

Article

Enhancing Review-Based Recommendations Through Local and Global Feature Fusion

Namhun Kim ¹, Haebin Lim ¹, Qinglong Li ², Xinzhe Li ¹, Seokkwan Kim ^{3,*} and Jaekyeong Kim ^{1,4}

¹ Department of Big Data Analytics, Kyung Hee University, 26, Kyungheedaero-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea; kimnh097@khu.ac.kr (N.K.); fkid009@khu.ac.kr (H.L.); lixz@khu.ac.kr (X.L.); jaek@khu.ac.kr (J.K.)

² Division of Computer Engineering, Hansung University, 116, Samseongyo-ro 16-gil, Seongbuk-gu, Seoul 02876, Republic of Korea; leecy@hansung.ac.kr

³ Department of Business Administration, Graduate School, Kyung Hee University, 26, Kyungheedaero-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea

⁴ School of Management, Kyung Hee University, 26, Kyungheedaero-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea

* Correspondence: yacall@khu.ac.kr

Abstract

With the rapid advancement of information and communication technology, the number of items users encounter increased exponentially. Consequently, the importance of recommendation systems emerged to reduce the time and effort required for users to make item selections. Recently, among various studies on recommendation systems, there has been significant interest in leveraging review text as auxiliary information. This study proposes a novel model to enhance recommendation performance by effectively analyzing review texts through the fusion of local and global features. By combining convolutional neural networks (CNN), which excel in extracting local features, and the RoBERTa model, renowned for capturing global contextual features, the proposed approach effectively uncovers users' latent preferences embedded within review texts. The proposed model comprises three key components: the user-item interaction module, which learns complex interactions between users and items; the feature extraction module, which extracts both local and global features using CNN and RoBERTa; and the preference prediction module, which combines the output vectors from the previous modules to predict user preferences for specific items. Extensive experiments conducted on three datasets collected from Amazon platform demonstrate that the proposed model significantly outperforms baseline models. These findings highlight the effectiveness of the proposed approach in considering both local and global features for extracting user preferences from review texts.

Keywords: recommender system; review texts; CNN; RoBERTa; local feature; global feature



Academic Editor: Xianzhi Wang

Received: 13 May 2025

Revised: 18 June 2025

Accepted: 19 June 2025

Published: 23 June 2025

Citation: Kim, N.; Lim, H.; Li, Q.; Li, X.; Kim, S.; Kim, J. Enhancing Review-Based Recommendations Through Local and Global Feature Fusion. *Electronics* **2025**, *14*, 2540. <https://doi.org/10.3390/electronics14132540>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the advancement of information and communication technology, the online e-commerce industry experienced continuous growth, leading to the constant introduction of new products and services [1]. While this expansion provides users with a broader range of choices, it also presents challenges in finding items that align with their preferences. To address this issue, e-commerce platforms such as Amazon and Netflix implemented recommendation systems [2]. By offering personalized item recommendations that reflect users' preferences, these systems help reduce the time and effort required for item selection while

simultaneously driving increased revenue for businesses [3]. As a result, recommendation systems have become an essential technology in e-commerce, playing a pivotal role in user decision making [4].

In recent years, there has been growing interest in recommendation systems that leverage user review texts [5]. Review texts, which contain users' detailed experiences and evaluations of items, serve as valuable auxiliary information that can alleviate data sparsity issues and improve recommendation performance [6]. For instance, Zheng et al. [7] proposed a model that effectively extracts user-specific features embedded in review texts, achieving superior recommendation accuracy. Similarly, Liu et al. [8] introduced a model that emphasizes critical segments of review texts through weighted analysis, enabling more effective text processing. These studies highlight that analyzing review texts effectively not only enhances the accuracy of capturing user preferences, but also addresses the limitations posed by data sparsity in recommendation systems [9].

There are two distinct perspectives for analyzing review texts, the local feature perspective and the global feature perspective [10]. The local feature perspective focuses on individual words, specific word combinations, or phrases, effectively capturing fine-grained semantic details [11]. For example, in the phrase "excellent picture quality," the combination of "picture" and "quality" highlights a key strength of the item. By addressing such specific word combinations, the local perspective identifies the aspects that users consider important within their reviews. On the other hand, the global feature perspective emphasizes the overall context of the review, which allows for a clearer understanding of its overarching intent [12]. For instance, in the review, "The item is expensive, but the quality is excellent," a local analysis might overemphasize the negative expression "expensive". However, considering the full context reveals that the user is expressing a positive overall evaluation of the item. By analyzing the text from a global perspective, the broader intent behind the review can be accurately identified, enabling the system to better understand latent user preferences and deliver more personalized recommendations [7].

To leverage the strengths of both perspectives, this study proposes the local and global feature fusion from review texts (LGFR) model, which effectively captures users' latent intent by considering both local and global features within review texts. The model integrates a convolutional neural network (CNN), well-suited for extracting local features, with the robustly optimized BERT approach (RoBERTa), which excels at capturing global contextual features. By combining these two approaches, the proposed model aims to provide highly personalized recommendations by uncovering users' implicit preferences embedded in reviews. The proposed model consists of three primary components: First, the user-item interaction module, which utilizes matrix factorization (MF) to model the interactions between users and items. Second, the feature extraction module, which extracts and integrates local and global features from reviews to construct a comprehensive textual representation. Third, the preference prediction module, which combines the outputs from the previous modules and employs the multi-layer perceptron (MLP) to predict the ratings that users are likely to assign to new items. The main contributions of this study are as follows:

- This study proposes the LGFR model, which incorporates both local and global features of review texts to fully leverage the strengths of each perspective. By considering these features together, the model enables the development of a recommendation system that more accurately reflects user preferences.
- By utilizing unstructured data, such as review texts, the proposed approach addresses the limitations of traditional models that rely solely on rating-based data. This not only alleviates the data sparsity problem, but also effectively extracts fine-grained user preferences embedded in reviews, thereby enhancing recommendation performance.

- The LGFR model was validated through experiments on multiple categories of datasets from a real-world e-commerce platform, Amazon. The results demonstrate that the proposed model outperforms several baseline models, proving its practicality and generalizability as a robust recommendation system.

The structure of this paper is as follows: Section 2 reviews related works, while Section 3 introduces the recommendation model proposed in this study. Section 4 describes the experimental datasets, evaluation methods, additional experiments, and baseline models. Section 5 summarizes the experimental results and evaluates the components of the proposed model. Finally, Section 6 presents the conclusions and directions for future research.

2. Related Work

2.1. Review-Based Recommender Systems

Recommendation systems emerged as critical technology for providing personalized recommendations to users, effectively addressing the problem of information overload [13]. Among various recommendation algorithms, MF is one of the most widely used methods. MF constructs a matrix based on user–item rating data, extracting latent user preferences and item characteristics [14]. However, since MF relies solely on rating data, it has limitations in capturing the complex, non-linear interactions between users and items. To overcome these shortcomings, He Liao [15] introduced neural matrix factorization (NeuMF), which incorporates non-linear modeling to address the limitations of linear combinations in traditional MF. By leveraging embedding techniques to represent latent factors of users and items, NeuMF demonstrated superior recommendation performance compared to MF. Despite these advancements, NeuMF also suffers from limitations, as it relies exclusively on rating data and struggles to capture specific user preferences. Moreover, due to the inherent sparsity in user–item matrices since most users only rate a few items, data sparsity remains a significant challenge for such approaches.

To address these limitations, recent studies increasingly utilized review texts as auxiliary information. Reviews, typically written by users after purchasing an item or receiving a service, capture users' personal opinions and satisfaction levels, making them highly effective for personalized recommendation services [16]. For instance, Zheng Noroozi [7] proposed a model that analyzes separate sets of reviews for users and items to predict ratings for new items. By independently analyzing these review sets, the model effectively captured their unique characteristics, achieving high performance. Similarly, Ghasemi and Momtazi [17] introduced a recommendation model that incorporates the similarity between user reviews as additional information to enhance traditional collaborative filtering (CF) models, which typically rely solely on rating data. By leveraging review texts as auxiliary information, their approach demonstrated superior performance compared to models that use only rating data. Meanwhile, Hong [18] aimed to enrich the high-dimensional user–item interactions by incorporating the review texts, leveraging a CNN to extract meaningful patterns from review texts and incorporating them into the two-dimensional user–item interaction map for greater precision. Additionally, Wang [19] applied ALBERT with personalized attention to enhance recommendation performance by aligning review texts with user preferences. Specifically, they introduced word- and review-level attention mechanisms to highlight salient content conditioned on user interest. Wang Yao [20] also introduced a multi-scale textual modeling approach to enhance review-based recommendation by leveraging hierarchical semantic representations. These studies highlight the growing trend of using review texts and other auxiliary information to alleviate the data sparsity issues for traditional rating-based methods [21].

2.2. Local and Global Features of Review Texts

In analyzing review texts, there exist local features that capture users' detailed preferences by considering individual words, word combinations, and phrases, and global features that capture users' intended meanings by considering the overall context of the review.

To effectively extract the local features embedded in reviews, a CNN model can be utilized. Although CNNs were initially proposed for image classification tasks, they demonstrated strong performance in natural language processing (NLP), making them widely used in review-based recommender systems for text analysis. CNNs operate using the n-gram approach, where words are grouped and slid incrementally during training [22]. Several studies focused on leveraging these local features of CNN models. For instance, Zhu Jiang [23] proposed a method that inputs text for sentiment analysis, capturing long-term dependencies and extracting detailed local features through two CNN layers. Moreover, Hu [24] indicated that the extraction of contextual information of short texts is limited by the Transformer-based model, and proposed a model that utilizes a CNN model that is effective in extracting local features to overcome this issue. Despite the effectiveness of CNNs in extracting local features, they face a limitation: when only the local aspects of reviews are considered, the broader contextual meaning may be overlooked, leading to ambiguity in understanding users' true intent. If only specific portions of the review are captured, users' actual intent within the overall context may be missed.

On the other hand, global feature extraction primarily employs pre-trained language models based on the Transformer architecture, such as bidirectional encoder representations from Transformers (BERT). BERT-based models effectively extract global features by deeply learning the contextual relationships between words and phrases in both directions. Various studies explored the advantages of global feature extraction. Feng and Zeng [25] utilized BERT to capture comprehensive contextual and word frequency information within review texts, allowing for holistic analysis. Their model went beyond simple word frequency analysis by capturing the overarching context within reviews to extract embedded global features. Additionally, Ge Zheng [26] proposed a model leveraging RoBERTa, which enhanced a version of BERT for global feature extraction. This model assigns weights to critical sections of the output vectors, effectively analyzing the sentiment trends within review texts. By weighting global features, it achieves more effective feature extraction. Zhang [27] proposed a dual-channel model combining BERT and TextCNN to extract sentiment and topic features from review texts, which are then fused and integrated into an NCF framework for personalized rating prediction. However, despite the advantages of bidirectional models in global feature extraction, focusing solely on global aspects may neglect the granular meanings embedded within the reviews. Although general context is considered, users' detailed preferences may be expressed only in specific segments, limiting the comprehensiveness of the extraction.

In summary, although approaches that rely solely on local features are effective in capturing fine-grained details in reviews, they cannot model broader contextual semantics. In contrast, approaches based only on global features can capture the overall contextual meaning, while they may overlook user-specific intents embedded in the texts. Therefore, leveraging the complementary strengths of both can more comprehensively capture user preferences. Jointly modeling local and global textual representations enables the system to capture both detailed expressions and overarching semantics to enhance personalized recommendation performance [28].

3. LGFR Framework

3.1. Problem Definition

The overall framework of the proposed LGFR model is illustrated in Figure 1. It consists of the user–item interaction module, feature extraction module, and preference prediction module. Specifically, the user–item interaction module utilizes user and item IDs as the input to learn the interactions between users and items. The feature extraction module utilizes both CNN and RoBERTa models to extract local and global features embedded in the review texts. The preference prediction module combines interaction and textual information to predict user ratings for specific items. Since this study formulates recommendations as a rating prediction task, given a tuple $T = (u, i, d)$, where u is the user, i is the item, and d is the associated review text, the LGFR model predicts the corresponding rating r .

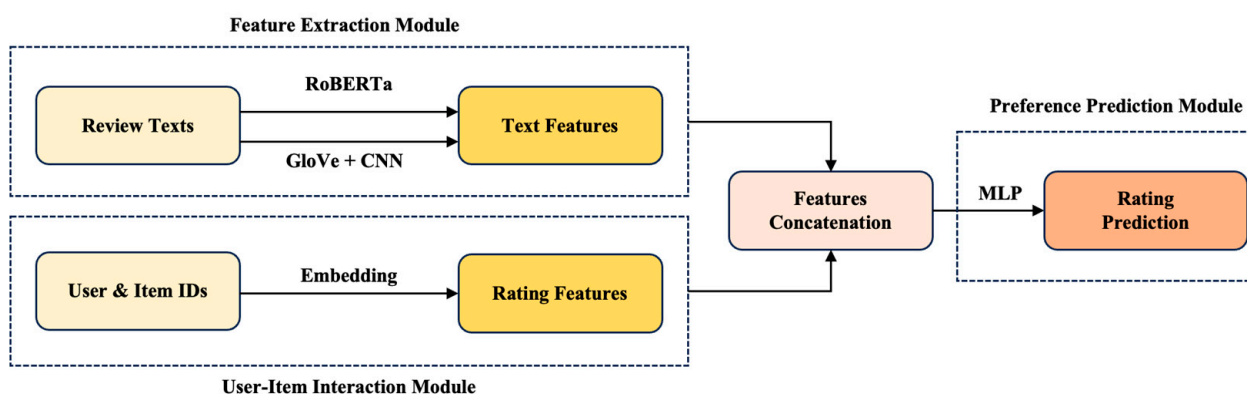


Figure 1. Framework of the LGFR model.

3.2. LGFR Architecture

This study proposes the LGFR model, which effectively analyzes review text information by combining CNN and RoBERTa models to introduce the local and global textual features in the recommendations. The detailed architecture is illustrated in Figure 2. The user–item interaction module follows the general recommendation strategy, which embeds user and item IDs into the high-dimensional vectors, concatenates them, and then employs the MLP to learn the complex user–item interactions. The feature extraction module processes from both local and global perspectives. For local features, review texts are first converted into word embeddings using global vectors for word representation (GloVe), then passed through a CNN with multiple kernel sizes to capture fine-grained n-gram patterns. For global features, the pre-trained RoBERTa model is applied as an encoder to generate contextualized representations, enabling a deeper semantic understanding. Finally, the preference prediction module concatenates the user–item vector with the extracted textual features and feeds the combined representation into the MLP to predict the rating for the given item. Further implementation details for each module are presented in the following sections.

3.3. User–Item Interaction Module

The user–item interaction module aims to learn the latent factors that reflect interactions between users and items. Specifically, each user ID and item ID is initially represented as a one-hot encoded vector as $v_u \in \mathbb{R}^n$ and $v_i \in \mathbb{R}^m$, where n and m represent the total number of users and items, respectively. These high-dimensional sparse vectors are then projected into a shared latent space of dimension k using trainable embedding matrices

$P \in \mathbb{R}^{n \times k}$ and $Q \in \mathbb{R}^{m \times k}$. This study determines the dimension k to be 32 based on fine-tuning for the experiments. This process is represented as follows:

$$p_u = P^T v_u, \quad (1)$$

$$q_i = Q^T v_i, \quad (2)$$

where $p_u \in \mathbb{R}^k$ and $q_i \in \mathbb{R}^k$ denote the latent factors of the user u and item i , respectively. Next, the user and latent factors are concatenated into a joint interaction vector and fed into the MLP to capture complex and non-linear interactions between the user and item. This process is described as follows:

$$\begin{aligned} I_0 &= p_u \oplus q_i, \\ I_1 &= \text{ReLU}(W_1^T I_0 + b_1), \\ &\vdots \\ O^{\text{Interaction}} &= I_l = \text{ReLU}(W_l^T I_{l-1} + b_l), \end{aligned} \quad (3)$$

where I_l denotes the output of the l -th layer, representing the learned non-linear interaction between the user and item. W_l^T and b_l refer to the weight matrix and bias of the l -th layer, respectively. The ReLU function is used as a non-linear activation function. The operator $\oplus$ indicates the concatenation of the two output feature vectors. The final interaction representation $O^{\text{Interaction}}$ is a 64-dimensional vector used for downstream prediction.

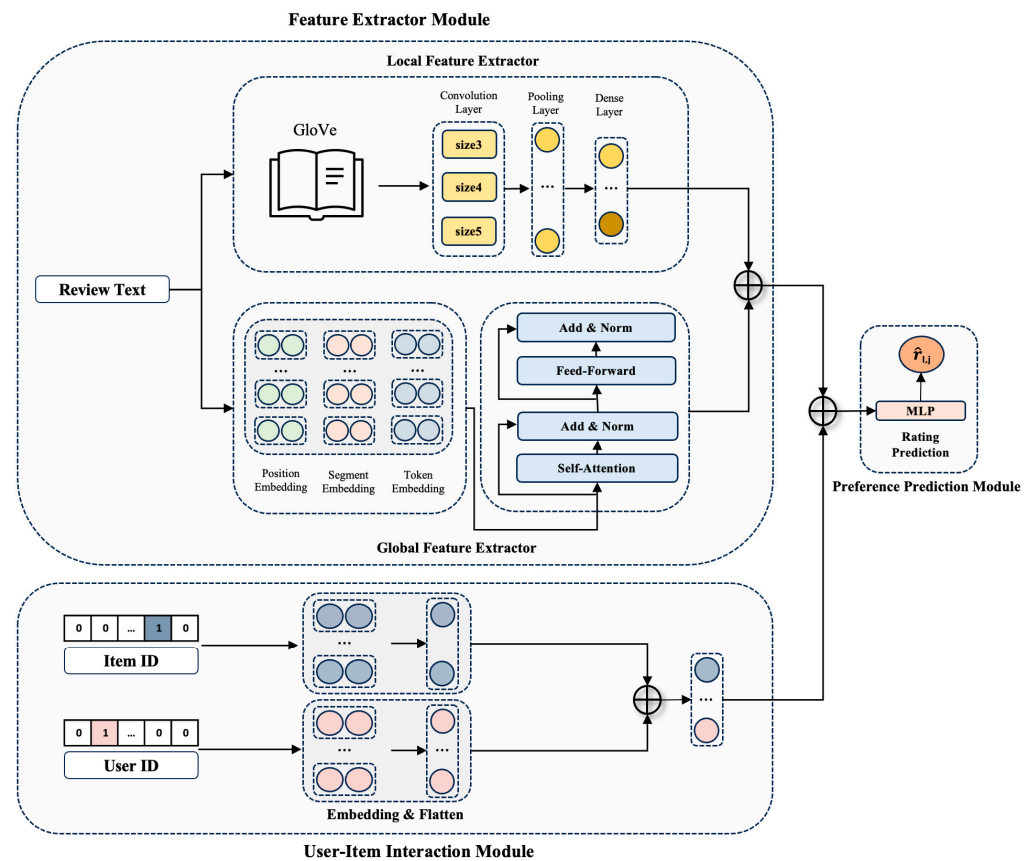


Figure 2. Architecture of the LGFR framework.

3.4. Feature Extraction Module

The feature extraction module aims to extract local and global features from the review text $D_{i,j}$ to effectively represent the text. To achieve this, it utilizes both the CNN and RoBERTa models. Previous recommendation studies primarily focused on capturing the

local features of review texts using CNN models. However, as previously mentioned, it is equally important to consider the global features present in the review text. Therefore, this study incorporates RoBERTa, which is effective for capturing global features, to better capture user preferences embedded in the reviews.

3.4.1. Local Feature Extractor

The local feature extractor focuses on capturing local features within the review using CNN, which is well-suited for n-gram-based learning of word groups [29]. This procedure facilitates the extraction of fine-grained, user-specific preference signals latent in the review texts. The review text is represented as $D_{i,j} = \{w_1, w_2, \dots, w_n\}$, where w_k represents the k -th word, and n denotes the length of the review text [30]. This study uses GloVe to extract semantic information from review texts, converting each word into a 300-dimensional vector $D_e \in \mathbb{R}^{n \times 300}$. The text embedding D_e extracted through this process is then fed into the convolution operation as shown below:

$$c_j = \text{ReLU}(D_e * K_j + b_j), \quad (4)$$

where $*$ denotes the convolution operation, and b_j and K_j denote the bias and the p -dimensional filter kernel matrix of the j -th filter, respectively. The ReLU function is used as the activation function. The output c_j represents the feature map generated after applying the convolutional filter. To extract the overall semantic meaning of the texts, average pooling is performed as shown in Equation (5).

$$O^{CNN} = \text{average}([c_1, c_2, \dots, c_{n-t+1}]), \quad (5)$$

where the feature vector O^{CNN} encapsulates the local features present in the review texts. This study applies three parallel convolutional layers with kernel sizes of 3, 4, and 5, each using 100 filters. After average pooling, the output feature vectors are concatenated to form the local representation $O^{CNN} \in \mathbb{R}^{300}$.

3.4.2. Global Feature Extractor

The global feature extractor focuses on capturing the global features within the review, using RoBERTa, which excels in capturing the contextual characteristics of text. RoBERTa is a Transformer-based model proposed to enhance BERT's performance through hyperparameter optimization, additional datasets, and various other techniques, enabling more effective learning of the contextual meanings of text. Notably, it simplifies the training process by eliminating inefficient tasks, such as next-sentence prediction. Additionally, RoBERTa incorporates dynamic masking and additional text data to understand the meaning of text within a broader context.

To input review text into RoBERTa, the text is first tokenized using WordPiece. The tokenized text is then fed into RoBERTa, producing 512 token embeddings with dimensions of 768, including special tokens such as [CLS] and [SEP]. In this study, the [CLS] token, which encapsulates the overall semantic characteristics of the text, is extracted as its embedding vector, a common approach in various NLP tasks [31]. Therefore, the tokenized $D_{i,j}$ is processed through RoBERTa, converting it into a 768-dimensional [CLS] embedding vector that captures the overall meaning of the text. This process is expressed as follows:

$$O^{RoBERTa} = \text{RoBERTa}(D_{i,j}), \quad (6)$$

where $O^{RoBERTa}$ denotes the global feature vector of the review output by RoBERTa. Finally, the feature extraction module combines the locally represented O^{CNN} vector obtained from

the CNN and the globally represented $O^{RoBERTa}$ vector through concatenation, as shown in the following equation:

$$O^{Text} = O^{CNN} \oplus O^{RoBERTa}, \quad (7)$$

where O^{Text} denotes the final review representation vector output by the feature extraction module, encompassing both the local and global features of the text. Since $O^{CNN} \in \mathbb{R}^{300}$ and $O^{RoBERTa} \in \mathbb{R}^{768}$, this study projects $O^{RoBERTa}$ to a 300-dimensional vector before concatenation to balance the contribution of local and global features and avoid overemphasis on any single representation.

3.5. Preference Prediction Module

In the final stage, the preference prediction module predicts the user's rating for an item based on the interaction vector between the user and the item, as well as the textual representation vector derived from the previous modules. The extracted vectors are concatenated as shown in Equation (8):

$$R_0 = O^{Interaction} \oplus O^{Text}, \quad (8)$$

where R_0 denotes the concatenated vector that combines the user-item feature vector extracted from the user-item interaction module and the review text feature vector extracted from the feature extraction module.

The vector derived from Equation (8) is input into a neural network that outputs a linear result to predict the rating as shown in the following equation:

$$\hat{r}_{i,j} = W_o R_0 + b_o, \quad (9)$$

in Equation (9), W_o and b_o denote the weight and bias of the final linear output layer, respectively. The predicted rating $\hat{r}_{i,j}$ represents the user's expected rating for the item as computed by this layer.

During training, gradient descent and backpropagation are used to minimize the difference between the predicted rating $\hat{r}_{i,j}$ and the actual rating $r_{i,j}$. The parameter optimization process employs the adaptive moment estimation (Adam) optimizer. The mean squared error (MSE) is used as the loss function to measure the difference between the predicted and actual ratings:

$$MSE = \frac{1}{n} \sum_{n=1}^n (r_{u,i} - \hat{r}_{u,i})^2 \quad (10)$$

where n denotes the number of data points, and $r_{u,i}$ and $\hat{r}_{u,i}$ represent the actual and predicted ratings, respectively.

4. Experiments

4.1. Datasets

To validate the performance of the proposed recommendation model, we used the widely adopted Amazon.com dataset in recommendation system research. The Amazon.com dataset (<https://amazon-reviews-2023.github.io/> (accessed on 3 December 2024)) contains various information, such as user-generated online reviews, item descriptions, and ratings [32]. For the experiments in this study, we utilized the Cell Phones and Accessories, Industrial and Scientific, and Video Games datasets from Amazon.com. From these datasets, we extracted user IDs, item IDs, ratings, and review texts. A five-core filtering technique was applied, including only users who had written at least five reviews in the datasets. Additionally, common text preprocessing steps were performed, such as stopword

removal and the handling of whitespace and characters. HTML tags, special characters, and other noise were removed during the data preprocessing step. For RoBERTa input, reviews shorter than 5 tokens were removed, and those exceeding 512 tokens were truncated to fit within the model's maximum input length. Among the filtered data, 70% were used as training data, 10% as validation data, and 20% as test data [33]. To ensure consistency across splits, the data were split based on user level, with each user's interactions allocated to training, validation, and test sets. This strategy ensures that all users are present in each subset, while preventing information leakage across user-specific behaviors. Table 1 shows the basic statistics of the preprocessed data.

Table 1. Basic statistics of datasets by category.

Dataset	Number of Reviews	Number of Users	Number of Items	Sparsity
Cell Phones and Accessories	2,112,651	344,454	119,144	99.994%
Industrial and Scientific	291,958	42,331	25,390	99.972%
Video Games	684,949	94,080	31,347	99.997%

4.2. Evaluation Metrics

In this study, mean absolute error (MAE) and root mean squared error (RMSE) were used to evaluate the performance of the recommendation model. These two metrics quantitatively measure the difference between predicted and actual ratings, effectively assessing the prediction accuracy of the recommendation system. MAE calculates the average of the absolute differences between predicted and actual values, while RMSE calculates the square root of the average squared differences. A lower value for both metrics indicates that the model's predictions are closer to the actual ratings [34]. By utilizing both metrics, this study aims to comprehensively evaluate the performance of the model from multiple perspectives.

$$MAE = \frac{1}{N} \sum_{i=1}^N |r_i - \hat{r}_i| \quad (11)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - \hat{r}_i)^2} \quad (12)$$

N denotes the number of predicted ratings, and r_i and $\hat{r}_i$ denote the actual and predicted ratings, respectively.

4.3. Baseline Models

In this study, the performance of the proposed LGFR model was compared with several baseline models. The baseline models include PMF and NeuMF, which are based on rating data, HFT, which incorporates review texts and latent Dirichlet allocation (LDA), and models that utilize both review texts and deep learning, including DeepCoNN, AENAR, UCAM, and RARV2. The selected baseline models allow us to evaluate (1) the benefit of incorporating review texts compared to using only rating data; (2) the effectiveness of deep learning methods over traditional statistical models; and (3) the impact of capturing both local and global textual features in comparison to other review-based models.

- PMF [35]: Probabilistic matrix factorization is a variant of matrix factorization that is effective for sparse and imbalanced rating data. This model decomposes the rating matrix into latent factors based on a Gaussian prior distribution.
- NeuMF [15]: Neural collaborative filtering is a deep learning-based model designed to overcome the limitations of MF, which only considers linear interactions. NeuMF introduces non-linear learning to capture complex interactions between users and items.

- HFT [36]: The hidden factors and hidden topics model employs LDA to extract hidden topics from aggregated reviews of users and items. It combines hidden factors obtained from matrix factorization with hidden topics.
- DeepCoNN [7]: the deep cooperative neural network is a deep learning-based recommendation system that processes the review texts of both users and items through two parallel CNNs to learn meaningful latent representations, which are then combined to predict ratings.
- UCAM [37]: The unstructured context-aware model is a deep learning model that utilizes unstructured text information extracted from reviews. It combines user–item interactions with the review information to make predictions.
- AENAR [38]: The aspect-aware explainable neural attentional recommender model extracts representation vectors from the review information of both users and items using CNNs. It then combines these vectors with an attention network to emphasize important parts of the review, predicting the rating the user will assign.
- RARV2 [39]: this model extracts text representations from review texts using RoBERTa and BERT, combining these with rating data to predict the user’s rating for the item.

4.4. Implementation Details

All hyperparameters of the proposed LGFR model were tuned based on performance on the validation set. The learning rate was selected from [0.001, 0.0001, and 0.00001], batch size from [64, 128, 256, 512, and 1024], embedding size from [8, 16, 32, and 64], and number of hidden layers from [1, 2, 3, 4, and 5]. A grid search strategy was used, varying one parameter at a time while keeping others fixed at baseline values. As a result of the optimization, the learning rate was set to 0.001, the batch size to 128, the embedding size to 32, and the number of hidden layers to 5. To prevent overfitting, early stopping was applied when validation loss did not improve for 10 consecutive epochs. To minimize experimental errors, the experiments were repeated five times, and the performance of each model was evaluated based on the average of the results. We used the RoBERTa-based model [40] as a frozen feature extractor, where all parameters were fixed, and no additional fine-tuning was performed during training. The model contains approximately 125 million parameters and supports a maximum sequence length of 512 tokens. Input texts longer than 512 tokens were truncated, and those shorter than 5 tokens were removed during preprocessing. The experiments were implemented using TensorFlow 2.18 and Transformers 4.52 in the Google Collaboratory Pro environment, using an NVIDIA A100 GPU.

For our proposed model, the embedding extraction phase took approximately 1 to 2.5 h depending on the dataset size. The training time per epoch ranged from 10 s to 1 min. These results demonstrate that LGFR maintains a reasonable computational cost and is scalable to real-world recommendation scenarios.

5. Results and Discussion

5.1. Comparison with Baseline Models

To validate the performance of the proposed LGFR model, three different datasets from Amazon.com were used, and the recommendation performance was compared with various baseline models, as shown in Table 2. The proposed LGFR model demonstrated superior performance compared to the baseline models in all experiments.

The conclusions derived from the results in Table 2 are as follows: First, the PMF model, which analyzes rating data through matrix factorization, exhibited lower performance compared to the NeuMF model. This suggests that learning non-linear interactions through the MLP is more effective in improving recommendation performance than representing user–item interactions through simple linear relationships.

Table 2. Performance comparison between the proposed model and baseline models.

Model	Cell Phones and Accessories		Industrial and Scientific		Video Games	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
PMF	1.690	2.026	2.239	2.591	1.198	1.517
NeuMF	1.025	1.387	0.818	1.284	0.946	1.195
HFT	0.926	1.324	0.697	1.177	0.725	1.056
DeepCoNN	0.791	0.977	0.673	0.928	0.677	0.946
AENAR	0.762	0.967	0.665	0.914	0.647	0.923
UCAM	0.524	0.745	0.450	0.717	0.558	0.790
RARV2	0.455	0.711	0.415	0.694	0.448	0.691
LGFR	0.429	0.692	0.402	0.662	0.421	0.673

Second, the NeuMF model, which uses only rating data, showed lower performance compared to the HFT model, which utilizes review text as auxiliary information. This indicates that addressing the data sparsity issue, a limitation of the NeuMF model, by incorporating review text can enhance the model's performance.

Third, models that apply deep learning to review texts outperformed the HFT model, which applies the statistical LDA technique. This implies that traditional statistical methods have limitations in fully capturing the expressions embedded in review texts. By leveraging deep learning models, the user's review intent can be captured more accurately.

Fourth, among the four models that use review texts and deep learning, models employing a single-review approach, such as UCAM and RARV2, outperformed models using a review aggregation approach, such as DeepCoNN and AENAR. The review aggregation models primarily reflect general trends by combining multiple reviews, which can lead to the dilution or loss of fine-grained information. On the other hand, single-review models fully utilize the detailed content and contextual characteristics of each review, allowing for the extraction of more specific and precise user or item preferences. Since review texts often vary in expression even for the same user, the single-review approach can reflect these differences and enable more personalized recommendations.

Lastly, the proposed model, which analyzes review texts through a single-review approach and applies deep learning, demonstrated the highest performance. The LGFR model alleviates the data sparsity issue by utilizing both rating data and review texts as auxiliary information. Additionally, by employing deep learning techniques instead of statistical methods, it can capture non-linear preferences and extract more detailed preferences and contextual information compared to review aggregation approaches. The use of CNN and RoBERTa, which are effective for extracting local and global features of review texts, further amplifies the strengths of the single-review approach by capturing users' latent review-writing intents.

5.2. Efficiency Analysis of Using Fused Features

The proposed LGFR model combines the CNN and RoBERTa models to capture both the local and global features of review texts. Previous studies primarily focused on capturing local features using CNN models. However, it is equally important to capture the global features within reviews, and this led to the development of the proposed model that incorporates both aspects. To verify whether the combined use of these two models is beneficial for improving recommendation performance, experiments were conducted comparing the performance of using only CNN for local feature extraction and only RoBERTa for global feature extraction. The experimental results are presented in Table 3.

Table 3. Performance comparison based on text features.

Model	Cell Phones and Accessories		Industrial and Scientific		Video Games	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
w/o local feature	0.624	0.962	0.558	0.933	0.604	0.957
w/o global feature	0.454	0.709	0.412	0.736	0.441	0.736
Local and global features (LGFR)	0.429	0.692	0.402	0.682	0.421	0.683

When only CNN was used to extract local features, it showed the lowest performance across all datasets. Additionally, when only RoBERTa was used to extract global features, its performance was better than that of the CNN-only model, but still lower compared to the proposed combined CNN and RoBERTa model. This indicates that the traditional single-model approach, where CNN was primarily used in recommendation systems to handle review texts, is limited in effectively capturing latent features due to its focus solely on local features. Furthermore, while the RoBERTa-only model outperformed the CNN-only model, it demonstrated lower performance compared to the proposed LGFR model, as its focus on extracting only global features made it less effective in capturing comprehensive textual information. In conclusion, combining the two models to consider both local and global features enhances recommendation performance, demonstrating the importance of this comprehensive approach for effectively analyzing review texts.

5.3. Efficiency Analysis of Feature Fusion Method

In this section, an experiment was conducted to determine whether the method of fusing feature vectors generated from each module affects performance. In this experiment, we tested different fusion methods commonly used for vector fusion, including Element-wise Sum, Element-wise Product, Element-wise Average, Attention Mechanism, gated multimodal unit (GMU), and Concatenation, to measure their impact on performance. The settings, except for the combination method, were kept the same as those of the proposed model. The experimental results are presented in Table 4.

Table 4. Performance comparison based on feature fusion methods.

Method	Cell Phones and Accessories		Industrial and Scientific		Video Games	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
LGFR (Sum)	0.424	0.700	0.483	0.706	0.434	0.688
LGFR (Product)	0.467	0.722	0.450	0.743	0.468	0.740
LGFR (Average)	0.433	0.705	0.424	0.688	0.429	0.711
LGFR (GMU)	0.428	0.703	0.407	0.688	0.469	0.699
LGFR (Attention)	0.438	0.718	0.404	0.743	0.449	0.704
LGFR (Concatenation)	0.429	0.692	0.402	0.682	0.421	0.683

The results from Table 4 demonstrate that the method used to fuse textual features influences recommendation performance, with the concatenation method used in the proposed model consistently showing the highest performance across all datasets. The other methods, which reduce the dimensional size of the fused vector through element-wise operations, showed lower performance. This suggests that reducing the dimension of the vector can lead to the loss of critical information regarding the features. Notably, the simple

concatenation of vectors from each extractor outperformed more complex fusion strategies involving additional weighting or transformation. This result indicates that the respective outputs of CNN and RoBERTa are already sufficiently expressive for capturing local and global contextual features within the proposed model. Therefore, concatenating the local and global features preserves the information from both perspectives, making it the most effective approach for extracting user preferences from review texts.

5.4. Analysis of Model Effectiveness by Review Length

The proposed model was designed to effectively capture both the local and global perspectives of review texts. To evaluate whether this design is genuinely effective, we conducted the following experiment. First, the review texts were categorized based on tokenized length: the bottom 25% were classified as short reviews, and the top 25% were classified as long reviews. We then compared the performance of the proposed model against the high-performing baseline models from experiment 5.1. The hypothesis was that shorter reviews might not exhibit a clear distinction between local and global features, whereas longer reviews would showcase these features more prominently, thereby allowing the proposed model to function more effectively. The experimental results validate this hypothesis, as shown in Table 5.

Table 5. Performance comparison based on review length.

Dataset	Model	Short Reviews		Long Reviews		Total Reviews	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Cell Phones and Accessories	AENAR	0.705	1.035	1.034	1.275	0.762	0.967
	UCAM	0.674	0.816	0.841	1.039	0.524	0.745
	RARV2	0.334	0.603	0.556	0.808	0.455	0.711
	LGFR	0.361	0.656	0.519	0.779	0.429	0.692
Video Games	AENAR	0.566	0.913	1	1.242	0.647	0.923
	UCAM	0.332	0.677	0.682	0.899	0.558	0.79
	RARV2	0.342	0.554	0.669	0.721	0.448	0.691
	LGFR	0.319	0.604	0.567	0.785	0.421	0.673
Industrial and Scientific	AENAR	0.699	1.09	1.054	1.367	0.665	0.914
	UCAM	0.372	0.763	0.639	0.864	0.45	0.717
	RARV2	0.339	0.548	0.539	0.738	0.415	0.694
	LGFR	0.355	0.621	0.522	0.775	0.402	0.662

For short reviews, the performance difference between the proposed model and the existing models is relatively small. The reason for this is that short reviews contain fewer words, making it difficult to clearly distinguish between local and global features within the text, thereby limiting the full realization of the model's advantage of considering both features. However, for long reviews, the proposed model demonstrates relatively higher performance. Long reviews include more diverse and detailed word groups and require greater contextual interpretation, allowing for a more effective analysis of both the local and global features of the text. As a result, the experimental findings indicate that the proposed model outperforms the existing models more significantly in long reviews. In other words, while the advantage of the proposed model is less prominent for short reviews where distinguishing local and global perspectives is challenging, it shows superior performance for long reviews, where both perspectives can be effectively analyzed. This highlights that the core design of the proposed model, which focuses on capturing both local and global features, is functioning effectively as intended.

6. Conclusions

With the development of the online e-commerce market, the number of items users need to manage also increased. To address this, research on recommendation systems that utilize user preferences and various pieces of information from online reviews has been actively conducted. There are two perspectives for viewing reviews: the local feature perspective and the global feature perspective. Local features focus on capturing the user's detailed intent and preferences within the review. On the other hand, global features allow for understanding the overall context of the review and capturing what the user generally intends to convey. To effectively analyze review texts, it is essential to consider both local and global features together. This study proposes the LGFR model to improve recommendation performance by integrating local and global features in review-based recommendation systems. To capture the local features in review texts, CNN, which learns sliding n-gram units, was used. To capture the global features, RoBERTa, which is useful for understanding overall context through bidirectional learning, was employed to effectively capture users' latent preferences within reviews. The proposed model was evaluated using data provided by the e-commerce platform Amazon, and it demonstrated superior performance compared to existing recommendation systems. Additionally, the results show that recommendation performance improves more when both local and global features are considered simultaneously, compared to when only local or global features are used independently. Further experiments were conducted to verify whether the design for extracting local and global features was effectively functioning. The results indicate that the proposed model performed relatively better for long reviews than short reviews. This is because long reviews contain more detailed word groups and require greater contextual interpretation, making it easier to differentiate between local and global features.

This study proposes a new approach to personalized recommendation systems in e-commerce, but acknowledges the following limitations: First, to mitigate the data sparsity issue, this study utilized review texts as auxiliary information. However, recent multi-modal studies explored the use of various forms of data, including images, videos, and audio, as auxiliary information. Therefore, future research should explore whether the LGFR model can achieve superior performance in multimodal recommendation systems. Second, to extend the scope of the research, it is necessary to validate the proposed model using datasets from different domains. Although this study used the Amazon dataset from the e-commerce field, various other domains also handle review texts. It would be beneficial to test the model on datasets from multiple domains to evaluate its generalizability. Particularly, comparing the performance of the LGFR model in domains with short, simple reviews versus those with longer, more complex reviews could be a promising area for future research. Third, there is a challenge in verifying whether the proposed model effectively reflects the local and global perspectives of reviews. Techniques such as explainable artificial intelligence (XAI) [41], including local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP), can be used to visualize which words or phrases the model considers important from either the local or global perspective within the review text. This visualization could enhance the practical utility of the model by offering more interpretable results. Finally, although the proposed LGFR model demonstrates strong empirical performance by combining local and global textual features, this study does not provide a theoretical or representation-level analysis (e.g., mutual information estimation or feature complementarity visualization) to explain the performance gain. Therefore, future work should further explore such analyses to understand the interaction between local and global representations.

Author Contributions: Conceptualization, N.K., H.L. and Q.L.; Methodology, N.K., H.L. and X.L.; Software, H.L. and Q.L.; Data Curation, N.K., H.L. and J.K.; Writing—Original Draft Preparation, N.K., H.L. and Q.L.; Writing—Review and Editing, X.L., S.K. and J.K.; Supervision, S.K. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available on <https://amazon-reviews-2023.github.io/> (accessed on 3 December 2024).

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

References

- Iwanaga, J.; Nishimura, N.; Sukegawa, N.; Takano, Y. Improving collaborative filtering recommendations by estimating user preferences from clickstream data. *Electron. Commer. Res. Appl.* **2019**, *37*, 100877. [CrossRef]
- 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]
- Duan, R.; Jiang, C.; Jain, H.K. Combining review-based collaborative filtering and matrix factorization: A solution to rating's sparsity problem. *Decis. Support Syst.* **2022**, *156*, 113748. [CrossRef]
- Kim, D.; Li, Q.; Jang, D.; Kim, J. AXCF: Aspect-based collaborative filtering for explainable recommendations. *Expert Syst.* **2024**, *41*, e13594. [CrossRef]
- Chen, C.; Zhang, M.; Liu, Y.; Ma, S. Neural attentional rating regression with review-level explanations. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1583–1592.
- Liu, H.; Wang, Y.; Peng, Q.; Wu, F.; Gan, L.; Pan, L.; Jiao, P. Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing* **2020**, *374*, 77–85. [CrossRef]
- Zheng, L.; Noroozi, V.; Yu, P.S. Joint deep modeling of users and items using reviews for recommendation. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 425–434.
- Liu, D.; Li, J.; Du, B.; Chang, J.; Gao, R. Daml: Dual attention mutual learning between ratings and reviews for item recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 344–352.
- Jang, D.; Li, Q.; Lee, C.; Kim, J. Attention-based multi attribute matrix factorization for enhanced recommendation performance. *Inf. Syst.* **2024**, *121*, 102334. [CrossRef]
- Niu, G.; Xu, H.; He, B.; Xiao, X.; Wu, H.; Gao, S. Enhancing local feature extraction with global representation for neural text classification. 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, 3–7 November 2019; pp. 496–506.
- Johnson, R.; Zhang, T. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 562–570.
- Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; Ruan, B. LGMRec: Local and global graph learning for multimodal recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence Vancouver, BC, Canada, 20–27 February 2024; pp. 8454–8462.
- Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
- Yang, S.; Li, Q.; Jang, D.; Kim, J. Deep learning mechanism and big data in hospitality and tourism: Developing personalized restaurant recommendation model to customer decision-making. *Int. J. Hosp. Manag.* **2024**, *121*, 103803. [CrossRef]
- He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.-S. Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 173–182.
- 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]
- Ghasemi, N.; Momtazi, S. Neural text similarity of user reviews for improving collaborative filtering recommender systems. *Electron. Commer. Res. Appl.* **2021**, *45*, 101019. [CrossRef]
- Hong, S.; Li, X.; Yang, S.; Kim, J. Based Recommender System Using Outer Product on CNN. *IEEE Access* **2024**, *12*, 65650–65659. [CrossRef]
- Wang, S.; Du, W.; Bhuiyan, A.; Chen, Z. Personalized Recommendation Method Based on Rating Matrix and Review Text. *Comput. Intell.* **2025**, *41*, e70024. [CrossRef]
- Wang, D.; Yao, H.; Yu, D.; Song, S.; Weng, H.; Xu, G.; Deng, S. Graph Intention Embedding Neural Network for tag-aware recommendation. *Neural Netw.* **2025**, *184*, 107062. [CrossRef]

21. Park, J.; Li, X.; Li, Q.; Kim, J. Impact on recommendation performance of online review helpfulness and consistency. *Data Technol. Appl.* **2023**, *57*, 199–221. [\[CrossRef\]](#)
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [\[CrossRef\]](#)
23. Zhu, Q.; Jiang, X.; Ye, R. Sentiment analysis of review text based on BiGRU-attention and hybrid CNN. *IEEE Access* **2021**, *9*, 149077–149088. [\[CrossRef\]](#)
24. Hu, H. A CNN-Transformer model for short Chinese texts sentiment analysis. In Proceedings of the 2024 5th International Conference on Computer Engineering and Application (ICCEA), Hangzhou, China, 12–14 April 2024; pp. 816–821.
25. Feng, X.; Zeng, Y. Neural collaborative embedding from reviews for recommendation. *IEEE Access* **2019**, *7*, 103263–103274. [\[CrossRef\]](#)
26. Ge, H.; Zheng, S.; Wang, Q. Based BERT-BiLSTM-ATT model of commodity commentary on the emotional tendency analysis. In Proceedings of the 2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI), Qingdao, China, 2–4 July 2021; pp. 130–133.
27. Zhang, L.; Xia, P.; Ma, X.; Yang, C.; Ding, X. Enhanced Chinese named entity recognition with multi-granularity BERT adapter and efficient global pointer. *Complex Intell. Syst.* **2024**, *10*, 4473–4491. [\[CrossRef\]](#)
28. Zhang, F.; Liang, T.; Wu, Z.; Yin, Y. PILL: Plug Into LLM with Adapter Expert and Attention Gate. *arXiv* **2023**, arXiv:2311.02126.
29. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)
30. Kim, G.; Choi, I.; Li, Q.; Kim, J. A CNN-based advertisement recommendation through real-time user face recognition. *Appl. Sci.* **2021**, *11*, 9705. [\[CrossRef\]](#)
31. Kenton, J.D.M.-W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; p. 2.
32. Jeong, E.; Li, X.; Kwon, A.E.; Park, S.; Li, Q.; Kim, J. A Multimodal Recommender System Using Deep Learning Techniques Combining Review Texts and Images. *Appl. Sci.* **2024**, *14*, 9206. [\[CrossRef\]](#)
33. Cheng, W.; Shen, Y.; Huang, L.; Zhu, Y. Dual-embedding based deep latent factor models for recommendation. *ACM Trans. Knowl. Discov. Data (TKDD)* **2021**, *15*, 1–24. [\[CrossRef\]](#)
34. Silveira, T.; Zhang, M.; Lin, X.; Liu, Y.; Ma, S. How good your recommender system is? A survey on evaluations in recommendation. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 813–831. [\[CrossRef\]](#)
35. Mnih, A.; Salakhutdinov, R.R. Probabilistic matrix factorization. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 1257–1264.
36. McAuley, J.; Leskovec, J. Hidden factors and hidden topics: Understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, China, 12–16 October 2013; pp. 165–172.
37. Unger, M.; Tuzhilin, A.; Livne, A. Context-aware recommendations based on deep learning frameworks. *ACM Trans. Manag. Inf. Syst. (TMIS)* **2020**, *11*, 1–15. [\[CrossRef\]](#)
38. Zhang, T.; Sun, C.; Cheng, Z.; Dong, X. AENAR: An aspect-aware explainable neural attentional recommender model for rating predication. *Expert Syst. Appl.* **2022**, *198*, 116717. [\[CrossRef\]](#)
39. Liu, Y.-H.; Chen, Y.-L.; Chang, P.-Y. A deep multi-embedding model for mobile application recommendation. *Decis. Support Syst.* **2023**, *173*, 114011. [\[CrossRef\]](#)
40. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning--based text classification: A comprehensive review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40. [\[CrossRef\]](#)
41. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.