Question-Attentive Review-Level Explanation for Neural Rating Regression

TRUNG-HOANG LE and HADY W. LAUW, School of Computing and Information Systems, Singapore Management University, Singapore

Recommendation explanations help to improve their acceptance by end users. Explanations come in many different forms. One that is of interest here is presenting an existing review of the recommended item as the explanation. The challenge is in selecting a suitable review, which is customarily addressed by assessing the relative importance or "attention" of each review to the recommendation objective. Our focus is improving review-level explanation by leveraging additional information in the form of questions and answers (QA). The proposed framework employs QA in an attention mechanism that aligns reviews to various QAs of an item and assesses their contribution jointly to the recommendation objective. The benefits are two-fold. For one, QA aids in selecting more useful reviews. For another, QA itself could accompany a well-aligned review in an expanded form of explanation. Experiments on datasets of ten product categories showcase the efficacies of our method as compared to comparable baselines in identifying useful reviews and QAs, while maintaining parity in recommendation performance.

CCS Concepts: • **Information systems → Recommender systems**; • **Computing methodologies → Neural networks**.

Additional Key Words and Phrases: neural rating regression, recommendation explanation, review-level explanation, question-level explanation

ACM Reference Format:
Trung-Hoang Le and Hady W. Lauw. 2024. Question-Attentive Review-Level Explanation for Neural Rating Regression. *ACM Trans. Intell. Syst. Technol.* 1, 1 (September 2024), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

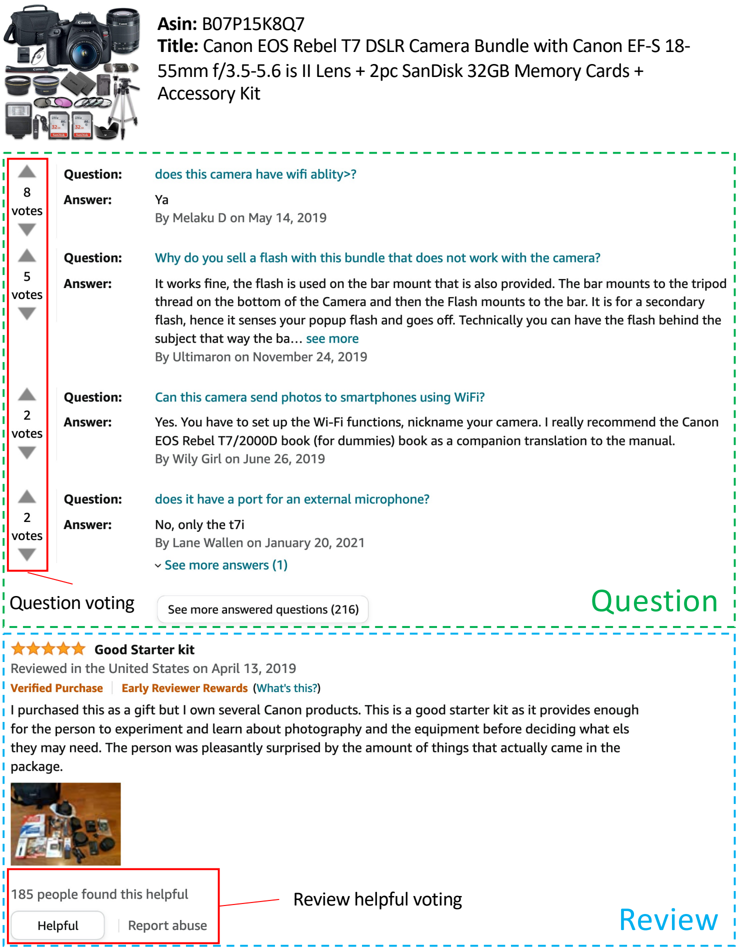
A ubiquitous feature of Web applications and e-commerce marketplaces is a recommender system that aids users in navigating the multitude of options available, be they products to purchase, social media posts to view, movies to watch, etc. The most common framework is that of collaborative filtering [18], predicting ratings or adoptions based on users' past interactions with various items.

Earlier in the evolution of recommender systems, the concern was predominantly on achieving higher accuracies [14, 38]. Of late, the concern shifts to greater interpretability and explainability, as ultimately the goal is to get users to adopt the recommendations. This gives rise to a plethora of explainable recommendation models [52], which seek to produce not only recommendations, but also accompanying explanations. There are diverse forms of explanations, leveraging different types of information associated with either users or items.

For a pertinent instance, we allude to *review-level explanation*, whereby the explanation to a recommendation takes the form of a review, selected from the existing reviews of the product.

Authors' address: Trung-Hoang Le, thle.2017@smu.edu.sg; Hady W. Lauw, hadywlauw@smu.edu.sg, School of Computing and Information Systems, Singapore Management University, Singapore.

2157-6904/2024/9-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX



Asin: B07P15K8Q7
Title: Canon EOS Rebel T7 DSLR Camera Bundle with Canon EF-S 18-55mm f/3.5-5.6 is II Lens + 2pc SanDisk 32GB Memory Cards + Accessory Kit

Question: does this camera have wifi ablity>?
Answer: Ya
By Melaku D on May 14, 2019

Question: Why do you sell a flash with this bundle that does not work with the camera?
Answer: It works fine, the flash is used on the bar mount that is also provided. The bar mounts to the tripod thread on the bottom of the Camera and then the Flash mounts to the bar. It is for a secondary flash, hence it senses your popup flash and goes off. Technically you can have the flash behind the subject that way the ba… see more
By Ultimaron on November 24, 2019

Question: Can this camera send photos to smartphones using WiFi?
Answer: Yes. You have to set up the Wi-Fi functions, nickname your camera. I really recommend the Canon EOS Rebel T7/2000D book (for dummies) book as a companion translation to the manual.
By Wily Girl on June 26, 2019

Question: does it have a port for an external microphone?
Answer: No, only the t7i
By Lane Wallen on January 20, 2021

Good Starter kit
Reviewed in the United States on April 13, 2019
Verified Purchase | Early Reviewer Rewards (What's this?)
I purchased this as a gift but I own several Canon products. This is a good starter kit as it provides enough for the person to experiment and learn about photography and the equipment before deciding what els they may need. The person was pleasantly surprised by the amount of things that actually came in the package.

185 people found this helpful

Fig. 1. An example product on Amazon.com with reviews, questions and answers

An insightful review, when presented with a recommended product, allows the recipient of the recommendation to empathize with the hands-on experience of the reviewer, thus anticipating what her own experience with the product would be. For instance, on Amazon.com, Canon EOS Rebel T7 Bundle[1] has more than 2800 ratings, more than 300 of which have reviews. One of these reviews is illustrated in Figure 1, relating to the quality of the starter kit. That popular products may have many reviews (some to the tune of tens of thousands) is a dual-edged sword. With a rich corpus for selection comes the problem of selecting which review to present as explanation. One existing paradigm [3, 28] is to weigh the contribution of various reviews to the recommendation.

Given the abundance of reviews, there is a proclivity to employ reviews to aid recommendations. Most of the works are intent on improving recommendation accuracy rather than to serve directly as explanations. These include content-based methods based on topic models [43], sentiments [8], social networks [39]. By using convolutional neural network, [55] encodes all reviews on an item to represent that item and all reviews written by a user to represent that user to enhance rating

[1] https://www.amazon.com/Canon-T7-18-55mm-3-5-5-6-Accessory/dp/B07P15K8Q7/

Table 1. Main Notations

Symbol	Description
$\mathcal{U}, \mathcal{P}$	set of all users and products
$\mathcal{T}, \mathcal{Q}, \mathcal{A}$	set of all reviews, questions, and answers
$t_{ij} \in \mathcal{T}$	a review of user i on product j
$\mathcal{Q}_j$	a set of all questions on product j
$q_{jk} \in \mathcal{Q}_j$	a question k of product j
$\mathcal{A}_{jk}$	a set of all answers on question q_{jk}
$a_{jkl} \in \mathcal{A}_{jk}$	an answer l of a question k on product j
$\xi(t_{ij}), \xi(q_{jk}), \xi(a_{jkl})$	embedded matrices of t_{ij}, q_{jk}, and a_{jkl}
$\zeta_u(i), \zeta_p(j)$	latent features of user i and product j
$O_{t_{ij}}, O_{q_{jk}}, O_{a_{jkl}}$	feature vectors extracted from t_{ij}, q_{jk}, and a_{jkl}
u_i, p_j	rating-based representation of user i and product j
α_{ij}	attention weights for $O_{t_{ij}}$
β_{ijk}	attention weight of review t_{ij} on question q_{jk}
δ_{jkl}	attention weight of answer a_{jkl} on question q_{jk}
ω_{jk}	QA representation of question q_{jk} after infusing answers
d_{jk}	document representation respecting to q_{jk} after infusing reviews
γ_{jk}	attention weight of document d_{jk}
b_u, b_i, μ	user bias, item bias, and global bias respectively

prediction. [45] learns to focus on a few reviews of users and items optimizing for rating prediction. In contrast to works that see reviews as content to help recommendation accuracy, we focus on the role of reviews as explanations.

In this work, we propose to go beyond reviews and incorporate other information associated with a product. One that is a focus of this work is a question posted by a user that in turn attracts answers from other users, hereinafter referred to in short form as QA. For instance, the same product Canon EOS Rebel T7 bundle featured in Figure 1 has more than 200 questions. Among them are whether the camera has wifi ability (*answer: yes*), whether there is a port for an external microphone (*answer: no, but another model T7i does*), and whether it is suitable for indoor sports (*answer: yes, it has a sport mode*). Similarly to reviews, QAs could also receive votes from users.

Interestingly, questions and their answers present a distinct yet complementary information to reviews. Where reviews tend to be subjective and replete with opinions, questions tend to be objective and inquisitive of factual concerns. Where a single review tends to be multi-faceted and comprehensive, each question tends to be concise and narrowly focused on a single aspect. Given this complementarity, we postulate that both QA and review could collectively serve as recommendation explanations. The former notifies the recommendee of relevant factual concern(s), while the latter gains the recommendee insights from a reviewer's experience.

QA as a feature is also increasingly prevalent across many platforms, with Amazon.com and Tripadvisor.com being a couple of prominent examples. For instance, across the ten product categories in our datasets (see Section 4), between 13% to 56% of products have QA information. Given the anticipated further increase in QA data over time, it is timely to consider how to leverage QA in addition to reviews for more informative recommendation explanations.

Problem. Let $\mathcal{U}$ be a set of users, and $\mathcal{P}$ be a set of products. A user $i \in \mathcal{U}$ assigns to a product $j \in \mathcal{P}$ a rating $r_{ij} \in \mathbb{R}_+$ along with a review t_{ij}. We denote the collection of ratings as $\mathcal{R}$, all reviews as $\mathcal{T}$, the subset of reviews concerning a product j as $\mathcal{T}_j$. Product j may have a set of questions $\mathcal{Q}_j = \{q_{j1}, q_{j2}, ..., q_{jK}\} \subset \mathcal{Q}$, where K is the total number of questions of product j. Each question q_{jk} has a collection of answers $\mathcal{A}_{jk} = \{a_{jk1}, a_{jk2}, \ldots, a_{jkL}\}$, where L is the total number of answers of question q_{jk}. Table 1 lists the notations (some to be introduced later). The problem can thus be stated as follows. Receiving as input users $\mathcal{U}$, products $\mathcal{P}$, ratings $\mathcal{R}$, reviews $\mathcal{T}$, and question-answer pairs Q, we seek a model capable of predicting a missing rating by a user i on product j for recommendation (rating regression), as well as identifying a question-answer pair (selected from Q_j) along with a review (selected from $\mathcal{T}_j$) to serve collectively as explanations accompanying the recommendation.

Due to the differing yet complementary natures of QA and reviews, we design a neural attention model, called QuestER, that operates at two levels. First, the concise QA serves as focal points of attention representing salient aspects to a product recommendation. Second, the multi-faceted nature of reviews means that they could be relevant to multiple aspects, and we model their relative importance to each QA. Together, QA and reviews serve dual roles in a hand-in-hand manner: to contribute content features to aid recommendation and to serve as explanations.

Contribution. We make several contributions. *First*, we incorporate product questions into an attention mechanism on reviews for recommendation. *Second*, we develop a neural model called QUESTion-attentive review-level Explanation for neural rating Regression or QuestER, which considers questions as a source of alignment to textual review. An important question would help to identify important reviews. *Third*, we conduct comprehensive experiments on ten product categories against comparable baselines. Importantly, we find that not only do QAs help in identifying useful reviews, but the expanded explanation that is the combination of QA and review also has value.

This manuscript is an extension of the conference version [22], and it differs with following additional contributions:

- Instead of treating answer as a part of question text, the model architecture of QuestER has been extended to include another attention layer for answers to contribute to the question representation. This is used in replacement of the question encoding.
- We include extensive experimental comparisons on ten product categories (Home, Health, Sport, Toy, Grocery, Baby, Office, Automotive, Patio, and Musical). In contrast, [22] reports only three product categories (Home, Sport, and Musical). These additional results provide comprehensive coverage of how the method applies across a wide range of product domains.
- We expand the experiments with discussions on additional metrics new to this manuscript, including the effect of the number of answers in each question towards the final rating predictions, as well as the performance of both review-level and question-level explanations. These give a more all-rounded coverage of the performance of the proposed method.
- In addition to quantitative experiments, we now include user studies that examine the quality of both review-level as well as question-level explanations, as well as comparison to top-rated reviews. For illustration, we also produce more case studies on more domains. The resulting analyses thoroughly examine the effectiveness of the proposed methods.

2 RELATED WORK

We survey related work that deals with questions or reviews in the context of recommendations.

QA-Based Recommendation. The use of QA for recommendation is still relatively rare in the literature. One is to detect a user's propensity to purchase a product based on the question that the user has submitted [4]. This is a distinct scenario from ours where the question does not have to be posed by the recipient of recommendations. Rather, we see questions as additional product information that may be relevant as explanation. QA-based recommendation is also orthogonal from question answering task. [54] selects relevant sentences in product reviews to answer a question. [5] identifies answers from product reviews for user questions by multi-task attentive networks. [51] incorporates aspect on reviews for predicting answer of a yes-no question. Our goal is not to answer questions, rather to select QA appropriate for recommendation explanations.

Review-Based Recommendation. Given the abundance of reviews, there is a proclivity to employ reviews to aid recommendations. Most of the works are intent on improving recommendation accuracy rather than to serve directly as explanations. These include content-based methods based on similarity metric [35], topic models [43], sentiments [8], social networks [39]. By using convolutional neural network, [55] encodes all reviews on an item to represent that item and all reviews written by a user to represent that user to enhance rating prediction. [45] learns to focus on a few reviews of users and items optimizing for rating prediction. [27] treats reviews with different polarities for rating prediction. In contrast to works that see reviews as content to help recommendation accuracy, we focus on the role of reviews as explanations.

Review-Based Recommendation Explanation. Our work belongs to a group that uses a whole review as explanation. We identify a few in this group and compare to them as baselines. NARRE [3] uses attention to weigh each individual review toward user and item representation and uses the most useful review(s) as review-level explanation. HRDR [28] uses multilayer perceptron to encode user's ratings (resp. item's ratings) as user features (resp. item's features) and use that as query for attention layer to weight the contribution of each review to rating prediction. HFT [32] could select the review with the closest topic distribution to the item's topic distribution. Our key distinction from these baselines is our unique incorporation of QA both for review selection and explanation. Another work uses three-tier attention [48] on word-level, sentence-level, and review-level for learning user text representation and item text representation towards the final rating prediction. [49] extends three-tier attention with graphs.

Rather than relying on review-level explanations, some works extract segments [29, 34, 42] or aspect-level sentiments [13, 21, 47, 53]. Another formulation is to select personalized review [2, 7, 10, 15], the selected personalized review for every user for a given item may vary, which is orthogonal to selecting useful reviews, the selected useful review for the same item are identical. [7] uses GRU as text encoder to encode word-level and review-level representation and learn the contribution of each word/review to the rating prediction. [15] selects personalized review based on extracted aspects. [10] employs bi-directional LSTM [37] to learn embedding for user and item's sentences in textual review then applies asymmetric attentive modules that the text on item side contribute to the text of user side.

Review Generation. A few try to predict ratings and generate reviews in multi-task learning manner [6, 23–25, 46, 50]. [25] uses the predicted rating as sentiment, along with user and item factors as context to generate explanation text. [6] extends with the attention on concepts from an oracle[2]. [46] further attends on visual aspects. [23] explicitly uses aspect keywords to generate explanation. [24] uses Transformer, a well-known language modeling technique, for personalized review generation. For further informative and factually explanation generation, [50] augments the review generator with external knowledge from another personalized retriever model that estimates the personalized review embedding for each user. We are concerned with the selection of existing reviews, rather than their generation.

Review Quality. We focus on the recommendation scenario. There are other formulations that seek to predict helpful reviews [9, 30, 31, 41, 44]. In those cases, the concern is with objective review quality. In contrast, this work concerns how a review is aligned to recommendation, and thus could serve as an explanatory device.

[2] https://concept.research.microsoft.com/

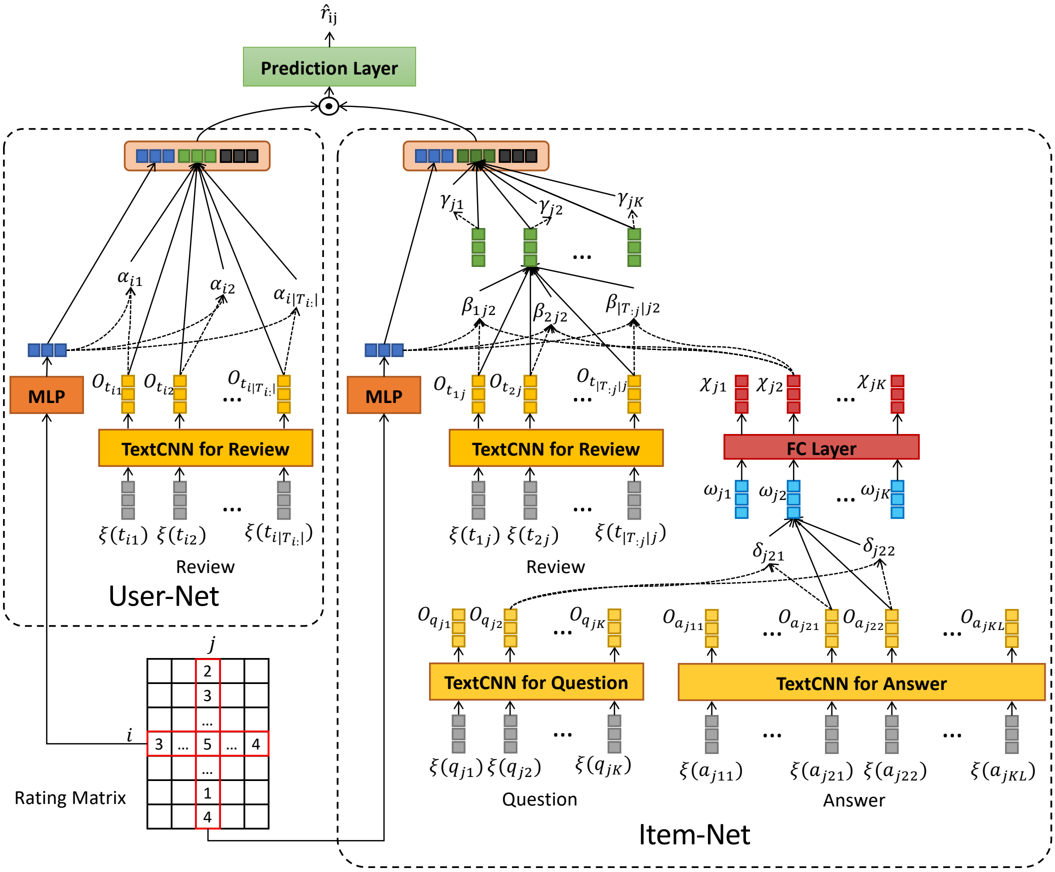


Fig. 2. The architecture of QuestER model

3 METHODOLOGY

Our formulation in having a pair of QA and review to accompany recommendation based on rating regression is novel. We hypothesize that the concise questions could serve as an attention mechanism in weighing the importance of reviews. This achieves an alignment between questions and reviews, potentially allowing expanded explanations that are more comprehensive and coherent.

The overall architecture of our proposed QuestER model is shown in Figure 2. Below we describe its various components.

Text Encoder. We use a widely adopted CNN text processor [2, 3, 28, 55], named TextCNN, for encoding to extract semantic features of text. TextCNN consists of a Convolutional Neural Network (CNN) followed by max pooling and a fully connected layer (see Figure 3). Particularly, we have a word embedding function $\xi : M \rightarrow \mathbb{R}^D$ to map each word in the text t into a D-dimensional vector, forming an embedded matrix $\xi(t)$ with fixed length W (padded zero for text with length $< W$). Following this embedding layer is a convolutional layer with m neurons, each associated with a filter $F \in \mathbb{R}^{w \times D}$, each k^{th} neuron produces features by applying convolution operator on the embedded matrix $\xi(t)$:

$$z_k = ReLU(\xi(t) * F_k + b_z) \qquad (1)$$

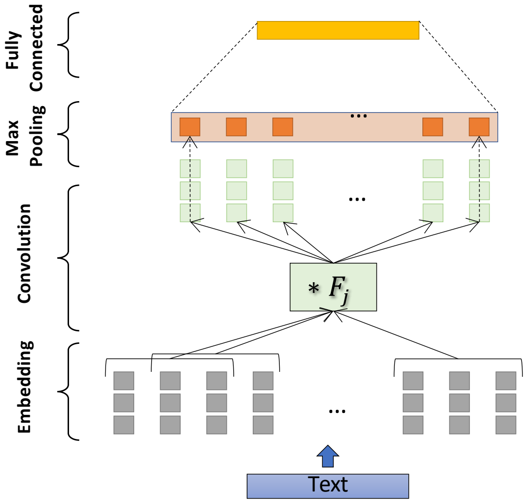


Fig. 3. The CNN Text Processor (TextCNN) architecture

where $ReLU(x) = \max(x, 0)$ is a nonlinear activation function and $*$ is the convolution operation. With sliding window w, the produced features would be $z_1, z_2, ..., z_k^{W-w+1}$, which are passed to a max pooling to capture the most important features having highest values, which is defined as:

$$o_k = \max(z_1, z_2, ..., z_k^{W-w+1}) \tag{2}$$

We get the final output of the convolutional layer by concatenating all output from m neurons, $O = [o_1, o_2, ..., o_m]$. A simple approach to get the final representation of the input text t is to pass O into a fully connected layer as follows:

$$X = WO + b \tag{3}$$

Besides TextCNN, there are other text processing methods based on deep learning technology that have been proposed and have claimed advantages over traditional methods, such as fastText [16], RNN and paragraph vector [20], etc. Especially, the recent Large-scale Pre-trained Language Models (PTM) [11], such as BERT, perform well on a variety of natural language processing (NLP) tasks. An analysis of both TextCNN and BERT as text encoder is reported in subsection 4.7.

Rating Encoder. Ratings are explicit features provided by users to indicate their interest on given items. The user ratings $r_{i:}$ form a rating pattern for user i, and the item ratings $r_{:j}$ form a rating pattern for item j. A reasonable choice is to use a multi-layer perceptron (MLP) network to learn the representation for the rating pattern [28] (see Figure 4). Specifically,

$$\begin{aligned} h_{i1} &= \tanh(W_{r_{i:}1} r_{i:} + b_{r_{i:}1}) \\ h_{i2} &= \tanh(W_{r_{i:}2} h_{i1} + b_{r_{i:}2}) \\ &... \\ u_i &= \tanh(W_{r_{i:}k} h_{i(k-1)} + b_{r_{i:}k}) \end{aligned} \tag{4}$$

The output u_i is the final rating-based representation of user i, h_{ik} is the output hidden representation at layer k of the MLP. Similarly, we can also get the rating-based representation p_j of product j from its input ratings $r_{:j}$ in similar manner. We use tanh as activation function to project the learned rating-based representation into the same range of text-based representations that will be discussed in the following paragraphs.

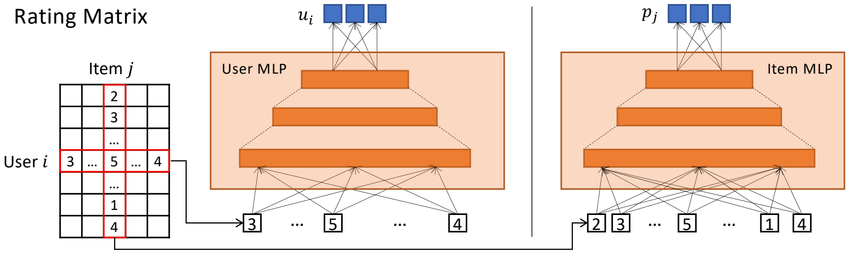


Fig. 4. Multi-Layer Perceptron for user ratings and item ratings encoder

User Attention-Based Review Pooling. Equation 3 presumes that the contribution of each review is the same towards the final representation. The importance of each individual review contributing to user final representation is learnt as follows:

$$\rho_{ij} = \tanh(W_{O_t}(O_{t_{ij}} \odot u_i) + b_\rho) \tag{5a}$$

$$\theta_{ij} = W_\rho \rho_{ij} + b_\theta \tag{5b}$$

$$\alpha_{ij} = \frac{e^{\theta_{ij}}}{\sum_j e^{\theta_{ij}}} \tag{5c}$$

where $\odot$ is element-wise multiplication operator, u_i is the rating-based representation of the user i, $O_{t_{ij}}$ is the feature vector extracted from review text t_{ij} by TextCNN, α_{ij} is the normalized attention score of the review t_{ij}, which can be interpreted as the contribution of that review to the feature profile O_i of user i, aggregating as follows:

$$O_i = \sum_j \alpha_{ij} O_{t_{ij}} \tag{6}$$

The final representation of user i is computed as follows:

$$X_i = W_{O_i} O_i + b_X \tag{7}$$

Item Question-Attentive Review-Level Explanations. Of particular importance is our modeling of product questions. A naive approach to model question on item side is to apply similar approach of modeling reviews. However, the connection between reviews and questions would have been overlooked. Here we presume that a product review may contain information that could be relevant to a question. We aggregate another attention layer based on item questions that help us to incorporate reviews based on their contribution towards item questions.

First, we use TextCNN to encode reviews, questions, answers. Let $O_{t_{ij}}$ be the review encoding, $O_{q_{jk}}$ be the question encoding of the question k on the product j, and $O_{a_{jkl}}$ be the answer encoding of the answer l of the question k. With respect to each question representation $O_{q_{jk}}$, we learn the attention weights δ_{jkl} for answer representation $O_{a_{jkl}}$ by projecting both question and answer representation onto an attention space followed by a non-linear activation function; the outputs are ϕ_{jk} and ψ_{jkl} respectively. We use tanh activation function to scale $O_{q_{jk}}$ and $O_{a_{jkl}}$ to the same range of values, so that neither component dominates the other. We let the question projection ϕ_{jk} interact with the answer projection ψ_{jkl} in two ways: element-wise multiplication and summation. The learned vector V plays the role of global attention context. This produces an attention value

v_{jkl}, which is normalized using softmax to obtain δ_{jkl}.

$$\phi_{jk} = \tanh\left(W_{O_q} O_{q_{jk}} + b_{\phi}\right) \tag{8a}$$

$$\psi_{jkl} = \tanh\left(W_{O_a} O_{a_{jkl}} + b_{\psi}\right) \tag{8b}$$

$$v_{jkl} = V^T \left(\psi_{jkl} \cdot \phi_{jk} + \psi_{jkl}\right) \tag{8c}$$

$$\delta_{jkl} = \frac{e^{\frac{v_{jkl}}{\tau}}}{\sum_l e^{\frac{v_{jkl}}{\tau}}} \tag{8d}$$

$$\tag{8e}$$

Where τ is a temperature parameter to adjust the probabilities in the softmax. We aggregate the answer representations $O_{a_{ijk}}$'s into each question representation ω_{jk} using the learned attention δ_{jkl}:

$$\omega_{jk} = \sum_l \delta_{jkl} O_{a_{jkl}} \tag{9}$$

Analogously, we learn the attention weight β_{ijk} for review representation $O_{t_{ij}}$ by projecting both question representation ω_{jk} and review representation onto an attention space followed by a non-linear activation function; the outputs are χ_{jk} and ρ'_{ij} respectively. To learn the question-specific attention weight of a review, we let the question projection χ_{jk} interact with the review projection ρ'_{ij} in two ways: element-wise multiplication and summation. The learned vector E plays the role of global attention context. This produces an attention value η_{ijk}, which is normalized using softmax to obtain β_{ijk}:

$$\chi_{jk} = \tanh\left(W_{\omega} \omega_{jk} + b_{\chi}\right) \tag{10a}$$

$$\rho'_{ij} = \tanh\left(W_{O_t} (O_{t_{ij}} \odot p_j) + b_{\rho'}\right) \tag{10b}$$

$$\eta_{ijk} = E^T \left(\chi_{jk} \odot \rho'_{ij} + \rho'_{ij}\right) \tag{10c}$$

$$\beta_{ijk} = \frac{e^{\frac{\eta_{ijk}}{\tau}}}{\sum_i e^{\frac{\eta_{ijk}}{\tau}}} \tag{10d}$$

Using the question-specific attention weights β_{ijk}, we aggregate the review representations $O_{t_{ij}}$'s into a question-specific representation d_{jk} as follows.

$$d_{jk} = \sum_i \beta_{ijk} O_{t_{ij}} \tag{11}$$

For a document (a product question with all of its reviews), we apply this attention mechanism for every product question, yielding a set of question-specific document representations d_{jk}, $k \in [1, |Q_j|]$. All the d_{jk}'s need to be aggregated into the final document representation O_j before incorporating to product representation. Thus, we seek to learn the importance weight γ_{jk}, signifying how each question-specific representation d_{jk} would contribute to O_j.

$$\kappa_{jk} = K^T \tanh(W_{d_{jk}} d_{jk} + b_{\kappa}) \tag{12a}$$

$$\gamma_{jk} = \frac{e^{\frac{\kappa_{jk}}{\tau}}}{\sum_k e^{\frac{\kappa_{jk}}{\tau}}} \tag{12b}$$

Question-specific representation d_{jk} is projected into attention space through a layer of neurons with non-linear activation function tanh. The scalar κ_{jk} indicates the importance of d_{jk}, obtained by multiplying with global attention context vector K (randomly initialized and learned during training). The representation d_{jk}'s due to the various questions are aggregated into the final product representation O_j using soft attention pooling with attention weight γ_{jk}'s.

$$O_j = \sum_k \gamma_{jk} d_{jk} \tag{13a}$$

$$Y_j = W_{O_j} O_j + b_Y \tag{13b}$$

Prediction Layer. The latent factors of user i and product j are mapped to a shared hidden space as follows:

$$h_{ij} = [u_i; X_i; \zeta_u(i)] \odot [p_j; Y_j; \zeta_p(j)] \tag{14}$$

where $\zeta_u(\cdot)$ and $\zeta_p(\cdot)$ are embedding function to map each user and each product into their embedding space respectively, X_i is user preferences and Y_j is item features obtained from user reviews and product reviews and questions, $[u_i; X_i; \zeta_u(i)]$ is the concatenation of user rating-based representation u_i, user text attention review pooling X_i, and user i embedding $\zeta_u(i)$. The final rating prediction is computed as follows:

$$\hat{r}_{ij} = W^T h_{ij} + b_i + b_j + \mu \tag{15}$$

Learning. Similar to prior works on rating prediction task [3, 28, 43], which is a regression problem, we adopt the squared loss function:

$$\mathcal{L} = \sum_{i,j \in \Omega} (\hat{r}_{ij} - r_{ij})^2 \tag{16}$$

Where Ω denotes the set of all training instances, r_{ij} is the ground truth rating that user i assigned on product j.

The most important question $\mathbb{Q}$ is selected by computing $\mathbb{Q} = argmax_k(\gamma_{jk})$ and the most useful review is selected by $argmax_i(\beta_{ij\mathbb{Q}})$. We use the selected question with its answer and the selected review collectively as explanation for a given recommendation.

A limitation of relying only on questions found within a product is that product features may not be captured completely, because some products do not have sufficient questions to cover all its important aspects. As a result, an important review may be overlooked because it does not correspond to any question. To address this limitation, in addition to the questions found in a product, we include one more global "General Question", which allows those important reviews to still be aligned. This additional question plays the role of "global" aspect, and also helps our model to potentially generalize to product without questions.

4 EXPERIMENTS

As this work is primarily about recommendation explanations, rather than rating prediction per se, and the two objectives are not necessarily directionally equivalent, our orientation is to improve explanations while maintaining parity in accuracy performance. In particular, our core contribution is in incorporating question and answer or QA for review-level explanation. The experimental objectives revolve around the utility of QA as part of explanation, the effectiveness of QA to aid the selection of review-level explanation, and the alignment of QA and review that are part of an explanation. Source code is available for reproducibility[3].

[3] https://github.com/PreferredAI/QuestER

Table 2. Data statistics

Dataset	#Item	#User	#Review (Rating)	#Question	#Answer	$\frac{\text{\#Item with Question}}{\text{\#Item}}$	$\frac{\text{\#Answer}}{\text{\#Question}}$
Home	28,169	66,295	549,895	368,904	1,079,983	0.3193	2.93
Health	18,464	38,416	344,888	105,814	207,330	0.1731	1.96
Sport	18,301	35,447	295,074	123,119	237,845	0.1940	1.93
Toy	11,870	19,322	166,821	35,520	75,276	0.1463	2.12
Grocery	8,690	14,632	150,802	18,134	42,779	0.1301	2.36
Baby	7,039	19,418	160,521	32,507	58,345	0.1301	1.79
Office	2,414	4,892	53,143	68,864	165,623	0.4544	2.41
Automotive	1,810	2,892	20,203	40,477	79,034	0.3470	1.95
Patio	951	1,667	13,133	22,454	53,550	0.3049	2.38
Musical	893	1,416	10,163	22,409	47,357	0.5622	2.11

Datasets. Towards reproducibility, we work with publicly available sources. While QA is a feature on many platforms, not many such datasets have both reviews and QA information. One that does is the Amazon Product Review Dataset[4] [12]. We experiment on ten product categories from this source as separate instances. These categories are selected for significant availability of QA information. Consistent performance across multiple categories with different statistics bolster the analysis. Table 2 summarizes basic statistics of the ten datasets.

For greater coverage, we collect item questions and acquire their helpful voting scores from the Amazon.com website[5]. These questions data are complement yet distinct from [33], as they do not include helpful voting scores for every question and answer. Too short reviews (less than 3 words), users and items with fewer than five reviews are filtered out. To aggregate overlapping questions, we cluster questions in each category with KMeans, keeping questions from big clusters which cover 80% of questions. For smaller clusters, we keep the nearest question to each cluster centroid and combine them into a single text, called General Question (all products have this by default). This is used solely for modeling to generalize to items without questions, but would not be used as a recommendation explanation. Moreover, a question will always be associated with at least an answer (when available). For questions without answer, the question content will be used as its own answer. In the subsequent experiments, we investigate QuestER that includes only one answer and QuestER+ that includes the maximum of five answers (an analysis of the maximum number of answers is reported in subsection 4.4).

Baselines. We evaluate our proposed QuestER and QuestER+ against the following baselines in terms of useful review and QA selection. Comparisons between methods are tested with one-tailed paired-sample Student's t-test at 0.05 level.

- **HRDR** [28] uses attention mechanism with the rating-based representation as features to weight the contribution of each individual review toward user/item final representation.
- **NARRE** [3] learns to predict ratings and the usefulness of each reviews by applying attention mechanism for reviews on users/items embedding.
- **HFT** [32] models the latent factors from user or item reviews by employing topic distributions. In this work, we employ item reviews and applied their proposed usefulness review retrieval approach for selecting useful reviews. The number of topics is $K = 50$.

Among the three selected baselines, HRDR and NARRE use similar TextCNN for learning text representation. There are other works that use other text processors [48, 49] (discussed more in

[4] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html
[5] The collected data is available at https://github.com/PreferredAI/QuestER

section 2), which we do not consider as direct baselines in this work. Note that our key distinction from the above mentioned baselines is that we further incorporate product questions. As there is no prior work on predicting ratings along with selecting useful question, when the evaluative task is to look into selecting questions (question retrieval and question similarity tasks, see Section 4.1 and Section 4.3), we would apply similar approach for each baseline such that item text will be item questions instead of item reviews.

Training Details. Each item's reviews are split randomly into train, validation, and test with ratio $0.8 : 0.1 : 0.1$. Unknown users are excluded from validation and test sets. Reviews in validation set and test set are excluded from training and will not be used for rating prediction on validation/test data. Answers are appended as additional text of the corresponding question. We employ the pretrained word embeddings from GloVe [36] to initialize the text embedding matrix with dimensionality of 100 in which the embedding matrix is shared for both reviews and questions. We use separate TextCNN for user reviews, item reviews, and item QAs. The maximum number of tokens for each text W is 128, the number of neurons in convolutional layer m is 64, the window size w is 3. The latent factor number was tested in $k \in \{8, 16, 32, 64\}$. After tuning, we set $k = 8$ for memory efficiency as using larger k does not improve the performance significantly. Dropout ratio is 0.5 as in [3], τ is 0.01. We apply 3-layers MLP for rating-based representation modeling as in [28], with the number of neural units in hidden layers to be $\{128, 64, m\}$. Using Adam optimizer [17] with an initial learning rate of 10^{-3} and mini-batch size of 64, we see models tend to converge before 20 epochs. We set a maximum of 20 epochs and report the test result from the best performing model (the lowest MSE) on validation, a uniform practice across methods.

Brief Comment on Running Time. Our focus in this work is recommendation explanation, rather than computational efficiency. The models can be run offline. For a sense of the running times, Table 3 reports the training time and testing time of all models on a machine with AMD EPYC 7742 64-Core Processor and NVIDIA Quadro RTX 8000. Increasing the maximum number of answers to 5 (QuestER+) slows down training time approximately 1.3 ∼ 1.5 times compared to training with only one answer. Inference time of all models are similarly fast.

Table 3. Running time (Train (seconds) / Test (seconds))

Model	Home	Health	Sport	Toy	Grocery	Baby	Office	Automotive	Patio	Musical
QuestER	19707 / 4.7	11892 / 2.9	9805 / 2.7	5453 / 1.5	5014 / 1.3	5360 / 1.4	2011 / 0.4	646 / 0.2	448 / 0.1	334 / 0.1
QuestER+	27493 / 4.8	16560 / 2.8	13966 / 2.7	7741 / 1.5	7202 / 1.2	7690 / 1.4	2651 / 0.4	931 / 0.2	649 / 0.2	503 / 0.1
HRDR	13603 / 4.3	10906 / 3.8	9770 / 2.5	3199 / 1.5	3048 / 1.2	3704 / 1.4	1158 / 0.4	424 / 0.2	267 / 0.1	180 / 0.1
NARRE	9855 / 5.1	6254 / 3.1	5034 / 2.8	2093 / 1.6	2755 / 1.3	2855 / 1.5	1067 / 0.4	329 / 0.2	249 / 0.1	172 / 0.1
HFT	9399 / 4.0	5806 / 2.3	5452 / 2.2	3305 / 1.2	2508 / 1.0	2665 / 1.2	1052 / 0.4	395 / 0.2	460 / 0.2	253 / 0.1

4.1 Question and Review Alignment

Our proposed recommendation explanation consists of a question-and-answer (QA) and a review. Ideally, these two components, QA on one hand, and review on the other hand, are well-aligned for a more coherent explanation. We measure this alignment using ROUGE [26] and METEOR [1], two well-known metrics for text matching and text summarization. To cater to words as well as phrases, we report F-Measure of ROUGE-1 measuring the overlapping unigrams, ROUGE-2 measuring the overlapping bigrams, and ROUGE-L measuring the longest common subsequence beween the reference summary and evaluated summary. We compute ROUGE and METEOR scores for the top-1 selected question and review and report them in Table 4.

The results show that the proposed QuestER and QuestER+ consistently outperform the baselines significantly across virtually all the datasets. This shows QuestER's QAs and reviews that

Table 4. Performance in question and review alignment

Data	Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Home	QuestER	**15.68**§	0.88§	**7.73**§	**9.56**§
	QuestER+	15.65§	**0.89**§	7.71§	9.55§
	HRDR	14.85	0.75	7.08	8.36
	NARRE	14.66	0.72	6.57	7.39
	HFT	13.55	0.66	6.40	7.53
Health	QuestER	19.54	**1.59**	7.99§	9.89§
	QuestER+	19.58	1.58	**8.01**§	**9.90**§
	HRDR	**19.59**	**1.59**	7.88	9.65
	NARRE	17.97	1.33	6.45	7.31
	HFT	17.13	1.28	6.59	7.93
Sport	QuestER	15.52§	0.72§	7.33§	9.04§
	QuestER+	**15.56**§	**0.73**§	**7.35**§	**9.07**§
	HRDR	15.25	0.64	7.14	8.35
	NARRE	14.52	0.56	6.21	7.00
	HFT	13.88	0.56	6.09	7.29
Toy	QuestER	**15.80**§	**1.17**§	**7.84**§	**9.41**§
	QuestER+	**15.80**§	**1.17**§	7.83§	**9.41**§
	HRDR	15.20	1.08	7.18	8.12
	NARRE	15.08	1.03	7.05	7.86
	HFT	14.05	0.96	6.53	7.39
Grocery	QuestER	**16.82**§	**0.74**§	7.04§	**8.15**§
	QuestER+	16.80§	**0.74**§	**7.05**§	8.13§
	HRDR	16.18	0.67	6.45	7.35
	NARRE	15.22	0.56	5.51	5.85
	HFT	14.68	0.57	5.71	6.46
Baby	QuestER	**18.82**§	**1.23**§	**7.84**§	**10.59**§
	QuestER+	18.80§	1.22§	7.81§	10.54§
	HRDR	18.51	1.15	7.39	9.75
	NARRE	17.64	1.04	6.79	8.50
	HFT	15.93	0.88	6.14	7.61
Office	QuestER	**18.00**§	**0.99**§	**7.89**§	**12.44**§
	QuestER+	17.82§	**0.99**§	7.76§	12.27§
	HRDR	17.53	0.76	7.36	11.36
	NARRE	17.14	0.70	6.76	9.13
	HFT	15.07	0.61	6.32	8.93
Automotive	QuestER	**17.94**	**1.22**	**7.98**§	**10.36**
	QuestER+	17.79	1.19	7.85	**10.36**
	HRDR	17.72	1.16	7.65	10.28
	NARRE	16.35	0.91	6.16	7.36
	HFT	15.27	0.88	6.41	8.12
Patio	QuestER	18.93	1.74	8.96	13.19
	QuestER+	**18.91**	**1.76**	**9.07**	**13.29**
	HRDR	18.55	1.73	8.94	**13.29**
	NARRE	16.90	1.32	7.12	9.42
	HFT	15.53	1.24	7.13	10.48
Musical	QuestER	**16.42**§	**0.96**§	**7.44**§	**11.16**§
	QuestER+	16.11§	0.91§	7.37§	10.71§
	HRDR	14.81	0.70	6.63	9.75
	NARRE	13.94	0.48	5.64	6.90
	HFT	12.98	0.55	5.96	8.73

§ denotes statistically significant improvements. Highest values are in **bold**.

are part of a collective explanation are better-aligned with each other, as compared to the respective pairings identified by the baselines. Note that HRDR, NARRE, and HFT had been designed solely to select helpful reviews. To be able to compare with these models, we ran each model twice, once with reviews and another time replacing item reviews with QA's. This approach essentially treats review and question in a disjoint manner, which contributes to why they are underperforming as compared to our proposed QuestER that jointly selects review and question that are well-aligned with each other.

Table 5. Performance in Review-Level Explanation Task

Data	Model	Prec@5	Rec@5	F1@5	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Home	QuestER	**0.145**§	**0.634**§	**0.231**§	34.27§	18.48§	24.56§	27.93§
	QuestER+	**0.145**§	0.632§	**0.231**§	**34.34**§	**18.58**§	**24.66**§	**27.97**§
	HRDR	0.136	0.588	0.216	32.00	16.21	22.28	25.46
	NARRE	0.129	0.557	0.204	27.53	11.84	17.79	21.06
	HFT	0.141	0.613	0.224	28.87	14.36	19.99	23.35
Health	QuestER	**0.152**§	**0.645**§	**0.239**§	34.13§	19.16§	24.93§	27.99
	QuestER+	**0.152**§	**0.645**§	**0.239**§	**34.20**§	**19.22**§	**25.00**§	**28.03**
	HRDR	0.142	0.601	0.224	33.17	17.65	23.62	28.02
	NARRE	0.137	0.574	0.215	26.46	11.63	17.27	20.62
	HFT	0.149	0.635	0.236	28.69	14.70	20.18	23.86
Sport	QuestER	**0.157**	**0.663**§	0.247	**34.65**§	**19.60**§	**25.37**§	28.50
	QuestER+	**0.157**	**0.663**§	**0.248**	34.64§	**19.60**§	25.36§	28.50
	HRDR	0.151	0.633	0.237	34.24	19.04	24.87	**28.73**
	NARRE	0.141	0.591	0.222	27.84	12.63	18.41	22.24
	HFT	0.155	0.656	0.245	29.63	15.48	20.95	24.83
Toy	QuestER	**0.158**§	**0.682**§	**0.250**§	36.72§	21.04§	26.68§	29.99§
	QuestER+	**0.158**§	0.681§	**0.250**§	**36.74**§	**21.08**§	**26.74**§	**30.02**§
	HRDR	0.143	0.611	0.226	31.67	15.20	21.14	25.72
	NARRE	0.143	0.611	0.226	30.35	14.13	19.98	24.19
	HFT	0.149	0.642	0.236	30.18	15.48	20.81	24.58
Grocery	QuestER	**0.165**§	0.695§	0.260§	**36.31**§	**21.36**§	**27.22**§	**30.23**§
	QuestER+	**0.165**§	**0.697**§	**0.261**§	36.13§	21.13§	27.00§	30.01§
	HRDR	0.155	0.649	0.244	32.49	16.83	22.90	28.08
	NARRE	0.152	0.635	0.239	28.66	13.33	19.26	23.03
	HFT	0.162	0.681	0.255	30.43	16.05	21.70	25.73
Baby	QuestER	0.138§	0.578§	0.217§	**35.05**§	**18.00**§	**24.15**§	**27.70**§
	QuestER+	**0.139**§	**0.583**§	**0.218**§	34.96§	17.87§	24.03§	**27.70**§
	HRDR	0.123	0.509	0.192	31.55	13.85	20.21	25.27
	NARRE	0.119	0.496	0.187	28.23	11.15	17.22	21.16
	HFT	0.128	0.537	0.201	27.77	12.49	18.00	21.39
Office	QuestER	0.144§	0.597§	0.222§	35.19§	18.40§	24.12§	28.62
	QuestER+	**0.145**§	**0.601**§	**0.224**§	**35.96**§	**19.32**§	**24.97**§	**29.81**
	HRDR	0.135	0.548	0.207	33.22	15.70	21.69	28.90
	NARRE	0.124	0.500	0.189	26.35	9.83	15.28	19.71
	HFT	0.126	0.516	0.193	27.04	12.00	17.04	21.48
Automotive	QuestER	**0.176**	**0.745**	**0.278**	**36.75**	**22.28**	**27.91**	31.11
	QuestER+	0.174	0.740	0.275	36.25	21.78	27.48	30.41
	HRDR	0.173	0.731	0.273	35.79	20.59	26.62	**31.94**
	NARRE	0.156	0.651	0.245	26.89	12.09	17.69	21.28
	HFT	0.168	0.710	0.265	29.94	15.95	21.44	25.04
Patio	QuestER	0.166	0.694	0.256	**38.10**	**21.94**	**27.58**	32.27
	QuestER+	0.164	0.685	0.253	37.14	20.87	26.61	31.07
	HRDR	0.165	0.679	0.252	37.01	20.13	25.93	**32.72**
	NARRE	0.152	0.629	0.233	28.25	11.65	17.23	22.38
	HFT	**0.168**	**0.704**	**0.260**	33.01	17.97	23.26	27.74
Musical	QuestER	0.145	0.617	0.230	35.40§	20.18§	25.88§	30.13§
	QuestER+	**0.149**	**0.633**	**0.236**	**36.21**§	**21.26**§	**26.82**§	**30.69**§
	HRDR	0.144	0.611	0.228	32.38	16.07	22.05	27.38
	NARRE	0.132	0.563	0.210	25.53	10.16	15.87	19.41
	HFT	0.144	0.613	0.228	27.40	12.56	18.08	22.32

§ denotes statistically significant improvements over the baselines. Highest values are in **bold**.

4.2 Review-Level Explanation

Here we assess whether incorporating questions would help in selecting reviews for the explanation. We take reviews that have the greatest positive helpfulness voting scores on every product to be the ground truth to study the performance of selecting useful reviews. We use Precision at 5 (Prec@5), Recall at 5 (Rec@5), and F1@5 as evaluation. As reported in Table 5 (left), our proposed QuestER and QuestER+ are the better-performing methods overall. Theirs outperformance over baseline

Table 6. Performance in Question-Level Explanation Task

Data	Model	Prec@5	Rec@5	F1@5	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Home	QuestER	**0.097**§	0.360§	0.146§	**21.09**§	**10.97**§	**17.82**§	**20.26**§
	QuestER+	**0.097**§	**0.365**§	**0.147**§	20.95§	10.88§	17.67§	**20.26**§
	HRDR	0.082	0.307	0.124	17.47	7.52	13.22	16.51
	NARRE	0.082	0.307	0.124	17.69	7.77	13.52	16.75
	HFT	0.082	0.309	0.125	17.72	8.14	14.91	16.33
Health	QuestER	**0.115**§	**0.447**§	**0.177**§	23.45§	**14.36**§	20.51§	**22.98**§
	QuestER+	0.114§	0.439§	0.175§	**23.65**§	14.24§	**20.72**§	22.87§
	HRDR	0.091	0.347	0.139	16.74	7.25	12.00	16.64
	NARRE	0.089	0.342	0.136	17.62	8.18	13.70	16.73
	HFT	0.092	0.353	0.140	18.36	8.95	15.63	17.33
Sport	QuestER	0.114§	0.443§	0.175§	**24.03**§	**14.04**§	**20.89**§	**23.24**§
	QuestER+	**0.116**§	**0.447**§	**0.178**§	23.40§	13.35§	20.18§	22.81§
	HRDR	0.085	0.329	0.131	13.13	3.65	7.79	12.71
	NARRE	0.088	0.335	0.134	18.14	8.08	13.83	17.04
	HFT	0.090	0.343	0.138	20.13	10.03	17.26	18.69
Toy	QuestER	**0.130**§	**0.485**§	**0.197**§	23.80§	**14.82**§	**20.77**§	**23.74**§
	QuestER+	0.126§	0.468§	0.191§	**23.85**§	14.02§	20.70§	23.70§
	HRDR	0.106	0.392	0.161	14.50	5.27	9.21	15.61
	NARRE	0.107	0.394	0.162	19.15	10.00	15.10	19.69
	HFT	0.110	0.404	0.166	21.16	11.80	18.49	20.79
Grocery	QuestER	**0.125**§	0.503§	**0.194**§	**26.92**§	**18.08**§	**24.08**§	**26.11**§
	QuestER+	0.124§	**0.504**§	0.193§	23.32	14.01	20.11	22.12§
	HRDR	0.105	0.427	0.164	20.16	10.79	15.79	19.17
	NARRE	0.103	0.425	0.161	17.66	8.28	13.18	17.48
	HFT	0.105	0.437	0.166	21.70	12.28	18.93	19.37
Baby	QuestER	**0.110**§	**0.399**§	**0.166**§	**23.70**§	**13.21**§	**20.16**§	**22.62**
	QuestER+	0.104§	0.384§	0.157§	22.52	11.43	18.72	21.30
	HRDR	0.085	0.317	0.129	15.07	4.22	9.78	15.32
	NARRE	0.086	0.327	0.132	20.63	9.58	16.28	20.18
	HFT	0.085	0.314	0.129	19.34	9.88	16.57	17.45
Office	QuestER	0.101§	0.399§	0.155§	**21.85**§	**11.98**§	**18.60**§	20.67§
	QuestER+	**0.107**§	**0.415**§	**0.164**§	21.63§	11.56§	18.13§	**20.75**§
	HRDR	0.075	0.291	0.115	14.08	4.00	8.61	12.84
	NARRE	0.072	0.273	0.109	13.57	3.78	8.53	12.72
	HFT	0.075	0.290	0.115	17.36	7.44	14.54	15.57
Automotive	QuestER	0.106§	0.416§	0.163§	26.50§	15.76§	23.23§	25.39§
	QuestER+	**0.107**§	**0.417**§	**0.164**§	**28.61**§	**18.39**§	**25.58**§	**27.49**§
	HRDR	0.063	0.251	0.097	14.57	3.65	10.36	12.25
	NARRE	0.063	0.253	0.098	16.15	5.31	11.01	14.82
	HFT	0.060	0.242	0.093	15.82	5.79	13.18	13.26
Patio	QuestER	0.094§	0.384§	0.147§	**23.29**§	**13.05**§	**20.10**§	**21.32**§
	QuestER+	**0.104**§	**0.422**§	**0.162**§	21.25§	10.59§	17.84§	20.17§
	HRDR	0.051	0.198	0.079	14.67	4.07	10.40	12.16
	NARRE	0.055	0.210	0.084	11.41	1.74	6.57	9.43
	HFT	0.054	0.212	0.083	14.55	5.42	12.27	10.83
Musical	QuestER	**0.118**§	**0.446**§	**0.179**§	**23.95**§	**13.01**§	**20.57**§	**23.43**§
	QuestER+	0.111§	0.427§	0.170§	22.58§	11.82§	19.38§	20.74§
	HRDR	0.075	0.293	0.116	18.48	7.66	13.84	16.79
	NARRE	0.087	0.339	0.134	12.86	2.19	7.00	12.29
	HFT	0.086	0.352	0.134	17.65	6.88	14.61	15.11

§ denotes statistically significant improvements over the baselines. Highest values are in **bold**.

models are statistically significant in the majority of cases. QuestER still outperforms NARRE (on Automotive, Patio, and Musical categories) and HFT (on Automotive category) significantly.

To further assess the quality of top-ranked reviews against top-rated helpful reviews, we again use ROUGE and METEOR as metrics. The results in Table 5 consistently show that our proposed QuestER and QuestER+ outperform all baseline models significantly in the majority of cases, i.e., the top-ranked reviews from QuestER and QuestER+ are more similar to the top-rated helpful reviews than those of HRDR, NARRE, and HFT. Overall, in addition to the reviews, our QuestER

Table 7. Rating prediction performance: Mean Square Error (MSE). Best values are in **bold**.

Data	HFT	NARRE	HRDR	QuestER	QuestER+
Home	1.2775	1.2654	1.2677	1.2670	1.2666
Health	1.2712	1.2853	1.2878	1.2862	1.2861
Sport	1.0251	1.0054	1.0072	1.0053	1.0047
Toy	0.9136	0.9971	0.9973	0.9974	0.9979
Grocery	1.2007	1.1987	1.1988	1.2011	1.2027
Baby	1.3719	1.3622	1.3639	1.3613	1.3614
Office	0.8948	0.9248	0.9267	0.9245	0.9250
Automotive	0.9570	0.9248	0.9250	0.9258	0.9236
Patio	1.1173	1.1537	1.1594	1.1588	1.1564
Musical	0.8846	0.8136	0.8102	0.8174	0.8155

and QuestER+ use additional product QA, achieving better results than the baseline methods those only use reviews as additional data, suggesting that using QA aids in selecting more useful reviews.

4.3 Question-Level Explanation

The novelty of the proposed QuestER and QuestER+ are in producing question-level explanation along with review-level explanation. We conduct a homologous quantitative evaluation as Review-Level Explanation above, but now with question votes as ground-truth and measure Prec@5, Rec@5, and F1@5. In addition, we measure the similarity between top ranked question by QuestER (or QuestER+) and top voted useful question using ROUGE and METEOR, only the question are being evaluated in this evaluation. As shown in Table 6, QuestER and QuestER+ are significant better than other baselines throughout. This result further highlights the improvement of the current version of QuestER (this work) in comparison to that of the previous version [22] in which this version achieves better results quantitatively for question-level explanation.

4.4 Rating Prediction

As previously established, our main focus in this work is on recommendation explanations, with an eye on improving the selection of reviews and incorporating questions in that endeavour. Nevertheless, while recommendation accuracy is not the main focus, we find that QuestER still maintains parity in this regard with the other methods.

We report the average of *Mean Square Error* (MSE) averaged across users on each category in Table 7. Our proposed QuestER and QuestER+ achieve comparable results when compared to the neural models HRDR and NARRE. HFT that is based on graphical model varies from the neural models. Depending on the reported domain, it is lower in some cases and higher in others. Such variant in performance between simpler and more complex models using neural networks in term of rating predictions is expected and has also been reported in [40].

In any case, as we see from the previous experiments as well, QuestER and QuestER+ stand out in having the better review-level and question-level explanations, which are the main focal points of this work.

Effect of Number of Answers in each Question. We now focus on analyzing the effect of using different maximum number of answers to be used in each question. We report the average of MSE averaged across users on each category when varying the maximum number of answers being used in the set {1,3,5,10} in Table 8. We observe relatively minor differences in rating prediction performance among the variants. The proposed method achieves best MSE w.r.t 5 answers (QuestER+) in the majority of cases, which motivates us further evaluating this variant in the previous experiments.

Table 8. Rating prediction performance (MSE) of QuestER w.r.t different maximum number of answers

Data	Mean Square Error (MSE) of QuestER			
	1 answer	3 answers	5 answers	10 answers
Home	1.2670	1.2667	1.2666	1.2669
Health	1.2862	1.2865	1.2861	1.2862
Sport	1.0053	1.0051	1.0047	1.0051
Toy	0.9974	0.9976	0.9979	0.9972
Grocery	1.2011	1.2006	1.2027	1.2001
Baby	1.3613	1.3624	1.3614	1.3622
Office	0.9245	0.9239	0.9250	0.9248
Automotive	0.9258	0.9253	0.9236	0.9238
Patio	1.1588	1.1560	1.1564	1.1551
Musical	0.8174	0.8123	0.8155	0.8178

Asin: B0009IQZ2K
Title: Meguiar's E7200 Mirror Glaze High-Tech Backing Pad

Top Rated Useful Question: What grit is this?
Answer: There is no grit. It is a flexible sanding pad that you wrap your wet/dry sandpaper around to allow for easier sanding.
Top Rated Useful Review: I'm not a Meguiar's fan when it comes to their polishes and cleaning supplies, but this pad seems to work well. I have better control of it when it's cut in half width wise, which gives me two square blocks.

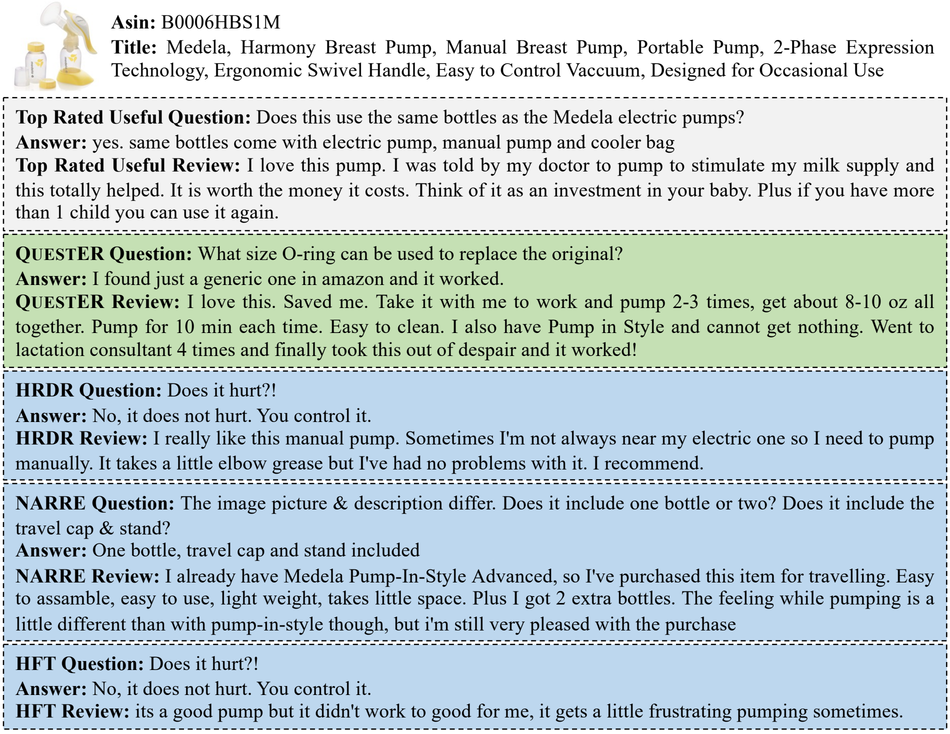
QuestER Question: Is this soft and flexible enough to be used for fine wet sanding? I basically need something that bends to the contours.
Answer: This is soft and flexible but I don't know if it is flexible enough to be able to bend to contours while you are placing even pressure across the entire sponge. If you were to purposefully distribute weight to the correct areas of the sponge then yes, but expecting the sponge to form fit to the area being sanded is not feasible.
QuestER Review: There's really not a lot to it. It's just a small 5-1/2" long X 2-1/2" wide foam block, but it worked fine as a backing pad along with 2000 grit paper for wet sanding clear coat before polishing. It seems to be holding up okay to use, so for what little it cost I think it was worth getting.

HRDR Question: How does the sand paper attach to this?? Thanks
Answer: It does not attach as with a sanding block. You wrap paper around it. I back my paper with duck tape then wrap it around the block and since it is flexible and semi soft you can really get into contours with it.
HRDR Review: I'm not a Meguiar's fan when it comes to their polishes and cleaning supplies, but this pad seems to work well. I have better control of it when it's cut in half width wise, which gives me two square blocks.

NARRE Question: What grit is this?
Answer: There is no grit. It is a flexible sanding pad that you wrap your wet/dry sandpaper around to allow for easier sanding.
NARRE Review: There's really not a lot to it. It's just a small 5-1/2" long X 2-1/2" wide foam block, but it worked fine as a backing pad along with 2000 grit paper for wet sanding clear coat before polishing. It seems to be holding up okay to use, so for what little it cost I think it was worth getting.

HFT Question: What grit is this?
Answer: There is no grit. It is a flexible sanding pad that you wrap your wet/dry sandpaper around to allow for easier sanding.
HFT Review: This is a great pad. Not to hard and not to Soft. Makes wet sanding much easier! I Love it!

Fig. 5. Example explanation: Meguiar's Sanding Pad (explanation by Top_Rated_Useful is in grey, that by QuestER is in green, and those by other baselines are in blue)



Asin: B0006HBS1M
Title: Medela, Harmony Breast Pump, Manual Breast Pump, Portable Pump, 2-Phase Expression Technology, Ergonomic Swivel Handle, Easy to Control Vaccuum, Designed for Occasional Use

Top Rated Useful Question: Does this use the same bottles as the Medela electric pumps?
Answer: yes. same bottles come with electric pump, manual pump and cooler bag
Top Rated Useful Review: I love this pump. I was told by my doctor to pump to stimulate my milk supply and this totally helped. It is worth the money it costs. Think of it as an investment in your baby. Plus if you have more than 1 child you can use it again.

QuestER Question: What size O-ring can be used to replace the original?
Answer: I found just a generic one in amazon and it worked.
QuestER Review: I love this. Saved me. Take it with me to work and pump 2-3 times, get about 8-10 oz all together. Pump for 10 min each time. Easy to clean. I also have Pump in Style and cannot get nothing. Went to lactation consultant 4 times and finally took this out of despair and it worked!

HRDR Question: Does it hurt?!
Answer: No, it does not hurt. You control it.
HRDR Review: I really like this manual pump. Sometimes I'm not always near my electric one so I need to pump manually. It takes a little elbow grease but I've had no problems with it. I recommend.

NARRE Question: The image picture & description differ. Does it include one bottle or two? Does it include the travel cap & stand?
Answer: One bottle, travel cap and stand included
NARRE Review: I already have Medela Pump-In-Style Advanced, so I've purchased this item for travelling. Easy to assamble, easy to use, light weight, takes little space. Plus I got 2 extra bottles. The feeling while pumping is a little different than with pump-in-style though, but i'm still very pleased with the purchase

HFT Question: Does it hurt?!
Answer: No, it does not hurt. You control it.
HFT Review: its a good pump but it didn't work to good for me, it gets a little frustrating pumping sometimes.

Fig. 6. Example explanation: Medela's Breast Pump (explanation by Top_Rated_Useful is in grey, that by QuestER is in green, and those by other baselines are in blue)

4.5 Case Studies

To investigate the usefulness of the recommendation explanation consisting of a QA as well as a review, we show a few case studies that benchmarks QuestER to the most voted question and the most voted review.

- Figure 5 shows five sets of explanations for a sanding pad product of *Meguiar's* brand. The first set (in grey box, above) comprise a QA and a review based on Top_Rated_Useful votes. The second set (in green box) comprise those selected by our QuestER. While both QuestER and Top_Rated_Useful provide useful information about the product, QuestER's explanation is notable in two respects. For one, QuestER's question with its answer is more aligned with its review than those of Top_Rated_Useful, ROUGE-L F-Measure for QuestER and Top_Rated_Useful are 10.61 and 8.37 respectively. For another, Top_Rated_Useful is based on explicit votes, which are not found on many products and therefore not universally available or applicable. The following three blue boxes comprise the explanations produced by the baseline methods. While NARRE has the same review explanation as QuestER, it produces a different QA explanation.
- Figure 6 shows explanation for a breast pump product of *Medela* brand. Both QuestER and Top_Rated_Useful provide further useful information about the product. QuestER's question with its answer is considered more aligned with its review than those of Top_Rated_Useful, ROUGE-L F-Measures are 12.59 and 9.02 respectively. In this case, the baselines HRDR and HFT pick the same question.

Asin: B004N0MKN8
Title: Planet Waves Guitar Rest

Top Rated Useful Question: What is the response to the numerous customer reviews that say that the thing keeps falling off unless the guitar is resting against it?
Answer: It does tend to fall off, like you say, but it really is great to lean the guitar on. Otherwise the guitar just falls over. Pick your poison! Sorry
Top Rated Useful Review: The Planet Waves Guitar Rest works for ukuleles! I just got one, and have used it for a few days, and it's the bomb! I can set my little ukuleles down now without fear of falling over. This product is a rubber disc with small "arms" in a gentle curve that nestles against the edge of any surface, and you can set your instrument against it, and voila, it doesn't fall over! Here at home, I use it on the second shelf of a bookcase, and my concert sized ukulele fits like a glove, heel on carpet, neck in Guitar Rest. I'm going to buy a couple more for my ukulele cases, because I can use them at one of my uke parties. If one sets a tiny ukulele on the floor, for instance, to take a whizz, they're just small enough to go unseen and have someone step on them. Here, I just find a spot near wherever I'm sitting, and it becomes my "lean" spot, and I can even set my beer can on the round part on the back! Coaster uke/guitar holder. It's quite immovable once it has some weight against it from the instrument. I could carry a metal stand with me, but it wouldn't fit in my ukulele case--this Planet Waves product does. A winner.

QuestER Question: Will this guitar rest work on a round table top?
Answer: That depends entirely on the dimensions of the table. Take the guitar and see if it can lay flat across the table. If it does, then it will work just fine. If it goes off the end a little bit, it should still be fine.
QuestER Review: I've had this thing for several weeks, and just now, when it fell off the table for the 100th time, I tossed it in the trash. The whole thing is one piece of soft floppy rubber, it's not stiff enough for the part that cradles the guitar neck, and it's not heavy enough to stay put. Even the force from the guitar neck makes it topple over. Unless you glue this thing to the table, or something like that, it's useless, even worse than useless, it's in the way.
Addendum: I raised the rating a bit, after hearing from the distributor/manufacturer ... at least these guys listen.

HRDR Question: What is the width of the cavity? I need it for a large Touch Style instrument that has a very wide neck.
Answer: Answer here...don't know the dimensions, but it could hold a very wide neck. Nice product, but you have to be careful where you place your guitar to protect it.
HRDR Review: I just can't trust it! I thought it'd be a good idea if I needed to prop my guitar up against the amp for a minute....but I just can't get myself to do it. I always end up using a little Fender portable stand. It does make a nice coaster for placing a beer on my amp.

NARRE Question: Just turn your guitar around and rest its strings (the fretboard side) against a table, so the back is facing you, but it works well without this thing. No?
Answer: Yea if you want to scratch your fretboard.. My cheap guitars I just lean back against a wall. I would never lean a nice guitar on its strings...
NARRE Review: It is a little light if you bump it anything that you store in the top goes everywhere when the rest falls on the floor. As long as the guitar is on it works great.

HFT Question: Lots of bad reviews is it as good as people say?
Answer: It's a good idea that needs refinement. The balance is off and without a guitar to hold it it falls off the counter top nearly 80% of the time. As a player, I don't want to have to gingerly lift the durn guitar off the thing to keep the holder off the floor.
HFT Review: This can work, but if you already use a guitar stand then it is a waste of money. It will work if you use your guitar next to the a table.

Fig. 7. Example explanation: Planet Waves Guitar Rest (explanation by Top_Rated_Useful is in grey, that by QuestER is in green, and those by other baselines are in blue)

- Figure 7 shows explanation for a guitar rest. Notably, the pairing by Top_Rated_Useful are not so coherent, with the QA discusses its use for guitars, while the review discusses its use for ukuleles. In contrast, the QA and the review by QuestER concentrate on the key issue of how well the item could hold a guitar in rest. QuestER's QA is more aligned with its review than those of Top_Rated_Useful, ROUGE-L F-Measures are 14.71 and 6.64 respectively.

4.6 User Studies

To evaluate the quality of questions and reviews selected by QuestER and Top_Rated_Useful (based on user votes on Amazon.com), we conduct a couple of user studies.

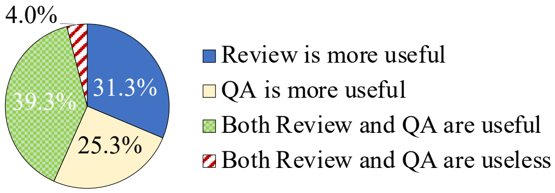


Fig. 8. Review vs Question-Answer annotation results

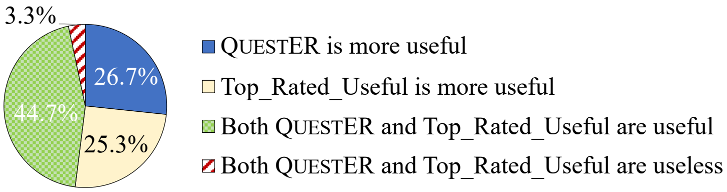


Fig. 9. QuestER vs Top_Rated_Useful annotation results

Reviews vs. QAs. In the first study, we seek to investigate whether users find questions and reviews helpful as part of a recommendation explanation. We conduct user studies concerning 30 examples (3 products from each category). We split these examples into 3 surveys, each containing 10 examples of different domains which are generated by QuestER. Each survey is done by 5 annotators, for a total of 15 annotators who are neither the authors nor having any knowledge of the objective of the study. Each product is presented with both question and review ordering randomly (review and question can be either group A or group B). We ask annotators to assess the pairwise quality with four options:

I. A is more useful than B
II. B is more useful than A
III. A and B are almost the same, both useful
IV. A and B are almost the same, both useless

The Fleiss' Kappa [19] for consistency for categorical ratings, $\kappa = 0.2955$, implies fair agreement.

Pairwise evaluation results are shown in Figure 8. As the key proposal is to have both review and question be part of an expanded explanation, it is gratifying that the most popular option is that both are useful, attaining 39.3%. While the percentage that finds reviews more useful is slightly higher than the percentage that finds questions more useful, this is less important as we are not seeking to replace reviews with questions. Excluding "both useless", 96% find at least one useful. We repeat the same study with explanations coming from Top_Rated_Useful and the conclusion still holds, i.e., the most popular option is that both reviews and QA are useful.

QuestER vs. Top_Rated_Useful. In the second user study, we would like to investigate the quality of the proposed *combined* explanation form consisting of a QA and a review. With the same set of examples and annotators, we split the examples into 3 other surveys, each containing 10 products from different categories. We present the explanations blindly by ordering survey's questions and explanations randomly (group A and group B are now either QuestER or Top_Rated_Useful). We ask similar questions as in the first study. Figure 9 shows the pairwise evaluation results between QuestER and Top_Rated_Useful. The Fleiss' Kappa score is 0.217 indicating fair agreement. In summary, when combining question and review as explanation, the overall quality of both QuestER and Top_Rated_Useful are useful (96.67%). Among those, question and review selected by QuestER are considered to be slightly more useful (26.7%) than those of top rated useful (25.3%).

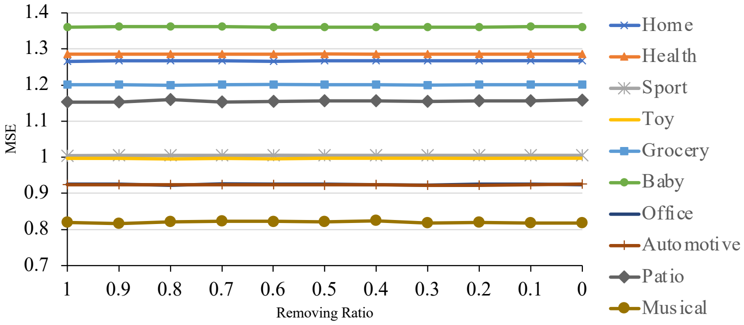


Fig. 10. Rating predictions performance (MSE) when removing reviews.

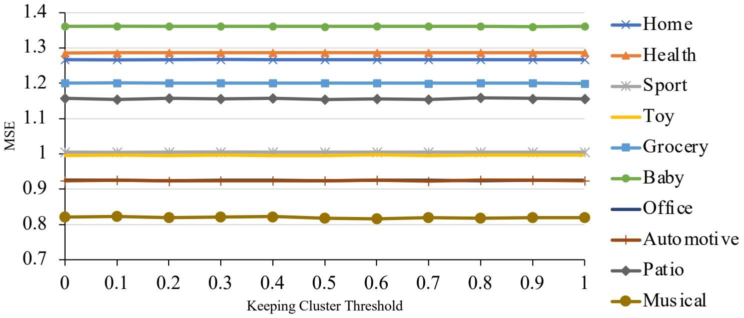


Fig. 11. Rating predictions performance (MSE) when varying the keeping big cluster threshold.

As important as the slight outperformance by QuestER over Top_Rated_Useful, or perhaps more so is that QuestER as a method is more widely applicable method. In contrast, Top_Rated_Useful relies on the existence of helpfulness votes, which are relatively rare, and therefore it stands more as a benchmark rather than a practical method for review and QA selection for explanation.

4.7 Discussion

Robust Rating Prediction Layer. Here we further explore the cold-start scenario by removing reviews as well as questions and answers. Keeping the available ratings, we randomly remove reviews with ratios in range [0,1] with step size 0.1. Results in Figure 10 for all datasets consistently show that the rating prediction performance of the proposed QuestER is quite stable regardless the amount of reviews. This can be explained using Equation 15, missing reviews only discard the contribution of user and item representations constructed using reviews and QAs while the rating-based representation as well as latent factors ζ_u and ζ_p are available. We further note here that using Equation 15 can produce rating prediction for users/items that have rating-only (u_i and p_j), content-only (X_i and Y_j), and known latent factors ($\zeta_u(i)$ and $\zeta_p(j)$). In addition, we investigate the overall rating prediction when vary the number of available questions by setting the keeping questions threshold to be covered by big clusters in the range of [0,1] with step size 0.1 (in the main experiment, this threshold is 0.8)[6]. We observe a similar trend that the rating prediction performance is quite stable (see Figure 11).

[6] When keeping all clusters (threshold is 1), the General Question is all the centroid questions.

Table 9. The overall performance of QuestER using BERT as Text Encoder on Musical data

Text Encoder	Train (s)	MSE	Text Alignment		Review-Level Explanation			Question-Level Explanation		
			ROUGE-L	METEOR	F1@5	ROUGE-L	METEOR	F1@5	ROUGE-L	METEOR
TextCNN	334	0.8174	7.44	11.16	0.230	25.88	30.13	0.179	20.57	23.43
BERT(L-2)	61164	0.7861	7.27	10.75	0.237	26.72	30.78	0.175	19.18	22.11
BERT(L-4)	66291	0.7915	7.29	10.73	0.240	25.77	30.06	0.171	18.94	21.71
BERT(L-6)	73332	0.7909	7.32	10.91	0.233	25.47	29.68	0.178	17.98	21.14
BERT(L-8)	74587	0.7929	7.33	10.89	0.240	25.96	30.45	0.176	20.16	22.26
BERT(L-10)	76953	0.7767	7.48	11.08	0.235	23.96	28.52	0.176	18.14	21.06
BERT(L-12)	79146	0.7892	7.31	10.69	0.233	25.96	30.14	0.174	15.45	17.63

Using BERT as Text Encoder. Here we investigate whether using other text encoder such as BERT can further enhance the overall performance. Table 9 reports the overall performance of different variants of QuestER based on its text encoder (default is TextCNN) including 6 different versions of small BERT model from TF Hub with 128 hidden dimension, from L-2 (2 Transformer blocks) to L-12 (12 Transformer blocks). Trivially, using a larger text encoder model consumes more time for training. Evidently, using BERT as text encoder does enhance the rating prediction performance. However, it does not clearly show that using BERT as text encoder enhances the explanation performance further in term of text alignment, review-level explanation, and question-level explanation.

5 CONCLUSION

QuestER is a framework for incorporating question-answer pair or QA into review-based recommendation explanation. We model QA in an attention mechanism to identify more useful reviews. Through joint modeling, we can collectively form an explanation in terms of QA and review. Comprehensive experiments on various product categories show that the QA and the review that are part of a collective explanation are more coherent with each other than those pairings found by the baselines. Review-level and question-level explanations identified by QuestER are also more consistent with top-rated ones based on helpfulness votes than those identified by the baselines. User studies further help to support that incorporating questions as part of a recommendation explanation is useful.

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-020).

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (Prague, Czech Republic). Association for Computational Linguistics, USA, 65–72.

[2] Rose Catherine and William Cohen. 2017. TransNets: Learning to Transform for Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 288–296. https://doi.org/10.1145/3109859.3109878

[3] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-Level Explanations. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1583–1592.

[4] Long Chen, Ziyu Guan, Qibin Xu, Qiong Zhang, Huan Sun, Guangyue Lu, and Deng Cai. 2020. Question-driven purchasing propensity analysis for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 35–42. https://doi.org/10.1609/aaai.v34i01.5331

[5] Long Chen, Ziyu Guan, Wei Zhao, Wanqing Zhao, Xiaopeng Wang, Zhou Zhao, and Huan Sun. 2019. Answer identification from product reviews for user questions by multi-task attentive networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19, Vol. 33)*. AAAI Press, 45–52. https://doi.org/10.1609/aaai.v33i01.330145

[6] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19* (Macao, China). International Joint Conferences on Artificial Intelligence Organization, 2137–2143. https://doi.org/10.24963/ijcai.2019/296

[7] Dawei Cong, Yanyan Zhao, Bing Qin, Yu Han, Murray Zhang, Alden Liu, and Nat Chen. 2019. Hierarchical Attention Based Neural Network for Explainable Recommendation. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (Ottawa ON, Canada) *(ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 373–381. https://doi.org/10.1145/3323873.3326592

[8] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, New York, USA) *(KDD '14)*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/2623330.2623758

[9] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 698–708.

[10] Xin Dong, Jingchao Ni, Wei Cheng, Zhengzhang Chen, Bo Zong, Dongjin Song, Yanchi Liu, Haifeng Chen, and Gerard de Melo. 2020. Asymmetrical Hierarchical Networks with Attentive Interactions for Interpretable Review-Based Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7667–7674. https://doi.org/10.1609/aaai.v34i05.6268

[11] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open* 2 (2021), 225–250.

[12] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[13] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. TriRank: Review-Aware Explainable Recommendation by Modeling Aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1661–1670. https://doi.org/10.1145/2806416.2806504

[14] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[15] Chunli Huang, Wenjun Jiang, Jie Wu, and Guojun Wang. 2020. Personalized Review Recommendation Based on Users' Aspect Sentiment. *ACM Trans. Internet Technol.* 20, 4, Article 42 (oct 2020), 26 pages. https://doi.org/10.1145/3414841

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[17] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

[18] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[19] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[20] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.

[21] Trung-Hoang Le and Hady W. Lauw. 2021. Explainable Recommendation with Comparative Constraints on Product Aspects. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 967–975. https://doi.org/10.1145/3437963.3441754

[22] Trung-Hoang Le and Hady W Lauw. 2022. Question-Attentive Review-Level Recommendation Explanation. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 756–761.

[23] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate Neural Template Explanations for Recommendation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 755–764. https://doi.org/10.1145/3340531.3411992

[24] 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4947–4957.

https://doi.org/10.18653/v1/2021.acl-long.383

[25] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. 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* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 345–354. https://doi.org/10.1145/3077136.3080822

[26] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (Edmonton, Canada) *(NAACL '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 71–78. https://doi.org/10.3115/1073445.1073465

[27] Han Liu, Yangyang Guo, Jianhua Yin, Zan Gao, and Liqiang Nie. 2022. Review polarity-wise recommender. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[28] Hongtao Liu, Yian Wang, Qiyao Peng, Fangzhao Wu, Lin Gan, Lin Pan, and Pengfei Jiao. 2020. Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing* 374 (2020), 77–85. https://doi.org/10.1016/j.neucom.2019.09.052

[29] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Coevolutionary Recommendation Model: Mutual Learning between Ratings and Reviews. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 773–782. https://doi.org/10.1145/3178876.3186158

[30] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting Social Context for Review Quality Prediction. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 691–700. https://doi.org/10.1145/1772690.1772761

[31] Lionel Martin and Pearl Pu. 2014. Prediction of helpful reviews using emotions extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[32] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 165–172. https://doi.org/10.1145/2507157.2507163

[33] Julian McAuley and Alex Yang. 2016. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 625–635. https://doi.org/10.1145/2872427.2883044

[34] Sicheng Pan, Dongsheng Li, Hansu Gu, Tun Lu, Xufang Luo, and Ning Gu. 2022. Accurate and Explainable Recommendation via Review Rationalization. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3092–3101. https://doi.org/10.1145/3485447.3512029

[35] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. *The adaptive web: methods and strategies of web personalization* (2007), 325–341.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[37] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[38] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[39] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social Collaborative Viewpoint Regression with Explainable Recommendations. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM'17)*. Association for Computing Machinery, New York, NY, USA, 485–494. https://doi.org/10.1145/3018661.3018686

[40] Noveen Sachdeva and Julian McAuley. 2020. *How Useful Are Reviews for Recommendation? A Critical Review and Potential Improvements*. Association for Computing Machinery, New York, NY, USA, 1845–1848. https://doi.org/10.1145/3397271.3401281

[41] Sunil Saumya, Jyoti Prakash Singh, and Yogesh K Dwivedi. 2020. Predicting the helpfulness score of online reviews using convolutional neural network. *Soft Computing* 24, 15 (2020), 10989–11005.

[42] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 297–305. https://doi.org/10.1145/3109859.3109890

[43] Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. 2016. Rating-Boosted Latent Topics: Understanding Users and Items with Ratings and Reviews. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2640–2646.

[44] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Context-Aware Review Helpfulness Rating Prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2507157.2507183

[45] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Multi-Pointer Co-Attention Networks for Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2309–2318. https://doi.org/10.1145/3219819.3220086

[46] Quoc-Tuan Truong and Hady Lauw. 2019. Multimodal Review Generation for Recommender Systems. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1864–1874. https://doi.org/10.1145/3308558.3313463

[47] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 165–174. https://doi.org/10.1145/3209978.3210010

[48] Chuhan Wu, Fangzhao Wu, Junxin Liu, and Yongfeng Huang. 2019. Hierarchical user and item representation with three-tier attention for recommendation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1818–1826.

[49] Chuhan Wu, Fangzhao Wu, Tao Qi, Suyu Ge, Yongfeng Huang, and Xing Xie. 2019. Reviews Meet Graphs: Enhancing User and Item Representations for Recommendation with Hierarchical Attentive Graph Neural Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4884–4893. https://doi.org/10.18653/v1/D19-1494

[50] Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder. 2023. Factual and informative review generation for explainable recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 13816–13824.

[51] Qian Yu and Wai Lam. 2018. Review-Aware Answer Prediction for Product-Related Questions Incorporating Aspects. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 691–699. https://doi.org/10.1145/3159652.3159718

[52] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval* 14, 1 (2020), 1–101.

[53] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. 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* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/2600428.2609579

[54] Jie Zhao, Ziyu Guan, and Huan Sun. 2019. Riker: Mining Rich Keyword Representations for Interpretable Product Question Answering. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 1389–1398. https://doi.org/10.1145/3292500.3330985

[55] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 425–434. https://doi.org/10.1145/3018661.3018665

Received 10 October 2023; revised 23 August 2024; accepted 22 September 2024