Disentangling Likes and Dislikes in Personalized Generative Explainable Recommendation

Ryotaro Shimizu (/profile?id=~Ryotaro_Shimizu1), Takashi Wada (/profile?id=~Takashi_Wada1), Yu Wang (/profile?id=~Yu_Wang24), Johannes Kruse (/profile?id=~Johannes_Kruse1), Sean O'Brien (/profile?id=~Sean_O%27Brien1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), Linxin Song (/profile?id=~Linxin_Song1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Yuki Saito (/profile?id=~Yuki_Saito1), Fugee Tsung (/profile?id=~Fugee_Tsung1), Masayuki Goto (/profile?id=~Masayuki_Goto1), Julian McAuley (/profile?id=~Julian_McAuley1)

06 Oct 2024 (modified: 30 Oct 2024) ACM TheWebConf 2025 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=UhPUR9cnRJ)

Track: User modeling, personalization and recommendation

Serve As Reviewer: Ryotaro Shimizu (/profile?id=~Ryotaro_Shimizu1)

Keywords: Explainable recommendation, Recommender systems, Large language model, Transformer, Personalization, Sentiment analysis

Confirmation: I certify that my OpenReview profile is up to date (including accurate first and last names, a valid preferred email, and current and past affiliations) and that this paper adheres to the guidelines in the Call for Papers, including policy on human participants, authorship, and limits on maximum authorship.

Abstract:

Recent research on explainable recommendation generally frames the task as a standard text generation problem, and evaluates models simply based on the textual similarity between the predicted and ground-truth explanations. However, this approach fails to consider one crucial aspect of the systems: whether their outputs accurately reflect the users' (post-purchase) sentiments, i.e., whether and why they would like and/or dislike the recommended items. To shed light on this issue, we introduce new datasets and evaluation methods that focus on the users' sentiments. Specifically, we construct the datasets by explicitly extracting users' positive and negative opinions from their post-purchase reviews using an LLM, and propose to evaluate systems based on whether the generated explanations 1) align well with the users' sentiments, and 2) accurately identify both positive and negative opinions of users on the target items. We benchmark several recent models on our datasets and demonstrate that achieving strong performance on existing metrics does not ensure that the generated explanations align well with the users' sentiments. Lastly, we find that existing models can provide more sentiment-aware explanations when the users' (predicted) ratings for the target items are directly fed into the models as input. We will release our code and datasets upon acceptance.

Submission Number: 1244

Discussion (/forum?id=UhPUR9cnRJ#discussion)

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Program Chairs | Submission1244 Authors | Submission1244... | Submission1244 Area... | Submission1244... | Submission1244... | Submission1244... | Submission1244... | Submission1244... | Submission1244...

37 / 37 replies shown

Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 20 Jan 2025, 04:24 (modified: 20 Jan 2025, 10:36) Program Chairs, Authors

Revisions (/revisions?id=ZINugt1oBA)

Decision: Accept (Poster)

The Two-Way Communication Period is Closing Soon

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

11 Dec 2024, 01:28 (modified: 11 Dec 2024, 16:33)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=LAwgYWKyyD)

Comment:

Dear reviewers (UUMe, rhoR, hUQ9, and Tsqz),

We sincerely appreciate your valuable feedback and have tried to address all of your concerns and questions. We kindly request your follow-up during the remaining two-way communication period and would greatly appreciate it if you could reconsider your score in light of our responses.

Official Review of Submission1244 by Reviewer UUMe

Official Review by Reviewer UUMe 02 Dec 2024, 12:04 (modified: 03 Dec 2024, 05:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors

Revisions (/revisions?id=StBWWjVY3g)

Review:

This work proposed a new benchmark (set of datasets) on the prediction of post purchase sentiments, which is kinda similar to the CTR task but with sentiments rather than simple score or regression. Having such benchmarks is definitely beneficial for the research community, and can inspire many new research under this topic. Nonetheless, I do have a couple concerns regarding this work:

1. The datasets are quite tiny, with at most 20K entities and half million interactions.

2. The ground-truth explanations are just simply outputs of LLMs via prompting.

Questions:
n/a

Ethics Review Flag: No
Scope: 3: The work is somewhat relevant to the Web and to the track, and is of narrow interest to a sub-community
Novelty: 2
Technical Quality: 3
Reviewer Confidence: 2: The reviewer is willing to defend the evaluation, but it is likely that the reviewer did not understand parts of the paper

(It seems that there is indeed an issue with the system.)

Official Comment
by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))
09 Dec 2024, 18:43 (modified: 09 Dec 2024, 21:14)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors
Revisions (/revisions?id=5paDdv1mFP)

Comment:
We sincerely appreciate your reply and decision to raise your score. We would greatly appreciate it if you could let us know which score you will raise and the revised score.

Additionally, if there are any remaining questions or concerns we could address to provide further clarification and improve the manuscript, please do not hesitate to let us know.

Official Comment by Authors

Official Comment
by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))
05 Dec 2024, 22:24 (modified: 06 Dec 2024, 01:39)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors
Revisions (/revisions?id=muJdxqRsNd)

Comment:
Thank you for your time and effort in reviewing our paper. We are sincerely grateful for your insightful feedback, which helps us refine our paper.

We have addressed all your questions with **A1** -- **A2**. Due to the character limit, we had to split our rebuttal comments into "rebuttal" and "official comment" text boxes below. We appreciate it if you take a look at all the responses and take them into account when updating your scores. Of course, if you have any further questions, please do not hesitate to ask.

Official Comment by Reviewer UUMe

Official Comment by Reviewer UUMe 09 Dec 2024, 17:56 (modified: 09 Dec 2024, 17:57)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors

Revisions (/revisions?id=M4Ow4jSJHY)

Comment:

I appreciate the rebuttal as well as the additional results, I'll raise my score when the system allows me to do so

Replying to Official Comment by Reviewer UUMe

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

09 Dec 2024, 22:30

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors

Comment:

It seems that the system now allows us to update the scores.

The Two-Way Communication Period is Closing Soon

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

10 Dec 2024, 20:17 (modified: 11 Dec 2024, 02:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors

Revisions (/revisions?id=FJsQKys9Ij)

Comment:

As the two-way communication period is closing soon, I would greatly appreciate it if you could update your current scores (from Novelty: 2, Technical Quality: 3), as you mentioned your intention.

Rebuttal by Authors 2

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:24 (modified: 06 Dec 2024, 14:17)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer UUMe, Authors

Revisions (/revisions?id=QwgMOeL08t)

Comment:

> The ground-truth explanations are just simply outputs of LLMs via prompting.

A2. To verify the dataset quality, we asked several human annotators to evaluate the subset of the datasets and confirmed that LLMs can correctly extract relevant information and users' sentiments from their reviews (see Section 3.2). Additionally, to address your concern, we further verified the dataset's quality using **Gemini-1.5-pro** and **Gemini-1.5-flash** as automatic evaluators, as well as **GPT-4o** (which we used in our paper). The results are shown below, ensuring the high quality of our datasets (the first table is reproduced from Table 6).

Table. The results of the dataset quality evaluation using **GPT-4o**. The numbers outside parentheses denote the scores estimated by GPT-4o, whereas those in parentheses indicate the percentage of the instances for which GPT-4o and human annotators make the same judgements.

Stage	Type	Amazon	Yelp	RateBeer
1	Factual	0.990 (0.95)	0.993 (0.98)	0.997 (0.95)
	Context-p	0.996 (0.98)	0.997 (0.96)	0.997 (0.98)
	Context-n	0.962 (0.97)	0.971 (0.95)	0.965 (0.97)
2	Factual-p	0.999 (1.00)	0.999 (1.00)	0.996 (1.00)
	Factual-n	0.998 (0.99)	0.998 (1.00)	0.998 (0.99)
	Complete-p	0.997 (0.99)	0.997 (1.00)	0.998 (1.00)
	Complete-n	0.998 (1.00)	0.996 (1.00)	0.998 (1.00)

Table. The results of the dataset quality evaluation using **Gemini-1.5-pro** (gemini-1.5-pro-002). The numbers outside parentheses denote the scores estimated by Gemini-1.5-pro, whereas those in parentheses indicate the percentage of the instances for which Gemini-1.5-pro and human annotators make the same judgements.

Stage	Type	Amazon	Yelp	RateBeer
1	Factual	0.994 (0.94)	0.996 (1.00)	0.997 (0.94)
	Context-p	0.998 (0.97)	0.998 (0.97)	0.998 (0.97)
	Context-n	0.995 (0.97)	0.997 (0.99)	0.994 (0.96)
2	Factual-p	0.999 (1.00)	1.000 (1.00)	1.000 (1.00)
	Factual-n	0.997 (0.97)	0.997 (1.00)	0.999 (0.98)
	Complete-p	0.997 (0.97)	0.997 (1.00)	0.998 (1.00)
	Complete-n	0.997 (0.98)	0.996 (1.00)	0.998 (1.00)

Table. The results of the dataset quality evaluation using **Gemini-1.5-flash** (gemini-1.5-flash-002). The numbers outside parentheses denote the scores estimated by Gemini-1.5-flash, whereas those in parentheses indicate the percentage of the instances for which Gemini-1.5-flash and human annotators make the same judgements.

Stage	Type	Amazon	Yelp	RateBeer
1	Factual	0.997 (0.94)	0.997 (1.00)	0.996 (0.94)
	Context-p	0.998 (0.98)	0.998 (0.97)	0.996 (0.98)
	Context-n	0.996 (0.98)	0.997 (0.99)	0.990 (0.95)
2	Factual-p	0.999 (1.00)	0.998 (1.00)	0.999 (1.00)
	Factual-n	0.993 (0.99)	0.995 (1.00)	0.996 (0.97)
	Complete-p	0.992 (0.99)	0.987 (0.99)	0.993 (1.00)
	Complete-n	0.990 (0.99)	0.983 (0.99)	0.990 (1.00)

Lastly, as we showed in Table 1, the ground-truth explanations in our datasets contain far less noise than existing datasets, which automatically generate explanations by retrieving sentences or phrases from reviews using rudimentary algorithms. For instance, the dataset used in [1] sometimes retrieves white spaces or just single characters such as "a", "b", ',' and "!" as features, and often retrieves very short phrases such as “great movie” as the explanations. On the other hand, our datasets provide more accurate and succinct explanations as well as relevant positive and negative features.

Rebuttal by Authors

Rebuttal

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:22 (modified: 06 Dec 2024, 01:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=lbu2uWsvsH)

Rebuttal:

We greatly appreciate your insightful and constructive feedback. We have addressed all your questions with **A1** -- **A2** below:

> The datasets are quite tiny, with at most 20K entities and half million interactions.

A1. Our datasets are of a similar size to existing ones in explainable recommendation, as shown in Table 14 in Appendix. The following tables compare the statistics of our datasets and previous ones, demonstrating that our datasets are not tiny:

Table: Statistics of the existing datasets [1].

	Amazon	**Yelp**	**Tripadvisor**
#users	7,506	27,147	9,765
#items	7,360	20,266	6,280
#interactions	**441,783**	**1,293,247**	**320,023**

Table: Statistics of the datasets constructed in our work.

	Amazon	**Yelp**	**RateBeer**
#users	7,445	11,780	2,743
#items	7,331	10,148	7,453
#interactions	**438,604**	**504,184**	**512,370**

Additionally, we show the statistics of other datasets used in previous works in the following tables:

Table: Statistics of the existing datasets [2].

	Amazon	**Yelp**	**Google**
#users	15,349	15,942	22,582
#items	15,247	14,085	16,557
#interactions	**360,839**	**393,680**	**411,840**

Table: Statistics of the existing datasets [3].

	Cellphones	**Clothings**	**CDs & Vinyls**
#users	27,879	39,387	75,258

	Cellphones	Clothings	CDs & Vinyls
#items	10,429	23,033	64,443
#interactions	**194,439**	**278,677**	**1,097,592**

Table: Statistics of the existing datasets [4].

	Google	Amazon	Yelp
#users	19,973	10,457	5,219
#items	15,863	17,076	15,500
#interactions	**167,242**	**95,855**	**37,751**

[1] Lei Li, et al. 2020. Generate Neural Template Explanations for Recommendation. In Proceedings of CIKM. 755–764.

[2] Qiyao Ma, et al. 2024. XRec: Large Language Models for Explainable Recommendation. In Findings of EMNLP, 391–402.

[3] Sung-Jun Park, et al. 2022. Reinforcement Learning over Sentiment-Augmented Knowledge Graphs towards Accurate and Explainable Recommendation. In Proceedings of WSDM, 784–793.

[4] F. Xie, et al. 2024. A Review-Level Sentiment Information Enhanced Multitask Learning Approach for Explainable Recommendation. IEEE Transactions on Computational Social Systems, 11 (5), 5925-5934.

*Due to the character limit, the further rebuttal comments will be shared as an "official comment."

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:23 (modified: 05 Dec 2024, 23:27)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=lDL5p24Lyw)

[Deleted]

Official Review of Submission1244 by Reviewer rhoR

Official Review by Reviewer rhoR 02 Dec 2024, 11:52 (modified: 03 Dec 2024, 05:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=DaQg42VEw9)

Review:

This paper highlights that current text-based evaluation methods for explainable recommendations are hard to capture users' sentimental preferences, such as liking or disliking specific aspects of items. To address this, the authors propose a new evaluation method that leverages Large Language Models (LLMs) to generate review summaries and identify personalized positive and negative opinions about items. The authors also plan to release a new dataset if the paper is accepted. Experiments show that their method outperforms commonly used metrics like BLEU, ROUGE, and BERTScore in distinguishing explainable recommendation models.

Pros:

- The authors promise to release the code and datasets upon acceptance and include the prompts in the Appendix, enhancing reproducibility.
- The paper involves a relevant and important topic: evaluating recommendation explanations with a focus on user preferences, which is likely to attract significant interest in the research community.
- Human annotators are invited in assessing the outputs generated by LLMs, ensuring the reliability of the experimental results.
- The authors conducted multiple experiments to demonstrate the effectiveness of each component in their proposed evaluation method.

Cons:

- The paper could benefit from addressing some specific questions and suggestions, as detailed in the Questions Section.

Questions:

- It would be beneficial to include human annotators in evaluating the "Informative" metric [1] in addition to Factual and Context-p/n. For example, in the case study, a statement like "user loves the delicious food" is too general for a generated explanation. The "Informative" metric ensures that the explanation is specific to the user-item pair. Ideally, the explanation should cover all relevant positive and negative features without being overly verbose.
- Since LLMs can produce different outputs for the same input, the authors should report the average results over multiple trials.
- Table 6 shows that GPT-4 and human annotators generally make similar judgments. The authors should confirm that the annotators did not rely on any LLMs for evaluation. Additionally, it may be necessary to explore alternative methods to assess annotator quality. Furthermore, as each review is evaluated by only one annotator, cross-validation between annotators is not possible.
- In the experiment setup, the authors use RoBERTa-large to calculate the Content-p/n metrics in a way similar to BERTScore. However, the performance table only reports BERTScore results, not RoBERTa-large. The authors should explain this omission. Moreover, they should clarify why this paper cites the P5 model but does not include it as a reference explainable recommendation model.

[1] Explainable and coherent complement recommendation based on large language models.

Ethics Review Flag: No
Scope: 4: The work is relevant to the Web and to the track, and is of broad interest to the community
Novelty: 5
Technical Quality: 5
Reviewer Confidence: 3: The reviewer is confident but not certain that the evaluation is correct

The Two-Way Communication Period is Closing Soon

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

10 Dec 2024, 18:38 (modified: 10 Dec 2024, 20:13)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=brvGC34JIZ)

Comment:

We sincerely appreciate your valuable feedback and have tried to address all of your concerns and questions. We kindly request your follow-up during the remaining two-way communication period and would greatly appreciate it if you could reconsider your score in light of our responses.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 23:32 (modified: 06 Dec 2024, 01:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=dbWc2pG2nU)

Comment:

Thank you for your time and effort in reviewing our paper. We are sincerely grateful for your insightful feedback, which helps us refine our paper.

We have addressed all your questions with **A1** -- **A6**. Due to the character limit, we had to split our rebuttal comments into "rebuttal" and "official comment" text boxes below. We appreciate it if you take a look at all the responses and take them into account when updating your scores. Of course, if you have any further questions, please do not hesitate to ask.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

10 Dec 2024, 13:18 (modified: 10 Dec 2024, 18:38)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=qvr2IYbyiE)

[Deleted]

Rebuttal by Authors 3

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 23:32 (modified: 06 Dec 2024, 00:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=oshiO7zY1y)

Comment:

> In the experiment setup, the authors use RoBERTa-large to calculate the Content-p/n metrics in a way similar to BERTScore. However, the performance table only reports BERTScore results, not RoBERTa-large. The authors should explain this omission.

A5. By BERTScore, we mean the methodology to calculate textual similarity using a language model, and we always used RoBERTa-large to report the scores. In the BERTScore library, (https://pypi.org/project/bert-score/ (https://pypi.org/project/bert-score/)), RoBERTa-large is set as the default model. As in our paper, previous studies also report this metric as BERTScore, rather than RoBERTaScore.

> Moreover, they should clarify why this paper cites the P5 model but does not include it as a reference explainable recommendation model. [1] Explainable and coherent complement recommendation based on large language models.

A6. We did not include P5 because we focused on evaluating models that generate explanations given user and item embeddings as input. On the other hand, P5 is a text2text model that takes a prompt as input such as "Help Hong "Old boy" generate a 5-star explanation about this product: OtterBox Defender Case for iPhone 3G, 3GS (Black) [Retail Packaging]", and generates the answer. It also differs from other models greatly in that it is trained for various tasks simultaneously, including sequential and direct recommendation, rating prediction, and review summarization. We will clarify this in our camera-ready paper.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 23:13 (modified: 05 Dec 2024, 23:32)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=jRMCeQQzw4)

[Deleted]

Rebuttal by Authors 2

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 23:13 (modified: 06 Dec 2024, 14:18)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer rhoR, Authors

Revisions (/revisions?id=BE0qfB3hK3)

Comment:

> Table 6 shows that GPT-4 and human annotators generally make similar judgments. The authors should confirm that the annotators did not rely on any LLMs for evaluation.

A3. We have contacted the annotators and comfirmed that they did not rely on LLMs for evaluation.

> Additionally, it may be necessary to explore alternative methods to assess annotator quality. Furthermore, as each review is evaluated by only one annotator, cross-validation between annotators is not possible.

A4. While we agree that our annotation process could be improved by performing annotator selection and cross-validation, it requires much more annotation cost. Therefore, we kept the annotation process simple and asked annotators to validate the dataset quality based on objective metrics rather than subjective ones such as "informativeness".

Additionally, to address your concern, we further verified the dataset's quality using **Gemini-1.5-pro** and **Gemini-1.5-flash** as automatic evaluators, as well as **GPT-4o** (which we used in our paper). The results are shown below, ensuring the high quality of our datasets (the first table is reproduced from Table 6).

Table. The results of the dataset quality evaluation using **GPT-4o**. The numbers outside parentheses denote the scores estimated by GPT-4o, whereas those in parentheses indicate the percentage of the instances for which GPT-4o and human annotators make the same judgements.

Stage	Type	Amazon	Yelp	RateBeer
1	Factual	0.990 (0.95)	0.993 (0.98)	0.997 (0.95)
	Context-p	0.996 (0.98)	0.997 (0.96)	0.997 (0.98)
	Context-n	0.962 (0.97)	0.971 (0.95)	0.965 (0.97)

Stage	Type	Amazon	Yelp	RateBeer
2	Factual-p	0.999 (1.00)	0.999 (1.00)	0.996 (1.00)
	Factual-n	0.998 (0.99)	0.998 (1.00)	0.998 (0.99)
	Complete-p	0.997 (0.99)	0.997 (1.00)	0.998 (1.00)
	Complete-n	0.998 (1.00)	0.996 (1.00)	0.998 (1.00)

Table. The results of the dataset quality evaluation using **Gemini-1.5-pro** (gemini-1.5-pro-002). The numbers outside parentheses denote the scores estimated by Gemini-1.5-pro, whereas those in parentheses indicate the percentage of the instances for which Gemini-1.5-pro and human annotators make the same judgements.

Stage	Type	Amazon	Yelp	RateBeer
1	Factual	0.994 (0.94)	0.996 (1.00)	0.997 (0.94)
	Context-p	0.998 (0.97)	0.998 (0.97)	0.998 (0.97)
	Context-n	0.995 (0.97)	0.997 (0.99)	0.994 (0.96)
2	Factual-p	0.999 (1.00)	1.000 (1.00)	1.000 (1.00)
	Factual-n	0.997 (0.97)	0.997 (1.00)	0.999 (0.98)
	Complete-p	0.997 (0.97)	0.997 (1.00)	0.998 (1.00)
	Complete-n	0.997 (0.98)	0.996 (1.00)	0.998 (1.00)

Table. The results of the dataset quality evaluation using **Gemini-1.5-flash** (gemini-1.5-flash-002). The numbers outside parentheses denote the scores estimated by Gemini-1.5-flash, whereas those in parentheses indicate the percentage of the instances for which Gemini-1.5-flash and human annotators make the same judgements.

Stage	Type	Amazon	Yelp	RateBeer
1	Factual	0.997 (0.94)	0.997 (1.00)	0.996 (0.94)
	Context-p	0.998 (0.98)	0.998 (0.97)	0.996 (0.98)
	Context-n	0.996 (0.98)	0.997 (0.99)	0.990 (0.95)
2	Factual-p	0.999 (1.00)	0.998 (1.00)	0.999 (1.00)
	Factual-n	0.993 (0.99)	0.995 (1.00)	0.996 (0.97)
	Complete-p	0.992 (0.99)	0.987 (0.99)	0.993 (1.00)
	Complete-n	0.990 (0.99)	0.983 (0.99)	0.990 (1.00)

Rebuttal by Authors

Rebuttal

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 23:11 (modified: 06 Dec 2024, 02:45)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=18m9c3JbV1)

Rebuttal:

We greatly appreciate your insightful and constructive feedback. We have addressed all your questions with **A1** -- **A6** below:

> It would be beneficial to include human annotators in evaluating the "Informative" metric [1] in addition to Factual and Context-p/n. For example, in the case study, a statement like "user loves the delicious food" is too general for a generated explanation. The "Informative" metric ensures that the explanation is specific to the user-item pair. Ideally, the explanation should cover all relevant positive and negative features without being overly verbose.

A1. Thank you for your interesting suggestion. While we agree that providing user-specific explanations is very important, it comes with a few challenges. First, to retrieve all relevant positive and negative features from reviews, we need to lax the token-length regulation of the LLM output, and that increases the risk of hallucination significantly. Therefore, we opted to set the length of the explanations to 15 words, following the average token length in existing datasets. In our camera-ready paper, we will stress that we prioritized reducing hallucination over providing more detailed explanations.

Another challenge is that the judgement of “informativeness” is very subjective. This means that we need to hire many annotators and cross-validate their judgements to get reliable scores, which however comes with large annotation cost. In this work, therefore, we focused on validating the dataset quality based on more objective criteria, e.g. whether the generated explanations contain hallucinations or not.

> Since LLMs can produce different outputs for the same input, the authors should report the average results over multiple trials.

A2. To mitigate randomness, we always set the temperature to 0 when using LLMs. We apologize for not mentioning this point. We will include it in the camera-ready manuscript. We have also verified our datasets using Gemini-1.5 in addition to GPT-4 (the results are shown in **A4**).

*Due to the character limit, the further rebuttal comments will be shared as an "official comment."

Official Review of Submission1244 by Reviewer hUQ9

Official Review by Reviewer hUQ9 01 Dec 2024, 15:34 (modified: 03 Dec 2024, 05:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer hUQ9, Authors

Revisions (/revisions?id=zvsErA4Him)

Review:

This paper addresses an important limitation in existing explainable recommender systems: the inability to align generated explanations with users’ positive and negative sentiments. The authors propose a new task focusing on sentiment alignment and introduce novel datasets specifically designed to evaluate this aspect. These datasets are constructed using a two-step process involving review summarization and aspect-based sentiment analysis, utilizing a large language model (LLM). The authors also propose new evaluation metrics to assess the alignment between generated and ground-truth explanations. Through extensive experiments, they benchmark several baseline models and demonstrate that existing metrics fail to capture sentiment alignment. They further show that providing ground-truth ratings as additional input improves the sentiment-aware performance of baseline models. The study opens new research directions for explainable recommender systems by focusing on the alignment of explanations with user sentiments.

Overall, this is a good paper, and the proposed method is novel and sound. If this paper is finally accepted, I suggest that the authors provide a more comprehensive discussion of related work on sentiment-related explainable recommendations and offer a deeper explanation of the rationale behind the dataset construction process.

1. The structure of this paper is clear and its core idea is easy to follow.
2. The existence of mixed feelings in user reviews is a well-established phenomenon. Without explicitly considering users' sentiments, explainable recommender systems may fail to align their explanations with the positive and negative features expressed in the reviews. By constructing a new dataset that explicitly includes positive and negative features, the authors provide a valuable benchmark for evaluating sentiment-aware explanations, addressing this critical gap effectively.

3. The paper successfully demonstrates through experiments that existing evaluation metrics are insufficient for assessing sentiment-aware explanations. Furthermore, the authors introduce new evaluation metrics specifically designed to address this limitation, providing a more accurate and comprehensive framework for evaluating sentiment alignment in explainable recommendations.
4. The case study effectively highlights the significance of incorporating users' sentiments into the evaluation process to accurately assess the quality of explanations.

Weaknesses.

1. A potential weakness is the decision to perform review summarization as a mandatory first step in dataset construction. This approach might risk omitting certain positive or negative features, potentially compromising the accuracy of subsequent feature extraction. It is unclear whether the authors have considered this issue or explored alternative approaches, such as multi-round extraction, to mitigate these limitations.
2. Requiring the extracted features to strictly match the exact words or phrases in the output of the review summarization task might result in the loss of implicit features that are not explicitly mentioned in the original review text but can be inferred from the context.
3. The authors may have overlooked some relevant related work on sentiment-aware explainable recommender systems, which also take users' sentiments into account, such as Ref. [1] Park S J, Chae D K, Bae H K, et al. Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation[C]//Proceedings of the fifteenth ACM international conference on web search and data mining. 2022: 784-793. [2] Xie F, Wang Y, Xu K, et al. A Review-Level Sentiment Information Enhanced Multitask Learning Approach for Explainable Recommendation[J]. IEEE Transactions on Computational Social Systems, 2024.
4. In the paper, references [21] and [22] refer to the same paper. The correct reference for PETER [21] should be as follows. [3] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 4947–4957.

Questions:

1. This approach might risk omitting certain positive or negative features, potentially compromising the accuracy of subsequent feature extraction. It is unclear whether the authors have considered this issue or explored alternative approaches, such as multi-round extraction, to mitigate these limitations.
2. Requiring the extracted features to strictly match the exact words or phrases in the output of the review summarization task might result in the loss of implicit features that are not explicitly mentioned in the original review text but can be inferred from the context.
3. Is it ensured that all positive and negative features have been extracted from the original review text?
4. Are the extracted positive and negative features non-redundant? For instance, has independence testing been conducted to verify their uniqueness?

Ethics Review Flag: No
Scope: 4: The work is relevant to the Web and to the track, and is of broad interest to the community
Novelty: 5
Technical Quality: 4
Reviewer Confidence: 3: The reviewer is confident but not certain that the evaluation is correct

The Two-Way Communication Period is Closing Soon

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

10 Dec 2024, 13:19 (modified: 10 Dec 2024, 20:13)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer hUQ9, Authors

Revisions (/revisions?id=WrBPaWdnJF)

Comment:

We sincerely appreciate your valuable feedback and have tried to address all of your concerns and questions. We kindly request your follow-up during the remaining two-way communication period and would greatly appreciate it if you could reconsider your score in light of our responses.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:26 (modified: 06 Dec 2024, 01:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer hUQ9, Authors

Revisions (/revisions?id=jYEUZRfSLR)

Comment:

Thank you for your time and effort in reviewing our paper. We are sincerely grateful for your insightful feedback, which helps us refine our paper.

We have addressed all your questions with **A1** -- **A5**. Due to the character limit, we had to split our rebuttal comments into "rebuttal" and "official comment" text boxes below. We appreciate it if you take a look at all the responses and take them into account when updating your scores. Of course, if you have any further questions, please do not hesitate to ask.

Rebuttal by Authors 2

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:26 (modified: 06 Dec 2024, 02:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer hUQ9, Authors

Revisions (/revisions?id=2zK2LlnlM7)

Comment:

> The authors may have overlooked some relevant related work on sentiment-aware explainable recommender systems, which also take users' sentiments into account, such as Ref. [1] Park S J, Chae D K, Bae H K, et al. Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation[C]//Proceedings of the fifteenth ACM international conference on web search and data mining. 2022: 784-793. [2] Xie F, Wang Y, Xu K, et al. A Review-Level Sentiment Information Enhanced Multitask Learning Approach for Explainable Recommendation[J]. IEEE Transactions on Computational Social Systems, 2024.

A3. Thank you for pointing out these relevant papers. Both share the idea that accurately capturing users' emotions is important to provide accurate explanations, and since they highlight the importance of our research, we would like to cite them in our camera-ready manuscript.

> In the paper, references [21] and [22] refer to the same paper. The correct reference for PETER [21] should be as follows. [3] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 4947–4957.

A4. Thank you for pointing them out. We apologize for the mistake and will fix them in our camera-ready paper.

> Are the extracted positive and negative features non-redundant? For instance, has independence testing been conducted to verify their uniqueness?

A5. Thank you for your insightful comment. We checked the uniquness of the features by calclullating the number of feature types that appear N times, devided by the total number of the unique feature types. The following tables show the results. They show that our datasets contain various types of features (e.g. 81.9% of the negative features appear only once on Amazon), and models cannot acheive good scores just by memorizing frequent features.

Table. Independence testing result for the features (phrase) on Amazon dataset (#interactions: 438,604, #unique positive features: 179,832, #unique negative features: 163,071)

	ratio - positive feature	**ratio - negative feature**
N = 1	0.7760	0.8191
N > 1	0.2240	0.1809
N > 5	0.0655	0.0456
N > 10	0.0373	0.0246
N > 25	0.0172	0.0107
N > 50	0.0092	0.0054
N > 100	0.0050	0.0027

Table. Independence testing result for the features (phrase) on Yelp dataset (#interactions: 504,166, #unique positive features: 207,949, #unique negative features: 173,034)

	ratio - positive feature	**ratio - negative feature**
N = 1	0.7623	0.8138
N > 1	0.2377	0.1862
N > 5	0.0699	0.0450
N > 10	0.0395	0.0233
N > 25	0.0179	0.0095
N > 50	0.0096	0.0046
N > 100	0.0053	0.0022

Table. Independence testing result for the features (phrase) on RateBeer dataset (#interactions: 512,370, #unique positive features: 76,440, #unique negative features: 108,676)

	ratio - positive feature	**ratio - negative feature**
N = 1	0.6865	0.7404
N > 1	0.3135	0.2596
N > 5	0.1115	0.0798
N > 10	0.0703	0.0469
N > 25	0.0367	0.0225
N > 50	0.0223	0.0127
N > 100	0.0135	0.0071

Rebuttal by Authors

Rebuttal

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:26 (modified: 06 Dec 2024, 01:41)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=aAlERUp2re)

Rebuttal:

We greatly appreciate your insightful and constructive feedback. We have addressed all your questions with **A1** -- **A5** below:

> A potential weakness is the decision to perform review summarization as a mandatory first step in dataset construction. This approach might risk omitting certain positive or negative features, potentially compromising the accuracy of subsequent feature extraction. It is unclear whether the authors have considered this issue or explored alternative approaches, such as multi-round extraction, to mitigate these limitations.

> This approach might risk omitting certain positive or negative features, potentially compromising the accuracy of subsequent feature extraction. It is unclear whether the authors have considered this issue or explored alternative approaches, such as multi-round extraction, to mitigate these limitations.

> Is it ensured that all positive and negative features have been extracted from the original review text?

A1. As you point out, our approach might omit some positive or negative features mentioned in the original reviews. However, if we prompt LLMs to extract all features from reviews, it increases the risk of hallucination and makes the evaluation process less trustworthy. Therefore, in this work, we opted to minimize the risk of hallucination by setting the explanation length to 15 words, and extract relevant features from reviews.

Regarding the accuracy of the subsequent feature extraction step, we showed that our method almost always extracts all features from the explanations, as shown in complete-p/n in Table 5 and 6.

> Requiring the extracted features to strictly match the exact words or phrases in the output of the review summarization task might result in the loss of implicit features that are not explicitly mentioned in the original review text but can be inferred from the context.

A2. As you point out, we prompted LLMs to extract features that are explicitly mentioned in the explanations, and we imposed this regulation to minimize the risk of hallucination. During our summarization step, we did not include such restrictions in a prompt (instead, we specified the word legnth to mitigate hallucination) and hence LLMs could extract implicit features that represent the users' sentiments in their reviews.

*Due to the character limit, the further rebuttal comments will be shared as an "official comment."

Official Review of Submission1244 by Reviewer FfXj

Official Review by Reviewer FfXj 21 Nov 2024, 13:45 (modified: 03 Dec 2024, 05:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer FfXj, Authors

Revisions (/revisions?id=6erjk1RodL)

Review:

This paper makes a meaningful contribution to the field of explainable recommendation systems by introducing datasets and metrics that focus on user sentiments. The paper also presents a modified version of PETER along with extensive experiments to validate their claims. However, although I can follow the claims and experiments in the paper, since I am not too familiar with explainable recommendations, I will have a confidence of 2.

Pros

- The authors introduce datasets that explicitly disentangle users' positive and negative opinions using an LLM, providing a more fine-grained understanding of user sentiment compared to existing datasets.
- Proposing sentiment-matching and content similarity metrics makes the evaluations more rigorous.
- The paper benchmarks state-of-the-art models and introduces variations that integrate predicted ratings as input.
- The method is well-documented, with detailed explanations of dataset construction and metrics.
- Human evaluations were conducted to validate the performance of the LLMs.

Cons

- The proposed methods could be written using notations to be more explicit.
- For the experiments, especially in Table 11 and 12, some of the improvements seem incremental (e.g. 0.2%). Doing some significance tests could help the quantitative analysis of the models.
- Some parts of the paper are not well discussed. Please see the questions below.

Questions:

- On line 208, the authors mention that the model's output is restrict to 15 words. Is this achieved through prompting or just filtering out output with longer than 15 words?
- Between line 445 and 448, the authors introduce the sentiment-matching score. I am actually not quite sure about how it is computed. Are you considering all the explanations together no matter the content and measure the percentage of explanations that have the same sentiment? Using math notations to writer this out explicit would help the readers to better understand.

Ethics Review Flag: No
Scope: 4: The work is relevant to the Web and to the track, and is of broad interest to the community
Novelty: 5
Technical Quality: 6
Reviewer Confidence: 2: The reviewer is willing to defend the evaluation, but it is likely that the reviewer did not understand parts of the paper

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:27 (modified: 06 Dec 2024, 01:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer FfXj, Authors

Revisions (/revisions?id=1m8WsWSby1)

Comment:

Thank you for your time and effort in reviewing our paper. We are sincerely grateful for your insightful feedback, which helps us refine our paper.

We have addressed all your questions with **A1** -- **A4**. Due to the character limit, we had to split our rebuttal comments into "rebuttal" and "official comment" text boxes below. We appreciate it if you take a look at all the responses and take them into account when updating your scores. Of course, if you have any further questions, please do not hesitate to ask.

Rebuttal by Authors 2

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:27 (modified: 06 Dec 2024, 00:43)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer FfXj, Authors

Revisions (/revisions?id=m9nn9okN1G)

Comment:

> Between line 445 and 448, the authors introduce the sentiment-matching score. I am actually not quite sure about how it is computed. Are you considering all the explanations together no matter the content and measure the percentage of explanations that have the same sentiment? Using math notations to writer this out explicit would help the readers to better understand.

A4. Thank you for your suggestion. We will make it clearer using math notations in our camera-ready paper.

To calculate the sentiment-matching score given generated and ground-truth explanations, we first extract positive and negative features from the generated explanation using GPT-4o-mini, in the same way as how we extract features from the ground-truth explanation during our feature extraction step. Next, we label each explanation as “positive” if it contains only positive features; “negative” if it contains only negative features; and “neutral” if it contains both positive and negative features. Finally, we compare the labels of the generated and ground-truth explanations and see whether they share the same label; if they do, it means that the generated and ground-truth explanations have the same sentiment, but the contents of the features might differ.

To evaluate the content similarity, we also calculate the content-similarity score, which compares the textual similarity of the extracted features between the generated and ground-truth explanations using BERTScore.

Rebuttal by Authors

Rebuttal

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:27 (modified: 06 Dec 2024, 01:41)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=oMkRLR6UWS)

Rebuttal:

We greatly appreciate your insightful and constructive feedback. We have addressed all your questions with **A1** -- **A4** below:

> The proposed methods could be written using notations to be more explicit.

A1. We will clarify our proposed methods using notations in our camera-ready paper.

> For the experiments, especially in Table 11 and 12, some of the improvements seem incremental (e.g. 0.2%). Doing some significance tests could help the quantitative analysis of the models.

A2. As you point out, improvements are marginal in some metrics in Table 11 and 12, but we see large improvements in many other metrics (e.g. +17.1% in sentiment on Yelp). Furthermore, these tables compare the performance of our proposed methods w/w.o using ground-truth ratings as input, rather than comparing our method with baselines.

In Table 9, we observed marginal improvements over baselines in existing metrics used in previous work. However, this suggests that popular metrics such as ROUGE cannot properly evaluate the alignment of the sentiments between the generated and ground-truth explanations (as discussed in L695), and this is one of the important findings of our work. On the other hand, when we evaluated models using our proposed metrics (which focus on users’ sentiments), we observed large improvements as we showed in Table 8.

> On line 208, the authors mention that the model's output is restrict to 15 words. Is this achieved through prompting or just filtering out output with longer than 15 words?

A3. We specified “within 15 words” in a prompt, and we also set the max_tokens to 50 in Open-AI API because one word can be segmented into multiple tokens.

*Due to the character limit, the further rebuttal comments will be shared as an "official comment."

Official Comment by Reviewer FfXj

Official Comment by Reviewer FfXj 09 Dec 2024, 08:22

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I appreciate the detailed response from the authors and my concerns are addressed. I will retain my scores.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

09 Dec 2024, 14:28 (modified: 09 Dec 2024, 16:43)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=1qfQ6pNwsS)

Comment:

We sincerely appreciate your thoughtful review and the score you have given us. If there are any remaining questions or concerns we could address to provide further clarification, improve the manuscript, or enhance your confidence, please do not hesitate to let us know.

Official Review of Submission1244 by Reviewer Tsqz

Official Review by Reviewer Tsqz 19 Nov 2024, 23:12 (modified: 03 Dec 2024, 05:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer Tsqz, Authors

Revisions (/revisions?id=iW64saZh1M)

Review:

This paper introduces a novel dataset and evaluation method designed to distinguish between users' positive and negative sentiments towards recommended items, providing new research materials for the field of explainable recommendation. The authors employ GPT-4 to automatically extract sentiment features from user reviews and generate lists of users' positive and negative features, resulting in a high-quality dataset with clear sentiment information.

In the dataset, users' ratings may not always align perfectly with the extracted sentiments. For instance, a user might give a high rating while mentioning some negative aspects. Such inconsistencies between ratings and sentiment features could potentially affect the model's learning performance. This contradictory relationship has not been explicitly addressed in the dataset, which might impact the emotional coherence of the generated model.

Moreover, and most importantly, since GPT-4 is used to generate explainable texts for creating the dataset, utilizing this dataset with GPT to generate explanations might inherently lead to optimal performance, as the dataset itself was generated by GPT. Therefore, this dataset may only be suitable for evaluating the performance of non-GPT models, significantly limiting its applicability. Additionally, when using this dataset for training, other models might imitate GPT's

specific output style, leading to seemingly better performance. However, such an "advantage" would primarily stem from the structural similarity between the generated data and GPT, rather than reflecting the model's ability to produce diverse and authentic explanations.

Some minor issues:

1. The sentiment-matching score only measures whether positive/negative features are mentioned but does not evaluate whether the predicted sentiment corresponds to the described aspects or facts. In other words, it only reflects the overall sentiment tendency of the sentence. In recommendation explanations, the influence of different features may vary in importance, but Content-p/n does not account for feature weights. As a result, features with low importance might significantly affect the score, potentially misaligning with users' actual perceptions.
2. Incorporating users' predicted ratings as input can enhance the performance of the explanation text generator by providing the model with additional information. Treating rating prediction as a subtask is also effective, as it helps prevent the sentiment tendency of the generated text from deviating from the user's overall attitude. I believe the two approaches can coexist, and using ratings as input while simultaneously treating rating prediction as a subtask could yield better performance.
3. Why do models treating ratings as discrete variables generally perform better than those treating them as continuous variables? This is likely due to the nonlinear relationship between users' emotions and their ratings of items. Moreover, from the observations in Table 8, it is evident that -d-emb does not consistently outperform -c-emb.
4. Line 743: This suggests that optimal models differ depending on the nature of the dataset. This lacks further analysis—what specific characteristics of the dataset are related to this observation?

Questions:
The dataset created by the authors only distinguishes between positive and negative sentiments, without addressing the multidimensional or nuanced emotional differences that users may hold (e.g., "slightly like" vs. "strongly like"). Given the use of LLMs for extraction, could such detailed variations be captured through prompt engineering?

Ethics Review Flag: No
Scope: 2: The connection to the Web is incidental, e.g., use of Web data or API
Novelty: 4
Technical Quality: 3
Reviewer Confidence: 3: The reviewer is confident but not certain that the evaluation is correct

The Two-Way Communication Period is Closing Soon

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

10 Dec 2024, 20:11

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer Tsqz, Authors

Comment:
We sincerely appreciate your valuable feedback and have tried to address all of your concerns and questions. We kindly request your follow-up during the remaining two-way communication period and would greatly appreciate it if you could reconsider your score in light of our responses.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:31 (modified: 06 Dec 2024, 01:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer Tsqz, Authors

Revisions (/revisions?id=p9PkOzWJgO)

Comment:

Thank you for your time and effort in reviewing our paper. We are sincerely grateful for your insightful feedback, which helps us refine our paper.

We have addressed all your questions with **A1** -- **A9**. Due to the character limit, we had to split our rebuttal comments into "rebuttal" and "official comment" text boxes below. We appreciate it if you take a look at all the responses and take them into account when updating your scores. Of course, if you have any further questions, please do not hesitate to ask.

Official Comment by Authors

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

10 Dec 2024, 13:19 (modified: 10 Dec 2024, 20:10)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer Tsqz, Authors

Revisions (/revisions?id=b4twFXczee)

[Deleted]

Rebuttal by Authors 3

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:31 (modified: 06 Dec 2024, 00:46)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer Tsqz, Authors

Revisions (/revisions?id=s0tZhS5rjS)

Comment:

> Incorporating users' predicted ratings as input can enhance the performance of the explanation text generator by providing the model with additional information. Treating rating prediction as a subtask is also effective, as it helps prevent the sentiment tendency of the generated text from deviating from the user's overall attitude. I believe the two approaches can coexist, and using ratings as input while simultaneously treating rating prediction as a subtask could yield better performance.

A5. In our preliminary experiments, we also tried training our models (*-d/c-emb) with rating prediction as a subtask, but it did not improve performance. For your reference, the following table shows the performance of PEPLER-d-emb trained with the rating prediction loss (we will include this in our camera-ready paper.)

Table. Results of the model performing rating prediction as a subtask while incorporating the predicted rating as input based on our proposed evaluation metrics. The best scores among all models are **boldfaced**.

Method	Amazon (sentiment)	Amazon (content-p)	Amazon (content-n)	Yelp (sentiment)	Yelp (content-p)	Yelp (content-n)	RateBeer (sentiment)	RateBeer (content-p)	RateBeer (content-n)
PEPLER	0.5691	0.7439	0.6187	0.5462	0.7888	0.5422	0.6445	0.7966	0.6504
PEPLER-d-emb	**0.5995**	**0.7624**	**0.6320**	**0.5539**	**0.7928**	**0.5537**	**0.6697**	**0.8043**	0.6580

Method	Amazon (sentiment)	Amazon (content-p)	Amazon (content-n)	Yelp (sentiment)	Yelp (content-p)	Yelp (content-n)	RateBeer (sentiment)	RateBeer (content-p)	RateBeer (content-n)
PEPLER-d-emb w/ subtask	0.5942	0.7620	0.6281	0.5513	0.7848	0.5471	0.6532	0.7880	**0.6600**

> Why do models treating ratings as discrete variables generally perform better than those treating them as continuous variables? This is likely due to the nonlinear relationship between users' emotions and their ratings of items.

A6. Yes, as you state, this is likely due to the nonlinear relationship between the users' emotions and their ratings of items, as we described in L. 680-682.

> Moreover, from the observations in Table 8, it is evident that -d-emb does not consistently outperform -c-emb.

A7. Yes, -d-emb does not always outperform -c-emb, but Table 8 shows that -d-emb performs the best overall. The difference between these models is minimal (i.e. whether we treat the users' ratings as continuous or discrete values), and the comparison of their performance is not the main focus of this study.

> Line 743: This suggests that optimal models differ depending on the nature of the dataset. This lacks further analysis—what specific characteristics of the dataset are related to this observation?

A8. As described in L739-742, we observed that PETER performs better than ERRA and PEPLER overall in our datasets, and that contradicts the previous findings that the latter models perform better on previous datasets. Given that previous datasets are automatically contructed using rudimentary algorithms and contain much noise, ERRA and PEPLER might be more noise-robust than PETER.

> The dataset created by the authors only distinguishes between positive and negative sentiments, without addressing the multidimensional or nuanced emotional differences that users may hold (e.g., "slightly like" vs. "strongly like"). Given the use of LLMs for extraction, could such detailed variations be captured through prompt engineering?

A9. Yes, we could assign more nuanced sentiments such as "strongly like" and "slightly like" to each feature. However, automatic judgements of such nuanced sentiments using LLMs can be difficult and they may assign inconsistent sentiment labels to different instances, making the evaluation process less trustworthy. Therefore, we opted to classify features into two labels: "positive" or "negative", which are easy to distinguish and unlikely to be confused.

Rebuttal by Authors 2

Official Comment

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:30 (modified: 06 Dec 2024, 00:46)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer Tsqz, Authors

Revisions (/revisions?id=bnMtQio2wP)

Comment:

> Moreover, and most importantly, since GPT-4 is used to generate explainable texts for creating the dataset, utilizing this dataset with GPT to generate explanations might inherently lead to optimal performance, as the dataset itself was generated by GPT. Therefore, this dataset may only be suitable for evaluating the performance of non-GPT models, significantly limiting its applicability. Additionally, when using this dataset for training, other models might imitate GPT's specific output style, leading to seemingly better performance. However, such an "advantage" would primarily stem from the structural similarity between the generated data and GPT, rather than reflecting the model's ability to produce diverse and authentic explanations.

A2. While our datasets might contain some biases specific to GPT-4, they are valuable for evaluating performance of existing explainable recommendation systems, most of which are non-GPT models. Furthermore, we evaluated models using various metrics, including our proposed metrics: “sentiment-matching score” and “Content-p/n”. These metrics measure whether a model correctly predicts the users’ sentiments and the content of the features, i.e. what they particularly like and dislike about the target item. In those metrics, a model performs poorly if it merely learns the structural or syntactic patterns of the GPT-4 outputs.

> The sentiment-matching score only measures whether positive/negative features are mentioned but does not evaluate whether the predicted sentiment corresponds to the described aspects or facts. In other words, it only reflects the overall sentiment tendency of the sentence.

A3. As you point out, the sentiment-matching score focuses on evaluating the sentiment alignment, and that is why we also propose Content-p/n, which measures the content similarity of the positive/negative features between the generated and ground-truth explanations.

> In recommendation explanations, the influence of different features may vary in importance, but Content-p/n does not account for feature weights. As a result, features with low importance might significantly affect the score, potentially misaligning with users' actual perceptions.

A4. To address your concern, we calculated Content-p/n by enabling the IDF (inverse document frequency) weighting option in BERTScore, which gives more weights to infrequent words than frequent ones. As a result, we observed very similar results to what we reported in our paper, as we show in the following table. This result further supports our findings, and we will include these results in our camera-ready paper.

Table. Results based on our content-similarity scores with idf-based weighting. The best scores among all models are **boldfaced**.

Method	Amazon (content-p)	Amazon (content-n)	Yelp (content-p)	Yelp (content-n)	RateBeer (content-p)	RateBeer (content-n)
CER	0.7050	0.6089	0.7476	0.5479	0.7799	0.6495
ERRA	0.7047	0.5971	0.7530	0.5387	0.7948	0.6509
PEPLER-D	0.4545	0.4927	0.5191	0.4549	0.5724	0.5986
PETER	0.7049	0.6165	0.7494	0.5443	0.7770	0.6499
PETER-c-emb	0.7257	0.6110	0.7547	0.5576	**0.8145**	0.6440
PETER-d-emb	0.7149	0.6215	0.7524	0.5483	0.7983	0.6536
PEPLER	0.7439	0.6187	0.7888	0.5422	0.7966	0.6504
PEPLER-c-emb	0.7589	0.6293	**0.7947**	0.5507	0.7978	0.6500
PEPLER-d-emb	**0.7624**	**0.6320**	0.7928	**0.5537**	0.8043	**0.6580**

Rebuttal by Authors

Rebuttal

by Authors (Takashi Wada (/profile?id=~Takashi_Wada1), Johannes Kruse (/profile?id=~Johannes_Kruse1), Yuya Yoshikawa (/profile?id=~Yuya_Yoshikawa1), Sai Htaung Kham (/profile?id=~Sai_Htaung_Kham1), +8 more (/group/info?id=ACM.org/TheWebConf/2025/Conference/Submission1244/Authors))

05 Dec 2024, 22:30 (modified: 06 Dec 2024, 01:41)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=hsDiPb4dmz)

Rebuttal:

We greatly appreciate your insightful and constructive feedback. We have addressed all your questions with **A1** -- **A9** below:

> In the dataset, users' ratings may not always align perfectly with the extracted sentiments. For instance, a user might give a high rating while mentioning some negative aspects. Such inconsistencies between ratings and sentiment features could potentially affect the model's learning performance. This contradictory relationship has not been explicitly addressed in the dataset, which might impact the emotional coherence of the generated model.

A1. As you point out, we showed in Figure 2, 4 and 7 that users can give high ratings while mentioning negative aspects. To address this gap, we could've discarded the features that contradict the users' overall ratings, but we decided to keep them in our datasets to make them as faithful as possible to the users' original reviews.

Regarding the model's learning process, we feed both the users' embeddings and their ratings into the mdoel, which allows it to learn whether the target user tends to mention mixed feelings in the review regardless of the rating, e.g. for users who often mention pros and cons in their reviews, the model would generate both positive and negative features even if their predicted ratings are very high or low.

*Due to the character limit, the further rebuttal comments will be shared as an "official comment."