complete observations in analysis, it can increase Type II error, lowering statistical power due to the smaller sample size (Graham, 2009). In pairwise deletion, known as available information analysis, all available observations are retained in the sample, and the variables with missing values in the observations are excluded

Table 1: Missing data techniques of survey-based studies, 2011-2016

Year	Studies	Averaged Sample Size	Incomplete responses	Missing rate	Deletion		Imputation	Not Indicated
					Listwise	Pairwise		
2011	19	580.6	29.2	4.8 %	8	1	0	10
2012	24	305.5	21.2	6.5 %	14	0	0	10
2013	25	396.0	28.7	6.8 %	8	1	1	15
2014	36	410.7	47.3	10.3%	15	0	0	21
2015	55	492.6	25.8	5.0 %	20	0	1	34
2016	72	539.7	23.6	4.2 %	34	2	1	35
Total	231	467.4	28.6	5.8 %	99	4	3	125

from estimation, minimizing information loss that may arise in listwise deletion (Hippel, 2004). Notwithstanding, pairwise deletion cannot be perfectly free from a problem of low statistical power because of the exclusion of incomplete observations. In addition, the use of different set of variables makes it difficult to apply the correlation or covariance matrix to multivariate analysis under pairwise deletion. Both deletion methods have another weakness; they are valid only when the data are MCAR and produce biased estimates when the data have MAR missing patterns (Buhi et al., 2008; Pigott, 2001). Moreover, such weaknesses of biased estimates and low statistical power can be consequential if missing values account for more than 5% of the total observations (Graham, 2009). As a result, deletion is not recommended for online review studies which are highly likely to deal with the data including more than 5% of missing values.

2.3 *Imputation*

Another missing data technique is imputation, which involves substituting plausible values for missing values. Compared to the deletion approach, imputation methods fully use observed data and thus help to improve statistical power (Schafer and Graham, 2002). In addition, estimates are generally consistent under the MAR assumption and even under the NMAR assumption if the response model is correctly specified. See Little and Rubin (2002) for a detailed discussion of complete case analysis (or listwise deletion) and multiple imputation.

Imputation can be broken down into two categories: single imputation and repeated imputation. In single imputation, a single plausible value is imputed into each missing value, whereas multiple values are assigned in repeated imputation. It is very difficult to

capture the imputation variance with single imputation because the imputed values are treated as known in data analysis (Rubin and Schenker, 1986). Due to this limitation, repeated imputation is preferred for general purpose estimation (Rubin, 1987). There are two types of repeated imputation: multiple imputation (Rubin, 1978, 1987; van Buuren et al., 1999) and fractional imputation (Fay, 1996; Kim and Fuller, 2004; Im et al., 2015). Multiple imputation is based on Bayesian estimation and generates multiply imputed datasets by filling in each missing value with one plausible value repeatedly. Meanwhile, fractional imputation constructs a single complete dataset that has multiple plausible values with different fractional weights for each missing value, which generally utilizes the expectation-maximization (EM) estimation method. Multiple imputation is more advantageous than fractional imputation in that the variance estimation process is simple and easy to implement compared to fractional imputation which requires advanced computation techniques (Im and Kim, 2017). In addition, from a practical standpoint, fractional imputation has not yet been widely used due to the lack of statistical packages offering this imputation function. We therefore focus on multiple imputation in this study.

A popular multiple imputation method for multivariate missing data is multiple imputation chained equation (MICE), proposed by van Buuren et al. (1999). MICE handles multivariate missing variables through the conditional imputation models, which are independently specified for each missing variable. In the MICE algorithm, a Gibbs sampler is often applied to generate the joint distribution for the missing variables. However, the MICE algorithm has limitations when applied to the unstructured data, such as text or image. Thus, this paper proposes a new MICE modified for unstructured online data.

3 Methodology

In this study, we develop a new multiple imputation method that can handle multivariate missing variables in online information. Specifically, we apply this imputation method to missing values of sub-ratings on product attributes in online reviews. Based on the sentiment analysis of text reviews and observed sub-ratings, our method creates multiply imputed values for missing values of sub-ratings.

3.1 *Basic Setup*

Let X be the overall rating score, Y_k be the kth attribute sub-rating score, $k = 1, \ldots, K$, and W be the customer's review. Assume that X and W are observed over the entire dataset but Y_k is subject to missingness. Now let R_k denote as a set of indicators of the missingness of Y_k. R_k takes the value of one if Y_k is observed and zero otherwise. Let Y_{obs} and Y_{mis} be the observed/missing values of $Y = (Y_1, \ldots, Y_k)$.

We assume MAR missing mechanism such that

$$P(R \mid Y, X, W) = P(R \mid X, W), \tag{1}$$

where $R = (R_1, \ldots, R_K)$. Thus, the missing values can be generated without specifying the response model.

3.2 *Construction of a latent variable Z*

Consider the predictive distribution of Y conditional on X and W as the imputation model,

$$P(Y \mid X, W). \tag{2}$$

One simple method for generating the imputed values for Y is to use only X because W is unstructured, making it difficult to estimate the predictive distribution (2). However, this method may cause efficiency loss or biased estimation results when the sub-ratings are strongly correlated with W given X. To avoid these problems, it needs to consider W, which makes the fuller use of the observed review data.

For this purpose, we need to structure the unstructured W by creating a latent variable Z that can substitute W. This property of W is formulated as follows,

$$P(Y, X \mid W, Z) = P(Y, X \mid Z), \tag{3}$$

where $Z = g(W)$ is a numerical variable constructed from W through a converting mechanism $g(\cdot)$. Text clustering, scoring analysis, or similar statistical methods can be applied to obtain Z (Aggarwal and Zhai, 2012; Liu, 2014). Since Condition (3) is not testable in practice, we generally assume that the newly created variable Z from the converting algorithm will be sufficient statistics of W, that is, Condition (3) is satisfied.

The latent variable Z could be continuous *or* categorical, but we want this latent variable Z to satisfy the following desirable property,

$$P(Y, X, R \mid Z = z_1) \neq P(Y, X, R \mid Z = z_2) \text{ for } z_1 \neq z_2. \tag{4}$$

Condition (4) implies that the latent variable Z is informative in explaining the joint distribution of (Y, X, R). Condition (4) can be decomposed into two properties as follows:

$$P(Y, X \mid Z = z_1) \neq P(Y, X \mid Z = z_2) \text{ for } z_1 \neq z_2, \tag{5}$$

$$P(R \mid X, Z = z_1) \neq P(R \mid X, Z = z_2) \text{ for } z_1 \neq z_2. \tag{6}$$

If at least one of two properties is satisfied, we can conclude that Z satisfies Condition (4). If Z fails to meet Conditions (5) or (6), the imputation estimates may have larger standard errors than the estimates obtained using a simpler model $P(Y \mid X)$. On the other hand, if Z satisfies one of two conditions, adding an extraneous variable will not lead to additional biases, and the predictive distribution (3) will provide more robust results compared to when using a simpler model $P(Y \mid X)$ which is possibly exposed to omitted variable problem. Thus, it is important to check if the created variable Z follows the required condition (5) or (6) to validate new imputation method.

We propose a new algorithm that does not require any training procedure or similarity score computation to construct a latent variable Z for our imputation. First, we select key feature words based on the parts of speech (POS) in text reviews and then compute reviewers' propensity scores based on sentiment analysis. The proposed method essentially uses a lexicon-based scoring algorithm that has the operation mechanism similar to Pand et al. (2002)'s classification on the movie reviews. Since reviewers' subjective opinions are well represented through specific POS such as adjective, adverb, and verb, we extract these POS and then make scores on these selected features. The detailed algorithm is given below:

(L.1) Assign POS on words within each text review using the natural language processing (NLP) toolkit.

(L.2) Select key feature words categorized to verb, adverb and adjective.

(L.3) Assign one of three values $\{positive, neutral, negative\}$ on each key word based on a dictionary of sentiment words, called a sentiment lexicon.

(L.4) Compute a propensity score for review i,

$$S_i = \frac{C_{pos,i} + C_{neg,i}}{\alpha + C_i}, \tag{7}$$

where $\alpha(> 0)$ is a tunning parameter, C_i is the number of key feature words, and $C_{pos,i}$ and $C_{neg,i}$ are the number of positive/negative key feature words.

The tunning parameter α in (L.4) adjusts the normalized degree. For short texts, the propensity scores will be more easily influenced by the sentiment of one key feature word than those of long texts. Thus, this tuning parameter is used to adjust the balance between short and long text reviews.

This procedure of constructing Z requires text mining knowledge. There are many packages in **R** or **Python** that provides the proposed scoring algorithm. Google SyntaxNet and NLP toolkit can be used to determine a POS for each word in text reviews. Bird et al. (2009), Hornik (2016), and SyntaxNet (2016) demonstrate how one can use Google SyntaxNet and NLP toolkit to obtain POS for key feature words with their open sources.

3.3 *Multiple Imputation*

Based on the predictive model (3), we employ widely used MICE algorithm as a multiple imputation method. This imputation alogrithm is offered by several statistical packages including **SAS**, **SPSS**, and **R**. See van Buuren and Groothuis-Oudshoorn (2011) for details of the MICE algorithm. A modified imputation procedure is summarized below:

(S.1) Convert W to a latent variable Z using sentiment analysis on a text review.

(S.2) Specify the imputation model for each missing variable with the fully observed data,

$$P(Y_k \mid Y_{-k}, X, Z, \delta = 1; \hat{\theta}_k^{*(0)}), \text{ for } k = 1, \ldots, K,$$

where Y_{-k} denotes the collection of Y except Y_k, $\delta = \prod_{k=1}^{K} R_k$, and $\hat{\theta}_k^{(0)}$ denotes the parameter estimates of imputation model for the k-th missing variable on the complete cases.

(S.3) For given imputed values $Y^{*(t-1)} = (Y_{obs}, Y_{mis}^{*(t-1)})$, the t-th iteration of MICE algorithm is a Gibbs sampler that sequentially generates

$$\begin{aligned}
\hat{\theta}_1^{(t)} &\sim P(\theta_1 \mid Y^{*(t-1)}, X, Z) \\
Y_1^{*(t)} &\sim P(Y_1 \mid Y_{-1}^{*(t-1)}, X, Z, \hat{\theta}_1^{(t)}) \\
\hat{\theta}_2^{(t)} &\sim P(\theta_2 \mid Y_1^{*(t)}, Y_{-1}^{*(t-1)}, X, Z) \\
Y_2^{*(t)} &\sim P(Y_2 \mid Y_1^{*(t)}, Y_3^{*(t-1)}, \ldots, Y_K^{*(t-1)}, X, Z, \hat{\theta}_2^{(t)}) \\
&\vdots \\
\hat{\theta}_K^{(t)} &\sim P(\theta_K \mid Y_{-K}^{*(t)}, Y_K^{*(t-1)}, X, Z) \\
Y_K^{*(t)} &\sim P(Y_K \mid Y_{-K}^{*(t)}, X, Z, \hat{\theta}_K^{(t)})
\end{aligned}$$

(S.4) Iterate (S.3) for large enough t until we have convergence.

(S.5) Independently repeat (S.3) and (S.4) for $M(>1)$ times so that we create M imputed datasets as the final imputation output.

We now consider an imputation estimator constructed from the multiply competed datasets. Let Q be an estimand defined with a known link function h, where $Q = h(Y, X)$. On the multiply completed datasets, the imputation estimator of Q is

$$\hat{Q} = M^{-1} \sum_{m=1}^{M} \hat{Q}_m = M^{-1} \sum_{m=1}^{M} h(Y^*, X)$$

where Y_m^* denotes a set of imputed values of the m-th completed dataset, and $\hat{Q}_m = h(Y_m^*, X)$ is the estimates obtained from the m-th completed dataset. The variance of imputation estimator can be estimated using the(Rubin, 1987)'s variance formula, where

$$\hat{V}(\hat{Q}) = \hat{A} + \frac{M+1}{M} \hat{B},$$

with the average of within-imputation variances,

$$\hat{A} = M^{-1} \sum_{m=1}^{M} \hat{V}(\hat{Q}_m),$$

and the between-imputation variance,

$$\hat{B} = (M-1)^{-1} \sum_{m=1}^{M} (\hat{Q}_m - \hat{Q})(\hat{Q}_m - \hat{Q})^T.$$

Table 2: Marginal distribution of Variables

		Value				
Case	Variable	1	2	3	4	5
$Z = 1$	X	0.3	0.2	0.2	0.2	0.1
	Y_1	0.3	0.2	0.2	0.2	0.1
	Y_2	0.3	0.2	0.2	0.2	0.1
$Z = 2$	X	0.1	0.2	0.2	0.2	0.3
	Y_1	0.1	0.2	0.2	0.2	0.3
	Y_2	0.1	0.2	0.2	0.2	0.3

4 Simulation Study

We conduct a simulation analysis to evaluate the performance of the proposed multiple imputation. In this simulation analysis, the overall score variable X and two sub-rating scores Y_1 and Y_2 are considered. Our imputation model uses text reviews to generate imputed values, but because it is very difficult to create artificial text data, we directly generate a propensity score Z that satisfies Condition (4).In other words, we presume that the text data are already converted to a propensity score variable which is a categorical data ($Z = 1, 2$) in this simulation analysis. The second group of Z ($Z = 2$) has the probability of having more positive rating scores than the first group ($Z = 1$).

To generate the dataset for the simulation, we adopt a random sampling algorithm based on Gaussian copula proposed by Ferrari and Barbiero (2012), which helps achieve the given correlation and marginal distribution of multivariate categorical variables. First, the variable Z is generated from a binomial distribution with the probability of 0.5 and the size of $n = 200$; three random variables of X, Y_1, and Y_2 are then separately generated within each group of Z. A **R** package **GenOrd** (Barbiero and Ferrari, 2015) is used to generate random samples of the three rating variables which have the same marginal distribution with the mean value of 3. The marginal distributions of X and Y conditional on Z are presented at Table 2. The correlation matrix of X, Y_1, Y_2, and Z is assumed to be

$$\Sigma = \begin{bmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}.$$

In this simulation analysis, we posit that missing values of Y_k can have three different

missing mechanisms as follows:

$$\text{Model 1} : P(R_k = 1) = 0.5,$$

$$\text{Model 2} : P(R_k = 1) = \begin{cases} (1 + \exp(0.5 - 0.1X))^{-1} & \text{if } Z = 1, \\ (1 + \exp(0.3 - 0.2X))^{-1} & \text{if } Z = 2, \end{cases}$$

$$\text{Model 3} : P(R_k = 1) = \begin{cases} 0.4 & \text{if } Z = 1, \\ 0.6 & \text{if } Z = 2. \end{cases}$$

In the case of Model 1, missing values of Y_k are randomly determined from a binomial distribution with the response probability of 0.5 indicating that randomly 50% of reviewers do not respond to sub-ratings. Models 2 and 3 assume missing at random mechanisms. While Model 2 depends on both X and Z, Model 3 only depends on the variable Z. In Models 2 and 3, the response rate is lower in the first group reviewers ($Z = 1$) whereas the second group reviewers ($Z = 2$) are more likely to respond to sub-ratings to evaluate product attributes than the first group reviewers ($Z = 1$). The marginal response rate is 0.5 for Models 1 and 3, and about 0.52 for Model 2.

In these three models, the simulation analysis investigates three missing data techniques including complete case analysis (CC) (or listwise deletion) and two multiple imputations: 1) multiple imputation without using the variable Z (MI-NZ), and 2) multiple imputation using the variable Z (MI-WZ). The $M(= 5)$ imputed datasets are created using the 'mice' function included in the **R** package **mice** (van Buuren and Groothuis-Oudshoorn, 2011).

Table 3: Monte Carlo biases (Bias) and standard errors (S.E.)

		CC		MI-NZ		MI-WZ	
Model	Parameter	Bias	S.E.	Bias	S.E.	Bias	S.E.
Model 1	$E(Y_1)$	0.00	0.15	0.00	0.11	0.00	0.10
	$E(Y_2)$	0.00	0.16	0.00	0.11	0.00	0.10
Model 2	$E(Y_1)$	0.20	0.14	0.03	0.11	0.00	0.11
	$E(Y_2)$	0.19	0.14	0.03	0.11	0.00	0.10
Model 3	$E(Y_1)$	0.16	0.14	0.04	0.11	0.00	0.11
	$E(Y_2)$	0.16	0.14	0.05	0.10	0.00	0.10

Table 3 the biases and standard errors of mean estimates with 1,000 Monte Carlo sam-

ples. In the case of Model 1, all three estimators of CC, MI-NZ, and MI-WZ produce nearly unbiased estimates. However, in the cases of Models 2 and 3, the CC estimator produces severely biased estimates, and MI-NZ estimates are also slightly biased. The biases of MI-NZ estimates can be understood as the omitted variable problem or response model misspecification. On the other hand, the MI-WZ estimator performs well over three different response models. In terms of standard errors, the MI-WZ estimator is the most efficient because it uses all available information during the imputation procedure. These simulation results imply that our multiple imputation using the available text review data will produce robust and efficient estimates when the joint distribution of Y, X, and R depends on the variable Z as Condition (4). Note that even if the latent variable Z does not satisfy Conditions (3) and (4), we only have to pay a small amount of efficiency loss compared to the conventional imputation estimator of MI-NZ.

5 Empirical Analysis

5.1 *Data*

We also test a new imputation method with the actual review data on hotels in Honolulu, Hawaii. Using the web scraping **R** package **rvest**, this study directly extracts all available review texts and relevant ratings from TripAdvisor. We collect 15,330 hotel reviews for 82 hotels between January 2017 and the middle of October 2017 in order to illustrate the proposed imputation method. The collected information includes the overall rating scores (X) and text reviews (W), six sub-rating scores: value (Y_1), location (Y_2), sleep quality (Y_3), rooms (Y_4), cleanliness (Y_5), and service (Y_6). Table 4 presents an example of review data. For brevity, posting dates and other miscellaneous information are removed from the table's representation.

Approximately 27% of the reviews (4,070 reviews) have the sub-rating scores on all hotel attributes while the remaining 73% have partial or no information on the sub-ratings. There is no missing value in the overall rating scores and text reviews, and the structured and unstructured information is used to create imputed values for missing sub-ratings. Table 5 presents the number of reviews categorized by the responses of six sub-ratings. The dominant

Table 4: An illustrative example of TripAdvisor hotel review data

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Part of W
4	NA	NA	NA	NA	NA	NA	$\cdots$ treated us like we were not valued customers. No upgrades and no add ins at all!
2	1	NA	NA	2	2	22	this over priced under serviced "resort" is terrible. $\cdots$ Very bad experience.
5	3	5	NA	4	NA	4	It's a great location $\cdots$ The fireworks show was great too.
4	3	NA	4	4	NA	4	$\cdots$ beautifully landscaped with a variety of pools $\cdots$ always a long line for guests waiting to check in.

patterns are $(0,0,0,0,0,0)$ and $(1,1,1,1,1,1)$, indicating that more than half of reviewers are likely to answer all or none at all. The marginal response rates of the six sub-ratings are all at about 43%, except for service (68%), thereby showing that customers are likely to report their evaluation on service rather than other attributes.

Table 5: Frequencies by response patterns

Response Patterns	Frequency (proportion)
(1,1,1,1,1,1)	4,059 (26%)
(0,0,0,0,0,0)	4,918 (32%)
(0,1,0,1,0,1)	657 (4%)
(0,1,0,1,1,1)	640 (4%)
(0,0,1,1,0,1)	639 (4%)
(1,0,0,1,1,1)	630 (4%)
(0,1,1,1,0,1)	610 (4%)
(0,0,0,1,1,1)	609 (4%)
(1,1,0,1,0,1)	609 (4%)
(0,0,1,1,1,1)	607 (4%)
(1,0,0,1,0,1)	593 (4%)
(1,0,1,1,0,1)	588 (4%)
Others	171 (2%)

$(r_1, r_2, r_3, r_4, r_5, r_6)$ denotes the of response of R_1, R_2, R_3, R_4, R_5 and R_6.

5.2 *Imputation for hotel review data*

We apply the proposed multiple imputation to the Honolulu hotel review data in 2017. Using text reviews, we create a latent variable Z that represents propensity of reviewers so that it explains the joint distribution of Y, X and R. The propensity score Z is created using the sentiment-based propensity scoring algorithm proposed in Methodology with a tuning

parameter $\alpha = 1$ and a dictionary of sentiment words included in **R** package **tidytext** (Silge and Robinson, 2016). The mean value of positive propensity scores is 0.31, and their standard deviation is 0.16. We fit the overall response indicator, defined by $\delta = \prod_{k=1}^{6} R_k$, into the overall rating X and the created score Z with a logistic regression:

$$P(\delta = 1 \mid X, Z) = \{1 + \exp(-\phi_0 - \phi_1 X - \phi_2 Z)\}^{-1}.$$

Table 6: Logistic regression result of δ given X and Z

Coefficient	Estimate	Standard Error	P-value
ϕ_0	-1.874	0.086	0.000
ϕ_1	0.047	0.019	0.016
ϕ_2	2.009	0.121	0.000

Table 6 presents the regression estimates, standard errors, and p-values obtained from the logistic regression of δ given X and Z. The regression coefficient ϕ_2 associated with the latent variable Z is statistically significant in the sense that Condition (6) is satisfied in this sample data, implying that the latent variable Z will be informative to identify reviewers' response behavior to sub-ratings.

We also investigate three estimators: CC, MI-NZ, and MI-WZ. For two multiple imputation estimators, $M = 30$ completed datasets are generated. The imputation estimates and their standard errors are computed using the formula presented in the Methodology section.

Table 7: Mean estimates and standard errors of three estimators: CC, MI-NZ, and MI-WZ

	CC		MI-NZ		MI-WZ	
Parameter	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
$E(X)$	4.30	0.015	4.22	0.008	4.22	0.008
$E(Y_1)$	4.19	0.016	4.09	0.015	4.09	0.011
$E(Y_2)$	4.62	0.011	4.61	0.010	4.60	0.009
$E(Y_3)$	4.31	0.016	4.26	0.014	4.26	0.012
$E(Y_4)$	4.24	0.016	4.17	0.012	4.17	0.011
$E(Y_5)$	4.44	0.014	4.41	0.013	4.41	0.011
$E(Y_6)$	4.46	0.015	4.40	0.011	4.40	0.009

Table 7 presents mean estimates with standard errors for the fully observed X and partially observed Y_k, $k = 1, \ldots, 6$. Given a significance level of 0.05, the 95% confidence interval (CI) of the CC mean estimate for X does not include the full sample mean of 4.22

(the imputation estimates): 95% CI lower limit 4.27 = 4.30 - (0.015*1.96), 95% CI upper limit 4.33 = 4.30 + (0.015*1.96). This significant difference in mean estimates between the CC and two imputation estimators indicates that the overall response indicator depends on the overall rating scores X, implying that the reviewers who fully respond to all sub-ratings have different overall rating scores from those of the reviewers who partially respond to sub-ratings. The systematic difference between these two reviewer groups can lead the CC estimator to yield biased results from the true values because the missing data technique completely ignores the partially responding reviewers who produce missing values in the data analysis. In a similar vein, the statistical difference in the mean estimates of Y_1, Y_3, Y_4, and Y_6 between the CC and imputation estimators also provide evidence of supporting biased results from the CC estimator.

Two multiple imputation estimators, MI-NZ and MI-WZ, have smaller estimated standard errors than the CC estimator, indicating that these imputation methods provide more efficient estimates than the CC estimator. Because the partial information is incorporated during the imputation, we may have efficient results in both imputation estimators. Furthermore, the MI-WZ estimator has more efficiency than the MI-NZ estimator according to our results. This implies that the text reviews are well summarized to the latent score Z which brings efficiency gains to our imputation mean estimator.

6 Discussion and conclusions

In consumer science, as one applied discipline of behavioral sciences, many studies including hospitality ones have investigated the data with missing values (Allison, 2001). It is thus recommended for researchers to report missing patterns and statistical methods used to handle missing problems (Schlomer et al., 2010). The report of missing data analysis has often been omitted in hospitality research mainly because the volume of missing values has been so low (less than 5%-10% of the total observations) that missing problems can be readily managed by deletion methods. However, missing problems may become much more serious for research on online reviews, which are often incomplete and unstructured. Thus, this study develops a new multiple imputation method that can effectively address the missing problems of online review data. Based on text reviews and observed sub-ratings on

attributes, we create a latent scores used for generating imputed values that replace missing values. To check the validity of this imputation method, we conduct Monte Carlo simulation and empirical tests and find that our imputation method yields more efficient and robust results compared to the listwise deletion method and the conventional imputation method.

6.1 *Theoretical implications*

One theoretical contribution of this study is the introduction of a new multiple imputation method for incomplete online information. To pre-process online information for research, many analytics studies have focused heavily on transforming unstructured online information into the structured data for analysis. They include topic extraction based on probabilistic latent semantic indexing and latent Dirichelet allocation (Blei et al., 2003; Hofmann, 1999) and sentiment classification using semantic analysis based on NLP techniques and N-gram analysis(Nasukawa and Yi, 2003). However, despite the significant impact of missing values on estimates' efficiency and biases, few studies have looked at the adequate handling of missing values of online information. This study is the first to discuss the details of how multiple imputation can be applied to incomplete online review data in order to address estimation problems.

Another contribution of this study is to help researchers and practitioners apply a big data approach to the analysis of various online reviews. Advantages of using online reviews include unlimited volume, easy access, real-time data, diverse topics, and spontaneous opinions from users Philander and Zhong (2016); Tirunillai and Tellis (2014) These advantages have generated considerable interest in the use of online reviews in academia and industries. However, unstructured online reviews are likely to produce a number of missing values that may limit the availability of online review data, thereby trading off their advantages. To address potential missing problems, we develop a new imputation model for online reviews that allows the fuller use of online reviews on a broader range of topics by using all observed data regardless of their completeness.

Moreover, this studys findings help easily apply a new imputation method to online review data by providing the information of used **R** packages and data mining techniques as well as the relevant studies that present the user with guidelines on the statistical and data

processing functions. We also provide the R code used to generate multiple imputed values for our proposed imputation (see online appendix at http://github.com/TAB-Research/Imputation). Such technical information facilitates the ease and convenient use of the proposed imputation method to manage missing values in online reviews properly. Furthermore, it can help imputation users generate valid analysis results.

In terms of missing data techniques, this study has a couple of implications. First, this study demonstrates how to use customers' textual opinions in handling the missing values of online review data. Text reviews are used as the auxiliary information for missing values of sub-ratings on hotel attributes. Our findings show that the text reviews are effective in explaining the missing values of sub-ratings, and our imputation method utilizing the text reviews outperforms other missing data techniques in terms of efficiency and biasness. Second, this study suggests the use of sentiment analysis to generate auxiliary information in a suitable form for data analysis. To transform unstructured texts into the auxiliary information, we propose a new propensity scoring algorithm based on sentiment analysis and use the resultant propensity scores as the basis for imputed values. Our findings prove the importance of text sentiment as auxiliary information in imputation procedures. This propensity scoring can be applied to creating imputed values for a variety of overall ratings on products and sub-ratings on product attributes in online review data.

6.2 *Limitations and future research*

This study has several limitations that provide avenues for potential future research. First, the imputation method developed in this study was tested using one type of online information. It is therefore an important area of future research to check a validation of the imputation method for different online information. Second, the findings indicate that multiple imputation using additional observed information representing characteristics of reviews better handle missing data than deletion, suggesting that the use of representative information increases the quality of multiple imputation. In this regard, we note that our imputation method can be improved by incorporating information from visual expressions in reviews, such as emoji and photos. Finally, our imputation method is tested based on the assumption that online reviews are MAR, which may restrict the generalizability of the method. Future

studies should apply the method to the review data with NMAR missing patterns.

References

Aggarwal, C. C. and C. Zhai (2012). *Mining Text Data*. New York, NY: Springer.

Allison, A. C. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Baka, V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management 53*, 148–162.

Barbiero, A. and P. A. Ferrari (2015). *Package 'GenOrd'*. https://cran.r-project.org/web/packages/GenOrd/GenOrd.pdf.

Bird, S., E. Loper, and E. Klein (2009). *Natural Language Processing with Python*. Sebastopol CA: O'Reilly Media Inc.

Blei, D., A. Ng, and M. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Buhi, E. R., P. Goodson, and T. B. Neilands (2008). Out of sight, not out of mind: strategies for handling missing data. *American Journal of Health Behavior 32*, 83–92.

Casaló, L. V., C. Flavián, M. Guinalíu, and Y. Ekinci (2016). Do online hotel rating schemes influence booking behavior? *International Journal of Hospitality Management 49*, 28–36.

Ding, Y. and J. S. Simonoff (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research 11*, 131–170.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association 91*, 490–498.

Ferrari, P. A. and A. Barbiero (2012). Simulating ordinal data. *Multivariate Behavioral Research 47*, 566–589.

Geetha, M., P. Singha, and S. Sinha (2017). Relationship between customer sentiment and online customer ratings for hotels - an empirical analysis. *Tourism Management 61*, 43–54.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *American Review of Psychology 60*, 549–576.

Hills, J. R. and G. Cairncross (2010). Small accommodation providers and ugc web sites: perceptions and practices. *International Journal of Contemporary Hospitality Management 23*(1), 26–43.

Hippel, P. T. V. (2004). Biases in spss 12.0 missing value analysis. *The American Statistician 58*, 160–164.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference*.

Hornik, K. (2016). *Apache OpenNLP Tools Interface*.

Horton, N. J. and K. P. Kleinman (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician 61*, 79–90.

Ibrahim, J. G., M. H. Chen, and S. R. Lipsitz (2005). Missing data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association 100*, 332–346.

Im, J., J. K. Kim, and W. A. Fuller (2015). Two-phase sampling approach to fractional hot deck imputation. In *Proceedings of Survey Research Methodology Section*, Seattle, WA, pp. 1030–1043. American Statistical Association.

Im, J. and S. Kim (2017). Multiple imputation for nonignorable missing data. *Journal of the Korean Statistical Society*.

Kim, J. K. and W. A. Fuller (2004). Fractional hot deck imputation. *Biometrika 91*, 559–578.

Kim, W. G., J. Li, and R. A. Brymer (2016). The impact of social media reviews on restaurant performance:the moderating role of excellence certificate. *International Journal of Hospitality Management 55*, 41–51.

Kwok, L. and L. Xie (2016). Factors contributing to the helpfulness of online hotel reviews: does manager responses play a role? *International Journal of Contemporary Hospitality Management 28*(10), 2156–2177.

Lee, S. H. and H. Ro (2016). The impact of online reviews on attitude changes: the differential effects of review attributes and consumer knowledge. *International Journal of Hospitality Management 56*, 1–9.

Levy, S. E., W. Duan, and S. Boo (2013). An analysis of one-star online reviews and responses in the washington d.c. lodging market. *Cornell Hospitality Quarterly 54*, 49–63.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

Liu, B. (2014). *Sentiment Analysis*. New York, NY: Cambridge.

Nasukawa, T. and J. Yi (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*.

Oxford Economics (2016). *The global economic contribution of TripAdvisor*. Retrived from https://d2bxpc4ajzxry0.cloudfront.net/TripAdvisorInsights/sites/default/files/downloads/2687/taoxford_tripadvisor_globalreport_2016.pdf.

Pand, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*, pp. 79–86.

Philander, K. and Y. Y. Zhong (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management 55*, 16–24.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation 7*, 353–383.

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological bayesian approach to nonresponse (with discussion). In *Proceedings of Survey Research Methods Section*, Alexandria, VA, pp. 20–34. American Statistical Association.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc.

Rubin, D. B. and N. Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association 81*, 366–374.

Saar-Tsechansky, M. and F. Provost (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research 8*, 1625–1657.

Schafer, J. L. and J. W. Graham (2002). Missing data:our view of the state of the art. *Psychological Methods 7*(2), 147–177.

Schlomer, G. L., S. Bauman, and N. A. Card (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology 57*, 1–10.

Silge, J. and D. Robinson (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software 1*(3). http://dx.doi.org/10.21105/joss.00037.

SyntaxNet (2016). Syntaxnet: Neural models of syntax.

Tirunillai, S. and G. J. Tellis (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research 51*, 463–479.

TripAdvisor (2017). *TripAdvisor network effect and benefits of total engagement*. Retrived from https://www.tripadvisor.com/TripAdvisorInsights/n2761/tripadvisor-network-effect-and-benefits-total-engagement.

van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine 18*, 681–694.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). **mice**: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45*, 1–67.

Viglia, G., R. Minazzi, and D. Buhalis (2016). The influence of e-word-of-mouth on hotel occupancy rate. *International Journal of Contemporary Hospitality Management 28*(9), 2035–2051.

Xiang, Z., Z. Schwartz, J. H. Gerdes., and M. Uysal (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management 44*, 120–130.