Generating Product Descriptions from User Reviews

Slava Novgorodov and Ido Guy
eBay Research
Netanya, Israel
snovgorodov@ebay.com,idoguy@acm.org

Guy Elad and Kira Radinsky
Technion, Israel Institute of Technology
Haifa, Israel
{sguyelad,kirar}@cs.technion.ac.il

ABSTRACT

Product descriptions play an important role in the e-commerce ecosystem, conveying to buyers information about a merchandise they may purchase. Yet, on leading e-commerce websites, with high volumes of new items offered for sale every day, product descriptions are often lacking or missing altogether. Moreover, many descriptions include information that holds little value and sometimes even disrupts buyers, in an attempt to draw attention and purchases. In this work, we suggest to mitigate these issues by generating short crowd-based product descriptions from user reviews. We apply an extractive approach, where review sentences are used in their original form to compose the product description. At the core of our method is a supervised approach to identify candidate review sentences suitable to be used as part of a description. Our analysis, based on data from both the Fashion and Motors domains, reveals the top reasons for review sentences being unsuitable for the product's description and these are used, in turn, as part of a deep multi-task learning architecture. We then diversify the set of candidates by removing redundancies and, at the final step, select the top candidates to be included in the description. We compare different methods for each step and also conduct an end-to-end evaluation, based on rating from professional annotators, showing the generated descriptions are of high quality.

CCS CONCEPTS

• **Information systems** → **Electronic commerce**; **Online shopping**; • **Applied computing** → **Electronic commerce**; • **Computing methodologies** → *Natural language generation*; *Multi-task learning*.

KEYWORDS

Deep multi-task leaning; electronic commerce; language generation; user-generated content.

ACM Reference Format:
Slava Novgorodov and Ido Guy and Guy Elad and Kira Radinsky. 2019. Generating Product Descriptions from User Reviews. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313532

WWW 2019, May 13-17, 2019, San Francisco, California

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313532

1 INTRODUCTION

The importance of content on e-commerce websites has been widely recognized. High-quality and trusted product content has been empirically shown to have a substantial influence on user behavior, which is manifested in conversion rates and the volume of sales [28, 35, 39]. Product descriptions are an important element of the content displayed on product pages, alongside the product's title, image, and key attributes, such as model name, color, or size. Yet, such descriptions are often lacking or missing; for example, the majority of the Fashion products on eBay have no description at all. Even when available, product descriptions are often long and tedious to read, containing a lot of information that is insignificant for potential buyers. Our own analysis indicates that substantial portions of the product description sentences include details specific to a single listing or seller, information about the brand as a whole, and pure marketing statements.

We refer to a *product description* as a written (textual) presentation of what the product is, how it can be used, and why it is worth purchasing. The purpose of a product description is to provide customers with details about the features and benefits of the product so they are compelled to buy.[1] In line with previous research on e-commerce content, we expect a good description to be informative, readable, objective, and relevant to the product (e.g., as opposed to a specific listing or a whole brand) [58]. We focus on concise descriptions of several sentences, which can be quickly consumed in their entirety and are especially suitable for small-screen devices, as e-commerce mobile applications have seen a remarkable growth and account for a major portion of the overall e-commerce traffic [29, 31]. Like other types of product content, credible descriptions have been shown to increase sales, while lacking descriptions withhold users from reaching a purchase decision or effectively searching for products [36].

In light of the aforementioned challenges, we propose to use the "crowd" to generate trustworthy product descriptions, by leveraging the products' user reviews [15, 27, 38]. Such reviews are often abundant on e-commerce websites and reflect the perspective of those who have already purchased the product. Therefore, potential buyers tend to trust reviews much more than they trust seller-provided content [30, 59].

User reviews primarily aim at reflecting a buyer's subjective perspective and include personal opinions, stories, experiences, and complaints, which are not suitable to include in a product description (e.g., "*My old shoes wore down and I needed a new pair*" or "*I can plug it to each of my three cars*"). The large volumes of user reviews accumulated for popular products, with each review typically containing multiple sentences, makes them practically

[1]This definition is largely based on that by the Shopify e-commerce platform: https://www.shopify.com

impossible to consume. As a result, users often read only a few reviews and may miss helpful information that appears in others. We observe that some portion of the review sentences are descriptive of the product [15], and suggest an extractive approach to generate *crowd-based* descriptions by combining original review sentences.

The transformation from reviews to descriptions is a challenging task, which to the best of our knowledge, is novel. While reviews aim to reflect the buyer's perspective, descriptions typically reflect the viewpoint of the seller. Moreover, while reviews are meant to reflect a variety of subjective opinions, descriptions are expected to provide objective fact-based information. Several prior studies have examined review summarization (e.g., [26, 45, 71]), however such summaries do not necessarily contain descriptions of the product.

Our extraction of candidate sentences from user reviews, to be included in the product depiction, is primarily supervised. We examine both classic machine learning models and deep learning approaches for the classification task, trained over thousands of sentences in two key e-commerce domains: Fashion and Motors. We also analyze the key reasons making review sentences unsuitable for a description. A deep multi-task learning classifier, which is based on mapping the top reasons to auxiliary tasks, is found to yield the best performance for the candidate identification task.

Following, we select the top sentences out of the candidate set for the final product description. To this end, we use a sentence similarity measure that helps diversify and avoid redundancies. Semantic similarity based on word embedding is found to be more effective than a bag-of-words approach. We experiment with several basic methods to produce the final description, which rely on the classification score from the candidate extraction process and the similarity measure. We perform a large-scale evaluation of the descriptions based on thousands of ratings from professional annotators, comparing the different methods and inspecting three description lengths: 3, 5, and 7 sentences.

The main contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first work to suggest the extraction of product descriptions from reviews.
- We provide analysis and examples of how review sentences can be used for descriptions (what portion, which kind of sentences) across two principal yet very different e-commerce domains.
- We develop a classifier for identifying review sentences suitable for product descriptions, reaching an AUC of over 0.92.
- We present an end-to-end system for description generation from reviews, comparing different approaches for sentence selection, reaching an average rating of 4.3 (out of 5) per description.

2 RELATED WORK

Textual product descriptions have been explored in the e-commerce literature along with other seller-provided product content types, such as titles [19, 52], images [6, 20], and attributes [52, 55]. Some of the studies refer to the product description in a broad sense, which encompasses the other content types, e.g., the list of structured attributes [2], title [66], or product image [48]. Other studies refer to a product description similarly to us, as the textual writeup that extends the title and attributes. Probst et al. [57] studied the extraction of attribute-value pairs from such product descriptions, in order to enrich the product's structured representation, used for tasks such as recommendation and matching. Shinzato and Sekine [62] proposed an unsupervised approach for the same task. Dumitru et al. [11] applied text mining and clustering techniques over product descriptions in order to recommend product features for a given domain. In a recent study, Pryzant et al. [58] showed that product descriptions of the type we study can help predict the product's business outcome. Experimentation was based on product descriptions and sales records from the Rakuten Japanese e-commerce website. None of these studies, however, provided a definition of a product description as presented in this work.

As mentioned in the Introduction, a related task to the one we explore is review summarization. Different from standard text summarization [18], where the goal is to generate a concise summary for a single [64] or multi-document [40], review summarization aims at extracting and summarizing opinions about a product from multiple reviews [45, 70]. Most studies focus on identifying the key attributes of an entity, such as a product, a movie, or a hotel, and then extracting key phrases that describe these attributes or the sentiment towards them (e.g., [26, 37, 56, 73]). Techniques used for this type of summarization include rule-based mining [42], topic modeling [51, 67], and neural networks [46, 68]. An overview of review summarization techniques can be found in several surveys [32, 41, 43, 54].

While summarization seeks a good coverage of the main topics within the set of reviews, sometimes revolving around key product attributes, we aim to identify a unique subset of the reviews' content that is descriptive of the product . A sentence that may be pivotal to the set of reviews, and thereby for its summary, might not be appropriate for a description. For example, the sentence "*I like this hat very much*" may be included in a review summary, but poses low value for a product description, as it is subjective and provides little information about the product. In our experiments, described later in this paper, we found that fewer than 10% of the review sentences were suitable "as is" to take part in the product description.

Another related body of research has focused on extracting experiences [50, 53] and tips [21, 72] from user reviews. Somewhat similarly to the motivation presented in this work, these studies aim at helping users sift through the large volumes of reviews by identifying a more specific type of information within the reviews. Nonetheless, extracting experiences and tips is each inherently different than extracting descriptive sentences: experiences are subjective in nature, reflecting the unique viewpoint of an individual user (or group) and are thus not suitable, almost by definition, to be part of a product description. Tips, on the other hand, are defined as concrete and typically actionable pieces of advice. Therefore, their extraction actually aims at excluding purely descriptive sentences, of the type pursued in this work.

3 DATASETS AND CHARACTERISTICS

In this section, we describe the datasets used for our analysis and experimentation and their characteristics.

Datasets. Our research is based on two product datasets from two principal yet very different e-commerce domains: Fashion (clothing, shoes, and jewelry) and Motors (automotive parts and

Table 1: Fashion and Motors dataset characteristics.

	Fashion				Motors			
	Avg	Std	Median	Max	Avg	Std	Median	Max
Reviews per product	1118	1371	596	8271	883	985	378	7352
Sentences per review	3.71	3.09	3	103	4.05	3.73	3	98
Words per sentence	8.04	6.43	6	58	9.24	7.26	7	63
Number of products	892				807			
Number of reviews	997,274				712,904			

Table 2: Most distinctive unigrams for reviews vs. descriptions ('Reviews') and vice versa ('Descriptions') in Fashion and Motors.

Fashion		Motors	
Reviews	Descriptions	Reviews	Descriptions
i	your	i	power
my	designed	my	generator
it	you	it	protection
they	features	was	your
was	comfort	great	fuel
but	jewelry	but	torqx
very	cotton	have	portable
great	imported	so	provides
good	inches	works	uego
these	fashion	this	economy
them	polyester	good	watts
bought	technology	they	designed
me	provides	very	advanced

accessories). Both datasets were obtained from a large e-commerce website in the United States, representing best-selling products in each of the two domains.[2] The datasets contain, per product, both its description and user reviews. Table 1 presents the characteristics of the two datasets. The number of products is rather similar in both datasets, while for Fashion there are more reviews per product and for Motors the number of sentences per review and their length in words is slightly higher. In addition, we used two larger datasets with over 10 million reviews of over 10,000 best-selling products in each domain (Fashion and Motors), for pre-training word embeddings, as will be described later in this paper.

Data Annotation. Labeling for training and evaluation in this work was performed by in-house professional editors (annotators), with domain expertise in both Fashion and Motors. The pool included a total of 20 editors, of whom different subsets were selected for different tasks, proportionally to the task's size. Unless otherwise stated, each evaluation was performed by a single editor.

Description Characteristics. The descriptions in our dataset are substantially longer than those we aim to generate. The average number of sentences per description in the Fashion dataset is 28.2 (std: 12.9, median: 27, min: 11, max: 65) and for motors it is 26.8 (std: 12.1, median: 29, min: 9, max: 68). To get a preliminary sense of the content of these descriptions, 50 descriptions from each dataset were annotated by two professional annotators. Only 45% of the sentences were labeled as suitable for a product description, with the key reasons for the sentences not being adequate including purely subjective marketing statements, accounting for slightly over 20% of the sentences (e.g., "*Give your clothes the luxury they deserve with these wonderful hangers!*"); information specific to a seller or a listing (18%; "*1 year limited warranty*"), and description of the brand as a whole (15%; "For over 35 years, we have been one of the largest sunglasses brands.")

Reviews vs. Descriptions. In essence, reviews and descriptions hold fundamentally contrasting characteristics: reviews reflect a subjective opinion based on an individual experience, while descriptions are expected to be "dryer", explaining what the product is and why it is worth purchasing. As a first step, we set out to examine the most prominent language differences between reviews and descriptions. To this end, we used Kullback-Leibler (KL) divergence, which is a non-symmetric distance measure between two given distributions [4]. Specifically, we calculated the terms, per each of the two domains, which contribute the most to the KL divergence between the language model of the reviews versus the language model of the descriptions and vice versa [22].

[2] As of July 2018.

Table 2 presents the most distinctive unigrams. It can be seen that the unigram list most characterizing reviews (relative to descriptions) is rather similar between Fashion and Motors. The first-person pronouns 'i' and 'my' are at the top of the two lists, both common on reviews yet hardly occurring on descriptions. For example, 'i' occurs on 3.43% of the Fashion review sentences and 3.11% of the Motors review sentences, whereas for descriptions it occurs on 0.03% and none of the sentences for both datasets, respectively. Other prominent unigrams on the review lists include 'was', which typically reflects a past-tense experience ('bought' can also be observed on the Fashion list); the third-person pronouns 'it' and 'they'; the adjectives 'good' and 'great', which often reflect a subjective opinion; and the emphasizing adverb 'very'. Further down the list, beyond Table 2, we also encountered unigrams that refer to aspects of a specific listing of the product, which may vary from one seller to another, such as 'price', 'cheap', and 'shipping'; other third-person references such as 'she' or 'son'; and verbs that reflect subjectivity ('recommend', 'like', 'love') or past tense ('arrived', 'purchased', 'got', 'ordered').

The most characterizing unigrams for descriptions relative to reviews are more dissimilar between Fashion and Motors. As can be seen in Table 2, each list includes its own domain-specific descriptive words, such as 'jewelry', and 'cotton' for Fashion, or 'power' and 'fuel' for Motors. The second-person pronoun 'your' ranks high on both lists, indicating that while first- and third-person language is used almost exclusively on reviews, second-person language is more characteristic of descriptions (e.g., 'your' occurs on 1.34% of the Fashion description sentences versus only 0.18% of the review sentences). The only other common words between the two lists are 'provides' and 'designed'.

Overall, the above analysis gives an indication of the key differences in the language of reviews versus descriptions. Following, review sentences to be used as part of a description need to be carefully selected. We elaborate on this process in the next section.

4 CANDIDATE SENTENCE EXTRACTION

In this section, we describe a key component of our description generation method: the extraction of candidate review sentences that can be used for the product's description. Given the set of all user reviews for the product, our goal is to identify a set of sentences that can be used in their original form for a description of the same product. We first apply rule-based filtering based on the analysis presented in the previous section. We then apply a supervised approach that learns to identify review sentences suitable for a description. We examine different types of classifiers for this task and compare their performance based on a large labeled set of review sentences.

4.1 Rule-Based Filtering

Considering our analysis of linguistic differences between descriptions and reviews, we established several simple rules to identify review sentences that cannot be used as part of a description: (1) **Short**: sentences of 3 words or fewer generally introduce little information and do not flow well as part of a product description. For example, "*Recommended*", "*Very good quality*", or "*no complaints*", as well as "*good fit*" or "*very soft*" for Fashion and "*easy to install*" or "*works as expected*" for Motors, were among the most common short review sentences in our datasets. Overall, short sentences accounted for 17.1% of all review sentences in the Fashion domain and 17.9% in Motors. (2) **Personal**: sentences with a first-person pronoun, such as 'i', 'my', 'our', or 'us', or a 3rd-person *personal* pronoun, such 'she', 'his', 'hers', but not 'it' or 'them'. As demonstrated in the previous section, such pronouns hardly ever occur on a product description. Examples include "*I like the color of these jeans*"; "*Perfect fit for our car*"; "*My husband makes good use of them*"; and "*best gift for his birthday*". Overall, 35.9% of the review sentences matched this filtering criterion in the Fashion domain and 37.8% in Motors. (3) **Listing-specific**: Some review sentences refer to listing-specific aspects, as observed in the previous section. Examples include "*Great value for a fair price*" or "*Delivery was smooth and fast*". Since our goal is to produce a description at the product level rather than the listing (item) level, such aspects are not suitable for referencing as part of the description, since they may vary according to the seller. Our blacklist for this rule included the unigrams 'price', 'cheap', 'expensive', 'delivery', 'shipping', 'seller', and 'warranty'. Overall, 5.7% of the review sentences matched this rule in the Fashion domain and 6.1% in Motors.

Our rules aim to filter out sentences that are not suitable for a description with a very high likelihood, almost by definition. We therefore did not filter out other potential candidates, such as sentences in past tense, since these can sometimes be appropriate (e.g., "Tested on several cars"). We also did not automatically filter out sentences from reviews with low ratings, because the vast majority of the reviews in our dataset had a positive rating, in line with past work that indicated online user reviews tend to the positive [8]. Overall, 53.7% and 55.2% of the review sentences were filtered out using the three rules above, for Fashion and Motors, respectively.[3]

4.2 Automatic Classification

After the initial rule-based filtering, we set out to explore a supervised approach, where we trained a classifier to predict whether a product review sentence is suitable as a description sentence for the same product. To this end, we sampled uniformly at random, out of all review sentences that were not filtered out by the rules, 25K sentences for each of the two datasets, Fashion and Motors. Each of the 25K sentences were then labeled by a group of 10 annotators, who were asked to mark them as either '*good*' or '*bad*', i.e., suitable to be part of a product description or not. In case the sentence was labeled *bad*, the annotators also selected a reason. The set of reasons was identified in an earlier round of labeling, and included 'other' in case none of the 10 reasons was appropriate. The annotators received detailed guidelines, explaining what makes a sentence suitable versus unsuitable for a description, with examples of *good* and *bad* sentences, as well as examples for each of the possible reasons for *bad*. They also performed qualification tests, i.e., an iterative process of labeling, followed by feedback from other annotators, until the quality was aligned among all. At the end of the process, the inter-annotator agreement for the task of *good* versus *bad* labeling, measured by Cohen's kappa [9], was 0.89 for Fashion and 0.9 for Motors, calculated over a set of 300 sentences labeled by two different annotators. Overall, 8.55% of the Fashion and 7.97% of the Motors sentences were labeled *good*.

Table 3: Reasons for review sentences labeled '*bad*' and their distribution (portion of all sentences marked *bad*) for Fashion and Motors.

Reason	% Fashion	% Motors	Example
Subjective	52.50%	52.43%	It was the easiest jumpstart ever.
Missing context	16.86%	16.82%	Otherwise it remains idle.
Refers to a listing's aspect	8.40%	6.73%	10 bucks for 3 pairs is a great deal.
Non-informative	7.95%	6.42%	This shirt is great.
Poor language and spelling	4.91%	5.17%	Extremly easy setup let's you pull you vehecle's code fast.
Negative sentence	3.90%	4.25%	Only issue is the pretty thin material.
Expresses explicit doubt	2.40%	2.30%	Probably good also for bicycle tires.
Refers to the description	1.83%	1.74%	The hat is exactly as described.
Other	1.49%	1.24%	Like others here have said, this gas can has a long rotating nozzle.
Too specific/detailed	0.64%	2.52%	Great for Honda 2003 2.0L.
Offensive language	0.12%	0.19%	Fantastic product, bright as sh*t.

Table 3 lists the different reasons for *bad* sentences. It can be seen that the distribution is similar for Fashion and Motors, with subjective sentences accounting for a little over half of the *bad* sentences in both, followed by sentences with a missing context. The only noticeable difference between the domains is for the too specific/narrow reason, which is generally uncommon, but occurred substantially more frequently in Motors.

4.2.1 Good vs. Bad Review Sentences. Before building our classifier, we performed a statistical analysis comparing the review sentences labeled *good* by our annotators with review sentences labeled *bad*.

Table 4 presents the most distinctive unigrams and bigrams for *good* review sentences versus *bad* review sentences and vice versa, for Fashion and Motors. Distinctive terms were calculated using KL divergence as described in Section 3. It can be seen that *good* sentences in both domains include positive adjectives and adverbs, such as 'easy' and 'very', which are at the top of both unigram lists, as well as 'great', 'sturdy', 'nice', and 'well', which are used to describe products as easy to use/install/assemble; being of good/high quality or well made; or being perfect/excellent/great for a specific use. While many of the terms are common for both domains, in Fashion traits such as comfortable, durable, or fitting

[3]The portions of all three rules do not sum up to the total number of filtered sentences, since some sentences matched more than one rule.

Table 4: Most distinctive unigrams and bigrams for *good* sentences vs. *bad* sentences ('*Good*') and vice versa ('*Bad*') in Fashion and Motors.

Fashion				Motors			
Good	*Bad*	*Good*	*Bad*	*Good*	*Bad*	*Good*	*Bad*
easy	was	easy to	a little	easy	was	easy to	it was
very	but	good quality	it was	very	but	to use	so far
sturdy	not	well made	so far	great	it	very easy	than the
quality	it	very sturdy	a bit	use	would	to install	better than
comfortable	love	perfect for	but the	install	not	works great	a little
nice	than	are very	as expected	quality	than	good quality	a bit
well	as	they fit	they were	works	had	well made	seems to
great	buy	great for	but it	well	as	to apply	as advertised
durable	had	high quality	happy with	nice	buy	very well	as described
is	would	to assemble	but they	your	this	comes with	happy with

Table 5: Percentage of *good* sentences in Fashion ('F') and Motors ('M') distributed by sentence length, review length, and sentence position within the review.

Sentence length (words)		4	5	6	7 – 8	9 – 10	11 – 12	13 – 15	16+
	F	9.22%	9.96%	10.21%	10.38%	8.32%	7.81%	6.69%	6.05%
	M	11.07%	9.91%	9.43%	8.86%	7.42%	6.86%	6.60%	6.63%

Review length (sentences)		1	2	3	4	5 – 6	7 – 9	10+
	F	7.49%	9.19%	10.27%	9.46%	8.33%	8.11%	6.81%
	M	9.15%	10.04%	10.30%	9.88%	8.24%	6.69%	5.66%

Sentence position in review			1st	2nd	3rd	4th	5th
	F	Review length 3	11.48%	9.24%	8.90%		
		Review length 4	10.67%	11.16%	7.02%	5.27%	
		Review length 5	8.92%	10.99%	8.72%	6.64%	4.35%
	M	Review length 3	12.25%	8.78%	7.56%		
		Review length 4	11.31%	10.76%	7.97%	6.82%	
		Review length 5	9.89%	11.50%	8.88%	5.59%	3.86%

well are salient, while in Motors aspects of installation, application, and work are more dominant. For *bad* sentences, there is also a substantial overlap between the domains, implying the language differences between *good* and *bad* sentences can be generalized. The list of terms reflects some of the key reasons for sentences not fitting into a description, such as subjective viewpoints (e.g., 'better than', 'happy with', 'love'), personal experiences in past tense ('it was'), expression of doubt ('seems to'), negative sense ('but', 'not') and a reference to the description itself ('as described', 'as advertised').

Table 5 shows the portion of sentences labeled as *good* according to various characteristics of the sentence and the review it originated from, for Fashion and Motors. Inspecting the sentence length, starting 4 words, it can generally be seen that shorter sentences are somewhat more likely to be labeled *good*. For Fashion, the "optimal" sentence length is 6-8 words, while for Motors the percentage of *good* sentences consistently decreases with the length. Long sentences are more likely to include expressions that would make them unsuitable for a description based on some of the reasons detailed in Table 3, particularly subjectivity, which is the most common. Looking at the length of the review, it can be observed that sentences originating from long reviews are less likely to be *good*. For both Fashion and Motors, the review length that yields the highest portion of *good* is 3 sentences. Finally, inspecting the position of the sentence within the review (while controlling for the review length), there is a consistent trend indicating that sentences that occur towards the beginning of the review (not necessarily the first sentence) have higher likelihood to be *good*. The last sentence of the review has a particularly low likelihood to fit a description. This trend is consistent for Fashion and Motors and persists for higher review lengths not shown in Table 5. Overall, it suggests that users are more likely to include descriptive details earlier in the review.

4.2.2 Supervised Approaches. For the *good* versus *bad* classification task, we experimented with both traditional machine learning models and deep learning approaches. For the latter, we examined the effect of an attention mechanism and the use of a multi-task learning approach that tries to predict a reason for a sentence being inadequate for a description. We used 5-fold cross-validation to tune the hyper-parameters and evaluate the classifiers. As evaluation metric, we used the area under the ROC curve (AUC).

Table 6: AUC performance results for classifying review sentences as *good* or *bad* for the product's description.

Classifier	Fashion	Motors
Naïve Bayes	0.779	0.798
XGBoost	0.831	0.839
LSTM	0.914	0.916
LSTM-Attention	0.915	0.916
LSTM-MTL	**0.924**	**0.924**

Naïve Bayes and XGBoost. We examined two common models for text classification: Naïve Bayes [60] and XGBoost [7]. Our features included textual features, specifically the unigrams, bigrams, and trigrams of the review sentence, and statistical features of the sentence and originating review, based on Table 5. For Naïve Bayes, we tuned the type of n-grams (unigrams, bigrams, or trigrams) and for XGBoost, in addition, the maximum depth of a tree, minimum child weight, as well as the learning rate and number of rounds (trees). Results for both classifiers are presented at the top rows of Table 6, indicating XGBoost achieved a higher AUC.

To better understand the contribution of non-textual features, which are based on the characteristics presented in Table 5, we trained the XGBoost classifier with different feature subsets (while always using textual features). Feature types included sentence length (in words and characters), review length (in words and sentences), and position (absolute position and normalized by the total number of review sentences). Results, presented in Table 7, indicate that position features yielded a slightly higher performance gain than sentence and review length. Overall, however, as can be observed in the 'All' row, the contribution of these features on top of the textual features was minor (+0.97% for Fashion and +0.96% for Motors). We therefore did not use these features in the remainder of our experiments and worked with the review text only.

While the XGBoost classifier demonstrated reasonable performance, we set out to explore how it can be further enhanced. In recent years, various applications in text classification have shown significant improvement by the use of deep learning models [17, 34].

Table 7: AUC performance of the XGBoost classifier with textual features when using ('Only') or disregarding ('Exclude') additional types of features. The 'All' row refers to all additional types jointly.

Additional Features	Fashion		Motors	
	Only	Exclude	Only	Exclude
Sentence Length	0.824	0.828	0.833	0.835
Review Length	0.828	0.826	0.835	0.834
Sentence Position	0.829	0.828	0.837	0.836
All	0.831	0.823	0.839	0.831

We therefore explored several deep learning approaches for our task:

LSTM. A recurrent neural network based on long short-term memory (LSTM) [25] architecture, with pre-trained word2vec embeddings using continuous bag of words (CBOW) with negative sampling [49]. We experimented with pre-training over Wikipedia and over our own datasets of more than 10 million product reviews for Fashion and Motors, respectively, as described in Section 3. Pre-training using our own data, with separate models for Fashion and Motors, consistently achieved a slightly better performance. We henceforth only report the results when using word2vec pre-trained based on our own data.

LSTM with Attention. Attention mechanisms enable the network to focus on relevant parts of the input [69]. The overall architecture of the "attention network" consists of two components: an LSTM-based word sequence encoder and a word-level attention layer. Given a review with the words $w_i, i \in [1, N]$, we first embed the words using pre-trained word2vec, as previously described. We then use the LSTM network to produce the hidden states $h_i, i \in [1, N]$. The attention mechanism is subsequently used to put more focus on certain words in the review sentence. For example, in the sentence "*this is a high quality product*" the words "*high quality*" should receive higher weight. We feed the word annotation h_i through a single-layer perceptron network to receive u_i, a latent representation of h_i. Then, we calculate the similarity of u_i with a word-level context vector, normalized by a softmax function, to produce the word's importance weight. We then construct the review vector as a weighted sum of the word annotations based on each word's weight.

Deep Multi-Task Learning. Multi-task learning (MTL) is based on the idea that features trained for one task can be useful for related tasks. Models for all tasks of interest are jointly trained with an additional linkage between their trainable parameters, aiming at improving the generalization error [5]. Multi-task learning 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 lead the model to prefer hypotheses that explain more than one task. MTL has been widely used for deep learning tasks, e.g., in computer vision [16] and natural language processing [23, 63].

Our network is based on an MTL hard parameter sharing architecture [61], which includes an LSTM layer shared among all tasks and separate feed-forward networks per task. We learned the task of predicting if a review sentence could be included as part of a description (*good*/*bad*) with the additional auxiliary tasks that

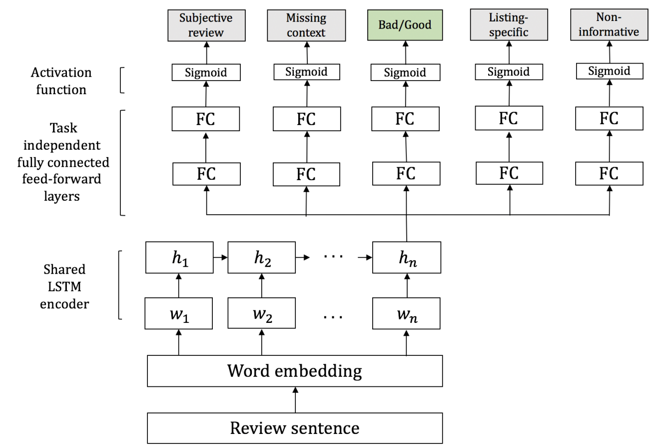

Figure 1: Hard parameter sharing for deep multi-task learning architecture.

specialize on a specific reason, particularly each of the top four reasons listed in Table 3, which cover over 80% of the *bad* sentences in both Fashion and Motors. Each specialized classifier is trained to predict if the sentence falls under the specific reason for not fitting in a description. For example, the 'subjectivity' classifier learns to predict subjective sentences. Overall, the specialized classifiers may help the main task of predicting *good* or *bad* sentences. Similarly to the LSTM with attention, given a review sentence, we first embed the words using our pre-trained word2vec, and then use the shared LSTM encoder to produce the latent representation of the review sentence. This representation is then fed in parallel into 2-5 fully-connected feed-forward networks (one per task), each with two hidden layers, trying to predict the output for that task independently. The ultimate loss function to be optimized assigns a weight to each of the task-specific loss functions. Notice the weighted loss function serves for optimization at training time, while testing remains for the *good* versus *bad* task only. In Figure 1, we demonstrate our MTL hard parameter sharing network, with a shared LSTM layer and 5 separate feed-forward layers, one for each task.

For each of the three deep learning methods, hyper-parameter tuning included the batch size, the dropout rates, and the number of hidden units in the LSTM layers. In addition, we experimented with both Adam [33] and RMSProp [65] optimizers. For the MTL architecture, we also tuned the subset of tasks to be included and their weight in the loss function, as well as the dropout rates and number of hidden units of the feed-forward layers.

4.2.3 Performance Results. Table 6 presents the AUC results for the *good* versus *bad* task. Evidently, the deep learning classifiers achieved substantially better results than the Naïve Bayes and XGBoost classifiers. The attention mechanism did not lead to any performance gain, possibly due to the sentences' length: as shown in Table 1, the median number of words per review sentence is 6 and 7, for Fashion and Motors, respectively. The effectiveness of the attention mechanism, however, typically lies in longer sentences [44]. The LSTM-based MTL model achieved the best AUC

Table 8: AUC performance of the deep MTL classifier with different subsets of the four auxiliary tasks: subjective ('subj'), missing context ('MC'), listing-specific ('LS'), and non-informative ('NI').

Fashion		Motors	
Subj, MC, LS, NI	0.921	Subj	0.920
LS, NI	0.922	MC, LI	0.922
Subj	0.922	MC, LS, NI	0.923
Subj, LS, NI	0.923	SUBJ, MC	0.923
Subj, MC	**0.924**	**Subj, MC, LS, NI**	**0.924**

results for both datasets, Motors and Fashion. Table 8 presents the auxiliary task subsets that yielded the best performance. For Motors, the combination of all four auxiliary tasks led to the best performance, while for Fashion it was the combination of the subjective and missing context tasks. Indeed, these are the two most common reasons for sentenced being labeled *bad*, as shown in Table 3. Other combinations listed in Table 8 yielded close performance to the top ones. Noticeably, using only one auxiliary task – for identifying subjective sentences – produced a substantial part of the performance gain compared to the vanilla LSTM, especially for Fashion. As previously shown (Table 3), 'subjective' is the most common reason, covering over half of the *bad* sentences for each of the two domains.

As mentioned before, we compared Adam and RMSProp optimizers. The latter was found to perform better, slightly but consistently across different number of tasks, for both Fashion and Motors.

5 DIVERSIFICATION

In order to create coherent and concise descriptions, we wished to avoid the inclusion of very similar sentences in the same description. After identifying candidate sentences for the description, our next step was therefore the identification of redundant sentences in order to increase the diversity in the content of the final description. The diversification phase was based on the computation of similarity between candidate sentences. To this end, we used the common cosine similarity measure, while experimenting with three sentence representation methods:

Weighted Bag-of-Words (BOW). Each sentence is represented as the TD-IDF weighted vector of its words [1].

Average Embedding. The weighted BOW approach measures similarity based on actual word overlap. In order to capture deeper semantic similarity, an embedded representation of the sentence can be used. Specifically, we apply word embedding using word2vec and average over all the sentence's words to produce the final sentence representation. The word2vec model is trained separately for Fashion and Motors, based on the dataset of over 10 million reviews per domain, as described in Section 3.

Weighted Embedding. This method applies the same word2vec embedding as described above, but instead of averaging the embeddings to generate the sentence representation, a common word weighting approach is applied [3]. This weighting is performed using TF-IDF scores, as it has been empirically shown to achieve similar results to learning the language model of the sentences.

Table 9: Accuracy for classifying sentence pairs as similar or not using different representations.

Sentence Representation	Fashion	Motors
Weighted bag of words	76.6	78.3
Average embedding	88.3	88.6
Weighted embedding	**92.2**	**91.7**

To decide whether two sentences are different enough to be included in the description, a similarity threshold θ for each method had to be determined. To avoid using a hard-coded threshold, we learned θ in an unsupervised manner from the data, for each of the two domains (Fashion and Motors). To this end, we considered the product descriptions in both of our datasets, which were manually curated by domain experts. We measured the similarity between each pair of sentences within each description and considered the 90-th percentile as the threshold θ. Intuitively, we set our threshold to allow a similarity level that is aligned with the typical degree of similarity that exists in professionally-written descriptions.

In order to compare the three representations, we conducted a small-scale experiment. For each domain (Fashion and Motors), we sampled 1000 pairs of sentences in the following way: we sampled 200 products from the respective dataset uniformly at random. For each product, we sampled 10 pairs of review sentences, uniformly at random out of all (unordered) pairs of sentences labeled *good*. We evaluated only *good* sentences, since the *bad* sentences were not considered for the final description. The sampled pairs were labeled by 3 professional annotators as either similar or not. The annotators were instructed to label a pair of sentences as similar in case one of them did not add any substantial piece of information to the reader on top of the other. The agreement between annotators, measured by Cohen's Kappa, was 0.84 for Fashion and 0.86 for Motors, calculated over 100 pairs for each domain, which were evaluated by two different annotators.

Overall, 35.2% of the sentences were labeled as similar, indicating there was indeed a high level of redundancy among *good* sentences. We tested each of the three methods against this labeled set by calculating its accuracy, i.e., the portion of pairs it correctly classified out of the 1000 pairs. Results are depicted in Table 9. It can be seen that both embedding-based methods reached a substantially higher accuracy than the weighted BOW method, for both domains. Weighted embedding slightly outperformed average embedding at over 90% for both domains, and was therefore our choice as sentence similarity method for the rest of our experiments.

6 FINAL DESCRIPTION GENERATION

In this section, we describe our methods for producing the final product description. Following, we present a detailed evaluation of the generated descriptions. Our evaluation compares the different methods for producing the descriptions and different description lengths by their overall quality ratings, as well as by ratings of more specific description aspects. In addition, we present results over publicly available data and experiment with cross-domain description generation.

6.1 Sentence Selection

The final step of our product description generation process is the selection of the final sentence set. From the previous steps, we have a list of candidate sentences, with their classification score reflecting their likelihood to be suitable to a product description, a similarity metric that can be applied to a pair of candidate sentences to measure their closeness, and a similarity threshold θ reflecting the desired similarity between sentences in a description. With these in hand, we examine four methods to generate the final description, given an integer K that determines the desired number of sentences in the description.

Greedy approach. This method traverses the list of candidates according to their *good*/*bad* classification score, from highest to lowest, and adds a candidate to the description if (and only if) it is not similar (i.e., has a similarity score of θ or higher) to any of the candidates already selected for the description. The process stops when K sentences have been added.

LexRank. This method uses LexRank, a common extractive summarization method that yielded high performance results for several text summarization tasks [12]. Specifically, LexRank assigns an importance score per sentence, using random walks and eigenvector centrality. We apply LexRank on top of the list of candidate sentences and use its score to rank the candidate sentences. Intuitively, the LexRank score indicates how well the candidate covers the information included in the other candidates. The method then operates similarly to the greedy approach: after ranking the candidate sentences by their LexRank score, the list is traversed by descending score and a sentence is added in case it is not similar to any of those previously selected for the description.

K-means classification score. This method partitions the candidate sentences into K clusters using the k-means algorithm [47] with k=K. The distance between sentences is calculated as $(1-s)$, where s is our sentence similarity measure described in Section 5. For the final product description, the clusters are traversed from the largest (representing the highest number of review sentences) to the smallest and from each cluster, the sentence with the highest classification score is added to the description.

K-means centroid. The method works as the previous one, but instead of selecting the candidate with the highest classification score from each of the K clusters, the candidate closest to the centroid is selected, assuming it best represents the cluster.

Note that smart ordering of the sentences within the final description is beyond the scope of this work. Currently, the sentences are ordered by classification score (greedy), summarization score (LexRank), or cluster size (both k-means methods).

6.2 End-to-End Description Evaluation

For an end-to-end evaluation of our approach, we set out to examine full descriptions generated using one of the four methods described above. For candidate sentence extraction, we used the LSTM-MTL classifier and for similarity measure we used weighted sentence embedding, since both were found to yield the best performance for their respective tasks, as previously reported. We experimented with descriptions of three lengths: 3, 5, and 7 sentences. These represent relatively concise content, which can be swiftly consumed in its entirety, and is especially suitable for small-screen devices,

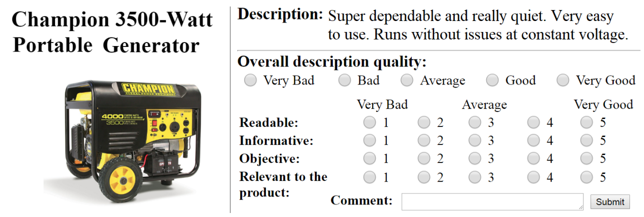

Figure 2: Product description evaluation interface.

such as mobile phones. We generated descriptions for all products in both the Fashion and Motors datasets. Using our candidate extraction method (Section 4), we identified review sentences that were suitable for a description for each product in the sample. On average, for each Fashion product we identified 46.1 *good* sentences (std: 9.3, median: 43, min: 24, max: 97) and for Motors 47.3 (std: 11.7, median: 45, min: 19, max: 103). For each such product, we then generated descriptions of length 3, 5, and 7, using each of the four methods described above (a total of 12 description versions). These descriptions were evaluated by 15 professional annotators. We ensured each annotator evaluated no more than one description per product.

Annotators were presented with the product's title, image, and generated description and were asked to rate the quality of the description's text for serving as the product's description, on a 5-point Likert scale, from 'very bad' to 'very good'. In addition, to examine finer-grained aspects of the description's quality, along the lines presented in the Introduction, annotators were asked to assess, on the same scale, to what extent the description's text was readable, informative, objective, and relevant to the product. They were provided with good and bad examples for each of the questions. Figure 2 demonstrates the user interface developed to collect annotators' input.

The inter-annotator agreement for the main quality evaluation question, measured using weighted Kappa [10], was 0.83 for Fashion and 0.84 for Motors, calculated over a set of 200 generated descriptions evaluated by two different annotators. Table 10 presents a few examples of our generated descriptions, which were rated 5.

6.2.1 Rating Results. Table 11 shows the ratings of the main question (overall description quality) in our end-to-end evaluation. Generally, the generated descriptions received high ratings, with an average of above 4 for all methods except for the greedy approach. Ratings for Fashion and Motors were similar, giving some indication that our method may be generalized to other e-commerce domains. Among the four methods, LexRank achieved the highest ratings, consistently for all three description lengths for both Fashion and Motors. At the other extreme, the greedy approach yielded substantially lower ratings than the other methods. K-means with the closest-to-centroid selection was consistently the second best, higher than k-means with highest-score selection. This suggests that selecting the sentence based on representation of the whole cluster rather than the individual prediction score for description suitability is preferable. The superiority of the LexRank approach implies that good representation of the whole set of candidates for final sentence selection works well. All differences between the

Table 10: Example descriptions rated 5.

Domain	Product	Description
Fashion	Socks	Perfect thickness for shoes or boots. Extra padding at toes. The quality is excellent. Easy to handle and very comfortable. No shrinkage in washer or dryer.
Fashion	Jacket	The cuffs on the sleeves are adjustable, which is perfect for keeping wind out when biking or for just a tighter fit around your wrists. It's very soft on the skin. The zipper works easily. The material is well made and the hood tucks away easily and hides well when not needed. Folds up to fit in bag.
Motors	Scratch removal system	Great for minor to moderate damage. Use it for removing scratches from your car. There is enough for several repairs. Coat twice for extra protection. Very recommended after wax.
Motors	Motor oil	It works well and lasts for a reasonable period of time. The oil is still very clean when changed. Engine runs smoother and pulls better. Good oil for any motorcycle. This oil is made for high temp engines.
Toys	Magnetic cubes	Works for all ages – toddlers to teens. The letters are a great addition and a fun way for young kids to learn their alphabet. Very colorful and high quality magnetic toy! Very easy to build. Perfect for family time.
Electr.	Mobile phone	This phone has a very good battery. The screen is big enough for watching movies. Includes power charger and headphones. The quality of photos is excellent. The phone is very smart and connects well with other devices.

Table 11: Average rating of description quality of 3, 5, and 7 sentences for products in Fashion and Motors.

	K=3		K=5		K=7	
	Fashion	Motors	Fashion	Motors	Fashion	Motors
Greedy	3.95	3.92	3.81	3.81	3.80	3.77
K-means score	4.10	4.09	4.08	4.06	4.04	4.01
K-means centroid	4.29	4.28	4.17	4.22	4.26	4.23
LexRank	**4.36**	**4.30**	**4.38**	**4.35**	**4.36**	**4.33**

methods were statistically significant, except between LexRank and k-means-centroid with K=3.[4] As for the length of the description, there was no consistent trend, but the highest ratings for LexRank were achieved for K=5 sentences by an insignificant difference from both K=3 and K=7 sentences.

Figure 3 presents the rating results for the additional questions in our evaluation, relating to the different quality aspects. Across all questions, LexRank consistently achieved the best ratings, with the rest of the methods ranked similarly to the general quality question. Among the four questions, ratings were highest for product relevance, with LexRank achieving an average of over 4.5 for all three description lengths. This indicates that as could be expected,

[4]Statistical significance was measured using one-way ANOVA with Tukey post-hoc comparisons for $p<0.01$.

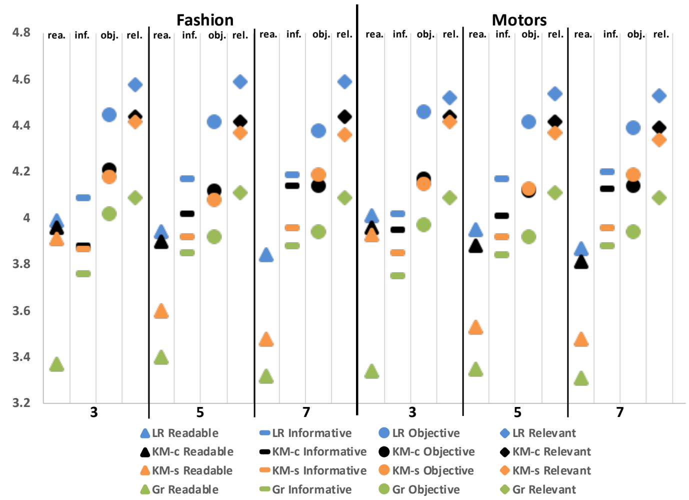

Figure 3: Average rating of four aspects (readable, informative, objective, relevant) for descriptions of length 3, 5, and 7 sentences, generated by four methods: LexRank (LR), K-Means centroid (KM-c), K-Means score (KM-s), and Greedy (Gr).

the reviews serve as a relevant source of information about the product. Objectivity was also rated high in general, indicating our method for filtering the objective parts out of the subjective reviews (step 1) works well. Readability received the lowest rating, with the greedy approach (and for K=5 and K=7 also k-means with highest-score selection) performing especially poorly. Inspecting the ratings by description length, objectivity and readability tend to decrease with description length, while informativeness increases. Product-relevance remains stable regardless of the description length. Overall, it is expected that as the description is longer, it is more likely to contain non-objective parts and become less readable due to connectivity or repetition issues, while it is also likely to contain more details and therefore become more informative. Inspecting the different ratings in Figure 3, descriptions of 5 sentences seem to yield the best trade-off between these two trends. For instance, consider the first example in Table 10. When the generated description with K=3 included the first 3 sentences, its 'informative' rating reduced. On the other hand, when the K=7 version included the extra sentence "*Best socks for sports, especially running*," the 'objective' rating reduced.

6.2.2 Evaluation over public data. The datasets used in this work are proprietary and cannot be shared due to business sensitivity. We therefore set out to get a sense how well our method performs on publicly available data. As a first step, we observe that 13.5% (120 in total) of the Fashion products and 16.7% (135) of the Motors products sent for our end-to-end evaluation are included in a large publicly-released e-commerce dataset [24]. Inspecting the ratings for these product subsets only, the results were almost completely identical to those reported for all products (Table 11 and Figure 3) and are thus not separately reported.

As a second (and final) step, we conducted another evaluation, in which all products in the public dataset [24] with at least 10

Table 12: Average rating of 5-sentence descriptions generated using the LexRank method over public data for Fashion and Motors (upper section) and for Toys and Electronics using cross-domain learning (lower section).

	Overall	Readable	Informative	Objective	Relevant
Fashion	4.34	3.88	4.20	4.40	4.53
Motors	4.32	3.95	4.22	4.43	4.50
Electronics	4.11	3.60	4.05	4.18	4.21
Toys	4.05	3.62	4.09	4.21	4.27

reviews between January and July 2014 (the most recent months in the dataset) were evaluated end-to-end using the LexRank method with k=5 sentences, by 3 professional annotators. This set included 140 Fashion products and 120 Motors products. Results, depicted in the upper section of Table 12, are very similar (±1.5%) to those reported for our own datasets in Table 11 and Figure 3.

6.2.3 Cross-Domain Generation. Finally, we set out to explore if our approach can be used across domains. To this end, we trained a model based on both the Fashion and Motors data and generated descriptions for two additional e-commerce domains: Toys and Electronics. As in the previous experiment, we generated descriptions of 5 sentences using the LexRank method for all products in these domains with at least 10 reviews between January and July 2014, resulting in a total of 55 products for Toys and 75 for Electronics [24]. These descriptions were evaluated by the same 3 annotators and the lower section of Table 12 presents their ratings. It can be seen that while the overall quality ratings are somewhat lower than for Fashion and Motors, they are still just above 4 on average. There is a rather similar rating decrease across all four quality aspects relative to Fashion and Motors, with the sharpest decline being for readability. The last two examples in Table 10 demonstrate generated descriptions for Toys and Electronics.

7 CONCLUSION AND FUTURE WORK

In this paper, we presented a method for automatic generation of product descriptions based on online user reviews. Product descriptions provided by sellers are often missing or lacking. Hence, customers turn to product reviews and often spend a significant amount of time sifting through them before making a purchase. However, users might still find it hard to extract the relevant information from thousands of reviews available online. Moreover, many reviews include personal and subjective opinions, while the users are sometimes only interested in the key product details before purchasing [15].

We first studied the main differences between the reviews and descriptions and adopted a supervised deep multi-task learning approach to identify appropriate review sentences. Afterwards, we introduced a similarity measure between review sentences that helped us eliminate redundancies when creating the descriptions. Finally, we used extractive text summarization to create a coherent and concise product description. We provided an extensive set of experiments that demonstrated our approach is productive.

Our experiments inspected two principally different e-commerce domains: Fashion and Motors. The similar results throughout our evaluation for both domains, as well as the analysis revealing common review language characterizing sentences that can be used for a description, suggest that our approach can be generalized to additional domains. Indeed, cross-domain description generation, trained over a combination of Fashion and Motors reviews, showed promising results when applied to two other primary e-commerce domains: Toys and Electronics. Our evaluation and examples also illustrate that our descriptions are somewhat different than seller-provided descriptions. A presentation of such descriptions on product pages may include an indication they are crowd-based or stem from user reviews. In addition, the descriptions can be presented as a bulleted list, as currently done on some e-commerce websites, to provide an alternative user experience and mitigate readability issues.

For future work, we plan to focus on four main directions. Currently, we select K sentences for the description without any special ordering. We believe that proper ordering may increase the readability of the descriptions. Second, automatically deriving the value of K based on the product and review content can further improve description quality. Third, abstractive approaches [13, 14], which adapt the original review text and combine content from different sentences, are also worth exploring, and can help generate descriptions of even higher quality based on fewer reviews. Finally, personalizing the generated descriptions for the individual consumer can help make them even more compelling.

REFERENCES

[1] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. 2008. The evaluation of sentence similarity measures. In *Proc. of DaWaK*. 305–316.
[2] Mathieu Acher, Anthony Cleve, Gilles Perrouin, Patrick Heymans, Charles Vanbeneden, Philippe Collet, and Philippe Lahire. 2012. On extracting feature models from product descriptions. In *Proc. of VaMoS*. 45–54.
[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proc. of ICLR*.
[4] Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proc. of SIGIR*. 222–229.
[5] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.
[6] Y. Chae, M. Nakazawa, and B. Stenger. 2018. Enhancing product images for click-through rate improvement. In *Proc. of ICIP*. 1428–1432.
[7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proc. of KDD*. 785–794.
[8] 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.
[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[10] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213–220.
[11] Horatiu Dumitru, Marek Gibiec, Negar Hariri, Jane Cleland-Huang, Bamshad Mobasher, Carlos Castro-Herrera, and Mehdi Mirakhorli. 2011. On-demand feature recommendations derived from mining public product descriptions. In *Proc. of ICSE*. 181–190.
[12] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.
[13] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proc. of COLING*. 340–348.
[14] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proc. of EMNLP*. 1602–1613.
[15] Anindya Ghose and Panagiotis G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23, 10 (2011), 1498–1512.
[16] Ross Girshick. 2015. Fast r-cnn. In *Proc. of ICCV*. 1440–1448.
[17] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proc. of ICML*. 513–520.

[18] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *Proc. of SIGIR*. 121–128.

[19] Vishrawas Gopalakrishnan, Suresh Parthasarathy Iyengar, Amit Madaan, Rajeev Rastogi, and Srinivasan Sengamedu. 2012. Matching product titles using web-based enrichment. In *Proc. of CIKM*. 605–614.

[20] Anjan Goswami, Naren Chittar, and Chung H. Sung. 2011. A study on the impact of product images on user clicks for online shopping. In *Proc. of WWW*. 45–46.

[21] 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.

[22] 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.

[23] 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).

[24] 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.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[26] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. of KDD*. 168–177.

[27] Nan Hu, Paul A. Pavlou, and Jennifer Zhang. 2006. Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication. In *Proc. of EC*. 324–330.

[28] Alice Jiang, Zhilin Yang, and Minjoon Jun. 2013. Measuring consumer perceptions of online shopping convenience. *Journal of Service Management* 24, 2 (2013), 191–214.

[29] Gagandeep Kaur and Gagandeep Kaur. 2016. Mobile applications are major players in the world of e-commerce. *International Journal of Advanced Research in IT and Engineering* 5, 2 (2016), 13–21.

[30] Zehra Kavasoğlu and Şule Gündüz Öğüdücü. 2013. Personalized summarization of customer reviews based on user's browsing history. *IADIS International Journal on Computer Science & Information Systems* 8, 2 (2013), 147–158.

[31] H. Khalid, E. Shihab, M. Nagappan, and A.E. Hassan. 2015. What do mobile app users complain about? *IEEE Software* 32, 3 (2015), 70–77.

[32] Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and Chengxiang Zhai. 2011. Comprehensive review of opinion summarization. *UIUC Technical Report* (2011).

[33] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint* abs/1412.6980 (2014).

[34] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification.. In *Proc. of AAAI*. 2267–2273.

[35] Eun-Ju Lee and Soo Yun Shin. 2014. When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo. *Comput. Hum. Behav.* 31 (2014), 356–366.

[36] Beibei Li, Anindya Ghose, and Panagiotis G Ipeirotis. 2011. Towards a theory model for product search. In *Proc. of WWW*. 327–336.

[37] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proc. of COLING*. 653–661.

[38] Xinxin Li and Lorin M Hitt. 2008. Self-selection and information role of online product reviews. *Information Systems Research* 19, 4 (2008), 456–474.

[39] Moez Limayem, Mohamed Khalifa, and A. Frini. 2000. What makes consumers buy from Internet? A longitudinal study of online shopping. *IEEE Trans. Systems, Man, and Cybernetics, Part A* 30, 4 (2000), 421–432.

[40] Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization: A prototype system and its evaluation. In *Proc. of ACL*. 457–464.

[41] Bing Liu. 2012. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[42] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Proc. of IJCAI*. 1291–1297.

[43] Roque Enrique Lpez Condori and Thiago Alexandre Salgueiro Pardo. 2017. Opinion summarization methods. *Expert Syst. Appl.* 78, C (July 2017), 124–134.

[44] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint* abs/1508.04025 (2015).

[45] Duy Khang Ly, Kazunari Sugiyama, Ziheng Lin, and Min-Yen Kan. 2011. Product review summarization from a deeper perspective. In *Proc. of JCDL*. 311–314.

[46] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proc. of IJCAI*. 4068–4074.

[47] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of Berkeley symposium on mathematical statistics and probability*, Vol. 1. 281–297.

[48] Deborah Brown McCabe and Stephen M Nowlis. 2003. The effect of examining actual products or product descriptions on consumer preference. *Journal of Consumer Psychology* 13, 4 (2003), 431–439.

[49] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint* abs/1301.37810 (2013).

[50] Hye-Jin Min and Jong C Park. 2012. Identifying helpful reviews based on customer's mentions about experiences. *Expert Systems with Applications* 39, 15 (2012), 11830–11838.

[51] Samaneh Moghaddam and Martin Ester. 2012. On the design of LDA models for aspect-based opinion mining. In *Proc. of CIKM*. 803–812.

[52] Ajinkya More. 2016. Attribute extraction from product titles in eCommerce. *arXiv preprint* abs/1608.04670 (2016).

[53] Quang Nguyen. 2012. *Detecting experience revealing sentences in product reviews*. Ph.D. Dissertation. University of Amsterdam.

[54] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135.

[55] Eun Joo Park, Eun Young Kim, Venessa Martin Funches, and William Foxx. 2012. Apparel product attributes, web browsing, and e-impulse buying on shopping websites. *Journal of Business Research* 65, 11 (2012), 1583–1589.

[56] A.M. Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. of HLT*. 339–346.

[57] Katharina Probst, Rayid Ghani, Marko Krema, Andrew Fano, and Yan Liu. 2007. Semi-supervised learning of attribute-value pairs from product descriptions. In *Proc. of IJCAI*. 2838–2843.

[58] Reid Pryzant, Young-Joo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *Proc. of ECOM (SIGIR Workshops)*.

[59] Pradeep Racherla, Munir Mandviwalla, and Daniel J Connolly. 2012. Factors affecting consumers' trust in online product reviews. *Journal of Consumer Behaviour* 11, 2 (2012), 94–104.

[60] 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.

[61] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint* abs/1706.05098 (2017).

[62] Keiji Shinzato and Satoshi Sekine. 2013. Unsupervised extraction of attributes and their values from product description. In *Proc. of ACL*. 1339–1347.

[63] 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.

[64] Krysta M. Svore, Lucy Vanderwende, and Chris J.C. Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proc. of EMNLP-CoNLL*.

[65] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.

[66] Damir Vandic, Flavius Frasincar, and Uzay Kaymak. 2018. A framework for product description classification in e-commerce. *Journal of Web Engineering* 17, 1&2 (2018), 001–027.

[67] Shuai Wang, Zhiyuan Chen, and Bing Liu. 2016. Mining aspect-specific opinion using a holistic lifelong topic model. In *Proc. of WWW*. 167–176.

[68] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proc. of AAAI*. 3316–3322.

[69] 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.

[70] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong. 2011. Analysis of adjective-noun word pair extraction methods for online review summarization. In *Proc. of IJACI*. 2771–2776.

[71] Naitong Yu, Minlie Huang, Yuanyuan Shi, et al. 2016. Product review summarization by exploiting phrase properties. In *Proc. of COLING*. 1113–1124.

[72] Di Zhu, Theodoros Lappas, and Juheng Zhang. 2018. Unsupervised tip-mining from customer reviews. *Decision Support Systems* 107 (2018), 116–124.

[73] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proc. of CIKM*. 43–50.