

The Tip of the Buyer: Extracting Product Tips from Reviews

SHARON HIRSCH, Ben-Gurion University of the Negev, Israel

SLAVA NOVGORODOV, Tel Aviv University, Israel

IDO GUY, Ben-Gurion University of the Negev, Israel

ALEXANDER NUS, eBay Research, Israel

Product reviews play a key role in e-commerce platforms. Studies show that many users read product reviews before a purchase and trust them to the same extent as personal recommendations. However, in many cases, the number of reviews per product is large and extracting useful information becomes a challenging task. Several websites have recently added an option to post *tips* – short, concise, practical, and self-contained pieces of advice about the products. These tips are complementary to the reviews and usually add a new non-trivial insight about the product, beyond its title, attributes, and description. Yet, most if not all major e-commerce platforms lack the notion of a tip as a first class citizen and customers typically express their advice through other means, such as reviews.

In this work, we propose an extractive method for tip generation from product reviews. We focus on five popular e-commerce domains whose reviews tend to contain useful non-trivial tips that are beneficial for potential customers. We formally define the task of tip extraction in e-commerce by providing the list of tip types, tip timing (before and/or after the purchase), and connection to the surrounding context sentences. To extract the tips, we propose a supervised approach and leverage a publicly-available dataset, annotated by human editors, containing 14,000 product reviews. To demonstrate the potential of our approach, we compare different tip generation methods and evaluate them both manually and over the labeled set. Our approach demonstrates particularly high performance for popular products in the Baby, Home Improvement and Sports & Outdoors domains, with precision of over 95% for the top 3 tips per product. In addition, we evaluate the performance of our methods on previously-unseen domains. Finally, we discuss the practical usage of our approach in real world applications. Concretely, we explain how tips generated from user reviews can be integrated in various use cases within e-commerce platforms and benefit both buyers and sellers.

Additional Key Words and Phrases: e-commerce, tips generation, product reviews, machine learning, deep learning

1 INTRODUCTION

The importance of product reviews for many e-commerce platforms has been proven empirically across different shopping domains [13, 18, 41, 68]. Recent studies have demonstrated that over 85% of the customers often read product reviews before making a purchase and trust them as much as personal recommendations [6]. Online shoppers read reviews for various reasons, such as seeking for other customers' opinions, looking to read about personal experiences, or obtaining buyers' point of view on product characteristics. In some cases, customers also read reviews to find *tips* - short, concise, practical and self-contained pieces of advice. Tips can provide complementary insights on top of the existing product information, such as title, attributes, and description. They can be useful both before the purchase, to learn more about the product, and after the purchase, when the

Authors' addresses: Sharon Hirsch, Ben-Gurion University of the Negev, Beer Sheva, Israel, shhirsch@ebay.com; Slava Novgorodov, Tel Aviv University, Tel Aviv, Israel, slavanov@post.tau.ac.il; Ido Guy, Ben-Gurion University of the Negev, Beer Sheva, Israel, idoguy@acm.org; Alexander Nus, eBay Research, Netanya, Israel, alnus@ebay.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

1533-5399/2022/7-ART \$15.00

<https://doi.org/10.1145/3547140>

product is already at hand. Each of these use cases holds its own value for e-commerce platforms: before the purchase, tips help customers make a more informed purchase decision, whereas after the purchase, tips can motivate customers to return to the site and increase their engagement.

The large number of reviews on e-commerce platforms, especially for popular products¹, makes the process of finding useful information challenging. In order to find relevant pieces of information, customers usually sort and filter the reviews by different parameters, such as date or review score, but eventually they often consume few of the reviews, and might therefore overlook many helpful pieces of advice. Tips are not typically enabled as first-class user-generated content type on major e-commerce websites. While a few small e-commerce websites allow tips as first-class user-generated content, this type of input is not widespread across the major e-commerce sites. Recently, some e-commerce websites added such functionality, but the most of the existing platforms are still missing this option. Therefore, customers have to provide their tips and advice (if they wish to) through other user-generated content options, such as reviews. In the travel domain, a few websites and applications (e.g., TripAdvisor, Yelp and Foursquare) enabled the tip functionality, but those were not prominent and did not demonstrate success or lasted for long [23].

In this work, we propose an automatic method for deriving such short and concise tips from customer reviews. We propose an extractive approach, where we are aiming to find several tips out of hundreds of review sentences per product. Extracting only few, yet informative and helpful sentences from a large number of reviews, can save a lot of effort to customers and can come in especially handy for mobile device users, who often seek for concise content. These tips may be of different types and each of these types is useful for different situation, some especially handy before the purchase (e.g., warnings), other can be applied right after the purchase (e.g., first time use) or at the long run (e.g., maintenance).

Tip extraction methods have been previously studied in other domains, especially travel [23, 62, 69]. A recent work has proposed an approach for extracting short practical and useful tips from developer answers written on the StackOverflow platform [61]. However, these tips are different in their applicability. For example, travel tips mainly focus on logistics, opening hours, discounts, or special attractions to notice. Tips for software developers usually include suggestions for coding best practices and specific libraries to use. In contrast, our tips focus on product aspects, such as usage, maintenance and workarounds. The techniques used in previous works that studied tip extraction in these domains included various methods, such as considering sentences that start with a verb in a base form (e.g., [62]), finding repeating patterns (templates) using regular expressions (e.g., [23]), and detecting “actionable clauses”, i.e., phrases that include action, target, method, time, and place (e.g., [51]). As part of our analysis, we consider all the aforementioned techniques and compare them to the methods introduced in this work.

In the e-commerce domain, closest in spirit to our research are the works by Li et al. [39, 40], which also studied tips generation for products. However, these works define a tip in a completely different way. While we extract short practical pieces of advice from the reviews, they merely consider the summary part of the review as a tip. While we require that the extracted sentence will add a new non-trivial information about the product on top of the existing information, the summary in many cases does not include such new information, but rather reflects an opinion overview (e.g., “*This is a great product for a great price*” and “*Not as good as I expected*”). Our strict definition makes tips quite rare in the reviews (as we later show, tip prevalence in review sentences is lower than 5%), while in their work, each product has at least one tip sentence (since each review has a summary). Finally, to the best of our knowledge, we are the first to define the various types of product tips and study their characteristics. As part of our study, we provide both analysis of tips “hidden” in reviews and an evaluation of methods for extracting them that attain high precision for popular products.

¹For instance, SENSO Bluetooth Headphones has over 36,000 reviews on Amazon.com

Our work uses a publicly available dataset of product reviews from one of the world's largest e-commerce platforms [26]. As the proposed approach is supervised, we extended the existing dataset with labels. First, we define a tip as a short, concise and self-contained piece of advice, in line with previous work in other domains [23, 62]. The data labeling is performed manually by human annotators via a dedicated tool designed for this task. We identify 10 main types of product tips that commonly appear in reviews and list them in the annotation tool. The list contain the following types: Alternative use, Complimentary product, First time use, Maintenance, Population segment, Size, Warning, Workaround, and Other. Additionally, the annotator has to select the tip's timing (i.e., if the tip is useful before and/or after the purchase), and connection to the surrounding context sentences (i.e., whether they could or need to be used as part of the tip). The additional labeled set contains 14,000 product reviews and is released for public use. Specifically, we focus on five popular e-commerce domains: Baby, Home Improvement, Musical Instruments, Sports & Outdoors, and Toys. These domains' product reviews tend to contain useful non-trivial tips that are beneficial for customers. The potential customers of these domains may often look for advice, such as recommended usage of new-born products for fresh parents in the Baby domain or "do it yourself" tips and tricks in Home Improvement.

As mentioned above, we apply a supervised approach and experiment with a wide range of well-known classifiers, from baselines such as Naïve Bayes and basic LSTM [29], to state-of-the-art methods, such as BERT [16]. For the BERT classifier, we also experimented with a multi-task learning approach [8]. In addition, we use a baseline method of taking sentences starting with a verb (e.g., "*Plug in the AUX cable first before turning it on*" and "*Make sure there is nothing behind the arm you're pulling with*"). As previously mentioned, this method has been used in key tip studies in domains other than e-commerce, either as the main approach [62] or as a baseline [23]. However, as our study shows, in e-commerce, only 5% of the tips start with a verb, hence this approach is not practically useful. Note that since the presentation area on product pages is usually very limited, especially on mobile devices [47], we aim at extracting a small number of high-quality tips per product.

For our evaluation, we use two main methods. First, we perform a standard multiple train/test evaluation via random re-sampling and cross-validation on the collected labeled data and report the precision/recall for each of the classes. We also experimented with a cross-domain approach, training the model on one domain and testing on another, in order to examine how well the model can generalize.

Second, in order to simulate the practical use-case of extracting the tips from a large set of reviews, we run our method on previously-unseen products from these domains. In addition, we test our model on products from previously-unseen domains and from another e-commerce platform. For this type of evaluation, we consider the top-k tip sentences identified by the model (ordered by classification score), and ask our annotators to manually assess the extracted tips. The second evaluation method allows us to estimate the quality of our algorithm in a real life scenario, and gain initial insights about the number of reviews needed to produce high-quality tips for a product. The results of the second evaluation demonstrated high precision, especially for the Baby, Home Improvement and Sports & Outdoors domains, with over 90% precision for the top 5 tips.

Also, in order to demonstrate the robustness of our method and the ability of the proposed model to generalize, we evaluate the model on datasets from additional domains from the same e-commerce platform, including Automotive, Cellphones, Fashion, Electronics, and Health, as well as domains from another e-commerce platform, including Shoes and Watches. Our method achieved high precision particularly for Automotive, Electronics, Health, and Watches domains, with over 82% precision for the top 5 tips.

Finally, we discuss practical usage of our approach in real world applications. Concretely, we explain how tips generated from user reviews can be integrated in various use cases within e-commerce platforms and benefit both buyers and sellers. As part of the discussion, we provide a visionary user interface prototype for tip integration within a product page. We also conduct a small-scale user study to assess the effect of tip presentation within the product page on potential buyers' experience.

Our contributions can be summarized as follows:

Table 1. Characteristics of the original datasets.

	Baby				Home Improvement				Musical Instruments				Sports & Outdoors				Toys			
	Avg	Std	Median	Max	Avg	Std	Median	Max	Avg	Std	Median	Max	Avg	Std	Median	Max	Avg	Std	Median	Max
Reviews per product	22.81	36.76	11	780	13.16	16.22	8	504	11.40	12.93	8	163	16.14	25.67	9	1042	14.06	15.85	9	309
Sentences per review	6.14	5.62	5	213	6.89	7.21	5	198	5.81	5.93	4	116	5.71	5.94	4	283	6.40	6.22	5	222
Words per sentence	16.19	10.70	14	425	16.08	10.54	14	573	15.66	10.99	14	230	15.39	10.53	13	829	15.74	10.25	14	586
Number of products	7,050				10,217				900				18,357				11,924			
Number of reviews	160,792				134,476				10,261				296,337				167,597			

- To the best of our knowledge, we are the first to introduce and study the tip extraction task in electronic commerce.
- We provide an extensive analysis of tips, their types, and distribution in reviews across different e-commerce domains.
- We present several supervised methods for detecting the tips and perform an extensive evaluation.
- We perform an extensive evaluation of the proposed methods on products from various domains and platforms to demonstrate the efficiency and robustness of these methods.
- We show how our proposed method can be integrated in a real world e-commerce platform and benefit both buyers and sellers.

The rest of the paper is organized as follows. Section 2 reviews related work. In Section 3, we describe the existing datasets and elaborate on the labeled data acquisition process. Section 4 describes the proposed tip extraction methods. In Section 5, we present the evaluation results of the suggested methods. In Section 6 we discuss how these methods can be applied in a real world e-commerce platform. Finally, we conclude in Section 7.

2 RELATED WORK

Previous work has shown that online reviews from customers have a strong effect on other customers' purchase decision process in e-commerce [13, 18, 41, 68]. The number of reviews in e-commerce sites has grown significantly in the recent years. The sharp increase in the number and variety of reviews brings new challenges to the table, such as review quality estimation [9, 34] and fabricated review detection [1, 31]. One of the primary challenges related to reviews is information overload. A number of studies have shown that information overload, due to the immense number of reviews, leads to an increase in the time required to make a decision and degrades decision quality [54, 57]. There are several proposed approaches to deal with this challenge. Some focus on selecting a compact and representative subset of reviews (e.g., [22, 36, 37, 45]), while others apply review summarization techniques and generate an aggregate statistics of negative and positive feedback about different product features (e.g., [14, 30, 53, 66]). Another related research direction deals with ranking the reviews according to different properties (e.g., helpfulness votes) [3, 58]. Extracting aspects and ranking the reviews accordingly has recently been studied by Huang et al. [32]. They analyze users' aspect preferences from reviews and define a review helpfulness score at the aspect level. Then, a recommendation algorithm is applied to provide each user with the top-k most useful reviews based on their helpfulness score. Finally, recent work focused on personalizing product descriptions and generating them from product reviews [19, 47, 48]. In contrast to most of the approaches mentioned above, our method's building block is a review sentence rather than the entire review. Previous approaches that worked on a sentence level fundamentally differ from our approach. On the one hand, we do not aim to cover all possible aspects contained in the reviews. On the other hand, the summaries and/or the descriptions generated by the above-mentioned methods do not necessarily contain any tips.

Extraction of tips is an established research direction that attracted much interest in the recent years. A few previous studies examined the identification of tips in domains other than e-commerce. A few of them have focused on extracting tips from sources such as question-answering websites. For example, a recent work has proposed an approach for extracting software development tips from answers on Q&A platforms such as StackOverflow by [61]. The work by Weber et al. [62] aimed at extracting tips from Yahoo Answers to address specific search queries. Similar to this work, they defined a tip as a “short, concrete and self-contained bits of non-obvious advice”. Their proposed extraction mechanism mainly used the question-answer structure. Specifically, they considered only “how-to” questions such as “how to”, “how do I”, or “how can I” and collected short answers that start with a verb. Afterwards, they performed a user study and annotated tip candidates manually. The annotated set was then used for training a binary classifier. The final tips were always of the form “X:Y”, where X is the tip’s goal, and Y is the suggestion. This approach is not applicable in our setting, since we are working with product reviews rather than question-answer pairs and search queries. Nonetheless, we consider sentences that start with a verb as a baseline to all other methods.

Li et al. [39, 40] study abstractive tip generation for products. These works are close to our research by their title, however, while the task seems to resemble our work, there is a major difference in the definition of the tip. Generally, these works used two datasets from the e-commerce and restaurants domains. The first dataset was from Amazon, where the extracted tips originated from the “summary” part of the review, for example “*One of our favorite games!*” or “*My son really loves this simple toy!*”, while we use the review content to extract tips. The second dataset was from the Yelp Challenge and included restaurant tips and reviews, for example “*Love their soup!*” or “*Pretty good local service!*”. These “tips” are subjective, far from our definition of a tip, and do not provide much non-trivial information or insights, but rather reflect an opinion summary. Another difference from their work is that in this study, we define and explore various tip characteristics, such as tip types, tip timing, and connection to the surrounding context.

The travel domain has been extensively researched in the recent years (e.g., [10–12]). For example, Chen et al. [10] proposed a novel technique for group itinerary recommendation. Their method was designed to schedule a list of interesting places to visit, while taking into account time constraints and other preferences of each of the group members. The travel domain is also the most popular domain for tip-related research, mostly because there are many available datasets (e.g., forums, blogs, questions and answers) and since the user is typically visiting an unfamiliar environment, where advice from knowledgeable individuals can be valuable. In contrast to tips in e-commerce, which are mostly about different usages of the products, travel tips focus mainly on logistics, opening hours, discounts, special attractions to be noticed, and so forth. Closest to our work is the research by Guy et al. [23] and by Zhu et al. [69]. The work by Guy et al. relies on 150 human-generated templates for travel tips. Examples of such templates are “make sure to *”, “check the * for” and “the * is closed on Mondays”, where the asterisk can represent any word. The work by Zhu et al. extends the work by Guy et al. and introduces an unsupervised approach that solves a similar task without relying on training data. The key difference from our work is the applicability of the proposed methods to the e-commerce domain. In this paper, we define the tip extraction task along with e-commerce specific tip types and their context, while in their works they focus on travel-specific language. Moreover, a template-based approach is not applicable in our setting, since we did not find any dominant repetitive n-gram patterns in our annotated tips.

Another closely related field of research focuses on detecting text units that include pieces of advice. Wicaksono and Myaeng [64] proposed to use conditional random fields, to extract advice sentences from travel forum entries. Some of the studies used linguistic features to detect tips. Ryu et al. [51] proposed a method to detect “actionable clauses” in how-to instructions using syntactic and modal characteristics. For example, “*Clean the bowl completely with mineral spirits on a rag*” can be converted into {ACTION: clean, TARGET: bowl, INSTRUMENT: mineral spirits on a rag}. However, in our work, the extracted tips are not always actionable or include specific instructions for the customer. A typical example for a product tip may look like “*The slipcovers come off easily to be machine-washed*”

Table 2. Characteristics of labeled tips across the five domains.

	Baby	Home Improvement	Musical Instruments	Sports & Outdoors	Toys
# of products	2,612	2,711	736	2,722	2,736
# of reviews	2,800	2,800	2,800	2,800	2,800
# of sentences	17,560	19,436	15,460	15,957	16,758
# of tips (% of sentences)	954 (5.43%)	880 (4.53%)	537 (3.47%)	805 (4.71%)	670 (4.00%)
Avg (median) words per tip	21.33 (19)	21.11 (19)	21.98 (19)	20.42 (19)	20.38 (18)
Before purchase	49.37%	35.34%	39.66%	39.13%	47.01%
After purchase	21.07%	37.27%	27.75%	27.83%	23.88%
Both	29.56%	27.39%	32.59%	33.04%	29.10%
Standalone	81.97%	79.20%	80.07%	84.10%	84.48%
Extend before tip	9.12%	9.31%	9.57%	11.25%	9.55%
Extend after tip	17.92%	16.20%	15.16%	19.43%	15.97%
Most common types	Warning (38.05%)	Usage (40.23%)	Usage (40.41%)	Warning (29.94%)	Warning (34.63%)
	Usage (23.69%)	Warning (29.55%)	Warning (30.17%)	Usage (29.81%)	Usage (23.88%)
	Size (8.07%)	Workaround (7.05%)	Workaround (8.38%)	Size (9.81%)	Population segment (13.13%)

or “*This toy contains small parts and is not recommended for children under 3*”. Wicaksono et al. [63] focused on finding advice sentences in travel blogs. They also proposed several linguistic features, mostly defined by hand-crafted rules that were looking for the appearance of terms such as “I suggest”, “I strongly recommend”, or “advice”, with an associated proper noun, representing a travel entity, such as a hotel. Our approach applies a preliminary rule-based step, to filter out sentences with very low likelihood of being tips. However, rules do not suffice in our case due to the scarcity of repetitive patterns in e-commerce tips. We therefore propose a supervised model as our main method for tip extraction.

Finally, our previous work [28] addresses the problem of tip generation from review sentences. The current version provides a diverse set of experiments along with extensive evaluation methods on additional datasets from a variety of e-commerce domains and platforms. Specifically, we evaluate another state-of-the-art method, Multi-task learning with BERT, and experiment with different tip types as separate tasks. Additionally, we evaluate how our proposed approach can generalize without adding much labeled data when facing new unseen domains, which demonstrates promising results. This study is especially important for large e-commerce platforms since they support thousands of different categories and acquiring training data for each new domain is costly. Finally, in this paper we suggest practical applications of the proposed method in these real world e-commerce platforms, and discuss how the generated tips can benefit both buyers and sellers.

3 DATASETS AND CHARACTERISTICS

In this section, we describe the datasets used for our analysis and experimental evaluation, their characteristics, and the annotation process we used in order to produce labeled data.

3.1 Datasets

Our research was conducted over five publicly available product datasets [26] from five e-commerce domains: Baby (baby clothing and supplementary products), Home Improvement (tools for home improvement), Musical Instruments (musical instruments, parts, and related accessories), Sports & Outdoors (equipment for sports and outdoor activities), and Toys (children’s toys and games). The datasets contain, per each product, its metadata (title, image, etc.) and all its associated user reviews. Table 1 depicts the main characteristics of the five datasets, including the number of total products and reviews, length of reviews in sentences, etc. The largest dataset is

Sports & Outdoors containing 18,357 products with nearly 300K reviews in total, while the smallest is Musical Instruments, with 900 products and a little over 10K reviews in total. The median number of reviews per product ranges from 8 to 11, while the median number of sentences per review is between 4 and 5. The average number of sentences per review is around 6, varying from 5.71 in Sports & Outdoors domain to 6.89 in Home Improvement domain. To demonstrate robustness of our methods and the ability of the proposed models to generalize, we use datasets from additional domains, including Automotive, Cellphones, Electronics, Fashion, and Health. For evaluation, we first consider all products above the 90th percentile according to their number of reviews in each of the five domains. Then, we randomly sample 50 products from each domain and select all review sentences of these products. Then, we apply our method on the sentences and present the top sentences (ordered by classification score) to be evaluated as a tip or not.

3.2 Tip Definition

We define a *tip* as a short, concise, practical, and self-contained piece of advice. In general, tips can be useful both before and after the purchase. Before-purchase tips are useful to learn more about the product, and after-purchase tips are helpful when the product is already in hand. Both of these use cases are important for e-commerce platforms: tips before the purchase help with the purchase decision by providing more information. Useful tips after purchases can increase customer satisfaction and motivate return to the site for additional shopping.

Despite the straightforward definition of a tip, there are some borderline cases that should be discussed. First, many review sentences may look like a tip, while they are very subjective and contain a personal experience (e.g., *“In addition, we live in a colder climate and do not heat the house above 62 degrees at night, so combined with a little heater, this sleep sack does the trick.”*) Such sentences are not considered as tips. Another type of a borderline case not considered as tip is a descriptive sentence (e.g., *“Highly recommend it for outside purposes especially in the cold weather.”* and *“The cards are high gloss with full color pictures on them.”*). Other non-tip sentences are obvious or trivial remarks (e.g., *“Just make sure you have a bunch of batteries to get started.”*) or those that repeat details already provided as part of the product description, without adding any new information (e.g., *“Perfect for tuning electric guitars.”* for a BROTOU Guitar Tuner that includes the following sentence in its description: *“Fits most electric guitars”*).

3.3 Data Annotation


Labeling for training and evaluation in this work was performed by in-house annotators, after three hours of training and qualification tests. The pool included a total of 10 annotators, who were assigned tasks from each of the five domains randomly.² Unless otherwise stated, each evaluation was performed by a single annotator. Labeling was performed using a dedicated tool developed for this task. The tool’s user interface is depicted in Figure 1. As shown in the figure, each review was split into sentences. For each sentence, the annotator was asked to select whether it is a tip (as defined in Section 3.2) or not. For each selected tip, the annotator was asked to select the tip type (out of 10 pre-defined types, presented at the first column of Table 3). Afterwards, annotators had to indicate if the sentence is a standalone tip or needs additional context. Moreover, annotators selected if this tip is useful before the purchase, after the purchase, or both. They could also (optionally) choose the previous and/or next sentence as useful information for extending the selected tip, regardless if it was marked as a standalone or not (see “extend tip” checkbox in the second column in the annotation tool screenshot). To measure the agreement between the annotators, we asked four of them to annotate the same 100 review sentences as tips or not. The Fleiss’ Kappa [20] among them was 0.815, indicating a high agreement level. All data annotated using the tool will be publicly released as an extension to the original public dataset.³

²Annotators were granted monetary compensation for their work.

³The dataset is available at http://proj.ise.bgu.ac.il/public/gen_tips.zip

R#B000TTX75W
Category: Sports_and_Outdoors

Tacstar Sidesaddle Fits Remington 870



Please mark all tips:

Tip? ☐ Extend tip

☐ I love the fact is comes with the allen wrenches too!

☐ Could use some lock-tite for the bolts going through the gun.

☐ ☐ Back plate is made from aluminum and the holder is plastic.

☒ If you have a long pump handle, this will not work.

☐ ☐ Holds the shells really well.

If you have a long pump handle, this will not work.

Tip type:

Standalone tip? ☐ Yes ☐ No

When useful? ☐ Before purchase ☐ After purchase ☐ Both

Comment:

Submit

Fig. 1. Annotation interface.

Table 3. Types of tips, their distribution (portion of all sentences marked as tips), and examples across all domains.

Warning	32.71%	The metal mounting clips scratched the edges of my trunk lid.
Usage	31.12%	Best used when replacing strings, so you can apply while they are off.
Workaround	6.66%	I needed a 5/8 female to 3/8 male adapter to get my mic to mount.
Complementary product	5.54%	You will need to buy fasteners for it, since the box only contains the vice.
Size	5.49%	But definitely order at least one size bigger than you wear.
Maintenance	4.60%	The slipcovers come off easily to be machine-washed.
Population segment	4.24%	Recommend for a 4-5 year old that likes cars and trucks.
First time use	3.98%	The wheels do need to be pumped with a bike pump prior to use.
Alternative use	2.99%	My baby doesn't need it anymore so now I use it as my neck pillow.
Other	2.68%	Her hair is much brighter blue than it appears in the photo.

3.4 Tip Characteristics

We sampled uniformly at random 14,000 product reviews across the five domains (2,800 per domain) from the original datasets (Table 1). The annotators labeled these 14,000 reviews, which included 85,171 sentences in total. Table 2 depicts the full statistics of the annotated dataset, including the number of labeled sentences and the collected tips across the five domains, along with the most common tip types. Overall, 3,846 sentences were annotated as tips, accounting for only 4.52% of all labeled sentences. This is a substantially lower percentage than the 23.3% reported for reviews of tourist attractions 3 depicts the distribution of the 10 different tip types in our labeled tips. The most popular tip types were ‘Warning’ and ‘Usage’, accounting each for slightly over 30% of all tips. The third most popular type was ‘Workaround’, followed by ‘Complementary product’ and ‘Size’. The least popular tip type was ‘Other’, while about half of these tips related to differences between the actual product received and the seller provided information (product image, title, or description; see example in Table 3). As depicted at the bottom of Table 2, some variance can be observed for the distribution of top tip types across the

Table 4. Tip timing (before and/or after purchase) distribution by type.

Type	Before purchase	After Purchase	Both
Warning	76.71%	13.04%	10.25%
Usage	17.88%	45.53%	36.59%
Workaround	3.91%	54.30%	41.80%
Complementary product	38.97%	11.74%	49.30%
Size	62.56%	1.42%	36.02%
Maintenance	3.95%	45.20%	50.85%
Population segment	78.53%	1.23%	20.25%
First time use	22.22%	49.67%	28.10%
Alternative use	6.09%	20.87%	73.04%
Other	43.69%	3.88%	52.43%

five domains. ‘Usage’ and ‘Warning’ are at the top of the list in each of the five domains, with ‘Usage’ the most common for Home Improvement and Musical Instruments and ‘Warning’ for Baby, Sports & Outdoors, and Toys.

Another interesting characteristic is the connection to the surrounding context sentences. Most of the tips (81.96%) were standalone, while only 18.04% were labeled as non-standalone sentences. As can be seen in Table 2, these results are rather consistent across the five domains. Overall, 26.91% of the tips could be extended to the adjacent sentence, with about two thirds of these to the succeeding sentence and a third to the preceding sentence. For example, for the standalone tip sentence: “*One note of caution, this is a very heavy router because it is a large plunge router*”, the succeeding sentence was marked as an extension: “*I mounted it on a Rockler X-Large router plate which is 1/4 inch thick aluminum, but it has a very slight bow in the middle.*” For the non-standalone tip sentence: “*You need a pipe cleaner to really get it*”, the preceding sentence was annotated as an extension: “*My one complaint is that there is an area that is hard to clean in the 4 piece nipple apparatus.*”. Overall, however, the low portions of non-standalone tips indicates that our choice to focus on single-sentence tips covers the majority of the cases. We leave the expansion to multi-sentence tips for future work.

The number of tips marked as being useful before the purchase was somewhat higher than those marked as useful after the purchase: 42.25% versus 27.61%, respectively. The rest, nearly a third (30.14%), were annotated as useful both before and after the purchase. These portions varied substantially across the different tip types, as depicted in Table 4. While ‘Population segment’, ‘Warning’, and ‘Size’ are typically useful before the purchase, ‘Workaround’, ‘First time use’, ‘Usage’, and ‘Maintenance’ tips are more often useful after the purchase. ‘Alternative use’ tips are prominently useful both before and after the purchase. We conjecture that alternate-use tips can both influence the purchase decision, as they reveal additional functionalities, and are also handy when the product is in possession, extending its potential use. The substantial differences across tip types are also reflected in differences across the domains, as can be seen in Table 2. For instance, domains with higher portion of ‘Warning’ tips, exhibit higher portions of tips that are useful before the purchase.

Tip Analysis. As a first step after obtaining the labeled data, we analyzed additional tip features. Table 5 presents the portion of tips according to different characteristics of the sentence in question, the originating review, and its authoring reviewer. It can be seen that longer sentences have higher likelihood of being tips: of the sentences consisting of 30 words or more, 8.6% were marked as tips, accounting for 23.1% of the tips in our dataset. At the other extreme, only 1.9% of the sentences composed of 6 to 9 words were labeled as tips. Inspecting review characteristics, it can be seen that sentences that originate from short (1-2 sentences) and especially long (over 15 sentences) reviews have somewhat lower likelihood of being tips. As can also be observed from the table, sentences that originate from reviews with two or more ‘helpful’ votes were more likely to be considered as tips,

Table 5. Tip characteristics (binned) within the complete dataset (all five domains). ‘%’ and ‘%T’ denote, per bin, the portion of tips out of all tips and the portion of sentences marked as tips, respectively.

Sentence Length (Number of Words)							
0	6-9	10-13	14-17	18-21	22-25	26-29	30+
%	6.6	12.4	16.1	16.8	13.9	11.1	23.1
%T	1.9	3.0	4.3	5.5	6.4	7.5	8.6
Review Length (Number of Sentences)							
	1-2	3	4	5-6	7-9	10-15	16+
%	5.8	8.8	10.6	17.6	17.6	20.0	19.8
%T	4.5	4.3	4.9	4.9	4.6	4.7	3.8
Tip Position within Review							
	First	Middle				Last	
%	10.3	74.6				15.1	
%T	2.9	5.0				4.2	
Review's Number of Helpful Votes							
	0	1				2+	
%	46.2	16.1				37.7	
%T	4.2	4.4				5.1	
Reviewer's Number of Past Reviews							
	0-19	20-39	40-79	80-129	130-199	200+	
%	13.3	21.0	24.1	12.7	8.6	20.3	
%T	4.8	4.8	4.5	4.2	4.4	4.1	
Reviewer's Portion of Past Reviews with Helpful Votes							
	0-20	20-30	30-40	40-50	50-60	60-100	
%	9.8	15.5	18.2	17.0	16.4	23.0	
%T	3.9	4.3	4.7	4.7	4.8	4.5	

whereas opening sentences (positioned first) and those originating from reviews with no ‘helpful’ votes had lower likelihood to be tips. Finally, we inspected characteristics of the reviewer who wrote the originating review. Interestingly, reviewers with many reviews on the site (130 and more) tend to include fewer tip sentences in their review. It could be that such “heavy” reviewers focus on other aspects in their reviews, such as personal experiences and opinions. Additionally, reviewers who had especially lower portion of past reviews with at least one helpful vote, produced a lower portion of tips, as can be seen in the last section of Table 5. We note that the results for each of the five domains demonstrated similar trends to those shown in Table 5.

Tip vs. Non-tip Language. We also set out to examine the most prominent language differences between the two classes of sentences – tips versus non-tips. To this end, we used Kullback-Leibler (KL) divergence, which is a non-symmetric distance measure between two given distributions [4, 7]. Specifically, we calculated the terms that contribute the most to the KL divergence between the language model of the tip sentences versus the language model of the non-tip sentences and vice versa [24].

Table 6 presents the most distinctive unigrams and bigrams. It is noticeable that the most characterizing unigram of tip sentences compared to non-tip sentences is the second-person pronoun *you*, while the first-person pronoun *i* tops the non-tip list. This indicates that tip sentences are usually phrased in a second-person language rather than first. The unigram *my*, which also reflects a first-person language is also high on the non-tip list, as

Table 6. Most distinctive unigrams and bigrams for tips vs. non-tip sentences.

Unigrams		Bigrams	
Tips	Non-tips	Tips	Non-tips
you	i	if you	i have
the	this	on the	the price
to	my	you have	i am
if	price	need to	i was
your	love	make sure	a great
or	our	you can	my son
off	was	if your	they are
on	had	is that	it was
use	his	sure you	a very
need	we	when you	i had
can	great	use the	in my
be	i'm	of the	this one
suggest	he	will need	i love
not	nice	have to	i bought
head	bought	that you	for my
put	like	you use	i got
note	very	you need	bought this
it	and	to the	and i
make	quality	be careful	i will

well as the plural we, while your is on the tip list. Other unigrams on the tip list include different prepositions and verbs, such as use and need, as well as the explicit suggest and note. The non-tip list includes was and had, which are often used to describe a past experience, the verbs like and love, which reflect subjective opinions, as well as adjectives and adverbs that reflect positive impressions, such as great, very, and nice and the noun quality, which is often associated with the reviewer's opinion on a certain feature of the product.

Inspecting the bigram lists, the tip list includes several expressions in second-person languages, such as if you, you can, you have, you need and you use and expressions typical to advice giving, such as make sure, will need, be careful. The non-tip bigram list, on the other hand, includes many first-person expressions, such as i have, and i, i am, i got, and i was, in addition to expressions that reflect personal experiences, such as my son or i bought. It also includes the price, referring to a listing-specific characteristic, which may change from one seller to another, and phrases that reflect subjective feelings, such as a great, a very, and i love.

4 TIP EXTRACTION

In this section, we describe the key components of our tip generation method. Given the set of all user reviews for a product, we go through the following steps to extract the tip sentences. We first apply rule-based filtering crafted based on analysis of the datasets. Following, we use a supervised approach that learns to identify tip sentences. To this end, we experiment with different types of classifiers, including state-of-the-art methods in language modeling, and compare their performance on a labeled set.

4.1 Rule-based Filtering

Since the data is very skewed, before applying any advanced classification approaches, we look for methods that would easily filter out non-tip sentences. After performing an analysis of basic features (e.g., sentence length) and words (KL) we could not observe simple rules that can massively be used for filtering, as done for the task of description generation [47]. This may indicate that our extraction task is more complex. Nevertheless, we

identify a few rules that could save up to 39% of the labeling task. The proposed rules decrease the total number of sentences from 85,171 to 51,929 and increase the total percentage of tips from 4.52% (as reported in Section 3.4) to 5.89%. To derive many of these rules, we observed the top KL n -grams for $n \in \{1, 2, 3\}$ and considered those that hardly appear in tips, i.e., in fewer than 5 sentences in our training set, but do appear frequently in non-tip sentences. Our rule list includes the following:

(1) **Short:** sentences of 5 words or fewer generally contain little information and rarely reflect any useful piece of advice. For example, “*Recommended*”, “*Very good quality*”, “*Useful*”, or “*Will buy again*” were among the most common short review sentences in our datasets. Analogously, in the travel domain it has been previously demonstrated that short sentences cannot serve as high-quality tips [23]. Overall, short sentences accounted for 11.2% of all review sentences across all domains in our dataset.

(2) **Enthusiastic:** some reviewers tend to describe the product using strong-sentiment positive adjectives, such as ‘wonderful’, ‘adorable’, ‘amazing’, ‘fantastic’ and verbs such as ‘love’ and ‘like’. Examples include “*I love this product and recommend it for everyone*” and “*Amazing quality, very useful for outdoor activities*”. Overall, 18.6% of the review sentences matched this filtering criterion.

(3) **Listing-specific:** review sentences that focus on listing-specific aspects, which may vary across different sellers of the product, such as price, shipping and return policy, warranty, product rating, and similar. Tokens used for filtering included ‘price’, ‘money’, ‘cheap’, ‘expensive’, ‘shipping’, ‘return’, ‘warranty’ and also the dollar sign ‘\$’. Examples of such sentences include “*Very useful product for just 20\$*”, “*Not very cheap but worth its money*”, and “*The shipping was almost as much as the panel itself*”. Overall, 14.7% of the review sentences matched this filtering criterion.

(4) **Personal:** sentences with a first-person pronoun, such as ‘i’m’, ‘i’ll’ and ‘i’ve’. As demonstrated in the previous section, such pronouns hardly ever occur on a product tip. Overall, 8.4% of the review sentences matched this filtering criterion.

We do not automatically filter out sentences from reviews with low ratings and do not run any sentiment analysis to filter out negative sentences, since these may hide useful tips, such as warnings, workarounds, or alternate use. For example, the sentence “*Tried using it lighting with a match which worked, but the off switch would not shut the flow of fuel off completely*” is a workaround tip for a Mini Jet Pencil Lighter and “*They fog up almost as quickly as my old Speedo goggles (several years old) which is after about one lap of the pool*” is a warning tip for a Speedo Baja Swim Goggle. Both tips were extracted from negative (one and two stars) reviews.

Our rules are designed to filter out sentences that are very likely not to contain a tip, hence we prefer to apply only high-precision rules. As already mentioned, our rules filtered out 39% of the review sentences, leading to a tip portion of 5.89% in the remaining set.⁴

4.2 Automatic Classification

After applying the initial rule-based filtering, we explore a supervised approach, by training a classifier to predict whether a product review sentence contains a tip. We use the labeled dataset described in Table 2 and experiment with various classifiers:

Naïve Bayes. We examine a common model for text classification - Naïve Bayes [50]. Our features include textual features, specifically the unigrams, bigrams, and trigrams of the review sentence. We also experiment with a variant that includes additional features, based on the characteristics described in Table 5.

LSTM. A recurrent neural network based on a long short-term memory (LSTM) [29] architecture, with Global Vectors for word representation (GloVe) [49] pre-trained on the Wikipedia 2014 and Gigaoword 5 corpora.

LSTM with Attention. The attention mechanism enables the network to focus on relevant parts of the input [67]. The overall architecture of the “attention network” consists of two components: an LSTM-based word

⁴The portions of all four rules do not sum up to the total portion of filtered sentences, since some sentences match more than one rule.

sequence encoder and a word-level attention layer. Given a review sentence split by words, we first embed each word using pre-trained GloVe embeddings, as previously described, and then use the LSTM network to produce the hidden states. The attention mechanism is often used to put more focus on certain words in the review sentence. For example, in the sentence “*Make sure to switch off the guitar*” the words “*make sure*” receive higher weight, and in the sentence “*Just be careful when opening the hood*”, the higher weight is given to “*just be careful*”. We feed the word annotations through a single-layer perceptron network to receive a latent representation. Then, we calculate the similarity of the latent representation with a word-level context vector, normalized by a softmax function, to produce the word’s importance weight. We then construct the sentence vector as a weighted sum of the word annotations based on each word’s weight.

For both LSTM methods, we performed hyper-parameter tuning, which included the batch size, number of epochs, learning rate, and the number of hidden units in the layers.

FastText. A library for learning word embeddings and text classification created by Facebook’s AI Research called FastText [5]. Each word is represented as a bag of character n-grams, and the final word embedding is the sum of character n-grams. This is useful for generalizing words with similar roots that appear in different forms (e.g., build and building). Hyper-parameter tuning was performed on the n-grams length, learning rate, and number of epochs.

BERT. A state-of-the-art technique for NLP pre-training called Bidirectional Encoder Representations from Transformers (BERT) [16]. This is a deep bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. In contrast to context-free models, such as word2vec [44] or GloVe [49], which generate a single word embedding representation for each word in the vocabulary, BERT generates a representation of each word based on the other words in the sentence. BERT is useful for extracting high-quality language features from text data. In addition, it is useful for fine-tuning a model on a specific task, as we did in our experiments. We used the pre-trained ‘BERT-Base, Uncased’⁵ model to train a binary classifier for our task. Our hyper-parameter tuning included the batch size, number of epochs, and learning rate.

For all deep learning methods, both LSTM methods and BERT, we used Adam [35] optimizer.

5 EVALUATION

In the following section, we present two evaluation methods for the extraction models. The first method is a standard train/test evaluation, which included both a 50-time repeated evaluation, in which the training and test sets are randomly re-sampled from the labeled data [65] and a 5-fold cross-validation. The second method simulates the practical scenario of extracting tips from a large set of reviews of previously-unseen products. We use the best performing model from the first evaluation method and apply post-processing human annotations for evaluation.

We compare the classifiers presented in Section 4.2, namely Naïve Bayes (including a variant with additional features), LSTM (including a variant with attention), FastText, and BERT. For the best performing method, we also experiment with multi-task approaches combining up to four most frequent tip types as different tasks. In addition, we experiment with the cross-domain approach, where the idea was to train the model only on one domain and evaluate its performance on another domain.

We use our labeled dataset containing a total of 3,059 tips and 48,870 non-tip sentences, obtained after the initial rule-based filtering described in Section 4.1. In addition, we use a simple baseline method that labels all sentences starting with verbs (VB or VBP part-of-speech tags [43]) as tips. This method has been used in key tip studies in domains other than e-commerce, either as the main approach [62] or as a baseline [23]. It yields low precision of 17.72% and low recall of 4.48% on our dataset, indicating that the task is far from trivial. This observation is aligned with previous findings in the travel domain, where 9.8% of the tips were reported to start with a verb [23].

⁵<https://github.com/google-research/bert>

Table 7. Recall results for classifying review sentences as tips at four different precision levels: 75%, 80%, 85%, and 90%.

Classifier	Recall@Precision=			
	75%	80%	85%	90%
Naïve Bayes	43.42%	26.71%	11.37%	6.18%
Naïve Bayes w/Features	40.71%	26.17%	12.07%	6.78%
LSTM	30.03%	25.25%	16.79%	15.30%
LSTM w/Attention	30.41%	22.40%	15.62%	12.38%
FastText	41.92%	29.01%	15.49%	8.88%
BERT	70.47%	58.05%	36.05%	19.33%

Table 8. Results of 5-fold cross-validation for classifying review sentences as tips. The best results in each column are boldfaced and mark statistically significant differences with the other models results.

Classifier	Precision	Recall	AUC
Naïve Bayes	70.44%	55.15%	0.66
Naïve Bayes w/Features	68.36%	58.84%	0.65
LSTM	67.52%	66.96%	0.66
LSTM w/Attention	62.88%	76.44%	0.66
FastText	67.61%	44.49%	0.65
BERT	71.83%	68.59%	0.73

As previously mentioned, the template-based technique proposed for travel tip extraction [23] is not applicable for our setting, since we could not detect dominant repetitive n-gram patterns in our data. The approach proposed by [62] is also not applicable, as we are working with product reviews rather than question-answer pairs.

5.1 Train/Test evaluation

We evaluate the algorithms over 6,118 sentences (a balanced set of 3,059 tips and 3,059 random non-tip sentences) and divide the labeled data to a 80% train and 20% test. We perform this evaluation 50 times, each with another random set of 3,059 sentences from the 48,870 labeled non-tips and randomly split the 6,118 sentences to train and test sets. As previously mentioned, our goal is to produce a small number of high-quality tips, especially suitable for presentation on mobile devices, with limited screen space. Specifically, we aim to produce up to 5 tips per product with high precision, even at the cost of compromising some of the recall. Table 7 depicts the average recall results of the different classifiers across the 50 train-test samples, while we set the desired precision threshold to 75%, 80%, 85%, and 90%. Note that when presenting 5 tips, 80% precision yields an average of one wrong tip per product.

As can be seen, BERT outperforms its competitors at all precision levels and achieves the best results, with 58.05% recall for 80% precision. It can also be observed that the addition of the non-textual features described in Table 5 to the Naïve Bayes classifier does not show a substantial performance gain; we therefore did not work with these features in the remainder of our experiments. The best performing configuration for the BERT classifier was with a batch size of 8, 5 epochs, and learning rate of 0.0002.

In order to further test the robustness and significance of the results, we perform a 5-fold cross-validation tuning and evaluation. Similarly to previous evaluation, we used a balanced set of 3,059 tips and 3,059 random non-tip sentences. In each iteration, 3 folds were used for training, 1 fold for hyper-parameter tuning and 1 fold for evaluation. Statistically significant differences of precision, recall and area under the ROC curve (AUC) were determined using the two-tailed approximate randomization test [46] at a 95% confidence level, with a random sample of 10,000 permutations, following the recommendations in [17]. We apply Bonferroni correction

for multiple comparisons. Hyper-parameter tuning was performed over the validation set, optimizing for the precision metric. Table 8 depicts the precision, recall and AUC results over the different classifiers. It can be observed that the BERT classifier outperforms its competitors by a statistically significant margin. Due to its high performance, we focused on BERT in our exploration of multi-task learning approaches, as described in the next section.

5.1.1 Multi-task learning. In addition to the regular BERT classifier, we also experiment with multi-task learning using BERT. Multi-task learning (MTL) is based on the idea that features trained for one task can be useful for related tasks. Multiple tasks are jointly trained by a shared model with an additional linkage between their trainable parameters, aiming at improving the generalization error [8].

MTL can be viewed as a form of inductive transfer learning, which can help improve the model by introducing an inductive bias. In the case of MTL, the inductive bias is provided by the auxiliary tasks, which are tasks conducted with the objective of improving the performance of the primary task. The inductive bias leads the model to prefer hypotheses that explain more than one task. Moreover, if the tasks share complementary information, they act as regularizers for each other. MTL has been widely used for deep learning tasks, e.g., in natural language processing [25, 55] and speech recognition [33].

We use the approach of MTL with hard parameter sharing, where the model includes a layer shared among all tasks (i.e., completely share parameters between tasks), with additional task-specific layers, that are learned independently for each task. We learn the task of predicting whether a review sentence is a tip with the additional auxiliary tasks that specialize on a specific type, particularly each of the top four types listed in Table 3 (Warning, Usage, Workaround, and Complementary Product), which cover over 76% of all the tips. Each specialized classifier is trained to predict if the sentence fits the specific tip type. For example, the Warning classifier learns to predict Warning sentences. Overall, the specialized classifiers may help the main task of predicting tip sentences.

Similarly to BERT, which feeds the output from the last layer into fully-connected feed-forward network, we change the output of the last layer such that it will be fed in parallel into 3-6 fully-connected feed-forward networks (one per task), trying to predict the output of the corresponding task independently. As mentioned above, we repeat this evaluation for 50 iterations, each with a balanced set of 6,118 sentences that consists of the same 3,059 labeled tips and a random set of 3,059 sentences from the 48,870 labeled non-tips. In each iteration, the 6,118 sentences are randomly split to train and test sets. In addition to the hyper parameters tuning of BERT, we tune the subset of auxiliary tasks considered for MTL and their weights in the loss function.

Table 9 depicts the results. Interestingly, the all-types model achieves the best performance, i.e., none of the BERT multi-task models could outperform the regular BERT model (that include all tip types). Except the all-types model, the best performing combinations in each precision level (marked in bold) were Warning + Workaround + Complementary product (for precision levels of 70%, 75%, 85%), Warning + Complementary product (for precision level of 80%) and Warning + Workaround (for precision level of 90%). It can be noted that the Usage type does not appear in none of the best performing combinations. We assume that this is due to the fact that usage is a broad type, hence less unique and therefore does not contribute additional performance. While the four types cover the vast majority of the tips, there are still 24% of the tips that belong to other types, hence it is hard to build a general model that captures all tip types using only four of them.

5.1.2 Cross-Domain evaluation. In this section, we present the results of a cross-domain evaluation, performed over the five domains introduced above. The idea is to train the model only on one domain and evaluate its performance on the other four domains to examine how well it can generalize.

The evaluation includes three setups (1) one-to-one cross-domain evaluation, where we train on one domain and test on another; (2) same-domain evaluation, where we train and test on the same domain; and (3) all-but-one cross-domain evaluation, where we train the model on four domains and test it on the remaining fifth domain. The

results of the evaluation (for the convenience of the presentation, we show the F1 score which is the geometric average of precision and recall scores) is depicted in Table 10.

For the one-to-one cross-domain evaluation, the training set includes all tip sentences from the source (train) domain and an equal number of randomly sampled non-tip sentences out of all the non-tips sentences from the same domain, so that the set is perfectly balanced. Similarly, the test set consists of a balanced set of all tip sentences from the test domain, and a randomly sampled non-tip sentences out of all non-tip sentences in the same domain. The set is split into 50% validation (for hyper-parameter tuning) and 50% “pure” test (for metric measurement).

For the same-domain evaluation (results are presented along the diagonal of 10), we use a balanced set of all tip sentences from the domain, and an equal-size random sample of non-tip sentences (same sentences that were sampled in the cross-domain setup for the test set). The evaluation was performed using 5-fold cross-validation to train, tune the hyper-parameters, and evaluate the classifier. In each iteration, 3 folds were used for training, one fold for validation, and one for test. We report the average results over the 5 test folds.

For the all-but-one cross-domain evaluation (last row of 10), we use a similar setup to the one-to-one cross-domain evaluation. The training set consists of a balanced set of all tip sentences from the four train domains and a random sample of equal size of non-tip sentences from each of the four domains. The test set includes all tip sentences from the target (fifth) domain and an equal number of non-tip sentences sampled randomly, same sample as in the one-to-one evaluation. The set is split to 50% validation (for hyper-parameter tuning) and 50% test.

Note that we use the same test set per each domain in all the experiments. In the same-domain setup, this test set was used for the 5-fold cross-validation.

Interestingly, for each test domain (columns in Table 10), except for Musical Instruments, the best performance is attained in the all-but-one setting rather than using same-domain training. This result implies that the all-but-one method has better ability to capture the differentiating tip sentence features and generalize to other domains. Models trained on a single domain (especially Home Improvement, Toys and Sports & Outdoors) often demonstrated comparable performance to the all-but-one model, but the most effective source domain is not consistent across all five target domains. We assume that these results indicate a high variance of the products (and hence the tips) in these domains, which helped to the generality of the model. Finally, the Musical Instruments domain was the most unique: the all-except-one method did not performed well on this domain. These results may stem from the fact that this domain has the smallest portion of tips (3.47% as shown in Table 2), but it could also imply that the Musical Instruments domain has its own specific (music related) language and unique types of products (e.g. “*The slight differences in tone can be easily manipulated with an EQ or just ordinary tone controls.*” or “*Individual strings may be intonated but action is done with two nuts, IE can raise one side higher or lower but not individual strings.*”).

5.2 Evaluation over Unseen Products from the Five Trained Domains

In order to simulate the practical use case of extracting the tips from large sets of reviews, we use the following evaluation method. We run the BERT model on previously-unseen products from the five domains, rank the tips by the model’s score, and select the top 5 tips. Then, we ask our in-house annotators to manually evaluate the generated tips. We specifically focus on popular products with many reviews, as for these we believe our method can work well even with low recall (as we mentioned before, we do not want to compromise precision). We check the quality for the scenario where we present to the user only a small number of tips: one, three, and five. This reflects the business need of extracting only few, but high-quality tips, which can fit within a limited user interface space on the product page. This evaluation method also allows us to gain insights about the number of reviews required to produce high-quality tips per each product. We start by considering all products above the

Table 9. Recall results of multi-task BERT classifier with four most prevalent types of tips at five different precision levels: 70%, 75%, 80%, 85%, and 90%.

Tip Type Auxiliary Tasks	Recall@Precision=				
	70%	75%	80%	85%	90%
Warning + Usage	77.71%	61.93%	47.50%	28.78%	15.11%
Warning + Complementary product	80.18%	61.56%	52.46%	27.37%	14.86%
Warning + Workaround	78.44%	64.31%	49.11%	29.16%	14.79%
Usage + Complementary product	79.99%	63.44%	49.35%	26.16%	14.53%
Usage + Workaround	79.97%	67.15%	48.84%	29.82%	12.83%
Workaround + Complementary product	78.19%	62.78%	49.28%	31.20%	16.09%
Warning + Usage + Workaround	80.04%	66.35%	50.55%	30.67%	13.71%
Warning + Usage + Complementary product	79.09%	65.24%	49.76%	29.67%	15.25%
Warning + Workaround + Complementary product	80.51%	68.54%	52.29%	31.82%	15.72%
Usage + Workaround + Complementary product	79.49%	64.75%	51.11%	29.57%	14.48%
Warning + Usage + Workaround + Complementary product	78.85%	64.49%	49.79%	30.64%	15.73%
All types	82.79%	70.47%	58.05%	36.05%	19.33%

Table 10. Cross-domain F1 score . ‘b’, ‘h’, ‘m’, ‘t’, and ‘s’ mark statistical differences with Baby, Home Improvement, Musical Instruments, Toys, and Sports & Outdoors, respectively. In each tested domain (column) the statistical differences is marked for the best two domains that achieved the highest performance compared to the other domains.

Train ↓ Test →	Baby	Home Improvement	Musical Instruments	Toys	Sports & Outdoors
Baby	67.96%	68.80%	69.29%	64.30%	70.40%
Home Improvement	64.71%	71.43%	72.94% ^s	62.58%	71.30% ^{mts}
Musical Instruments	62.41%	69.15%	72.20%	65.90%	66.56%
Toys	69.13%	72.08%	72.61% ^s	69.47% ^{bhs}	70.20%
Sports & Outdoors	69.75% ^{hm}	72.26% ^{bm}	67.59%	63.70%	67.18%
All except one	69.48% ^{hm}	73.05% ^{bm}	70.42%	72.83% ^{bhms}	72.73% ^{mts}

Table 11. Reviews per product by percentile and average.

Domain	10th	30th	50th	70th	90th	Average
Baby	5	6	9	13	29	22.81
Home Improvement	5	7	11	19	49	13.16
Musical Instruments	5	6	9	14	31	11.40
Sports & Outdoors	5	6	8	12	25	16.14
Toys	5	6	8	10	20	14.06

90th percentile according to their number of reviews in each of the five domains (Table 11 presents the statistics of review number per product). Then, we randomly sample 50 products from each domain and apply the BERT classifier on all review sentences of these products. Finally, we present the top sentences (ordered by classification score) and ask the annotators to evaluate if they are tips. Each sentence is reviewed by two annotators and considered as a tip only if both agree on it. The results of the top 1, 3, and 5 sentences are depicted in Table 12. Inspecting precision@1 (i.e., the portion of sentences with highest classification score that are deemed as tips), Home Improvement demonstrates the highest performance with a perfect 100%, followed by Baby and Sports

Table 12. Precision of top-k predicted sentences by BERT.

Domain	Top 1	Top 3	Top 5
Baby	98.00%	96.00%	97.20%
Home Improvement	100.00%	96.00%	92.80%
Musical Instruments	90.00%	84.33%	80.40%
Sports & Outdoors	98.00%	95.33%	94.00%
Toys	94.00%	90.67%	89.20%

Table 13. Examples of automatically-extracted tips.

Domain	Product	Tip	Tip Type
Baby	SoundSpa On-The-Go White Noise Machine	Be prepared to purchase a battery charger for AAA NiMH batteries if you want to run this thing every night.	Complementary product
Baby	Safety 1st Simple Step Diapering Disposal System	Be aware though, that if you are using a reusable liner, the mechanism for the foot pedal/opening the lid does not operate that great.	Warning
Baby	HALO Early Walker Sleepsack Wearable Blanket	However, you need to keep in mind that the feet at the bottom are not designed for running around playing really but more for sleeping in.	Usage
Home Improvement	PORTER-CABLE 7-Amp Plate Joiner Kit	If you keep your fingers above and on the opposite side of the board, there is no danger.	Warning
Home Improvement	8-LED Motion-sensing Night Light	If installing flat (like on a ceiling or under a shelf) the adhesive tape won't hold the weight of the unit with batteries .. you will have to use the mounting screws.	Workaround
Home Improvement	Culligan FM-15A Faucet Mount Advanced Filter	If you tighten the cartridge too tight it leaks.	Warning
Musical Instruments	Fender 351 Shape Premium Picks for Electric Guitar	Celluloid is highly flammable, so be aware of that if you smoke (I don't).	Warning
Musical Instruments	Dunlop Acoustic Trigger Gold Guitar Capo	Be careful about using capos like this, because if you're using a fine guitar, it may damage the finish.	Maintenance
Musical Instruments	String Swing Metal Home & Studio Wide Guitar Hanger	The yoke width is adjustable, and a combination of slope and two keeper rings prevent the instrument from coming off the holder.	Usage
Sports & Outdoors	Manduka PRO Eco Friendly Yoga Mat - 6mm Thick Mat	If you do find the mat is becoming sticky, use a Manduka Mat Renew or any non-solvent household cleaner and a damp cloth or sponge.	Maintenance
Sports & Outdoors	Invicta Men's Pro Diver Stainless Steel Watch	The directions that come with the watch are not very helpful and do not indicate that you have to unscrew the stem in a counterclockwise direction to pull it out to set the time and date.	First time use
Sports & Outdoors	Park Tool CT-5 Mini Chain Brute Bicycle Chain Tool	Be sure to read the directions as placing the chain in the wrong slot to break it or re-unite the chain can bend the links out of shape.	Warning
Toys	Crayola 48 Count Washable Sidewalk Chalk	To remove, wash surface with water pressure from garden hose.	Maintenance
Toys	Pretend & Play Teaching Cash Register	Take out one or two of the screws that hold that transparent piece of plastic to the top of the cash drawer.	First time use
Toys	Melissa & Doug Shapes Chunky Puzzle	They could also be used for tracing with paper and a crayon/marker.	Alternative use

Table 14. Precision of top-k predicted sentences by BERT on new unseen domains (domains that were not included in the training phase).

Domain	Top 1	Top 3	Top 5
Automotive	94.00%	90.67%	86.80%
Cellphones	90.00%	84.00%	80.40%
Electronics	90.00%	89.33%	88.00%
Fashion	90.00%	82.67%	75.20%
Health	94.00%	88.00%	82.80%

& Outdoors domains that attain a high 98%, and Toys with 94%. The lowest precision@1 is yielded for Musical Instruments, with 90%. Precision remains high in the Baby, Home Improvement and Sports & Outdoors domains for the top 3 and top 5. For Toys, it is down to around 90% for the top 3 and top 5. For Musical Instruments, the sharpest drop is recorded, down to around 84% for the top 3 and 80% for the top 5. These results show that our method can produce top tips at high precision: the precision@1, precision@3, and precision@5 across the five domains are 96%, 92.47%, and 90.75%, respectively. The Baby, Home Improvement, and Sports & Outdoors domains include relatively higher portion of tips within their reviews (Table 2) and yield the highest tip precision (Table 12), implying they are especially suitable for tip extraction. Table 13 presents examples of extracted tips, including the product’s domain, title, and tip’s type.

5.3 Evaluation over Unseen Products from New Domains

To test the generalization of our approach and its applicability to other domains using the same training data, we trained our model on all the tagged data from the five domains mentioned above and applied them on five new (previously-unseen) e-commerce domains: Automotive, Cellphones, Electronics, Fashion, and Health, using the same evaluation technique as in Section 5.2. The results of Precision at the top 1, 3, and 5 sentences are depicted in Table 14. Inspecting the top 1 tip, the best performing domains were Automotive and Health with 94% precision. The other three domains achieved 90% precision. The Electronics domain was stable around 90% also in top 3 and top 5, in contrast to the Cellphones domain which perform substantially worse (84% in top 3 and 80.4% in top 5). After an error analysis for the latter domain, we found that the tips are often longer and consists of several sentences, which were successfully detected by the model. However, since we evaluated only one sentence per tip, most of the tips were hard to understand without the surrounding context while presented standalone, and hence achieved lower performance.

The biggest drop was in the Fashion domain (82.67% in top 3 and 75.2% in top 5). We conjecture that the relatively low performance in the Fashion domain is due to the fact that the domain contains many tips of type Size (e.g., “*Fits true to your normal size, so don’t order up.*” for a Maternity Belly Band and “*If you wear size large, purchase size XL.*” for Womens Stretch Cotton Yoga Pants), specifically 34.4% of the total number of tips, compared to only 5.49% in the original set of domains (as mentioned in Table 3), hence it is harder for the general purpose model to recognize those specific tips.

The Health domain also exhibited a drop between the top 1, to the top 3, and to the top 5. Ultimately, for top 5, the best performing domain is Electronics, followed by Automotive. Overall, we see that cross-domain learning works even for previously-unseen domains. This findings is consistent across 5 different unseen domains, with precision equal or greater than 90% precision for the top 1 tip, and between 75% and 88% for the top 5 tips. Note that despite the rather good performance of the model in the Health domain (and existence of quite big amount of tips), this domain may be somewhat problematic for our application due to its sensitivity. For instance, consider the following tip extracted for Melatonin Tablets: “*Do not use this as a replacement for visiting your doctor, but it is*

Table 15. Examples of automatically-extracted tips from unseen domains.

Domain	Product	Tip	Tip Type
Automotive	Zwipes Microfiber Waffle Weave Kitchen Dish Towel	A word of caution, though - the color bleeds when washing quite a bit at first, so I suggest washing them in the sink the first couple of times.	Warning
Automotive	Griot's Garage 11023 Ultra-Thick Microfiber Towel	To get the most out of your towel, be sure to not use fabric softener when you wash it (true of any towel), and don't use it on anything that might have any grease or oil (it will permanently reduce the towel's absorbency, also something that is true of all towels).	Maintenance
Automotive	Camco RV Vent Insulator And Skylight Cover with Reflective Surface, Fits Standard 14" RV Vents (45192)	Be sure to crack open the vent a little (1/4") to allow hot air to escape, as the foil reflects the sun back under the inside the vent, much like a greenhouse and it may warp or damage the plastic vent.	Warning
Electronics	NETGEAR GS108NA ProSafe 8-Port Gigabit Ethernet Desktop Switch	Please note that this device does require it's own power source, so you'll need a wall outlet or a power strip wherever you intend to install it.	Complementary product
Electronics	Logitech Squeezebox Boom All-in-One Network Music Player / Wi-Fi Internet Radio	Be sure the battery is fully charged when you start, and disconnect the battery when its output drops by a volt (down to 11.6 volts); this will usually take a couple of hours.	First time use
Electronics	Logitech Z-5500 THX-Certified 5.1 Digital Surround Sound Speaker System	Just keep in mind it needs AIR, to cool itself so don't pile stuff all over and block the air flow around the heat sink, you'll just cook and kill it!	Warning
Cellphones	Generiks TM iPhone 4 / 4S *CLEAR* Screen Protectors	If you do get dust trapped underneath then lift it up with a piece of scotch tape and insert another piece of scotch tape to stick the dust particles.	Workaround
Cellphones	Jabra HALO2 Wireless Bluetooth Stereo Headset, Black	Close head phones up when done to avoid draining battery as battery life is about 6 hours.	Usage
Cellphones	TYLT TUNZ Rechargeable Bluetooth Speaker with NFC	Note that the internal battery is 2800 mAh in size, and you can only charge an attached device until the internal battery is half empty.	Usage
Health	Abreva Cream Tube 2gm	I suggest that you do not apply it directly to the sore this may contaminate the tube.	Usage
Health	Afrin PureSea Medium Stream, 4-Ounce Bottles	One word of warning: When you put the top on the bottle, don't press down too hard as this is what releases the fluid from the bottle.	Warning
Health	Old Spice Deodorant High Endurance, Smooth Blast 3.25-Ounce Sticks	*Disclaimer* careful if you wear white t-shirts as the blue color of this deodorant does come off sometimes onto the shirt.	Warning
Fashion	Dockers Men's Jean Cut Straight-Fit Pant	You may want to think about adding an inch to the waist from what you normally wear.	Size
Fashion	Clarks Women's Wave Trek Sneaker	Because they are a waterproofed suede leather, the leather does not stretch as easily as other shoes, so be sure to take time to get used to them, before wearing for a long term.	First time use
Fashion	Champion Men's Tech Performance Long Boxer Brief, Pack of 2	Another important care tip is to make sure you wash these separately, in the gentle cycle.	Maintenance

good enough for a general sense of your vitals between visits.". This example includes a medical advice extracted from a review that was possibly written by a non-expert and hence should not necessarily be trusted.

Table 15 depicts the examples of the extracted tips (3 examples per each domain). Note that even though the model was trained on completely different domains, it was general enough to find good tips in the new unseen domains. Nevertheless, there was a difference in the performance across the domains, since some domains are broader and contain a wide spectrum of tip types, while other domains include only specific tip types which are less popular in the broader domains (e.g., Size type in Fashion).

Table 16. Precision of top-k predicted sentences by BERT on eBay domains.

Domain	Top 1	Top 3	Top 5
Shoes	86.00%	85.33%	81.60%
Watches	96.00%	87.33%	86.40%

5.4 Evaluation over eBay Products

To further explore the generality of our approach, we want to assess if our findings can be expanded to other e-commerce platforms. Specifically, after training the tips model on the 5 domains from a publicly available reviews dataset, we set out to evaluate it over eBay, one of the largest e-commerce marketplaces. As opposed to other leading e-commerce platforms, such as Amazon and Walmart, eBay’s “first-class” entity is a listed item (‘listing’) offered for sale by a specific seller, rather than a catalog product. The catalog products are created by matching listings to existing products and creating new products for unmatched listings. The product notion is therefore somewhat weaker, and the results returned by eBay’s main search engine are listings rather than products. Having said all that, product pages, which include customer reviews, still play a key role in the overall user experience on eBay. Our experiments focus on two different domains, which have been especially popular on the platform during 2020 and 2021: Shoes and Watches. These domains are somewhat narrower than the five domains used for training and evaluation of our model, as described in the previous section. These differences allow us to examine the generalizability of our approach even further. We note that the Shoes and Watches domain are different in their properties. The popular products (exact definition is given below) in the Watches domain have twice⁶ the number of reviews compared to the popular products in the Shoes domain. Moreover, the reviews in the Shoes domain are shorter than the reviews in the Watches domain (average of 16.27 vs 20.23 words per review and median of 9 vs 11 words per review, respectively).

Similarly to the method described in Section 5.3 we consider the most popular products (products in the 90th percentile according to their number of reviews). Then, we randomly sample 50 products from each domain and apply the BERT classifier on all the review sentences of these products. Finally, we order the sentences by the classification score, present the top 5 and ask the annotators to evaluate if they are tips. As previously mentioned, each sentence is reviewed by two annotators and considered as a tip only if both agree on it. The results of the top 1, 3, and 5 sentences are depicted in Table 16. Generally, the results are comparable to these reported for other domains (Table 14), indicating our model can generalize to new e-commerce platforms, while maintaining good precision (over 80% for the top 5 tips). Across all the top-k values, but especially for the top 1, the model performed much better on the Watches domain than the Shoes domain. We conjecture that due to the fact that the Watches domain have on average more reviews per product and longer reviews compared to the Shoes domain, the probability to find a tip sentence in their reviews is greater. Interestingly, while performing error analysis for the results, we identified a noticeable difference from the results of the five domains from the public dataset: most of the wrongly identified tips in the eBay reviews data were seller-related. As explained above, since the notion of a listing by a specific seller is strong on eBay, the reviews also contain feedback about the seller, such as the shipping time, responsiveness, money-back policy, price, and so forth. We therefore hypothesize that adding to the training process eBay specific data can further improve the model performance and its overall generalizability. Nonetheless, the high performance results, especially for the Watches domain, indicate that there are common characteristics to the language of e-commerce reviews and tip expression across different e-commerce platforms. Table 17 presents examples from the tips extracted using our model for the Shoes and Watches domains on eBay.

⁶We omit the exact number due to business sensitivity

Table 17. Examples of automatically-extracted tips in eBay.

Domain	Product	Tip	Tip Type
Shoes	FILA Mens WeatherTec Black Boot 10 Men US	I should add that you need to have a good heavy winter sock on for all day comfort, I use an extra heavy all wool hunting sock.	Complementary product
Shoes	Dearfoams Women's Velour Ballerina Slippers Large Pewter	This is one of the best styles, you can wash them in the washing machine and put them in the dryer for a couple minutes- leaving them out overnight to finish drying safely.	Maintenance
Shoes	Women Casual Straps Warm Plush Lining Mid Calf Snow Boots	Definitely order at least one size larger than you'd normally wear as they run small.	Size
Shoes	PUMA Redon Move Black White	Only drawbacks to look out for are that these shoes are extremely narrow.	Warning
Shoes	Specialized Comp Road Shoe Black/red	In order to anchor your heel down and back, you need to crank the BOA system which is harder than the ratchet system.	Workaround
Watches	Casio Classic A158WA-1DF Wrist Watch for Men - Silver	It's nice and comfortable fits good on men but girls might have to remove some links to fit.	Size
Watches	Casio Waveceptor WV58A-1AV Wrist Watch for Men	One trick I found is to remove the band and the back cover to remove the battery, but usually it just needs to be lifted to disconnect it and cause the watch to reboot.	Workaround
Watches	Citizen Eco-Drive BJ8050-08E Wrist Watch for Men	The only other problem is that you need adapters to change the strap.	Complementary product
Watches	Casio Waveceptor WV58DA-1AV Wrist Watch for Men	Have to manually reset it to the correct time.	First time use
Watches	Citizen Eco-Drive JY0000-53E Wrist Watch for Men	You just need to make sure it's exposed to the Sun from time to time.	Usage

BERT was the best performing method in all of our evaluations, in particular for the unseen domains (from various e-commerce platforms) and the cross-domain evaluation. We conjecture that BERT outperformed its competitors due to its ability to learn implicit representations that capture a variety of linguistics characteristics. Our analysis indeed shows that there are domain-agnostic characteristics of tip sentences, such as the use of second person pronoun and special phrases typical of advice. In addition, since BERT can be pre-trained over a large corpus and then fine-tuned using small data for a specific task, it can effectively generalize to new domains.

5.5 Limitations

The main limitation in the described approach is the low recall. In addition, tip sentences are scarce. Particularly, every 22 review sentences contain only one tip sentence, on average. Even after applying the rule-based filtering, a tip appears every 17 review sentences, or every 4-5 reviews on average. These two limitations make our approach applicable to products that have accumulated a large number of reviews. Having said that, popular products are broadly exposed, and so their effect on a large number of users can be high and henceforth pave the way for tips receiving more prominence on e-commerce platforms.

Another limitation is the potential repetition of tips extracted for a given product. As the number of desired tips per product grows, diversification should be applied to avoid redundant tips. For example, the following usage tips, “*To remove the marker designs from the screen you need to use a damp cloth then allow to completely dry - just a bit of a pain*” and “*To clean the screen off, you just run a damp paper towel or cloth over the screen*”, were generated for the same Widescreen Light Designer. Semantic similarity can be used to cluster similar tips, so that the final list is diverse. The clusters' size can also assist in selecting the top tips to be presented. Such an approach was found productive in previous work on product description generation from reviews [47] and can be similarly applied in our case.

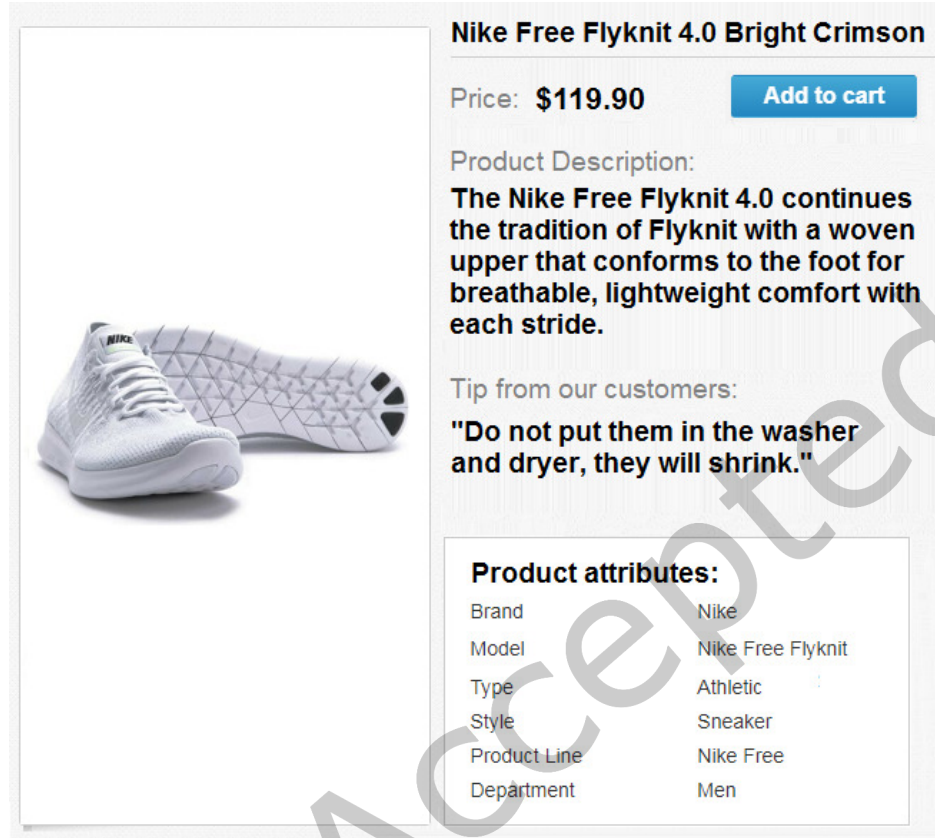


Fig. 2. Visionary UI that demonstrates usage of a tip along with product title, attributes, description and image.

6 APPLICABILITY IN REAL-WORLD PLATFORMS

In this section, we focus on the applicability of our approach in real-world e-commerce platforms. First, we discuss potential applications of the proposed method for both sellers and buyers. Second, we present a user study with 10 experienced shoppers, designed to assess the effect of tip presentation on product pages on potential buyers' experience.

6.1 Potential Applications

Tips generated from user reviews can be integrated in various use cases within e-commerce platforms, both for buyers and sellers.

Tips' applications for buyers. The tips can be presented for buyers as part of the product page or for certain search results. In both cases the presented tips can support the purchase decision (e.g., Size, Complementary product, Warning tip types) or provide additional useful information to an already purchased product (e.g., First time use, Maintenance, Workaround tip types). Figure 2 depicts a visionary user interface of a product page that presents tips. The UI contains the regular product information, such as its title, image, description and attributes together with a newly added section that contains a tip automatically extracted from the reviews.

The number of tips may change, e.g., only one tip will be presented for users who browse using mobile devices and up to 5 tips for web browser users. As already mentioned, another potential use case is showing the tips on the search results page [15, 27, 56]. For example, when a user searches for running shoes, we can show the Size tips (e.g., “I suggest to purchase one size larger than usual cause they are small”) or Complementary product tips (e.g., “Buy extra batteries if you want to use them often”). Such tips may help customers make more informed purchase decision and reduce return rates [52, 59, 70]. Moreover, the tips can be also used for diversified recommendations, e.g., the Alternative use tips can expose buyers to different types of product and avoid “more-of-the-same” recommendations. This may increase the satisfaction and engagement of the buyers and encourage them to return to the platform for their next online shopping. Another significant reason to return to the platform after the purchase may be the “after purchase” tips. In this list, buyers can view their purchased products together with the relevant First time use, Maintenance, Workaround and Alternative use tips.

Tips’ applications for sellers. Sellers can also benefit from the extracted tips. First, sellers may consider selling products suggested in the Complementary Product tips, as separate products or by composing a bundle and/or lot [38, 60] (e.g., batteries, chargers, and additional accessories). Other types of tips such as Size, Alternative Use and Workaround may help sellers better describe their products, provide additional useful information and gain more trust from the potential buyers. Warning tips may help the sellers to avoid offering dangerous products or suggesting risky products to the wrong population, e.g., a toy with small pieces to a baby. The platform may also suggest sellers to print the First time use tips on the package as a nice service for buyers who are about to use for first time. Since the tips are generated automatically, the platform should notify the sellers about newly added tips to the product they offer for sale. The seller should be able to see all the tips for each of their products in one place and the suggested action items as described above (e.g., printing instructions for first time use, state the warnings upfront, or better describe their products in terms of size, maintenance, population segment, and alternative uses).

In summary, automatically-extracted tips can be useful for both buyers and sellers in online marketplaces and their real world uses should be further explored on different e-commerce platforms.

6.2 User Study

In order to examine the effect of our extracted tips in a real-world scenario, we designed a user study that compares product presentation with and without tips. We examined tips extracted from products in the following five domains : Baby, Home Improvement, Musical Instruments, Sports & Outdoors, and Toys. The goal of the study was to evaluate the usefulness of showing tips on product pages and get a sense of the potential impact of tip presentation on buyers’ experience and purchase decision. The study included 10 participants (6 female and 4 male, with ages ranging between 24 and 41), who were experienced online shoppers, i.e., each of them performed at least five online purchases during the last quarter of 2021. The study compared the effect of the presentation of products with tips (actual tips extracted using our method), to the presentation of products without tips. It included 25 randomly selected products (5 from each domain) that were presented to all of the participants. For each of the products, we extracted the top tip from its corresponding reviews using our previously described method. Each product was presented to half of the participants with the tip and to the other half without a tip. For each participant, half of the products (12 or 13) were presented with a tip and the other half without a tip. The order of the products was randomized per participant.

In the study, the products were presented to the participants using the user interface shown in Figure 2. Specifically, for each product, we presented the product’s title, image, description, and attributes. As explained above, in half of the cases the product also included an automatically extracted tip, which was presented under the “Tip from our customers” section. In order to examine the effect of tip presentation, the participants were asked to answer two questions: “Do you find this product attractive?” and “Would you consider to buy this product?”.

Table 18. Rating results for the question “Do you find this product attractive?”. Boldfaced average ratings are significantly higher based on a one-tailed paired t-test with $p < .01$.

Domain	Average Rating		% Rated 5	
	With tips	Without tips	With tips	Without tips
Baby	4.56	3.88	60%	20%
Home Improvement	4.80	3.80	80%	28%
Musical Instruments	4.40	4.04	56%	20%
Sports & Outdoors	4.56	3.88	68%	12%
Toys	4.64	4.12	76%	24%
All	4.59	3.94	68%	20.8%

Table 19. Rating results for the question “Would you consider to buy this product?”. Boldfaced average ratings are significantly higher based on a one-tailed paired t-test with $p < .01$.

Domain	Average Rating		% Rated 5	
	With tips	Without tips	With tips	Without tips
Baby	3.96	3.48	16%	16%
Home Improvement	3.88	3.20	8%	4%
Musical Instruments	3.80	3.64	20%	16%
Sports & Outdoors	3.72	3.36	12%	4%
Toys	3.96	3.44	32%	12%
All	3.86	3.42	17.6%	10.4%

Participants were asked to rate their answers to both question on a 5-point Likert scale, with 1 representing “not at all” and 5 representing “very much”. We did not disclose to the participants the topic of the study.

Tables 18 and 19 show the rating results for both questions in our user study. For each question, the average rating and the portion of 5 ratings out of all ratings are shown across all products presented with a tip versus all products presented without a tip. This comparison is shown for each of the five domains and in total, across all five domains. In general, as could be expected, the ratings for the first question (product attractiveness) are substantially higher than for the second question (consider buying).

Inspecting the average rating results in Tables 18 and 19, a consistent trend can be observed, where products presented with a tip receive higher average ratings than products presented without a tip. For the product attractiveness question, all differences are significant, while for the consider-buying question, they are significant for 3 of the 5 domains. For both questions, the largest rating difference in favor of products with tips can be observed for the Home Improvement domain, while the smallest gap can be observed for the Musical Instruments domain. This is consistent with our findings when evaluating the quality of tips, presented in Section 5, which indicated that the top tip is of highest precision for Home Improvement and of lowest precision for Musical Instruments (see particularly Table 12).

Examining the portion of products rated 5 indicates similar trends. For the product attractiveness question, large gaps can be observed across all five domains, as products presented with tips receive much more often the

highest attractiveness rating. Noticeably, the smaller gap is attained for the Musical Instruments domain compared to all other domains, consistent with the generally lower quality of tips for this domain. For the consider-buying question, the portions of products rated 5 are generally low and likely affected by participants' specific interests and needs (e.g., baby products are especially relevant to young parents, sports products are relevant to those practicing particular activities). The total difference across all domains in this case is still considerable between products presented with tips (17.6% rated 5) compared to products presented without tips (10.4%), but the gaps vary substantially across domains.

Overall, the results of our user study suggest that the presentation of a tip as part of the product page has a significant effect on whether potential buyers perceive the product as attractive and would consider buying it. These results are consistent across all inspected e-commerce domains and are more significant for domains for which higher quality tips could be extracted. Future work should experiment with in-vivo presentation of tips on product pages in e-commerce platforms to validate the findings presented in our study.

7 CONCLUSION AND FUTURE WORK

In this work, we propose a tip extraction method from product reviews. We train various models on five domains that naturally contain useful and non-trivial tips across the reviews and are likely to be beneficial for potential customers.

We formally define the task of tip extraction in e-commerce by providing the list of tip types, tip timing (before and/or after the purchase), and connection to the surrounding context sentences. We evaluate different approaches of supervised tip extraction that are trained on labeled data from 14,000 product reviews. Tips are labeled using a dedicated tool and released for public use, as part of a dataset's extension. In addition, we use a variety of evaluation methods on several domains that include, among others, cross-domain and cross-platform experimental study to show the robustness of our approach.

For the evaluation of the five domains that contain labeled data, the best performing method, BERT, achieves recall of 58.05% at 80% precision on a balanced test set. Moreover, when the method is applied to unseen products, the precision@1 is 90% for the lowest domain (Musical Instruments) and 100% for the highest (Home Improvement). Precision@5 is 80.4% for the lowest domain (Musical Instruments) and 97.2% for the highest (Baby). Our method is not specific to any of the five domains and can therefore be potentially applicable to other e-commerce areas. For the five additional (unseen) domains from the same e-commerce platform the lowest precision@1 (90%) was achieved in Cellphones, Electronics, Fashion domains while the highest precision@1 (94%) was achieved in Automotive and Health domains. For the two additional domains from another platform, the Watches domain outperformed Shoes domain in both precision@1, precision@3 and precision@5 (96%, 87.33% and 86.4% vs 86%, 85.33% and 81.6% respectively). The performance of the model over the unseen domains (both from the same and from the different platforms) clearly shows the ability of the model to generalize. Hence, acquiring training data for a new domain is not necessarily required due to these promising results demonstrated on the unseen domains. Moreover, since labeled data is a costly resource, this outcome is specifically important for large e-commerce platforms that support thousands of different categories [2]. Finally, we provide various types of potential applications of the proposed method for the e-commerce platforms. These application would improve both buyers and sellers experience in the platform. We demonstrate a visionary user interface for one of such applications and discuss the other possible usages in details. Our small-scale user study indicates that a presentation of a tip as part of the product's page can have a significant effect on its attractiveness to potential buyers, as well as their decision whether to purchase the product.

For future work, we plan to focus on five main directions. Currently, we focus on extracting single-sentence tips, but as discussed in Section 3, over 25% of the tips can be extended to include adjacent sentences; hence, extending our approach to support multi-sentence tips is a key direction. Second, we plan to explore abstractive

approaches that combine content from different sentences and adapt them [21, 42]. This can help increase the number of extracted tips and also deal with sentences that contain irrelevant information in addition to the tips. Third, tip diversification is an important step in providing multiple useful and non-repetitive tips. Fourth, we plan to investigate how to elevate characteristics specific to different tip types in order to improve the overall tip extraction quality. Finally, implementing the proposed applications on the e-commerce platform and studying its effect on user behavior is another intriguing research direction.

REFERENCES

- [1] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. 2013. Opinion fraud detection in online reviews by network effects. In *Proc. of ICWSM*, Vol. 7.
- [2] Uri Avron, Shay Gershtein, Ido Guy, Tova Milo, and Slava Novgorodov. 2022. Automated Category Tree Construction in E-Commerce. In *Proc. of SIGMOD*. 1770–1783.
- [3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet rating of product reviews. In *European conference on information retrieval*. Springer, 461–472.
- [4] Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proc. of SIGIR*. 222–229.
- [5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL* 5 (2017), 135–146.
- [6] BrightLocal. 2018. *Local Consumer Review Survey*. <https://www.brightlocal.com/research/local-consumer-review-survey/>
- [7] David Carmel, Erel Uziel, Ido Guy, Yosi Mass, and Haggai Roitman. 2012. Folksonomy-Based Term Extraction for Word Cloud Generation. *ACM Trans. Intell. Syst. Technol.* 3, 4, Article 60 (sep 2012), 20 pages.
- [8] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.
- [9] Chien Chin Chen and You-De Tseng. 2011. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems* 50, 4 (2011), 755–768.
- [10] Lei Chen, Jie Cao, Huanhuan Chen, Weichao Liang, Haicheng Tao, and Guixiang Zhu. 2021. Attentive multi-task learning for group itinerary recommendation. *Knowledge and Information Systems* 63, 7 (2021), 1687–1716.
- [11] Lei Chen, Jie Cao, Youquan Wang, Weichao Liang, and Guixiang Zhu. 2022. Multi-view Graph Attention Network for Travel Recommendation. *Expert Systems with Applications* 191 (2022), 116234.
- [12] Lei Chen, Jie Cao, Guixiang Zhu, Youquan Wang, and Weichao Liang. 2021. A multi-task learning approach for improving travel recommendation with keywords generation. *Knowledge-Based Systems* 233 (2021), 107521.
- [13] Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.
- [14] Paolo Cremonesi, Raffaele Facendola, Franca Garzotto, Matteo Guarnerio, Mattia Natali, and Roberto Pagano. 2014. Polarized review summarization as decision making tool. In *Proc. of AVI*. 355–356.
- [15] Arnon Dagan, Ido Guy, and Slava Novgorodov. 2021. An image is worth a thousand terms? analysis of visual e-commerce search. In *Proc. of SIGIR*. 102–112.
- [16] 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* (2018).
- [17] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proc. of ACL*. 1383–1392.
- [18] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. Do online reviews matter? An empirical investigation of panel data. *Decision support systems* 45, 4 (2008), 1007–1016.
- [19] Guy Elad, Ido Guy, Slava Novgorodov, Benny Kimelfeld, and Kira Radinsky. 2019. Learning to generate personalized product descriptions. In *Proc. of CIKM*. 389–398.
- [20] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378–382.
- [21] Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Abstractive text summarization by incorporating reader comments. In *Proc. of the AAAI Conference*, Vol. 33. 6399–6406.
- [22] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bitu Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proc. of EMNLP*. 1602–1613.
- [23] Ido Guy, Avihai Mejer, Alexander Nus, and Fiana Raiber. 2017. Extracting and ranking travel tips from user-generated reviews. In *Proc. of WWW*. 987–996.
- [24] Ido Guy and Bracha Shapira. 2018. From Royals to Vegans: Characterizing Question Trolling on a Community Question Answering Website. In *Proc. of SIGIR*. 835–844.

- [25] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple NLP tasks. *arXiv preprint abs/1611.01587* (2016).
- [26] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proc. of WWW*. 507–517.
- [27] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query reformulation in E-commerce search. In *Proc. of SIGIR*. 1319–1328.
- [28] Sharon Hirsch, Slava Novgorodov, Ido Guy, and Alexander Nus. 2021. Generating Tips from Product Reviews. In *Proc. of WSDM*. 310–318.
- [29] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [30] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. of KDD*. 168–177.
- [31] Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. 2012. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems* 52, 3 (2012), 674–684.
- [32] Chunli Huang, Wenjun Jiang, Jie Wu, and Guojun Wang. 2020. Personalized review recommendation based on users’ aspect sentiment. *ACM Transactions on Internet Technology (TOIT)* 20, 4 (2020), 1–26.
- [33] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7304–7308.
- [34] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proc. of EMNLP*. 423–430.
- [35] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint abs/1412.6980* (2014).
- [36] Theodoros Lappas, Mark Crovella, and Evimaria Terzi. 2012. Selecting a characteristic set of reviews. In *Proc. of KDD*. 832–840.
- [37] Theodoros Lappas and Dimitrios Gunopulos. 2010. Efficient confident search in large review corpora. In *ECML PKDD*. Springer, 195–210.
- [38] Gal Lavee and Ido Guy. 2022. Lot or Not: Identifying Multi-Quantity Offerings in E-Commerce. In *Proc. of ECNLP* 5. 250–262.
- [39] Piji Li, Zihao Wang, Lidong Bing, and Wai Lam. 2019. Persona-Aware Tips Generation. In *The World Wide Web Conference*. 1006–1016.
- [40] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proc. of SIGIR*. 345–354.
- [41] Stephen W Litvin, Ronald E Goldsmith, and Bing Pan. 2008. Electronic word-of-mouth in hospitality and tourism management. *Tourism management* 29, 3 (2008), 458–468.
- [42] Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2018. Generative adversarial network for abstractive text summarization. In *Thirty-second AAAI conference on artificial intelligence*.
- [43] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. (1993).
- [44] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint abs/1301.3781* (2013).
- [45] Thanh-Son Nguyen, Hady W Lauw, and Panayiotis Tsaparas. 2013. Using micro-reviews to select an efficient set of reviews. In *Proc. of CIKM*. 1067–1076.
- [46] Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.
- [47] Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. 2019. Generating product descriptions from user reviews. In *Proc. of WWW*. 1354–1364.
- [48] Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. 2020. Descriptions from the Customers: Comparative Analysis of Review-based Product Description Generation Methods. *ACM Transactions on Internet Technology (TOIT)* 20, 4 (2020), 1–31.
- [49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*, Vol. 14. 1532–1543.
- [50] Irina Rish. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. 41–46.
- [51] Jihee Ryu, Yuchul Jung, and Sung-Hyon Myaeng. 2012. Actionable clause detection from non-imperative sentences in howto instructions: A step for actionable information extraction. In *TSD*. Springer, 272–281.
- [52] Hannu Saarijärvi, Ulla-Maija Sutinen, and Lloyd C Harris. 2017. Uncovering consumers’ returning behaviour: a study of fashion e-commerce. *The International Review of Retail, Distribution and Consumer Research* 27, 3 (2017), 284–299.
- [53] Ruben Sipos and Thorsten Joachims. 2013. Generating comparative summaries from reviews. In *Proc. of CIKM*. 1853–1856.
- [54] Doug Snowball. 1980. Some effects of accounting expertise and information load: An empirical study. *Accounting, Organizations and Society* 5, 3 (1980), 323–338.
- [55] Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proc. of ACL*, Vol. 2. 231–235.
- [56] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A taxonomy of queries for e-commerce search. In *Proc. of SIGIR*. 1245–1248.

- [57] Cheri Speier, Joseph S Valacich, and Iris Vessey. 1999. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences* 30, 2 (1999), 337–360.
- [58] Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proc. of ICWSM*.
- [59] David Tsurel, Michael Doron, Alexander Nus, Arnon Dagan, Ido Guy, and Dafna Shahaf. 2020. E-commerce dispute resolution prediction. In *Proc. of CIKM*. 1465–1474.
- [60] Hen Tzaban, Ido Guy, Asnat Greenstein-Messica, Arnon Dagan, Lior Rokach, and Bracha Shapira. 2020. Product bundle identification using semi-supervised learning. In *Proc. of SIGIR*. 791–800.
- [61] Shaohua Wang, NhatHai Phan, Yan Wang, and Yong Zhao. 2019. Extracting API tips from developer question and answer websites. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 321–332.
- [62] Ingmar Weber, Antti Ukkonen, and Aris Gionis. 2012. Answers, not links: extracting tips from yahoo! answers to address how-to web queries. In *Proc. of WSDM*. 613–622.
- [63] Alfian Farizki Wicaksono and Sung-Hyon Myaeng. 2012. Mining Advices from Weblogs. In *Proc. of CIKM*. 2347–2350.
- [64] Alfian Farizki Wicaksono and Sung-Hyon Myaeng. 2013. Toward advice mining: Conditional random fields for extracting advice-revealing text units. In *Proc. of CIKM*. 2039–2048.
- [65] Qing-Song Xu and Yi-Zeng Liang. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* 56, 1 (2001), 1–11.
- [66] Cheng Yang, Lingang Wu, Kun Tan, Chunyang Yu, Yuliang Zhou, Ye Tao, and Yu Song. 2021. Online User Review Analysis for Product Evaluation and Improvement. *Journal of Theoretical and Applied Electronic Commerce Research* 16, 5 (2021), 1598–1611.
- [67] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proc. of NAACL*. 1480–1489.
- [68] Qiang Ye, Rob Law, and Bin Gu. 2009. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management* 28, 1 (2009), 180–182.
- [69] Di Zhu, Theodoros Lappas, and Juheng Zhang. 2018. Unsupervised tip-mining from customer reviews. *Decision Support Systems* 107 (2018), 116–124.
- [70] Yada Zhu, Jianbo Li, Jingrui He, Brian Leo Quanz, and Ajay A Deshpande. 2018. A Local Algorithm for Product Return Prediction in E-Commerce.. In *Proc. of IJCAI*. 3718–3724.