Learning to Rank Aspects and Opinions for Comparative Explanations

Trung-Hoang Le[1†] and Hady W. Lauw[1]

[1]School of Computing and Information Systems, Singapore Management University, 80 Stamford Rd, 178902, Singapore.

Contributing authors: thle.2017@smu.edu.sg; hadywlauw@smu.edu.sg;
†Corresponding author: Trung-Hoang Le thle.2017@smu.edu.sg

Abstract

Comparative recommendation explanations help to make sense of recommendations by comparing a recommended item along some aspects of interest with one or many items being considered. This work extends the notion of comparative explanations, by going beyond merely better/worse statements, to further incorporate aspect-level opinions for more informative comparisons. To enhance the quality of both the personalized recommendation and the explanation, we incorporate optimization objectives that preserve relative rankings of aspects and opinions, in addition to the classical rankings of overall preferences for items. We integrate the multiple ranking objectives and multi-tensor factorization together. Experiments on datasets of different domains validate the efficacy of our proposed framework in both recommendation and comparative explanation against comparable explainable recommendation baselines.

Keywords: Recommender Systems, Multi-Tensor Factorization, Comparative Explanation, Aspect-Level Opinion

1 Introduction

Recommendation is a prevalent feature in many online applications. While the initial aim is primarily on accurate recommendations, increasingly a greater emphasis is placed on providing some explanations to help users better comprehend the recommendations. Of the prior works that have been proposed to tackle recommendation explanations Zhang et al (2020), a vast majority attempt to explain a single item.

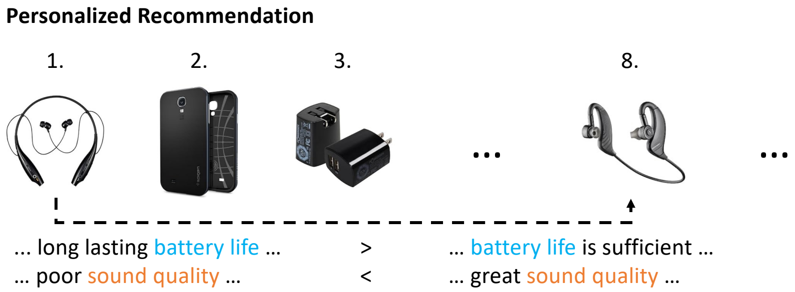


Fig. 1: A comparative explanation between two items

For instance, Wang et al (2018a) explains an aspect for a recommended item using opinion phrase, e.g., *"Its [aspect] is [opinion]"*. The idea is to contextualize the latent factors along known aspects towards recommendations Zhang et al (2014a) as well as aspect-level explanations associated with opinion phrases Wang et al (2018a) by reconstructing observed scores.

There is a nascent yet growing interest in comparative explanations, as exemplified by Le and Lauw (2021) that assesses a recommended item in comparison to another reference item (a previously adopted items), i.e., *"[recommended item] is better at [aspect] than [reference item], but worse at [another aspect]"*. However, some limitations remain. For one, relying on previous adoptions as reference is limiting as some users have little purchasing history. For another, the preoccupation on *better/worse*, but not *how*, yields inadequately informative explanations. We posit that users are inherently interested in perusing relative comparisons for choice-making purposes Park and Gretzel (2010). The pertinent comparison involving two items can be distilled into the trade-offs between their common aspects. The fine-grained scores for these aspects may be too abstract for direct consumption. Instead, by focusing on the opinion words describing the aspects in common between a recommended item and a reference item, we can gain further insights into the comparison.

In this work, we seek better comparative recommendation explanations via a learning to rank framework for comparative aspects and opinions, and extend the reference items to the recommendation list, facilitating comparisons among the recommended items. For illustration, we would compare a recommendation item with other items in the same ranking list (see Figure 1) via a template containing opinions for comparative explanation, as below:

> `[recommended item] [aspect], which is [opinion], is [better than/worse than/similar to] that of [reference item], which is [opinion].`

Contributions. First, as exemplified above, we introduce a novel form of comparative explainable recommendation using comparative aspects and opinions across pairs of items. Second, we develop Companion, a learning to rank framework based on multi-tensor factorization, which not only optimizes for personalized item recommendation objective, but also has a particular novelty in further incorporating ranking objectives for aspects and opinions. Third, we conduct comprehensive experiments against baselines that showcase the efficacies of our approach both in quantitative and qualitative terms.

Table 1: Main Notations

Symbol	Description
$\mathcal{U}, \mathcal{P}, \mathcal{A}, \mathcal{O}$	sets of users, products, aspects, and opinions
(a, w, ρ)	a tuple of aspect a, opinion phrase w, and sentiment polarity ρ
Q	aspect-level quality tensor (user-item-aspect quality scores)
X	user-item-aspect tensor with ratings
$\gamma_{(ij)kw}$	frequency of opinion w describing aspect k on review t_{ij} with positive sentiment ($\rho > 0$)
Y	positive user-item-aspect-opinion tensor
$\zeta_{(ij)kw}$	frequency of opinion w describing aspect k on review t_{ij} with negative sentiment ($\rho < 0$)
Z	negative user-item-aspect-opinion tensor
U, P, A, O	low rank matrices of user, item, aspect, and opinion
M, N, L, V	number of latent factors of U, P, A, O
$\mathcal{G}, \mathcal{E}, \mathcal{E}', \mathcal{F}, \mathcal{F}'$	core tensors in Tucker decomposition of $X, Y,$ and Z

2 Problem Formulation

Table 1 lists the main notations using in this work. $\mathcal{P}$ and $\mathcal{U}$ denote the universal sets of products and users respectively. Let $\mathcal{A}$ be the set of aspects and $\mathcal{O}$ be the set of opinions extracted from reviews. A user $i \in \mathcal{U}$ can rate a product $j \in \mathcal{P}$ a score $r_{ij} \in [1, \eta]$, associated with a review text $t_{ij} \in \mathcal{T}$ that can be represented by a list of tuples (a, w, ρ), where $a \in \mathcal{A}$, $w \in \mathcal{O}$, $\rho \in \{-1, 1\}$. The collection of tuples extracted from reviews make up a contextual lexicon $\mathcal{L}$. The formal problem is stated below:

Problem 1. *Given the sets of users $\mathcal{U}$, products $\mathcal{P}$, ratings $\mathcal{R}$, aspects $\mathcal{A}$, opinions $\mathcal{O}$, and contextual lexicon $\mathcal{L}$. We seek a model that can produce top-n personalized ranking list of items as well as a comparative explanation associated with each recommended item. The explanation will express the tradeoff in aspects between the recommended item and other items (i.e., from top-n personalized items), detailed with opinion phrases.*

3 Comparative Aspects and Opinions Ranking

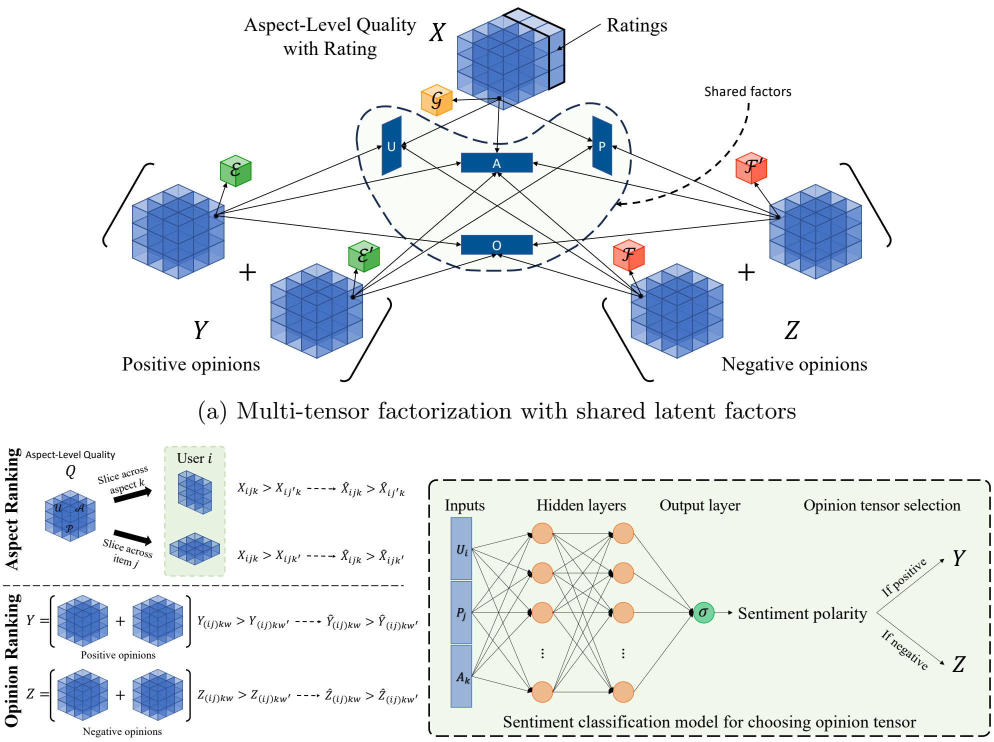
In this section, we propose Comparative Aspects and Opinions Ranking for Recommendation Explanations (Companion). Figure 2 illustrates the overall architecture of Companion. Below we describe its various objectives.

3.1 Bayesian Personalized Ranking on Explicit Ratings

We formulate the recommendation task as a ranking problem, rank a given list of items based on the user preference on those items. We address this by using Bayesian Personalized Ranking (BPR) principle Rendle et al (2009) on the explicit ratings in which an item should be ranked higher if it is preferred (higher rating score) by the user than the other items, minimizing the following loss:

$$L_{BPR} = -\sum_{i \in \mathcal{U}} \sum_{\{(i,j,j') | r_{ij} > r_{ij'}\}} \ln \sigma(\hat{r}_{ij} - \hat{r}_{ij'}) \tag{1}$$

where item j is preferred than item j' (either unobserved or $r_{ij} > r_{ij'}$) and $\sigma(\mathrm{x}) = \frac{1}{1+e^{-\mathrm{x}}}$ is Sigmoid function.



(a) Multi-tensor factorization with shared latent factors

(b) Maintaining relative ranking orders

(c) MLP classifier for sentiment classification

Fig. 2: COMPANION architecture

3.2 User Preference on Product Aspects

Presuming the explicit quality scores were given, e.g., they can be estimated based on sentiments extracted from textual reviews (i.e., via sentiment analysis Zhang et al (2014b)). We can use an aspect-level quality tensor $Q \in \mathbb{R}_+^{|\mathcal{U}|\times|\mathcal{P}|\times|\mathcal{A}|}$ to represent the preference of user on product aspects, where $Q_{ijk} \in Q$ is the quality score of aspect $k \in \mathcal{A}$ that user $i \in \mathcal{U}$ assigns to product $j \in \mathcal{P}$:

$$Q_{ijk} = \begin{cases} 0, \text{ if aspect } k \text{ is not mentioned in } t_{ij} \\ 1 + \frac{\eta-1}{1+e^{-s_{ijk}}}, \text{ otherwise} \end{cases} \tag{2}$$

where s_{ijk} is the total of the sentiment polarity scores (the aforementioned ρ) for aspect k extracted from user i's review of item j. The rating scores in the target domain fall within the range $r_{ij} \in [1, \eta]$, e.g., $[1, 5]$ on Amazon.com. This formula projects the sentiment scores to quality scores into the same scale as rating scores.

Following Wang et al (2018a), we append the rating r_{ij} to the tensor Q as an additional aspect to form tensor X including user-item-aspect interactions with ratings, i.e., $X_{ijk} \equiv Q_{ijk}$ if $k \in \mathcal{A} \Leftrightarrow k < |\mathcal{A}|$ and $X_{ij|\mathcal{A}|} \equiv r_{ij}$, which can be factorized by using Tucker decomposition Kolda and Bader (2009), i.e., minimizing the loss:

$$\begin{aligned} L_X &= ||\hat{X} - X||_F \\ \text{s.t. } \hat{X} &= \mathcal{G} \times_1 U \times_2 P \times_3 A = \textstyle\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \mathcal{G}_{mnl} U_{:m} \circ P_{:n} \circ A_{:l} \\ \mathcal{G} &\in \mathbb{R}_+^{M \times N \times L}, U \in \mathbb{R}_+^{|\mathcal{U}| \times M}, P \in \mathbb{R}_+^{|\mathcal{P}| \times N}, A \in \mathbb{R}_+^{(|\mathcal{A}|+1) \times L} \end{aligned} \tag{3}$$

where $\mathcal{G}$ is a core tensor in Tucker decomposition. U, P, A are the low rank factor matrices of user, item, and aspect respectively, with the corresponding number of latent dimensions are M, N, L. The operation $\mathcal{G} \times_n U$ denotes the n-mode product between tensor $\mathcal{G}$ and matrix U, i.e., Each mode-n fiber of $\mathcal{G}$ is multiplied by matrix U. Each estimation score for tensor X for a specific user i's preference on aspect k (including rating, $k = |\mathcal{A}|$) of item j is computed by:

$$\hat{X}_{ijk} = \textstyle\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \mathcal{G}_{mnl} U_{im} P_{jn} A_{kl} \tag{4}$$

Thus, the objective in Equation 1 could be rewritten as:

$$L_{BPR} = -\textstyle\sum_{i \in \mathcal{U}} \sum_{\{(i,j,j')|r_{ij} > r_{ij'}\}} \ln \sigma(\hat{X}_{ij|\mathcal{A}|} - \hat{X}_{ij'|\mathcal{A}|}) \tag{5}$$

To maintain the relative order of aspect quality scores for explanation in which preferred aspects should be ranked higher, we apply BPR principle on the aspect quality scores to rank aspects on the same item as well as across multiple items. Preference of a user for a single item varies depending on various aspects, preferred aspects should be ranked higher, minimizing:

$$L_p = -\textstyle\sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{P}} \sum_{\{(k,k') \in \mathcal{A} | X_{ijk} > X_{ijk'}\}} \ln \sigma(\hat{X}_{ijk} - \hat{X}_{ijk'}) \tag{6}$$

Also, aspect quality across items should reflect the user preference. If one aspect of an item is better than that of another item in observed data, the reconstructed scores after optimization should also reflect that relative order. We minimize:

$$L_a = -\textstyle\sum_{i \in \mathcal{U}} \sum_{k \in \mathcal{A}} \sum_{\{(j,j') \in \mathcal{P} | X_{ijk} > X_{ij'k}\}} \ln \sigma(\hat{X}_{ijk} - \hat{X}_{ij'k}) \tag{7}$$

Towards joint modeling, we use Lagrange multiplier to integrate the loss as follows:

$$L = L_X + \lambda_b L_{BPR} + \lambda_p L_p + \lambda_a L_a + \lambda_\Theta ||\Theta||_F^2 \tag{8}$$

where $\lambda_b, \lambda_p, \lambda_a$ control the contribution of L_{BPR}, L_p, and L_a respectively, λ_Θ controls $l2$-regularization on model parameters Θ.

Top-n Recommendation. Along with the overall rating scores, we should also recommend user with items that have high quality aspects. Similar to Le and Lauw (2021), we use the following formula to incorporate both ratings and top aspects into the final ranking score of a target user i on a given item j:

$$\text{RankingScore}_{ij} = \alpha \cdot \frac{\sum_{k \in \Lambda_{ij}} \hat{q}_{ijk}}{|\Lambda_{ij}|} + (1 - \alpha) \cdot \hat{r}_{ij}, \text{ s.t., } \; 0 \le \alpha \le 1 \tag{9}$$

where Λ_{ij} is the set of top aspects, α controls the trade-off between aspects and ratings.

3.3 Explaining Aspects using Opinion Phrases

The opinion that a user uses to describe an item's aspect reflects his or her preference (i.e., sentiment) on that aspect. The usage of opinion phrases is dependent on the user's preference on the target product Amarouche et al (2015), as well as the aspects it describes Feldman et al (2007). Intuitively, users may use different set of opinions to describe positive and negative sentiments. In addition, negation words may be used to describe a contrast sentiment. Without loss of generality, let $\gamma_{(ij)kw}$ (resp. $\zeta_{(ij)kw}$) be the frequency of user i uses opinion phrase w express aspect k on product j with positive sentiment polarity (resp. negative sentiment polarity). Two tensors $Y, Z \in \mathbb{R}^{(|\mathcal{U}|\times|\mathcal{P}|)\times|\mathcal{A}|\times|\mathcal{O}|}$ are constructed to model positive and negative opinions separately as:

$$
\begin{aligned}
Y_{(ij)kw} &= \begin{cases} 0, \text{ if opinion } w \text{ with } \rho = +1 \text{ is not in } t_{ij} \\ 1 + (\eta - 1)\left(\frac{2}{1+e^{-\gamma_{(ij)kw}}} - 1\right), \text{ otherwise} \end{cases} \\
Z_{(ij)kw} &= \begin{cases} 0, \text{ if opinion } w \text{ with } \rho = -1 \text{ is not in } t_{ij} \\ 1 + (\eta - 1)\left(\frac{2}{1+e^{-\zeta_{(ij)kw}}} - 1\right), \text{ otherwise} \end{cases}
\end{aligned} \tag{10}
$$

We also use Tucker decomposition to factorize tensor Y, minimizing:

$$
\begin{gathered}
L_Y = ||\hat{Y} - Y||_F \\
\text{s.t. } \hat{Y}_{(ij)kw} = \left(\sum_{m=0}^{M-1} \mathcal{E}_{ikw} U_{im} + \sum_{n=0}^{N-1} \mathcal{E}'_{jkw} P_{in}\right) \sum_{l=0}^{L-1} \sum_{v=0}^{V-1} A_{kl} O_{wv} \\
\mathcal{E} \in \mathbb{R}_+^{M\times L\times V}, \mathcal{E}' \in \mathbb{R}_+^{N\times L\times V}, O \in \mathbb{R}_+^{|\mathcal{O}|\times V}
\end{gathered} \tag{11}
$$

and similarly for tensor Z, we minimize:

$$
\begin{gathered}
L_Z = ||\hat{Z} - Z||_F \\
\text{s.t. } \hat{Z}_{(ij)kw} = \left(\sum_{m=0}^{M-1} \mathcal{F}_{ikw} U_{im} + \sum_{n=0}^{N-1} \mathcal{F}'_{jkw} P_{in}\right) \sum_{l=0}^{L-1} \sum_{v=0}^{V-1} A_{kl} O_{wv} \\
\mathcal{F} \in \mathbb{R}_+^{M\times L\times V}, \mathcal{F}' \in \mathbb{R}_+^{N\times L\times V}, O \in \mathbb{R}_+^{|\mathcal{O}|\times V}
\end{gathered} \tag{12}
$$

where $\mathcal{E}, \mathcal{E}', \mathcal{F}, \mathcal{F}'$ are core tensors. O is the low rank factor matrix of opinion that is shared across the decomposition of both positive and negative opinion tensors.

The decomposition in both Equation 11 and Equation 12 focus on minimizing element-wise reconstruction error. We further formulate the selection of opinion phrases for explanation as a ranking problem, applying BPR principle, which preferred opinion phrases should be ranked higher, minimizing the loss:

$$
L_y = -\sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{P}} \sum_{k\in\mathcal{A}} \sum_{\{(w,w')\in\mathcal{O}|Y_{(ij)kw} > Y_{(ij)jw'}\}} \ln \sigma(\hat{Y}_{(ij)kw} - \hat{Y}_{(ij)kw'}) \tag{13}
$$

and similarly, for negative tensor Z, we minimize:

$$
L_z = -\sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{P}} \sum_{k\in\mathcal{A}} \sum_{\{(w,w')\in\mathcal{O}|Z_{(ij)kw} > Z_{(ij)jw'}\}} \ln \sigma(\hat{Z}_{(ij)kw} - \hat{Z}_{(ij)kw'}) \tag{14}
$$

Overall Objective. We jointly model the recommendation and explanation objectives by minimizing,

$$L = L_X + \lambda_Y L_Y + \lambda_Z L_Z + \lambda_b L_{BPR} + \lambda_p L_p + \lambda_a L_a + \lambda_y L_y + \lambda_z L_z + \lambda_\Theta ||\Theta||_F^2 \quad (15)$$

where Θ is the model parameters, λ_Θ controls the strength of $l2$-regularization, λ_Y and λ_Z control the contribution of tensor reconstruction losses L_Y and L_Z respectively, λ_y and λ_z control the contribution of positive and negative opinions ranking losses L_y and L_z respectively.

3.4 Aspect-Sentiment Polarity Prediction

We use opinion phrases to explain an aspect for a target user on a given item. The overall objective in Equation 15 helps us to jointly model both positive and negative opinion ranking altogether. However, when providing explanation for a certain aspect, we should be able to identify what is the sentiment polarity for a specific aspect k of a user i on target item j for selecting the appropriate positive opinion from tensor Y or negative opinion from tensor Z. This could be formulated as a binary classification task. We take the user-item-aspect triple (i, j, k) as input and seek for a model that can classify sentiment polarity of that triple to either positive or negative. Where $s_{ijk} > 0$ is positive label and $s_{ijk} < 0$ is negative label. To address this, we leverage the learned representation $U_{i:}, P_{j:}, A_{k:}$ of user i, item j, and aspect k as inputs and use a Multi-Layer Perceptron model as binary classification model. Specifically,

$$h_1 = \text{ReLU}(W_1[U_{i:}; P_{j:}; A_{k:}] + b_1) \quad (16a)$$

$$h_2 = \text{ReLU}(W_2 h_1 + b_2) \quad (16b)$$

$$\dots$$

$$\hat{s} = \sigma(W_k h_{k-1} + b_k) \quad (16c)$$

where $\text{ReLU}(x) = \max(x, 0)$, W_i is the learning parameters at i-th layer of the MLP. The last layer (Equation 16c) produces a single score for binary classification, learning by minimizing Binary Cross Entropy loss Good (1963),

$$L_s = -\frac{1}{|\Omega|} \sum_{\omega \in \Omega} s_\omega \log(\hat{s}_\omega) + (1 - s_\omega) \log(1 - \hat{s}_\omega) \quad (17)$$

where $s_\omega \in \{0, 1\}$ is the truth label of the sample ω, and $\hat{s}_\omega$ is the probability of positive class, Ω is training instances.

3.5 Explanation Generation

A pertinent issue is how to choose items for comparative explanation. One could be items from user's purchased history Le and Lauw (2021) as reference for comparison, which could be limited if users had purchased only a few. In this work, we explore a novel approach for selecting comparative items, which retrieve directly from the list of top-n personalized item recommendation $\tau \subset \mathcal{P}$. We select the first item $j' \in \tau$ in the list of recommendation items that is not dominated by the recommended item j, i.e., $\exists k, k' \in \mathcal{A}$:$\hat{x}_{ijk} > \hat{x}_{ij'k} \wedge \hat{x}_{ijk'} < \hat{x}_{ij'k'}$, where item j is better than item j' at aspect k but worse at aspect k'. This explanation will be more detailed by accompanying opinion phrases describing how a certain aspect on a given item is. The opinion phrases are selected using the estimated scores of positive tensor $\hat{Y}$ or negative tensor $\hat{Z}$ depending on the sentiment polarity classified by the classifier introduced in section 3.4. Moreover, the selected opinion phrases of two items may be identical, making comparison simply from aspect-level quality scores insufficient, which can be fixed by post-correction using equivalent comparison such as "similar to". Furthermore, we also use negation word (i.e., *not*) to switch the sentiment of an opinion phrase in case it reflects the opposite sentiment, e.g., opinion phrase "*bad*" associated with positive sentiment will be updated to "*not bad*" in the explanation template.

Guaranteeing the selected items are truly comparable is beyond the scope of this work. In the case item j' cannot be found in the top-n recommendations, we can always expand the recommendation list or fall back to an evaluative explanation for such scenarios.

3.6 Evaluation

First, we note that recommendation is essentially a ranking problem, which needs to determine the relative relevance among a set of candidates to indicate which item is more relevant to a user, and this can be effectively evaluated via ranking metrics such as Area Under the ROC Curve (AUC) and Normalized Discounted Cumulative Gain (NDCG), that have been used in many literatures Rendle et al (2009); Wang et al (2018a); Le and Lauw (2021). For these metrics, higher values (maximum value is 1) reflect better ranking performance. A challenge arises on how to evaluate a comparative recommendation explanation. We formulate a comparison from one to another item via their relative aspect qualities. Even if the ground truth aspect qualities can be extracted from the held-out reviews, it is often the case that the qualities of certain aspects on the reference item are not available due to sparsity. Thus, we can instead evaluate the comparative explanations via ranking metrics (homogeneously as recommendation objective) by measuring ranking performance of the aspect recovery as well as item retrieval from aspect. Furthermore, retrieving opinions for explaining item's aspect is also formulated as a ranking problem, and thus similar ranking evaluation metrics can be applied. In summary, we use AUC and NDCG to evaluate both recommendations and comparative explanations. Specifically, we evaluate the ranking performance of recommendations, aspect recovery, item retrieval from aspect, and opinion retrieval for explaining aspect.

4 Related Work

Here we survey works that deal with explainable recommendation and comparative explainable recommendation.

Explainable Recommendation. There are several explainable recommendation methods relying on aspect-level sentiment. Zhang et al (2014a); Bauman et al (2017) use matrix factorization approaches. He et al (2015) enhances explainable recommendation using tripartite graphs and Gao et al (2019) uses trees. Chen et al (2016); Wang et al (2018a); Le and Lauw (2021) are tensor factorization approaches. Zhang et al (2014a); He et al (2015); Bauman et al (2017); Chen et al (2016); Gao et al (2019) focus on only modeling aspects. Wang et al (2018a) extends to model opinions. These methods are distinct from our proposed method in the way they explain a single product on its own, without any comparative explanations.

Beside aspect-level sentiment, there are alternatives for recommendation explanation, e.g., similar content Herlocker et al (2004), topics McAuley and Leskovec (2013); Wu and Ester (2015); Tan et al (2016); Chen et al (2019b), rules Wang et al (2018b); Ma et al (2019); Tao et al (2019); Xian et al (2019); Chen et al (2021); Geng et al (2022); Balloccu et al (2023), trusted social relation Ren et al (2017), visual attention Chen et al (2019a), helpful reviews Chen et al (2018), helpful questions Le and Lauw (2022), ranked attributes/sentences extracted from reviews Li et al (2023), generated reviews Li et al (2017); Lu et al (2018); Chen et al (2019c); Truong and Lauw (2019); Li et al (2021); Hada et al (2021); Xie et al (2023), etc. Comparisons with these different forms of explanation are beyond the scope of this study.

Comparative Explainable Recommendation. Comparative explainable recommendation, assessing a recommended item in comparison to one or multiple items, attracts more attention recently. Chen and Wang (2017) studies tradeoff-oriented explanation comparing multiple products using an interface. Moraes et al (2020) highlights the important of attributes in product quality comparisons. Our method is among those with comparative explanations using template. Le and Lauw (2021) produces comparative explanation using a template comparing one item to a previously adopted item as reference.

Additionally, there are text generation approaches for comparative explanations, which are orthogonal to our work. Yang et al (2022) develops a text generation model using extracted candidate sentences from recommended item's profile as a prototype and rewrites that sentence to optimize the desired quality metric for comparative explanation. Vedula et al (2023) makes use of multi-decoder for generating product-level or attribute-level comparison texts. Echterhoff et al (2023) generates comparative sentences to contrast one item from another by highlighting important aspects.

5 Experiment

Datasets. We verify the efficacies of our models in different domains, we select products of *Baby (Baby)* category and *Cellphones and Accessories (Cellphone)* category from Amazon Review Dataset[1] He and McAuley (2016), and businesses from Yelp Dataset Challenge[2] of *Cambridge city - Massachusetts acquired in 2021 (Cambridge)* and *Philadelphia city - Pennsylvania acquired in 2022 (Philadelphia)*. We use 5-core data, each user and item have at least 5 ratings. We first extract aspect-level sentiment using Sentires[3] Zhang et al (2014b), an open sourced toolkit for Phrase-level Sentiment Analysis. Using textblob[4] to estimate the sentiment of each opinion phrases, we eliminate all neutral sentiment opinion phrases. It is possible to apply other approaches to extract aspect sentiment from product reviews, which

[1] http://jmcauley.ucsd.edu/data/amazon/
[2] https://www.yelp.com/dataset
[3] https://github.com/evison/Sentires
[4] https://textblob.readthedocs.io

Table 2: Datasets

Dataset	#User	#Item	#Rating	#Aspect	#Opinion	#Aspect-Opinion-Sentiment
Baby	19,445	7,050	160,792	1,634	439	108,154
Cellphone	27,879	10,429	194,439	1,727	511	149,631
Cambridge	8,196	1,757	100,741	1,909	561	88,741
Philadelphia	35,865	10,940	562,150	5,463	828	708,065

Table 3: Overview of selected baselines

Method	Year	Tensor-based	Aspect	Opinion	Aspect Ranking	Opinion Ranking
EFM	2014	✗	✔	✗	✗	✗
TriRank	2015	✗	✔	✗	✔	✗
LRPPM	2016	✔	✔	✗	✔	✗
MTER	2018	✔	✔	✔	✗	✗
ComparER	2022	✔	✔	✔	✔	✗
Companion	2024	✔	✔	✔	✔	✔

in any case, we consider these as given. Table 2 summarizes basic statistics of the datasets. #Aspect-Opinion-Sentiment column shows the total number of aspect-opinion-sentiment tuples extracted from reviews. We apply user-based splitting to split interactions into training, validation and testing sets. Specifically, the last two items of each user's purchased sequence (sort in chronological order) belong to either validation or testing sets randomly. We exclude unknown items, aspects, and opinions, i.e., those are not appear in training set. Code and data are available for reproducibility at https://github.com/PreferredAI/Companion.

Baselines. We evaluate our proposed Companion against several *explainable recommendation baselines*, chronologically ordered by the publication year (see Table 3).

EFM Zhang et al (2014a): A non-negative multi-matrix factorization approach for aspect-level explanation that optimizes for rating prediction, modeling objective aspect-level quality on item side.

TriRank He et al (2015): A personalized ranking algorithm on tripartite graphs (user-item-aspect relations).

LRPPM Chen et al (2016): An aspect ranking model that also based on BPR principle, using Pairwise Interaction Tensor Factorization (PITF) Rendle and Schmidt-Thieme (2010), optimized for rating prediction.

MTER Wang et al (2018a): A tensor factorization approach for explainable recommendation that uses Tucker decomposition, learning to model user, item, aspect, and opinion jointly.

ComparER Le and Lauw (2021): A comparative explainable recommendation model, also a tensor factorization model which uses Tucker decomposition, that leverages the *skyline* aspects of items in user purchased sequences to model the comparative constraints on product aspects.

Except for Companion, none of the baselines incorporate opinion ranking. TriRank and LRPPM have aspect ranking without opinion modeling. MTER has opinion modeling but no aspect and opinion ranking. ComparER has aspect ranking and opinion modeling without opinion ranking. EFM models aspects without opinions.

Evaluation Metrics. We employ multiple standard ranking metrics, such as Area Under the ROC Curve (AUC) and Normalized Discounted Cumulative Gain at 10 (NDCG@10) and 50 (NDCG@50). For these, higher values mean better performance.

Learning Details. As suggested in Le and Lauw (2021), using a large dimension for latent factors in tensor factorization models such as MTER and ComparER requires much more time for training, and these models often achieve better AUC using a smaller dimensionality for number of latent factors such as 8 (searched in $\{8, 16, 32, 64, 128\}$). We also observed similar trends on our datasets for LRPPM, MTER, ComparER, and our proposed COMPANION. Thus, we set the latent factors to be 8, $\lambda_b = 10$, $\lambda_Y = \lambda_Z = 1$, and the regularization $\lambda_\Theta = 0.1$. For simplicity, we set $\lambda = \lambda_p = \lambda_a = \lambda_y = \lambda_z$, tune λ in the candidate set of $\{0.01, 0.1, 1, 10, 100, 1000\}$, and achieve best performance with $\lambda = 10$ across datasets, which is consistent with $\lambda_b = 10$ controlling ranking objective on explicit ratings. Using a smaller λ may overlook ranking objectives while using a larger λ may increase their error rates on observed ratings which diversify the learned factors from the reconstructing the observed scores. For sentiment classification, we use a 3-layer MLP (section 3.4) with the number of neural units in hidden layers to be $\{16, 8, 1\}$, using regularizer $l2 = 0.001$ and dropout layers with ratio 0.5 to mitigate overfitting. EFM achieves best performance with latent dimensions of 128 (searched in $\{8, 16, 32, 64, 128\}$), consistent with Le and Lauw (2021). For each method, the setting with the best AUC on validation set is selected.

5.1 Quantitative Results

In this section, we look into the performance in quantitative aspects along several dimensions. We measure ranking performance on personalized recommendation to evaluate the effectiveness of COMPANION against comparable baselines. Also, we measure ranking performance of aspect recovery, item retrieval from aspect, and opinion ranking to quantitatively evaluate the performance of comparative explanation.

5.1.1 Personalized Item Ranking Performance

We first investigate whether adding ranking objectives on aspects and opinions enhances the personalized item ranking performance. Table 4 shows the results against baselines. Overall, the proposed COMPANION achieves the best performance on almost metrics across datasets in a statistically significant manner, the only exception is NDCG@10 on Baby dataset. Comparisons between methods are tested with one-tailed paired-sample Student's t-test at 0.05 level. Although ComparER achieves a better results than MTER as reported in Le and Lauw (2021) due to a denser dataset setting (each user has at least 10 ratings), sparsity severely hurts ComparER's effectiveness, explaining its lower ranking performance. Another note is that NDCG values are small due to the significant higher number of items on each dataset. It is often the case that other papers have higher NDCG due to a less reproducible evaluation approach[5] of only ranking a small sample set of items to speed up their evaluation.

[5]We rank the whole set of items to evaluate ranking performance, which is easier to reproduce but expensive for large number of items. To speed up the evaluation process, other papers use negative sampling to sample a subset of items that need to be provided for reproducibility.

Table 4: Personalized item ranking performance

	Model	AUC	NDCG@10	NDCG@50
Baby	EFM	0.6572	0.0063	0.0129
	TriRank	0.6918	0.0046	0.0104
	LRPPM	0.6833	**0.0084**	0.0156
	MTER	<u>0.7137</u>	0.0080	<u>0.0179</u>
	ComparER	0.7007	0.0065	0.0149
	Companion	**0.7194**§	<u>0.0083</u>	**0.0184**§
	Improvement %	0.80%	-1.20%	2.79%
Cellphone	EFM	0.6594	0.0085	0.0168
	TriRank	0.7093	0.0085	0.0171
	LRPPM	0.6758	0.0097	0.0210
	MTER	<u>0.7627</u>	<u>0.0154</u>	<u>0.0293</u>
	ComparER	0.7552	0.0128	0.0250
	Companion	**0.7752**§	**0.0167**§	**0.0316**§
	Improvement %	1.64%	8.44%	7.85%
Cambridge	EFM	0.7609	0.0266	0.0550
	TriRank	0.8024	0.0277	0.0609
	LRPPM	0.7917	0.0302	0.0592
	MTER	<u>0.8198</u>	<u>0.0329</u>	<u>0.0694</u>
	ComparER	0.8017	0.0260	0.0555
	Companion	**0.8274**§	**0.0361**§	**0.0737**§
	Improvement %	0.93%	9.73%	6.20%
Philadelphia	EFM	0.7661	0.0117	0.0222
	TriRank	0.7694	0.0089	0.0200
	LRPPM	0.7868	0.0173	0.0292
	MTER	<u>0.8464</u>	<u>0.0197</u>	<u>0.0367</u>
	ComparER	0.8412	0.0145	0.0287
	Companion	**0.8576**§	**0.0220**§	**0.0403**§
	Improvement %	1.32%	11.68%	9.81%

§p-value < 0.05. Best values are **bolded**. Second best values are <u>underlined</u>.

Table 5: Ablation study on personalized item ranking performance (AUC)

Model Variants	Baby	Cellphone	Cambridge	Philadelphia
Companion	**0.7194**	<u>0.7752</u>	**0.8274**	**0.8576**
Companion w/o personalized ranking ($\lambda_b = 0$)	0.6477	0.6915	0.7846	0.7894
Companion w/o opinion rankings ($\lambda_y = \lambda_z = 0$)	0.7182	0.7749	0.8245	<u>0.8544</u>
Companion w/o aspect ranking on item ($\lambda_p = 0$)	0.7173	0.7726	<u>0.8265</u>	0.8504
Companion w/o item ranking by aspect ($\lambda_a = 0$)	<u>0.7184</u>	<u>0.7752</u>	0.8198	0.8492
Companion w/o aspect rankings ($\lambda_p = \lambda_a = 0$)	0.7167	**0.7757**	0.8234	0.8499

Best values are **bolded**. Second best values are <u>underlined</u>.

Ablation Study. To further evaluate the effectiveness of each proposed component in Companion, we systematically remove key components in Equation 15: including personalized ranking ($\lambda_b = 0$), opinion rankings ($\lambda_y = \lambda_z = 0$), aspect rankings[6] ($\lambda_p = \lambda_a = 0$), aspect ranking on item ($\lambda_p = 0$), and item ranking by aspect ($\lambda_a = 0$). Results in Table 5 clearly show that removing the personalized ranking component results in a noticeable performance decline across datasets, highlighting

[6]Opinion rankings are implicitly discarded ($\lambda_y = \lambda_z = 0$) when either $\lambda_p = 0$ or $\lambda_a = 0$.

the need for personalized ranking objective. Although the aspect rankings and opinion rankings have a minor contribution to the overall performance, they enhance the overall ranking performance (the only exception is on Cellphone data that removing aspect rankings results in a slightly higher AUC, COMPANION is a runner up with a small margin), indicating additional ranking constraints on aspects and opinions helps address the problem of data sparsity. Excluding either aspect ranking on item ($\lambda_p = 0$) or item ranking by aspect ($\lambda_a = 0$) leads in reduced performance, eliminating both components ($\lambda_p = \lambda_a = 0$) may cause further performance degradation.

5.1.2 Aspect Recovery for Item Explanation

Here we assess whether incorporating aspect ranking objectives would help in selecting aspects for the explanation. Retrieving aspects from reviews in held-out test data as ground truth for evaluation, we rank aspects based on the scores of the reconstruction tensor $\hat{X}$ and use AUC as evaluation metric. Table 6 shows methods that incorporate ranking objective in their optimization objectives are better in aspect ranking, as expected. COMPANION is competitive throughout, achieving best performance on Baby and Cellphone datasets and is the runner up on Cambridge and Philadelphia datasets with small margins. Among those that do not optimize for aspect ranking objective, EFM, a matrix factorization model, achieves better aspect ranking performance than MTER and ComparER, tensor factorization models, indicating aspect recovery is more difficult due to sparsity of tensor modeling approaches. By incorporating aspect ranking objective, we can overcome the sparsity issue for better aspect ranking performance.

Table 6: Aspect ranking performance (AUC)

Model	Baby	Cellphone	Cambridge	Philadelphia
EFM[†]	0.8868	0.8967	0.8975	0.9263
TriRank	0.9539	<u>0.9707</u>	0.9526	0.9712
LRPPM	<u>0.9546</u>	0.9659	**0.9631**	**0.9752**
MTER[†]	0.5886	0.6097	0.6174	0.5877
ComparER[†]	0.7329	0.5934	0.6744	0.5956
COMPANION	**0.9606**	**0.9740**	<u>0.9559</u>	<u>0.9717</u>

[†]Method without aspect ranking objective. Best values are **bolded**. Second best values are <u>underlined</u>.

5.1.3 Item Retrieval from Aspect as Recommendation

In our objective, we also optimize for ranking items based on aspect quality scores (see Equation 7). Here we define another item recommendation problem as follows: Given a target user with a specified aspect, we seek to retrieve a personalized ranking list of products as recommendation. To evaluate, we extract aspects from reviews of test data as inputs. From the reconstruction aspect quality scores, we rank items based on the given aspect. Among the selected baselines, TriRank is excluded because it cannot perform ranking items from a given aspect. As reported in Table 7, COMPANION is competitive throughout, the second highest AUC on Baby and Cambridge datasets,

Table 7: Item ranking based on aspect performance (AUC)

Model	Baby	Cellphone	Cambridge	Philadelphia
EFM	0.6983	0.6986	0.7754	0.7858
LRPPM	0.6667	0.6728	0.6386	0.6000
MTER	**0.7249**	<u>0.7669</u>	**0.8615**	<u>0.8397</u>
ComparER	0.6209	0.6380	0.5997	0.5796
Companion	<u>0.7200</u>	**0.7794**	<u>0.8582</u>	**0.8750**

Best values are **bolded**. Second best values are <u>underlined</u>.

Table 8: Sentiment polarity classification accuracy

	Baby	Cellphone	Cambridge	Philadelphia
Accuracy	90.94%	88.11%	86.72%	88.84%

and the best AUC on Cellphone and Philadelphia. Notably, the improvement upon the second best model MTER on Philadelphia data is 4.2%. Which shows the advantage of using Companion to pick comparative items for comparative explanations.

5.1.4 Opinion Ranking for Aspect Explanation

Table 8 reports the performance of our proposed binary sentiment classification in section 3.4. The proposed MLP classifier achieves high accuracy ($> 86\%$) across datasets. Among selected baselines, only MTER and ComparER are able to rank opinions as they also model opinions. After determining sentiment for an aspect of a given user on a target item, we rank the opinion phrases based on their estimated scores on positive opinion tensor $\hat{Y}$ (for positive sentiment) and negative opinion tensor $\hat{Z}$ (for negative sentiment). Table 9 shows Companion consistently rank opinion phrases better than baselines, with much higher AUC scores as the baselines do not optimize for opinion ranking objective. There are two main reasons affecting the low AUC on negative opinion: (1) imbalanced distribution of positive opinions and negative opinions samples; (2) both MTER and ComparER only model positive opinions. If we allow MTER and ComparER model negative opinion separately, the performance of ranking negative opinions can be enhanced. However, they do not model opinion rankings in their objectives, which limits their performance. This further emphasizes the need to model opinion ranking objectives for both positive and negative opinions.

5.2 Qualitative Study

To assess the quality of the generated comparative explanation, we present a few examples in Figure 3 as case studies and further discuss a user study.

5.2.1 Case Studies

For case study, we compare top-1 recommended item with another reference item from the same ranking list using the method described in section 3.5. Using aspects and opinions produced by the underlying Companion for parity, we apply the template

Table 9: Opinion ranking performance (AUC)

Dataset	Model	Opinion		
		Positive	Negative	All
Baby	MTER	0.6060	0.5078	0.5965
	ComparER	0.5876	0.4954	0.5786
	COMPANION	**0.9792**	**0.9446**	**0.9758**
Cellphone	MTER	0.5984	0.5381	0.5904
	ComparER	0.5833	0.4942	0.5716
	COMPANION	**0.9705**	**0.9427**	**0.9669**
Cambridge	MTER	0.6901	0.5743	0.6741
	ComparER	0.6081	0.5231	0.5963
	COMPANION	**0.9751**	**0.9364**	**0.9697**
Philadelphia	MTER	0.7442	0.6024	0.7271
	ComparER	0.6958	0.5929	0.6834
	COMPANION	**0.9778**	**0.9457**	**0.9739**

Best values are **bolded**.

Fig. 3: Comparative explanations by COMPANION on various templates

proposed by the baselines MTER Wang et al (2018a) (no comparison) and ComparER Le and Lauw (2021) (comparison with aspect-only). Examples in Figure 3 show that explanations from baselines are lacking details while that of COMPANION is more detailed with comparative aspects and opinions. MTER explains given aspects using opinions without comparison and does not explicitly state the comparisons, leaving the comparison to the end users interpretation. ComparER compares two items by using simply *better/worse* for comparison. Comparative explanations from COMPANION provide both comparison and detail using opinions describing certain aspects.

5.2.2 User Study

Our focus in this study is to validate the effectiveness of the template format for comparative explanations. We compare a template with aspects and opinions but *no comparison* Wang et al (2018a), a template has *comparison with aspect-only* Le and Lauw (2021), and a template has *comparison with aspects and opinions* (this work). We apply the same set of aspects and opinions produced by COMPANION for parity, conducting a user study comprising 20 examples selected across the four datasets (5 examples from each dataset), asking one question *"How does the explanation help you to compare the recommendation items?"* to be answered by rating the explanation from 1 (poor) to 5 (excellent). Using 3 independent surveys in which each contains 20 explanations using different templates (selected randomly and presented blindly), involving a total of 9 participants who are not the authors, an explanation is seen by 3 different people.

Table 10: User study on comparative explanations

Explanation template	Avg. score	Standard deviation	Krippendorff's $\alpha^\dagger$
No comparison	2.617	0.958	0.002
Comparison w/ aspect-only	**3.933**	1.103	-0.028
Comparison w/ aspects & opinions	**3.933**	**0.899**	**0.238**

†Higher value means higher agreement. Best values are **bolded**.

Table 10 reports the summary results. The template having no comparison explicitly stated within the explanation received low ratings. Both templates with comparison achieve the highest average scores. Furthermore, the template having comparison with aspects and opinions achieves the lowest standard deviation, implying more consistent ratings. This is validated further by measuring the Krippendorff's Alpha-Reliability coefficient Krippendorff (2011), assessing the agreement among different annotators. The template having comparison with aspects and opinions achieve higher $\alpha = 0.238$, indicating some level of reliability, while that of template having comparison with aspect-only α is -0.028, showing some disagreements among annotators. To make sense of these disagreements, we looked into user feedback (which was optional) on how certain ratings were assigned:

- *It is not clear how similar items being compared on this basis.* (assigned 1 score)
- *The sentence is simple but the aspect words inside brackets are not quite clear.* (assigned 4 score)
- *Some aspects are better understood for comparisons, while others are not as useful and subjective for making comparisons.* (assigned 2 score)

This further demonstrates the desire for detailed explanations in comparative explanations and COMPANION works toward this goal.

6 Discussion

We aware that the comparative template explanations are repetitive and limit in their expressiveness. Here we discuss an option to enhance the explanation further. In recent years, the advancement of Large Language Models (LLMs) demonstrate that LLMs are very good in generating text. However, LLMs are known to be hallucinate. Since our proposed Companion are transparent in term of comparative explanation, we ask ChatGPT[7], a pretrained Large Language Model by OpenAI, to re-write the explanation text to make it reads more naturally given the comparisons achieved by Companion. For example, we can use the following prompt:

> Given 2 items "[Title of item 1]" and "[Title of item 2]". Write a short comparison between the two items if you know that the first item is better than the second item in "[better aspect]" aspect and worse in "[worse aspect]" aspect.

Figure 4 illustrates an example[8] of generating comparison text to compare two items of *Plantronics* brand. The generated text reflects well the provided comparisons. Although LLMs are quite powerful, to evaluate the effectiveness of generated comparative explanations thoroughly, we also need a reliable source of ground truth for evaluation, which is not yet available during the time we conduct the study. Without going beyond the scope of this work, this is considered as a new line of research to explore in the future.

User: Given 2 items "Plantronics MX150 Flexible Boom Headset with 2.5mm Plug" and "Plantronics Voyager 510 Bluetooth Headset". Write a short comparison between the two items if you know that the first item is better than the second item in "fits" aspect and worse in "microphone" aspect.

ChatGPT: When comparing the Plantronics MX150 Flexible Boom Headset with the Plantronics Voyager 510 Bluetooth Headset, the MX150 excels in the aspect of fit, offering a comfortable and secure feel. Its flexible boom design ensures a snug fit for most users. However, in terms of microphone performance, the MX150 falls short compared to the Voyager 510. The Voyager 510 boasts superior microphone quality, delivering clear and crisp audio during calls. Therefore, if fit is a priority, the MX150 is the better choice, but for superior microphone performance, the Voyager 510 stands out.

Fig. 4: Comparisons by ChatGPT-3.5 using Companion's output

7 Conclusion

We proposed a novel approach for comparative explanations. Leveraging learning to rank framework into multi-tensor factorization model, ranking items for recommendation along with ranking aspects and opinions for explanation, Companion performs well on personalized recommendation task against selected explainable recommendation baselines and produces good comparative explanations associating comparative aspects and opinions.

References

Amarouche K, Benbrahim H, Kassou I (2015) Product opinion mining for competitive intelligence. Procedia Computer Science 73:358–365

[7] https://chat.openai.com
[8] More examples can be found at https://chat.openai.com/share/a11a41f0-26c1-4775-8928-a0c17e6c4423

Balloccu G, Boratto L, Fenu G, et al (2023) Reinforcement recommendation reasoning through knowledge graphs for explanation path quality. KBS 260:110098

Bauman K, Liu B, Tuzhilin A (2017) Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In: KDD, KDD '17

Chen C, Zhang M, Liu Y, et al (2018) Neural attentional rating regression with review-level explanations. In: Proceedings of the 2018 World Wide Web Conference, WWW '18, p 1583–1592, https://doi.org/10.1145/3178876.3186070

Chen H, Li Y, Sun X, et al (2021) Temporal meta-path guided explainable recommendation. In: WSDM. ACM, WSDM '21, p 1056–1064

Chen L, Wang F (2017) Explaining recommendations based on feature sentiments in product reviews. In: IUI, IUI '17, p 17–28, https://doi.org/10.1145/3025171.3025173

Chen X, Qin Z, Zhang Y, et al (2016) Learning to rank features for recommendation over multiple categories. In: SIGIR. ACM, SIGIR '16, p 305–314

Chen X, Chen H, Xu H, et al (2019a) Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In: SIGIR, SIGIR'19, https://doi.org/10.1145/3331184.3331254

Chen X, Zhang Y, Qin Z (2019b) Dynamic explainable recommendation based on neural attentive models. In: AAAI, pp 53–60

Chen Z, Wang X, Xie X, et al (2019c) Co-attentive multi-task learning for explainable recommendation. In: IJCAI, IJCAI '19, pp 2137–2143

Echterhoff J, Yan A, McAuley J (2023) Comparing apples to apples: Generating aspect-aware comparative sentences from user review. arXiv preprint arXiv:230703691

Feldman R, Fresko M, Goldenberg J, et al (2007) Extracting product comparisons from discussion boards. In: ICDM, IEEE, ICDM'07, pp 469–474

Gao J, Wang X, Wang Y, et al (2019) Explainable recommendation through attentive multi-view learning. In: AAAI, pp 3622–3629

Geng S, Fu Z, Tan J, et al (2022) Path language modeling over knowledge graphsfor explainable recommendation. In: WWW. ACM, New York, NY, USA, WWW '22, p 946–955, https://doi.org/10.1145/3485447.3511937

Good IJ (1963) Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. Ann Math Statist 34(3):911–934

Hada DV, M. V, Shevade SK (2021) Rexplug: Explainable recommendation using plug-and-play language model. In: SIGIR. ACM, New York, NY, USA, SIGIR '21, p 81–91, https://doi.org/10.1145/3404835.3462939

He R, McAuley J (2016) Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: WWW, WWW '16, p 507–517, https://doi.org/10.1145/2872427.2883037

He X, Chen T, Kan MY, et al (2015) Trirank: Review-aware explainable recommendation by modeling aspects. In: CIKM, CIKM '15, p 1661–1670

Herlocker JL, Konstan JA, Terveen LG, et al (2004) Evaluating collaborative filtering recommender systems. TOIS 22(1):5–53

Kolda TG, Bader BW (2009) Tensor decompositions and applications. SIAM review 51(3):455–500

Krippendorff K (2011) Computing krippendorff's alpha-reliability

Le TH, Lauw HW (2021) Explainable recommendation with comparative constraints on product aspects. In: WSDM, WSDM '21, p 967–975

Le TH, Lauw HW (2022) Question-attentive review-level recommendation explanation. In: BigData, IEEE, BigData'22, pp 756–761

Li L, Zhang Y, Chen L (2021) Personalized transformer for explainable recommendation. In: ACL, pp 4947–4957

Li L, Zhang Y, Chen L (2023) On the relationship between explanation and recommendation: Learning to rank explanations for improved performance. ACM Trans Intell Syst Technol 14(2). https://doi.org/10.1145/3569423

Li P, Wang Z, Ren Z, et al (2017) Neural rating regression with abstractive tips generation for recommendation. In: SIGIR. ACM, SIGIR '17, p 345–354

Lu Y, Dong R, Smyth B (2018) Why i like it: Multi-task learning for recommendation and explanation. In: RecSys. ACM, RecSys '18, p 4–12

Ma W, Zhang M, Cao Y, et al (2019) Jointly learning explainable rules for recommendation with knowledge graph. In: WWW. ACM, WWW '19, p 1210–1221

McAuley J, Leskovec J (2013) Hidden factors and hidden topics: Understanding rating dimensions with review text. In: RecSys, RecSys '13, p 165–172

Moraes F, Yang J, Zhang R, et al (2020) The role of attributes in product quality comparisons. In: CHIIR, CHIIR '20, p 253–262

Park YA, Gretzel U (2010) Influence of consumers' online decision-making style on comparison shopping proneness and perceived usefulness of comparison shopping tools. Journal of Electronic Commerce Research 11(4):342–354

Ren Z, Liang S, Li P, et al (2017) Social collaborative viewpoint regression with explainable recommendations. In: WSDM. ACM, WSDM '17, p 485–494

Rendle S, Schmidt-Thieme L (2010) Pairwise interaction tensor factorization for personalized tag recommendation. In: WSDM. ACM, WSDM '10, p 81–90

Rendle S, Freudenthaler C, Gantner Z, et al (2009) Bpr: Bayesian personalized ranking from implicit feedback. In: UAI, UAI'09, pp 452–461

Tan Y, Zhang M, Liu Y, et al (2016) Rating-boosted latent topics: Understanding users and items with ratings and reviews. In: IJCAI, pp 2640–2646

Tao Y, Jia Y, Wang N, et al (2019) The fact: Taming latent factor models for explainability with factorization trees. In: SIGIR. ACM, SIGIR'19, p 295–304

Truong QT, Lauw H (2019) Multimodal review generation for recommender systems. In: WWW, WWW '19, p 1864–1874, https://doi.org/10.1145/3308558.3313463

Vedula N, Collins M, Agichtein E, et al (2023) Generating explainable product comparisons for online shopping. In: WSDM. ACM, WSDM '23, p 949–957

Wang N, Wang H, Jia Y, et al (2018a) Explainable recommendation via multi-task learning in opinionated text data. In: SIGIR. ACM, SIGIR '18, p 165–174

Wang X, He X, Feng F, et al (2018b) Tem: Tree-enhanced embedding model for explainable recommendation. In: WWW, WWW '18, p 1543–1552

Wu Y, Ester M (2015) Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In: WSDM. ACM, WSDM '15, p 199–208

Xian Y, Fu Z, Muthukrishnan S, et al (2019) Reinforcement knowledge graph reasoning for explainable recommendation. In: SIGIR. ACM, SIGIR'19, p 285–294

Xie Z, Singh S, McAuley J, et al (2023) Factual and informative review generation for explainable recommendation. In: AAAI, AAAI'23

Yang A, Wang N, Cai R, et al (2022) Comparative explanations of recommendations. In: WWW, WWW '22, p 3113–3123, https://doi.org/10.1145/3485447.3512031

Zhang Y, Lai G, Zhang M, et al (2014a) Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: SIGIR. ACM, SIGIR '14, p 83–92, https://doi.org/10.1145/2600428.2609579

Zhang Y, Zhang H, Zhang M, et al (2014b) Do users rate or review? boost phrase-level sentiment labeling with review-level sentiment classification. In: SIGIR, SIGIR '14

Zhang Y, Chen X, et al (2020) Explainable recommendation: A survey and new perspectives. Foundations and Trends® in Information Retrieval 14(1):1–101