Appendix for “REASONER: An Explainable Recommendation Dataset with Multi-aspect Ground Truths”

Xu Chen[1,2], **Jingsen Zhang**[1,2,*], **Lei Wang**[1,2,*], **Quanyu Dai**[3], **Zhenhua Dong**[3], **Ruiming Tang**[3], **Rui Zhang**[4], **Li Chen**[5], **Wayne Xin Zhao**[1,2], **Ji-Rong Wen**[1,2]
[1]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
[2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[3]Huawei Noah's Ark Lab, China [4]www.ruizhang.info
[5]Department of Computer Science, Hong Kong Baptist University, Hong Kong

Contents

[*]Equal contribution.

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

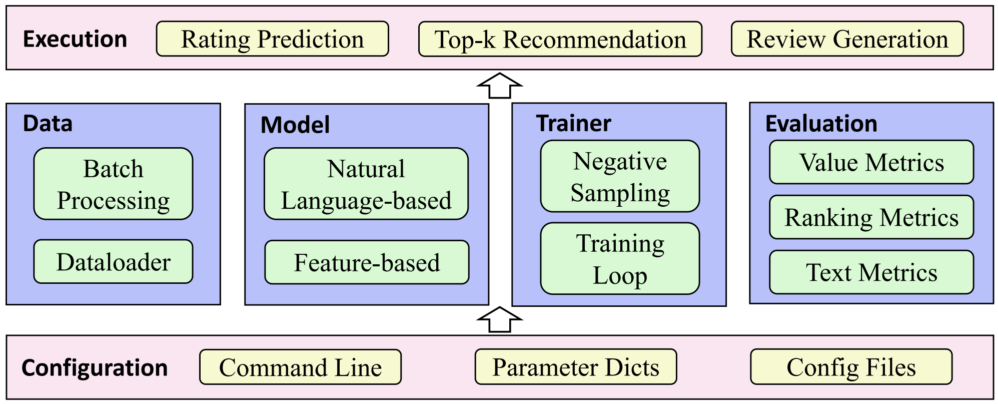

Figure 1: The structure of our library. There are six parts including the configuration, data, model, trainer, evaluation and execution modules.

A Accessibility

All the resources in our work are accessible at https://reasoner2023.github.io, providing comprehensive information about the dataset, library, related documents and the repository with long-term availability. Our licensing for the dataset is under a CC BY-NC 4.0 (Creative Commons Attribution-NonCommercial 4.0). See official instructions here[2]. We will consistently maintain and update the resources to ensure the long-term usability.

B Library

B.1 The Structure of the Library

We show the structure of our library in Figure 1. The configuration module is the base part of the library and responsible for initializing all the parameters. We support three methods to specify the parameters, that is, the command line, parameter dictionary and configuration file. Based on the configuration module, there are four higher layer modules, that is,

Data module: this module converts the raw data into the model inputs. There are two components: the first one is responsible for loading the data and building vocabularies for the user reviews. The second part aims to process the data into the formats required by the model inputs, and generate the sample batches for model optimization.

Model module: this module implements the explainable recommender models. There are two types of methods in our library. The first one includes the feature-based explainable recommender models, and the second one contains the models with natural language explanations. We delay the detailed introduction of these models in the next section.

Trainer module: this module is leveraged to implement the training losses, such as the Bayesian Personalized Ranking (BPR) and Binary Cross Entropy (BCE). In addition, this module can also record the complete model training process.

Evaluation module: this module is designed to evaluate different models, and there are three types of evaluation tasks, that is, rating prediction, top-k recommendation and review generation.

Upon the above four modules, there is an execution module on the upper-most layer. It is responsible for optimizing the recommender model for different tasks, such as rating prediction, tag prediction and review generation. For more detailed introduction on our library architecture, we refer the readers to our project at https://reasoner2023.github.io/.

[2]https://creativecommons.org/licenses/by-nc/4.0/

B.2 The Implemented Models

In our library, we implement two types of explainable recommender models, which are widely studied in the research community. The first one are feature-based explainable recommender models, where the features can be the tags, item aspects and so on. The second one are the models with natural language explanations.

More specifically, we implement the following representative feature-based explainable recommender models:

EFM [15] predicts the user preferences and generates explainable recommendations based on explicit product features and user opinions from the review information.

TriRank [5] models the user-item-aspect ternary relation as a heterogeneous tripartite graph based on user ratings and reviews, and it devises a vertex ranking algorithm for recommendation.

LRPPM [3] is a tensor-matrix factorization algorithm which captures the user preferences using ranking-based optimization objective over various item aspects.

SULM [1] enhances recommendations by recommending not only item but also the specific aspects by using aspect-level sentiment analysis.

MTER [14] is a tensor factorization method which models the task of item recommendation using a three-way tensor over the users, items and features. We omit the modeling of the opinions in the original implementation for adapting our data.

AMF [6] improves the recommendation accuracy by using the auxiliary information extracted from the user review aspects.

TRDM [18] introduces a two-stage approach to generate accurate item recommendations and effective tag-based potential features simultaneously for enhancing recommendation accuracy and diversity.

TRAL [17] proposes attention-based learning to capture diverse tag-based features, and compress these features with an attention pooling layer to enhance recommendation accuracy.

HPTR [16] employs hyperbolic distance to measure semantic relevance between entities, which better captures hierarchical structures presented in tag information.

AIRec [2] enhances tag-aware recommender system by employing a hierarchical attention network to capture multi-aspect preferences and leveraging tag intersection to improve conjunct feature learning.

HAN-TR [13] captures distinct user preferences and informative elements by employing separate attention networks for element-level influence and information-level attentiveness.

TNAM [7] addresses the issues of tag weight assignment in recommender systems by introducing a tag-based neural attention network that captures users' specific tag attention.

BPR-T [8] addresses high dimension and sparsity issues of tagging information by integrating tag mapping into a Bayesian personalized ranking collaborative filtering model.

In addition to the above shallow models based on matrix factorization, we also implement the following deep feature-based explainable recommender models (called **DERM** for short):

DERM-MLP is a deep recommender model for jointly predicting the ratings and tags. The two tasks share the set of user/item/tag embeddings. The hidden states as well as the tag embeddings are put into different layers corresponding to the different tasks.

DERM-MF firstly obtains a hidden state based on the user/item embeddings using matrix factorization, and then the outputs are computed by a neural network.

DERM-C combines matrix factorization and Multi-Layer Perceptron (MLP) to derive the hidden states, and the outputs are merged in a concatenated manner.

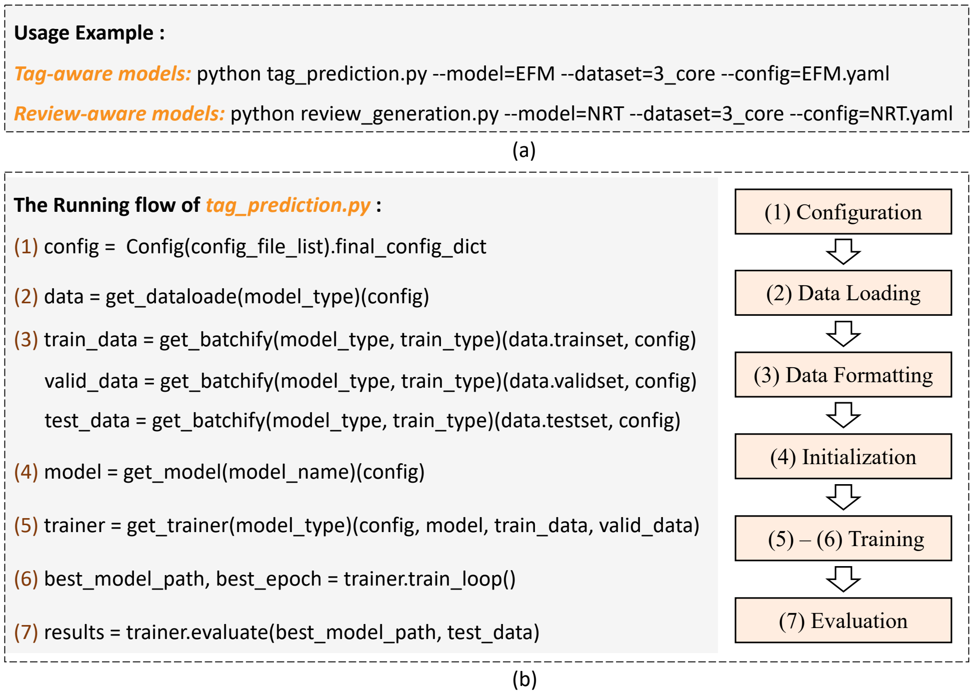

Figure 2: (a) Examples of the feature- and review-based models. (b) The running flow of our library.

DERM-H leverages the tags to profile the users and items, and then use the same architecture as DERM-MLP for predicting the ratings and tags.

For the models with natural language explanations, we implement the following representative methods:

Att2Seq [4] is a review generation model which uses LSTM as the decoder, and output the texts directly based on the user/item IDs and rating information.

NRT [10] simultaneously predicts the reviews and ratings based on the input user-item pair, where the two tasks share the same embedding and hidden layers.

PETER [9] leverages Transformer to generate the user reviews, which is a state-of-the-art review generation model.

B.3 Examples for Using Our Library

In this section, we introduce how to use our library. We present a simple example in Figure 2(a), where one can directly execute *tag_prediction.py* or *review_generate.py* to run a feature-based or review-based model, respectively. In each of these commands, one needs to specify three parameters to indicate the names of the model, dataset and configuration file, respectively.

Take *tag_prediction.py* for example, it sequentially executes the following steps: (1) Configuration. In this step, the parameters related to the model architecture and optimization process from different sources (commend line, configuration dictionary and files) are integrated into a dictionary. (2) Data loading. In this step, the dataloader is selected according to model type. For the review-aware models, this step reads all the records and build the vocabulary. (3) Data Formatting. The training, validation and test sets are processed into the formats required by the model input in a sample batch manner. (4) Initialization. The corresponding model class will be defined and initialized according to the parameter values in configuration. (5)-(6) Training. Selecting the optimization approach to train the model. (7) Evaluation. Measuring the model performance on different tasks.

Table 1: Statistics of the datasets.

Dataset	REASONER
# Users	2,997
# Items	4,672
# Tags	6,115
# Interactions	58,497
Avg. # words / review	17

Our library is highly extensible, and there are three steps to realize a new model: (1) implementing the basic functions of the model, including the model architecture, preference score prediction etc. (2) Customizing the training approaches in *train.py*. (3) Indicating the parameters in the config file.

C Benchmark

C.1 Experiment Setup

Considering that we have three types of ground truths for the explanations, we evaluate the model performance by predicting the tags for Q1, Q2, Q3, Q1+Q2, Q2+Q3, Q1+Q3 and Q1+Q2+Q3, respectively. When we have to predict multiple types of ground truths, we extend the original models to their multi-task versions by sharing the embedding parameters. We randomly split the dataset into the training, validation and testing sets according to the ratio of 8:1:1. For the review generation task, we use the most 20,000 frequently mentioned words to construct the vocabulary, and the maximum length of the generated sentences is set to 17, which is equal to the average length of reviews in dataset. The dataset statistics are presented in Table 1. For all the models, the batch size is set as 256. We tune the other key hyper-parameters by grid search. In specific, we tune the learning rate and the weight of L2 regularization in the range of [0.1, 0.01, 0.001, 0.0001] and [0.001, 0.0001, 0] respectively. For the deep models, we tune the hidden size and the layer number in the range of [32, 64, 128, 256] and [1, 2, 3, 4] respectively. More details of the experiment setting are shown in our project, which has been released at https://reasoner2023.github.io/. We use RMSE and MAE as the metrics to evaluate the performance of the rating prediction task. For the task of tag prediction, F1 and NDCG are selected to evaluate the model performance. To evaluate the quality of the generated reviews, we leverage the metrics including BLEU [12] and ROUGE [11] for model comparison.

C.2 Experiment Results

The comparison results of the feature-based explainable recommender models on the tasks of tag and rating predictions are presented in Table 2-8. The comparison results of the models with natural language explanations on the task of review generation and rating prediction are presented in Table 9. We use the tags with top 10 prediction scores to calculate F1 and NDCG, and the results are percentage values with "%" omitted. For RMSE and MAE, a lower value indicates better performance. For each evaluation metric, we use bold fonts to label the best performance. Since the TriRank we implemented does not support to predict multiple types of tags simultaneously, we omit it in the corresponding tables.

Table 2: The benchmarking results of the feature-based explainable recommender models on predicting the tags for persuasiveness and ratings.

Metrics	Persuasiveness		Rating Prediction	
	F1	NDCG	RMSE	MAE
EFM	$26.99_{\pm 0.35}$	$18.89_{\pm 0.23}$	$1.68_{\pm 0.00}$	$1.24_{\pm 0.01}$
TriRank	$18.36_{\pm 0.07}$	$13.98_{\pm 0.06}$	$2.90_{\pm 0.00}$	$2.58_{\pm 0.00}$
LRPPM	$37.31_{\pm 0.23}$	$23.25_{\pm 0.10}$	$1.22_{\pm 0.00}$	$0.96_{\pm 0.00}$
SULM	$\mathbf{41.68}_{\pm 0.63}$	$\mathbf{25.77}_{\pm 0.22}$	$1.65_{\pm 0.08}$	$1.30_{\pm 0.06}$
MTER	$5.66_{\pm 1.74}$	$2.65_{\pm 1.13}$	$2.27_{\pm 0.64}$	$1.96_{\pm 0.62}$
AMF	$27.93_{\pm 0.08}$	$17.62_{\pm 0.17}$	$2.28_{\pm 0.00}$	$1.86_{\pm 0.00}$
DERM-MLP	$37.74_{\pm 0.12}$	$23.45_{\pm 0.02}$	$1.30_{\pm 0.01}$	$1.06_{\pm 0.01}$
DERM-MF	$36.57_{\pm 0.14}$	$21.49_{\pm 0.17}$	$1.32_{\pm 0.00}$	$1.14_{\pm 0.00}$
DERM-C	$37.17_{\pm 0.19}$	$23.21_{\pm 0.01}$	$1.30_{\pm 0.00}$	$1.07_{\pm 0.01}$
DERM-H	$35.58_{\pm 0.07}$	$22.02_{\pm 0.18}$	$1.27_{\pm 0.01}$	$1.04_{\pm 0.01}$
TRDM	$30.24_{\pm 1.57}$	$13.44_{\pm 1.11}$	$\mathbf{1.19}_{\pm 0.00}$	$\mathbf{0.95}_{\pm 0.01}$
TRAL	$5.93_{\pm 0.05}$	$2.61_{\pm 0.02}$	$1.29_{\pm 0.00}$	$1.08_{\pm 0.00}$
HPTR	$38.61_{\pm 0.20}$	$23.05_{\pm 1.28}$	$2.16_{\pm 0.71}$	$1.88_{\pm 0.68}$
AIRec	$38.05_{\pm 0.07}$	$23.07_{\pm 0.03}$	$1.31_{\pm 0.00}$	$1.08_{\pm 0.01}$
HAN-TR	$34.96_{\pm 0.60}$	$18.58_{\pm 3.37}$	$2.10_{\pm 0.80}$	$1.83_{\pm 0.75}$
TNAM	$5.97_{\pm 0.01}$	$2.60_{\pm 0.01}$	$1.37_{\pm 0.01}$	$1.17_{\pm 0.01}$
BPR-T	$32.70_{\pm 0.62}$	$18.32_{\pm 0.82}$	$1.23_{\pm 0.00}$	$0.98_{\pm 0.01}$

Table 3: The benchmarking results of the feature-based explainable recommender models on predicting the tags for informativeness and ratings.

Metrics	Informativeness		Rating Prediction	
	F1	NDCG	RMSE	MAE
EFM	$5.38_{\pm 0.28}$	$3.97_{\pm 0.19}$	$1.68_{\pm 0.00}$	$1.24_{\pm 0.01}$
TriRank	$18.78_{\pm 0.10}$	$14.50_{\pm 0.09}$	$2.90_{\pm 0.00}$	$2.58_{\pm 0.00}$
LRPPM	$37.85_{\pm 0.22}$	$38.35_{\pm 0.15}$	$1.22_{\pm 0.00}$	$0.96_{\pm 0.00}$
SULM	$\mathbf{43.25}_{\pm 0.59}$	$\mathbf{42.97}_{\pm 0.40}$	$1.65_{\pm 0.08}$	$1.30_{\pm 0.06}$
MTER	$8.40_{\pm 0.86}$	$6.13_{\pm 0.68}$	$2.04_{\pm 0.68}$	$1.74_{\pm 0.66}$
AMF	$28.63_{\pm 0.24}$	$28.95_{\pm 0.29}$	$2.28_{\pm 0.00}$	$1.86_{\pm 0.00}$
DERM-MLP	$38.60_{\pm 0.06}$	$38.99_{\pm 0.03}$	$1.30_{\pm 0.01}$	$1.06_{\pm 0.01}$
DERM-MF	$37.10_{\pm 0.09}$	$35.37_{\pm 0.14}$	$1.32_{\pm 0.00}$	$1.14_{\pm 0.00}$
DERM-C	$37.96_{\pm 0.10}$	$38.43_{\pm 0.05}$	$1.30_{\pm 0.00}$	$1.07_{\pm 0.01}$
DERM-H	$36.36_{\pm 0.55}$	$35.85_{\pm 0.41}$	$1.29_{\pm 0.01}$	$1.06_{\pm 0.01}$
TRDM	$31.49_{\pm 2.84}$	$24.09_{\pm 2.80}$	$\mathbf{1.19}_{\pm 0.00}$	$\mathbf{0.95}_{\pm 0.01}$
TRAL	$6.04_{\pm 0.06}$	$4.59_{\pm 0.01}$	$1.28_{\pm 0.00}$	$1.08_{\pm 0.00}$
HPTR	$39.55_{\pm 0.09}$	$37.73_{\pm 1.88}$	$1.27_{\pm 0.07}$	$1.05_{\pm 0.09}$
AIRec	$38.90_{\pm 0.06}$	$39.33_{\pm 0.06}$	$1.31_{\pm 0.00}$	$1.08_{\pm 0.01}$
HAN-TR	$34.95_{\pm 1.70}$	$30.69_{\pm 5.89}$	$2.10_{\pm 0.79}$	$1.83_{\pm 0.75}$
TNAM	$37.77_{\pm 0.35}$	$37.97_{\pm 0.20}$	$1.36_{\pm 0.00}$	$1.16_{\pm 0.00}$
BPR-T	$33.83_{\pm 0.43}$	$31.15_{\pm 0.33}$	$1.23_{\pm 0.01}$	$0.98_{\pm 0.01}$

Table 4: The benchmarking results of the feature-based explainable recommender models on predicting the tags for satisfaction and ratings.

Metrics	Satisfaction		Rating Prediction	
	F1	NDCG	RMSE	MAE
EFM	$4.57_{\pm 0.46}$	$1.79_{\pm 0.19}$	$1.68_{\pm 0.00}$	$1.24_{\pm 0.01}$
TriRank	$16.82_{\pm 0.03}$	$13.16_{\pm 0.02}$	$2.90_{\pm 0.00}$	$2.58_{\pm 0.00}$
LRPPM	$36.04_{\pm 0.16}$	$22.35_{\pm 0.08}$	$1.22_{\pm 0.00}$	$0.96_{\pm 0.00}$
SULM	$\mathbf{40.46}_{\pm 0.62}$	$\mathbf{24.80}_{\pm 0.19}$	$1.64_{\pm 0.09}$	$1.29_{\pm 0.07}$
MTER	$5.97_{\pm 1.92}$	$2.85_{\pm 1.07}$	$2.26_{\pm 0.65}$	$1.96_{\pm 0.62}$
AMF	$27.16_{\pm 0.19}$	$17.05_{\pm 0.21}$	$2.28_{\pm 0.00}$	$1.86_{\pm 0.00}$
DERM-MLP	$36.76_{\pm 0.07}$	$22.53_{\pm 0.12}$	$1.30_{\pm 0.01}$	$1.06_{\pm 0.01}$
DERM-MF	$35.40_{\pm 0.23}$	$20.59_{\pm 0.35}$	$1.32_{\pm 0.00}$	$1.14_{\pm 0.00}$
DERM-C	$36.20_{\pm 0.28}$	$22.28_{\pm 0.22}$	$1.29_{\pm 0.01}$	$1.07_{\pm 0.01}$
DERM-H	$34.65_{\pm 0.43}$	$21.33_{\pm 0.47}$	$1.28_{\pm 0.01}$	$1.05_{\pm 0.02}$
TRDM	$31.29_{\pm 0.63}$	$14.74_{\pm 0.69}$	$\mathbf{1.19}_{\pm 0.00}$	$\mathbf{0.95}_{\pm 0.01}$
TRAL	$5.89_{\pm 0.05}$	$2.52_{\pm 0.01}$	$1.29_{\pm 0.00}$	$1.08_{\pm 0.00}$
HPTR	$35.35_{\pm 2.90}$	$17.78_{\pm 3.92}$	$1.81_{\pm 0.77}$	$1.56_{\pm 0.72}$
AIRec	$37.10_{\pm 0.09}$	$22.86_{\pm 0.08}$	$1.30_{\pm 0.01}$	$1.08_{\pm 0.01}$
HAN-TR	$33.95_{\pm 1.80}$	$18.06_{\pm 4.15}$	$2.10_{\pm 0.80}$	$1.83_{\pm 0.75}$
TNAM	$5.89_{\pm 0.05}$	$2.52_{\pm 0.01}$	$1.37_{\pm 0.00}$	$1.17_{\pm 0.00}$
BPR-T	$33.82_{\pm 0.29}$	$19.52_{\pm 0.25}$	$1.23_{\pm 0.00}$	$0.98_{\pm 0.00}$

Table 5: The benchmarking results of the feature-based explainable recommender models on jointly predicting the tags for persuasiveness, informativeness and ratings.

Metrics	Persuasiveness		Informativeness		Rating	
	F1	NDCG	F1	NDCG	RMSE	MAE
EFM	$15.69_{\pm 0.03}$	$12.74_{\pm 0.07}$	$5.38_{\pm 0.73}$	$3.94_{\pm 0.42}$	$1.66_{\pm 0.00}$	$1.23_{\pm 0.00}$
LRPPM	$37.32_{\pm 0.21}$	$23.26_{\pm 0.09}$	$37.89_{\pm 0.19}$	$38.37_{\pm 0.13}$	$1.22_{\pm 0.00}$	$0.96_{\pm 0.00}$
SULM	$\mathbf{41.34}_{\pm 0.53}$	$\mathbf{25.68}_{\pm 0.20}$	$\mathbf{42.70}_{\pm 0.50}$	$\mathbf{42.82}_{\pm 0.35}$	$1.67_{\pm 0.06}$	$1.31_{\pm 0.05}$
MTER	$35.53_{\pm 0.21}$	$21.88_{\pm 0.35}$	$36.22_{\pm 0.30}$	$36.09_{\pm 0.61}$	$1.36_{\pm 0.01}$	$1.09_{\pm 0.01}$
AMF	$27.67_{\pm 0.16}$	$17.45_{\pm 0.15}$	$28.23_{\pm 0.33}$	$28.57_{\pm 0.34}$	$2.28_{\pm 0.00}$	$1.86_{\pm 0.00}$
DERM-MLP	$38.49_{\pm 0.15}$	$23.80_{\pm 0.10}$	$39.14_{\pm 0.10}$	$39.38_{\pm 0.10}$	$1.30_{\pm 0.01}$	$1.06_{\pm 0.01}$
DERM-MF	$36.84_{\pm 0.04}$	$22.55_{\pm 0.05}$	$37.58_{\pm 0.11}$	$37.44_{\pm 0.08}$	$1.32_{\pm 0.00}$	$1.15_{\pm 0.00}$
DERM-C	$37.85_{\pm 0.26}$	$23.43_{\pm 0.21}$	$38.82_{\pm 0.06}$	$39.10_{\pm 0.09}$	$1.30_{\pm 0.01}$	$1.08_{\pm 0.01}$
DERM-H	$37.47_{\pm 0.26}$	$23.23_{\pm 0.22}$	$38.17_{\pm 0.22}$	$38.32_{\pm 0.36}$	$1.28_{\pm 0.00}$	$1.04_{\pm 0.01}$
TRDM	$32.91_{\pm 1.06}$	$15.51_{\pm 0.72}$	$33.01_{\pm 0.50}$	$25.02_{\pm 0.37}$	$\mathbf{1.19}_{\pm 0.00}$	$\mathbf{0.94}_{\pm 0.01}$
TRAL	$5.93_{\pm 0.05}$	$2.61_{\pm 0.02}$	$6.06_{\pm 0.06}$	$4.60_{\pm 0.02}$	$1.29_{\pm 0.00}$	$1.08_{\pm 0.00}$
HPTR	$38.68_{\pm 0.25}$	$23.11_{\pm 1.20}$	$36.20_{\pm 4.79}$	$31.40_{\pm 9.27}$	$2.16_{\pm 0.71}$	$1.88_{\pm 0.68}$
AIRec	$38.73_{\pm 0.10}$	$23.97_{\pm 0.07}$	$39.43_{\pm 0.08}$	$39.66_{\pm 0.09}$	$1.31_{\pm 0.00}$	$1.08_{\pm 0.01}$
HAN-TR	$33.32_{\pm 0.75}$	$16.83_{\pm 0.53}$	$33.31_{\pm 0.93}$	$27.72_{\pm 0.46}$	$1.29_{\pm 0.02}$	$1.06_{\pm 0.03}$
TNAM	$5.91_{\pm 0.05}$	$2.63_{\pm 0.02}$	$37.47_{\pm 0.46}$	$37.35_{\pm 0.58}$	$1.36_{\pm 0.01}$	$1.17_{\pm 0.01}$
BPR-T	$33.15_{\pm 0.22}$	$18.90_{\pm 0.44}$	$34.05_{\pm 0.20}$	$31.51_{\pm 0.26}$	$1.24_{\pm 0.01}$	$0.99_{\pm 0.01}$

Table 6: The benchmarking results of the feature-based explainable recommender models on jointly predicting the tags for persuasiveness, satisfaction and ratings.

Metrics	Persuasiveness		Satisfaction		Rating	
	F1	NDCG	F1	NDCG	RMSE	MAE
EFM	$15.58_{\pm 0.03}$	$12.84_{\pm 0.07}$	$4.58_{\pm 0.45}$	$1.77_{\pm 0.18}$	$1.66_{\pm 0.00}$	$1.23_{\pm 0.00}$
LRPPM	$37.32_{\pm 0.21}$	$23.26_{\pm 0.09}$	$36.06_{\pm 0.14}$	$22.37_{\pm 0.07}$	$1.22_{\pm 0.00}$	$0.96_{\pm 0.00}$
SULM	$\mathbf{41.36}_{\pm 0.56}$	$\mathbf{35.71}_{\pm 0.22}$	$\mathbf{40.15}_{\pm 0.54}$	$\mathbf{24.70}_{\pm 0.18}$	$1.67_{\pm 0.06}$	$1.31_{\pm 0.05}$
MTER	$5.83_{\pm 0.52}$	$2.56_{\pm 0.18}$	$5.36_{\pm 0.17}$	$2.23_{\pm 0.12}$	$2.04_{\pm 0.68}$	$1.74_{\pm 0.66}$
AMF	$27.71_{\pm 0.21}$	$17.52_{\pm 0.21}$	$26.90_{\pm 0.18}$	$16.94_{\pm 0.16}$	$2.28_{\pm 0.00}$	$1.86_{\pm 0.00}$
DERM-MLP	$38.37_{\pm 0.09}$	$23.71_{\pm 0.12}$	$37.32_{\pm 0.02}$	$22.87_{\pm 0.03}$	$1.30_{\pm 0.01}$	$1.06_{\pm 0.01}$
DERM-MF	$36.90_{\pm 0.12}$	$22.63_{\pm 0.12}$	$35.78_{\pm 0.10}$	$21.74_{\pm 0.12}$	$1.32_{\pm 0.00}$	$1.15_{\pm 0.00}$
DERM-C	$38.03_{\pm 0.11}$	$23.60_{\pm 0.06}$	$36.95_{\pm 0.10}$	$22.72_{\pm 0.06}$	$1.31_{\pm 0.01}$	$1.08_{\pm 0.00}$
DERM-H	$37.49_{\pm 0.24}$	$23.32_{\pm 0.18}$	$36.11_{\pm 0.29}$	$22.34_{\pm 0.17}$	$1.29_{\pm 0.01}$	$1.04_{\pm 0.01}$
TRDM	$32.57_{\pm 1.92}$	$15.11_{\pm 1.21}$	$30.91_{\pm 1.77}$	$13.86_{\pm 1.40}$	$\mathbf{1.19}_{\pm 0.00}$	$\mathbf{0.94}_{\pm 0.00}$
TRAL	$5.91_{\pm 0.05}$	$2.63_{\pm 0.02}$	$5.89_{\pm 0.05}$	$2.52_{\pm 0.01}$	$1.29_{\pm 0.00}$	$1.08_{\pm 0.00}$
HPTR	$38.64_{\pm 0.20}$	$22.93_{\pm 1.46}$	$32.96_{\pm 4.03}$	$14.78_{\pm 4.06}$	$2.16_{\pm 0.71}$	$1.88_{\pm 0.68}$
AIRec	$38.69_{\pm 0.06}$	$23.94_{\pm 0.05}$	$37.63_{\pm 0.05}$	$23.08_{\pm 0.02}$	$1.30_{\pm 0.01}$	$1.07_{\pm 0.01}$
HAN-TR	$35.95_{\pm 2.15}$	$19.75_{\pm 3.93}$	$34.95_{\pm 2.02}$	$19.13_{\pm 3.74}$	$2.09_{\pm 0.80}$	$1.83_{\pm 0.75}$
TNAM	$5.91_{\pm 0.05}$	$2.63_{\pm 0.02}$	$5.89_{\pm 0.05}$	$2.52_{\pm 0.01}$	$1.39_{\pm 0.02}$	$1.17_{\pm 0.02}$
BPR-T	$33.07_{\pm 0.18}$	$18.93_{\pm 0.48}$	$33.75_{\pm 0.26}$	$19.57_{\pm 0.28}$	$1.23_{\pm 0.00}$	$0.99_{\pm 0.01}$

Table 7: The benchmarking results of the feature-based explainable recommender models on jointly predicting the tags for informativeness, satisfaction and ratings.

Metrics	Informativeness		Satisfaction		Rating	
	F1	NDCG	F1	NDCG	RMSE	MAE
EFM	$5.15_{\pm 0.79}$	$3.75_{\pm 0.38}$	$4.57_{\pm 0.54}$	$1.75_{\pm 0.21}$	$1.66_{\pm 0.00}$	$1.23_{\pm 0.00}$
LRPPM	$37.89_{\pm 0.19}$	$38.37_{\pm 0.14}$	$36.06_{\pm 0.14}$	$22.37_{\pm 0.07}$	$1.22_{\pm 0.00}$	$0.96_{\pm 0.00}$
SULM	$\mathbf{42.70}_{\pm 0.49}$	$\mathbf{42.84}_{\pm 0.37}$	$\mathbf{40.12}_{\pm 0.51}$	$\mathbf{24.70}_{\pm 0.17}$	$1.67_{\pm 0.06}$	$1.31_{\pm 0.05}$
MTER	$6.13_{\pm 0.03}$	$4.56_{\pm 0.18}$	$5.64_{\pm 0.66}$	$2.37_{\pm 0.30}$	$2.04_{\pm 0.68}$	$1.74_{\pm 0.66}$
AMF	$28.17_{\pm 0.28}$	$28.53_{\pm 0.32}$	$26.79_{\pm 0.06}$	$16.89_{\pm 0.09}$	$2.28_{\pm 0.00}$	$1.86_{\pm 0.00}$
DERM-MLP	$39.23_{\pm 0.08}$	$39.45_{\pm 0.02}$	$37.40_{\pm 0.07}$	$22.93_{\pm 0.06}$	$1.30_{\pm 0.01}$	$1.06_{\pm 0.01}$
DERM-MF	$37.60_{\pm 0.13}$	$37.49_{\pm 0.17}$	$35.77_{\pm 0.16}$	$21.76_{\pm 0.16}$	$1.32_{\pm 0.00}$	$1.15_{\pm 0.00}$
DERM-C	$38.77_{\pm 0.13}$	$39.08_{\pm 0.18}$	$36.84_{\pm 0.23}$	$22.59_{\pm 0.16}$	$1.30_{\pm 0.01}$	$1.07_{\pm 0.01}$
DERM-H	$38.13_{\pm 0.50}$	$38.45_{\pm 0.47}$	$36.44_{\pm 0.28}$	$22.55_{\pm 0.21}$	$1.27_{\pm 0.01}$	$1.04_{\pm 0.01}$
TRDM	$33.15_{\pm 0.98}$	$25.24_{\pm 1.92}$	$30.42_{\pm 1.50}$	$13.83_{\pm 1.22}$	$\mathbf{1.19}_{\pm 0.00}$	$\mathbf{0.94}_{\pm 0.01}$
TRAL	$6.04_{\pm 0.06}$	$4.56_{\pm 0.02}$	$5.84_{\pm 0.01}$	$2.43_{\pm 0.08}$	$1.29_{\pm 0.00}$	$1.08_{\pm 0.00}$
HPTR	$37.36_{\pm 3.15}$	$32.22_{\pm 7.55}$	$35.35_{\pm 2.93}$	$17.79_{\pm 3.92}$	$1.81_{\pm 0.77}$	$1.56_{\pm 0.72}$
AIRec	$39.46_{\pm 0.11}$	$39.72_{\pm 0.07}$	$37.65_{\pm 0.11}$	$23.13_{\pm 0.07}$	$1.30_{\pm 0.00}$	$1.08_{\pm 0.01}$
HAN-TR	$35.79_{\pm 3.09}$	$32.11_{\pm 6.85}$	$34.04_{\pm 2.94}$	$18.48_{\pm 4.06}$	$1.31_{\pm 0.01}$	$1.10_{\pm 0.02}$
TNAM	$37.01_{\pm 1.35}$	$37.01_{\pm 1.33}$	$5.89_{\pm 0.05}$	$2.52_{\pm 0.01}$	$1.36_{\pm 0.01}$	$1.15_{\pm 0.01}$
BPR-T	$34.11_{\pm 0.32}$	$31.57_{\pm 0.50}$	$33.94_{\pm 0.18}$	$19.75_{\pm 0.28}$	$1.23_{\pm 0.00}$	$0.98_{\pm 0.01}$

Table 8: The benchmarking results of the feature-based explainable recommender models on jointly predicting the tags for persuasiveness, informativeness and satisfaction and ratings.

Metrics	Persuasiveness		Informativeness		Satisfaction		Rating Prediction	
	F1	NDCG	F1	NDCG	F1	NDCG	RMSE	MAE
EFM	$11.66_{\pm0.15}$	$8.52_{\pm0.10}$	$4.97_{\pm0.61}$	$3.70_{\pm0.48}$	$5.33_{\pm1.10}$	$2.24_{\pm0.62}$	$1.66_{\pm0.01}$	$1.23_{\pm0.01}$
LRPPM	$37.32_{\pm0.24}$	$23.26_{\pm0.11}$	$37.94_{\pm0.23}$	$38.46_{\pm0.18}$	$36.08_{\pm0.17}$	$22.35_{\pm0.10}$	$1.22_{\pm0.00}$	$0.96_{\pm0.00}$
SULM	$\mathbf{41.12}_{\pm0.50}$	$\mathbf{25.35}_{\pm0.21}$	$\mathbf{42.35}_{\pm0.45}$	$\mathbf{42.66}_{\pm0.34}$	$\mathbf{40.00}_{\pm0.49}$	$\mathbf{24.66}_{\pm0.16}$	$1.69_{\pm0.09}$	$1.33_{\pm0.08}$
MTER	$36.16_{\pm0.06}$	$22.38_{\pm0.15}$	$36.75_{\pm0.09}$	$36.94_{\pm0.24}$	$34.84_{\pm0.09}$	$21.52_{\pm0.03}$	$1.34_{\pm0.04}$	$1.08_{\pm0.03}$
AMF	$27.83_{\pm0.37}$	$17.47_{\pm0.19}$	$28.35_{\pm0.33}$	$28.66_{\pm0.34}$	$27.09_{\pm0.34}$	$17.03_{\pm0.16}$	$2.28_{\pm0.00}$	$1.86_{\pm0.00}$
DERM-MLP	$38.60_{\pm0.08}$	$23.81_{\pm0.07}$	$39.33_{\pm0.09}$	$39.57_{\pm0.05}$	$37.52_{\pm0.09}$	$22.97_{\pm0.09}$	$1.31_{\pm0.02}$	$1.07_{\pm0.02}$
DERM-MF	$37.42_{\pm0.21}$	$23.16_{\pm0.08}$	$38.26_{\pm0.13}$	$38.46_{\pm0.16}$	$36.60_{\pm1.01}$	$22.18_{\pm0.19}$	$1.33_{\pm0.00}$	$1.16_{\pm0.00}$
DERM-C	$38.05_{\pm0.22}$	$23.53_{\pm0.07}$	$39.03_{\pm0.15}$	$39.29_{\pm0.11}$	$37.19_{\pm0.15}$	$22.79_{\pm0.08}$	$1.30_{\pm0.01}$	$1.08_{\pm0.01}$
DERM-H	$37.64_{\pm0.24}$	$23.36_{\pm0.18}$	$38.52_{\pm0.44}$	$38.83_{\pm0.39}$	$36.70_{\pm0.40}$	$22.60_{\pm0.16}$	$1.28_{\pm0.01}$	$1.05_{\pm0.02}$
TRDM	$33.50_{\pm1.85}$	$15.64_{\pm1.83}$	$33.94_{\pm1.06}$	$26.47_{\pm1.89}$	$31.79_{\pm1.17}$	$14.77_{\pm1.09}$	$\mathbf{1.19}_{\pm0.00}$	$\mathbf{0.94}_{\pm0.01}$
TRAL	$5.88_{\pm0.14}$	$2.56_{\pm0.04}$	$5.91_{\pm0.09}$	$4.48_{\pm0.17}$	$35.42_{\pm0.20}$	$21.09_{\pm0.08}$	$1.29_{\pm0.00}$	$1.07_{\pm0.00}$
HPTR	$38.82_{\pm0.38}$	$22.14_{\pm1.49}$	$39.31_{\pm0.50}$	$38.01_{\pm2.21}$	$37.64_{\pm0.26}$	$21.46_{\pm1.51}$	$1.77_{\pm0.98}$	$1.50_{\pm0.93}$
AIRec	$38.89_{\pm0.09}$	$23.98_{\pm0.05}$	$39.61_{\pm0.11}$	$39.82_{\pm0.09}$	$37.85_{\pm0.05}$	$23.16_{\pm0.06}$	$1.30_{\pm0.00}$	$1.07_{\pm0.01}$
HAN-TR	$37.95_{\pm0.63}$	$23.46_{\pm0.02}$	$38.65_{\pm0.70}$	$39.01_{\pm0.39}$	$37.07_{\pm0.82}$	$22.74_{\pm0.04}$	$1.80_{\pm0.67}$	$1.57_{\pm0.62}$
TNAM	$37.96_{\pm0.13}$	$23.51_{\pm0.02}$	$38.74_{\pm0.23}$	$38.81_{\pm0.17}$	$5.89_{\pm0.08}$	$2.44_{\pm0.12}$	$1.37_{\pm0.01}$	$1.17_{\pm0.01}$
BPR-T	$33.41_{\pm0.44}$	$19.39_{\pm0.48}$	$34.37_{\pm0.28}$	$32.15_{\pm0.36}$	$34.29_{\pm0.15}$	$20.11_{\pm0.22}$	$1.24_{\pm0.01}$	$1.00_{\pm0.02}$

Table 9: The benchmarking results of the models with natural language explanations in our library. For BLEU and ROUGE, the results are percentage values with "%" omitted. "-" means the evaluation metric is not available for the model.

Metrics	BLEU (%)		ROUGE-1 (%)			ROUGE-2 (%)		
	B-1	B-4	F1	R	P	F1	R	P
Att2Seq	$\mathbf{19.96}_{\pm0.27}$	$\mathbf{3.25}_{\pm0.19}$	$\mathbf{22.13}_{\pm0.27}$	$\mathbf{19.73}_{\pm0.44}$	$26.40_{\pm0.78}$	$\mathbf{5.56}_{\pm0.08}$	$\mathbf{5.19}_{\pm0.16}$	$6.26_{\pm0.24}$
NRT	$17.67_{\pm1.10}$	$2.92_{\pm0.65}$	$20.60_{\pm0.57}$	$16.04_{\pm1.27}$	$\mathbf{30.02}_{\pm2.40}$	$5.23_{\pm0.56}$	$4.20_{\pm0.77}$	$\mathbf{7.33}_{\pm0.42}$
PETER	$17.65_{\pm1.18}$	$2.35_{\pm0.35}$	$20.00_{\pm1.07}$	$15.68_{\pm1.62}$	$28.59_{\pm1.42}$	$4.99_{\pm0.48}$	$3.95_{\pm0.58}$	$7.00_{\pm0.57}$

References

[1] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–725, 2017.

[2] Bo Chen, Yue Ding, Xin Xin, Yunzhe Li, Yule Wang, and Dong Wang. Airec: Attentive intersection model for tag-aware recommendation. *Neurocomputing*, 421:105–114, 2021.

[3] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. Learning to rank features for recommendation over multiple categories. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 305–314, 2016.

[4] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, 2017.

[5] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1661–1670, 2015.

[6] Yunfeng Hou, Ning Yang, Yi Wu, and Philip S Yu. Explainable recommendation with fusion of aspect information. *World Wide Web*, 22:221–240, 2019.

[7] Ruoran Huang, Nian Wang, Chuanqi Han, Fang Yu, and Li Cui. Tnam: A tag-aware neural attention model for top-n recommendation. *Neurocomputing*, 385:1–12, 2020.

[8] Hongmei Li, Xingchun Diao, Jianjun Cao, Lei Zhang, and Qin Feng. Tag-aware recommendation based on bayesian personalized ranking and feature mapping. *Intelligent data analysis*, 23(3):641–659, 2019.

[9] Lei Li, Yongfeng Zhang, and Li Chen. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*, 2021.

[10] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 345–354, 2017.

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[13] Jianshan Sun, Mingyue Zhu, Yuanchun Jiang, Yezheng Liu, and Le Wu. Hierarchical attention model for personalized tag recommendation. *Journal of the Association for Information Science and Technology*, 72(2):173–189, 2021.

[14] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 165–174, 2018.

[15] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92, 2014.

[16] Weibin Zhao, Aoran Zhang, Lin Shang, Yonghong Yu, Li Zhang, Can Wang, Jiajun Chen, and Hongzhi Yin. Hyperbolic personalized tag recommendation. In *International Conference on Database Systems for Advanced Applications*, pages 216–231. Springer, 2022.

[17] Yi Zuo, Shengzong Liu, Yun Zhou, and Huanhua Liu. Tral: A tag-aware recommendation algorithm based on attention learning. *Applied Sciences*, 13(2):814, 2023.

[18] Yi Zuo, Yun Zhou, Shengzong Liu, and Yupeng Liu. A tag-aware recommendation algorithm based on deep learning and multi-objective optimization. In *2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA)*, pages 42–46. IEEE, 2023.