Identifying Helpful Sentences in Product Reviews

Iftah Gamzu[1] **Hila Gonen**[1] **Gilad Kutiel**[1] **Ran Levy**[1] **Eugene Agichtein**[2,3]
{iftah,gonenhi,gkutiel,ranlevy,eugeneag}@amazon.com

[1]Amazon, Tel-Aviv, Israel
[2]Amazon, Seattle, WA, USA
[3]Emory University, GA, USA

Abstract

In recent years online shopping has gained momentum and became an important venue for customers wishing to save time and simplify their shopping process. A key advantage of shopping online is the ability to read what other customers are saying about products of interest. In this work, we aim to maintain this advantage in situations where extreme brevity is needed, for example, when shopping by voice. We suggest a novel task of extracting a single representative helpful sentence from a set of reviews for a given product. The selected sentence should meet two conditions: first, it should be helpful for a purchase decision and second, the opinion it expresses should be supported by multiple reviewers. This task is closely related to the task of Multi Document Summarization in the product reviews domain but differs in its objective and its level of conciseness. We collect a dataset in English of sentence helpfulness scores via crowd-sourcing and demonstrate its reliability despite the inherent subjectivity involved. Next, we describe a complete model that extracts representative helpful sentences with positive and negative sentiment towards the product and demonstrate that it outperforms several baselines.

1 Introduction

Customer reviews are known to be a valuable source of information for potential buyers. This is evident from the high engagement of customers with reviews, for example by up-voting a review for its helpfulness.[1] As online shopping platforms attract more traffic it is becoming increasingly difficult to consume the wealth of information customers share. For this reason, helpful reviews (defined as such by costumers) are made more visible than those that are less helpful.

[1]At the time of writing this paper, a review for the Echo Dot 3rd generation received more than 10K votes.

The topic of review helpfulness has attracted a lot of academic interest in which reviews were always considered as a whole (see Diaz and Ng (2018) for a survey). However, in some scenarios, such as the limited real-estate in mobile screens, or in voice interactions with a virtual assistant, presenting a full review is impractical and the need to automatically extract helpful excerpts arises. While in the mobile scenario, a persistent customer may still be able to read the entire review, the voice scenario is inherently more challenging as it demands patience and focus from the customer, while the assistant reads the text out loud. As a result, the need for extreme brevity and the ability to understand what matters most to customers becomes crucial.

In addition to brevity and helpfulness, another desirable property from the extracted content is being faithful to the reviews as a whole. Indeed, a costumer looking for relevant and helpful reviews, often interacts with more than one review before making their decision, trying to pinpoint those helpful bits of information that are shared by multiple reviewers. This process is tedious because of the sheer amount of reviews and biased because of the order they appear in. A system that aims to replace this process while maintaining trust in the content it provides should be able to extract concise helpful texts that repeat across multiple reviews, indicating that they are faithful to the reviews' content (from here onward we shall refer to such texts as "faithful").

Our goal is to extract such sentences, i.e., sentences that are both **helpful** for a purchase decision and **faithful**. To this end, we first define two new notions: A *Helpful Sentence* is a sentence which is considered helpful by the average customer in their purchase decision process. A *Representative Helpful Sentence (RHS)* is a helpful sentence that is also highly supported, that is, the ideas it expresses appear in multiple reviews for the given product (not necessarily in the exact same wording).

It is traditionally assumed that judging the importance of a text excerpt requires reading the entire text. We challenge this assumption, at least in the domain of product reviews, and collect a dataset of single review sentences with their helpfulness scores by averaging the scores assigned to them by multiple crowd workers. We show that despite the highly subjective nature of this task, and despite the fact that workers are exposed to sentences without their surrounding context, the resulting scores are reliable. Using the data we collected, from 6 different categories, ranging from Electronics to Books, we train and evaluate several supervised algorithms to predict helpfulness score, which achieve promising results. Finally, we present an initial implementation of a model that given a set of product reviews, extracts a single positive RHS (supports the purchase) and a single negative RHS (opposes the purchase).

In summary, the main contributions of this work are: (1) We propose a novel task that given a set of reviews for a product, outputs a single sentence that is both helpful for a purchase decision and supported by multiple reviewers; (2) We show that the helpfulness of a sentence can be reliably rated based solely on the sentence, allowing for an efficient dataset creation. These helpfulness scores can be leveraged for other tasks such as highlighting important parts of a review; (3) We publish a novel dataset of sentences taken from customer reviews along with their helpfulness score;[2] (4) We develop an end-to-end model for our task that shows promising results and outperforms several baselines.

2 Related Work

Review Helpfulness Modeling and Prediction Customer reviews are a valuable source of information for customers researching a product before making a purchase (Zhu and Zhang, 2010). Diaz and Ng (2018) survey recent work on the tasks of modeling and predicting review helpfulness. While some researchers treat helpfulness votes as ground-truth, others have argued that these votes are not good indicators for actual review helpfulness (Liu et al., 2007; Tsur and Rappoport, 2009; Yang et al., 2015).

Some general observations have been made based on helpfulness votes, e.g., review length has been shown to be strongly correlated to helpfulness (Kim et al., 2006; Liu et al., 2007; Otterbacher, 2009; Mudambi and Schuff, 2010; Pan and Zhang, 2011; Yang et al., 2015). Another widely-agreed indication for review helpfulness is the review star rating (Kim et al., 2006; Mudambi and Schuff, 2010; Pan and Zhang, 2011).

A related dataset was presented in Almagrabi et al. (2018). The main advantages of the dataset we create over this previously suggested one are: (1) **Binary vs. continuous scores** – We use continuous scores rather than binary scores. Our aim is to surface the most helpful sentences, which is not possible if many of the sentences are annotated as equally helpful; (2) **Range of products/domains** – The previous dataset includes only 5 products, all from the Electronics domain. Our dataset is significantly more diverse, providing annotations for 123 products from 6 different domains, allowing to evaluate a model's ability to generalize across domains.

Product Review Summarization The most common approach for product review summarization, which centers the summary around a set of extracted aspects and their respective sentiment, is termed *aspect based summarization*. One of the early abstractive works, by Hu and Liu (2004), was designed to output lists of aspects and sentiments. Other works target a traditional summarization output and at times somewhat simplify the task by assuming aspects or seed words are provided as input (Gerani et al., 2014; Angelidis and Lapata, 2018; Yu et al., 2016). Recently advances were made on unsupervised abstractive reviews summarization, by leveraging neural networks (Chu and Liu, 2019; Bražinskas et al., 2020b) followed by a few shot variant (Bražinskas et al., 2020a).

Extractive summarization include earlier works such as Carenini et al. (2006); Lerman et al. (2009) and Xiong and Litman (2014) who suggested to use review helpfulness votes as means to improve the content extraction process. More recently, Tan et al. (2017) suggested a novel generative topic aspect sentiment model.

3 Representative Helpful Sentences

Task Definition In this work, we focus on summarization of reviews in the setting of shopping over voice with the help of a virtual assistant. Our goal is to provide users with content that is both **helpful** and **faithful** in this challenging setting

[2]The dataset is available at https://registry.opendata.aws/.

where the information the user can absorb is extremely limited. First, we aim to maximize the informativeness, while maintaining brevity. To this end, we introduce a new notion of *helpful sentences* – sentences which the average costumer will consider as helpful for making a purchase decision. Next, to ensure faithfulness, we introduce the notion of *support* for a given sentence – the number of review sentences with a highly similar content. We seek to **automatically** identify a helpful sentence with a wide support, which we term *representative helpful sentence (RHS)*. Note that Representative Helpful Sentences, being supported by many similar sentences, are by construction faithful to the review pool from which they are extracted. We restrict ourselves to single sentences that are extracted as-is from product reviews, as this serves as another mechanism to ensure faithfulness. We do not restrict the number of reviews in the input. Table 1 presents a few helpful sentences for example, as extracted by our model (see Section 5).

Product	Representative Helpful Sentence
Toy Pirate Ship	It was easy to put together, is the perfect height, and very durable.
Headphones	They fit well and produce good sound quality.
Speakers	Quality good, price okay, sound output great.

Table 1: Example Representative Helpful Sentences.

Our task resembles the well known (extractive) customer review summarization task (Hu and Liu, 2004) but differs in several important aspects. First, its output is very concise due to the extreme space constraint, resembling the extreme summarization task (Narayan et al., 2018), which however, deals with news articles and outputs an abstractive summary. In our application there is low tolerance for factually incorrect summaries, so we choose extraction over abstraction. Second, we do not restrict the system's output to aspect based opinions, as we find that sometimes factual content may also be quite helpful. Third, while traditional summarization systems favor information that appears frequently in the source documents, we target information that is both frequent and helpful.

Subjectivity As mentioned above, review helpfulness scores are derived from votes of actual customers. Deciding on whether or not to up-vote a review is a subjective decision as different customers may value different product qualities. However, the underlying assumption of the voting mechanism is that reviews with many up-votes are indeed helpful for the average customer. Restricting the user to a single sentence makes matters even more challenging as it cannot possibly discuss all the product merits and shortcomings. To emphasize the subjectivity involved in assigning a helpfulness score for a standalone sentence, consider the examples in Table 2. The first example may be helpful for parents looking to buy a book for their children but entirely unhelpful for adults who wish to purchase the book for themselves. Similarly, the second one is more helpful to readers of extreme height (short or tall) than to those of medium height.

Product	Sentence
Harry Potter book	Finding 1 book that keeps your child intrigued and helps him or her develop a love for reading is amazing.
Jump rope	It's a pretty standard jump rope but it's really nice and you can adjust the length which is perfect because I'm really short.

Table 2: Review sentence examples.

Despite the evident subjectivity, we assume that there exists an "average helpfulness" score for every sentence, which can be estimated by averaging the ratings of multiple crowd workers. In the following section we establish this assumption by compiling a new dataset of sentences along with their helpfulness scores, and showing quantitatively that the annotations in our dataset are consistent and reliable.

4 Helpful Sentences Annotation

Our main contribution in this work lies in the notion of *helpful sentences* and the ability to identify such sentences without observing entire reviews. In what follows, we describe the process of compiling a dataset of sentences along with their helpfulness scores using crowdsourcing. Note that this dataset is intended solely for scoring helpfulness of sentences. Faithfulness is ensured by other means which are not reflected in the dataset, i.e. by requiring a RHS to have a wide support of similar sentences, as discussed in section 3 and implemented in our model, as described in Section 5.

4.1 Annotation Task

We consider a subset of 123 products arbitrarily selected from the Amazon.com website, so that

each has at least 100 customer reviews and they (approximately) equally represent 6 different categories (Toys, Books, Movies, Music, Camera and Electronics). We started with 45,091 reviews, split them into 210,121 sentences and randomly selected a train set with 20,000 sentences, and a test set with 2,000 sentences. We asked annotators to rate each sentence according to how helpful it is for reaching a purchase decision, using the Appen platform.[3] Ratings were provided on a 3-level scale of **Not Helpful** (0), **Somewhat Helpful** (1), or **Very Helpful** (2). The final helpfulness score of a given sentence was set to the average rating. See Section A in the Appendix for more details on the annotation task guidelines.

Each example was rated by 10 different annotators in the training set and 30 different annotators in the test set. Initial experiments revealed that 10 annotations per sentence, while noisy, are still sufficient to train a model. We observed that the number of annotators used to calculate the test set affects the evaluation. This is due to the subjective nature of this task and the observed helpfulness score that becomes closer to its "real" score as the number of votes collected for each sentence increases. Table 3 demonstrates the effect the number of annotators used to rate each example in the test set has on the final evaluation. It shows that after fixing the model and predictions, the evaluation score (Pearson correlation in this case) increases as we average more votes. From our experience, there is no gain beyond 30 votes per sentence for this particular task.

# of votes	1	10	20	25	30
Pearson	0.523	0.776	0.822	0.831	0.838

Table 3: For a fixed prediction the correlation between the prediction and the scores obtained by averaging individual scores increases as we consider more votes per sentence. The phenomenon is not unique to correlation.

We observe a skewed helpfulness distribution with a fairly high mode of 1.3 which shows that the raters did not provide random answers. Furthermore, under the assumption that most review authors aim for their reviews to be helpful, we should expect a distribution that is skewed towards higher scores. See Section A in the Appendix for a depiction of the helpfulness distribution within the train set.

[3] www.appen.com

Table 4 presents the most helpful sentence, a sentence that is somewhat helpful (with a median score) and the least helpful sentence from the test set for particular headphones as perceived by the annotators.

Sentence	Helpfulness
Really great headphones, especially for $25, but honestly, they sound better than my gaming headset and my DJ headphones in many respects.	1.97
Call quality just can't be beat for ear buds	1.47
Any thoughts from others?	0

Table 4: The most helpful, a somewhat helpful and the least helpful sentences for particular headphones.

4.2 Annotation Analysis

As mentioned earlier, rating sentence helpfulness is a highly subjective task, and some disagreement is expected. Nevertheless, we argue that the data we collected is reliable and demonstrate it through the three following experiments.

Inter-annotator Agreement We compute agreement in the spirit of the analysis performed in (Snow et al., 2008). For each annotator, we restrict the data to the set of rows that they completed and compute the Pearson correlation between their answers against the average of all other annotators. Finally, we take the average across all annotators after removing the worst 10% annotators according to the method of (Dawid and Skene, 1979). We get an average of 0.44 ± 0.01 Pearson correlation on the train set (10 annotators per row) and 0.57 ± 0.02 on the test set (30 annotators per row), which demonstrates good agreement given the subjective nature of this task.[4] We also randomly split the annotators into two disjoint sets and calculated the correlation between the corresponding scores. There was a correlation of 0.49 for the train set and 0.81 for the test set.

Internal Consistency A necessary condition for ensuring reliability is that similar sentences get similar helpfulness scores. We verify that our crowdsourced test data meets this requirement by measuring the standard deviation of the helpfulness scores within groups of similar sentences. We use

[4] This scores are comparable, for example, with the scores reported in Snow et al. (2008) for the highly subjective Affective Text Analysis task.

the sentence-transformers embeddings of Reimers and Gurevych (2019) which were optimized for computing semantic similarity. For each sentence in the test set, we construct its semantic neighborhood by grouping together all sentences with high similarity. For each non-singleton group, we measure the standard deviation of the helpfulness score and compare it with the standard deviation of a similarly sized group of random sentences from the test set. We expect to get a tighter distribution of helpfulness scores within the similarity groups (compared to the random groups) if the data is internally consistent. Indeed, we found 217 groups with an average standard deviation of 0.16 while the average standard deviation of the corresponding random groups was 0.29.[5]

Sentence Helpfulness vs. Review Helpfulness As the third and final reliability analysis, we compare the crowd helpfulness scores with review helpfulness votes taken from the Amazon.com website. We consider reviews for the 123 products selected earlier, and extract two subsets. The first (the helpful set) is the set of all reviews with at least 50 helpful votes. The second (the unhelpful set) is the set of all reviews with no helpful votes. See Section B in the Appendix for statistics on the two subsets. We randomly select 500 sentences from each set and collect crowd helpfulness ratings. For each set we calculate the mean helpfulness score and the ratio of sentences with helpfulness score greater than 1 and 1.5 respectively. Table 5 shows the results which demonstrate a higher mean helpfulness score in the helpful set.[6]

	Mean Score	Ratio Score > 1	Ratio Score > 1.5
Helpful Set	1.21	70%	18%
Unhelpful Set	1.15	64%	14%

Table 5: Contrasting Sentence Helpfulness Score with Review Helpfulness votes — Results.

These results indicate that helpful reviews tend to include more helpful sentences on average. However, as can be expected, the differences are not dramatic. Looking at the average length of reviews sheds some more light on the differences: a helpful review is almost 10 times longer than a non helpful review on average. This means that in order for a review to be helpful it must provide details, a requirement that a single sentence simply cannot meet. Therefore, we conjecture that a helpful sentence captures the most essential statements made in the review while a helpful review is one that includes details and justifies its rating.

4.3 Analysis of Helpful Sentences

A brief examination of the crowd-sourced data reveals two sentence characteristics that contribute to the helpfulness of a sentence: the length of the sentence and the sentiment, which is more strongly correlated with helpfulness.

Length The Pearson correlation between the length (in characters) and the helpfulness score on the test set is 0.37. This correlation is expected, since longer sentences can potentially convey more information and thus tend to be more helpful.

Sentiment We use Amazon AWS comprehend sentiment analysis tool[7] to classify each sentence into one of four sentiment classes: positive, negative, neutral and mixed. We got a negative Pearson correlation of -0.53 between the helpfulness scores of the sentences and the scores assigned to the neutral class. To better understand this relationship, we define a helpful sentence as one with score greater or equal to 1.5 and a sentence with sentiment as one that is not in the neutral class, and estimate two conditional probabilities:

$$P(\text{Helpful} \mid \text{Sentiment}) = 0.15$$
$$P(\text{Sentiment} \mid \text{Helpful}) = 0.68$$

This shows that having sentiment is an important condition for a sentence to be helpful, but it is not a sufficient condition. We indeed observed that sentences with sentiment that do not provide additional reasoning or details do not get high helpfulness scores. Some related examples from reviews can be found in Section C in the Appendix.

5 Surfacing Representative Helpful Sentences

We now turn to create an end-to-end model for surfacing representative helpful sentences (RHS): given a set of reviews for a certain product, we aim

[5]The differences were statistically significant with a p-value of $7E-20$ using a paired two-tailed t-test.

[6]The difference is statistically significant with a p-value of approximately 0.0079 using a t-test with an equal variance assumption as well as t-test with different variance assumption, a.k.a Welch's t-test.

[7]https://aws.amazon.com/comprehend/

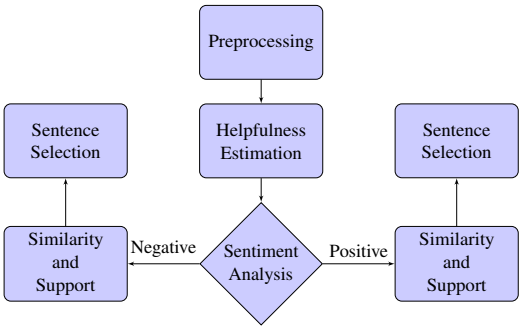

Figure 1: High-level overview of our model.

to output a single RHS with positive sentiment and a single RHS with negative sentiment. Figure 1 depicts the different sub-components of our model. Given a set of reviews, we preprocess the input and predict helpfulness scores for each of the sentences. Next, we analyze the sentiment of each sentence and separate into positive and negative sets. Following that, the support of each sentence is determined, and finally we select the RHS sentence based on its helpfulness score and its support. In what follows, we describe each of the components in details.

Preprocessing We remove HTML tags and split the cleaned reviews into sentences. The sentences are then filtered by removing sentences of extreme length (both short and long). See Section D in the Appendix for additional details.

Helpfulness Estimation This component assigns a helpfulness score for each sentence and removes all sentences with score below 1. This filtering serves two purposes: First, it ensures that we do not output any sentence in case there is no helpful sentence in the product reviews. Second, it reduces the runtime of the downstream Similarity and Support component which is quadratic in the number of sentences.

We experiment with three helpfulness models and find that a pre-trained BERT (Devlin et al., 2018) fine-tuned on our training data performs best. The two other models we compare are: (1) TF-IDF: a model that treats each sentence as a bag-of-words. We use `TfidfVectorizer` from the `sklearn` package to convert each sentence into a vector and then fit a Ridge regression model on top of it; (2) ST-RIDGE: a model that fits a Ridge regression on top of the Sentence-Transformers embedding (Reimers and Gurevych, 2019).

We use 3 measures for evaluation: Mean Squared Error (MSE), which is the traditional measure for regression, Pearson correlation between the predicted score and the ground-truth score, and finally a ranking measure that evaluates the quality of the top ranked sentence (NDCG@1). The results are depicted in Table 6. The TF-IDF model has an acceptable performance but it suffers from out-of-vocabulary problem and ignores the sentence as a whole, for example, the model predicts a higher score than that of the annotators to the sentence "fantastic brilliant amazing superb good". In order to gain some understanding into what constitutes a helpful sentence, we checked the top positive and negative features of this model.[8] We observed that the top positive words include sentiment words and product aspects. The results, however, indicate that these features are not sufficient to evaluate the helpfulness in a more fine-grained manner. The ST-RIDGE model significantly outperforms the TF-IDF model in all metrics. Finally, the BERT model is significantly better than the ST-RIDGE model in terms of MSE and Pearson correlation.

	MSE	Pearson	NDCG@1
RANDOM	0.5 ± 0.026	0.018 ± 0.065	0.68 ± 0.044
TF-IDF	0.09 ± 0.006	0.63 ± 0.055	0.91 ± 0.022
ST-RIDGE	0.062 ± 0.0042	0.78 ± 0.037	0.94 ± 0.015
BERT	0.053 ± 0.0037	0.84 ± 0.022	0.95 ± 0.015

Table 6: Evaluation of Helpfulness Prediction (with confidence intervals).

Sentiment Analysis In this step, we employ the Amazon AWS comprehend sentiment analysis tool to assign each sentence a sentiment class and a score for each of the four classes: positive, negative, neutral and mixed. Sentences with a neutral or mixed classes are removed and all the rest are divided into a positive set and a negative set. The purpose of this step is twofold: first, the separation allows us to output a final sentence for both positive and negative sentiments. Second, we gain more confidence that semantically similar sentences (as measured in the downstream Similarity and Support component) have indeed the same meaning (and not the exact opposite).

Similarity and Support At this stage we aim to compute the support of each sentence, which we define as the size of the set of highly similar sentences. Formally, for a given sentence s_i, its support is $|\{s_{j\neq i} | sim(s_i, s_j) > \sigma\}|$, where σ is a predefined threshold.

[8] Top-10 positive features: great, sound, quality, good, excellent, price, easy, lens, recommend, perfect. Top-10 negative features: bought, review, know, don, got, amazon, gift, reviews, christmas, order.

To compute the similarity, we convert each sentence pair to the corresponding representations and compute the cosine similarity. In order to get the most accurate results, we compare several sentence representations on the semantic similarity task: the Sentence Transformers (Reimers and Gurevych, 2019), the Universal Sentence Encoder (USE) (Cer et al., 2018), FastText (Mikolov et al., 2018), and a bag-of-words representation weighted by the inverse document frequency. We find that the Sentence Transformers embeddings perform best.

To compare the methods, we sample 300,000 sentence pairs from the reviews of our 123 products, compute the similarity scores on this sample and select the top 500 pairs using each of the methods. We next consider the union of the above pairs to form a dataset of 2,035 pairs. We ask human annotators to determine if the sentences of each pair have a roughly similar meaning or not. We then calculate the precision at K (for K between 1 and 2,035) for each of the methods. As can be seen from Figure 2, Sentence-Transformers is superior to the other methods.

Finally, we derived a precision-oriented similarity score threshold ($\sigma = 0.876$) for Sentence Transformers that achieves a precision of 0.9 ± 0.286 and a recall of 0.46 ± 0.022 where the recall is estimated based on the set of 2,035 pairs.

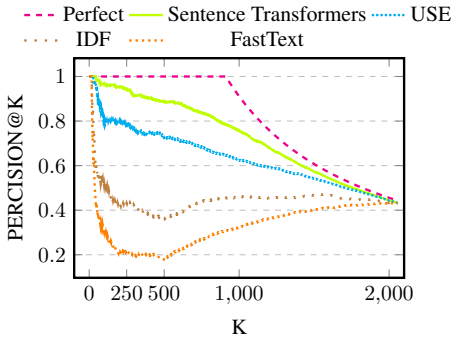

Figure 2: Comparison of Similarity Measures.

Sentence Selection The Sentence Selection component is in charge of selecting a single sentence that is both helpful and well supported. We enforce a minimum support of 5, as we observed that such a limit increases the overall quality of the sentence and avoids surfacing esoteric opinions. After applying this threshold, we rank the remaining sentences according to the formula: $\text{support} \times \text{helpful}^{\alpha}$, where α is a boosting parameter. To derive an appropriate value for α we conducted another annotation task and obtained a value of $\alpha = 38.8$ that gives a lot of emphasis to the helpfulness score. We describe this in detail in Section D in the Appendix.

6 Evaluation

The evaluation of our end-to-end model is challenging and does not have a natural scheme. Recall that we do not restrict our input to small random samples of the review set, as commonly done in review summarization, and was shown to produce biased results (Shapira and Levy, 2020). Instead, we allow for dozens or hundreds of reviews per product. Thus, we cannot expect annotators to carefully read the full input before choosing an RHS. Nonetheless, we show that our notion of helpfulness is indeed useful for surfacing important review content by comparing our models to previous summarization works in two different settings.

Single Review Summarization In this evaluation we only consider the helpfulness component, as means to create an extractive summary comprised of a single sentence.

Abstractive single review summarizers (Ma et al., 2018; Isonuma et al., 2019; Wang and Ren, 2018) are not suitable for comparison as these works are trained on header-like summaries of 4.36 words on average, much shorter than our extractive, one-sentence output. Instead, we consider the unsupervised single document summarization algorithm Textrank[9] (Mihalcea and Tarau, 2004). Textrank, which is extractive and can output any number of sentences, is a viable candidate for comparison as our goal is not to achieve SOTA results on this task, but rather to demonstrate that the helpfulness model can produce good extractive summaries without being trained on reference summaries.

We selected a sample of 300 reviews, in which the prediction of the two algorithms differed (the output was exactly the same on 28% of the reviews), and asked crowd workers to rate each of the selected sentences in a 5-level scale according to how helpful the selected sentence was for a purchase decision (our objective) and according to how well the selected sentence summarized the review (the traditional objective). Each sentence was annotated by 5 workers, where the sentences of the two algorithms appeared one next to the other but in random order. Table 7 summarizes the results,

[9]We used the implementation of https://pypi.org/project/sumy/

showing our method is superior in both aspects.[10]

	Helpfulness		Summarization	
	Mean	Std	Mean	Std
Helpful Sentence	3.41	1.11	3.34	1.05
Textrank	3.28	1.16	3.27	1.10

Table 7: Single Document Summarization - comparison to TextRank.

End-To-End Evaluation Our complete model resembles the task of Multi-Document Summarization (MDS) which ideally consumes the entire set of reviews related to a specific product and outputs a single summary or a single sentence in our case. In practice, MDS is applied to document sets of relatively small sizes, which significantly reduces the potential impact of our Similarity and Support sub-component. In order to put our evaluation in context of prior work, we evaluate our model with two minor modifications tailored for small review sets: we relax the similarity threshold to 0.75 and remove the minimal support constraint. We only consider the positive sentences in this evaluation, as the majority of the reviews are positive.

We use the dataset published in Bražinskas et al. (2020b) which covers 60 products from 4 different categories (Cloth, Electronics, Health Personal Care and Home Kitchen)[11] of which only 1 category is included in our own data. Each product has 8 reviews and 3 reference summaries written by humans. We evaluate our model in a straight forward manner by comparing the sentences selected by our model to sentence rankings provided by humans.

We ask expert annotators (one annotator per example) to read the reviews and rate each sentence from the reviews on a scale of 1 to 5. A score of 1 means that the sentence does not help to make a purchase decision or it does not reflect the overall theme of the reviews where a score of 5 means that the sentence is both helpful and aligns well with the common opinions expressed in the reviews. The mean score of the top sentence for each product is 4.31, which means that even for products with only 8 reviews it is common to find a sentence that is both helpful and supported by the reviews. We evaluate our model by averaging NDCG@K over all products for $K \in \{1, 10\}$. We compare the performance of our model with two baselines: ranking the sentences in a random order and from the longest to the shortest. Our method outperforms the baselines by a large margin, see Table 8.

	K=1	K=10
Our Model	0.87	0.94
From Longest to Shortest	0.60	0.68
Random	0.54	0.62

Table 8: Mean NDCG@K score.

For the sake of completeness we also report the common MDS evaluation metric, ROUGE (Lin, 2004), which does not fully suit our setting, as it is based on n-gram comparisons between the output and golden summaries written by humans, which are typically much longer than a single sentence. In Table 9 we compare the ROUGE scores of 3 sentence selection variants: our model, a random sentence and an Oracle, i.e., the sentence that maximizes the ROUGE-L score. We also report the results of Copycat (Bražinskas et al., 2020b),[12] a state-of-the-art review MDS model. We note that Copycat is not truly comparable to our model due to the significantly different summary length requirement (in this dataset an average sentence contains 74 characters while an average reference summary contains 293 characters). Note, however, that in terms of precision, which is what we aim for with such an extreme "summary", the RHS is almost as good as the Oracle and much better than Copycat.

	Random	RHS	Oracle	Copycat
Rouge-1 f	0.127	0.166	0.250	0.297
Rouge-1 p	0.329	0.420	0.440	0.247
Rouge-1 r	0.084	0.109	0.185	0.386
Rouge-2 f	0.014	0.028	0.054	0.055
Rouge-2 p	0.045	0.084	0.106	0.045
Rouge-2 r	0.008	0.018	0.039	0.073
Rouge-L f	0.094	0.120	0.177	0.192
Rouge-L p	0.251	0.309	0.320	0.160
Rouge-L r	0.061	0.078	0.129	0.249

Table 9: ROUGE scores.

Examples of RHS We pick two examples from Bražinskas et al. (2020a), depicted in Table 10, and use our model to extract a single sentence for each. Each of the examples consists of 8 reviews and a reference summary written by a

[10] The results are statistically significant using a 1-tail paired t-test, with a p-value of $1.05E-06$ for helpfulness and 0.005 for summarization.

[11] 4 of the products in this dataset are no longer available on amazon.com and we omitted them from the evaluation.

[12] Results are based on our own computation using https://pypi.org/project/py-rouge/

Reviews from Yelp!	
RHS	They have ridiculous price also bad customer services.
Human	Bennett Medical has poor customer service. Phone calls can take a long time to get answered and leaving voice mails tend to be fruitless. The products are overpriced and take a long time to be refilled. Using this medical supply company can be a hassle.
Summarizer	This medical supply is the worst medical supply company in the valley. The customer service is horrible, the staff is rude, the wait times are long, and the service reps are not helpful at all. Do not recommend this company to anyone.
Reviews from Amazon.com	
RHS	These thongs are fun, festive, flexible and surprisingly comfortable.
Human	These are a very comfortable and cute sandal. This thong sandal goes with a variety of outfits and the cushy sole allows for all day comfort. However, they do run a little small, sizing up provides a better fit. Overall, a reasonably priced shoe that will last for years to come.
Summarizer	These sandals are very cute and comfortable. They fit true to size and are very comfortable to wear. They look great with a variety of outfits and can be dressed up or down depending on the occasion.

Table 10: Helpful Sentences vs. Abstractive Summarization.

human.[13]

The first example deals with Yelp reviews and our model seems to do well even though it was only trained on product reviews from Amazon. While our extracted sentence is less elaborative compared to the human and abstractive summaries, it gives enough information to make a decision. Note also, that the abstractive summary does not refer to the high pricing. As for the second example, while not covering all aspects of the product, the helpful sentence is faithful to the reviews and aligns with the overall sentiment. The summarizer, on the other hand, contradicts the reviews with regarding to the sandals size.

Recall that these examples are constructed from 8 reviews only, while our model benefits considerably from large number of reviews, which is often the case for popular products. This is due to the greater sentence variety it can choose from and the fact that the support becomes more meaningful as more reviews are available. See Section E in the Appendix for additional examples and some statistics of our model outputs.

7 Conclusion

In this paper we address the challenge of summarizing product reviews with limited space, like when using a virtual assistant. We define a new notion that fits the needs of this setting, a *representative helpful sentence*, and propose a new task accordingly: given a set of product reviews, extract a sentence that is both helpful for a purchase decision and well supported by the opinions expressed in the reviews.

As a first step, we collect and annotate a new dataset of review sentences with their helpfulness scores, and make this dataset available to facilitate further research. Next, we develop an end-to-end model for surfacing representative helpful sentences. Our model combines several necessary components which are optimized for our goal. In order to get a feeling for the performance of our model, we compare our results to summarization tasks that are similar in nature, and show that our model performs better in the aspects we target.

8 Ethical Considerations

In this work, we make use of customer reviews published on Amazon.com. The reviews must comply with Amazon's Community Guidelines[14] which prohibit offensive, infringing, or illegal content. Amazon encourages anyone who suspects that content manipulation is taking place or that its Guidelines are being violated to notify Amazon. Amazon investigates concerns thoroughly and takes appropriate actions, including removal of reviews that violate these Guidelines, including reviews that contain hatred or intolerance for people on the basis of race, ethnicity, nationality, gender or gender identity, religion, sexual orientation, age, or disability. Among other things, Amazon has a broad license to use, reproduce, publish, and create derivative works from the customer reviews on Amazon.com. The authors of this paper are employees of Amazon and are authorized to use customer reviews in this work.

A small sample of annotated review sentences is released for research purposes according to the provided license.[15] Annotations were conducted by a service provider pursuant to a Service Agreement with Amazon. Under that Service Agreement, the service provider represents and warrants that it complies with all applicable laws, regulations, and ordinances when performing those services.

[13] We only show the summaries, the complete set of reviews are available in Bražinskas et al. (2020a).

[14] https://www.amazon.com/gp/help/customer/display.html?nodeId=GLHXEX85MENUE4XF

[15] https://cdla.dev/sharing-1-0/

References

Hana Almagrabi, Areej Malibari, and John McNaught. 2018. Corpus analysis and annotation for helpful sentences in product. *Computer and Information Science*, 11(2).

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169.

Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. 2006. Multi-document summarization of evaluative text. *Comput. Intell.*, 29:545–576.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer errorrates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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)*, pages 698–708.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *EMNLP*.

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*, pages 168–177.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, page 423–430, USA. Association for Computational Linguistics.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan T. McDonald. 2009. Sentiment summarization: Evaluating and learning user preferences. In *EACL*.

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

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, Prague, Czech Republic. Association for Computational Linguistics.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4251–4257. AAAI Press.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Susan M. Mudambi and David Schuff. 2010. Research note: What makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly*, 34(1):185–200.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Jahna Otterbacher. 2009. "helpfulness" in online communities: A measure of message quality. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 955–964, New York, NY, USA. Association for Computing Machinery.

Yue Pan and Jason Q. Zhang. 2011. Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, 87(4):598 – 612.

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.

Ori Shapira and Ran Levy. 2020. Massive multi-document summarization of product reviews with weak supervision.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Jiaxing Tan, Alexander Kotov, Rojiar Pir Mohammadiani, and Yumei Huo. 2017. Sentence retrieval with sentiment-specific topical anchoring for review summarization. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*.

Hongli Wang and Jiangtao Ren. 2018. A self-attentive hierarchical model for jointly improving text summarization and sentiment classification. In *Asian Conference on Machine Learning*, pages 630–645.

Wenting Xiong and Diane Litman. 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1985–1995, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 38–44, Beijing, China. Association for Computational Linguistics.

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*, pages 1113–1124. The COLING 2016 Organizing Committee.

Feng Zhu and Xiaoquan Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics.

A Annotation Task

Figure 3 displays a single annotation task as presented to the annotators.

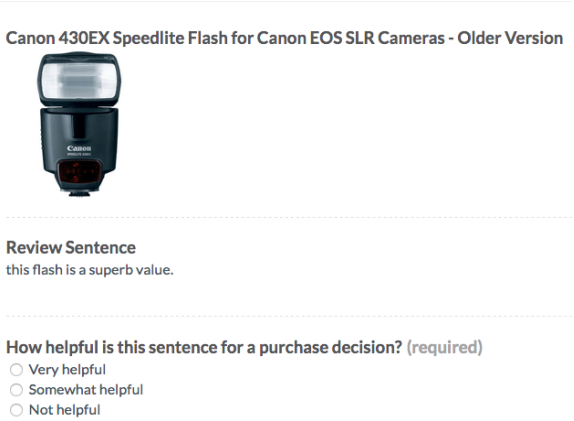

Figure 3: A single annotation as presented to the annotators.

Helpfulness Distribution within the Train Set Figure 4 depicts the helpfulness distribution within the train set.

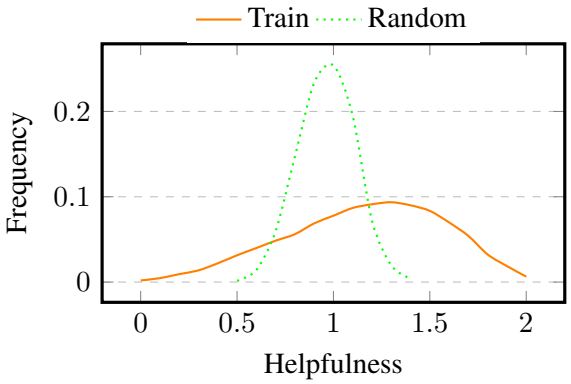

Figure 4: Helpfulness distribution among the train set. The green, dotted line represents the expected distribution if the annotators would have chosen an answer uniformly at random. The most frequent helpful score on the training set is 1.3.

B Annotation Analysis

Table 11 provides some statistics for the helpful and unhelpful subsets.

	# Helpful Votes	# Reviews	# Sentences	Mean Sentences
Helpful Set	≥ 50	101	5032	49.82
Unhelpful Set	0	3803	22030	5.79

Table 11: Contrasting Sentence Helpfulness Score with Review Helpfulness votes — Helpful and unhelpful sets statistics.

C Analysis of Helpful Sentences

Length Sometimes a sentence may be long but provide information that is not very helpful for customers. As an example for a long sentence with low helpfulness score, consider the sentence "As engineers, you must learn to handle the technical aspects, the social and political gamesmanship, the public reactions to our work and the daunting challenge of staying at pace with rapid developments in our fields" that was taken from a review about the book Mastery which deals with becoming a master in one's field. Indeed, this sentence is long and informative but it talks about engineers while the book itself is meant for everyone interested in mastery, not necessarily engineers.

Sentiment Consider the following sentence, demonstrating that sentiment is not always necessary for helpfulness: "It teaches parents to use twelve key strategies in everyday scenarios to help them in providing experiences to promote healthy connections in the brain." This sentence was deemed helpful by the annotators but does not express any sentiment, it merely states a fact about the product.

D Surfacing Representative Helpful Sentences

Preprocessing First, HTML markup is removed from each review using the BeautifulSoup[16] package and then the cleaned texts are split into sentences using the Spacy[17] package. Next, sentences with character length outside the range of $[30, 200]$ are removed. We chose these thresholds based on manual inspection under the assumption that extremely short sentences will not be very helpful for customers while extremely long sentences will result in a frustrating customer experience (especially in the scenario of voice interactions). In the movies domain, for example, a typical short sentence would state that *"This movie was really good."* or that *"It's a must see film."*, statements that do not contribute much on top of the star rating. In our dataset, long sentences are quite rare while short sentences, on the other hand, are more common and form 10% of the sentences.

Sentence Selection The Sentence Selection component is in charge of selecting a single sentence that is both helpful and well supported. We enforce a minimum support of 5, as we observed that such a limit increases the overall quality of the sentence and avoids surfacing esoteric opinions. After applying this threshold, we rank the remaining sentences according to the formula:

$$\text{support} \times \text{helpful}^{\alpha} \quad (1)$$

where α is a boosting parameter. To derive an appropriate value for α we conducted another annotation task in which annotators were asked again to score the helpfulness of the sentences presented to them. This time we consider all the sentences that are not dominated by any other sentence, i.e. we consider sentence s if and only if there is no sentence s' such that both $\text{helpful}(s') > \text{helpful}(s)$ and $\text{support}(s') > \text{support}(s)$, in other words, we asked to annotate all the sentences from the Pareto front with respect to helpfulness and support. Each sentence was joined with a prefix that quantifies its support as in *20 customers agreed that: has very good pic quality and extremely easy to use*. We optimized Formula 1 by minimizing the Kullback–Leibler divergence between the score distribution (softmax) from the annotators and the score distribution from the formula (softmax) and obtained $\alpha = 38.8$. While this value may seem enormous at a first glance we note that the helpfulness score obtained from our model for the sentences in the Pareto set tend to be very close to each other while their support may vary considerably. To put this number into proportion, consider two sentences with support and helpfulness $(20, 1.5)$ and $(10, 1.52)$ respectively, then the final score of the first sentence will only be slightly better than the final score of the second sentence (which is the expected behavior as its support is twice as large).

Interestingly, the experiment confirmed our hypothesis that customers perceive highly supported sentences as more helpful compared to the case when no support information is given.[18]

[16] https://pypi.org/project/beautifulsoup4/

[17] https://spacy.io/

[18] We experimented with another prefix, that only states *"one customer thought that"*.

E Model Output Statistics

As for end-to-end results on our 123 products, our model found a positive helpful sentence for 114 products, and a negative helpful sentence for 16 products. The low coverage for negative helpful sentences might be explained by our choice to concentrate on popular products (having more than 100 reviews) which are probably of better quality than random products.

In Table 12 we present the selected sentence, its helpfulness score and its top supported sentences along with their similarity scores for 3 products.

	It was easy to put together, is the perfect height, and very durable.	1.6
	It was easy to put together and is sturdy.	0.94
	Sturdy and easy to put together.	0.92
	Also, it was very easy to put together.	0.92
	It's sturdy and cleans easily.	0.91
	Pretty sturdy, too, and easy to handle.	0.91
	They fit well and produce good sound quality.	1.7
	fits great and has great sound	0.93
	The sound is pretty good as well.	0.93
	Great Pair works great with excellent sound	0.93
	They do have good sound, when I can keep them in.	0.91
	These sound great and fit very well.	0.9
	Quality good, price okay, sound output great.	1.7
	For the price point, this set delivers good sound.	0.92
	Very good sound for the cost!!	0.92
	It works and the sound quality is good.	0.9
	Good quality speakers that provide great sound.	0.9
	For me ... the sound was very good.	0.88

Table 12: The selected sentence, its helpfulness score and its top supported sentences along with their similarity scores for 3 products.