electronics

Article

A Multimodal Deep Learning Framework for Consistency-Aware Review Helpfulness Prediction

Seonu Park [1], Xinzhe Li [1], Qinglong Li [2,*] and Jaekyeong Kim [1,3,*]

[1] Department of Big Data Analytics, Kyung Hee University, Seoul 02447, Republic of Korea; sunu0087@khu.ac.kr (S.P.); lixz@khu.ac.kr (X.L.)
[2] Division of Computer Engineering, Hansung University, Seoul 02876, Republic of Korea
[3] School of Management, Kyung Hee University, Seoul 02447, Republic of Korea
* Correspondence: leecy@hansung.ac.kr (Q.L.); jaek@khu.ac.kr (J.K.)

Abstract

Multimodal review helpfulness prediction (MRHP) aims to identify the most helpful reviews by leveraging both textual and visual information. However, prior studies have primarily focused on modeling interactions between these modalities, often overlooking the consistency between review content and ratings, which is a key indicator of review credibility. To address this limitation, we propose CRCNet (Content–Rating Consistency Network), a novel MRHP model that jointly captures the semantic consistency between review content and ratings while modeling the complementary characteristics of text and image modalities. CRCNet employs RoBERTa and VGG-16 to extract semantic and visual features, respectively. A co-attention mechanism is applied to capture the consistency between content and rating, and a Gated Multimodal Unit (GMU) is adopted to integrate consistency-aware representations. Experimental results on two large-scale Amazon review datasets demonstrate that CRCNet outperforms both unimodal and multimodal baselines in terms of MAE, MSE, RMSE, and MAPE. Further analysis confirms the effectiveness of content–rating consistency modeling and the superiority of the proposed fusion strategy. These findings suggest that incorporating semantic consistency into multimodal architectures can substantially improve the accuracy and trustworthiness of review helpfulness prediction.

Keywords: review helpfulness prediction; multimodal; consistency; co-attention mechanism; gated multimodal unit

Academic Editor: Xianzhi Wang

Received: 7 July 2025
Revised: 26 July 2025
Accepted: 31 July 2025
Published: 1 August 2025

Citation: Park, S.; Li, X.; Li, Q.; Kim, J. A Multimodal Deep Learning Framework for Consistency-Aware Review Helpfulness Prediction. *Electronics* **2025**, *14*, 3089. https://doi.org/10.3390/electronics14153089

1. Introduction

E-commerce has become an increasingly important component of modern life due to its cost efficiency, convenience, and rapid growth [1]. As a result, new products are continually being introduced, and various products are actively traded. Most platforms have introduced online review systems that allow consumers to share experiences and assess product quality [2]. Online reviews reduce product uncertainty and provide critical information for consumer decision-making [3,4]. As product variety continues to grow, consumers increasingly depend on online reviews to make informed purchasing decisions [5,6].

However, the growing volume and diversity of reviews make it difficult to identify trustworthy feedback [5,7]. Moreover, not all reviews offer the same value, and consumers tend to rely more on reviews that they find helpful for making purchase decisions [8,9]. To

address this issue, many e-commerce platforms employ helpfulness voting mechanisms to highlight more informative and trustworthy reviews [3,4]. Nevertheless, such mechanisms often prioritize older reviews with more votes, resulting in the underrepresentation of recent but potentially useful feedback [10]. Review Helpfulness Prediction (RHP) has been actively studied to address these issues and to provide consumers with more helpful reviews.

Early studies on RHP primarily utilized traditional machine learning models that incorporated various helpfulness determinants related to the review, reviewer, product, and business [11,12]. These studies identified textual variables as the most influential factors among the diverse determinants. With the advancement of deep learning, there has been a growing interest in capturing the semantic features of review text more effectively. For instance, Mitra and Jenamani [13] explored various approaches such as CNN, LSTM, and traditional machine learning methods to effectively capture the unstructured nature of review text. Furthermore, Saumya et al. [14] proposed a CNN-based model that effectively learns complex local semantic structures embedded in textual content. These studies have demonstrated that the semantic features in review text can effectively enhance the performance of helpfulness prediction. However, they have largely overlooked the visual information provided by review images, which limits the potential performance of helpfulness prediction models.

Text conveys detailed product descriptions and subjective opinions, while images visually show how the product is actually used. To address the challenges of integrating such semantically distinct modalities, Huang et al. [15] proposed an attention-based fusion method designed to combine textual and visual information without loss. Recent studies on RHP have also proposed multimodal prediction models that integrate text and images, leveraging the complementary characteristics of these modalities to improve prediction performance [5,8]. For example, Xiao, Chen, Zhang and Li [8] introduced a model that explicitly differentiates between complementation and substitution relationships between text and images, incorporating them into the loss function. Similarly, Ren et al. [16] enhanced both prediction performance and interpretability by disentangling and integrating internal and intra-modal interactions between the two modalities. These studies consistently report that combining information from both modalities results in superior performance compared to using textual data alone.

Despite these advances, existing studies have largely overlooked the semantic consistency between review content and rating. Consumers consider both the review text and the rating, and both elements influence the perceived helpfulness of a review [17]. While review text and images convey rich and descriptive information, ratings offer a quantifiable summary of the reviewer's overall evaluation. In this study, semantic consistency refers to the degree to which multimodal review content composed of text and image is semantically consistent with the numerical rating.

Figure 1 presents two 5-star reviews from Amazon that differ in their semantic consistency between content and rating. In (a), the review content provides positive and descriptive feedback consistent with the 5-star rating. In contrast, (b) conveys dissatisfaction and shows packaging-related issues, which are misaligned with the high rating. Notably, the consistent review in (a) received a significantly higher number of helpfulness votes, whereas the inconsistent review in (b) received none. This suggests that semantic consistency between content and rating can strongly influence how helpful a review is perceived by other users.

We model semantic consistency between the embeddings of review content and the scalar rating using a co-attention mechanism grounded in vector-space alignment. When review ratings and textual content are consistent, consumers are more likely to trust the review and perceive it as helpful [18]. In contrast, inconsistency between review ratings and

textual content can cause confusion and reduce review credibility [18,19]. Similarly, review images that convey vivid visual cues should be semantically aligned with the ratings to enable accurate helpfulness evaluation. However, studies that consider the consistency between review content and ratings are still limited, and this has rarely been explored in the context of multimodal review helpfulness prediction (MRHP).

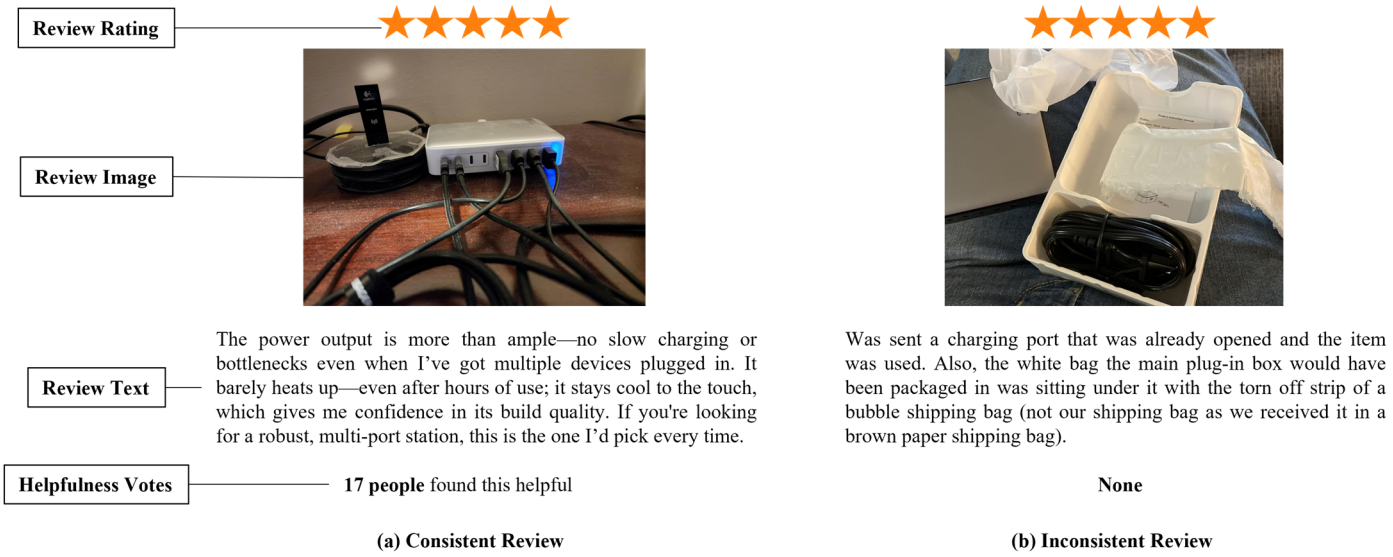

Figure 1. Example of consistent and inconsistent reviews from Amazon.com (accessed on 6 July 2025) (**a**) Consistent Review and (**b**) Inconsistent Review.

To address this limitation, we propose a novel MRHP model called CRCNet (Content–Rating Consistency Network). It captures interactions between textual and visual modalities while modeling the consistency between review content and ratings. Specifically, CRCNet utilizes RoBERTa (Robustly optimized BERT approach) and VGG-16 to extract semantic and visual features from review text and images, respectively. Then, a co-attention mechanism captures the consistency between review content and ratings, and a Gated Multimodal Unit (GMU) is applied to learn the semantic consistency between review text and images. The main contributions of this study are as follows:

- This study introduces CRCNet that incorporates not only the interaction between text and images but also the consistency between ratings and review content. This approach extends prior work by emphasizing the role of consistency between ratings and review content in improving prediction performance.
- CRCNet applies a co-attention mechanism to capture the consistency between review content and ratings and leverages a GMU to integrate text and image features. This approach extends beyond simple feature fusion by effectively modeling relationships across modalities, resulting in better prediction performance.
- We conduct extensive experiments using a large-scale Amazon review dataset, not only comparing baseline models but also validating the effectiveness of the proposed components. The results demonstrate that incorporating the GMU, multimodal integration, and consistency between review content and rating significantly enhances prediction performance across diverse product categories.

This paper is organized as follows. The related studies on text-based and multimodal approaches are reviewed in Section 2. The architecture of CRCNet is detailed in Section 3. The dataset and experimental design are described in Section 4. Experimental results and their interpretation are reported in Section 5. Section 6 concludes the study and discusses future research directions.

2. Related Works

2.1. Review Helpfulness Prediction

RHP aims to identify the most valuable reviews from various available reviews to assist consumers in making informed purchase decisions [16,20]. Although e-commerce platforms generate large volumes of user reviews, not all offer substantial value to consumers [10]. RHP has emerged as a critical technique for automatically identifying reliable and informative reviews, thus enabling consumers to make faster and more accurate purchase decisions [4].

Early studies on RHP primarily utilized traditional machine learning methods. One of the first approaches was based on Support Vector Machines (SVM), which leverages review text and metadata to automatically evaluate review helpfulness [11,12]. Kim, Pantel, Chklovski and Pennacchiotti [11] proposed a supervised RHP model that incorporated structural and linguistic features, whereas Tsur and Rappoport [12] introduced an unsupervised ranking method that identified helpful reviews based on product-related lexical content. Both studies focus on the critical role of textual features in review helpfulness prediction. Furthermore, Lee and Choeh [21] demonstrated that a deep neural network (DNN) outperformed traditional linear regression models, leading to the application of various machine learning models in RHP [4,22,23].

With advancements in deep learning, there has been a surge of interest in employing deep learning models for RHP. Malik [24] provided empirical evidence that DNN outperforms traditional machine learning models by effectively learning nonlinear relationships. Mitra and Jenamani [13] proposed a comprehensive approach that combined lexical, sequential, and structural aspects. Their model employed a deep convolutional neural network (D-CNN) to extract semantic features, a long short-term memory (LSTM) network to capture sequential dependencies, and statistical and syntactic features to enhance predictive power. They also manually evaluated helpfulness scores to improve annotation reliability. Saumya, Singh and Dwivedi [14] argued that previous studies often relied on handcrafted features and failed to fully incorporate contextual information in the review text. To address this, they proposed a two-layer CNN model that effectively learned complex semantic structures from text. This model outperformed existing approaches by learning complex semantic structures in text.

However, most of these studies have focused primarily on textual features, overlooking the visual information in review images. Currently, review images are frequently included alongside text, and such multimodal reviews tend to convey more direct and concrete information compared to text-only reviews [25,26]. In recent years, there has been a growing interest in examining the role of multimodal information in RHP [8,27]. Therefore, relying solely on textual information through a unimodal approach is insufficient for accurate review helpfulness prediction, as it fails to incorporate visual information.

2.2. Multimodal Review Helpfulness Prediction

In recent years, consumer experience sharing has shifted from text-based to image-oriented [25]. This trend reflects a growing preference for vivid and easily produced images over lengthy textual descriptions when expressing personal opinions. Recognizing this change, recent studies on RHP have begun to examine how visual information can contribute to helpfulness prediction rather than relying solely on textual analysis [8]. MRHP seeks to improve predictive accuracy by modeling the interactions between textual and visual modalities, thereby enabling more informative and trustworthy review assessments [27].

Ma, Xiang, Du and Fan [25] was one of the first studies to explore the impact of user-provided images on review helpfulness. They found that while images alone are

insufficient for assessing helpfulness, they can complement or reinforce the effects of review texts. Huang, Zhang, Zhao, Xu and Li [15] pointed out that traditional multimodal fusion methods struggle to effectively integrate the semantic information from inherently heterogeneous visual content and textual descriptions. To address this, they proposed Deep Multimodal Attentive Fusion (DMAF) which leverages attention mechanisms and deep fusion techniques based on intermediate fusion strategies. These studies highlighted the potential of multimodal approaches and the importance of images in review helpfulness prediction.

Xiao, Chen, Zhang and Li [8] pointed out that although many studies have examined the complementarity between text and images, specific approaches for MRHP remain limited. In response, they introduced the complementation-substitution enhanced interactive multimodal deep learning method (CS-IMD), which explicitly incorporates both complementation and substitution effects between modalities. Furthermore, Ren, Diao and Kim [16] proposed the Disentangled Multi-level Fusion Network (DMFN), which captures the complementarity between text and images at multiple semantic levels. This model provided valuable insights into how each modality interacts in specific contexts.

Meanwhile, several studies have extended MRHP tasks to include not only text descriptions and images but also metadata. Zheng, Lin, Zhang, Jiao, Su, Tan, Fan, Xu and Law [3] focused on the impact of reviews, reviewer profiles, and business-/product-related attributes on MRHP performance. Their results demonstrated that traditional determinants of helpfulness can effectively complement textual content and images, thereby contributing to improved MRHP performance. In addition, Ren, Diao, Guo and Hong [27] enhanced the accuracy and interpretability of MRHP by utilizing both deep and handcrafted features from text and images.

These studies have proposed various multimodal approaches to improve MRHP performance. These models primarily aim to improve the accuracy of review helpfulness by enhancing the informativeness of reviews through the integration of text and images. However, despite the advancements in MRHP, the ratings provided alongside review content have often been overlooked. A previous study [3] has partially examined the influence of ratings on review helpfulness, but they did not consider the relationship between ratings and review content.

While textual and visual data play a crucial role in conveying detailed and perceptual aspects of the review, the rating serves as a quantitative summary of the consumer's overall satisfaction and sentiment. Inconsistencies between review content and ratings can create confusion and undermine the perceived credibility of the review [18,19]. In contrast, when the textual and visual content aligns with the rating, consumers are more likely to perceive the review as trustworthy and helpful. To address this limitation, this study proposes CRCNet, which comprehensively incorporates the semantic consistency between review content (text and images) and ratings. By modeling this consistency, CRCNet aims to enhance the accuracy and reliability of review helpfulness prediction.

3. CRCNet Framework

This study proposes CRCNet, which integrates textual, visual, and rating information to accurately predict review helpfulness. It captures not only the complementary characteristics of text and images but also the semantic consistency between review content and ratings.

The following sections describe the overall architecture of CRCNet, which consists of four main modules: (1) a multimodal representation module, which encodes semantic and visual features from review text and images using RoBERTa and VGG-16, respectively; (2) a rating embedding module, which transforms scalar rating values into dense vectors to

enable semantic-level comparison with review content; (3) a semantic consistency module, which models the consistency between review content and rating using co-attention and fuses the consistency-aware features via a GMU; and (4) a helpfulness prediction module, which estimates the helpfulness score based on the fused representation. An overview of CRCNet's framework is illustrated in Figure 2.

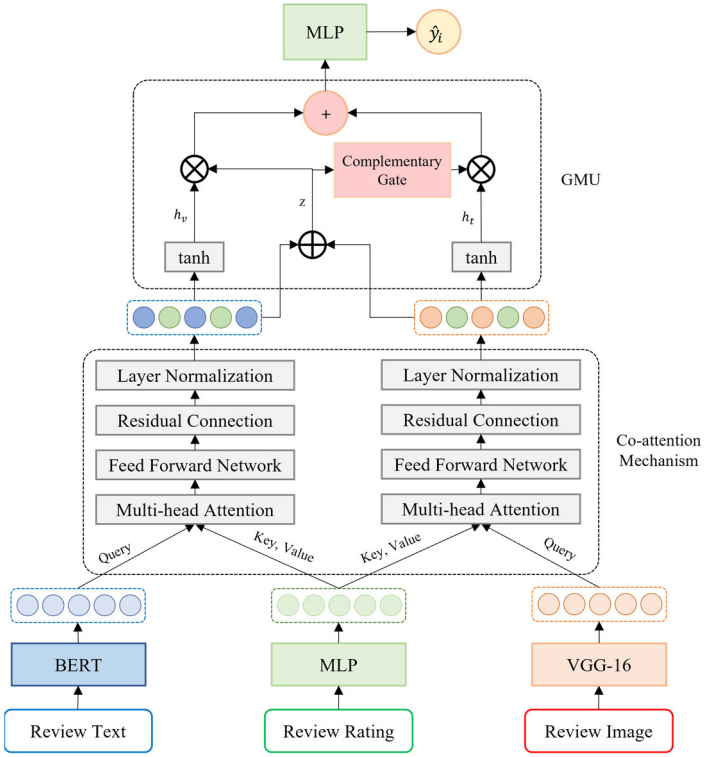

Figure 2. Proposed CRCNet framework.

3.1. Multimodal Representation Module

The multimodal representation module extracts feature representations from both review text and review images. For textual information, we employ RoBERTa to generate contextual embeddings of the review text. RoBERTa [28] is a robust variant of BERT that removes the next sentence prediction task and is trained on larger corpora. Compared to the original BERT, RoBERTa has shown improved performance across a variety of downstream NLP tasks due to its enhanced pretraining strategy and training stability. In our study, it is used to capture rich contextual semantics from the review content.

We extract the final-layer [CLS] token of RoBERTa as a textual representation of the review. This token captures a global representation of the input sequence and results in a 768-dimensional vector. The process of extracting textual representation from the review text T using RoBERTa is defined as follows:

$$h_{text} = RoBERTa(T), \tag{1}$$

where the h_{text} denotes the resulting 768-dimensional feature vector. For visual information, we employ VGG-16 to extract visual embeddings from review images. VGG-16 [29] is a deep convolutional neural network that has been widely used in various computer vision tasks due to its strong capability in capturing fine-grained visual patterns. It has also been adopted in recent MRHP studies to extract rich visual cues embedded in review images [8,16,27].

The review image is first converted into a 4096-dimensional feature vector using a pre-trained VGG-16. Specifically, we use the output of the penultimate fully connected

layer, which is known to capture high-level visual representations. This vector is then linearly projected to a 1024-dimensional space for subsequent processing. The process of extracting visual representation from the review image I using VGG-16 is defined as follows:

$$h_{image} = VGG16(I), \tag{2}$$

3.2. Rating Embedding Module

In this study, we consider not only the complementarity between text and images but also the consistency between review content and ratings. However, scalar rating scores differ in structure and representation format compared to multimodal review content. To address this, we follow a vector-space alignment strategy that enables the integration of heterogeneous modalities by transforming different input formats into a shared representational space [30,31].

Following this approach, we apply a multilayer perceptron (MLP) to project the rating into a dense vector with the same dimensionality as the review representation. This allows the rating and the review content to have equivalent representation capabilities [20]. The transformation of the scalar rating score $r \in R$ into a dense embedding via an MLP is defined as follows:

$$\begin{aligned} a_1 &= \sigma(W_1 r + b_1), \\ &\vdots \\ h_{rating} &= \sigma(W_{n-1} a_{n-1} + b_{n-1}), \end{aligned} \tag{3}$$

where a_n denotes the output of the n-th layer, W_n and b_n represent the corresponding weight matrix and bias. σ is the activation function applied in each layer. The resulting vector h_{rating} represents the rating embedding, which is subsequently passed to the semantic consistency module to model consistency with the review content.

3.3. Semantic Consistency Module

The semantic consistency module focuses on modeling the consistency between review content and ratings, as well as capturing the complementarity across modalities. First, we adopt a co-attention mechanism to effectively model the consistency between review content and ratings. In particular, we independently capture the consistency between text–rating and image–rating pairs.

Inspired by prior work, we implement a co-attention mechanism based on the multi-head attention structure of the Transformer [27,32]. This design allows the model to attend to the aspects of review text and images that semantically correspond to the numeric rating from multiple perspectives, thereby enhancing consistency modeling. The attention mechanism is formally defined as follows:

$$Attention(Q,\ K,\ V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{4}$$

where Q, K, V denote the query, key, and value matrices, respectively. The attention mechanism computes a weighted sum of the value vectors, where the weights are determined by the scaled dot-product similarity between queries and keys. This allows the model to selectively focus on the most relevant parts of the input sequence.

Multi-head attention builds upon the standard attention mechanism by employing multiple attention heads in parallel, each with its own set of projection parameters. This allows the model to capture diverse semantic relationships from different subspaces, thereby

enriching the representational capacity of the attention mechanism. The multi-head attention mechanism is formally defined as follows:

$$\begin{aligned} MultiHead(Q,\ K,\ V) &= [head_1\ ;\ \ldots\ ;\ head_h]W^O, \\ where\ head_i &= Attention\left(QW_i^Q,\ KW_i^K,\ VW_i^K\right), \end{aligned} \tag{5}$$

where QW_i^Q, KW_i^K, and VW_i^K denote the projection matrices for the query, key, and value in the i-th attention head, respectively. W^O is the output projection matrix used to combine the outputs from all heads. The operator ; indicates vector concatenation across attention heads, and i indicates the head index.

Building on this, we implement co-attention by treating the review content (text or image) as the query and the rating representation as the key and value. Co-attention enables mutual referencing between the two modalities, allowing each representation to selectively attend to the most relevant features in the other [32]. While the attention weights reflect semantic relevance, the resulting representations allow the model to compare and reason about the degree of consistency between the modalities. This architecture therefore supports learning semantic consistency by emphasizing mutually informative components across modalities [31]. To allow each modality to interact independently with the rating representation and to capture modality-specific aspects of semantic consistency, we design separate co-attention modules for text and image inputs.

Through the preceding embedding process, the text representation h_{text}, the image representation h_{image}, and the rating representation h_{rating} are placed in the same representational space. Therefore, the process of modeling the consistency between each review modality and the rating through the co-attention module is defined as follows:

$$\begin{aligned} \widetilde{T} &= MultiHead\left(h_{text},\ h_{rating},\ h_{rating}\right), \\ \widetilde{I} &= MultiHead\left(h_{image},\ h_{rating},\ h_{rating}\right), \end{aligned} \tag{6}$$

where $\widetilde{T}$ and $\widetilde{I}$ denote the text and image representations that attend to aspects semantically aligned with the rating, respectively. Subsequently, the intermediate representations $\widetilde{T}$ and $\widetilde{I}$ are passed through a Feed-Forward Network (FFN), followed by a residual connection and Layer Normalization (LN). The FFN allows the model to capture complex feature interactions that may be important for expressing consistency. LN helps the model train more stably and maintain the representations balanced across different inputs. This transformation makes the attended representations more expressive and structured. It also enables more stable and accurate modeling of consistency. This process is defined as follows:

$$\begin{aligned} \widetilde{T}_{co} &= LN(FFN(\widetilde{T}) + \widetilde{T}), \\ \widetilde{I}_{co} &= LN(FFN(\widetilde{I}) + \widetilde{I}), \end{aligned} \tag{7}$$

where $\widetilde{T}_{co}$ and $\widetilde{I}_{co}$ denote the final co-attentive representations of the text and image modalities, respectively, each reflecting the consistency with the rating representation.

Subsequently, to capture the complementarity between the consistency-aware representations, we adopt the GMU [33]. It employs a sigmoid-based gating mechanism to dynamically adjust the contribution of each modality and selectively fuse their features. To facilitate this process, each consistency-aware representation is first transformed using a tanh activation function, resulting in modality-specific vectors $\widetilde{h}_t$ and $\widetilde{h}_v$. A gating vector

is then computed from these vectors and used to adaptively fuse them into a unified representation. The GMU process is defined as follows:

$$\begin{aligned} z &= \sigma(W_z[\widetilde{h}_t; \widetilde{h}_v] + b_z), \\ h &= z \odot \widetilde{h}_t + (1 - z) \odot \widetilde{h}_v, \end{aligned} \tag{8}$$

where z denotes the computed gating vector, W_z and b_z represent the weight matrix and bias used to generate the gate. The gating vector z adjusts the relative contribution of each modality by assigning a higher weight to the more informative modality. The operator $\odot$ denotes the element-wise product, while ; denotes the concatenate operation. The final fused representation h denotes the consistency-aware multimodal representation obtained through the GMU process, which is passed to the final prediction module.

3.4. Helpfulness Prediction Module

The helpfulness prediction module focuses on predicting review helpfulness. In our study, we not only consider the complementarity across modalities but also incorporate the consistency between review content and rating into the prediction process. To this end, the consistency-aware multimodal representation h obtained from the semantic consistency module is fed into an MLP to predict the helpfulness score. This process is defined as follows:

$$\begin{aligned} f_1 &= \sigma(W_1 h \ + \ b_1), \\ &\vdots \\ \hat{y} &= \sigma(W_{n-1} f_{n-1} \ + \ b_{n-1}). \end{aligned} \tag{9}$$

where f_n denotes the output of the n-th layer, W_n and b_n represent its corresponding weight matrix and bias, respectively. We define review helpfulness prediction as a regression task, where the actual number of helpful votes is a non-negative value. Therefore, the activation function σ is chosen as ReLU (Rectified Linear Unit) to ensure non-negativity of the output. $\hat{y}$ is the final predicted helpfulness score, which is trained to minimize the error with respect to the actual helpfulness score.

4. Experiments

4.1. Datasets

This study employs the Amazon review dataset to develop and evaluate the review helpfulness prediction model [34,35]. The dataset consists of product reviews written by consumers after purchasing products on the platform, offering a large volume of data across various categories. For this study, we focused on datasets from the Cell Phones and Accessories, and Electronics categories, which are frequently used in prior RHP research. The dataset contains multiple attributes, such as review text, image, ratings, and helpfulness votes, making it suitable for training models to predict review helpfulness.

To ensure data quality, the following preprocessing steps were applied. First, reviews without text or images were excluded from the analysis, as their absence could introduce noise into the model training process. All text was converted to lowercase, and irrelevant characters, such as special symbols and numbers, were removed. Stop words that do not contribute to the semantic meaning were also eliminated, leaving only the most relevant information from the reviews [5,20]. Additionally, reviews with no helpfulness votes were excluded, as reviews that users have not evaluated could adversely affect the performance of the prediction model [13,36]. Thus, only reviews with at least one helpful vote were included in the training process. A logarithmic transformation was applied to the helpfulness vote counts to mitigate the issue of excessive variance in helpfulness votes and improve model stability. Following the previous study, we added one to each helpfulness

vote before applying the logarithmic transformation, ensuring the avoidance of zero values and producing a more balanced data distribution [3,37,38].

After preprocessing, we obtained 64,590 reviews from the Cell Phones and Accessories category and 163,449 reviews from the Electronics category. Descriptive statistics for the datasets are summarized in Table 1.

Table 1. Descriptive statistics of the Amazon review dataset.

Dataset	Component	Min	Max	Mean	Std. Dev.
Cell Phones and Accessories	Number of helpfulness votes	1	4149	12.334	39.068
	Number of helpfulness votes (logarithm)	1.099	8.820	2.106	0.991
Electronics	Number of helpfulness votes	1	6770	17.256	64.381
	Number of helpfulness votes (logarithm)	1.099	8.331	1.956	0.898

4.2. Evaluation Metrics

In this study, the predictive performance of the model was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) [13,27]. MAE measures the average of the absolute differences between predicted and actual values. MSE measures the mean of the squared differences between predicted and actual values, which penalizes larger errors more heavily. This metric is particularly useful for assessing the overall error magnitude in the model's predictions. RMSE is the square root of MSE and represents the prediction error in the same unit as the target variable. MAPE expresses the error as a percentage of the actual values and evaluates the model's relative accuracy.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{10}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{12}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100. \tag{13}$$

Here, N denotes the number of reviews in the test set, y_i represents the log-transformed actual helpfulness score of the i-th review, and $\hat{y}_i$ is the predicted helpfulness score. By using these four metrics, we can evaluate the model's performance in terms of absolute error magnitude, sensitivity to outliers, and relative accuracy.

4.3. Baseline Models

Unlike prior studies that mainly focused on interactions between modalities, this study introduces ratings as an additional modality and incorporates the consistency between ratings and review content into the prediction process. To validate the effectiveness of CRCNet, we conducted comparative experiments with baseline models. Specifically, the baseline models include three text-based models and three multimodal models. A detailed description of the baseline models used in the comparison is as follows:

- D-CNN [13]: A dual-layer CNN model designed to extract semantic features from review text. It applies convolutional kernels of sizes 2, 3, and 4 in the first layer to capture n-gram-level patterns, followed by a second convolutional layer that generates document-level representations. The final prediction is obtained by averaging the

outputs from each kernel size, allowing the model to effectively capture local semantic patterns within reviews.

- LSTM [13]: A sequential model designed to capture long-range contextual dependencies in review text. Similarly to the D-CNN model, it uses 50-dimensional GloVe embeddings as input, followed by a 100-cell LSTM layer and a fully connected layer for prediction. A dropout rate of 0.09 is applied to reduce overfitting during training. In contrast to the D-CNN, which captures local semantic patterns, the LSTM captures the overall semantic flow.
- TNN [39]: A multi-channel CNN model that captures local semantic patterns from multiple perspectives. It uses 100-dimensional GloVe embeddings and applies three parallel 1D convolutional layers with kernel sizes of 3, 4, and 5, each with 100 filters. The max-pooled outputs are concatenated and passed through a fully connected layer. This model learns complementary semantic features from multiple n-gram perspectives, which contribute to effective review helpfulness prediction.
- DMAF [15]: A multimodal fusion model that emphasizes the most discriminative features within each modality through visual and semantic attention mechanisms. In addition, it incorporates a multimodal attention mechanism to integrate complementary multimodal features. The model adopts a deep intermediate fusion strategy to jointly learn modality-specific and unified representations.
- CS-IMD [8]: An MRHP model that explicitly distinguishes between complementary and substitutive relationships between text and images. Multimodal features are extracted using pre-trained BERT and VGG-16 and then processed through attention-based modules designed to capture the relative importance of each modality component. This approach captures both shared and modality-specific contributions to review helpfulness prediction by jointly optimizing complementation and substitution loss terms.
- MFRHP [27]: An MRHP model that incorporates both hand-crafted and deep features from text and images. For textual input, it uses review length, readability scores, and BERT embeddings; for visual input, it adopts pixel brightness and VGG-16 features. These are fused using a co-attention mechanism to capture multimodal interactions, thereby enhancing both prediction accuracy and interpretability.

4.4. Hyperparameter Settings and Experimental Environment

In this study, the dataset was split into training, validation, and test sets at a 7:1:2 ratio. The training set was used to fit the model, the validation set for hyperparameter tuning, and the test set for final performance evaluation. For the hyperparameter settings, the learning rate was tuned within the range of $\{0.001, 0.005, 0.01, 0.05\}$, and the batch size was tuned within the range of $\{64, 128, 256, 512\}$. To optimize the model's loss function, we adopted the Adaptive Moment Estimation (Adam) optimizer. Additionally, in the attention mechanism, the dimensions of the query (Q), key (K), and value (V) vectors were set to 64 to maintain computational efficiency while providing sufficient representational power [40]. As a result, we set the learning rate to 0.001, the batch size to 128, and the number of attention heads to 4.

The number of epochs was set to 100 for all experiments, and early stopping was applied to prevent overfitting, with training halted if the loss function did not decrease for five consecutive epochs. To reduce the randomness of the experimental results, each experiment was repeated five times, and the average of the results was used as the final performance metric. The hyperparameters for the baseline models used in the performance comparison were maintained as defined in the original studies. All experiments were

conducted on a workstation equipped with an Intel(R) Core(TM) i9-10900K CPU, 128GB RAM, and an NVIDIA GeForce RTX 3090 GPU.

5. Experimental Results and Discussions

5.1. Model Performance Comparison

To demonstrate the effectiveness of CRCNet, we conducted a comparative evaluation against baseline models. As a result, CRCNet outperformed all the baseline models across all evaluation metrics. Specifically, it achieved an average improvement of 2.97% in MAE, 4.07% in MSE, 2.07% in RMSE, and 4.82% in MAPE compared to the baseline models. The detailed experimental results are presented in Table 2.

Table 2. Performance comparison of MRHP models on the Amazon dataset.

Dataset	Model	MAE	MSE	RMSE	MAPE
Cell Phones and Accessories	D-CNN	0.704 ± 0.002	0.792 ± 0.002	0.890 ± 0.001	40.483 ± 0.203
	LSTM	0.706 ± 0.000	0.806 ± 0.000	0.898 ± 0.000	40.053 ± 0.061
	TNN	0.697 ± 0.007	0.788 ± 0.008	0.888 ± 0.004	39.836 ± 0.816
	DMAF	0.686 ± 0.001	0.786 ± 0.002	0.887 ± 0.001	38.110 ± 0.166
	CS-IMD	0.683 ± 0.001	0.777 ± 0.000	0.882 ± 0.000	37.942 ± 0.081
	MFRHP	0.677 ± 0.004	0.767 ± 0.008	0.876 ± 0.005	37.359 ± 0.287
	CRCNet	0.675 ± 0.001	0.765 ± 0.001	0.875 ± 0.001	37.236 ± 0.163
Electronics	D-CNN	0.750 ± 0.003	0.926 ± 0.002	0.962 ± 0.001	40.375 ± 0.459
	LSTM	0.745 ± 0.005	0.930 ± 0.002	0.964 ± 0.001	39.255 ± 0.758
	TNN	0.752 ± 0.004	0.923 ± 0.010	0.961 ± 0.005	40.843 ± 0.745
	DMAF	0.738 ± 0.000	0.930 ± 0.000	0.964 ± 0.000	38.612 ± 0.043
	CS-IMD	0.730 ± 0.004	0.888 ± 0.001	0.943 ± 0.001	38.725 ± 0.598
	MFRHP	0.728 ± 0.004	0.883 ± 0.001	0.940 ± 0.001	38.562 ± 0.519
	CRCNet	0.720 ± 0.004	0.877 ± 0.002	0.936 ± 0.001	37.634 ± 0.605

Notably, text-based models such as D-CNN, TNN, and LSTM showed inferior performance compared to multimodal approaches. Although these models are based on different architectures, their performance remained similar. This suggests that approaches focusing solely on local patterns or sequential dependencies in the text have limited capacity for improvement. Moreover, these models fail to incorporate complementary modalities such as images or ratings. As a result, unimodal approaches based only on review text are unable to reflect critical visual cues or quantitative assessments essential for consumer decision-making, leading to constrained predictive performance.

The multimodal models DMAF, CS-IMD, and MFRHP demonstrated improved performance by learning the interactions between text and images. Among them, MFRHP achieved the highest performance, particularly due to its integration of both deep and shallow features from reviews. This finding suggests that considering the relationships between review content and auxiliary information can enhance predictive accuracy. However, these models do not consider the semantic consistency between review content and ratings, limiting their ability to model consumer trust dynamics.

Based on prior research indicating that consumers are more likely to trust reviews when the ratings and content are consistent [9,19,41], CRCNet enhances prediction performance by capturing the interactions between review content and ratings. By jointly leveraging review text, images, and rating consistency, CRCNet achieved superior accuracy across all evaluation metrics. These findings demonstrate that a multimodal approach incorporating not only textual and visual information but also rating-content consistency significantly improves the effectiveness of review helpfulness prediction.

5.2. Effectiveness Analysis of Consistency Modeling in RHP

In this study, we focused on the consistency between review content and rating from a multimodal perspective. To evaluate the effectiveness of this approach, we compared three consistency models: TR (review text and rating), IR (review image and rating), and MR (multimodal review and rating). The results show that the IR model exhibited the lowest performance, while the MR model (CRCNet) achieved the highest performance. The detailed results are presented in Figure 3.

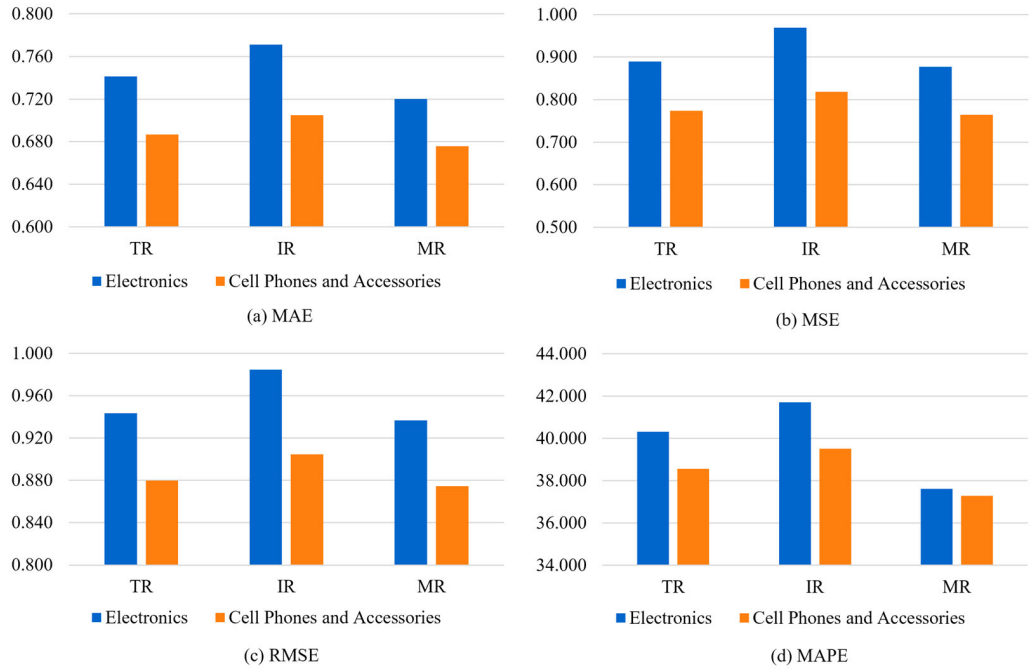

Figure 3. Performance comparison of content–rating consistency models (**a**) MAE, (**b**) MSE, (**c**) RMSE and (**d**) MAPE.

The TR model predicts review helpfulness based on the consistency between review text and ratings. This approach shows moderate performance, as textual descriptions and emotional expressions often align with the rating score. However, it utilizes only textual information and thus cannot leverage the additional cues provided by visual content in the reviews.

In contrast, the IR model focuses on consistency between review images and ratings. However, it faced challenges in predicting consistency based solely on visual characteristics. While review images provide visual context, they are limited in conveying detailed information for predicting review helpfulness. This limitation resulted in lower performance compared to the TR model.

The MR model, which considers the interactions among text, images, and ratings, demonstrated the best performance. This can be explained by the complementary relationship between text and images, as well as the role of ratings. Text provides detailed descriptions and emotional content, while images offer visual context. In contrast, ratings reflect consumers' overall experiences as quantitative summaries. Since these elements are interrelated, considering their consistency can significantly improve prediction accuracy. These results demonstrate that modeling consistency between review content and ratings from a multimodal perspective is an effective strategy for improving RHP performance.

5.3. Effectiveness Analysis of Multimodal Fusion Methods

In this study, we adopted the GMU to integrate consistency-aware representations from different modalities. To evaluate its effectiveness, we compared three fusion methods:

concatenation, element-wise product, and GMU. The experimental results show that the GMU achieves the best performance across all metrics. The detailed experimental results are presented in Table 3.

Table 3. Performance comparison for multimodal fusion methods.

Dataset	Method	MAE	MSE	RMSE	MAPE
Cell Phones and Accessories	Concatenation	0.698 ± 0.000	0.794 ± 0.000	0.891 ± 0.000	39.400 ± 0.006
	Element-wise product	0.703 ± 0.000	0.792 ± 0.000	0.890 ± 0.000	40.335 ± 0.005
	GMU	0.676 ± 0.001	0.765 ± 0.001	0.875 ± 0.001	37.279 ± 0.127
Electronics	Concatenation	0.733 ± 0.000	0.894 ± 0.000	0.946 ± 0.000	38.948 ± 0.001
	Element-wise product	0.739 ± 0.000	0.895 ± 0.000	0.946 ± 0.000	39.829 ± 0.003
	GMU	0.720 ± 0.004	0.877 ± 0.002	0.936 ± 0.001	37.610 ± 0.521

The concatenation simply stacks feature vectors from each modality. While this preserves modality-specific information and leads to a moderate performance improvement, it lacks the capacity to capture multimodal interactions, resulting in lower performance than GMU. These results suggest that although simple fusion methods can offer gains, modeling modality interactions is essential for improved predictive accuracy.

The element-wise product fuses modalities by aligning corresponding features across dimensions. Although it captures some linear-level interaction, its ability to model complex relationships is limited. As a result, it underperforms in comparison to more expressive fusion techniques.

By contrast, the GMU outperforms all fusion methods. Through a gating mechanism that dynamically adjusts the contribution of each modality, GMU effectively captures the complex interactions between text, images, and ratings. Each modality contributes complementary information, but their importance is not necessarily equal. As such, simple fusion methods that treat all modalities equally or fail to account for their interactions are often insufficient when dealing with complex representations. This advantage becomes more pronounced when integrating complex features that reflect the consistency with review ratings. Future research could extend this approach and explore its application in other multimodal data scenarios.

5.4. Efficiency Analysis of the Rating Embedding Mechanism

We leveraged the semantic consistency between review content and ratings as a key component for enhancing the accuracy of review helpfulness prediction. To empirically validate this architecture, we conducted an ablation study comparing CRCNet with a variant (w/o rating) that excludes rating information. CRCNet is the complete model that integrates a co-attention mechanism to capture consistency between the review content and the rating representation. In contrast, the variant removes the rating embedding and relies solely on the interaction between text and image modalities. The detailed experimental results are presented in Table 4.

Table 4. Performance comparison for rating embedding ablation.

Dataset	Model	MAE	MSE	RMSE	MAPE
Cell Phones and Accessories	CRCNet	0.676 ± 0.001	0.765 ± 0.001	0.875 ± 0.001	37.279 ± 0.127
	w/o rating	0.696 ± 0.001	0.819 ± 0.000	0.905 ± 0.000	38.297 ± 0.073
Electronics	CRCNet	0.720 ± 0.004	0.877 ± 0.002	0.936 ± 0.001	37.610 ± 0.521
	w/o rating	0.737 ± 0.001	0.901 ± 0.000	0.949 ± 0.000	39.267 ± 0.001

As a result, CRCNet consistently outperformed the w/o rating across all evaluation metrics in both product categories. This performance gap indicates that removing rating information hinders the model's ability to accurately predict review helpfulness. When reading reviews, consumers not only consider the content but also refer to the accompanying rating [17]. If the review text and rating are semantically inconsistent, users may become confused, question the credibility of the review, and find it difficult to assess its helpfulness [18,19]. Without modeling this consistency, the system may incorrectly interpret incoherent or misleading reviews as helpful. These results support our central argument that modeling semantic consistency between review content and ratings enhances the performance of multimodal review helpfulness prediction.

5.5. Efficiency Analysis of the MRHP Model

In this section, we examine the training efficiency of CRCNet in comparison with other state-of-the-art MRHP models. While achieving high prediction accuracy is important, models with excessive training time may not be suitable for real-world deployment. To assess efficiency, we measured the average training time per epoch and predictive performance for each model on both datasets. CRCNet achieved the highest prediction accuracy across both datasets while also maintaining competitive training time compared to other multimodal baselines. The results are summarized in Table 5.

Table 5. Efficiency and performance comparison of MRHP models.

Dataset	Model	MAE	MSE	RMSE	MAPE	Training Time (s)
Cell Phones and Accessories	CS-IMD	0.683 ± 0.001	0.777 ± 0.000	0.882 ± 0.000	37.942 ± 0.081	4
	DMAF	0.686 ± 0.001	0.786 ± 0.002	0.887 ± 0.001	38.110 ± 0.166	9
	MFRHP	0.677 ± 0.004	0.767 ± 0.008	0.876 ± 0.005	37.359 ± 0.287	13
	CRCNet	0.675 ± 0.001	0.765 ± 0.001	0.875 ± 0.001	37.236 ± 0.163	8
Electronics	CS-IMD	0.730 ± 0.004	0.888 ± 0.001	0.943 ± 0.001	38.725 ± 0.598	8
	DMAF	0.738 ± 0.000	0.930 ± 0.000	0.964 ± 0.000	38.612 ± 0.043	19
	MFRHP	0.728 ± 0.004	0.883 ± 0.001	0.940 ± 0.001	38.562 ± 0.519	34
	CRCNet	0.720 ± 0.004	0.877 ± 0.002	0.936 ± 0.001	37.634 ± 0.605	14

CRCNet consistently outperformed all baselines on both datasets. Specifically, it outperformed MFRHP and DMAF in both prediction accuracy and training efficiency, reducing training time per epoch by 38.5% and 58.8% in the Cell Phones and Accessories and Electronics datasets, respectively. Although its training time was slightly longer than that of CS-IMD, CRCNet still achieved higher accuracy, with MAE improvements of 1.17% and 1.37%, respectively. These findings highlight that CRCNet achieves an effective balance between predictive accuracy and computational efficiency. By delivering the highest accuracy without incurring excessive training costs, CRCNet demonstrates strong potential for practical deployment in real-world multimodal review processing tasks.

6. Conclusions and Future Work

This study aims to enhance existing MRHP approaches, which mainly focus on the interaction between text and images while overlooking the role of associated review ratings. To address this limitation, we proposed CRCNet, which explicitly models the semantic consistency between review content and ratings during prediction. The model employs a co-attention mechanism to capture the consistency between review content and ratings and incorporates a GMU to capture complex interactions between text and images. This architecture effectively captures the complex multimodal features that reflect rating consistency, leading to improved predictive performance. CRCNet outperformed both unimodal

models such as D-CNN, LSTM, and TNN, as well as multimodal models such as DMAF, CS-IMD, and MFRHP across all evaluation metrics. These results demonstrate that a multimodal approach that integrates consistency information can effectively enhance the performance of review helpfulness prediction.

CRCNet makes significant contributions to the RHP; however, there remain potential directions for further research. First, this study evaluated CRCNet on two Amazon product categories, which mainly represent search goods. However, review characteristics and the perceived helpfulness can vary by product type. Therefore, CRCNet's generalizability to experience goods such as hospitality or tourism remains unverified. Future research should evaluate CRCNet across a broader range of product domains. Second, we adopted BERT and VGG-16 for feature extraction to enable fair comparisons with existing MRHP studies. However, recent multimodal models based on large language models (LLMs), demonstrate stronger capabilities in capturing complex cross-modal interactions. Future work will explore these LLM-based encoders to improve consistency modeling and benchmark CRCNet against more advanced baselines. Third, CRCNet focuses on modeling rating-content consistency through a co-attention mechanism. However, consistency can also be examined across reviews for the same item or written by the same reviewer. In addition, ratings may be influenced by external factors that are not directly observable in the review content, such as delivery speed or customer service. Future work could incorporate metadata or infer such external signals to develop more comprehensive and robust RHP models. Finally, this study focused on reviews with at least one vote and did not account for temporal or social biases in helpfulness votes. Early reviews may receive more votes due to longer exposure, and reviewer reputation may affect perceived helpfulness. Future work should address such biases during modeling or preprocessing.

Author Contributions: Conceptualization, S.P., X.L. and Q.L.; methodology, S.P. and X.L.; software, X.L. and Q.L.; data curation, S.P., X.L. and J.K.; writing—original draft preparation, S.P. and Q.L.; writing—review and editing, X.L., Q.L. and J.K.; supervision, Q.L. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by Hansung University.

Data Availability Statement: The data are available on https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/ (accessed on 12 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Q.; Li, X.; Lee, B.; Kim, J. A hybrid CNN-based review helpfulness filtering model for improving e-commerce recommendation Service. *Appl. Sci.* **2021**, *11*, 8613. [CrossRef]
2. Siering, M.; Muntermann, J.; Rajagopalan, B. Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. *Decis. Support Syst.* **2018**, *108*, 1–12. [CrossRef]
3. Zheng, T.; Lin, Z.; Zhang, Y.; Jiao, Q.; Su, T.; Tan, H.; Fan, Z.; Xu, D.; Law, R. Revisiting review helpfulness prediction: An advanced deep learning model with multimodal input from Yelp. *Int. J. Hosp. Manag.* **2023**, *114*, 103579. [CrossRef]
4. Lee, M.; Kwon, W.; Back, K.-J. Artificial intelligence for hospitality big data analytics: Developing a prediction model of restaurant review helpfulness for customer decision-making. *Int. J. Contemp. Hosp. Manag.* **2021**, *33*, 2117–2136. [CrossRef]
5. Li, X.; Li, Q.; Kim, J. A Review Helpfulness Modeling Mechanism for Online E-commerce: Multi-Channel CNN End-to-End Approach. *Appl. Artif. Intell.* **2023**, *37*, 2166226. [CrossRef]
6. Kwon, W.; Lee, M.; Back, K.-J.; Lee, K.Y. Assessing restaurant review helpfulness through big data: Dual-process and social influence theory. *J. Hosp. Tour. Technol.* **2021**, *12*, 177–195. [CrossRef]
7. Moro, S.; Esmerado, J. An integrated model to explain online review helpfulness in hospitality. *J. Hosp. Tour. Technol.* **2021**, *12*, 239–253. [CrossRef]
8. Xiao, S.; Chen, G.; Zhang, C.; Li, X. Complementary or substitutive? A novel deep learning method to leverage text-image interactions for multimodal review helpfulness prediction. *Expert Syst. Appl.* **2022**, *208*, 118138. [CrossRef]

9. Baek, H.; Ahn, J.; Choi, Y. Helpfulness of online consumer reviews: Readers' objectives and review cues. *Int. J. Electron. Commer.* **2012**, *17*, 99–126. [CrossRef]
10. Wang, S.; Qiu, J. Utilizing a feature-aware external memory network for helpfulness prediction in e-commerce reviews. *Appl. Soft Comput.* **2023**, *148*, 110923. [CrossRef]
11. Kim, S.-M.; Pantel, P.; Chklovski, T.; Pennacchiotti, M. Automatically assessing review helpfulness. In Proceedings of the 2006 Conference on empirical methods in natural language processing, Sydney, Australia, 22–23 July 2006; pp. 423–430.
12. Tsur, O.; Rappoport, A. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In Proceedings of the International AAAI Conference on Web and Social Media, San Jose, CA, USA, 17–20 May 2009; pp. 154–161.
13. Mitra, S.; Jenamani, M. Helpfulness of online consumer reviews: A multi-perspective approach. *Inf. Process. Manag.* **2021**, *58*, 102538. [CrossRef]
14. Saumya, S.; Singh, J.P.; Dwivedi, Y.K. Predicting the helpfulness score of online reviews using convolutional neural network. *Soft Comput.* **2020**, *24*, 10989–11005. [CrossRef]
15. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowl. Based Syst.* **2019**, *167*, 26–37. [CrossRef]
16. Ren, G.; Diao, L.; Kim, J. DMFN: A disentangled multi-level fusion network for review helpfulness prediction. *Expert Syst. Appl.* **2023**, *228*, 120344. [CrossRef]
17. Mudambi, S.M.; Schuff, D. Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS Q.* **2010**, *34*, 185–200. [CrossRef]
18. Li, Q.; Park, J.; Kim, J. Impact of information consistency in online reviews on consumer behavior in the e-commerce industry: A text mining approach. *Data Technol. Appl.* **2024**, *58*, 132–149. [CrossRef]
19. Aghakhani, N.; Oh, O.; Gregg, D.G.; Karimi, J. Online review consistency matters: An elaboration likelihood model perspective. *Inf. Syst. Front.* **2021**, *23*, 1287–1301. [CrossRef]
20. Li, X.; Li, Q.; Jeong, D.; Kim, J. A novel deep learning method to use feature complementarity for review helpfulness prediction. *J. Hosp. Tour. Technol.* **2024**, *15*, 534–555. [CrossRef]
21. Lee, S.; Choeh, J.Y. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Syst. Appl.* **2014**, *41*, 3041–3046. [CrossRef]
22. Krishnamoorthy, S. Linguistic features for review helpfulness prediction. *Expert Syst. Appl.* **2015**, *42*, 3751–3759. [CrossRef]
23. Hu, Y.-H.; Chen, K.; Lee, P.-J. The effect of user-controllable filters on the prediction of online hotel reviews. *Inf. Manag.* **2017**, *54*, 728–744. [CrossRef]
24. Malik, M.S.I. Predicting users' review helpfulness: The role of significant review and reviewer characteristics. *Soft Comput.* **2020**, *24*, 13913–13928. [CrossRef]
25. Ma, Y.; Xiang, Z.; Du, Q.; Fan, W. Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. *Int. J. Hosp. Manag.* **2018**, *71*, 120–131. [CrossRef]
26. Li, Y.; Xie, Y. Is a picture worth a thousand words? An empirical study of image content and social media engagement. *J. Mark. Res.* **2020**, *57*, 1–19. [CrossRef]
27. Ren, G.; Diao, L.; Guo, F.; Hong, T. A co-attention based multi-modal fusion network for review helpfulness prediction. *Inf. Process Manag.* **2024**, *61*, 103573. [CrossRef]
28. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; Ye, H.-J. Ovis: Structural embedding alignment for multimodal large language model. *arXiv* **2024**, arXiv:2405.20797. [CrossRef]
31. Chen, S.; Song, B.; Guo, J. Attention alignment multimodal LSTM for fine-gained common space learning. *IEEE Access* **2018**, *6*, 20195–20208. [CrossRef]
32. Liu, M.; Liu, L.; Cao, J.; Du, Q. Co-attention network with label embedding for text classification. *Neurocomputing* **2022**, *471*, 61–69. [CrossRef]
33. Arevalo, J.; Solorio, T.; Montes-y-Gómez, M.; González, F.A. Gated multimodal units for information fusion. *arXiv* **2017**, arXiv:1702.01992. [CrossRef]
34. He, R.; McAuley, J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11 April 2016; pp. 507–517.
35. Ni, J.; Li, J.; McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–19 November 2019; pp. 188–197.
36. Lee, M.; Jeong, M.; Lee, J. Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. *Int. J. Contemp. Hosp. Manag.* **2017**, *29*, 762–783. [CrossRef]

37. Bilal, M.; Marjani, M.; Hashem, I.A.T.; Malik, N.; Lali, M.I.U.; Gani, A. Profiling reviewers' social network strength and predicting the "Helpfulness" of online customer reviews. *Electron. Commer. Res. Appl.* **2021**, *45*, 101026. [CrossRef]
38. Zhang, Y.; Lin, Z. Predicting the helpfulness of online product reviews: A multilingual approach. *Electron. Commer. Res. Appl.* **2018**, *27*, 1–10. [CrossRef]
39. Olmedilla, M.; Martínez-Torres, M.R.; Toral, S. Prediction and modelling online reviews helpfulness using 1D Convolutional Neural Networks. *Expert Syst. Appl.* **2022**, *198*, 116787. [CrossRef]
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
41. Lee, S.; Lee, S.; Baek, H. Does the dispersion of online review ratings affect review helpfulness? *Comput. Hum. Behav.* **2021**, *117*, 106670. [CrossRef]