Expert Systems With Applications 198 (2022) 116717

Contents lists available at ScienceDirect

Expert Systems With Applications

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

AENAR: An aspect-aware explainable neural attentional recommender model for rating predication

Tianwei Zhang [a,1], Chuanhou Sun [a,1], Zhiyong Cheng [b,*], Xiangjun Dong [a,*]

[a] Faculty of Computing, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[b] Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

ARTICLE INFO

Keywords:
Recommender systems
Neural networks
Attention mechanism
Deep learning

ABSTRACT

Explainable rating predication becomes challenging with the largely growing number of information and items. Of particular interest is to capture users' preferences for various items by using textual reviews to achieve accurate and interpretable recommendations. In this paper, we report an aspect-aware explainable neural attentional recommender model for rating predication (AENAR) and this model enables intelligent predication and recommendation by capturing the varying aspect attentions that users pay to different items. The experimental results based on six public datasets reveals that the designed model consistently outperforms five existing state-of-the-art alternatives. Furthermore, the designed attention network allows to highlight the context-aware information in textual reviews that unambiguously suggest users' aspect-level preference for their desired items, improving the interpretability of the rating prediction.

1. Introduction

1.1. Overview introduction

As the rapid development of e-commerce and the tremendous growth of digital information, it becomes a big challenge to search useful and desired information. Recently, recommendation system has attracted a great deal of attention due to its intelligent predication and recommending. Among various recommendation systems, Collaborative Filtering (CF) is undoubtedly considered as the most promising one by modeling users' preferences for items based on their interactions (such as ratings, clicks, etc.) in the past. As a classical algorithm of CF, Matrix Factorization (MF) (Bell & Koren, 2007; Koren, Bell, & Volinsky, 2009) has been demonstrated effective to transform both users and items into latent vectors in a common latent space for rating predication. However, this modeling will encounter the problem of cold starting and in particular the difficulty in presenting the reasons why the items are preferred (Cheng, Ding, Zhu and Mohan, 2018; McAuley & Leskovec, 2013; Wang, He, Feng, Nie, & Chua, 2018).

In order to address these issues, textual review, a general part in most e-commerce websites, has become a hot topic for explainable recommendation. In general, reviews include users' preference and items' characteristics such as appearance, quality and price, allowing for a preliminary reference for new and potential users. Therefore, a number of recommendation systems by utilizing textual reviews with aspect-based models have emerged (Chen, Zhang, Liu, & Ma, 2018; Cheng et al., 2018; Cheng, Ding, Zhu and Kankanhalli, 2018; Chin, Zhao, Joty, & Cong, 2018; Guan et al., 2019; Le & Lauw, 2020; Liu, Wang, Xu, Peng, & Jiao, 2020; Mukherjee et al., 2020; Peña et al., 2020; Wu et al., 2019). Usually, the aspect-based models extract the aspect information in the textual reviews to depict users' preference. For instance, the topic models such as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) have been utilized to automatically extract aspect information. Recent years, with the widespread application of deep learning technology, deep neural networks have been demonstrated to be powerful due to their learning capabilities to extract desired information(features) in textual reviews for effective and explainable rating predication.

Apparently, the aspects in the textual review are not equally important to reflect the interaction between users and items. For example, when watching a suspense movie, users will pay more attention to the "plot", while they tend to pay more attention to the "special effects" when watching a superhero movie (Liu et al., 2019). Additionally, the preference will vary according to users' application. A user may care more about the displaying performance when selecting a gaming phone, while the portability and battery life may become more important when purchasing a phone for daily use. Therefore, it is beneficial to utilize textual reviews to extract aspect-level features of users and items for guiding the learning of the implicit relevance between them. In

* Corresponding authors.
E-mail addresses: tw3369@163.com (T. Zhang), sch96100@163.com (C. Sun), jason.zy.cheng@gmail.com (Z. Cheng), d-xj@163.com (X. Dong).
[1] Tianwei Zhang and Chuanhou Sun contributed equally to this work.

https://doi.org/10.1016/j.eswa.2022.116717
Received 17 April 2021; Received in revised form 13 January 2022; Accepted 20 February 2022
Available online 6 March 2022

particular, the specific aspects may be of significant importance, even not independent of each other between users and items, but closely integrated with their interaction. Motivated by this observation, we propose an aspect-aware explainable neural attentional recommender model (AENAR) and it can adaptively capture users' varying attention to different aspects of items for a more explainable recommending.

In AENAR, the Convolution Neural Network (CNN) is used to extract features from the context window of the textual reviews (Kim, 2014), based on which the joint feature vectors can be constructed for users and items. In details, for each pair of user and item, while embedding their respective id into latent vectors, we will also utilize CNN to extract the feature vectors from their own review texts, so that we can (1) fusion the textual features with their id latent vectors to help learn the final joint latent factor of this pair of user and item and (2) by an attention network, capture the user's attention vector to this particular item. It should be noted that each node in the attention vector obtained in step (2) is regarded as an aspect, and then, by getting the location of the node with the largest value in the attention vector, we can capture which aspect this user is most concerned about when interacting with the current item. At the same time, based on the location of the largest node, we can also trace back to the original user's textual review and find the keywords that best reflect the user's concern in this aspect. After the steps above, finally, based on the attentive interaction of the user's and item's final latent factors, the rating that we predict can be got by a rating prediction multi-layer perceptron.

In order to verify the effectiveness of our model, we conducted comprehensive experiments on six Amazon product datasets to compare our model with several state-of-the-art methods. The results show that our model outperforms the competitors. In summary, the main contributions of our work are as follows:

- We proposed an aspect-aware explainable rating prediction recommender approach based on a novel designed adaptive attention network. In particular, we apply CNN to extract abstract features from the textual reviews to assist in the construction of the final latent vectors, which not only maps the latent vectors in the traditional collaborative filtering to a higher non-linear dimension, thus improving the recommendation effect, but also enables the subsequent attention network to trace back to the keywords in original textual reviews through the position of the maximum node.
- We designed an attention network to adaptively capture the specific attention vector of the current user for the current item in each specific user–item pair. At the same time, as mentioned earlier, we can trace back to specific keywords in the original textual reviews according to the position of the maximum node in the attention vector we get, thus greatly improving the interpretability of our recommendation model.
- We conducted comprehensive experiments on publicly accessible Amazon datasets to compare and evaluate the effectiveness of our model.

1.2. Problem definition

Let D be a collection of reviews and ratings for a particular dataset (e.g., clothes), and there are also a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$. Among them, each user has built interactions with its corresponding items, which includes the ratings and the textual reviews made by the user for the items. Here we represent each user-item interaction as a tuple $(u,\ i,\ r_{u,i},\ t_{u,i})$ where u and i respectively represent the user id and item id in the current interaction, $r_{u,i}$ is a numerical rating denoting user u's overall satisfaction towards item i, and $t_{u,i}$ is the corresponding textual review that user u wrote. The primary goal is to predict the unknown rating of a given user u to the item i with which he or she has not previously interacted.

2. Related work

Recently, a number of efforts have demonstrated the feasibility of integrating textual reviews with recommendations and promisingly acceptable performance have been achieved. In this section, the most relevant works are reviewed including *Context-Based Feature Learning*, *Diverse Preference Modeling* and *Diverse Preference Modeling*.

2.1. Context-based feature learning

A general idea for combining reviews with recommendations is to apply topic models to extract latent topics for users and items from reviews, such as HFT (McAuley & Leskovec, 2013), RMR (Ling, Lyu, & King, 2014), EFM (Zhang et al., 2014), TriRank (He, Chen, Kan, & Chen, 2015), RBLT (Tan, Zhang, Liu, & Ma, 2016), sCVR (Ren, Liang, Li, Wang, & de Rijke, 2017), A^{3}NCF (Cheng, Ding, He et al., 2018), ALFM (Cheng, Ding, Zhu, Mohan, 2018) and MMALFM (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019). As some representative work, HFT (McAuley & Leskovec, 2013) and CTR (Wang & Blei, 2011) exploited latent topics from textual reviews by techniques similar to the LDA topic model. In addition, RBLT tried to detect the topic features from rating-boost review text as latent factors. However, these work only extract features of the items. As an improvement, A^{3}NCF proposed by Cheng et al. applies a new topic model that can directly extract user preferences and item features from comments. Most of the above-mentioned work based on topic models use the bag of words (BOW) method (Collobert et al., 2011), which treats each document as a word frequency vector, thereby transforming text information into digital information that is easy to model. But the BOW method does not consider the order of words, which simplifies the complexity of the problem and also provides an opportunity for the improvement of the model. Each document represents a probability distribution composed of some topics, and each topic represents a probability distribution composed of many words. Due to the weak correlation between the components of the Dirichlet distribution random vector, our hypothetical potential topics are almost irrelevant, which is not consistent with many practical problems, which causes another remaining problem of LDA.

With the rapid development of deep learning technology, some work begins to extract additional features with the help of the powerful representation ability of deep network. Kim et al. proposed ConvMF (Kim, Park, Oh, Lee, & Yu, 2016) that they utilize a Convolutional Neural Networks (CNN) to extract the latent representations from textual reviews by considering word order and local context. Zheng et al. proposed DeepCoNN (Zheng, Noroozi, & Yu, 2017), a parallel CNN model which can separately learn the latent features of users and items from their reviews. Based on DeepCoNN, TransNets (Catherine & Cohen, 2017) extended it by attaching an extra layer to learn the representation of the user–target item review during training, and this representation could be utilized to make regularization for the output of the source network. Our work also adopts a similar idea, using CNN to extract features from the review text as auxiliary information to help learn the representation of users and items better.

2.2. Diverse preference modeling

In recent years, some researchers have tried to improve the interpretability of recommender systems by modeling the varying user preferences towards different items (Liu et al., 2019). A commonly used idea is to obtain each user's attention on different aspects of a target item from reviews. In particular, Chin et al. (2018) developed a recommendation model based on an end-to-end attention neural network, which utilized reviews and ratings to learn the different preferences of users for different aspects of items. A^{3}NCF (Cheng, Ding, He et al., 2018) applies a novel topic model which can simultaneously extract both users' preferences and items' characteristics on different aspects, and then the authors use a neural attention network to learn

the user's attention from the feature they extracted. While in CARL (Wu et al., 2019), Wu et al. use CNN to extract different aspects from review documents, and they also applied an attention network to find out which aspect of the item the user is more concerned about in the current user–item interaction. After obtaining the user's concerned aspect information and other required features, the final score prediction step falls back to two common ideas: perform calculations according to the MF idea; or feed these features into a deep neural network, using the deep network's powerful representation learning ability to directly predict the rating.

2.3. DNN-based recommender systems

Deep learning technology has already achieved great success in fields such as computer vision, pattern recognition and natural language processing. In recent years, there have also been many works applying various neural network structures to recommender systems and they have improved the recommendation performance. As one of the representative work to introduce deep neural networks into the field of recommendation, He et al. presented a neural framework to learn the nonlinear interactions between users and items, named Neural Collaborative Filtering (NCF) (He et al., 2017). The proposal of NCF is a breakthrough for the traditional collaborative filtering technology, Later, He et al. developed Neural Factorization Machines (NFM) (He & Chua, 2017), which combined the deep network with the traditional factorization machine (FM) (Rendle, 2010), and enhanced the performance of FM by modeling the interaction of high-order and nonlinear features. NRT (Li, Wang, Ren, Bing, & Lam, 2017) combined traditional collaborative filtering with gated recurrent neural networks and realized a function of simultaneously predicting ratings and generating abstract tips that simulate user experience and feelings. In our work, we not only use the Convolution Neural Network (CNN) to extract the feature of the textual reviews, but also feed the final fusion latent vector into a multi-layer perceptron to predict the rating in the final output part.

3. Intuition

When faced with different kinds of items, users may give different attention in all aspects. Even in the face of the same category, users are likely to pay different attention to various aspects of different items. This attention to different aspects is not only reflected in the user's rating of the item, but also more intuitively in the textual reviews that the user wrote. Therefore, we could extract the preferences of user u and the characteristics of item i in different aspects from the textual reviews, which are expressed as u and i respectively. The extracted features could not only be used for auxiliary modeling to represent the latent vector of the interaction between the user and the item, so as to realize the final rating prediction function, but also the interaction between u and i can be inputted into a special attentional neural network to estimate the attention vector $\boldsymbol{a}_{u,i}$ of user u for item i.

4. Aspect-aware Explainable Neural Attentional Recommender Model

In this section, we introduce our Aspect-aware Explainable Neural Attentional Recommender model(AENAR). We firstly outline the architecture of our model and the motivations behind some key components, following that we will elaborate on our attention-based module for learning the aspect-level importance of a given user–item pair. Finally, we will go through the learning and optimization details for AENAR.

Fig. 1 shows the structure of our AENAR, which consists of four components: *The Input and Text feature extraction part*, *Feature Fusion part*, *Attentive Interaction part* and the final *Rating Prediction part*.

4.1. The input and text feature extraction part

This part takes the ID of target user u, all the textual reviews written by user u (regardless of the item), the ID of target item i, all the textual reviews written for item i (regardless of user) as inputs. The ID of user u and item i are converted to one-hot encodings, and then projected to a K-dimensional dense vector via an embedding layer, where K represents the number of aspects. Since the principle of text feature extraction is the same for both users and items, we will describe it based on a specific user u in the following discussion.

Let D_u be the user document after processed, we transform D_u into a word matrix by a embedding function:

$$f : D_u \rightarrow \mathbf{M}_u \in \mathbb{R}^{l \times d} \tag{1}$$

where l is the length of this matrix (i.e., the number of words in D_u), and d denotes the dimension of each word's embedding vector. This embedding can be any pre-trained embeddings such as $word2vec$ or $GloVe$. The same process is applied to item i. Based on the processed reviews, we apply a convolution neural network (CNN) to extract the preferences of specific user u and the characteristics of specific item i respectively, represented as $\boldsymbol{\theta}_u \in \mathbb{R}^K$ and $\boldsymbol{\varphi}_i \in \mathbb{R}^K$. In this process, considering the CNN's size change characteristics, we have:

$$\frac{l - k + 2 \cdot p}{s} + 1 = K \tag{2}$$

where p is the number of layers padded around the input matrix, s is the stride of each slip of the filter in the convolution process, and k denotes the width of the filter. It should be noted that the width of the filter in the vertical direction is k, and the width in the horizontal direction should be consistent with d.

Finally, the identity-based embedding and the review-based features of users and items are passed to the next layer.

4.2. Feature fusion part

The purpose of the feature fusion part is to fuse the features we got in the previous part for better representation learning, i.e., the identity-based embedding features and review-based features. Taking user u as an example, we fuse the K-dimensional identity-based dense vector of user u with its corresponding review-based feature $\boldsymbol{\theta}_u \in \mathbb{R}^K$ obtained in the previous part (the same for items). Here, same to A^{3}NCF, we also adopt the $addition$ fusion method, and add a fully-connected neural layer directly after the fusion step. This layer also adopts the Rectified Linear Unit($ReLU$) (Nair & Hinton, 2010) activation function to introduce nonlinear factors, thereby improving the effect of its representation learning. The feature vector outputted in this part is also K-dimensional.

4.3. Attentive interaction part

The attentive interaction part aims to capture the targeted user's attention vector for different aspects of the target item. For a particular user–item pair, let $\boldsymbol{p}_u \in \mathbb{R}^K$ and $\boldsymbol{q}_i \in \mathbb{R}^K$ be the representation vectors of user u and term i learned from the previous part, respectively. We first perform element-wise production on $\boldsymbol{p}_u$ and $\boldsymbol{q}_i$, then fuse $\boldsymbol{p}_u$, $\boldsymbol{q}_i$ with the review-based features $\boldsymbol{\theta}_u$, $\boldsymbol{\varphi}_i$ extracted from CNN together and feed them into the attention network to obtain an attention vector $\boldsymbol{a}_{u,i} \in \mathbb{R}^K$ of user u for item i. As mentioned above, K is the number of aspects, so the value of the kth node in $\boldsymbol{a}_{u,i}$ reflects the weight of the kth aspect when user u pays attention to item i. Since $\boldsymbol{a}_{u,i}$, $\boldsymbol{p}_u$ and $\boldsymbol{\theta}_u$ are all K-dimensional, assuming that the maximum value of the $\boldsymbol{a}_{u,i}$ is the jth node, based on this location coordinate j, we can: (1) understand that the current user u values the jth aspect most when facing the item i, (2) trace back to the original reviews and deduce the position of the high importance word (and its context) in the comment, which could be considered as the words that can best reflect the user's attention to the jth aspect. The details of this part will be elaborated in $Section\ E$.

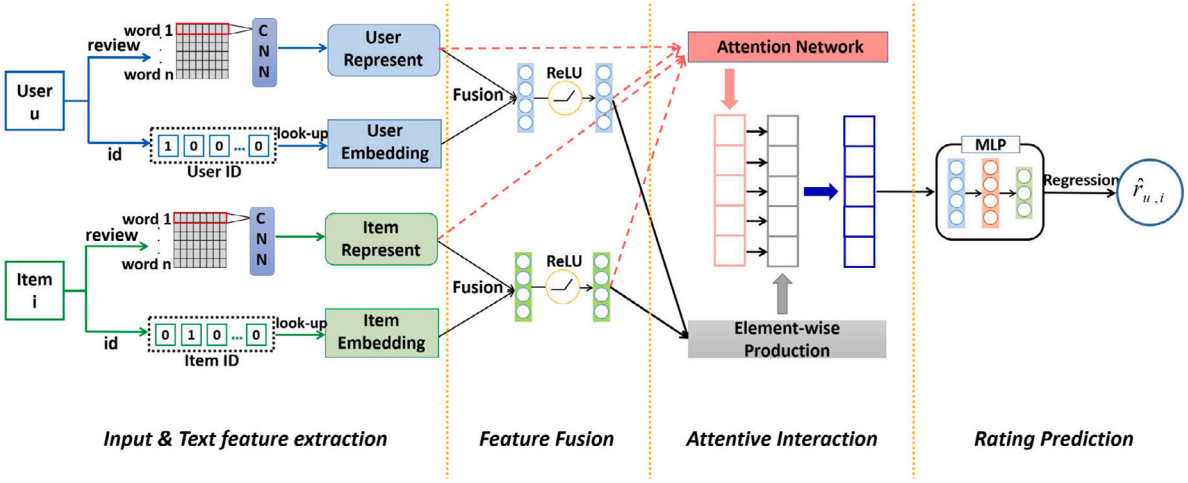

Fig. 1. The structure of our AENAR model.

At the end of this part, we merge $\boldsymbol{p}_u$, $\boldsymbol{q}_i$ and $\boldsymbol{a}_{u,i}$ to obtain a final representation $\boldsymbol{V}_{u,i} \in \mathbb{R}^K$ of the user–item pair:

$$\boldsymbol{V}_{u,i} = \boldsymbol{p}_u \odot \boldsymbol{q}_i \odot \boldsymbol{a}_{u,i} \tag{3}$$

where $\odot$ denotes the element-wise production, and the representation vector $\boldsymbol{V}_{u,i}$ will be passed into the next part for the rating prediction.

4.4. Rating prediction part

This is the final output part, we apply a multilayer perceptron (MLP) as the final rating prediction network. As mentioned above, the interaction feature vector $\boldsymbol{V}_{u,i}$ will be passed as the input of this part. Then the predicted rating rui is obtained as follows:

$$\hat{\boldsymbol{r}}_{u,i} = \sigma\left(\boldsymbol{W} \cdot \boldsymbol{V}_{u,i} + \boldsymbol{b}\right) \tag{4}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ denote the weight matrix and bias vector, respectively. And σ is the activation function, here we adopt the $ReLU$ function for all the layers in this network.

4.5. Attention mechanism

In this section, we introduce the attention mechanism in our AENAR which can capture the targeted user's attention vector for different aspects of the target item. For a particular user–item pair, as mentioned above, we have got the representation vectors $\boldsymbol{p}_u \in \mathbb{R}^K$ and $\boldsymbol{q}_i \in \mathbb{R}^K$ of user u and term i, as well as their review-based features $\boldsymbol{\theta}_u \in \mathbb{R}^K$, $\boldsymbol{\varphi}_i \in \mathbb{R}^K$. The attention vector is computed as:

$$\hat{\boldsymbol{a}}_{u,i} = \sigma\left(\boldsymbol{W}\left[\boldsymbol{p}_u; \boldsymbol{q}_i; \boldsymbol{\theta}_u; \boldsymbol{\varphi}_i\right] + \boldsymbol{b}\right) \tag{5}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ denotes the weight matrix and bias vector that project the input into a hidden layer, respectively. $\boldsymbol{\sigma}$ is the activation function for this attention network, and $\left[\boldsymbol{p}_u; \boldsymbol{q}_i; \boldsymbol{\theta}_u; \boldsymbol{\varphi}_i\right]$ denotes the *concatenation* of all the four vectors.

The final weight of attention are obtained by normalizing the above attention scores with the softmax function, which can be interpreted as the importance of the kth node (i.e., the kth aspect) in this attention vector for this interaction:

$$a_{u,i,k} = \frac{\exp\left(\hat{a}_{u,i,k}\right)}{\Sigma_{k=1}^{K}\left(\hat{a}_{u,i,k}\right)} \tag{6}$$

4.6. Learning

Since our task is rating prediction, which can be treated as a regression problem. For which the squared loss is commonly used as the objective function:

$$L = \sum_{(u,i)\in D}\left(\hat{r}_{u,i} - r_{u,i}\right)^2 \tag{7}$$

where $r_{u,i}$ is the ground truth rating made by the user u to the item i, $\hat{r}_{u,i}$ is the rating predicted from our model. To optimize the objective function, we adopt the Adaptive Moment Estimation ($Adam$) as the optimizer.

5. Experiments

5.1. Experimental settings

Datasets: The six publicly accessible datasets from the Amazon Product Review dataset are selected to cover different domains and scales to evaluate our proposed model. For each interaction in these datasets, *user ID*, *item ID*, *overall rating* and *review* were extracted for the experiments. All the reviews in these datasets were cleaned by removing stop words and punctuation. Then all the comment texts in each dataset were merged into a review document. For the six evaluation datasets, six comment documents were generated. Considering that the words with too low-frequency (less than 3 times) may be misspelled, and conversely those too high-frequency (more than 30,000 times) words mentioned by most users cannot reflect the individual characteristics, thus these words are removed. The characteristics of the chosen datasets (after processed) are summarized in Table 1.

Experiments details: We randomly split the dataset into training (80%), validation (10%), and testing (10%) set for each user. In addition, for the textual reviews after being cleaned in our datasets, we integrate all the reviews made by each user u (regardless of the item) into one comment document D_u, and by the same way, all the reviews written for each item i (regardless of the user) are also merged into one document D_i. For the documents D_u and D_i, we adjust the total number of words (i.e., the length of the document) to l by *truncating* or *padding*. Here, for the convenience of subsequent experiments, we set l to 256. That is, let $\mathcal{T}$ denotes the collection of all the words and their contextual order in one comment document, if the number of words in the original document is greater than or equal to 256, we will truncate at the 256th word in $\mathcal{T}$ and discard the words after it. Conversely, if the number of words in the document is less than 256, we will copy $\mathcal{T}$ and paste it from the end of the original document. Repeat the above

Table 1
Statistics of the evaluation datasets (after processed).

Datasets	# Users	# Items	# Ratings	# Sparsity	# Words per user	# Words per item	# Words per interaction
Amazon Instant Video	5130	1685	37,126	0.9957	223.948	681.811	30.945
Tools and Home Improvement	16,638	10,217	134,476	0.9992	170.903	278.309	21.145
Video Games	24,303	10,672	231,780	0.9991	246.938	562.345	25.892
Clothing Shoes and Jewelry	39,387	23,033	278,677	0.9997	37.173	63.566	5.254
Health and Personal Care	38,609	18,534	346,355	0.9995	67.999	141.653	7.580
Home and Kitchen	66,519	28,237	551,682	0.9997	77.944	183.615	9.398

Table 2
Overall performance comparisons of adopted methods in terms of RMSE. For each dataset, the best performance is highlighted in boldface while the second best result is highlighted in underlined.

Datasets	MF	NARRE	A^{3}NCF	ANR	CARL	AENAR
Amazon Instant Video	1.131	<u>0.968</u>	0.977	0.982	0.978	**0.941**
Tools and Home Improvement	1.038	0.979	0.979	<u>0.974</u>	0.976	**0.949**
Video Games	1.207	1.057	<u>1.053</u>	1.099	1.054	**1.033**
Clothing Shoes and Jewelry	1.109	1.050	1.067	1.058	<u>1.048</u>	**1.023**
Health and Personal Care	1.122	1.033	1.037	1.046	<u>1.023</u>	**1.008**
Home and Kitchen	1.104	1.034	1.041	1.045	<u>1.033</u>	**1.011**

process until the number of words in the document is greater than 256, and then we can truncate at the 256th word in $\mathcal{T}$ and discard the words after it like what we mentioned above.

Since we have set l to 256, as shown in Eq. (6), we can change the dimension K of the final output vector of CNN by changing parameters k, p, s. For example, by setting $k = 2$ and $s = 2$ with a padding size $p = 0$, we can get the output dimension $K = 128$. Which means, we extract 128 aspects from the original 256 words, and each context window (non-overlapping) formed by every 2 words in the original text corresponds to one aspect. In the subsequent experiments, we tried $K = 256, 128, 64, 32, 16, 8$, the result and discussion are shown in Section 6.

Baselines: To evaluate our proposed AENAR's performance of the rating prediction task, we compare the proposed model to several state-of-the-art methods which utilize both textual reviews and ratings to improve the recommendation performance:

- **NARRE** (Chen et al., 2018): Neural Attentional Rating Regression with Review-level Explanations. It built a recommendation model which can select highly useful reviews by introducing a neural attention mechanism.
- **A^{3}NCF** (Cheng, Ding, He et al., 2018): An Adaptive Aspect Attention Model for Rating Prediction. It developed a novel topic model to extract both user and item features from reviews and proposed an aspect-aware rating prediction method based on an adaptive aspect attention modeling design.
- **ANR** (Chin et al., 2018): Aspect-based Neural Recommender. By designing an attention mechanism, it focused on the relevant parts of these reviews while learning the representation of aspects in the task, so as to implement aspect-based representation learning for users and projects. In addition, it utilized the idea of common attention to jointly estimate the importance of aspect-level users and projects, thus allowing more fine-grained interactions between users and projects to be modeled.
- **CARL** (Wu et al., 2019): A Context-Aware User-Item Representation Learning for Item Recommendation. For each user–item pair, it learns context-aware representations on their individual characteristics with their interactions together by exploiting both textual reviews and user–item interaction data.

In addition to the above state-of-the-art methods, we also introduced matrix factorization (MF), the most classic collaborative filtering algorithm, to run on these datasets to verify the effectiveness of attention mechanism, deep learning and other modules. A detailed analysis of the ablation experiments is given in Section 6.

Evaluation metric: In order to evaluate the performance of our proposed model and these baselines, we adopt the Root Mean Square Error (RMSE) as the evaluation metric. Given a ground-truth rating $r_{u,i}$ and the rating $\hat{r}_{u,i}$ predicted from our model, the RMSE score can be calculated as:

$$RMSE = \sqrt{\frac{1}{|C_t|}\sum_{(u,i)\in C_t}\left(\hat{r}_{u,i} - r_{u,i}\right)^2} \tag{8}$$

Where C_t denotes the collection of all the user–item pairs obtained from the testing set, and a lower RMSE score indicates a better performance.

5.2. Performance evaluation

The overall performance of the proposed AENAR and baseline models on the six evaluation datasets is presented in Table 2.

As a traditional collaborative filtering algorithm proposed more than a decade, MF did not rely on additional information from textual reviews, and no tools such as deep learning and attention mechanism were applied to learn abstract features. Therefore, compared with others, the worst rating prediction performance was obtained by MF. The huge gap between the performance of MF and other algorithms unambiguously displayed the advantage of rating predication by introducing textual information, deep learning and attention mechanism.

Based on MF, ANR constructed word projection matrix from reviews and utilized attention mechanism for learning aspect-level representations of both users and items, and therefore significantly improved the rating predication performance. As another algorithm which utilized the reviews, A3NCF extracted the topical feature from reviews via a novel designed topic model, and applied MLP instead of the element-wise production in the final rating predicating process to achieve better recommending performance. As for the rest three models (NARRE, CARL and our proposed AENAR), all of them utilized their respective designed deep network to extract features from reviews, which retains the orders between words and avoids problems of losing of some important contextual information like what may happen in some topic models.

Apparently, the proposed AENAR model exhibited decent improvements compared to all the five state-of-the-art methods owing to that our model applies more complicate interactions between the latent vectors of users and items via a designed attention network.

5.3. Model analysis

Number of predictive factors: The performance of AENAR by varying the numbers of latent dimensions K is shown in Fig. 2. Owing to the similarity of Instant Video and Clothing Shoes and Jewelry with other datasets, these two datasets were omitted. The Health and Home datasets possess very close RMSE values over the studied range of K. Compared with them, the Video Games displays slightly enhanced RMSE values except the first point. In contrast, the Tools exhibits significantly reduced RMSE. Except the first two points, AENAR work

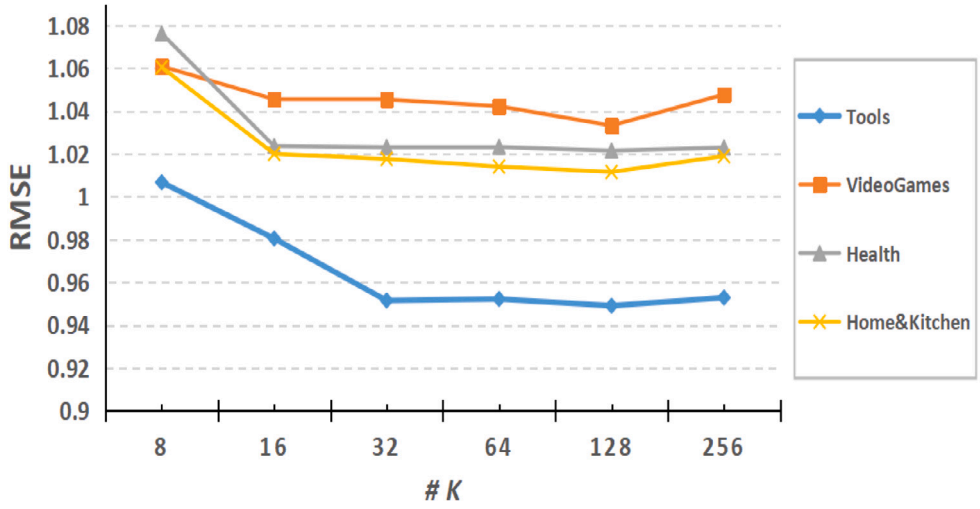

Fig. 2. Impact of latent dimensions K across the four datasets.

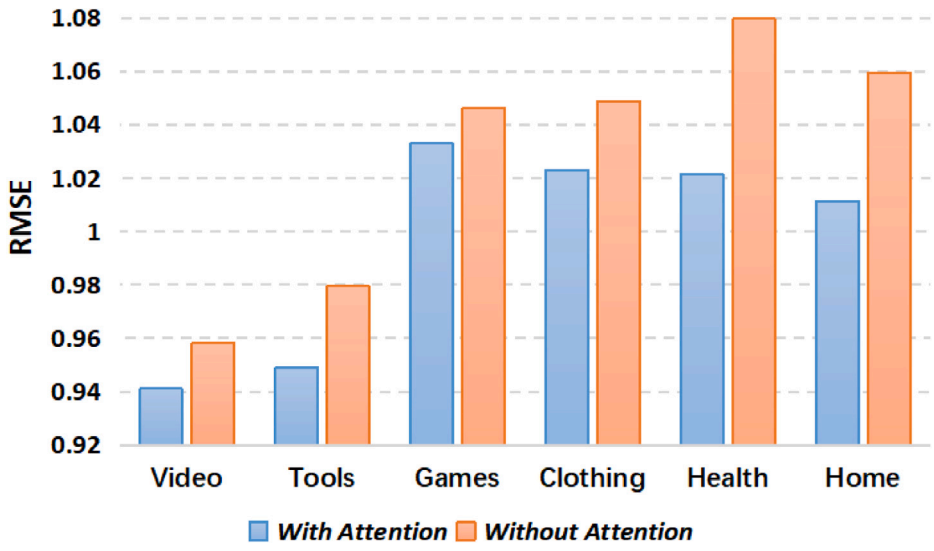

Fig. 3. Impact of attention component across the six datasets.

consistently on all the four datasets within a wide range of K values from 32 to 256 with negligible variation, indicating good stability.

The impact of attention component: In our AENAR model, the attention component is used to capture the users' attention vectors for different aspects of items. Here, we conducted an ablation experiment to determine the effectiveness of the attention mechanism. The attention layer in our model was removed, and instead the feature vectors obtained from the feature fusion layer were directly inputted into the final rating prediction network, and the experiment was run again on the same six datasets.

The impact of attention component across the six datasets is displayed in Fig. 3. For all the datasets, the RMSE values are apparently increased without attention component by up to about 6%. Interestingly, with large amount of interactions and high sparsity, the results on the datasets of *Health and Personal Care* and *Home and Kitchen* exhibited significantly reduced performance. This observation suggests that the attention mechanism can effectively help learn the representations from the interactions between the users and items, and thus improve the performance of rating prediction, especially on sparse and large datasets.

Cold-start setting: As shown in Table 1, the datasets in practical recommendation systems are often extremely sparse. With such a limited number of interactions, it is challenging to do well for training the model. Therefore, recommendation system models represented by matrix factorization often suffer from the cold-start problem. In our model, since the textual review information of users and items is also integrated into the learning process, i.e., in addition to the original interaction information (ratings), user preferences and item characteristics are additionally mined, so our model could alleviate the cold-start problem to a certain extent. To demonstrate the capability of our model on dealing with users with very limited interactions, we designed and conducted the following experiment:

In order to simulate the phenomenon of cold-start due to a limit of interactions corresponding to users, we sampled and removed some interactions from the original dataset. For the processed dataset, it was randomly split into training, validation, and testing sets in the ratio of 8:1:1 for each user. With this setting, 80% of the interactions corresponding to each user will be fed into the training set, e.g., a user with 10 interactions will keep 8 samples in the training set. According to our design, all samples in the training set will eventually be retained

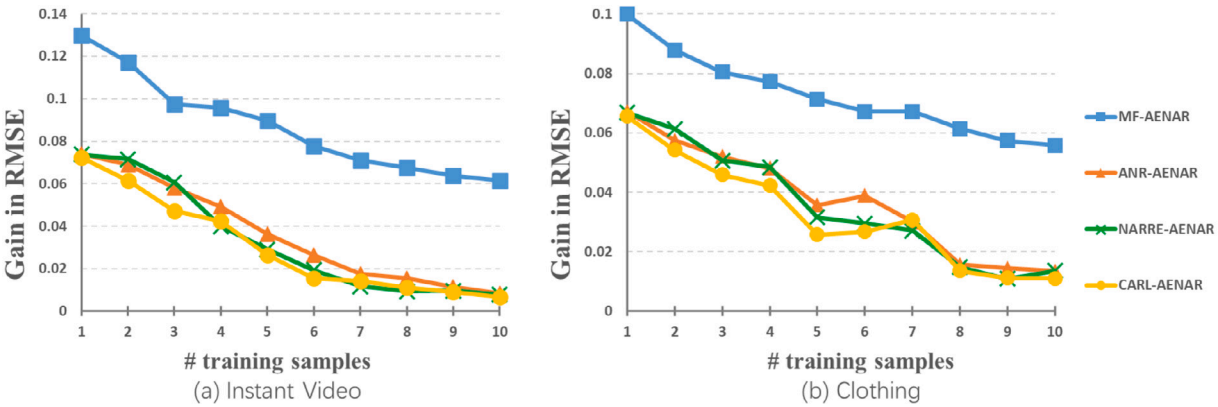

Fig. 4. Gain in RMSE of AENAR over baselines in the cold-start setting.

Table 3
Highlighted words by attentive weights in the review documents.

Pair1 (user *u*, item *1*, Rating = 5.0)	
user *u*' s review document (*after processed*):	**user *u*' s review for item *1*:**
shortly introduced thinking sliced bread man handheld stiff cumbersome unbalanced worse except height pain spend step hp yes tightly loose twice tighten solves blower blow jigsaw nicely primarily miter stationary dc basement moved disc pick roughly pvc thru fault favorite siding removal vinyl equally impressed slip tug j channel window installing narrow snap properly portable replacing double measure quot sheet plywood clamping raised panel strength sold borrowed friend blew nicer balance buck process grab truck nailing pneumatic compressor finding tripping lubed rarely trim woodworking brad nailer collecting everyone jobsites knocked purr kitten fell happened broke outfeed wheel scratch tinkering	This is one of, if not, the best portable tablesaw you can buy. It's fairly light, powerful, accurate, and the included base just makes it that much better. I have had mine for about 9 months and the only problem I've had was replacing the blade. The fence is so accurate that you don't need to double check it with a tape measure. If the tape on the fence say's 19 15/16", that's exactly what you'll get. The saw will cut sheets of plywood in half without a problem. If you want a good tablesaw for home or work, or if you don't have room for a larger saw, this is the saw to get.

Pair2 (user *u*, item *2*, Rating = 1.0)	
user *u*' s review document (*after processed*):	**user *u*' s review for item *2*:**
shortly introduced thinking sliced bread man handheld stiff cumbersome unbalanced worse except height pain spend step hp yes tightly loose twice tighten solves blower blow jigsaw nicely primarily miter stationary dc basement moved disc pick roughly pvc thru fault favorite siding removal vinyl equally impressed slip tug j channel window installing narrow snap properly portable replacing double measure quot sheet plywood clamping raised panel strength sold borrowed friend blew nicer balance buck process grab truck nailing pneumatic compressor finding tripping lubed rarely trim woodworking brad nailer collecting everyone jobsites knocked purr kitten fell happened broke outfeed wheel scratch tinkering	I bought this router shortly after it was introduced thinking it was the next best thing since sliced bread. Man was I wrong! For handheld use, it is stiff, cumbersome, and unbalanced. It's worse in a router table, which was the main reason I bought it. I thought the second power switch would be handy for table use, except you have to lock the handle switch in the on position. And in order to do that, I had to tape mine just so the switch would lock into place. The height adjustment is a pain to work with. Overall, this is one of the worse routers I have used. If you want a router for a router table, spend the extra money and step up to a 3 HP router.

in the range of 1 to 10, and the same for the validation and test sets. In this way, those entries in the training set with a very small number of samples (e.g., users with only 1 interaction or 2 interactions) simulate the realistic cold-start situation. Then, we evaluate the performance of users who have the number of interactions from 1 to 10 in the training set, and compare the evaluation results of our model with baselines to verify the ability of our model to alleviate the cold-start problem. In Fig. 4, we show the gain of RMSE (y-axis) grouped by the number of samples of users in the training set(x-axis), where the gain is defined as the average RMSE of the baseline minus that of our model corresponding to the number of samples (e.g., "MF-AENAR"). Obviously, the positive value of this gain indicates that our model has better prediction results for this grouping of training samples. As shown in Fig. 4, our model achieves better results against cold-starting compared to other baselines, and the larger gain between AENAR and MF indicates that integrating the review information into the training process can effectively alleviate the cold-start problem.

Visualize attention keywords: To further understand the process of aspect-level feature learning of our AENAR model, two user–item pairs were selected randomly from the real world dataset of *Tools and Home Improvement* to demonstrate the efficiency of capturing users' preferences for different items. The two interacted items for the selected user u are denoted as item1 and item2, respectively. The scores rated by user u for these two items are 5.0 and 1.0, respectively. The processed results by our model is shown in Table 3. The user's review document (after processed) is shown in the left part of the table, while the right part are the textual reviews made by user u on the two items respectively. These two user–item pairs were inputted into AENAR and

the attention vectors are obtained after the running processes. For each vector, the position of the node with the largest value is recorded, the user's review document is traced back and the corresponding words are highlighted accordingly. Subsequently, the context of the reviews in the right can be marked. Promisingly, the words highlighted by AENAR are substantially different for the two items. For item 1, the words captured by AENAR are positive or neutral, while most of the captured words for item 2 are negative. This result indicates that AENAR can unambiguously identify users' preference for different items.

6. Conclusion

In this work, an aspect-aware recommender model based on a novel designed adaptive attention network has been proposed to make explainable rating prediction. Both the rating scores and textual reviews of each user–item pair are exploited in our proposed AENAR. By evaluating an existing user–item pair in light of their individual characteristics and their interactions together, a better rating prediction performance was achieved by the proposed model. The experimental results indicate that AENAR consistently exhibits smaller RMSE values than each of the five state-of-the-art alternatives. Furthermore, AENAR can trace back to the keywords in the original textual reviews for visible and direct illustration. This work offers an effective method for rating prediction task and may provide guidance for other researchers in design and develop efficient models for explainable recommendation.

In this work, we choose the review text as the side information of the recommendation system, so as to improve the recommendation performance and interpretability of the recommendation system. Meanwhile, there are also many other information that can also be used as side information for recommendation systems (e.g., images or labels of products, etc.). As for the further research work, we have the idea to try to apply the combination of multiple side information to improve the performance of the recommendation systems.

CRediT authorship contribution statement

Tianwei Zhang: Data Processing, Conduct experiments and compile results, Paper Writing. **Chuanhou Sun:** Data Processing, Conduct experiments, Paper Writing. **Zhiyong Cheng:** Paper reviewing and editing, Experimental results analysis. **Xiangjun Dong:** Experimental results analysis, paper reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62076143.

References

Bell, R. M., & Koren, Y. (2007). Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, *9*(2), 75–79.

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

Catherine, R., & Cohen, W. (2017). Transnets: Learning to transform for recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 288–296).

Chen, C., Zhang, M., Liu, Y., & Ma, S. (2018). Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1583–1592).

Cheng, Z., Chang, X., Zhu, L., Kanjirathinkal, R. C., & Kankanhalli, M. (2019). MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems (TOIS)*, *37*(2), 1–28.

Cheng, Z., Ding, Y., He, X., Zhu, L., Song, X., & Kankanhalli, M. S. (2018). A^3ncf: An adaptive aspect attention model for rating prediction. In *IJCAI* (pp. 3748–3754).

Cheng, Z., Ding, Y., Zhu, L., & Kankanhalli, M. (2018). Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 2018 world wide web conference* (pp. 639–648).

Cheng, Z., Ding, Y., Zhu, L., & Mohan, K. (2018). Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 27th International Conference on World Wide Web* (pp. 639–648). IW3C2.

Chin, J. Y., Zhao, K., Joty, S., & Cong, G. (2018). ANR: Aspect-based neural recommender. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 147–156).

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Guan, X., Cheng, Z., He, X., Zhang, Y., Zhu, Z., Peng, Q., et al. (2019). Attentive aspect modeling for review-aware recommendation. *ACM Transactions on Information and System*, *37*(3), 28:1–28:27.

He, X., Chen, T., Kan, M. Y., & Chen, X. (2015). TriRank: Review-aware explainable recommendation by modeling aspects. In *The 24th ACM International*.

He, X., & Chua, T.-S. (2017). Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 355–364).

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web* (pp. 173–182).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1181, URL: https://www.aclweb.org/anthology/D14-1181.

Kim, D., Park, C., Oh, J., Lee, S., & Yu, H. (2016). Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 233–240).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *42*, In *IEEE Computer* (08), (pp. 42–49).

Le, T.-H., & Lauw, H. W. (2020). Synthesizing aspect-driven recommendation explanations from reviews. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence (IJCAI-20)*.

Li, P., Wang, Z., Ren, Z., Bing, L., & Lam, W. (2017). Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 345–354).

Ling, G., Lyu, M. R., & King, I. (2014). Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM conference on recommender systems* (pp. 105–112).

Liu, F., Cheng, Z., Sun, C., Wang, Y., Nie, L., & Kankanhalli, M. (2019). User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1526–1534).

Liu, H., Wang, W., Xu, H., Peng, Q., & Jiao, P. (2020). Neural unified review recommendation with cross attention. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1789–1792).

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on recommender systems* (pp. 165–172).

Mukherjee, R., Peruri, H. C., Vishnu, U., Goyal, P., Bhattacharya, S., & Ganguly, N. (2020). Read what you need: Controllable aspect-based opinion summarization of tourist reviews. arXiv preprint arXiv:2006.04660.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Icml*.

Peña, F. J., O'Reilly-Morgan, D., Tragos, E. Z., Hurley, N., Duriakova, E., Smyth, B., et al. (2020). Combining rating and review data by initializing latent factor models with topic models for top-n recommendation. In *Fourteenth ACM Conference on Recommender Systems (RecSys)* (pp. 438–443).

Ren, Z., Liang, S., Li, P., Wang, S., & de Rijke, M. (2017). Social collaborative viewpoint regression with explainable recommendations. In *Proceedings of the tenth ACM international conference on web search and data mining* (pp. 485–494).

Rendle, S. (2010). Factorization machines. In *2010 IEEE International Conference on Data Mining* (pp. 995–1000). IEEE.

Tan, Y., Zhang, M., Liu, Y., & Ma, S. (2016). Rating-boosted latent topics: Understanding users and items with ratings and reviews. In *IJCAI, Vol. 16* (pp. 2640–2646).

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 448–456).

Wang, X., He, X., Feng, F., Nie, L., & Chua, T.-S. (2018). Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 world wide web conference* (pp. 1543–1552).

Wu, L., Quan, C., Li, C., Wang, Q., Zheng, B., & Luo, X. (2019). A context-aware user-item representation learning for item recommendation. *ACM Transactions on Information Systems (TOIS)*, *37*(2), 1–29.

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., & Ma, S. (2014). Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 83–92).

Zheng, L., Noroozi, V., & Yu, P. S. (2017). 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* (pp. 425–434).