Recommendation with Causality enhanced Natural Language Explanations

Jingsen Zhang†
xu.chen@ruc.edu.cn
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Xu Chen‡
xu.chen@ruc.edu.cn
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Jiakai Tang†
tangjiakai5704@gmail.com
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Weiqi Shao†
shaoweiqi@ruc.edu.cn
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Quanyu Dai
daiquanyu@huawei.com
Huawei Noah’s Ark Lab

Zhenhua Dong
dongzhenhua@huawei.com
Huawei Noah’s Ark Lab

Rui Zhang
rayteam@yeah.net
www.ruizhang.info

ABSTRACT

Explainable recommendation has recently attracted increasing attention from both academic and industry communities. Among different explainable strategies, generating natural language explanations is an important method, which can deliver more informative, flexible and readable explanations to facilitate better user decisions. Despite the effectiveness, existing models are mostly optimized based on the observed datasets, which can be skewed due to the selection or exposure bias. To alleviate this problem, in this paper, we formulate the task of explainable recommendation with a causal graph, and design a causality enhanced framework to generate unbiased explanations. More specifically, we firstly define an ideal unbiased learning objective, and then derive a tractable loss for the observational data based on the inverse propensity score (IPS), where the key is a sample re-weighting strategy for equalizing the loss and ideal objective in expectation. Considering that the IPS estimated from the sparse and noisy recommendation datasets can be inaccurate, we introduce a fault tolerant mechanism by minimizing the maximum loss induced by the sample weights near the IPS. For more comprehensive modeling, we further analyze and infer the potential latent confounders induced by the complex and diverse user personalities. We conduct extensive experiments by comparing with the state-of-the-art methods based on three real-world datasets to demonstrate the effectiveness of our method.

CCS CONCEPTS

- Information systems → Recommender systems.

† Corresponding author.
‡ Beijing Key Laboratory of Big Data Management and Analysis Methods.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '23, May 1–5, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/authors. Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9416-1/23/04...
https://doi.org/10.1145/3543507.3583260

KEYWORDS

Explainable Recommendation, Natural Language Explanations

ACM Reference Format:

1 INTRODUCTION

Explainable recommendation has been recognized as an important problem in both of the academic and industry communities. It basically aims to solve the problem of “why an item is recommended to a user?”, which can help to enhance the recommendation persuasiveness, user satisfaction and system transparency [36]. To achieve explainable recommendation, people have designed a large amount of models [7, 15, 16, 18, 24, 35, 38], among which the methods for generating natural language explanations are becoming more and more popular due to their capabilities on delivering richer and more accessible information [36]. In general, these methods regard user reviews as the explanations [16, 19], and the task of explainable recommendation is converted to the review generation problem. More specifically, early models like NRT [19] and Att2Seq [9] generate explanations completely based on the user/item IDs (or additionally the rating information). Due to the lack of informative guidance, these explanations usually contain a large amount of general words [16], which are less effective for assisting user decisions. To solve this problem, recent models, such as NETE [16] and PE-TER [18] firstly extract item features (e.g., product quality, clothing style) from the user reviews, and then regard them as inputs to generate more specific and informative explanations.

While the above models have achieved many successes, they are optimized based on the observational review datasets (e.g., Amazon and Yelp), which may have been biased by the user intrinsic preference or previous recommender systems (RecSys). For example, in Figure 1(a), the user is a digital fan. Thus we observe that most of her interactions and reviews are about items like computers, phones and digital cameras. However, for the other products (e.g., clothes), only a few interactions and reviews can be observed. Based on such
a dataset, the estimated explanations for the clothes can be not well. In the user reviews, we may observe that the feature “item quality” is mentioned with a much higher frequency due to the user reviewing habits. However, providing explanations should not be related with the reviewing habits. The system should have the capability of choosing any feature for explanation. When the system selects appearance or price as the explanation feature, the model may not work well due to the insufficient training corpus. To further verify such item- and feature-level biases, we conduct a preliminary study based on three real-world datasets including the TripAdvisor\(^1\), Amazon\(^2\) and Yelp\(^3\). We focus on the relation between the observation\(^4\) frequency and the number of items (or features) with this frequency. The results are presented in Figure 1(b). We can see a small number of items (or features) are observed with much higher frequency than the other ones. If one directly learns models based on such datasets, the parameters would be biased towards the frequently observed items/features, and the explanations generated for the underrepresented items/features can be unsatisfied.

To alleviate the above problem, we propose to formulate the explainable recommendation task with a causal graph to understand the bias formation mechanism and design an unbiased explainable recommender framework based on the inverse propensity score (IPS). While this seems to be an interesting idea, it is not easy due to the following challenges: to begin with, while recent years have witnessed many promising unbiased recommender models \(^27\), they are mostly designed for the item-level bias. However, in our problem, the biases come from both of the item- and feature-level. How to jointly correct them in a unified framework is still not clear. Then, the recommendation datasets can by quite sparse and noisy, the estimated IPS may deviate from the real one, how to handle the estimation error to guarantee the final performance needs our careful designs. In addition, user personalities in real-world scenarios can be quite diverse. Thus, there may exist latent confounders, which may invalidate the basic causal assumptions. How to model and infer them may also challenge our idea.

To overcome the above challenges, we firstly define an unbiased learning objective considering both of the item- and feature-level biases, and then derive a tractable loss for the observational data based on IPS, where the basic idea is to impose smaller weights to the items/features with higher observational frequencies, and assign larger weights to the long-tail items/features. For handling the prediction error of IPS, we firstly estimate it from the noisy data, then assume that the real IPS should be not far from the estimated one. At last, we minimize the maximum loss induced by the sample weights near the estimated IPS. For handling the potential latent confounders, we leverage neural networks to model and infer them, and the obtained results are incorporated into the IPS and user preference estimation processes. Based on all the above designs, we finally propose an unbiased explainable recommender framework, where we call it as USER for short.

In a summary, the main contributions of this paper can be concluded as follows: (1) we propose to build an unbiased explainable recommender framework based on causal inference, which, to the best of our knowledge, is the first time in the recommendation domain. (2) To achieve the above idea, we design a framework to jointly correct the item- and feature-level biases, where we propose a fault tolerant IPS estimation strategy, and also model and infer the potential latent confounders. (3) We conduct extensive experiments to demonstrate the effectiveness of our model based on three real-world datasets. To benefit the research community, we have released our framework at https://gitee.com/mindspore/models/tree/master/recommendation.

2 PRELIMINARIES

2.1 RecSys with Natural Language Explanations

Natural language explanations hold the promise of explaining recommendations according to the user preference in a flexible and informative manner. In practice, it is hard to obtain the ground truth of the explanations. Thus, people leverage user reviews, which contain rich user preferences, to approximate the real explanations \(^8, 16, 18\). Formally, suppose we have a user set \(\mathcal{U}\) and an item set \(\mathcal{I}\), the observed ratings and reviews\(^2\) from the users to the items are collected in \(\mathcal{R} = \{(r_{u, i}, s_{u, i}) | u \in \mathcal{U}, i \in \mathcal{I} \}\), where \(r_{u, i}\) is the rating falling into the range of \([1, 5]\), \(s_{u, i} = (s_{u, i}^1, s_{u, i}^2, \ldots, s_{u, i}^{|\mathcal{S}|})\) is the review posted by user \(u\) on item \(i\), \(s_{u, i}^k\) is the \(k\)th word in the review, the word vocabulary is defined as \(\mathcal{V}\), and \(l_{u, i}\) is the review length. For each review \(s_{u, i}\), there is a feature set \(\mathcal{F}\) associated with it, indicating the review contents. For example, in Figure 1(a), for the review of “The quality of the phone is very good”, the feature is “quality”. We define by \(\mathcal{F}\) the set of all features. Given \(\mathcal{U}, \mathcal{I}, \mathcal{F}\) and \(\mathcal{R}\), the task of natural language explainable recommendation aims to learn a model, such that for a give user-item pair and a feature set, the model can accurately predict the review and rating.

To accomplish the above task, people have designed a lot of models \(^8, 9, 16, 18, 19\). Generally speaking, there are usually two parts in these models: (1) review prediction and (2) rating prediction. For the first part, suppose we have a review \(s_{u, i} = (s_{u, i}^1, s_{u, i}^2, \ldots, s_{u, i}^{|\mathcal{S}|})\),

\(^1\)https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews
\(^2\)https://jmcauley.ucsd.edu/data/amazon/
\(^3\)https://www.yelp.com/dataset
\(^4\)Here, an observation can be an item interaction or a feature mention.
we only consider the user, item and rating, the ideal loss function can be adopted. In this section, we provide a causal perspective to understand this problem. For general recommendation, where the observed datasets do not follow this causal graph, since the review and rating can be associated (e.g., on the sentiment polarity), $g^S$ and $g^R$ are usually designed by sharing the architectures and parameters \cite{16, 18}, which unifies $L^S_{u, i, f}$ and $L^R_{u, i}$ into a multi-task learning framework.

2.2 Causal Understanding of Debiased RecSys

Debiased recommendation is becoming more and more popular \cite{4, 12, 26, 27, 32}. In this section, we provide a causal perspective to understand this problem. For general recommendation, where we only consider the user, item and rating, the ideal loss function should evaluate all the user-item pairs \cite{8, 27}, that is: $L_{\text{ideal}} = \frac{1}{|U||I|} \sum_{u \in U} \sum_{i \in I} L^R_{u, i} = \frac{1}{|U||I|} \sum_{u \in U} \sum_{i \in I} E_{i} - p^f(i)[L^R_{u, i}]$, where $p^f(i) = \frac{1}{|I|} \sum_{i' \in I} E_{i'}$ is the uniform distribution on the item set. This ideal objective is equal to optimizing the model with the dataset generated according to the causal graph in Figure 2(a). However, in practice, the observed datasets do not follow this causal graph, since the observation of an item should be related with the user due to the diverse personal characteristics or the personalization requirements of the previous recsys (see Figure 2(b)). In order to obtain debiased recommender models based on the observed datasets, the following loss function can be adopted $L_{\text{debias}} = \frac{1}{|U||I|} \sum_{u \in U} \frac{1}{|I|} \sum_{i \in I} \frac{p^f(i)[L^R_{u, i}]}{p^f(i)[L^S_{u, i, f}]} t^R_{u, i}$.

Figure 2: (a) The causal graph for generating ideal datasets in general recommendation. (b) The causal graph for generating observational datasets in general recommendation. (c) The causal graph for generating ideal datasets in explainable recommendation. (d) The causal graph for generating observational datasets in explainable recommendation. (e) The causal graph for explainable recommendation with confounders.

\[ p^f(i)[L^R_{u, i}] \] is the probability of observing the interaction of item $i$ given user $u$ (a.k.a. propensity score). $I_u$ is the set of items interacted by user $u$. For the unbiasedness of this objective, we have the following theory, where we present the proof in Appendix A.

**Theorem 1** (Unbiasedness of $L_{\text{debias}}$). Suppose each item in $I_u$ is a random variable, and independently sampled from the observational probability $p^f(i)[L^S_{u, i, f}]$, then $E[L_{\text{debias}}] = L_{\text{ideal}}$.

Remark. Above, we provide a causal understanding of the debiased recommendation by analyzing the causal graphs leveraged for generating the ideal and observed datasets. These causal graphs enable us to more intuitively understand the bias formation mechanisms, and inspire us to improve debiased recommender models from the causal perspective (e.g., capturing the potential latent confounders).

3 THE USER FRAMEWORK

In this section, we firstly define the ideal unbiased learning objective for explainable recommendation, and correspondingly show the causal graph for ideal dataset generation. Then, we analyze the causal graph for generating the observed datasets, and propose a debiased loss function for the observed datasets based on IPS. In the next, we design a fault tolerant mechanism for the IPS estimated from the noisy and sparse recommendation datasets. At last, we model and infer the potential latent confounders, and incorporate them into the IPS and review/rating prediction processes. In the following, we introduce our framework more in detail.

3.1 Ideal Learning Objective

In our explainable recommendation task, the ideal objective should evaluate the reviews and ratings for all the users, items and features, thus we have the following ideal loss:

\[
L^e_{\text{ideal}} = \frac{1}{|U||I||F|} \sum_{u \in U} \sum_{i \in I} \sum_{f \in F} t^S_{u, i, f} p^f(i)[L^S_{u, i, f}] + \frac{\alpha}{|U||I|} \sum_{u \in U} \sum_{i \in I} t^R_{u, i} + \frac{1}{|U|} \sum_{u \in U} E_{i-p^f(i)[L^S_{u, i, f}]} + \frac{\alpha}{|U|} \sum_{u \in U} E_{i-p^f(i)[L^R_{u, i}]},
\]
where $\alpha$ is a hyper-parameter balancing the importances between the review and rating predictions. $p^F(f) = \frac{1}{|F|}$ ($\forall f \in F$) is the uniform distribution on the feature set. This objective needs to be optimized based on the datasets generated according to the causal graph in Figure 2(c), where the observation probabilities of the items and features are not influenced by any other factors, but follow uniform distributions. However, such datasets can only be obtained by conducting a large amount of online experiments, which are too expensive [27]. Thus, we derive a tractable unbiased loss for the observed datasets, which are much more accessible.

3.2 Unbiased Loss for the Observed Datasets

To derive the unbiased loss function, we assume that the observed datasets are generated according to the causal graph in Figure 2(d). The rationalities of this causal graph are presented as follows: $U \rightarrow I$: the items are observed due to the user active selections or system recommendations, which are naturally influenced by the user personalized preferences. $U \rightarrow F$ and $I \rightarrow F$: since the reviews are written by the users for the items, the mentioned features in the review are determined by the user preferences and observed items. $U, I \rightarrow Ra$: the ratings are given by the users on the items, thus they are determined by the user preferences and observed items. $U, I, F \rightarrow Re$: the reviews are posted by the users for the items around the features, which form the edges from $U, I, F$ to $Re$.

Based on the above causal graph, we design the following unbiased loss function:

$$L_{debias}^F = \frac{1}{|U|} \sum_{u \in U} \left\{ \frac{1}{|I_u|} \sum_{i \in I_u} \frac{p^I(i)}{p^O(i,u)} \left[ \frac{1}{|F_u|} \sum_{f \in F_u} p^F(f) \left( \sum_{i \in I} \frac{p^F(f)}{p^O(f,u,i)} L^S_{u,i,f} \right) \right] \right\} + \frac{\alpha}{|U|} \sum_{u \in U} \frac{1}{|I_u|} \sum_{i \in I_u} \frac{p^I(i)}{p^O(i,u)} L^R_{u,i},$$

where $|F_u|$ is the set of features mentioned in the review of user $u$ for item $i$. $p^I(i) / p^O(i,u)$ is the probability of observing feature $f$ given the user-item pair $(u, i)$. The unbiasedness of $L_{debias}^F$ is shown in the following theorem, and the proof is presented in Appendix B.

**Theorem 2** (Unbiasedness of $L_{debias}^F$). Suppose each item in $I_u$ and each feature in $F_u$ are independently sampled from $p^I_0(i)$ and $p^O_0(f,u,i)$, respectively, then $E[L_{debias}^F] = L_{debias}^F$.

3.3 IPS with Fault Tolerant Mechanism

In most of the previous works [4, 12, 26, 27, 32], IPS is estimated from the recommendation datasets, which can be highly sparse and noisy, making it hard to obtain the real IPS. To alleviate this problem, we introduce a fault tolerant mechanism to handle the estimation error. More specifically, we firstly predict an initial IPS (which can be inaccurate), and then assume that the real IPS fall into an $\epsilon$-ball centered at the initial IPS. At last, we minimize the maximum loss induced by the IPS in this $\epsilon$-ball. Such optimization is equal to minimize the upper bound of the loss function with the real IPS, which is demonstrated to be effective in our experiments.

Formally, we solve the following optimization problem:

$$\min_{\Theta_p} \max_{\theta_p} \frac{1}{|U||I||F|} \sum_{u \in U} \sum_{i \in I_u} \frac{1}{p^O(i,u)} L^S_{u,i,f} + \frac{\alpha}{|U||I|} \sum_{u \in U} \frac{1}{|I_u|} \sum_{i \in I_u} p^I(i) L^R_{u,i},$$

subject to $|p^I_0(i,u) - p^I(i)| \leq \epsilon, |p^F_0(f,u,i) - p^F(f,u,i)| \leq \epsilon, \forall u \in U, i \in I_u, f \in F_u.$

where $p^I_0(i,u)$ and $p^F_0(f,u,i)$ are the initially estimated IPS. Straightforwardly, $p^I_0(i,u)$ (or $p^F_0(f,u,i)$) can be realized with free parameters for each user-item pair (or user-item-feature triplet). However, this requires nearly $|U||I|$ (or $|U||I||F|$) parameters, which may easily over-fit the training data, and be hard to generalize. In practice, we learn models to estimate $p^I_0(i,u)$ (or $p^F_0(f,u,i)$), where the model parameters are collected in $\Theta_p$, and we introduce their specifications later. $\Theta_p$ is the set of model parameters for predicting the reviews and ratings. $\epsilon$ defines the tolerance level. Smaller $\epsilon$ requires more accurate initial IPS to quickly discover the real IPS within a small range. Larger $\epsilon$ indicates lower confidence on the initial IPS, and our model needs to find the real IPS in a wider space.

Actually, the above problem has close relation with the ideal learning objective $L_{ideal}^F$. To reveal such relation, we rewrite the constraints in (1) as regularizers, and define the following objective:

$$L_{debias} = \max_{\Theta_p} \frac{1}{|U||I||F|} \sum_{u \in U} \sum_{i \in I_u} \frac{1}{p^O(i,u)} L^S_{u,i,f} + \frac{\alpha}{|U||I|} \sum_{u \in U} \frac{1}{|I_u|} \sum_{i \in I_u} p^I(i) L^R_{u,i} - \lambda_S \sum_{u \in U} \sum_{i \in I_u} |p^I_0 - p^I| - \lambda_R \sum_{u \in U} \sum_{i \in I_u} |p^F_0 - p^F|,$$

where $p^I_0 = p^I_0(i,u)$ and $p^F_0 = p^F_0(f,u,i)$, $\lambda_R$ and $\lambda_S$ are regularization parameters, and we slightly change (1) by constraining all the user-item pairs. Then we have the following theorem:

**Theorem 3** (Theoretical justification). Suppose: (1) $p^I_0$ and $p^F_0$ are the real propensity scores, $p^I_0$ and $p^F_0$ are the propensity scores used in our objective and $p^I_0$ and $p^F_0$ are the initially estimated propensity scores. All the propensity scores fall into the range of $[\kappa_1, \kappa_2]$, where $0 < \kappa_1 < \kappa_2 < 1$. (2) $\frac{L_{debias}^F}{p^O(i,u)} \leq \Delta$ and $\frac{L_{ideal}^F}{p^O(i,u)} \leq \Delta_1$, if we set $\lambda_S = \frac{\Delta_1}{\Delta}, \lambda_R = \frac{\Delta_1}{\Delta}$. Then the following inequality holds with probability as least $1 - \eta$.

$$L_{ideal}^F \leq L_{debias}^F + Const,$$

where $Const$ is a constant related with $\eta$.

This theory relates our objective with the ideal loss, which theoretically reveals the rationality of our framework. The proof of this theory is presented in the Appendix C.

3.4 Modeling the Latent Confounders

In our problem, there can be three types of latent confounders (see Figure 2(e)), that is, (i) the confounder between $I$ and $Ra$ (e.g., the promotion of the items), (ii) the confounder between $I$ and
we leverage similar architectures as used for \( g \), the behaviors of following the other reviews). We assume that the above latent confounders can be fully represented by three embeddings \( z_1, z_2 \) and \( z_3 \), respectively. Similar to [2, 22], we infer the latent confounders by neural networks based on the user, item and feature information.

In specific, we let \( z_1 = g_1(u, i) \), \( z_2 = g_2(u, i) \) and \( z_3 = g_3(u, i, f) \). Intuitively, if the latent confounders are well represented, then they should lead to larger likelihoods of the item/feature observations. Thus we have the following cross-entropy loss:

\[
L \left( \theta, \theta, \theta, \alpha \right) = \sum_{u \in U} \sum_{i \in I_u} \left( \frac{1}{N[F]} \sum_{f \in F} p^S_{O\hat{O}}(f|u, i, z_1) + p^S_{O\hat{O}}(f|u, i, z_2) + p^S_{O\hat{O}}(f|u, i, z_3) \right) \]
biased training and unbiased evaluation”. To this end, we use 50%
We apply our framework on
NETE†WWW ’23, May 1–5, 2023, Austin, TX, USA Jingsen Zhang
MF with the following simple but effective models:
Table 2: Overall comparison between our framework and baselines. For BLEU and ROUGE, the results are percentage values with “%” omitted. For each dataset and evaluation metric, we use bold fonts to label the best performance. “-” means the evaluation metric is not available for the model. The performance improvements of our framework are significant under paired t-test with p < 0.05.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>BLEU (%)</th>
<th>ROUGE-1 (%)</th>
<th>ROUGE-2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-1</td>
<td>BLEU-4</td>
<td>F1 Recall</td>
</tr>
<tr>
<td>MF</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVD++</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Att2Seq</td>
<td>13.788</td>
<td>0.622</td>
<td>13.981</td>
</tr>
<tr>
<td>NRT</td>
<td>10.853</td>
<td>0.306</td>
<td>14.830</td>
</tr>
<tr>
<td>NETE</td>
<td>14.960</td>
<td>0.641</td>
<td>15.690</td>
</tr>
<tr>
<td>NETE-USER</td>
<td>15.923</td>
<td>0.769</td>
<td>26.060</td>
</tr>
<tr>
<td>PETER</td>
<td>18.718</td>
<td>2.148</td>
<td>22.344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AZ-MT dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
</tr>
<tr>
<td>SVD++</td>
</tr>
<tr>
<td>Att2Seq</td>
</tr>
<tr>
<td>NRT</td>
</tr>
<tr>
<td>NETE</td>
</tr>
<tr>
<td>NETE-USER</td>
</tr>
<tr>
<td>PETER-USER</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>YELP dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
</tr>
<tr>
<td>SVD++</td>
</tr>
<tr>
<td>Att2Seq</td>
</tr>
<tr>
<td>NRT</td>
</tr>
<tr>
<td>NETE</td>
</tr>
<tr>
<td>NETE-USER</td>
</tr>
<tr>
<td>PETER</td>
</tr>
</tbody>
</table>

Baselines. We compare our framework with the following repre-
sentative baselines: Att2Seq [9] is an LSTM model for generating
user reviews directly based on the user/item ID and rating informa-
tion. NRT [19] is also a review generation model based user/item
ID, but its backbone is the gated recurrent unit (GRU). NETE [16] is
a controllable review generation model, where the text sequence is
decoded by taking the user, item and feature information as input.
PETER [18] also leverages features as input to generate more
informative explanations, but it uses transformer as the main archi-
tecture, which can usually achieve the state-of-the-art performance.
We apply our framework on NETE and PETER, which are both feature enhanced explainable recommender models. The obtained
methods are called NETE-USER and PETER-USER, respectively.
For the task of rating prediction, we also compare our framework
with the following simple but effective models: MF [14] is the well-
known matrix factorization model, where the users and items are
represented by latent vectors, and the user-item preferences are
estimated by inner-product based on these vectors. SVD++ [13] is
a variant of MF, where the preference is estimated by taking the
user history information into consideration.
Implementation details. In general, the experiments in the do-
main of debiased recommendation should follow the paradigm of
“biased training and unbiased evaluation”. To this end, we use 50%
of each user interactions as the biased training set. For building the
validation/testing sets, we follow the previous work [3, 20, 25, 26,
34, 39] to sample from the other interactions based on the inverse
item/feature observation frequencies, where the more frequently
observed items/features are sampled with lower probabilities. We
set the splitting ratio between the validation and testing sets as 1:1.
In the experiments, we set the maximum length of the generated
explanations as 15, and the vocabulary ‘V’ is constructed by 20000
most frequently mentioned words. We tune the hyper-parameters
of our model by grid search. In specific, we tune the learning rate
and hidden size in the ranges of [0.1, 0.01, 0.001] and [32, 64, 128,
256], respectively. The batch size for all models is set as 128 and the
weight of $L_{\text{reg}}$ is tuned in [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000].
We implement the baselines with the codes released in [16, 18] at
https://github.com/lileipisces/NLG4RS. The parameters are set as
the optimal values reported in the original paper or tuned in the
same ranges as our model’s. To evaluate the explanation quality, two
commonly used metrics including BLEU [23] and ROUGE [21] are
leveraged for model comparisons. To evaluate the recommendation
performance, we use MAE [19] as the evaluation metric.

5.2 Overall Performance
The overall comparison results are presented in Table 2. We can see:
for the explanation task, the winner between Att2Seq and
NRT varies on different datasets and evaluation metrics, and the
performance gap is not large, which suggests that these models
may have similar explanation capabilities. By introducing item
features as input, NETE and PETER can achieve better performances
than Att2Seq and NRT in most cases. This is as expected, since the
features can provide much valuable information to describe the
review content and better guide the sentence decoding process. It
Table 3: Performance comparison between our framework and its variants, where “ROUGE” represents “ROUGE-1-F1 (%)”. We use bold fonts to label the best performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>TA-HK ROUGE MAE</th>
<th>AZ-MT ROUGE MAE</th>
<th>YELP ROUGE MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETER-USER-T</td>
<td>22.344 0.710</td>
<td>21.664 0.732</td>
<td>22.078 0.877</td>
</tr>
<tr>
<td>PETER-USER-L</td>
<td>19.848 0.714</td>
<td>19.421 0.917</td>
<td>12.354 0.958</td>
</tr>
<tr>
<td>PETER-USER-F</td>
<td>22.897 0.696</td>
<td>22.016 0.711</td>
<td>21.560 0.855</td>
</tr>
<tr>
<td>PETER-USER-I</td>
<td>26.636 0.718</td>
<td>22.336 0.744</td>
<td>21.606 0.964</td>
</tr>
<tr>
<td>PETER-USER-F</td>
<td>22.237 0.706</td>
<td>21.292 0.710</td>
<td>21.940 0.847</td>
</tr>
<tr>
<td>PETER-USER-I</td>
<td>26.897 0.692</td>
<td>23.234 0.705</td>
<td>22.747 0.838</td>
</tr>
</tbody>
</table>

is encouraging to see that, by imposing our framework on NETE and PETER, the performance can be significantly enhanced, which are consistent on all the datasets and metrics. Considering that the evaluation datasets are unbiased, this result demonstrates that our framework can provide better explanations if we treat different items and features equally in the testing phase. In the original models, the parameters are directly learned based on the biased datasets, thus they cannot perform well when the testing sample distribution is changed to be unbiased. However, our framework can effectively correct the item- and feature-level biases, which achieves much better performance than the original models. For the rating prediction task, we find that by incorporating the history information, SVD++ can achieve much better performance than MF on all the datasets. In most cases, NRT, NETE and PETER perform better than MF and SVD++. This is not surprising, because in the first three models, the rating and review prediction problems are formulated as a multi-task learning framework, which makes them can be mutually enhanced by each other. By applying our framework on NETE and PETER, the performances are improved, which demonstrates the debiasing effect of our framework.

5.3 Ablation Studies

In this section, we conduct ablation studies to study the contributions of different components of our framework. More specifically, suppose the original model is X, then we compare our framework with its four variants: in X-USER-T, we remove the IPS fault tolerant mechanism, and only retain the initially estimated IPS. In X-USER-L, we do not model the latent confounders. In X-USER-I, we do not correct the item-level bias (i.e., removing \( p_{I1} \) and \( p_{I2} \) in \( L_{\text{final}} \)). In X-USER-F, we do not correct the feature-level bias (i.e., removing \( p_{F1} \) in \( L_{\text{final}} \)). We use PETER as the original model, and the conclusions on NETE are similar and omitted. We set the parameters as their optimal values tuned in the above section. The comparison results are presented in Table 3. We can see: in most cases, dropping the IPS fault tolerant mechanism lowers the performance more severely than ignoring the latent confounders. We speculate that the recommendation datasets can be too sparse and noisy to obtain accurate IPS, thus introducing fault tolerant mechanism for the estimated IPS can well bound the real objective function, and help to generate better explanations. While modeling latent confounders can indeed bring better performance, the improvements are not large. We argue that the confounder structures can be very complicated, representing them with unified embeddings can be too coarse, and fail to provide enough priors for effective parameter learning. Correcting the item- and feature-level biases are both important, which is evidenced by the lowered performances of X-USER-I and X-USER-F against X-USER. For all datasets, X-USER-F can achieve better performance than X-USER-I on MAE, because the feature-level bias is irrelevant with the rating prediction task. However, for the explanation task, the performance of X-USER-F is comparable or worse than that of X-USER-I, which suggests that removing the feature-level bias is important to enhance the explanation quality. The best performance is achieved when we combine all the components, which demonstrates that they are all necessary.

5.4 Influence of the Item/Feature Popularity

To better understand the debiasing effect of our framework, we report the performance on the items/features with different popularity levels. In specific, we firstly sort the items and features in the training set according to their frequencies, where more popular items are ranked higher. Then we cluster the items/features ranked between 0%-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100% into five groups. According to the item groups, we separate the original testing set into five subsets, and we denote them as \( I_1, I_2, I_3, I_4 \) and \( I_5 \), for example, in \( I_1 \), the items are all ranked between 0%-20% in the training set. Similarly, based on the feature groups, the testing set can also be divided into five subsets, and we denote them as \( F_1, F_2, F_3, F_4 \) and \( F_5 \), respectively. In the experiments, we remain the training set the same as the previous experiments, while the final performances are reported on the above ten sub-testing sets. We use YELP as the experiment dataset, and the model parameters are set as their optimal values tuned in Table 2. From the results presented in Figure 3, we can see: from the item perspective, the performance improvement brought by our framework is larger on the items with lower observation frequencies. The conclusion is consistent on both of the explanation generation and rating prediction tasks. The reason can be that the items with lower frequencies are higher weighted in our framework. They can be optimized more sufficiently, and thus the performance on them can be improved larger. From the feature perspective, the performance improvement is small when the features are too unpopular. We find that the feature frequencies in the last groups (e.g., \( F_5 \)) are mostly 1”. In such a scenario, while we have higher weighted these samples, the information provided for each feature is too limited and not diverse enough, which impacts the model generalization capability and performance improvements.

5.5 Studies on the Generated Explanations

5.5.1 Qualitative studies. For more intuitively understanding our framework, we present many examples from the TA-HK dataset. In specific, we use PETER as the original model, and compare the explanations generated from PETER and PETER-USER. We set the model parameters as their optimal values tuned in Table 2. For references, we also present (i) the ground truth of the explanations, and (ii) the observation frequencies of the items and features, which are computed as the ratio between the current and maximum item interaction (or feature mention) times. From the results shown in Table 4, we can see, in the first case, when the item and feature have high observation frequencies, both PETER and PETER-USER can capture the key word “amenities”. For the second and third cases, where the items and features are underrepresented in the dataset (i.e., with lower observation frequencies), PETER-USER
can generate more accurate and informative explanations around the given features, while PETER only produces general words for the explanations. More specifically, in the second case, PETER-USER can accurately generate the explanation around the feature “steak”, but PETER fails to capture such information. The reasons behind the above phenomena can be that in our framework, the items/features with lower observation frequencies are higher weighted to enhance their importances in the training phase. However, in the original model, these items/features are mostly neglected, thus the performance is not satisfied.

5.5.2 Quantitative studies. To study whether the generated explanations can indeed help users, we further design questionnaires to ask the feelings of the users on the explanations. In specific, we focus on two aspects of the explanations [28], that is, (i) effectiveness: whether the explanations can communicate useful information on the items, and (ii) persuasiveness: whether the explanations can persuade users to make decisions. Based on these aspects, we design the following two questions [29]: Q1: Does the explanation help you to learn more about the recommended item? Q2: Does the explanation help you to make fast decisions? For each of these questions, the annotator is required to give a rating ranging from 1 to 5 to indicate her agreement (i.e., 1-strongly disagree, 2-disagree, 3-neutral, 4-agree and 5-strongly agree) on the question. In the experiments, we randomly select 50 samples from the testing set of each dataset, and 12 annotators with different backgrounds are employed from a university. We leverage NETE and PETER as the original models, and compare the explanations generated from them with the ones produced by imposing our framework. The results are reported as the average rating across different annotators and samples, which are presented in Table 5. We can see: for both questions, our framework can lead to better performance than the original model on all the datasets, and the results are consistent for both NETE and PETER. On average, our framework can improve the performance of the original model by about 17.0% and 16.6% on Q1 and Q2, respectively. This observation demonstrates the effectiveness of our framework on improving the explanation effectiveness and persuasiveness, which demonstrates its potential in real-world settings. While we have noticed that there can be general limitations for the questionnaire studies (e.g., discrepancies between different annotator understandings on the explanations), under the same limitations, our framework can always achieve better performances, which suggests its superiority.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we propose the task of debiased explainable recommendation for the first time. For solving this task, we build a principled framework to jointly correct the item- and feature-level biases, and design fault tolerant IPS mechanism and latent confounder modeling strategy to improve this framework. Extensive experiments demonstrate that our framework can bring improved explanation and recommendation performances for the state-of-the-art models. This paper opens a novel direction on explainable recommendation, and we believe there still left much room for improvement. For example, one can leverage IP’s normalization or doubly robust methods to lower the variance of the loss function. In addition, this paper mainly focuses on natural language explanations, it is interesting to extend the debiasing idea to the other types of explanations. At last, the user preferences in real-world scenarios can be dynamic, how to design debiased models considering the temporal influence is also important.
ACKNOWLEDGMENTS
This work is supported in part by National Natural Science Foundation of China (No. 62102420 and No. 61832017), Beijing Outstanding Young Scientist Program NO. BJJWZYHJ12019100020098, Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class Initiative," Renmin University of China, and Public Computing Cloud, Renmin University of China. The work is sponsored by Huawei Innovation Research Programs. We gratefully acknowledge the support from Mindspore\footnote{https://www.mindspore.cn}, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

REFERENCES
A PROOF OF THEORY 1

Proof. By taking expectation on $L_{\text{debias}}$, we have:

$$E[L_{\text{debias}}] = E[\frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} \frac{p^f(i)}{p^o(i)} L^R_{u,i}]$$

$$= \frac{1}{|U|} \sum_{u \in U} [1 \sum_{i \in I_u} \frac{E_i p^f(i)}{p^o(i)} L^R_{u,i}]$$

$$= \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} \frac{p^f(i)}{p^o(i)} \left( \frac{p^f(i)}{p^o(i)} L^R_{u,i} \right)$$

$$= \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} \frac{p^f(i)}{p^o(i)} \left( L^R_{u,i} \right)$$

$$= \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} E_i p^f(i) L^R_{u,i}$$

(3)

$$= L_{\text{ideal}}$$

where the second equation holds because $i$ is the only random variable and independently sampled from $p^f(i)$. The second last equation holds because $E_i p^f(i) L^R_{u,i}$ is irrelevant with $i$. □

B PROOF OF THEORY 2

Proof. By taking expectation on $L^e_{\text{debias}}$, we have:

$$E[L^e_{\text{debias}}] = \left( \frac{1}{|U|} \sum_{u \in U} \left[ \frac{1}{|I_u|} \sum_{i \in I_u} \frac{p^f(i)}{p^o(i)} \left( \frac{1}{|T_{u,i}|} \sum_{f \in T_{u,i}} p^f(f) L^S_{u,i,f} \right) \right] \right)$$

$$= \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} E_i p^f(i) L^S_{u,i}$$

$$= \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} E_i p^f(i) L^S_{u,i}$$

$$= \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} \left( L^S_{u,i} \right)$$

where the second equation holds because $|T_{u,i}| = 1$. □

C PROOF OF THEORY 3

For easy analysis, we introduce some notations. Let

$$L^S_{\text{ideal}} = \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} E_i p^f(i) L^S_{u,i}$$

$$L^R_{\text{ideal}} = \frac{1}{|U|} \sum_{u \in U} \sum_{i \in I_u} E_i p^f(i) L^R_{u,i}$$

then

$$L^e_{\text{ideal}} = L^S_{\text{ideal}} + aL^R_{\text{ideal}}$$

Let

$$L^S_{\text{debias}} = \frac{1}{|U||I||F|} \sum_{u \in U} \sum_{i \in I_u} \frac{1}{|T_{u,i}|} p^f(f) L^S_{u,i,f}$$

$$L^R_{\text{debias}} = \frac{1}{|U||I|} \sum_{u \in U} \sum_{i \in I_u} \frac{1}{|T_{u,i}|} L^R_{u,i}$$

then

$$L_{\text{debias}} = L^S_{\text{debias}} + aL^R_{\text{debias}}$$

To prove this theory, we firstly bound $L^R_{\text{ideal}}$ and $L^S_{\text{ideal}}$ with $L^R_{\text{debias}}$ and $L^S_{\text{debias}}$, respectively. Then we combine these results to obtain the inequality (2).

C.1 Bounding $L^R_{\text{ideal}}$ with $L^R_{\text{debias}}$

Let

$$L^R_{\text{ideal}} - E_{P^o} [L^R_{\text{debias}}] = \frac{1}{|U||I|} \sum_{u \in U} \sum_{i \in I_u} \left( 1 - \frac{P^f_{o}}{P^o} \right) L^R_{u,i}$$

(4)

Let $s_u = \frac{1}{|U|} \sum_{i \in I_u} \frac{p^f_{i}}{p^o_{i}} L^R_{u,i}$, then

$$L^R_{\text{debias}} = \frac{1}{|U|} \sum_{u \in U} s_u.$$

According to the Hoeffding’s inequality, if $s_u \in [a_u, b_u]$ and $|a_u - b_u| \leq B_1$, then the expectation of $L^R_{\text{debias}}$ is bounded by the following value with probability at least $1 - \eta$:

$$E_P [L^R_{\text{debias}}] \leq L^R_{\text{ideal}} + B_1 \sqrt{\frac{1}{2|U|} \log \frac{2|U||I||F|}{\eta}}.$$
Let $\Delta_2 = \sum_{u \in U} 3u$. According the Hoeffding’s inequality, suppose $3u \in [cu, du]$, $|cu - du| \leq B_2$, then the expectation of $L_{\text{debias}}^S$ is bounded by the following value with probability at least $1 - \eta_2$:

$$
E_{p_O^s,p_O^s}[L_{\text{debias}}^S] \leq L_{\text{debias}}^S + B_2\sqrt{\frac{1}{2|U|}\log\left(\frac{2|I||F|}{\eta_2}\right)}.
$$  (6)

Let $\Delta_3 = B_2\sqrt{\frac{1}{2|U|}\log\left(\frac{2|I||F|}{\eta_2}\right)}$, $\beta_{u,i,f} = \frac{L_{u,i,f}^S}{|U||I||F|}$, then we have the following inequality holds with probability at least $1 - \eta_2$:

$$
L_{\text{ideal}}^S = L_{\text{ideal}}^S - E_{p_O^S,p_O^S}[L_{\text{debias}}^S] + E_{p_O^S,p_O^s}[L_{\text{debias}}^S] \\
\leq \frac{1}{|U||I||F|} \sum_{u \in U} \sum_{i \in I} \left(1 - \frac{p_{I}^s}{p_{O}^s}\right) L_{u,i,f}^S + L_{\text{debias}}^S + C_3 \\
= \sum_{u \in U} \sum_{i \in I} \beta_{u,i,f}(p_{I}^s - p_{O}^s) + L_{\text{debias}}^S + C_3 \\
= \sum_{u \in U} \sum_{i \in I} \beta_{u,i,f}(p_{O}^s - p_{I}^s) + L_{\text{debias}}^S + C_3 \\
+ \sum_{u \in U} \sum_{i \in I} \beta_{u,i,f}(p_{I}^s - p_{O}^s) + \sum_{u \in U} \sum_{i \in I} \beta_{u,i,f}(p_{O}^s - p_{I}^s) \\
+ L_{\text{debias}}^S + C_4 \\
\leq \sum_{u \in U} \sum_{i \in I} \beta_{u,i,f}(p_{I}^s - p_{O}^s) + \sum_{u \in U} \sum_{i \in I} \beta_{u,i,f}(p_{O}^s - p_{I}^s) \\
+ L_{\text{debias}}^S + C_4 \\
\leq \frac{4\Delta_2}{|F|} \sum_{u \in U} \sum_{i \in I} (p_{I}^s - p_{O}^s) + (p_{O}^s - p_{I}^s) \\
+ L_{\text{debias}}^S + C_4 \\
= -L_{\text{debias}}^S - \frac{\Delta \kappa_2}{|U||I||F|} \sum_{u \in U} \sum_{i \in I} (p_{I}^s - p_{O}^s) + (p_{O}^s - p_{I}^s) + C_5
$$  (7)

where $C_4 = \Lambda_1(\kappa_2^2 - \kappa_1^2) + C_3$ and $C_5 = \frac{4\Lambda_1\kappa_2^2}{|F|} + C_4$.

### C.3 Bounding $L_{\text{ideal}}$ with $L_{\text{debias}}$

Suppose we define $X_1$ as $L_{\text{ideal}}^R \leq L_{\text{debias}}^R - \frac{\Lambda_1\kappa_2}{|U||I||F|} \sum_{u \in U} \sum_{i \in I} |p_{I}^s - p_{O}^s| + C_2$, $X_2$ as $L_{\text{ideal}}^S \leq L_{\text{debias}}^S - \frac{\Lambda_2\kappa_2}{|U||I||F|} \sum_{u \in U} \sum_{i \in I} (|p_{I}^s - p_{O}^s|) + (|p_{O}^s - p_{I}^s|) + C_5$ and $Y$ as $L_{\text{ideal}}^S \leq L_{\text{debias}}^S + \text{Const}$. Based on the above results, we know $P(X_1) \geq 1 - \eta_1$ and $P(X_2) \geq 1 - \eta_2$. Since $X_1 \cap X_2 \rightarrow Y$, we have: $P(Y) \geq P(X_1 \cap X_2) = 1 - P(X_1 \cup X_2) \geq 1 - P(X_1) - P(X_2) \geq 1 - \eta_1 - \eta_2$. Let $\eta = \eta_1 + \eta_2$, we have $Y$ holds with probability at least $1 - \eta$. 

---

**Recommendation with Causality enhanced**

Natural Language Explanations

WWW ’23, May 1–5, 2023, Austin, TX, USA