A novel deep learning method to use feature complementarity for review helpfulness prediction

Xinzhe Li, Qinglong Li and Dasom Jeong
department of Big Data Analytics, Kyung Hee University, Seoul, Republic of Korea, and Jaekyeong Kim
Department of Big Data Analytics and School of Management, Kyung Hee University, Seoul, Republic of Korea

Abstract

Purpose – Most previous studies predicting review helpfulness ignored the significance of deep features embedded in review text and instead relied on hand-crafted features. Hand-crafted and deep features have the advantages of high interpretability and predictive accuracy. This study aims to propose a novel review helpfulness prediction model that uses deep learning (DL) techniques to consider the complementarity between hand-crafted and deep features.

Design/methodology/approach – First, an advanced convolutional neural network was applied to extract deep features from unstructured review text. Second, this study used previous studies to extract hand-crafted features that impact the helpfulness of reviews and enhance their interpretability. Third, this study incorporated deep and hand-crafted features into a review helpfulness prediction model and evaluated its performance using the Yelp.com data set. To measure the performance of the proposed model, this study used 2,417,796 restaurant reviews.

Findings – Extensive experiments confirmed that the proposed methodology performs better than traditional machine learning methods. Moreover, this study confirms through an empirical analysis that combining hand-crafted and deep features demonstrates better prediction performance.

Originality/value – To the best of the authors’ knowledge, this is one of the first studies to apply DL techniques and use structured and unstructured data to predict review helpfulness in the restaurant context. In addition, an advanced feature-fusion method was adopted to better use the extracted feature information and identify the complementarity between features.

Keywords Review helpfulness prediction, Convolutional neural network, Deep learning, Hospitality industry, Online reviews, Feature complementarity

Paper type Research paper

This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).
1. Introduction

The development of information and communication technology has allowed the hospitality and tourism industries to grow (Chatterjee, 2020; Oh et al., 2022). Therefore, massive amounts of consumer-generated content (e.g. online consumer reviews) have accumulated. Online reviews significantly affect consumer purchase decision-making and tourism product evaluation (Kwon et al., 2021; Ren and Hong, 2019). As online tourism communities extensively recommend products, online consumer reviews have become an essential product information resource, and hospitality and tourism websites have become reliable. However, the exponential growth in online reviews has made it difficult for consumers to identify valuable reviews for purchase decision-making (Hu and Chen, 2016). Most hospitality and tourism websites collect review helpfulness voting information to evaluate consumer review quality. Consumers filter online reviews based on helpful vote system information to reduce information exploration efforts and time spent in purchase decision-making processes. However, only partial online reviews receive helpful voting, and less popular products lack voting information (Chen et al., 2022).

Because helpful reviews play an essential role in consumer decision-making, hospitality and tourism websites should automatically identify high-quality online reviews and provide them to consumers. Hospitality and tourism scholars have proposed prediction methodologies for identifying helpful reviews (Lee et al., 2021; Lee et al., 2018; Tsai et al., 2020). The core idea is to design feature patterns and rules based on a traditional machine learning (ML) model using hand-crafted features. Specifically, they extracted the most relevant intuitive numerical features that affect review helpfulness through domain knowledge and used them as inputs for prediction. For example, reviewer-generated features (e.g. reviewer disclosure and expertise) have been shown to affect review helpfulness (Liang et al., 2019). However, because ML models cannot directly process unstructured review texts, most studies extract a wide range of hand-crafted features (e.g. sentiment, subjectivity and readability) from review texts (Hu and Chen, 2016; Tsai et al., 2020). There is an advantage in using hand-crafted features, as they support many theories, resulting in highly interpretable prediction results; however, they represent only shallow information (Ren et al., 2024). In addition, applying hand-crafted features is practical and easy but requires domain knowledge and a considerable amount of time to extract and select features.

With the emergence and development of deep learning (DL) techniques, advanced feature learning abilities have become widely used to assess review texts (Luo and Xu, 2021; Ren et al., 2024; Zheng et al., 2021). Such DL techniques can automatically extract high-dimensional deep features from text that cannot be detected by humans. For example, a convolutional neural network (CNN) can represent high-level semantic representations using convolutional operations in a sentence context (Ren et al., 2024). Several scholars of hospitality and tourism have applied CNN to identify new knowledge and consumer
behavior patterns embedded in review texts (Q. Li et al., 2023; Zheng et al., 2021). Many studies have shown that CNN performs outperform in predicting review text (Q. Li et al., 2023; Puh and Bagić Babac, 2023). DL techniques have better prediction performance but lack transparency in the learning process, which makes the model less interpretable (Ren et al., 2024). Therefore, integrating hand-crafted and deep features represents a promising approach to enhance the accuracy and interpretability of review helpfulness predictions. To the best of our knowledge, few studies have examined these features simultaneously, especially in the context of review helpfulness prediction in the hospitality and tourism industry. This study explores the potential advantages of combining these disparate features, providing theoretical and practical implications for more effectively predicting the helpfulness of online reviews.

This study proposes a novel review helpfulness prediction model that uses DL techniques to consider the complementarity between hand-crafted and deep features. First, an advanced CNN is introduced to extract deep and contextual features from textual information. Second, this study follows previous studies to identify hand-crafted features that affect review helpfulness and improve interpretability. Third, this study uses the feature-fusion methodology to incorporate deep and hand-crafted features obtained in the first two phases into a review helpfulness prediction model and evaluates its performance using the Yelp.com data set.

This study makes three main contributions. First, we propose a novel DL-based online review helpfulness prediction model using hand-crafted and deep features. The proposed model extracts deep and contextual features from unstructured review texts using a CNN and improves prediction performance by considering the complementarity between hand-crafted and deep features. This study expands the literature on predicting review helpfulness in hospitality and tourism by emphasizing the importance of evaluating review helpfulness using structured and unstructured information. Second, this study uses an advanced feature-fusion methodology for each input feature to better use the extracted feature information and identify the complementarity between features. This study maps review texts and star ratings as embedding vectors with the same dimensions to ensure that they have equivalent representation capabilities. In addition, the hand-crafted features are normalized and converted into feature vectors, and fusion features were obtained for each attribute. Then, a multilayer perceptron (MLP) is applied to enhance the learning ability and capture complex nonlinear relationships. Third, to evaluate the performance of the proposed prediction model, this study empirically compares its performance with ML-based baseline models using a real-world data set provided by Yelp.com and confirms the efficiency of the proposed model.

2. Literature review and research background
2.1 Traditional techniques for review helpfulness prediction
In the purchase decision-making process in online shopping, review helpfulness represents the consumers’ perceived value of reviews (Lee et al., 2021). Owing to the uncertainty of information, consumers seek trustworthy and valuable information to make informed decisions. Many scholars have found that helpful reviews positively affect consumer purchase behavior for specific products or services and significantly affect consumer decision-making (Hu and Yang, 2021; Tsai et al., 2020). Helpful reviews contain detailed and reliable information that allows consumers to explore various aspects of their products and services. Therefore, hospitality and tourism websites provide helpful reviews on review pages to support potential consumer decision-making.
Previous studies have identified features that affect review helpfulness through extensive experimentation and investigation. Many scholars have investigated the roles of various numeric (i.e. star rating and review length) and text features. Many studies have found that review length positively affects review helpfulness (Liu and Park, 2015; Moro and Esmerado, 2021). In addition, several studies have found that star ratings and extremities are essential factors affecting review helpfulness (Lee et al., 2018; Park and Nicolau, 2015).

In addition to numerical factors, several studies have investigated the effects of textual features on review helpfulness. Scholars have found that readability and subjectivity are essential factors affecting review helpfulness (Fang et al., 2016; Ghose and Ipeirotis, 2010). This suggests that consumers tend to trust reviews that are easy to read and contain lucid information. Meanwhile, some scholars have found that review sentiments can effectively reflect the features of hotels and restaurants to improve review helpfulness quality (Bagherzadeh et al., 2021). For example, many studies agree that negative emotions in online reviews positively affect review helpfulness (Filieri et al., 2021; Huang et al., 2020).

Despite the importance of review content features, scholars have found that providing reviewer-generated features (e.g. number of friends, number of reviews and elapsed months) affects review helpfulness (Kwon et al., 2021; Lee et al., 2021; Park and Nicolau, 2015). In addition, some scholars have investigated the effects of seller-generated features on review helpfulness. Scholars have found that popular products attract more consumers to read and vote on helpful reviews (Chen et al., 2022). Some scholars have used average star ratings and review counts to investigate how restaurant or hotel quality affects review helpfulness (Liang et al., 2019; Zhang and Lin, 2018).

In summary, such studies provided a theoretical foundation for extracting hand-crafted features to predict review helpfulness. Therefore, incorporating hand-crafted features into the review helpfulness prediction model enhances interpretability.

A helpfulness prediction model that automatically provides helpful reviews to address information overload and reduce consumers’ efforts to explore information is gaining significant attention in the hospitality and tourism domains. Early studies on review helpfulness prediction in the hospitality and tourism domains focused on traditional ML approaches. For example, Hu and Chen (2016) used ML methods such as multilinear regression (MLR), model tree (i.e. M5P) and support vector regression to predict hotel review helpfulness using review content, sentiment, reviewers and visibility features. Hu et al. (2017) used ML methods such as linear regression (LR), reduced error-pruning tree and random forest (RF) to automatically evaluate hotel review helpfulness using both review content and reviewer features. Lee et al. (2021) found that integrating review, reviewer, restaurant and linguistic content features into ML methods, such as MLR, RF, support vector machine (SVM) and extreme gradient boosting (XGBoost), yields better predictions.

Extensive studies emphasize the critical role of hand-crafted features (e.g. reviewer, review content and restaurant) in effectively predicting review helpfulness (Kwon et al., 2021; Moro and Esmerado, 2021). Typically, such studies extract hand-crafted features from unstructured review texts, such as sentiment, subjectivity assessment and readability (Lee et al., 2017). However, compressing unstructured text into simplified numeric scores can dilute important information, thus diluting the depth of analysis required for nuanced interpretation. On the other hand, the advent of DL techniques introduces a paradigm shift, enabling the extraction of deep features that capture semantic and contextual information more holistically. These non-hand-crafted features offer a sophisticated mechanism for interpreting the intricacies of review content, thus addressing the limitations inherent in traditional feature extraction methodologies. Integrating hand-crafted and deep features in constructing a review helpfulness prediction model represents a methodological
advancement, marrying the detailed specificity of hand-crafted features with the comprehensive insight of DL approaches. Such an approach cannot only overcome the individual limitations of each feature type but also enhance the model's overall interpretability and prediction performance.

2.2 Deep learning techniques for review helpfulness prediction

The DL technique has recently gained significant attention owing to its outstanding performance in classifying and predicting hospitality and tourism. The DL technique must be applied to extract significant information on various features that affect review helpfulness. The hospitality and tourism industries have widely applied the DL technique, which offers numerous advantages (Essien and Chukwukelu, 2022; Liu et al., 2023). First, unlike the traditional ML approach, the DL technique is a data-driven approach that does not rely on assumptions regarding the data. However, traditional ML approaches are parametric models that require specific assumptions, such as a normal distribution and a linear relationship between independent and dependent features. The DL technique is an advanced method that is capable of approximating continuous functions. It can effectively process unstructured problems of complex relationships between features, rather than processes under value-based human reasoning. In summary, the DL technique has an advantage over statistical models in capturing complex patterns of relationships between features owing to its better ability to capture nonlinear relationships from data. Therefore, a DL technique can capture complex patterns between features, providing an opportunity to predict review helpfulness (Lee and Choeh, 2014).

For natural language processing (NLP) tasks, various models using the DL technique have been proposed to extract deep features from textual information. A CNN is a typical model that has shown better performance in NLP tasks such as text classification and sentiment analysis. A CNN focuses on extracting local information and discards insufficiently significant information. In general, they are divided into character-level CNN and word-level CNN, where the difference between them is whether extracting local information from review texts is based on characters or words. Regardless of the method used, many hospitality and tourism studies have applied CNNs to analyze texts from various perspectives because they can extract high-level semantic representations from review texts. For example, Arenas-Márquez et al. (2021) extracted unique classes/topics associated with each traveler using a CNN to analyze traveler reviews. Moreover, Puh and Bagić Babac (2023), Liu and Zhao (2023) also applied CNN in sentiment analysis. In particular, Liu and Zhao (2023) proposed a CNN-based DL model that can integrate multigranular semantic features that contain radical and part-of-speech features in addition to character and word features. Meanwhile, Li et al. (2023) proposed a restaurant recommendation model using a CNN to extract consumer preferences embedded in online review texts. Consequently, previous studies have attempted to extract deep and contextual features embedded in review texts by building a CNN.

In addition to CNN, recurrent neural networks (RNN) are widely used in DL techniques. This enables scholars to conveniently process sequential review text data; it retains the memory of what came before the current sequence is processed. Unlike CNN, RNN imitates human reading behavior by extracting semantic representations from left to right of the entire review text. However, as RNN are not ideal for capturing long-distance semantic connections and face vanishing gradient problems (Yang et al., 2022), their variation, known as long short-term memory (LSTM), is more commonly applied. For example, Luo and Xu (2021) applied LSTM to predict the sentiments of online restaurant reviews and demonstrated that the DL technique has a better ability to perform prediction tasks. Oh et al.
JHTT (2022) predicted consumer satisfaction with hospitality services using multimodal data, specifically, by applying LSTM to process review comments. They also used a basic RNN for comparison and demonstrated that LSTM is more appropriate for processing long reviews that include both positive and negative comments. Zheng et al. (2021) applied both CNN and LSTM to investigate the reliability of reviews and indicated that they are beneficial for sequence data generation.

In summary, because CNN and RNN efficiently extract high-level semantics from review texts, they are commonly used in the hospitality and tourism domains (Essien and Chukwukelu, 2022; Zheng et al., 2021). However, an RNN considers every piece of information and its order; it may learn too much unnecessary information, causing noise and further limiting its performance (Oh et al., 2022). On the other hand, from the perspective of how consumers evaluate reviews, consumers tend to focus on important information (i.e. keywords) rather than reading the entire review text. Therefore, this study applies a CNN to capture the most significant local information to predict review helpfulness. Specifically, this study applied a word-level CNN with filters of different sizes to capture multigram constituent semantic representations embedded in review texts. It can fully capture the interactions between constituent semantic fragments to use the semantic relations of online reviews, which is consistent with our goal of extracting high-level semantic representations.

3. Research methodology
This study aims to develop a review helpfulness prediction model by applying DL techniques; the specific methodology is shown in Figure 1. In the data collection phase, this study used restaurant reviews collected from Yelp.com in five states in the USA[1]. The data preprocessing phase includes data cleaning processes such as filtering, integration and text cleaning. The determinant features were identified and extracted to develop a model. The final task of the data preprocessing phase was to divide the experimental data set into training and test sets for model development and performance comparison. Finally, we conducted model comparison experiments and evaluated the prediction performance based on the attribute sets.

3.1 Data collection
This study used the Yelp open data set provided by Yelp.com to conduct experiments. Yelp.com is one of the most popular restaurant review websites, with 6,990,280 consumer reviews (Luo et al., 2020; Nakayama and Wan, 2019). The Yelp data set is widely used in various research domains. Using the Yelp data set not only provides a fair comparison with the models proposed in previous studies but also provides significant implications for the hospitality and tourism domains and consumer-generated content quality evaluation. This study used restaurant reviews in the top five states based on the number of review helpfulness votes: Pennsylvania, Florida, Louisiana, Missouri and Nevada. Finally, this study used 2,417,796 restaurant reviews of 828,214 consumers from 20,130 restaurants collected between February 2005 and January 2022.

This study adopted the Natural Language Toolkit (NLTK) package in Python to conduct text preprocessing to ensure data quality and analysis efficiency (X. Li et al., 2023). First, spaces and non-English reviews were removed. Second, the review text was converted to lowercase to avoid duplicate variations. Third, stop words, punctuation marks, numbers and special characters were removed. Fourth, stemming and lemmatizing were adopted, and words that appeared fewer than three times were removed. After text preprocessing, the data set was randomly split into training and test sets at an 8:2 ratio to evaluate model performance. In this study, 20% of the training set was used for validation to avoid
overfitting and to select the optimized parameters. Moreover, following previous suggestions, this study adopts a classification approach that labels a review as helpful if it receives at least one helpful vote and otherwise unhelpful (Krishnamoorthy, 2015; Lee et al., 2018; Olmedilla et al., 2022). The number of reviews from one label was randomly sampled to match the number of reviews from another to ensure label balance. This study randomly sampled the same number of 983,208 reviews based on label type (i.e. helpful or unhelpful) to prevent prediction bias.

3.2 Feature selection

Table 1 summarizes the seven independent and one dependent variables used in this study. Following the investigation and integration of previous studies, this study identifies seven features: review content, reviewers and restaurant attributes. As shown in Table 1, this study measured the remaining variables, excluding unstructured review texts, based on previous studies. This study attempted to apply a CNN model to address issues in which it is difficult to directly extract semantic representations embedded in unstructured review texts. Finally, the model was developed and evaluated by applying a DL mechanism to develop an advanced helpfulness prediction model.

3.3 Methodology

3.3.1 Overview. The purpose of performing review helpfulness prediction is to identify and provide high-quality online reviews to consumers. Previous studies have primarily used hand-crafted features that affect review helpfulness and input them into ML models for prediction. However, these features are extracted through domain knowledge and represent only shallow information. This study emphasizes the efficiency of the DL technique and the importance of considering the complementarity between hand-crafted and deep features to ensure the accuracy and interpretability of helpful predictions. Figure 2 illustrates the overall framework of the proposed model.

As the impact on review helpfulness varies for each attribute, this study processes them independently. Input attributes can be summarized as review content, reviewer or restaurant attributes. As for the review content attribute, this study applies CNN models to extract deep and high-level semantic representations embedded in review texts and uses an element-wise operation to capture the consistency between review texts and star ratings to
### Table 1. Descriptions and summary statistics of variables

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Feature</th>
<th>Description</th>
<th>Type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review</td>
<td>Star rating</td>
<td>Extracted from the reviewer-left rating of the review</td>
<td>Numeric</td>
<td>Lee et al. (2018)</td>
</tr>
<tr>
<td>content</td>
<td>Review text</td>
<td>Extracted by applying CNN model and represented as a vector</td>
<td>Object</td>
<td>Liu and Park (2015)</td>
</tr>
<tr>
<td>Reviewer</td>
<td>Number of Friends</td>
<td>Extracted by counting the number of friends of the reviewer</td>
<td>Numeric</td>
<td>Liu and Zhao (2023)</td>
</tr>
<tr>
<td></td>
<td>Number of Reviews</td>
<td>Extracted by counting the number of reviews of the reviewer</td>
<td>Numeric</td>
<td>Puh and Bagić Babac (2023)</td>
</tr>
<tr>
<td></td>
<td>Elapsed months</td>
<td>Extracted by calculating the elapsed months between reviewer enrollment and review writing</td>
<td>Numeric</td>
<td>Kwon et al. (2021)</td>
</tr>
<tr>
<td></td>
<td>Average rating</td>
<td>Extracted by calculating the average rating of the restaurant</td>
<td>Numeric</td>
<td>Zhang and Lin (2018)</td>
</tr>
<tr>
<td>Restaurant</td>
<td>Review count</td>
<td>Extracted by counting the number of reviews of the restaurant</td>
<td>Numeric</td>
<td>Lee et al. (2021)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Zhang and Lin (2018)</td>
</tr>
</tbody>
</table>

**Dependent Variable**:
- Review helpfulness: Extracted by classifying the helpfulness of review based on received vote

*Source: Created by authors*

---

**Figure 2.**
The framework of the proposed model

*Source: Created by authors*
introduce it to the prediction task (X. Li et al., 2023). For the reviewer and restaurant attributes, this study normalizes and converts each feature into a vector and concatenates them. Then, MLP is applied to ensure that concatenated vectors can capture complex nonlinearity and enhance learning ability. Finally, this study concatenates all vectors and applies the MLP to capture the relationships and dependencies for predicting review helpfulness.

3.3.2 Feature extraction of review content attribute. As review texts are unstructured, scholars tend to represent them using various numerical scores, such as sentiment, subjectivity and readability (Hu and Chen, 2016; Tsai et al., 2020). However, as this study emphasizes, such features are represented based on particular perspectives of review texts, which may dilute important information as unstructured data are compressed into numeric scores. They also ignore the semantic and contextual information embedded in the textual information. Therefore, this study aims to apply a CNN model to address these issues.

CNN, as one of the typical DL techniques, has been widely applied in various studies about hospitality and tourism domains in processing textual information (Liu and Zhao, 2023; Puh and Bagić Babac, 2023). Before applying the CNN, each word in the review text is mapped to a fixed-size embedding vector. Word embedding vectors are pre-trained using large text corpora and used to capture the semantic relationships between words. This process is commonly used for preparation before applying DL-based models. The main components of a CNN are the convolutional and pooling layers. The convolutional layer extracts local information, and the pooling layer extracts the most significant information. Therefore, a CNN aims to extract local information and discard information that is not sufficiently significant. During the text feature extraction process, semantic representations are first extracted by applying multiple filters to the convolutional layer. For example, assume we analyze the sentence “The service was good, but it’s not the best taco from Taco Bell” and set the filter size to 3 (3 grams). The output of the convolutional layer would be “The service was,” “service was good,” “was good but,” etc. Then, the pooling layer extracts the most significant information, and the outputs would highlight “good service,” “contrast in quality” and “Taco Bell.” However, because of the inexplicability of the DL technique, the actual result tends to be represented as a vector.

Regarding star ratings, this study only embedded them into a feature vector because they contain relatively simple and direct information. Aghakhani et al. (2021) found that the consistency between review texts and star ratings significantly affects consumers’ evaluations of review helpfulness. Inspired by these findings, this study introduced the concept of consistency between review texts and star ratings in prediction tasks. Therefore, this study first applied the MLP to convert them into the same dimension to ensure that review texts and star ratings have equivalent representation capabilities. This study then uses the element-wise operation to learn consistency and finally obtains a vector representing the extracted feature of the review content attribute.

3.3.3 Feature extraction of reviewer and restaurant attributes. For reviewer and restaurant attributes, each selected feature is hand-crafted, contains numerical scores and is highly interpretable. As each feature represents different meanings and is measured on different scales, this study normalizes and concatenates them based on the attributes to which they belong. The primary role of the MLP is not simply to convert the vector’s dimensions but to provide a high level of nonlinearity (Q. Li et al., 2023). Therefore, this study adds an MLP to concatenated reviewer and restaurant vectors to enhance their learning ability and capture complex nonlinear relationships.

As mentioned above, this study proposes a novel review helpfulness prediction model that considers the complementarity between hand-crafted and deep features. Therefore, this
study integrates review content, reviewers and restaurant attributes as a fusion of deep and hand-crafted features. We entrust the learning process of complementarity to MLP, which is the most widely applied and practical methodology for capturing information relationships and dependencies. Finally, the output feature reflects complementarity between hand-crafted and deep features for predicting the review helpfulness type as “Helpful” or “Unhelpful.” The proposed model predicts review helpfulness with high accuracy and interpretability.

This study developed a model using TensorFlow 2.6.0 in Python 3.7. Meanwhile, this study set the kernel size of the CNN to three to extract the semantic features from the review texts. For the embedding dimensions of the review texts and star ratings, we set the number of dimensions to 300 to conduct various experiments. Moreover, L2 regularization was added to the proposed model to prevent overfitting. Regarding the review length used in the CNN, we calculated the length of each review and set a maximum length of 90% of the overall reviews as the review length to conduct the experiments effectively (Du et al., 2020).

3.4 Baselines
We applied various prediction models that have been widely used in previous studies to compare the prediction performance of the proposed model. The baseline models used in this study were six traditional ML techniques and one DL technique, which included multilinear regression (MLR), RF, SVM, XGBoost and a deep neural network (DNN). To ensure a fair comparison, the input features were consistent with those listed in Table 1. However, because such models are limited to automatic feature extraction from review texts, the input features of review texts are represented by review length, sentiment, subjectivity and readability, which have been commonly used in previous studies (Hu and Chen, 2016; Hu et al., 2017; Lee et al., 2021). Review length was extracted by counting the number of words in the review, and sentiment was extracted by calculating and summing the sentiment scores of the review. Subjectivity and readability were also extracted by calculating the subjectivity and Flesch–Kincaid Reading Ease scores of the reviews. This study extracted such scores with the help of NLTK and the TextBlob library in Python.

3.5 Prediction performance evaluation
This study adopts accuracy, precision, recall and F1-score as evaluation metrics widely used in classification tasks to evaluate the classification performance (Yoo et al., 2024). These metrics are instrumental in assessing binary classification tasks. Accuracy is the ratio of actual reviews (both helpful and unhelpful) to the total number of reviews. Precision is the ratio of actual helpful reviews to classified helpful reviews, and recall is the ratio of classified helpful reviews to actual helpful reviews. F1-score represents the balanced weight average of the precision and recall values, which can be used for classification comparisons.

4. Experimental results
4.1 Comparison analysis
This study collected 1,966,416 reviews from five states to conduct a comprehensive series of experiments. An equal number of reviews (983,208) were selected for each label to mitigate any potential bias toward a specific review helpfulness type. This balanced approach to sampling was crucial in ensuring the reliability and generalizability of our findings. This study first conducted a model comparison experiment to compare the review helpfulness prediction performance of the proposed model with that of the baseline models, as shown in Table 2.
The results show that the proposed model outperforms baseline models. Compared with the traditional ML-based models, the proposed model improved accuracy by 8.772% to 17.789% and F1-score by 11.755% to 25.465%. Compared to the DL-based model (i.e. DNN), the proposed model improved accuracy by 3.963% and F1-score by 4.328%. This finding effectively demonstrates our methodology and highlights the necessity of adopting DL techniques for review helpfulness prediction tasks.

Moreover, the DNN, as a typical DL-based model, outperformed traditional ML-based models. This is because of their DNN structure, which can perform nonlinear transformations of inputs with advanced operations to enhance the learning ability. However, compared with the proposed model, the prediction performance of a DNN is still limited. The DNN inputs the features by simply converting them into feature vectors, similar to the processing of reviewer and restaurant attributes in the proposed model. However, because the proposed model applies a CNN as an encoder to extract high-level semantic representations from review texts, the extracted review text features contain contextual information between each word and text. Therefore, the better of the proposed extraction method is reflected in the prediction performance, which further demonstrates the limitations of manual extraction methods that ignore the surrounding words and sentence contexts.

From an application perspective, although previous studies indicate that the purpose of developing review helpfulness prediction models is to provide consumers with helpful reviews, they are less likely to be applied to e-commerce websites. This is because traditional ML-based models are predicted based on manually extracted features, which require scholars to process feature extraction first. However, the proposed methodology was developed with an end-to-end structure that can automatically predict review helpfulness. For example, if a review is posted, the proposed model can directly predict its helpfulness and provide it to the consumers. Moreover, previous studies have been limited in that they require prior knowledge of specific domains. However, because the features determined by the proposed methodology are commonly used in any study, they can be effectively applied to various domains and provide better prediction services.

### 4.2 Sensitivity analysis

This section further evaluates the robustness of the proposed model and reports the experimental results that compare the prediction performance based on attribute sets. Because review helpfulness prediction is conducted based on review content, this section evaluates the performance of review content attributes and combined attribute sets to compare the effects of attribute importance on prediction performance, as shown in Table 3. Based on these results, this study concludes that predicting review helpfulness using a combined attribute set improves the prediction performance compared to using one or two attribute sets. Meanwhile,

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>% change (Accuracy)</th>
<th>% change (F1-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline MLR</td>
<td>0.579</td>
<td>0.648</td>
<td>0.460</td>
<td>0.538</td>
<td>17.789</td>
<td>25.465</td>
</tr>
<tr>
<td>RF</td>
<td>0.612</td>
<td>0.706</td>
<td>0.501</td>
<td>0.586</td>
<td>11.428</td>
<td>15.188</td>
</tr>
<tr>
<td>SVM</td>
<td>0.595</td>
<td>0.539</td>
<td>0.626</td>
<td>0.579</td>
<td>14.622</td>
<td>16.580</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.627</td>
<td>0.644</td>
<td>0.567</td>
<td>0.604</td>
<td>8.772</td>
<td>11.755</td>
</tr>
<tr>
<td>DNN</td>
<td>0.656</td>
<td>0.688</td>
<td>0.610</td>
<td>0.647</td>
<td>3.963</td>
<td>4.328</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.682</td>
<td>0.689</td>
<td>0.662</td>
<td>0.675</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**Source:** Created by authors

Table 2. Results of prediction performance comparison between proposed and baseline models
the prediction performance of combining review content and reviewer attributes outperforms that of combining review content and restaurant attributes. This result is consistent with that of Lee et al. (2021), who indicate that features related to reviewer attributes have the most significant impact on predicting helpfulness. Lascu et al. (1995) also indicated that consumers prefer expert suggestions before the decision-making process because experts are more valuable and trustworthy. Based on these results, this study concludes that using reviewer attributes to predict review helpfulness is more significant than using restaurant attributes. Nonetheless, when using the overall attributes, the proposed model exhibited the best performance, demonstrating the correctness of our ideas.

5. Discussion and conclusions

5.1 Conclusion

Review helpfulness prediction is an essential topic in the hospitality and tourism industries and has gained significant attention from scholars. Most studies have reviewed helpfulness prediction by relying on simple and static hand-crafted features. Although this methodology is practical and easy, it requires area-specific knowledge and considerable time to extract and select features. In addition, because the extraction process ignores the semantic and contextual information embedded in review texts, textual information only reflects shallow information. This study aims to incorporate structured and unstructured data by applying DL techniques to predict review helpfulness with high accuracy and interpretability. Our findings highlight the feasibility of comprehensiveness in revealing the informativeness of reviews and predicting helpfulness. The experimental results also demonstrated the effectiveness of the feature-fusion methodology in prediction tasks.

This study makes several important contributions to the existing literature. First, this is one of the first studies to apply DL techniques and consider the complementarity between structured and unstructured data. Second, the proposed model fully captures the dependencies of different types of information and significantly improves the review helpfulness prediction performance. Third, because the determined features were commonly used in previous studies, the proposed model can be easily applied to various domains.

5.2 Theoretical implications

This study has the following theoretical implications. First, this study advances the theoretical framework for online review analysis by integrating structured and unstructured

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review content</td>
<td>Predict review helpfulness based on review content attribute</td>
<td>0.624</td>
<td>0.674</td>
<td>0.559</td>
<td>0.611</td>
</tr>
<tr>
<td>Review content + Reviewer</td>
<td>Predict review helpfulness based on review content and reviewer attributes</td>
<td>0.668</td>
<td>0.702</td>
<td>0.622</td>
<td>0.660</td>
</tr>
<tr>
<td>Review content + Restaurant</td>
<td>Predict review helpfulness based on review content and restaurant attributes</td>
<td>0.647</td>
<td>0.650</td>
<td>0.635</td>
<td>0.642</td>
</tr>
<tr>
<td>Review content + Reviewer + Restaurant</td>
<td>Predict review helpfulness based on overall attributes</td>
<td>0.682</td>
<td>0.689</td>
<td>0.662</td>
<td>0.675</td>
</tr>
</tbody>
</table>

Table 3. Comparisons of prediction performance for different attribute sets

Source: Created by authors
data through DL techniques. Previous studies in the hospitality and tourism domains have mainly focused on either data type separately. Our DL-based approach, which uniquely applies to structured and unstructured data for review helpfulness prediction, not only addresses this gap but also offers a comprehensive model that outperforms traditional ML-based methodologies. This approach underscores the various aspects of review helpfulness and presents a direction for future studies exploring different feature integrations in prediction modeling.

Second, our empirical analysis reaffirms the significance of hand-crafted features identified in previous hospitality and tourism studies, such as sentiment, subjectivity and readability, within the prediction model. By systematically verifying the impact of these factors, the study enriches the theoretical foundation on review helpfulness by providing empirical evidence of their prediction power, thereby highlighting their continued relevance in the era of DL.

Third, diverging from the existing literature’s focus on hand-crafted features, this study emphasizes the utility of deep features extracted using CNN. This methodology challenges the conventional approach, which relies on numeric or dummy variables, and theoretically enriches our understanding of how semantic and contextual nuances embedded in review texts impact helpfulness perceptions. By demonstrating the excellent prediction performance achieved through this approach, the study contributes a theoretical insight into the complex dynamics of review text analysis.

Finally, this study addresses a critical challenge in previous studies by innovatively combining hand-crafted and deep features to enhance predictability and interpretability. This integration tackles the trade-off between these two aspects, contributing to a new theoretical model that leverages the strengths of econometric and predictive study findings. This paves the way for future research to harmonize these complementary features in online content analysis.

5.3 Practical implications
This study has the following practical implications. First, our model suggests innovative strategies to refine review ranking systems, significantly reducing search costs and cognitive effort for consumers. By accurately predicting and surfacing high-quality reviews, online platforms can facilitate more informed decision-making, enabling consumers to assess restaurants from diverse perspectives based on the most helpful reviews. This enhanced review helpfulness system supports thoughtful decision-making, improving consumer experience and satisfaction.

Second, the findings empower restaurant marketers to identify factors that significantly influence consumer experiences during meals automatically. As helpful reviews reflect a wide range of consumer experiences, marketers are provided with data-driven information to identify and facilitate restaurant shortcomings. Restaurant marketers can leverage insights from helpful reviews to customize dining experiences and marketing strategies, better meeting consumer expectations and increasing brand value and competitiveness in the market.

Finally, our findings can help platform managers introduce novel review-ranking systems to display reviews on hospitality and tourism websites. Most websites sort reviews according to star ratings or the newest. The proposed model can provide insights into improving existing ranking systems. Platform managers could develop helpfulness-based ranking systems to highlight helpful reviews when sorting them. Moreover, even though specific websites have not introduced a review helpfulness voting system, they can benefit from our proposed model. Overall, such an automatic review helpfulness prediction system

Review helpfulness prediction
could be an essential source of competitive advantage for platforms. A review helpfulness prediction service with better performance can improve a platform's reputation and enhance consumer trust.

5.4 Limitations and future research
Review helpfulness prediction is an essential topic in the hospitality and tourism industries and has gained significant attention from scholars. Most studies have reviewed helpfulness prediction by relying on simple and static hand-crafted features. Although this methodology is practical and easy, it requires area-specific knowledge and considerable time to extract and select features. In addition, because the extraction process ignores the semantic and contextual information embedded in review texts, textual information only reflects shallow information. This study aims to incorporate structured and unstructured data by applying DL techniques to predict review helpfulness with high accuracy and interpretability. Our findings highlight the feasibility of comprehensiveness in revealing the informativeness of reviews and predicting helpfulness. The experimental results also demonstrated the effectiveness of the feature-fusion methodology in prediction tasks.

This study makes several important contributions to the existing literature. First, this is one of the first studies to apply DL techniques and consider the complementarity between structured and unstructured data. Second, the proposed model fully captures the dependencies of different types of information and significantly improves the review helpfulness prediction performance. Third, because the determined features were commonly used in previous studies, the proposed model can be easily applied to various domains.

Although the proposed model exhibits better prediction performance, some limitations still exist. The limitations of this study and future research directions are summarized as follows. First, because this study applied a single-channel CNN, the prediction performance was limited. Recently, various studies have applied multichannel CNN to NLP tasks to improve prediction performance. Therefore, future research could apply a multichannel CNN to verify whether such a model improves review helpfulness prediction performance. Second, in addition to the selected features, image features have been analyzed in recent studies that have approached review helpfulness prediction using a multimodal fusion strategy. Therefore, future research could use imaging features to improve the significance and contributions of this study. Finally, although this study considers the dependencies of review content, reviewers and restaurant attributes in predicting review helpfulness, the concept of relative importance is not considered. This study has a limitation in that it assigned weights to attributes based on self-importance. Therefore, future research could apply an attention mechanism that assigns the highest weights to the most relevant attributes to improve prediction performance.

Note
1. [https://www.yelp.com/dataset](https://www.yelp.com/dataset)

References


Corresponding author
Qinglong Li can be contacted at: leecy@khu.ac.kr