Operations Research and Fuzziology 运筹与模糊学, 2025, 15(2), 528-541
Published Online April 2025 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2025.152103

Hans汉斯

基于历史评论反馈建模的可解释推荐算法

徐戈辉，艾　均*，苏　湛，苏　杭

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年3月7日；录用日期：2025年4月3日；发布日期：2025年4月11日

摘　要

可解释推荐系统在提供个性化推荐的同时揭示推荐逻辑，已成为推荐系统领域的重要研究方向。尽管基于Transformer的文本生成技术能够产生流畅的自然语言解释，但其性能高度依赖于数据集的密度，在现实场景中普遍存在的稀疏数据条件下表现受限。针对这一挑战，本文提出一种融合历史评论反馈建模的可解释推荐框架。该框架采用编码器－解码器架构，在特征提取阶段，对用户历史评论和物品关联评论进行深度语义建模，通过注意力机制提取评论中的隐含偏好特征与属性特征。在解释生成阶段，使用多任务联合学习范式，将评分预测任务与文本生成任务进行联合优化。通过在Amazon、Yelp等公开数据集上的实验验证，本模型在评分预测和解释生成任务中均达到了最先进水平，在多数指标上取得了显著的提升。本研究为平衡推荐系统可解释性与数据稀疏性问题提供了新的解决方案。

关键词

可解释推荐，历史评论，反馈建模，表示学习，Transformer

Explainable Recommendation Algorithms Based on Historical Reviews Feedback Modelling

Gehui Xu, Jun Ai*, Zhan Su, Hang Su

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 7th, 2025; accepted: Apr. 3rd, 2025; published: Apr. 11th, 2025

Abstract

Explainable recommendation systems, which deliver personalized suggestions while revealing

*通讯作者。

文章引用: 徐戈辉, 艾均, 苏湛, 苏杭. 基于历史评论反馈建模的可解释推荐算法[J]. 运筹与模糊学, 2025, 15(2): 528-541. DOI: 10.12677/orf.2025.152103

their rationale, have become pivotal in recommendation research. Although Transformer-based text generation techniques can produce fluent natural language explanations, their performance heavily relies on dataset density and becomes constrained under sparse data conditions prevalent in real-world scenarios. To address this challenge, this paper proposes a novel explainable recommendation algorithm via historical reviews feedback modeling. The framework adopts an encoder-decoder architecture: during the feature extraction phase, it performs deep semantic modeling of user historical reviews and item-related reviews, extracting implicit preference features and attribute characteristics from reviews through attention mechanisms. In the explanation generation phase, a multi-task joint learning paradigm is employed to jointly optimize the rating prediction task and text generation task. Experimental validation on public datasets including Amazon and Yelp demonstrates that our model achieves state-of-the-art performance in both rating prediction and explanation generation tasks, showing significant improvements across most evaluation metrics. This research provides a new solution for balancing recommendation system explainability with data sparsity challenges.

Keywords

Explainable Recommendation, Historical Reviews, Feedback Modeling, Representation Learning, Transformer

1. 引言

推荐系统(Recommender System)作为个性化服务的核心技术，通过分析用户行为数据和群体相似性，构建智能化的内容筛选机制。其核心功能在于预测用户潜在需求，并广泛应用于电子商务、社交网络及数字内容平台，有效提升用户体验与商业转化率。在此基础上衍生的可解释推荐系统(Explainable Recommender System)，通过提供透明的推荐依据，实现了算法决策过程的可视化，成为当前推荐领域的重要研究方向。

可解释推荐系统的研究进展主要体现在三大技术路径的突破。首先，自然语言生成技术的突破推动了推荐解释的智能化生成。以 Transformer 为核心的序列建模技术通过自注意力机制生成连贯的文本解释[1]-[3]，变分自编码器(VAE)利用潜在变量控制生成内容的多样性[4] [5]，而扩散模型则通过迭代优化机制提升了解释文本的质量[6] [7]。这些技术的协同应用使得推荐解释具备了拟人化表达能力。

其次，大语言模型(LLMs [8])的兴起重构了解释生成的范式架构。以 GPT 系列[9]-[11]和 BERT [12]为代表的预训练模型，通过海量语料库习得的语言表征能力，能够生成符合用户认知逻辑的解释文本[13]-[18]。特别是其上下文理解能力，使得复杂场景下的推荐依据呈现多维度解析特征[19] [20]，在解释深度和语义连贯性层面实现了跨越式发展。

最后，多源信息融合策略显著增强了解释的精准性与可信度。知识图谱技术通过结构化知识网络为解释提供实体关联支持[21]-[23]，方面分析方法[24] [27]聚焦用户偏好细粒度特征，而多模态融合技术[28]则整合文本、视觉等多维度信息，构建起立体化的解释体系。这种融合创新不仅提升了推荐系统的透明度，更形成了用户需求与算法决策之间的可验证映射关系。

尽管可解释推荐领域已取得显著进展，但当前仍面临以下挑战需要突破：

a) 计算成本与模型效果的权衡：高效的推荐模型通常需要大量的计算资源，而复杂模型虽然能够提

供更佳的推荐效果，却会显著增加计算成本。因此，如何在确保推荐效果的同时降低计算成本，是当前亟待解决的关键问题[29]。

b) 数据稀疏场景下的生成瓶颈：基于 Transformer 的文本生成技术虽能产生流畅的自然语言解释，但其性能表现与训练数据密度呈现强相关性。在现实应用中普遍存在的长尾分布和稀疏交互场景下，该技术的解释生成质量存在显著退化风险，这已成为制约实际部署的重要障碍。

c) 评估体系的标准化缺失：传统评价指标(如 BLEU 等)在可解释推荐场景中存在适用性局限，难以全面评估解释的语义合理性、逻辑连贯性及用户感知价值。建立多维度、细粒度的评估框架已成为推进领域发展的迫切需求[30]。

d) 多模态需求：面对文本、图像、音频等多模态数据的处理需求，推荐系统需要构建跨模态的可解释机制，这种多模态需求增加了系统的复杂性，并使得解释过程变得更加困难[29]。

本文针对可解释推荐系统中存在的模型性能不足及数据稀疏场景下解释生成受限等挑战，做出了如下贡献：

a) 本文通过整合历史评论文本数据来增强生成模型的方法，构建基于评论编码的用户－项目潜在语义表征，有效增强推荐解释的语义相关性和个性化程度。

b) 本文设计并实现了一种基于历史评论反馈建模的可解释推荐框架 PER-HR (Personalized Explainable Recommendation via Historical Reviews)，该框架通过建立历史评论反馈机制，实现了评分预测与解释生成的协同优化。

c) 本文评估验证了所提方法的有效性。实验结果表明，与现有基线模型相比，PER-HR 在评分预测和解释生成任务上分别取得显著提升，特别是在大规模数据场景下，能有效缓解了数据稀疏性对模型性能的制约。

d) 本文揭示了通过潜评论表征和在语义空间的迁移学习，可以增强模型生成效果，有效缓解数据稀疏性对模型性能的制约。

2. 相关工作

为实现历史评论文本数据与生成模型的深度融合，本研究构建了基于反馈建模、表示学习与 Transformer 技术的增强框架。

反馈建模(Feedback Modeling)是指通过解析用户显式/隐式反馈数据构建动态偏好表征。传统上，推荐系统根据显式反馈(如评分、点击等)或隐式反馈(如浏览、购买历史等)来进行训练。然而，最新的研究扩展了反馈建模的范围，涵盖了更多细粒度的反馈信号，如情感分析结果、用户评论、时序变化等。这些反馈不仅能反映用户的偏好，还可以揭示用户对推荐结果的态度与反应。研究表明，结合用户的情感反馈与个性化的上下文信息能够进一步提升推荐的准确性和多样性。在可解释推荐系统中，理解用户的反馈有助于生成符合用户期望和需求的推荐理由。因此，本文将用户和物品的历史评论文本作为反馈信息，通过增强的反馈建模框架来进一步提高推荐解释的个性化和相关性，确保推荐理由更贴合用户的实际需求和情感。

表示学习(Representation Learning)是一种机器学习技术，旨在从原始数据中提取有效的低维特征表示，捕捉数据的潜在结构和语义。近年来，表示学习和多模态融合方法显著提升了语义表征的细粒度建模能力，如 CLIP 模型通过跨模态对齐实现了文本－图像联合嵌入。本文通过表示学习对用户和物品的历史交互进行建模，通过训练一个联合表示学习框架，捕捉用户和物品的语义关系，从而为生成推荐解释提供更加精准的信息基础。

Transformer 模型是一种基于自注意力机制的深度学习架构，可有效建模长距离依赖关系，在序列生成任务中展现出显著优势。最新的研究表明，基于 Transformer 的深度生成模型(如 GPT)可用于生成个性

化的推荐解释，该方法不仅能提高推荐系统的精准度，还能够为每个推荐提供清晰的理由和背景信息。在本文中，基于 Transformer 架构，结合用户和物品的历史评论信息，通过一个额外的编码器来处理和融合这些评论数据，提升生成推荐的质量，包括在评分预测和解释生成任务上。

上述三项核心技术在本研究中形成有机整体：引入用户–物品的历史评论文本作为反馈信息，使用表示学习捕捉用户–物品特征和文本语义特征，增强 Transformer 生成模型在评分预测和解释生成任务中的效果。

3. 研究方法

在可解释推荐系统研究中，评分预测与解释生成构成核心研究维度。形式化定义如下：给定用户集合 $\mathcal{U}$ 与物品集合 $\mathcal{I}$，评分预测任务致力于构建模型 $\mathcal{M}_r$ 以建模用户 $u \in \mathcal{U}$ 与未交互物品 $i \in \mathcal{I}$ 的潜在关联，进而预测离散评分 $\hat{r}_{u,i} \in \{1,2,3,4,5\}$，其优化目标通常为最小化预测值 $\hat{r}_{u,i}$ 与真实评分 $r_{u,i}$ 之间的差异，如公式(1)所示：

$$\min \sum_{(u,i)\in\mathcal{D}} \left|\hat{r}_{u,i} - r_{u,i}\right| \tag{1}$$

另一方面，解释生成任务需要构建模型 $\mathcal{M}_e$ 以产生自然语言解释序列 $\hat{E}_{u,i} = \langle \hat{e}_1, \hat{e}_2, \cdots, \hat{e}_n \rangle$。该过程需满足语义连贯性与文本可读性双重约束，同时要求解释文本 $\hat{E}_{u,i}$ 能够有效涵盖用户–物品特征子集 $F_{u,i} \subseteq \mathcal{V}$，其中 $\mathcal{V}$ 为预定义词汇表。具体而言，特征集合 $F_{u,i}$ 中的每个特征项 $f_k \in F_{u,i}$ 以及解释序列中的每个词汇单元 $\hat{e}_t \in \hat{E}_{u,i}$ 均需满足 $f_k, \hat{e}_t \in \mathcal{V}$，如式(2)所示：

$$F_{u,i} \cup \{\hat{e}_t\}_{t=1}^{n} \subseteq \mathcal{V} \tag{2}$$

本研究所涉关键符号体系及其语义定义详见表 1。需特别说明的是，评分预测与解释生成虽为独立建模任务，但通过共享特征空间与协同优化机制，可实现推荐系统准确性(accuracy)与可解释性(interpretability)的联合提升。

Table 1. Key symbols and their descriptions
表 1. 关键符号及其说明

符号	说明
$\mathcal{T}$	训练集
$\mathcal{T}_e$	测试集
$\mathcal{U}$	用户集合
$\mathcal{I}$	物品集合
$\mathcal{V}$	词汇表
$\mathcal{F}$	特征集合
$\mathcal{E}$	解释集合
u，i	一组用户–物品对，u 代表用户，i 代表物品
$r_{u,i}$	用户 u 对物品 i 的真实评分
$c_{u,i}$	在整个词汇表中，与用户 u 对物品 i 相关联词汇的概率分布
$F_{u,i}$	与用户 u 对物品 i 相关的特征集合
$E_{u,i}$	用户 u 对物品 i 的解释性文本序列
$\mathrm{softmax}(\cdot)$	Softmax 函数
W	权重矩阵
b	偏置参数
$\mathcal{L}$	损失

3.1. 算法架构

本研究所提出的算法框架架构整体由编码器驱动的特征提取模块和解码器驱动的生成模块构成。现有研究[1] [2]已证实，通过引入上下文任务的联合训练机制，可显著增强解释文本的生成质量。

在特征提取模块中，分别对用户 u 的历史评论文本 $review_u$ 与物品 i 的历史评论文本 $review_i$ 进行特征建模。首先，将 $review_u$ 和 $review_i$ 中的记录拼接起来，对评论文本序列执行动态截断或填充<pad>标记以保持统一维度。通过堆叠式编码器对融合特征进行深度表征学习得到 mem，其数学表达如公式(3) (4)所示。

$$info = \text{Fuse}\left(\left[review_u, review_i\right]\right) \tag{3}$$

$$mem = \text{Encoder}\left(info\right) \tag{4}$$

在生成模块中，采用多任务协同架构，整合用户特征嵌入 u、物品特征嵌入 i 及历史词元序列 $e_1, e_2, \cdots, e_t$ 进行联合建模。基于自回归生成机制(Autoregressive Generation)，该模块通过分解式解码过程实现评分预测与文本生成：首先，拼接用户/物品嵌入与词元序列，并注入位置编码信息，然后利用论文[1]中的掩码机制 $mask$ 实现时序感知解码，输出三通道特征 x_r、x_c 和 x_e。

x_r 经由多层感知机(MLP)实现非线性映射得到预测评分 $\hat{r}_{u,i}$，x_c 经由单层线性层和 Softmax 层，来估计解释中出现的所有单词的概率分布，x_e 经由单层线性层、Softmax 层，生成词汇表 V 中每个单词作为下一单词 e_{t+1} 的概率分布。其数学表达如公式(5)~(9)所示。

$$tgt = \text{PosEncoder}\left(\left[u, i, \langle bos\rangle, E_{u,i}^{t}\right]\right) \tag{5}$$

$$\left[x_r, x_c, x_e\right] = \text{Decoder}\left(tgt, mem, mask\right) \tag{6}$$

$$\hat{r}_{u,i} = \text{MLP}\left(x_r\right) \tag{7}$$

$$\hat{c}_{u,i} = \text{softmax}\left(W_c x_c + b_c\right) \tag{8}$$

$$\hat{E}_{u,i}^{t+1} = \text{softmax}\left(W_c x_e + b_c\right) \tag{9}$$

其中，$\langle bos\rangle$ 是起始标志，W_c、W_e、b_c 和 b_e 是可训练参数。

整个过程的伪代码如算法 1 所示。

算法 1 基于历史评论反馈建模的可解释推荐算法

输入：历史评 $review_u$，$review_i$，用户 u，物品 i，输入文本序列 $E_{u,i}^{t} = [e_1, e_2, \cdots, e_t]$，掩码矩阵 $mask$，模型参数 W_c，b_c，W_e，b_e

输出：评分 $r_{u,i}$，上下文分布 $c_{u,i}$，输出文本序列 $E_{u,i}^{t+1} = [e_1, e_2, \cdots, e_t, e_{t+1}]$

$info \leftarrow \text{Fuse}\left(\left[review_u, review_i\right]\right)$

$src \leftarrow \text{Embedding}_w\left(info\right)$

$mem \leftarrow \text{Encoder}\left(src\right)$

$tgt \leftarrow \left[\text{Embedding}_u\left(u\right), \text{Embedding}_i\left(i\right), \text{Embedding}_w\left(\left[\langle bos\rangle, E_{u,i}^{t}\right]\right)\right]$

$tgt \leftarrow \text{PosEncoder}\left(tgt\right)$

$out \leftarrow \text{Decoder}\left(tgt, mem, mask\right)$

续表

7: $r_{u,i} \leftarrow \mathrm{MLP}(out[0])$

8: $c_{u,i} \leftarrow \mathrm{softmax}(W_c out[1] + b_c)$

9: $E_{u,i}^{t+1} \leftarrow \mathrm{argmax}\left(\mathrm{softmax}\left(W_e out[2..] + b_e\right)\right)$

10: RETURN $r_{u,i}, c_{u,i}, E_{u,i}^{t+1}$

3.2. 损失函数

本研究的模型训练采用基于标准反向传播算法的优化策略。为实现多任务协同学习，构建如公式(10)~(13)所示的多任务联合损失函数，该函数由三个预测目标的加权组合构成：

$$\mathcal{L}_{\mathrm{r}} = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} \left(\hat{r}_{u,i} - r_{u,i}\right)^2 \tag{10}$$

$$\mathcal{L}_{\mathrm{c}} = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} -\log \hat{c}_{u,i}^{e_t} \tag{11}$$

$$\mathcal{L}_{\mathrm{e}} = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} -\log \hat{e}_t^{e_t} \tag{12}$$

$$\mathcal{L} = \lambda_r \mathcal{L}_{\mathrm{r}} + \lambda_c \mathcal{L}_{\mathrm{c}} + \lambda_e \mathcal{L}_{\mathrm{e}} \tag{13}$$

其中，$\mathcal{T}$ 表示训练样本集合，$|\cdot|$ 表示集合中元素的数量，e_t 对应真实解释文本序列 $E_{u,i}$ 中的第 t 个词元。$\mathcal{L}_{\mathrm{r}}$ 为评分预测任务的均方误差损失项，$\mathcal{L}_{\mathrm{c}}$ 度量解释文本生成的交叉熵损失，$\mathcal{L}_{\mathrm{e}}$ 则计算文本元素预测的负对数似然损失。超参数 λ_r 、λ_c 和 λ_e 分别对应各任务的加权系数，通过动态调整这些超参数实现多任务间的平衡优化，从而提升模型的综合性能表现。

4. 实验设计

4.1. 数据集与实验细节

本文实验采用来自 Yelp (餐饮)、Amazon (电影与电视)及 TripAdvisor (酒店)三大领域的公开可解释推荐基准数据集[31]。在数据预处理阶段，通过筛选确保用户与物品实体均具备不少于一条交互记录，有效保障数据完整性。每条数据样本包含五元组结构：用户 ID、物品 ID、用户评分(1~5 星)、解释文本和特征信息。其中解释文本来源于用户评论的语义片段，经人工标注验证至少包含一个与推荐决策相关的特征项。

Table 2. Key symbols and their descriptions
表 2. 三个数据集的统计

	Yelp	Amazon	TripAdvisor
用户总数	27,147	7,506	9,765
物品总数	20,266	7,360	6,280
记录总数	1,293,247	441,783	320,023
特征总数	7,340	5,399	5,069
用户平均记录数	47.64	58.86	32.77
物品平均记录数	63.81	60.02	50.96
解释文本平均长度	12.32	14.14	13.01
稀疏率(%)	99.76	99.20	99.48

如表 2 所示，三大数据集呈现差异化统计特性。以 Yelp 数据集为例，其用户与物品基数较大，但平均交互频次较低，呈现出显著的数据稀疏性特征，适用于验证模型在长尾分布场景下的鲁棒性。Amazon 数据集则表现出较高的交互密度，稀疏性指标优于其他数据集。TripAdvisor 数据集在用户评分分布与解释文本的聚焦性方面具有均衡特性，为模型性能评估提供了中间态测试环境。

实验采用分层随机划分策略，按 8:1:1 比例划分为训练集、验证集与测试集。具体而言，训练集用于参数优化，验证集负责超参数调优与早停机制触发，测试集则用于最终性能评估。为降低随机性影响，每个实验配置均执行五次独立重复实验，结果取均值。

在超参数选择上，嵌入向量维度设为 512，评分预测、上下文预测和解释生成任务的权重分别设置为 1.0、1.0 和 0.1。纳入编码的评论个数设为 20，生成的句子长度设为 15，初始学习率为 1.0。每经过一个 epoch 达到最小损失时，学习率将调整为当前值的 0.25。当连续 5 个 epoch 的总损失未出现显著下降时，训练将提前终止。

4.2. 评估指标

为了全面评估可解释推荐的性能，本文采用 RMSE 和 MAE 两项指标来衡量评分预测任务的效果。MAE 用于评估模型预测值与实际值之间的绝对误差。其计算方式如公式(14)所示：

$$\mathrm{MAE}=\frac{1}{N}\sum_{u,i\in\mathcal{T}_e}\left|r_{u,i}-\hat{r}_{u,i}\right| \tag{14}$$

其中，N 为测试样本的总数。

RMSE 通过对误差的平方取平均后开方，强调较大误差的影响。其计算方式如公式(15)所示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{u,i\in\mathcal{T}_e}\left(r_{u,i}-\hat{r}_{u,i}\right)^2} \tag{15}$$

在解释生成任务中，本文从可解释性和文本质量两个角度评估模型的性能。在可解释性方面，采用特征匹配率(FMR)、特征覆盖率(FCR)和特征多样性(DIV) [31]。

FMR 衡量生成的解释文本中包含真实文本特征的比率，计算方式如公式(16) (17)所示：

$$\mathrm{FMR}=\frac{1}{N}\sum_{u,i}\delta\left(f_{u,i}\in\hat{E}_{u,i}\right) \tag{16}$$

$$\delta(x)=\begin{cases}1 & \text{if } x=\text{true}\\ 0 & \text{if } x=\text{false}\end{cases} \tag{17}$$

其中，对于每个用户–物品对u,i，$\hat{E}_{u,i}$表示模型生成的解释文本，$f_{u,i}$表示真实特征。

FCR 衡量生成的解释文本中所包含特征的数量占真实特征总量的比例，计算方式如公式(18)所示：

$$\mathrm{FCR}=\frac{\left|\hat{\mathcal{F}}\cap\mathcal{F}\right|}{\left|\mathcal{F}\right|} \tag{18}$$

其中，$\mathcal{F}$ 表示真实解释文本中特征的集合，$\hat{\mathcal{F}}$ 表示生成解释文本中特征的集合。

DIV 衡量不同生成文本之间特征的交集，其计算方式如公式(19)所示：

$$\mathrm{DIV}=\frac{2}{N\times(N-1)}\sum_{u,u',i,i'}\left|\hat{F}_{u,i}\cap\hat{F}_{u',i'}\right| \tag{19}$$

其中，$\hat{F}_{u,i}$和$\hat{F}_{u',i'}$表示两个生成的文本包含的特征集合。

在文本质量评估方面，本论文采用了多种指标，包括 USR [31]、BLEU-1、BLEU-4 [32]、ROUGE-1 和 ROUGE-2 [33]的精确率(Precision)、召回率(Recall)、F1 分数来衡量解释生成的效果。

BLEU (Bilingual Evaluation Understudy) [32]是一种广泛应用于生成句子质量评估的指标，其核心思想是通过计算生成句子与参考句子之间 n-gram 的匹配程度来衡量翻译或生成任务的表现。BLEU 的计算过程主要包括以下几个步骤：

a) 计算 n-gram 精确度(Precision)：首先，针对每个 n-gram，计算生成句子 $\hat{E}$ 中每个 n-gram 与参考句子 E 中 n-gram 的匹配数量。然后，计算生成句子中所有 n-gram 的精确度。

b) 为避免生成句子过短导致 n-gram 精确度偏高，BLEU 引入了惩罚因子 BP (Brevity Penalty)，其计算方式如公式(20)所示：

$$\mathrm{BP}=\begin{cases}\mathrm{e}^{1-\frac{|\hat{E}|}{|E|}}, & \text{if } \dfrac{|\hat{E}|}{|E|}\le 1\\ 1, & \text{if } \dfrac{|\hat{E}|}{|E|}>1\end{cases} \tag{2}$$

c) 计算 BLEU 分数：将不同 n-gram 的精确度和 BP 惩罚因子结合，BLEU 的计算方式如公式(21)所示：

$$\mathrm{BLEU}=\mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N}w_n\log P_n\right) \tag{21}$$

其中，P_n 表示第 n 阶 n-gram 的精确度，w_n 为权重，通常取 $w_n=1/N$ 。

BLEU-1 和 BLEU-4 分别代表仅考虑 1-gram 精确度和考虑 1-gram 到 4-gram 精确度的 BLEU 得分，其计算方式如公式(22) (23)所示：

$$\mathrm{BLEU\text{-}1}=\mathrm{BP}\cdot P_1 \tag{22}$$

$$\mathrm{BLEU\text{-}4}=\mathrm{BP}\cdot\exp\left(\frac{1}{4}\sum_{i=1}^{4}\log P_i\right) \tag{23}$$

BLEU 分数越高，表示生成句子 $\hat{E}$ 与参考句子 E 之间的相似度越高，生成质量也越好。

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [33]是文本摘要任务中常用的评价指标，用于衡量生成的句子 $\hat{E}$ 与原始句子 E 之间的相似度。ROUGE 的主要度量包括精确度(Precision)、召回率(Recall)和 F1 分数。它们的计算过程如下：

a) 精确度(Precision)：精确度衡量生成的句子 $\hat{E}$ 中与原始句子 E 相匹配的部分占 $\hat{E}$ 总长度的比例，计算方式如公式(24)所示：

$$\mathrm{Precision}=\frac{|\hat{E}\cap E|}{|\hat{E}|} \tag{24}$$

b) 召回率(Recall)：召回率衡量原始句子 E 中与生成句子 $\hat{E}$ 相匹配的部分占 E 总长度的比例，计算方式如公式(25)所示：

$$\mathrm{Recall}=\frac{|\hat{E}\cap E|}{|E|} \tag{25}$$

c) F1 分数(F1-Score)：F1 分数是精确度和召回率的调和平均，用于综合评估这两者的平衡性，计算方式如公式(26)所示：

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

其中，$\left|\hat{E} \cap E\right|$表示生成的句子 $\hat{E}$ 和原始句子 E 之间匹配的词语或片段数量，$\left|\hat{E}\right|$和$|E|$分别表示生成句子和原始句子的长度(通常以词或 n-gram 的数量来衡量)。

然而，BLEU 与 ROUGE 主要依靠词汇匹配来评估生成文本的质量，这使得它们在检测相同句子方面存在一定局限性。因此，在评估生成文本的个性化程度时，BLEU 与 ROUGE 的表现较为有限。为了更好地评估个性化生成文本，本论文引入了 USR (Unique Sentence Ratio)指标来计算生成句子的唯一性比率，计算方式如公式(27)所示：

$$\text{USR} = \frac{\left|\hat{\mathcal{E}}\right|}{N} \tag{27}$$

其中 $\hat{\mathcal{E}}$ 表示模型生成的句子集合。

4.3. 基线算法

本文选择了两组基线算法，分别用于评分预测任务和解释生成任务的自动评估。在评分预测任务中，采用基于分解的 PMF [34]、SVD++ [35]和两种基于深度神经网络的方法 NRT [36]和 PETER [1]作为比较基准。在解释生成方面，采用 NRT、Att2Seq [37]、PETER 和 PEPLER [18]。以下是几种基线算法的介绍。

a) PMF：一种基于概率矩阵分解的协同过滤方法，通过分解用户－物品评分矩阵来预测评分。

b) SVD++：在传统 SVD 的基础上，引入了用户隐式反馈，增强了对用户偏好的捕捉能力。

c) NRT：一种神经网络模型，用于生成个性化评分预测和相关提示。

d) Att2Seq：一种生成个性化推荐文本的序列到序列模型，常用于生成推荐解释。

e) PETER：一种基于 Transformer 的多任务学习模型，能够同时进行评分预测和解释生成。

f) PEPLER：一种基于 GPT-2 的个性化推荐生成模型，通过微调 GPT-2 来生成个性化解释。

5. 实验结果与分析

5.1. 评分预测

Table 3. Rating prediction results, with the best performance values in bold
表 3. 评分预测结果，最佳性能值以**粗体**标出

	Yelp		Amazon		TripAdvisor	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
PMF	1.09	0.88	1.03	0.81	0.87	0.70
SVD++	1.01	0.78	0.96	0.72	0.80	**0.61**
NRT	1.01	0.78	0.95	0.70	**0.79**	**0.61**
PETER	1.01	0.78	0.95	0.71	0.81	0.63
PER-HR	**0.99**	**0.76**	**0.88**	**0.66**	0.81	0.63

本文使用了均方根误差(RMSE)和平均绝对误差(MAE)两个常见指标对多种推荐算法的性能进行了评估。表 3 展示了 PMF、SVD++、NRT、PETER 和 PER-HR 五种模型在 Yelp、Amazon 和 TripAdvisor

三个数据集上的表现，最佳性能值以粗体标出。

在 Yelp 数据集上，PER-HR 模型在 RMSE (0.99)和 MAE (0.76)方面均表现最佳，明显优于其他模型。相比之下，PMF 模型的性能较差，RMSE 为 1.09，MAE 为 0.88，误差较大。这表明，PER-HR 模型在该数据集上能够更好地拟合用户偏好，提供更准确的推荐。

在 Amazon 数据集上，PER-HR 模型再次表现出色，RMSE 为 0.88，MAE 为 0.66，均为最低值，优于其他所有模型。NRT 和 PETER 模型的表现相似，RMSE 分别为 0.96 和 0.95，MAE 值也较高。PER-HR 模型在此数据集上的优势尤其明显，表明它能够有效捕捉用户的需求和物品的特征。

在 TripAdvisor 数据集上，尽管 PER-HR 并未在 RMSE 和 MAE 上都取得最优成绩，但其表现依然非常竞争力，RMSE 为 0.81，MAE 为 0.63，与其他模型相差无几。

综上所述，PER-HR 模型在 Yelp 和 Amazon 数据集上表现突出，无论是在 RMSE 还是 MAE 指标上均为最佳。尽管在 TripAdvisor 数据集上它并非绝对最好，但其在 MAE 上依然领先，表明 PER-HR 在推荐任务中的表现稳定且可靠。

5.2. 解释生成

表 4 展示了在三个数据集(Yelp、Amazon 和 TripAdvisor)上，对不同推荐算法在可解释性和文本质量方面的性能进行了比较。评估指标包括 FMR (可解释性评分)、FCR (可解释性评分)、DIV (多样性，越小越好)、USR (用户满意度)，以及 BLEU-1 (B1)、BLEU-4 (B4)、ROUGE-1 (R1)和 ROUGE-2 (R2)的精度(P)、召回率(R)和 F1 值(F)。其中，BLEU 和 ROUGE 指标的结果以百分比表示，其他指标为绝对值。

在可解释性方面，PER-HR 在所有三个数据集上均展现了优异的表现。具体来说，FMR 和 FCR 两个指标显示了 PER-HR 在生成高质量解释方面的显著优势。特别是在 Yelp 和 Amazon 数据集上，PER-HR 的 FMR 分别为 0.23 和 0.20，FCR 分别为 0.66 和 0.73，明显超过了其他算法。此外，PER-HR 在多样性(DIV)指标上也表现突出，尤其在 Yelp (1.10)和 Amazon (1.43)数据集上，其生成的推荐解释具有较高的多样性，避免了重复或单一的解释。

在文本质量方面，PER-HR 在 BLEU 和 ROUGE 指标上均取得了最佳成绩。作为衡量生成文本与参考文本相似度的标准，BLEU-1 和 BLEU-4 显示了 PER-HR 在文本生成方面的强大能力。在 Yelp 数据集上，PER-HR 的 BLEU-1 为 28.66，远高于其他方法，而 BLEU-4 为 12.18，进一步显示了其卓越的生成性能。此外，ROUGE-1 和 ROUGE-2 指标用于衡量生成文本与参考文本的重叠程度，PER-HR 在这两个指标上也表现优异。特别是在 Yelp 和 Amazon 数据集上，PER-HR 的 ROUGE 得分远超其他方法，表明其生成的文本质量较高，具备较强的表达能力。

虽然其他算法如 NRT、Att2Seq、PETER 和 PEPLER 在某些指标上也取得了一定成绩，但总体而言，PER-HR 在可解释性和文本质量的多个维度上均表现出色。在 Yelp 数据集上，PER-HR 在可解释性和文本质量方面的提升尤其显著，特别是在 FCR (0.66)和 DIV (1.10)上，明显高于其他方法，并且 BLEU 和 ROUGE 得分均有明显提高。在 Amazon 数据集上，尽管 PER-HR 的 BLEU-4 (8.08)较低，但其在其他指标，尤其是 FCR 和 ROUGE 指标上的优异表现，证明其生成的推荐解释具有较高的实用价值。在 TripAdvisor 数据集上，PER-HR 同样在 FCR (0.66)和 DIV (1.63)上表现出色，尽管其 BLEU-4 (6.47)相较于 Yelp 略逊一筹，说明其生成文本的多样性和质量在不同数据集上有所变化。

综合来看，PER-HR 在三个数据集上的整体表现均优于其他对比算法，特别是在 Yelp 和 Amazon 数据集上，展现了最为卓越的可解释性和文本质量。无论是在 FMR、FCR 还是 BLEU 和 ROUGE 等指标上，PER-HR 都表现出了明显的优势，证明其不仅能够生成高质量的推荐文本，还能提供丰富且多样的个性化解释。因此，PER-HR 在可解释推荐任务中具有较大的应用潜力，尤其适用于那些需要高质量推荐解

释的实际场景。

Table 4. Comparison of text quality, with the best performance values in bold. B1 and B4 represent BLEU-1 and BLEU-4. R1-P, R1-R, R1-F, R2-P, R2-R, and R2-F represent the precision, recall, and F1 value of ROUGE-1 and ROUGE-2. The results of BLEU and ROUGE are expressed in percentages (% omitted), and the other indicators are expressed in absolute values
表 4. 解释生成结果。评估指标 B1 和 B4 分别代表 BLEU-1 和 BLEU-4。R1-P、R1-R、R1-F、R2-P、R2-R 和 R2-F 分别表示 ROUGE-1 和 ROUGE-2 的精度(Precision)、召回率(Recall)和 F1 值。BLEU 和 ROUGE 的结果以百分比表示(省略%)，其他指标以绝对值表示。最佳性能值用**粗体**标出

	可解释性			文本质量								
	FMR	FCR	DIV ↓	USR	B1	B4	R1-P	R1-R	R1-F	R2-P	R2-R	R2-F
						Yelp						
NRT	0.07	0.11	2.37	0.12	11.66	0.65	17.69	12.11	13.55	1.76	1.22	1.33
Att2Seq	0.07	0.12	2.41	0.13	10.29	0.58	18.73	11.28	13.29	1.85	1.14	1.31
PETER	0.08	0.19	1.54	0.12	10.64	0.71	18.57	12.1	13.7	2.02	1.35	1.47
PEPLER	0.08	0.3	1.52	0.35	11.23	0.73	17.51	12.55	13.53	1.86	1.42	1.46
PER-HR	**0.23**	**0.66**	**1.10**	**0.42**	**28.66**	**12.18**	**33.43**	**28.31**	**29.73**	**14.25**	**13.58**	**13.74**
						Amazon						
NRT	0.12	0.07	2.93	0.17	12.93	0.96	21.03	13.57	15.56	2.71	1.84	2.05
Att2Seq	0.12	0.2	2.74	0.33	12.56	0.95	20.79	13.31	15.35	2.62	1.78	1.99
PETER	0.12	0.15	1.89	0.22	12.97	1.15	20.08	13.95	15.43	2.82	2.1	2.22
PEPLER	0.11	0.27	2.06	0.38	13.19	1.05	18.51	14.16	14.87	2.36	1.88	1.91
PER-HR	**0.20**	**0.73**	**1.43**	**0.44**	**24.32**	**8.08**	**29.68**	**24.09**	**25.6**	**10.25**	**9.27**	**9.51**
						TripAdvisor						
NRT	0.06	0.09	4.27	0.08	15.05	0.99	18.22	14.39	15.4	2.29	1.98	2.01
Att2Seq	0.06	0.15	4.32	0.17	15.27	1.03	18.97	14.72	15.92	2.4	2.03	2.09
PETER	0.07	0.11	3.1	0.06	15.98	1.1	18.9	16.02	16.37	2.33	2.17	2.1
PEPLER	0.07	0.21	2.71	0.24	15.49	1.09	19.48	15.67	16.24	2.48	2.21	2.16
PER-HR	**0.13**	**0.66**	**1.63**	**0.30**	**21.01**	**6.47**	**23.53**	**21.12**	**21.37**	**7.43**	**7.13**	**7.14**

5.3. 结果分析

PER-HR 算法在评分预测和解释生成方面展现了卓越的性能。具体来说，PER-HR 在三种数据集(Yelp、Amazon 和 TripAdvisor)上都表现出了显著的优势，不仅能够提供准确的评分预测，还能生成高质量的推荐解释。其评分预测效果和解释生成效果在多个指标上均超过了对比算法。

PER-HR 在不同数据集上的表现呈现出与其他算法不同的规律。其他算法通常表现出与数据稀疏性成反比的规律，即数据越稀疏，预测效果越差。然而，PER-HR 的表现与数据规模成正比，数据记录越多，其评分预测和解释生成效果越好。这一现象表明，PER-HR 能够充分利用大规模数据中的信息，从而生成更加精准和多样的推荐解释。

在 Yelp 和 Amazon 数据集上，PER-HR 充分展现了其在大规模数据上的优势，随着数据量的增加，算法的效果得到了显著提升。相比之下，其他算法的表现随着数据量的增多并未呈现出相同的正向变化，尤其在数据规模大，数据稀疏时，效果下降明显。

6. 总结与展望

本研究提出了 PER-HR 模型，并通过在多个数据集上的实验验证了其在推荐任务中的优越性能。通过综合评估评分预测与解释生成两个方面，本文得出以下几个主要结论。

首先，PER-HR 在评分预测方面的表现显著优于其他对比算法。在 Yelp 和 Amazon 数据集上，PER-HR 在 RMSE 和 MAE 指标上均取得了最佳成绩，表明该模型能够更准确地预测用户的评分偏好。在 TripAdvisor 数据集上，虽然其在评分预测方面未能绝对领先，但依然表现稳健，展示了其在不同数据集中的稳定性和可靠性。

其次，PER-HR 在解释生成方面的优势尤为突出。通过多种评估指标的对比，PER-HR 在可解释性和文本质量方面均表现出了领先地位。在生成推荐解释时，PER-HR 不仅能够提供个性化且多样化的解释，还能在 BLEU 和 ROUGE 指标上取得显著优势，证明其生成的文本质量具有较高的参考价值，能够有效满足用户对推荐解释的需求。

此外，PER-HR 的表现与数据规模密切相关。在大规模数据集上，PER-HR 的评分预测和解释生成能力均得到了明显提升，这一特性使得该模型在处理大规模用户数据时具有显著优势。相较于其他算法，PER-HR 能够更好地利用数据中蕴含的信息，从而在数据丰富的场景中实现更优的推荐效果。

综上所述，PER-HR 在多个维度上均展示了其作为可解释推荐系统的潜力。其优异的性能不仅提升了推荐的准确性，还增强了推荐解释的可读性和实用性。因此，PER-HR 模型在实际应用中，尤其是需要高质量个性化推荐和解释的场景中，具有广阔的应用前景。

未来工作可以从以下几个方向展开深入研究：首先，探讨如何有效处理历史评论数据中的噪声、偏差与主观性问题，通过引入更为精细的数据预处理、噪声过滤和偏差校正方法，进一步提升解释生成的准确性和可信度。其次，针对当前生成文本在长度、风格和内容上的控制不足，未来将设计灵活的生成控制策略，旨在实现解释内容的个性化和多样化，以更好地满足用户不同场景下的需求。此外，还将探索跨域知识迁移和多模态数据融合等新技术，以增强模型在大规模及多样化数据环境中的泛化能力。

致　谢

感谢国家自然科学基金项目(61803264)对本研究的支持与资助。

基金项目

国家自然科学基金项目(61803264)。

参考文献

[1] Li, L., Zhang, Y. and Chen, L. (2021) Personalized Transformer for Explainable Recommendation. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing* (*Volume* 1: *Long Papers*), 1-6 August 2021, 4947-4957. https://doi.org/10.18653/v1/2021.acl-long.383

[2] Raczynski, J., Lango, M. and Stefanowski, J. (2023) The Problem of Coherence in Natural Language Explanations of Recommendations. *ECAI* 2023—26*th European Conference on Artificial Intelligence*, Kraków, 30 September-4 October 2023, 1922-1929.

[3] Liu, Z., Ma, Y., Schubert, M., Ouyang, Y., Rong, W. and Xiong, Z. (2024) Multimodal Contrastive Transformer for Explainable Recommendation. *IEEE Transactions on Computational Social Systems*, **11**, 2632-2643. https://doi.org/10.1109/tcss.2023.3276273

[4] Wang, L., Cai, Z., De Melo, G., Cao, Z. and He, L. (2023) Disentangled CVAEs with Contrastive Learning for Explainable Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 13691-13699. https://doi.org/10.1609/aaai.v37i11.26604

[5] He, M., An, B., Wang, J. and Wen, H. (2023) CETD: Counterfactual Explanations by Considering Temporal Dependencies in Sequential Recommendation. *Applied Sciences*, **13**, Article 11176. https://doi.org/10.3390/app132011176

[6] Guo, Y., Cai, F., Chen, H., Chen, C., Zhang, X. and Zhang, M. (2023) An Explainable Recommendation Method Based on Diffusion Model. 2023 9*th International Conference on Big Data and Information Analytics* (*BigDIA*), Haikou, 15-17 December 2023, 802-806. https://doi.org/10.1109/bigdia60676.2023.10429319

[7] Li, L., Li, S., Marantika, W., *et al*. (2024) Diffusion-EXR: Controllable Review Generation for Explainable Recommendation via Diffusion Models. arXiv: 2312.15490.

[8] Minaee, S., Mikolov, T., Nikzad, N., *et al*. (2024) Large Language Models: A Survey. arXiv: 2402.06196.

[9] Radford, A., Wu, J., Child, R., *et al*. (2019) Language Models Are Unsupervised Multitask Learners. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

[10] Brown, T.B., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners. *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, Red Hook, 6-12 December 2020, 1-25.

[11] Openai, A.J., Adler, S., *et al*. (2024) GPT-4 Technical Report. arXiv: 2303.08774. https://arxiv.org/abs/2303.08774

[12] Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Under-Standing. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, June 2019, 4171-4186. https://doi.org/10.18653/v1/n19-1423

[13] Luo, Y., Cheng, M., Zhang, H., Lu, J. and Chen, E. (2024) Unlocking the Potential of Large Language Models for Explainable Recommendations. In: Onizuka, M., *et al*., Eds., *Database Systems for Advanced Applications*, Springer, 286-303. https://doi.org/10.1007/978-981-97-5569-1_18

[14] Chu, Z., Wang, Y., Cui, Q., *et al*. (2024) LLM-Guided Multi-View Hypergraph Learning for Human-Centric Explainable Recommendation. arXiv: 2401.08217.

[15] Lin, C., Tsai, C., Su, S., Jwo, J., Lee, C. and Wang, X. (2023) Predictive Prompts with Joint Training of Large Language Models for Explainable Recommendation. *Mathematics*, **11**, Article 4230. https://doi.org/10.3390/math11204230

[16] Peng, Q., Liu, H., Xu, H., *et al*. (2024) Review-LLM: Harnessing Large Language Models for Personalized Review Generation. arXiv: 2407.07487.

[17] Peng, Y., Chen, H., Lin, C., Huang, G., Hu, J., Guo, H., *et al*. (2024) Uncertainty-Aware Explainable Recommendation with Large Language Models. 2024 *International Joint Conference on Neural Networks* (*IJCNN*), Yokohama, 30 June-5 July 2024, 1-8. https://doi.org/10.1109/ijcnn60899.2024.10651104

[18] Li, L., Zhang, Y. and Chen, L. (2023) Personalized Prompt Learning for Explainable Recommendation. *ACM Transactions on Information Systems*, **41**, 1-26. https://doi.org/10.1145/3580488

[19] Ma, Q., Ren, X. and Huang, C. (2024) XRec: Large Language Models for Explainable Recommendation. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, 12-16 November 2024, 391-402. https://doi.org/10.18653/v1/2024.findings-emnlp.22

[20] Yang, M., Zhu, M., Wang, Y., Chen, L., Zhao, Y., Wang, X., *et al*. (2024) Fine-Tuning Large Language Model Based Explainable Recommendation with Explainable Quality Reward. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 9250-9259. https://doi.org/10.1609/aaai.v38i8.28777

[21] Lyu, Z., Wu, Y., Lai, J., *et al*. (2023) Knowledge Enhanced Graph Neural Networks for Explainable Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 4954-4968.

[22] Bastola, R. and Shakya, S. (2024) Knowledge-enriched Graph Convolution Network for Hybrid Explainable Recommendation from Review Texts and Reasoning Path. 2024 *International Conference on Inventive Computation Technologies* (*ICICT*), Lalitpur, 24-26 April 2024, 590-599. https://doi.org/10.1109/icict60155.2024.10544384

[23] Qu, X., Wang, Y., Li, Z., *et al*. (2024) Graph-Enhanced Prompt Learning for Personalized Review Generation. *Data Science and Engineering*, **9**, 309-324.

[24] Li, J., He, Z., Shang, J. and McAuley, J. (2023) UCEpic: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation. *Proceedings of the* 29*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, 6-10 August 2023, 1248-1257. https://doi.org/10.1145/3580305.3599535

[25] Cheng, H., Wang, S., Lu, W., Zhang, W., Zhou, M., Lu, K., *et al*. (2023) Explainable Recommendation with Personalized Review Retrieval and Aspect Learning. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 51-64. https://doi.org/10.18653/v1/2023.acl-long.4

[26] Hu, Y., Liu, Y., Miao, C., *et al*. (2023) Aspect-Guided Syntax Graph Learning for Explainable Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 7768-7781.

[27] Liao, H., Wang, S., Cheng, H., Zhang, W., Zhang, J., Zhou, M., *et al*. (2025) Aspect-Enhanced Explainable Recommendation with Multi-Modal Contrastive Learning. *ACM Transactions on Intelligent Systems and Technology*, **16**, 1-24. https://doi.org/10.1145/3673234

[28] Yan, A., He, Z., Li, J., Zhang, T. and McAuley, J. (2023) Personalized Showcases: Generating Multi-Modal Explanations for Recommendations. *Proceedings of the* 46*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 23-27 July 2023, 2251-2255. https://doi.org/10.1145/3539618.3592036

[29] Chatti, M.A., Guesmi, M. and Muslim, A. (2024) Visualization for Recommendation Explainability: A Survey and New Perspectives. *ACM Transactions on Interactive Intelligent Systems*, **14**, 1-40. https://doi.org/10.1145/3672276

[30] Ma, B., Yang, T. and Ren, B. (2024) A Survey on Explainable Course Recommendation Systems. In: Streitz, N.A. and Konomi, S., Eds., *Distributed, Ambient and Pervasive Interactions*, Springer, 273-287. https://doi.org/10.1007/978-3-031-60012-8_17

[31] Li, L., Zhang, Y. and Chen, L. (2020) Generate Neural Template Explanations for Recommendation. *Proceedings of the* 29*th ACM International Conference on Information & Knowledge Management*, 19-23 October 2020, 755-764. https://doi.org/10.1145/3340531.3411992

[32] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics—ACL*'02, Philadelphia, 7-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[33] Lin, C.Y. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In: Moens, M.-F. and Szpakowicz, S., Eds., *Text Summarization Branches out*, Association for Computational Linguistics, 74-81. https://www.aclweb.org/anthology/W04-1013

[34] Salakhutdinov, R. and Mnih, A. (2007) Probabilistic Matrix Factorization. *Proceedings of the* 20*th International Conference on Neural Information Processing Systems*, Vancouver, 3-6 December 2007, 1257-1264.

[35] Koren, Y. (2008) Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, 24-27 August 2008, 426-434. https://doi.org/10.1145/1401890.1401944

[36] Li, P., Wang, Z., Ren, Z., Bing, L. and Lam, W. (2017) Neural Rating Regression with Abstractive Tips Generation for Recommendation. *Proceedings of the* 40*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, 7-11 August 2017, 345-354. https://doi.org/10.1145/3077136.3080822

[37] Dong, L., Huang, S., Wei, F., Lapata, M., Zhou, M. and Xu, K. (2017) Learning to Generate Product Reviews from Attributes. *Proceedings of the* 15*th Conference of the European Chapter of the Association for Computational Linguistics*: *Volume* 1, *Long Papers*, Valencia, 3-7 April 2017, 623-632. https://doi.org/10.18653/v1/e17-1059