Generating Product Insights from Community Q&A

Lital Kuchy
Amazon, Israel
litalku@amazon.com

Ran Levy
Amazon, Israel
ranlevy@amazon.com

Avihai Mejer
Amazon, Israel
amejer@amazon.com

Noam Segev
Amazon, Israel
nsegev@amazon.com

Shunit Agmon*
Technion – Israel Institute of Technology
shunit.agmon@gmail.com

Miriam Farber†
Israel
mashafarber@gmail.com

ABSTRACT

In e-commerce sites, customer questions on the product detail page express the customers' information needs about the product. The answers to these questions often provide the necessary information. In this work, we present and address the novel task of generating product insights from community questions and answers (Q&A). These insights can be presented to customers to assist them in their shopping journey. Our method first generates concise, self-contained sentences based on the information in the Q&A. Then insights are selected based on the prominence of their associated questions. Empirical evaluation attests to the effectiveness of our approach in generating well-formed, objective, and helpful insights that are often not available in the product description or in summaries of customer reviews.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**.

KEYWORDS

Natural language generation, Community question answering

ACM Reference Format:
Lital Kuchy, Ran Levy, Avihai Mejer, Noam Segev, Shunit Agmon, and Miriam Farber. 2023. Generating Product Insights from Community Q&A. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3583780.3615480

1 INTRODUCTION

Product descriptions on e-commerce sites such as *amazon.com* or *ebay.com* have been shown to play an important role during customers' shopping journey [10, 13, 21]. While information provided in the description is rich and useful, many products have short description or no description at all [15]. Moreover, product descriptions often suffer from inherent bias, as they present the seller's point of view of the product, and reflect the seller's incentive to sell. Another source of important information that does not suffer from the seller's bias is the customer reviews section. In fact for popular products, the reviews section is so large that potential buyers cannot read through them and rely instead on smart ranking of the reviews [4, 27] or on automatic summaries [2, 5, 16]. However, reviews tend to focus on subjective information, sometimes leaving out important objective product characteristics. For this reason, some websites have a Questions and Answers (Q&A) section, in which customers can post questions of interest to the community. Customer questions express information needs about the product, and the answers to these questions often provide the necessary information. We observed that customers often ask for information that is missing from the product description or not prominent in customer reviews. In other cases, customers wish to verify the seller-provided information.

Motivated by this observation, we develop a method to distill useful product information from community Q&A. One way to present this information is to identify prominent Q&A, and simply surface them, as is, to the customer. However, as questions are often lengthy and may receive several answers, this may require the customer to read through large amounts of information. Another approach is to generate concise snippets using information present in the Q&A, and surface those to the customers. This approach presents the required information to the customer in a summarized and easy to read format. In this work, we focus on the latter approach, which can contribute to a wide range of use cases, such as enriching product description (Fig. 1), voice interface experiences, and side-by-side product comparison.

We define an *insight* as a concise, self-contained sentence that is likely helpful to many customers during their shopping journey. As the source of information for insights, we focus on yes-no questions, which constitute 54% of the Q&A in Amazon.com according to the Amazon-PQA dataset [26].

The proposed process for insight generation from community Q&A constitutes of two stages. First, an insight is generated from a yes-no question and its answers, reflecting the information provided in them. In the second stage, we select a subset of the generated insights to be presented to the customers, based on the prominence of the associated questions and the diversity of the final insight set. Our main contributions are: (1) presenting a novel task for generation of product insights from community Q&A; (2) an end-to-end pipeline consisting of insight generation and selection stages; and (3) a new annotated dataset of 20K yes-no questions, their answers, and the generated insights.

*This work was done while Shunit interned at Amazon.
†This work was done while Miriam worked at Amazon.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615480

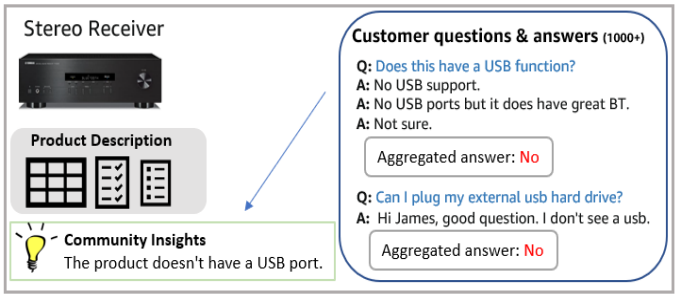

Figure 1: Product insights generated from community Q&A by our method.

2 RELATED WORK

Many works have been done on extracting useful insights from product *reviews*. Some works [9, 14, 30] focus on aspect-based summarization: extraction of prominent product aspects, and aggregation of review sentences by aspect and sentiment into a concise representation. Other summarization tasks include identifying helpful sentences [5], generating tips from reviews [8], and generating entire descriptions based on reviews [15]. Other works aim to generate product insights based on matching reviews to questions [12], or using reviews of similar products [20]. In contrast, community Q&A are an untapped source of information, with a great potential to yield valuable insights, different from review-based insights, as we show in our analysis (Section 6). A question is usually more focused on a single aspect of the product than a review, which may contain multiple aspects. Therefore, the challenge in summarizing Q&A is not aspect extraction, but rather an abstractive transformation of question and answer into a standalone readable sentence.

Previous works have studied the task of generating a full, standalone answer based on a question and a factoid short answer [17], or a question and a paragraph [1]. In [17], a Q&A dataset based on Wikipedia was used to create training data for this task, and a web-based, manually curated dataset was used in [1]. However, full answer generation in the product Q&A domain is different than in those domains: the answer is not always contained in the product description or the customer answers, so they cannot be used as a paragraph as in [1]; customer answers may be multiple and contradicting; and both customer questions and answers are noisy and contain grammar and spelling errors. Additionally, in this work we focus on yes-no questions, which are different from open ended [1] or factoid questions [17], as demonstrated in §4.1.

Finally, our work presents a possible approach for Q&A summarization. We note that previous works exist about summarization or processing of multiple answers to a single question [18, 28]. However, to the best of our knowledge, there are currently no works on summarization of the information presented in multiple product questions and their answers.

3 MINING INSIGHTS FROM Q&A

In this section, we describe our proposed method to generate helpful insights out of a Q&A collection which consists of multiple processing steps. At a high level the process can be divided to two stages. The first, illustrated in Figure 2, is an insight generation stage from a yes-no question and its answers, reflecting the information provided in them. The second is an insight selection stage, based on the prominence of the associated questions and the diversity of the final insight set.

3.1 Product insight generation

The input to our method is a Q&A collection of a specific product. Since information is divided between the question and the answer, and usually, neither of them constitute a standalone sentence, extractive methods are not suitable for our use-case. Therefore, we first transform each question and answer into a standalone sentence (insight).

Processing questions and answers. In this work we focus on yes-no questions, which constitute more than 50% of the community Q&A in Amazon according to the Amazon-PQA dataset [26]. Therefore, the first step is to identify and retain only the yes-no questions. Next, as community questions typically have free-text answers, a second required step is to map each answer to a yes, no, or neutral (unknown) label. For example, the answer *of course it does* is mapped to a "yes", *No USB support* is mapped to a "no", and *I'm not sure* is mapped to "neutral". Finally, community questions are often answered independently by multiple users (51% of yes-no questions in the Amazon-PQA dataset [26] have multiple answers), and the answers may not be unanimous. Therefore, a third required step is aggregating the diverse answers, and determining a final "yes", "no", or "neutral" answer, which is used for generating the insight. We note that the multiple answers (e.g. two "Yes" answers, one "Neutral", and one "No") can be used for estimating the confidence in each of the insights. This information could be used for insight selection, and may even be explicitly exposed to the customers to increase their confidence in the generated insights.

To perform the three mentioned steps, we adopt the solutions proposed in [26]. Namely, identifying yes-no questions is performed using the heuristics proposed in [7]. Mapping the free-text answers to yes/neutral/no answers is performed using a RoBERTa-based classifier trained for this task. Finally, in case of multiple answers, these are aggregated using a simple heuristic: when an answer is provided by a verified seller, it is considered as the final label; otherwise, the majority vote answer is assigned as final yes/no label, or neutral in case of a tie. This process was applied to generate the Amazon-PQA dataset that we utilize in this work, which contains an aggregated answer per each question.

Sentence generation. In this step, given a yes-no question and its aggregated answer, the goal is to generate a concise standalone sentence (insight) representing that information. We framed the insight generation task as a neural machine translation task, where the inputs are the yes-no question and the short answer, and the output is the concise standalone sentence. We leveraged a transformer-based model, and experimented with several training sets: a small scale dataset of human generated insights; a large scale out-of-domain dataset; transfer learning between the two; and a few-shot setting.

3.2 Product insight selection

The Q&A section often contains more than one question and for popular products the number of questions can become quite large.

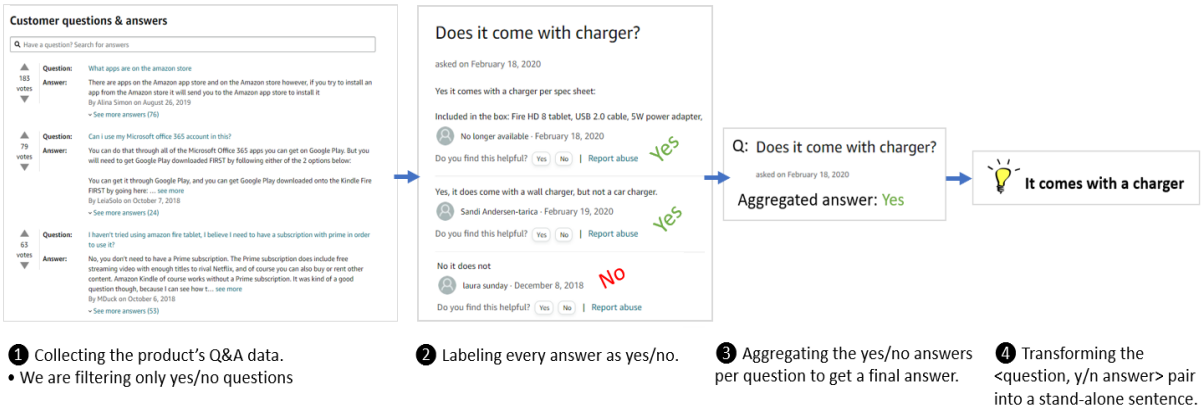

Figure 2: First stage of our pipeline - insight generation

In the Amazon-PQA dataset, over 13% of the products have at least 5 answered yes-no questions. Since the available space for displaying product insights is often limited, a selection mechanism that promotes the most helpful insights is needed. We rely on the following hypothesis (verified in experiments in §4.2): a helpful insight is based on a prominent question that expresses an information need common to many customers.

In order to estimate question prominence, we can simply search for the most popular product questions posted in the customer Q&A widget. However, based on our analysis almost 90% of the questions are asked only ones. Another way customers can seek information on the detail page is to utilize the search widget (located above the Q&A section) to type a query. We found that for 50% of the products[1], the number of search queries on the detail page was at least 4 times greater than the number of questions available. This makes the search history log a valuable resource for understanding customers' information needs. Therefore, we consider two techniques for estimating question prominence:

Log Popularity (*LogPop*). In the Q&A search widget, customers typically type very short queries to express their information needs. In fact, more than 90% of queries have at most 2 words. In order to match between the product questions and the queries, we leverage the existing production algorithm that retrieves a set of existing questions in response to the customer query. We rank questions based on the number of queries they were retrieved for, as a proxy for question popularity.

Category Popularity (*CatPop*). While there are almost no duplicate questions for the same product, we can rank them by their popularity in similar products. For each product, we rank the questions (and their corresponding insights) in descending order by the number of similar questions we find within the product category. Two questions are considered similar if the cosine similarity between their embeddings is above a predefined threshold of 0.87. [2] To embed the questions, we use a Sentence-Transformers model [24], pretrained to find similar questions on Quora dataset.[3]

Diversification. To avoid duplicate questions and select a final diverse set of insights for a given product, we apply a greedy diversity mechanism, iterating over the ranked questions in descending order and selecting a question only if its cosine similarity with previously selected questions is below 0.5.[4] We used the same embeddings as in the selection step to represent the questions. The process ends when a predefined number of questions are selected.

4 EXPERIMENTS

4.1 Insight generation

4.1.1 Datasets. As a source data for our experiments, we used *Amazon-PQA* [26], a publicly available dataset of product Q&A. This dataset contains over 9M questions for over 1.4M unique products, divided to 100 narrow categories. 54% of the questions are yes-no questions, and for each such question an aggregated yes-no answer is provided (36% yes, 16% no, 48% neutral). We filtered out questions where the aggregated answer is neutral, since they represent uncertainty in the answer.

In-Domain dataset (ID) – [5] A small scale manually generated dataset. 50K yes-no questions were randomly sampled from *Amazon-PQA*. These questions and their aggregated yes-no answers were presented to annotators via the Appen annotation platform [6], who transformed each (question, answer) pair into a stand-alone sentence reflecting the information provided in the pair. The annotation underwent grammar correction [25] and cleaning steps, resulting in 19,470 triplets (e.g., *<Is it waterproof?, Yes, The product is waterproof.>*). We split the dataset into train (70%), validation (15%) and test (15%).

Out-Of-Domain dataset (OOD) – A large scale dataset of 315K question, factoid answer and full answer triplets. The dataset was

[1]The analysis was performed on the search log of 551 most popular products
[2]Tuned on the Amazon-PQSim Dataset [26] to achieve a precision of 90% for similar product questions
[3]https://huggingface.co/sentence-transformers/distilbert-base-nli-stsb-quora-ranking
[4]Corresponds to a precision of 77% for non-similar questions, based on the Amazon-PQSim [26] dataset
[5]The dataset is available at https://registry.opendata.aws/amazon-pqa-insights/
[6]www.appen.com

Table 1: Experimental results of the generation models (all numbers are in percents). 'i' and 't' marks statistically significant differences (using a two-tailed paired t-test with p-value ≤ 0.05) with T5-ID and T5-transfer, respectively. Boldface: the largest value in a column.

Training setup	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Readable	Consistent
T5 - OOD	49	33	49	19	–	–
T5 - ID	**67**	51	**66**	**38**	74	95
T5 - OOD→ ID (transfer)	**67**	**52**	**66**	**38**	76	96
GPT-J (few-shot)	61	43	60	30	$\mathbf{85}^{it}$	$\mathbf{99}^{it}$

created [17] using SQuAD [23], a Wikipedia-based Q&A dataset, by matching each question and factoid short answer with the original sentence containing the answer.

4.1.2 Experimental details. We tested several approaches for insight generation, involving supervised direct and transfer learning over the aforementioned datasets, as well as a few-shot setting. For the supervised sequence-to-sequence generation task we used a T5 model [22], initialized with the pretrained 'T5-base'[7] checkpoint which has 220 million parameters. For each dataset, the first two entries (question and short answer) in the triplet were concatenated and fed to the model with the prefix "summarize:", and the last entry (stand-alone sentence) was used as the target. We used an AdamW optimizer with batch size of 8 and learning rate of 1e-4, greedy decoding and a maximal length of 80 tokens. We used repetition penalty of 2.5 to avoid repetition of a phrase in the generated sentence.

In our experiments, we either trained the T5 model directly on one of the datasets for 5 epochs, or performed transfer learning. In the latter approach, we first trained the model on the OOD dataset for 10 epochs, and then further fine-tuned it on the smaller ID dataset for 5 epochs. In both setups, we applied early stopping according to the loss on the validation set, to choose the best model checkpoint.

In the few-shot setting, we used the open-source GPT-J [29] model which has 6 billion parameters and was trained on the Pile dataset [6]. GPT-J has shown impressive performance compared to the 6 billion GPT-3 [3] model on various zero-shot NLP tasks. At inference time the model gets a prompt that consists of the task instruction ("Generate a factually correct statement from the following question and answer pair"), followed by a representative set of 13 examples and one new example (a question and answer pair) that we want the model to generate the insight for (see Figure 3). The model predicts the most probable next tokens that are used as the generated insight. We use a greedy decoding strategy as it yields the best result according to our experiments. The maximum length of the generated text was set to 25 tokens.

4.1.3 Evaluation and results. We compared the generation performance of the model in various training schemes over the test portion of the *ID* dataset (Table 1). We measured the performance via automatic scores (ROUGE [11] and BLEU [19]) and manual evaluation (2 rightmost columns). In the automatic evaluation, we observed that a supervised T5 model trained or fine-tuned on the target dataset outperformed both the OOD T5 model and the GPT-J model. Additionally, transfer learning from OOD (3rd row) led to a

[7]https://huggingface.co/t5-base

Generate a factually correct statement from the following question and answer pair.

[Question] will this work on a 240w power supply? [Answer]: yes [Statement]: This will work on a 240w power supply.
[Question] can these be used for watching tv? [Answer]: yes [Statement]: These can be used for watching TV.
[Question] Does this tv have an optical port? [Answer]: no [Statement]: This TV does not have an optical port.
[Question] Is remote included? [Answer]: yes [Statement]: The remote is included.
[Question] does it come with a phone number and servise? [Answer]: no [Statement]: It does not come with a phone number and service.
[Question] Does it work with the Microsoft surface RT [Answer]: yes [Statement]: It works with the Microsoft surface RT.}
[Question] does this clip player have the fm radio [Answer]: yes [Statement]: This clip player has FM radio.
[Question] Does it charge a lenovo yoga 11s? [Answer]: yes [Statement]: It charges a Lenovo Yoga 11s.
[Question] do i need to purchse the Alexa? [Answer]: no [Statement]: You dont have to purchase Alexa.
[Question] I have a galaxy j7 prime is it for that model ??? [Answer]: yes [Statement]: It is compatible with Galaxy j7 prime.
[Question] Will these fit my 2018 jeep wrangler JL [Answer]: yes [Statement]: These fit 2018 Jeep Warngler JL
[Question] Can we get Whatsapp on this stereo? [Answer]: no [Statement]: You cant get Whatsapp on this stereo.
[Question] Is this watch waterprof? [Answer]: no [Statement]: This watch is not waterproof.

[Question] Is this compatible with Android? [Answer]: no [Statement]:

Figure 3: Few-Shot prompt for GPT-J; typos and grammatical errors are introduced to teach the model how to handle them.

boost in performance, showing the benefit of pre-training on a large scale dataset prior to fine-tuning on the target domain. The Rouge and Bleu metrics measure how well the generated text matches the ground-truth text in terms of overlapping n-grams. Therefore, in-domain training allows the model to learn the specific vocabulary and patterns of that domain, resulting in a higher number of overlapping n-grams with the ground truth sentences. Moreover, the GPT-J model outperformed the out-of-domain T5 model. GPT-J is pre-trained on a large amount of general text data, which may enable it to quickly adapt to a new domain with just a few examples.

We further compared the methods via manual evaluation of insights generated from 300 Q&A pairs (we excluded the worst performing method from this evaluation). Four in-house expert annotators were asked whether each generated sentence was: (i) readable – clear and has no significant grammar mistakes, (ii) consistent with the yes-no answer, and (iii) hallucinating any information that was not present in the (question, answer) pair. Readability and consistency rates are shown in Table 1. We used 200 sentences to calculate annotator agreement between annotator pairs. Cohen's Kappa scores were 0.59, 0.89 and 0.49 for tasks (i), (ii) and (iii) respectively (between fair to excellent agreement). We found that 99.9% of the generated sentences (via all methods) did not hallucinate irrelevant information. For T5, the transfer learning setting reached higher readability and consistency scores than direct training. The best performing model was GPT-J, suggesting that while the insights generated by GPT-J differed from the ground truth insights, this model is better at generating text that is more readable and accurate for humans. An analysis of the error cases shows that the major cause for inconsistency was wrong model/item name (e.g, *chevy tahoe* in the question was replaced with *chevy tie* in

the generated insight). We additionally observe that unreadable insights are mostly (65%) due to minor grammar or spelling errors, and only 35% of them are completely unclear.

Henceforth, we utilize the T5 model for insight generation, as its complexity is an order of magnitude smaller compared to GPT-J. Moreover, despite slightly lower readability and consistency scores, T5 outperforms GPT-J in terms of similarity to human-written insights.

4.2 Insight selection - evaluation and results

We evaluate the insight selection procedure on a set of 551 products with the highest number of detail page views in headphones, laptops, mobile phones and televisions categories. As a baseline, we ranked the insights completely at random and compare the results to our popularity based methods. For each product, the top 5 insights retrieved by each selection model were annotated. Ten in-house annotators labeled the insights as helpful/non-helpful. A helpful insight should contain valuable information that helps the customer make a purchase or usage decision.

The random baseline reached 73% helpfulness, supporting the assumption at the basis of our work, that Q&A contain valuable product information. The results also showed that CatPop and LogPop achieved impressive helpfulness scores of 93.6% and 92.2%, respectively. Further analysis of the top 5 ranked insights revealed that 70% of these insights were consistent across both methods, which may explain the similar results we observed. These results attest to the merits of selecting insights by their corresponding question prominence.

We next analyzed the key reasons the annotators marked insights as unhelpful. The leading cause (31% of errors) was lack of clarity, usually as a result of grammar mistakes or lack of context in the originating question (e.g., *This phone has 6.0*, based on the question *does the phone have 6.0?*). Other errors were over-specificity (20%), i.e., insights about a feature that is too niche, e.g., *These headphones block out really loud snoring.*, or overly general insights (15%), e.g., *This TV is really worth buying.*

For the remainder of the paper, LogPop will be used as the primary ranking method because it is derived from customers actual searches for the specific product rather than for products in the same category. We do however, view CatPop as an important alternative as it better handles the cold start issue, i.e. products for which search information has not been accumulated yet.

5 ONLINE EXPERIMENT

In order to evaluate the helpfulness of the insights, we set up an experiment on "Alexa's Insights" widget on the retail website. The widget is placed in the product detail page and aims to help customers with their shopping decisions by providing a snapshot view of customer review aspects and product information as presented in Figure 4 (a).

We conducted a user study to learn more about how customers perceive the helpfulness of Alexa's Insights. Ten customers were asked to imagine they were shopping for a pair of wireless headphones, and that they go to Amazon and find the new Apple AirPods Pro. The customers had a full length detail page prototype with the Alexa's Insights widget included. Testers were asked how helpful

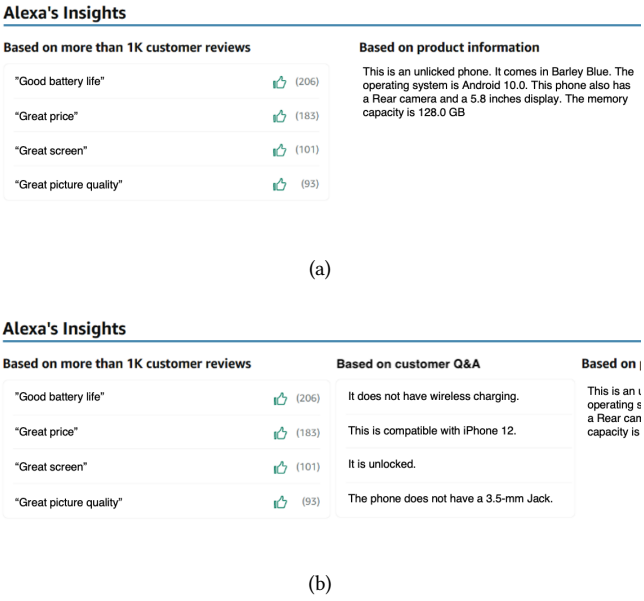

Figure 4: Alexa's Insights widget on the desktop. (a) Control - existing widget in the product detail page; (b) Treatment - our insights are added.

would they find Q&A insights on a 5-point rating scale. The majority of testers thought that the insights were helpful and would save them time of having to looking through all of the Q&A ("Very Helpful": 5 votes, "Helpful": 4 votes, "Neither": 1 vote).

Encouraged by the user study results, we conducted an online A/B test to measure the potential impact of our insights. When customers viewed the detail page of a supported product (in US), and Q&A insights were present, they were allocated to one of the following groups: (1) Control - existing widget with Review Aspects and product information. (2) Treatment - the Q&A insights are shown between Review Aspects and product information (see Figure 4 (b)).

During the experiment we measured a *LongTermBenefit* metric which is an estimate for long term customer activity on the e-commerce platform. The estimation is derived from customer's purchases and other activities, such as searches, performed on the platform. Based on a 28-day analysis of the experiment we observed a statistically significant improvement of 0.14% in LongTermBenefit in the treatment group compared to the control group. This positive result demonstrates that the new insights provide useful information to customers that help them make more informed and confident purchasing decisions.

6 DATA ANALYSIS

So far our results show that our method successfully generates helpful, well-formed and relevant insights. In this section we examine whether the generated insights add new information beyond customer reviews and beyond the product description provided by the seller.

Comparison to customer reviews. As mentioned in Section 1, a key hypothesis motivating our work is that product reviews

Table 2: Examples for insights generated from Q&A and human review summaries.

Product	Top Insight generated from Q&A	Human summary from reviews
Laptop Backpack	1. This fits the size restriction to be used as carry on during air travel. 2. It can carry 60 pounds of material. 3. This is also good for grocery. 4. The piping on this bag is reflective. 5. This can pass as a carry-on baggage of 10 H x 17 W x 24 L.	Verdict: The biggest of the backpacks on our list makes it a good choice for school or adventure. Pro: Features a multi-compartment design and a laptop sleeve that fits most 17-inch laptops for school as well as tuck away waist belt and rain cover for adventures Con: Less durable than some of the other models on our list
Headphones	1. The cord has a volume control adjustment. 2. The cord is removable. 3. It will work with the Samsung Galaxy S4. 4. The headphones deliver stereo sound. 5. There is a difference between the 300 and the 500 in terms of sound quality.	Verdict: These headphones produce quality sound, especially for this price range, and they're comfortable enough to wear for hours at a time. Pro: Adjustable headband with comfortable ear cups that you can wear for long periods at a time Good sound quality Long cable gives you a good range of motion Lightweight, yet durable Con: There's no padding on the headband Can't be used wirelessly

do not cover all information needs of online shoppers, and that this information gap is covered by product Q&A. To verify this hypothesis, we demonstrate that product insights generated from Q&A complement human written summaries of product reviews, i.e., the information they provide is not covered by the review based summaries. For the sake of this experiment we rely on the recent AmaSum dataset[2], which contains human-written summaries for 31, 483 Amazon products where each summary consists of verdict, pros, and cons. We selected 1,624 products based on the availability of Q&A data for these items. Table 2 presents some examples of insights generated by our method along with the corresponding AmaSum human summaries.

We first perform a qualitative comparison by presenting annotators with 50 summaries paired with top-5 insights for the corresponding products. Annotators were asked to count the number of insights in the top-5 list, which overlap the content of the associated summary. We found that 31 products had no overlap between the top insights and the summary and that the average overlap across all 50 products was 0.54 where 0 indicates no overlap and 5 indicates full overlap.

We also opted for a quantitative comparison on a larger scale (the entire subset of 1624 products) and found a similar trend, namely that the overlap between insights and summaries is small. For each product we computed the mean similarity between summary sentences and top-5 insights (using cosine similarity over the Sentence-Transformers embeddings of both sentence and insight [8]). As a reference for these similarities we compute the intra-similarities of sentences within the summary and of insights within the top-5 set. The average cross similarity between the top-5 insights and summary was 0.176 compared with 0.265 and 0.289 for the intra-summary and intra-insights respectively (both differences are significant using a paired t-test). We observe a similar trend when taking maximal similarity instead of average.

To get a better understating of the different nature of insights generated from Q&A and those extracted from reviews, we applied a subjectivity classifier[9] on each of the top insights and on every sentence from the AmaSum summaries. We found that 55% of the review-based summaries provide subjective opinions about the product, as opposed to more than 80% objective insights generated from Q&A. These analyses support our view that Q&A provide complementary information to reviews. Together with the fact that many customers interact with the Q&A section we conclude that such insights are both complementary and necessary building blocks for product summarization systems that truly handle customers' information needs.

Comparison to description. We analyzed the top insights and the corresponding product descriptions for the same set of products we used in the previous analysis. We found that 81% of the top insights add new information which is not covered by the existing product description.[10] In the remaining 19%, the customer may wish to corroborate information with other customers, e.g., *Does it really work outdoors?*, or to verify claims such as wide model compatibility (e.g., The Charger is compatible with all Samsung models) for a specific model (*Does it charge S8?*). We additionally found that 15% of the insights refer to a missing property or feature in the product, e.g., *You can't record FM radio*, and *The product is not compatible with Alexa.*

7 CONCLUSIONS

We addressed the novel challenge of generating product insights based on community Q&A. We presented a method that converts the Q&A into a self-contained insight, and then selects the most helpful insights by leveraging the question popularity. Both annotator-based evaluation and an online experiment on Amazon's website demonstrate the merits of our method for customers. Moreover, we found that our method retrieves helpful information that is often missing from the product description, and that Q&A-based insights are different from insights extracted from customer reviews. Future directions include integrating the generated insights in additional use-cases and examining how customers engage and benefit from them. We also plan to expand our method beyond yes-no questions to open-ended questions.

[8]We used the all-mpnet-base-v2 model
[9]Part of sentiment-analysis tool in https://textblob.readthedocs.io
[10]We refer to the product's title, bullet-points, free text description and the table of product features as "product description"

REFERENCES

[1] Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2020. Generating Well-Formed Answers by Machine Reading with Stochastic Selector Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7424–7431. https://doi.org/10.1609/aaai.v34i05.6238

[2] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning Opinion Summarizers by Selecting Informative Reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] 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.

[5] Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying Helpful Sentences in Product Reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 678–691.

[6] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[7] Jing He and Decheng Dai. 2011. Summarization of yes/no questions using a feature function model. In *Asian Conference on Machine Learning*. PMLR, 351–366.

[8] Sharon Hirsch, Slava Novgorodov, Ido Guy, and Alexander Nus. 2021. Generating Tips from Product Reviews. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 310–318.

[9] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 168–177.

[10] Moez Limayem, Mohamed Khalifa, and Anissa Frini. 2000. What makes consumers buy from Internet? A longitudinal study of online shopping. *IEEE Transactions on systems, man, and Cybernetics-Part A: Systems and Humans* 30, 4 (2000), 421–432.

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

[12] Mengwen Liu, Yi Fang, Alexander G Choulos, Dae Hoon Park, and Xiaohua Hu. 2017. Product review summarization through question retrieval and diversification. *Information Retrieval Journal* 20, 6 (2017), 575–605.

[13] Gerald L Lohse and Peter Spiller. 1998. Quantifying the effect of user interface design features on cyberstore traffic and sales. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 211–218.

[14] Duy Khang Ly, Kazunari Sugiyama, Ziheng Lin, and Min-Yen Kan. 2011. Product review summarization from a deeper perspective. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 311–314.

[15] Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. 2019. Generating product descriptions from user reviews. In *The World Wide Web Conference*. 1354–1364.

[16] Nadav Oved and Ran Levy. 2021. Pass: Perturb-and-select summarizer for product reviews. 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)*. 351–365.

[17] Vaishali Pal, Manish Shrivastava, and Irshad Bhat. 2019. Answering naturally: Factoid to full length answer generation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. 1–9.

[18] Vinay Pande, Tanmoy Mukherjee, and Vasudeva Varma. 2013. Summarizing answers for community question answer services. In *Language Processing and Knowledge in the Web*. Springer, 151–161.

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

[20] Dae Hoon Park, Hyun Duk Kim, ChengXiang Zhai, and Lifan Guo. 2015. Retrieval of relevant opinion sentences for new products. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 393–402.

[21] Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*.

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 1–67.

[23] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[25] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. *arXiv preprint arXiv:2106.03830* (2021).

[26] Ohad Rozen, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering Product-Questions by Utilizing Questions from Other Contextually Similar Products. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 242–253.

[27] Sunil Saumya, Jyoti Prakash Singh, Abdullah Mohammed Baabdullah, Nripendra P Rana, and Yogesh K Dwivedi. 2018. Ranking online consumer reviews. *Electronic commerce research and applications* 29 (2018), 78–89.

[28] Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing answers in non-factoid community question-answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 405–414.

[29] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

[30] Naitong Yu, Minlie Huang, Yuanyuan Shi, and Xiaoyan Zhu. 2016. Product review summarization by exploiting phrase properties. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 1113–1124.