Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210175

Interpretable Dual-Feature Recommender System Using Reviews[1]

Jing Sheng LEI [a], Chen Si Cong ZHU [a], Sheng Ying YANG [a,2], Guan Mian LIANG [b], Cong HU [a] and Wei SONG [a]

[a] *Department of School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, PR China*

[b] *Cancer Hospital of the University of Chinese Academy of Science, Hangzhou, PR China*

Abstract. Reviews have been commonly used to alleviate the sparsity problem in recommender systems, which has significantly improved the recommender performance. The review-based recommender systems can extract users features and items from review texts. The existing models such as D-Attn and NARRE employ convolutional neural networks and a coarse-grained attention mechanism to code reviews that have been embedded using the static word embedding, ignoring the long distance text information and lacks interpretability. To overcome these problems, this paper proposes the DNRDR (Dual-feature Neural Recommender with Dual-attention using Reviews) model, which can extract dual features of review text and can also enhance the interpretability using the word-level and review-level attention mechanisms. The proposed model is verified by experiments and compared with the state-of-the-art models. Besides, the dual-level attention mechanism can be visualized to improve interpretability.

Keywords. Neural networks, recommender systems, review text, multiple features

1. Introduction

As an effective method to alleviate information sparsity in the e-commerce field, recommender systems enable consumers to find their interested information. Collaborative filtering technology that can find out users' preferences and accordingly predicts products that users may like using the mining of users' historical behavior has been the most widely used recommendation algorithm currently [1, 2, 3, 4, 5, 6]. However, it is difficult to produce reliable recommendations for users with few ratings [7]. In addition, contextual information cannot be clearly reflected in the rating matrix. These problems can be solved using textual reviews [8]. Specifically, users will explain the reason for his rating directly, and then, based on the review text, the corresponding recommendation is given. The existing review-based recommender systems have been using neural networks to ex-

[1]This work is supported by Natural Science Foundation of China (No. 61672337, 61972357), Zhejiang Key R&D Program (No. 2019C03135), Basic Public Welfare Research Project of Zhejiang(LGF19F020003).

[2]Corresponding author: Shengying Yang, Department of School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, 310023, PR China; E-mail: syyang@zust.edu.cn.

tract features, achieving good recommendation results [7, 9, 10]. It should be mentioned that although CNNs(Convolutional Neural Networks) have the ability of local semantic feature extraction, the fixed length of convolution kernels limits the extraction performance of the word order and contextual connection, which could easily result in an understanding deviation of review texts. Furthermore, although CNNs have been proven to be effective in decreasing prediction error, learned filters in a convolutional layer provide little help in interpreting the features of users or items. Attention mechanisms have been proposed to enhance interpretability since attention scores can quantify the importance of words or reviews.

To overcome those problems, in this study, a dual-feature processing module consisting of a CNN and a bidirectional Long Short-Term Memory (LSTM) [11] network is proposed to reduce information loss. In the proposed module, a CNN is used to extract short-distance features from textual reviews and supplemented with long-distance features obtained by an LSTM. Besides, every comment is modeled by a dual-dimension attention mechanism to improve the interpretability and to judge the importance of reviews and words. In addition, a dynamic word vector BERT(Bidirectional Encoder Representations from Transformers) [12] is exploited, and the word vector is calculated according to the context.

The contributions of this paper can be summarized as follows:

1. A neural network model DNRDR, which uses word vectors embedded with BERT as input and a dual-feature processing module to extract short- and long-distance features of users or items review texts, is proposed.
2. Interpretable attention mechanisms, i.e., the word-level attention mechanism used to differentiate the importance of every word and the review-level attention mechanism used to judge the usefulness of every review, which can enhance both interpretability and visualization of a recommender, are introduced.
3. Comparative experiments on four benchmark datasets from Amazon are conducted, and the experimental results have shown that the proposed DNRDR outperforms the baseline recommendation methods. In addition, the proposed text processing module has been proven to be effective in other review-based models, and the interpretability and visualization benefits from the proposed dual-dimension attention mechanism have also been demonstrated.

2. Related Work

There are two research branches closely related to the subject studied in this work. One includes the review-based recommender systems proposed in recent years, and another includes methods for extracting review features.

2.1. Review-based recommender systems in recent years

The Matrix Factorization (MF) [2] that simulates a user's explicit feedback, such as a score, by mapping users and items to a potential space has attained great attention in recent years [2, 4, 5, 13], but it is strongly influenced by sparsity. This limitation can be overcome by introducing review-based models [7, 9, 14, 15, 16], where reviews are mapped to the hidden space so that the neural networks can effectively extract advanced

features. However, most recommenders combine all reviews of users (items) into one long document that is used as an input. Although this approach lowers the prediction error, it neglects differences between reviews. For instance, users comments can refer to a variety of goods, so it would be absurd to predict a score of a movie based on reviews of electronic products. In the 2018 World Wide Web Conference, the NARRE [10] used a review-level attention mechanism to determine the usefulness of every review and further improved the prediction accuracy.

2.2. Review features extraction

The core of the review-based recommender systems is feature extraction from review texts. Many text-processing methods based on deep learning technology have been proposed recently, and good performances have been achieved. The TextCNN [17] conveys word embeddings into a one-dimensional convolution layer, obtaining sentences of a fixed length by pooling operation. In [18], it has been proved that CNNs can extract local features but not long-distance features. Compared to the CNNs, RNNs adopt a linear sequence structure to collect input information from the front to the back continuously, which could easily cause a severe gradient disappearance or gradient explosion [11]. In order to solve this problem, in the proposed model, an LSTM is introduced, adding intermediate state information, so as to alleviate the gradient disappearance problem. With respect to the word embedding, the Transformer [19] was proposed in 2017 by Google in the in the research article about the machine translation task. It abandons the traditional neural network architectures and completely relies on the attention mechanism. The BERT [12] model based on the Transformer can effectively extract context information and thus provides significant progress in the field of natural language processing [20, 21, 22, 23].

3. OUR PROPOSED MODEL

3.1. Structure of DNRDR

This section presents the proposed DNRDR model. As shown in Fig 1, our proposed model consists of two similar networks connected in parallel. Since the network structure of the item part is the same as that of the user part, this section focuses on the left part of the network that processes user input data.

The DNRDR input is a group of reviews, including j user reviews of d words $\{R_{u1}, R_{u2}, \ldots, R_{un}\}$, which are then mapped by BERT [12] to a comment matrix $W_{u,r} \in \mathbb{R}^{j\times d\times n}$ where n refers to the word vector dimension. The BERT mapping is performed on each item review to obtain an item comment matrix $W_{i,r} \in \mathbb{R}^{j^{'}\times d\times n}$, where $j^{'}$ refers to the review number of each item.

In the proposed model, the word-level attention module is first used to learn the most informative words. Assume $w_{w-level} \in \mathbb{R}^{t\times n}$ is a weight matrix of the input word vector; then, the weighting scores for a word are computed as follows:

$$S_i = softmax(W_a \times \sigma(w_{w-level} \times W_{u,r_i}), i \in [1, j] \tag{1}$$

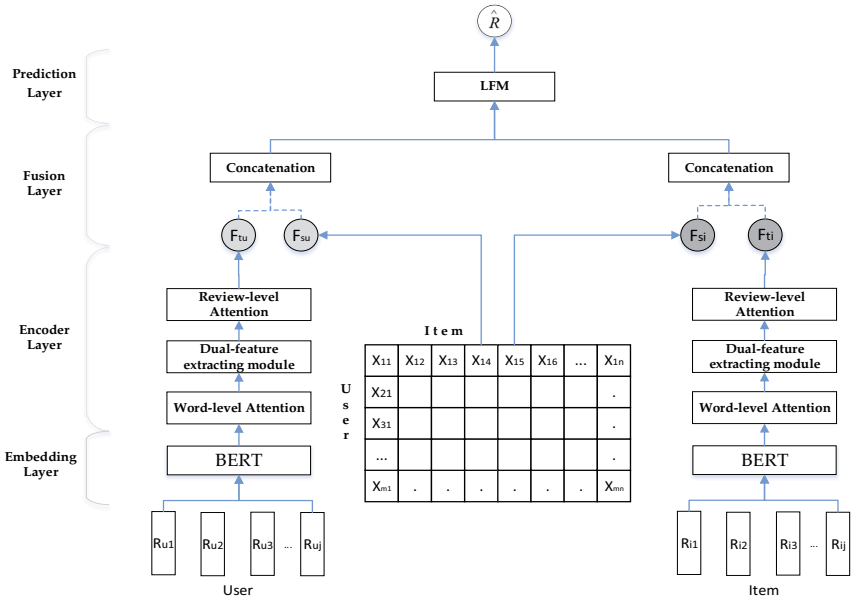

Figure 1. The structure of Dual-feature Neural Recommender with Dual-attention using Reviews is divided into User part(Left), Rating Matrix(Middle) and Item part(Right) from a horizontal perspective. From a longitudinal perspective, the model consists of Embedding Layer, Encoder Layer, Fusion Layer and Prediction Layer.

where $W_a \in \mathbb{R}^{1\times t}$ are the attention parameters and t is the hyperparameter of the attention dimension. σ represents the tanh function for the activation function. $S_i \in \mathbb{R}^{1\times d}$ are the attention scores for each word. Then let $O_{ui} = S_i \times W_{u,r_i}, i \in [1, j]$ be the weighted word vectors.

The encoder layer is the core of DNRDR, consisting of two parallel networks that extract the long and short distance features of the review texts, respectively. The long-distance review features for a word h_{ui} are obtained through a two-way LSTM. Then, the $2l$-dimensional features of the d words are used to obtain the hidden state of the LSTM network $h_{u,L} \in \mathbb{R}^{d\times 2dl}$:

$$h_{u,L} = h_{u1} \oplus h_{u2} \oplus \ldots \oplus h_{ui}, i \in [1, j] \tag{2}$$

Then a fully connected layer is used to prepare the long- and short-distance features for merging by setting them to the same dimension. The short-distance reviews features are obtained by a convolutional layer composed of m neurons where each neuron is associated with a convolution kernel as follows::

$$o_i = ReLU(O_{u,i} * K_i + b_i) \tag{3}$$

where b_i is the paranoid term, and $*$ represents the convolution operation.

Suppose $\{o_1, o_2, \ldots, o_{d-t+1}\}$ are the features obtained by (5), the final feature of the convolution kernel can be calculated using the maximum pooling [24] operation so the short-distance review features $Review_C$ are merged with the features obtained by m

neurons. Then the long- and short-distance features of the user's reviews are combined to obtain the user's text features $Review_{tu} \in \mathbb{R}^{1\times 2m}$.

The review-level attention machanism [25] is utilized to judge which reviews are the most useful. The attention score vector describes the usefulness of reviews, and it is used to weigh each of the comments as follows:

$$\alpha_L = softmax(\tanh(w_{r-level} \times Review_{tu}{}^{\mathrm{T}})) \tag{4}$$

where $\alpha_L \in \mathbb{R}^{1\times j}$, $w_{r-level} \in \mathbb{R}^{1\times 2j}$ are the parameters. The attention score vector describes the usefulness of the reviews, and it is used to weight each of the comments$F_{tu} \in \mathbb{R}^{1\times j}$ as follows:

$$F_{tu} = \alpha_L \times Review_{tu} \tag{5}$$

Similarly, the weighted features F_{ti} of an item's review texts can be obtained. Then the proposed model uses LFM [2] which is an algorithm based on matrix factorization technology, to extract the hidden factors of users F_{su} and items F_{si} in the rating matrix.

In the fusion layer, user review features and user rating featuresare are concatnated to a unified vector F_U and the item features F_I can be gained in the same way. Then the prediction layer predicts the users rating for item by using the dot product to map user and item features to the same space as follows:

$$\hat{R}_{u,i} = \omega \times (F_U \otimes F_I) + b_u + b_i + \mu \tag{6}$$

where μ is the fully connected layer parameter.

3.2. Model Learning

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

Γ represents the user-item datasets, $\hat{r}_{u,i}$ is the predicted score of user u for product i, and $r_{u,i}$ is the true rating in the training set. In order to minimize the loss function, the Adam (Adaptive Moment Estimation) [26] estimator is used as an optimizer. In order to prevent overfitting, the dropout [27] is used to solve this problem, whose main idea is to randomly discard some neurons during the training process.

4. EXPERIMENTAL VERIFICATION

4.1. Datasets

In the experiments, four public datasets from the Amazon 5-core [28] review datasets were used including Digital_Music (DM), Movies_and_TV (MT), Kindle_Store (KS), Toys_and_Games (TG). Every user and item in these sets contains at least five reviews. The smallest dataset is TG, which contains 167,597 samples; thus, they all have enough samples to build and verify the model, i.e., they contain enough semantic information to guide modeling the nerual networks.

Table 1. Prediction Performance of different models

	NMF	LFM	SVD++	HFT	DeepCoNN	D-Attn	NARRE	DNRDR
Digital_Music	1.228	0.851	0.837	0.849	0.828	0.819	0.816	0.796*
Movies_and_TV	1.415	1.281	1.273	1.211	1.019	1.048	0.998	0.981*
Kindle_Store	0.995	0.623	0.616	0.622	0.614	0.611	0.606	0.593*
Toys_and_Games	1.429	0.869	0.856	0.852	0.849	0.840	0.835	0.827*

4.2. Baselines

We compare our model with the state-of-the-art recommendation models including the NMF [4], LFM [2], SVD++ [5], HFT [29], DeepCoNN [9], D-Attn [7] and NARRE [10]. The NMF, LFM, and SVD++ models ultilized only the rating matirx in the model training phase, while the DeepCoNN and D-Attn models used only the reviews; the NARRE used the review texts and incorporated the information on ratings.

4.3. Data Preprocessing

Particularly, the preprocessing task included converting all characters to lowercase, detecting and correcting spelling errors, deleting punctuation marks, stop words and numbers. The stop word frequency was set to 0.7. To prevent the long tail effect, the length and number of reviews covering 80% of users (items) were used in the experiment; every user (item) retained 12 comments, and every comment retained 200 words. The overall dataset were randomly divided into training set (80%), validation set (10%) and test set (10%), and the reviews in the test and validation set were removed.

5. RESULTS ANALYSIS

5.1. Prediction Performance

We use the mean square error (MSE) as the evaluation index of model performance. The prediction performance of our model and the comparison models are shown in Table 1, where it can be seen that the proposed model outperformed all the baseline models on all datasets, which proved the effectiveness of the proposed model. It can be concluded that, compared with the traditional NMF, LFM, SVD++, and HFT models, the DeepCoNN, D-Attn, and NARRE models and the proposed model that used the deep learning methods performed better, demonstrating that the neural networks could extract user preferences and item features contained in the review texts. Among the models considering reviews, the prediction errors of the HFT, DeepCoNN, and D-Attn models were larger than those of the NARRE model and the proposed model. This was expected since the HFT, DeepCoNN, and D-Attn model integrated all review texts into one document for further processing, ignoring differences between the reviews. Besides, they did not consider the rating matrix, so part of the information was lost.

5.2. Attention Mechanism Visualization

To demonstrate the word- and review-level attention mechanisms, the words with a high word-level attention score are highlighted in Table 2; the words with the highest scores

Table 2. Visualization of the attention mechanisms

$\alpha_L = 0.221$	A classically-styled and introverted album. The Memory of Trees is a masterpiece of subtlety. Many of the songs have an endearing shyness to them - soft piano and a lovely, quiet voice. For certain, The Memory Of Trees is melodic, romantic and sensuous.
$\alpha_L = 0.096$	I got this album on vinyl when it came out in the 70s. A lot of people didn't know about this cd when it came out. This was and still is a great buy. The singing is priceless.This is a must album with such iconic hits. Seems like they never get old.

are colored in gray, the middle ones in light gray, and the lowest ones are not colored. A group of reviews for an item were randomly selected. The review-level attention scores that indicate the importance of the review are presented on the left side in Table 2.

The results showed that the two attention mechanisms greatly improved model interpretability. In terms of the word-level attention, the trivial pronoun words were neglected by the model and words rich in the features of users or items were considered. For instance, in Table 2, the conjunction "and" in "For certain, The Memory Of Trees is melodic, romantic and sensuous." is not colored, which indicates the model considered it unimportant. On the contrary, "melodic", "romantic" and "sensuous" that describe the characteristics of the music are marked. With respect to the review-level attention mechanism, the scoring $alpha_L$ of the first review was higher than those of the other mechanisms, which proves the effectiveness of the proposed attention mechanisms because it is easy to find that the first review contains more instructive information about the item. In contrast, only general opinions and irrelevant information can be obtained from the second review.

6. CONCLUSIONS

This paper proposes a deep learning model DNRDR based on the rating matrix and review texts, which can learn long- and short-distance features of review texts. The proposed model can visualize the learned features by the word-level attention mechanism and can improve the interpretability of recommender systems by the review-level attention mechanism. The proposed model is verified by experiments and compared with a few state-of-the-art models. The experimental results show that the proposed method can reduce the error of prediction scores. Additional experiments are conducted to demonstrate that the proposed DNRDR can deal with the cold start problem more effectively than the related models by fusing the comment text with the ratings rather than using the rating matrix alone, hence could achieve better recommendation results on e-commerce websites.

References

[1] Silvana Aciar, Debbie Zhang, Simeon Simoff, and John Debenham. Informed recommender: Basing recommendations on consumer product reviews. *IEEE intelligent systems*, 22(3):p.39–47, 2007.

[2] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[3] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. E-commerce in your inbox: Product recommendations at scale. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1809–1818, 2015.

[4] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the Advances in neural information processing systemss, Vancouver, Canada*, pages 556–562, 2001.

[5] Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA*, pages 24–27, 2008.

[6] Soroush Ojagh, Mohammad Reza Malek, Sara Saeedi, and Steve Liang. A location-based orientation-aware recommender system using iot smart devices and social networks. *Future Generation Computer Systems*, 108:97–118, 2020.

[7] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 297–305, 2017.

[8] Huibing Zhang, Hao Zhong, Weihua Bai, and Fang Pan. Cross-platform rating prediction method based on review topic. *Future Generation Computer Systems*, 101:236–245, 2019.

[9] Lei Zheng, Vahid Noroozi, and Philip S Yu. 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*, pages 425–434, 2017.

[10] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1583–1592, 2018.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural computation*, volume 9, pages 1735–1780. MIT Press, 1997.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[14] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, pages 233–240, 2016.

[15] Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. Anr: Aspect-based neural recommender. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 147–156, 2018.

[16] Libing Wu, Cong Quan, Chenliang Li, Qian Wang, Bolong Zheng, and Xiangyang Luo. A context-aware user-item representation learning for item recommendation. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–29, 2019.

[17] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[18] Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. In *arXiv preprint arXiv:1808.08946*, 2018.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[20] Jianquan Li, Xiaokang Liu, Honghong Zhao, Ruifeng Xu, Min Yang, and Yaohong Jin. Bert-emd: Many-to-many layer mapping for bert compression with earth mover's distance. *arXiv preprint arXiv:2010.06133*, 2020.

[21] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv preprint arXiv:2002.02925*, 2020.

[22] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[23] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer, 2019.

[24] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(1):2493–2537, 2011.

[25] Kamal Al-Sabahi, Zhang Zuping, and Mohammed Nadher. A hierarchical structured self-attentive model for extractive document summarization (hssas). *IEEE Access*, PP(99):1–1, 2018.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[28] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

[29] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.