2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Aspect-Based Helpfulness Prediction for Online Product Reviews

Yinfei Yang
Redfin, Inc.
Seattle, WA 98121
yangyin7@gmail.com

Cen Chen
School of Information Systems
Singapore Management University
Singapore, 188065
cenchen.2012@phdis.smu.edu.sg

Forrest Sheng Bao
University of Akron
Akron, OH 44325
forrest.bao@gmail.com

***Abstract*—Product reviews greatly influence purchase decisions in online shopping. A common burden of online shopping is that consumers have to search for the right answers through massive reviews, especially on popular products. Hence, estimating and predicting the helpfulness of reviews become important tasks to directly improve shopping experience. In this paper, we propose a new approach to helpfulness prediction by leveraging aspect analysis of reviews. Our hypothesis is that a helpful review will cover many aspects of a product at different emphasis levels. The first step to tackle this problem is to extract proper aspects. Because related products share common aspects to different degrees, we propose an aspect extraction model making use of product category information to balance the aspects of a general category and those of subcategories under it. On top of this model, a two-layer regressor is trained for helpfulness prediction. Experiment results show that we can improve helpfulness prediction by 7% than the baseline on 5 popular product categories from Amazon.com.**

I. Introduction

The Internet is changing the way that people do many things, including shopping. E-commerce provides a whole new way for shopping that consumers can post their reviews about products and others can make purchase decisions based on reviews of products of their interests. Due to the vast amount of reviews for each product, reviews need to be ranked, recommended and/or selectively presented to consumers. Hence, predicting the helpfulness of reviews is an important task to improve shopping experience and promote sales [1]–[4]. The simple question "Was this review helpful to you?" increases an estimated $2.7B revenue to Amazon.com annually[1].

We pioneered the semantic approach to review helpfulness prediction, hypothesizing that helpfulness is an underlying feature of the text itself [5]. Leveraging semantic features, our result shows that helpful reviews reflect more cognitive and psychological processes of the thoughts embedded in the reviews. By using semantic analysis, not only can we answer the question *what* makes a review helpful, but also the question *how*.

In this paper, we answer the two questions from a new angle, by focusing on the aspect coverage of reviews. We hypothesize that the helpfulness of a review text is strongly related to *what* and *how* aspects of the product are mentioned. This hypothesis is based on our observation that a helpful review will cover multiple aspects of a product, with different sentiments, e.g., a laptop can have light weight but poor battery life.

Aspect analysis from online reviews has been addressed in previous work [6], often with the application of sentiment analysis [7]–[9] or summarization [10], [11]. In this paper, we expand its application to helpfulness prediction. Extracting various aspects of products from consumer feedback allows us to provide more targeted helpfulness prediction. For example, to customers caring about services, reviews focusing on the service aspects are more helpful than reviews focusing on other aspects, such as quality.

A challenge for aspect-based review helpfulness analysis is extracting the right aspects for various products. The aspects need to be general in order to build a transferrable helpfulness prediction model. Previous work on aspect extraction is usually on the product/item level or fine-grained category level. The item-level aspects are suitable for sentiment analysis and summarization because different products can have very distinct aspects which are directly presented to the customers. But in the context of review helpfulness analysis, item-level aspect extraction is not scalable as it is impractical to build product-tailored models for the vast types of products and many products do not have enough reviews for aspect extraction, i.e., the cold-start problem [12].

Therefore, we focus on the category-level aspect extraction by making use of the category relationship (e.g.., "gaming laptop" is a subcategory of "electronics") and assume an aspect distribution for each product category. Our inspiration is that reviews of all subcategories will talk about something common among them, along with something specific for each of them. For example, all laptop computers have the aspects "keyboards" and "portability," but reviews of gaming laptops will talk more about keyboards while those of mobile workstations will talk more about portability. A topic generative model (topic modeling) inspired from Twitter-LDA [13] is used in this paper to extract aspects, where a user in Twitter-LDA is modeled as a product category.

After extracting aspects, a two-layer regression model is built to predict helpfulness score. In order to increase the transferability of the regression model, we associate the aspects from topic modeling to 8 high-level aspects, namely, *brand,*

[1] http://www.uie.com/articles/magicbehindamazon/

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.127

IEEE computer society

appearance, functionality, price, quality, usability, service, and others. Presumably, each sentence, the building block of a review, talks about one aspect. We then build a set of regressors (Layer 1), each of which predicts the helpfulness on one high-level aspect or a special label *multiple* – for sentences talking about multiple high-level aspects, using the aspect features along with four other features, STR, UGR, LIWC, and INQUIRER. Each review is assigned with a high-level aspect label if and only if the majority of its sentences are labeled of this aspect. Finally, the outputs from all Layer-1 regressors are combined to estimate the final/overall helpfulness score, forming the Layer-2 regressor.

The purpose of building one regressor for each aspect (Layer-1) and then combining their results (Layer-2) is due to our hypothesis that a helpful review often covers many aspects and the final voting is based on the helpfulness for all aspects. Another motivation of this ensemble approach is to boost the performance. Experimental results show that our approach can improve the review helpfulness prediction by 7.27% in terms of correlation coefficient than the baseline [5], on 5 popular product categories from Amazon.com Review Dataset [14].

The rest of the paper is organized as follows. Section II briefly overviews the related work. Our methodology is explained in Section III. Experimental result is presented in Section IV and Section V concludes the paper.

II. Related Work

The task of estimating and predicting text helpfulness has been addressed in several papers [1], [2], [4], [5], [15]. [15] used regression to predict the helpfulness ranking of product reviews based on various classes of text and non-text features. The models were further improved by measuring the readability, subjectivity and emotion [2], [4]. [5] introduced two semantic features to help understand helpful reviews and demonstrated a model built purely from text that is transferable between different product categories. In this paper, we solve the task at a deeper level by understanding the content of reviews. To our best knowledge, this work is the first model considering product aspects into helpfulness prediction.

The first part of our work is aspect extraction using topic modeling, which has been intensively studied recently [13], [16], [17]. The drawback of applying original LDA directly is that many discovered topics are not user ratable aspects [16] as they are not understandable or meaningful to users [17]. Recent studies on topic modeling suggest several improved approaches. [16] proposes a multi-grain LDA (MG-LDA) introducing global and local topics. They argue that the discovered global topics are global properties of an object and local topics are user ratable aspects. The proposed model improves the aspect extraction on the general category level by integrating the product category information.

Furthermore, we search for the answer to the question what makes a review helpful from the content level. Early works tried to understand this question from the decision-making perspective [1], [18], where a helpful review was defined as a peer-generated product evaluation that could facilitate the consumers' purchase decision process. In particular, [1] revealed that review rating, product type, word counts are three important factors for a helpful review and modeled the helpfulness score as a combination of these three factors. [18] studied the option evaluation from Amazon.com book reviews and found that the final helpfulness voting for a review is affected by many external factors, such as the ratio of its helpfulness score to those of other reviews. Our model understands the content of reviews more thoroughly by integrating aspects information into helpfulness estimation, enabling us to study what information the review writers want to convey through text and how.

A task similar to our work is review ranking. [19] and [20] modeled the helpfulness ranking problem as part of a recommendation system. Then the recommendation algorithm, like collaborative filtering, can be applied to predict the helpfulness. We think our work should be a part of the ranking/recommendation system by providing the helpfulness scores from text. The prediction of helpfulness score from text can be an ideal input of a review recommendation system, which combines the helpfulness score and other features to make the final recommendation.

III. Methodology

In this section, we first introduce a topic generative model that leverages product category information to extract product aspects. Topics generated in this way will align the best with the aspects in our commonsense. Therefore, a topic can represent an aspect. Then we propose a two-layer regressor to predict the review helpfulness based on the topic modeling.

A. The Generative Model with Categorical Information

Most aspect extraction methods in the literature [7], [16], [17], [21] work well for fine-grained product categories or item products, like MP3 players, cameras, etc. However, we want to build a model can work on relatively larger and/or general categories, such as electronics, or home tools, each of which covering many subcategories. Inspired by Twitter-LDA [13], we introduce a new generative model to extract aspects from reviews, by leveraging product category information.

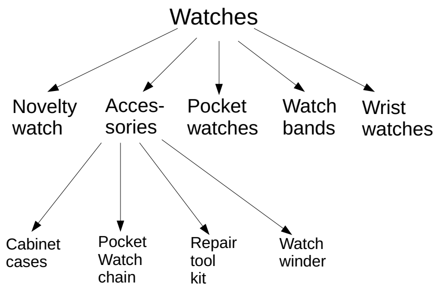

Fig. 1. Product tree for the category Watches. In this paper we treat all nodes of a common ancestor (e.g., Watches) as subcategories of the ancestor. Details about this flattened tree is to be discussed later.

When consumers write reviews, they do not write for a general category but product items belonging to different subcategories under the general category. The relationship between subcategories and general categories can be acquired by many means, such as the catalogs of online shopping websites. An example for the product category "watches" are given in Fig. 1. A set of topics is assumed for one general category. Its subcategories have different emphases across those topics. Such a generative process is represented in the graphical model in Fig. 2.

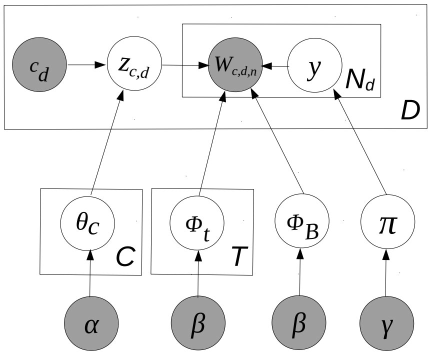

Fig. 2. A graphical model representation of review generation.

For each general product category, we assume that it has a set T of topics where aspects are associated with. Like in standard topic modeling, each topic t is a distribution over words, denoted as ϕ_t and $\sum_{t\in T}\phi_t = 1$. There is a general background topic shared by all categories, whose distribution is denoted as ϕ_B. The choice between background words and topic words follow Bernoulli distribution π. Assuming that the general category has C subcategories, in the second level of our generative model, each subcategory has a distribution over topics. Let θ_c be the topic distribution of the fine category c such that $\sum\theta_c = 1$. The distribution of topic words ϕ_t over all topics $t \in T$ and background words ϕ_B, the topic distribution θ_c over all fine categories $c \in C$, and the coupling π between topic words and background words are generated from Dirichlet priors with hyperparameters α, β, and γ respectively. The hyperparameter c_d controls category selection.

When a sentence is being generated, given the product category it first picks a topic based on the topic distribution and then generates the sentence following word distribution in this topic, along with some background words. Some subcategories may have higher chances to talk about certain topics while the other subcategories mention other topics more often. The generation process is detailed as follows:

1) For every category $c \in C$, choose $\theta_c \sim$ Dirichlet(α)
2) Choose $\phi_B \sim$ Dirichlet(β) and $\pi \sim$ Dirichlet(γ)
3) For every topic $t \in T$, choose $\phi_t \sim$ Dirichlet(β)
4) For each sentence (a document) $d \in D$,
   a) Get category c_d from meta data and choose a topic $z_{c,d} \sim$ Multinomial(θ_c)
   b) For each word $n \in [1..N_d]$,
      Choose a balance $y \sim$ Bernoulli(π)
      if $y = 1$, choose a topic word $w_{c,d,n} \sim$ Multinomial($\phi_{z_{c,d}}$); else choose a background word $w_{c,d} \sim$ Multinomial(ϕ_B)

where N_d means the number of words in document d. Each multinomial distribution is governed by some symmetric Dirichlet distribution. Gibbs sampling is used to perform model inference. We collapse out all the $\theta_{c(\cdot)}$, $\phi_{t(\cdot)}$, and $\phi_{B(\cdot)}$. Let τ be the set of hyperparameters $\{\alpha, \beta, \gamma\}$. We sample each document d's n-th aspect label z as:

$$p(z_{d,n} = t \mid Z_{d,\neg n}, Y, W, \tau)$$
$$\propto \frac{n^t + \alpha}{n^{\cdot} + T\alpha} \cdot \frac{\prod_w \prod_{p=0}^{n_{d,n}^w - 1}(n_w^{t,y=1} + \beta + p)}{\prod_{q=0}^{n_{d,n}^{\cdot} - 1}(n_{\cdot}^{t,y=1} + V\beta + q)}.$$

Similarly, we sample $y_{d,n}$ as:

$$p(y_{c,d,n} = y \mid Y_{\neg c,d,n}, Z, W, \tau)$$
$$\propto \frac{n_y + \gamma}{n_{\cdot} + 2\gamma} \cdot \left[\frac{n_w^{t,y=1} + \beta}{n_{\cdot}^{t,y=1} + V\beta}\right]^{y=1} \cdot \left[\frac{n_w^{y=0} + \gamma}{n_{\cdot}^{y=0} + V\gamma}\right]^{y=0}.$$

For new reviews, we apply the trained model on each sentence, compute the most liked topic, and assign the topic to the sentence. Again, topics generated in this way will align the best with the aspects in our commonsense. Therefore, a topic can represent an aspect, referred to as a *low-level aspect* in this paper.

Note that our model is adapted from Twitter-LDA [13]. A category in our model corresponds to a user in Twitter-LDA, and reviews of the category corresponds to a document from the user.

B. Aspect-Based Helpfulness Prediction

A generative model to detect product aspects at sentence level is described above. Here we discuss how to use the aspect information for review helpfulness prediction.

We hypothesize that people have different criteria when rating helpfulness for different aspects. Therefore, we propose an aspect-based two-layer model for review helpfulness prediction. In each review, we first extract the aspects for each of its sentences. If an aspect is assigned to more than half of the sentences in this review, we assign the aspect to this review. Otherwise we assign a special aspect *multiple* to this review, meaning that it covers multiple aspects. In this fashion, we form a set of reviews for each aspect and the special aspect *multiple*.

To predict the helpfulness, we first train a regressor (Layer 1) for each aspect individually, using features to be mentioned below. Then another regressor (Layer 2) is trained, whose features are the outputs from the Layer-1 regressors. The intuition behind this 2-layer architecture is that the overall helpfulness of a review is a combination of the helpfulness on different aspects. In the experiment, the corpus is split into training set, development set and testing set. The Layer-1 classifiers are trained on the training set while the Layer-2 classifier is trained on the development set.

Inspired by previous work, we select the Structure (STR) [15], [22], Unigram (UGR) [15], [20], [22] and Semantic dictionary [5] as features. We also introduce a feature, short as ASP, based on the aspects detected from our generative model.

a) STR: Following the [22], we use the following structural features: total number of tokens, total number of sentences, average length of sentences, number of exclamation marks, and the percentage of question sentences.

b) UGR: Unigram feature has been demonstrated as a very reliable feature for review helpfulness prediction in previous work. We build a vocabulary with all stopwords and non-frequent words ($df < 3$) removed. Each review is represented by the vocabulary with $tf - idf$ weighting for each appeared term.

c) LIWC (Linguistic Inquiry and Word Count): [23] is a dictionary which helps users to determine the degree that any text uses positive or negative emotions, self-references and other language dimensions. Each word in LIWC is assigned 1 or 0 for each language dimension. For each review, we sum the values of all words for each dimension. Eventually each review is represented by a histogram of language dimensions. We employ the LIWC2007 English dictionary which contains 4,553 words with 64 dimensions in our experiments.

d) INQUIRER: General Inquirer [24] is a dictionary in which words are grouped in categories. It is basically a mapping tool which maps each word to some semantic tags, e.g., *absurd* is mapped to tags NEG and VICE. The dictionary contains 182 categories and a total of 7,444 words. Like for LIWC representation, we compute the histogram of categories for each review.

e) ASP: For each review, we count the number of sentences assigned to each aspect, and simply concatenate the numbers of all aspects to form the feature vector. The dimension of the feature vector is equal to the number of topics ($|T|$) in the model proposed in Section III-A. The aspect feature can assist helpfulness prediction in two folds. First, a word or writing style may appear in a positive review about one aspect and also in a negative review about another aspect. In order to capture this, we need to use aspect features in conjunction with other features such as UGR and STR. Second, aspects contribute unequally to the helpfulness, as to be discussed later. For example, a review about shipping or return is unlikely to be considered helpful at Amazon.com. Introducing the aspect feature gives us a new modality to predict helpfulness.

C. High-Level Aspects

There are several drawbacks of using the low-level topic-represented aspects extracted according to Section III-A directly. First, the data size for each aspect is relative small so some of the topics will not get enough training data. Second, topics generated by LDA-like methods are usually subjective and product-dependent. A predictive model built using such topics is less transferable.

To address these problems, we group the low-level aspects into high-level aspects. We argue that the high-level aspects are semantically meaningful and can be product-independent in most cases. Results later (Table I) will show that different categories of products have different focuses on these high-level aspects. The 8 high-level aspects are:

1) **Brand**: discussing the brands
2) **Appearance**: anything people can perceive about the product without using.
3) **Functionality**: what the product can do
4) **Price**: anything related to cost
5) **Quality**: how well the product perform its functions
6) **Usability**: how well can users interact with the product
7) **Service**: anything between the users and the providers
8) **Other**.

Any low-level aspect should belong to one of these high-level aspects. For example, those about shipping/return belong to *Service* while those about color/size belong to *Appearance*. One low-level aspect may correspond to many high-level aspects, but here we only associate it with the most dominant one for simplification.

IV. Experiments

A. Data and Setup

The dataset used in this paper is constructed from Amazon.com Review Dataset [14] which includes nearly 35 million reviews from Amazon.com between 1995 and 2013. Category information for each product is also available in this dataset. Reviews from 2 fine-grained categories, Watches and Cellphones, and 3 coarse-grained categories, Home & Kitchen (*Home* for simplification), Outdoor and, Electronics are used as the corpora. These 5 categories are treated as general categories.

For each general category, the product tree (e.g., Fig. 1) is flattened into two layers, where the root is the general category itself and the rest of the product tree under it are treated as its subcategories becoming the children of the root. The number of subcategories under each general category is given in Column 2 of Table II.

The setup for the generative model is described as follows. For each fine-grained category, Watches or Cellphones, 30 topics are generated. For each coarse-grained category, because of the relatively more diverse subcategories, 100 topics are generated. We run 1000 iterations of Gibbs sampling for all categories.

The helpfulness prediction, defined as predicting helpfulness scores, is modeled as a regression problem. The scores can be extracted using the "X of Y approach" from the votes of users (i.e., "X of Y users think this review is helpful") that come with the dataset. Following the previous work, only the reviews with at least 5 helpfulness votes are used. We use SVM regressor of RBF kernel provided by LibSVM [25] in both of the first layer and the second layer. Performance is evaluated by correlation coefficients.

The dataset are split into training set, development set and test set with the ratio 5:4:1. Ten-fold cross-validation is performed for all experiments.

TABLE I
TOP ASPECTS AND TOPIC WORDS FOR EACH CATEGORY

Category	High-Level Label	Aspect Label	Top Words
Watches	Service	Shipping	product, amazon, item, received, return, service, order, seller, shipping, customer...
		Warranty	amazon, warranty, return, service, shipping, repair, customer, refund, seller, replacement...
	Function	Accuracy	accurate, time, day, seconds, days, hours, week, month, fast, stopped...
		Display	light, face, read, easy, numbers, digital, bright, day, glow, time...
	Price	Price	price, bought, buy, paid, cost, dollars, purchased, expensive, bucks, money...
Cellphone	Service	Shipping	product, amazon, item, received, back, return, order, seller, shipping, battery...
		Plan	phone, service, plan, month, year, minutes, contract, free, cingular, t-mobile...
	Function	Functions	phone, email, web, internet, text, games, messaging, access, features, camera...
		Bluetooth	bluetooth, headset, phone, easy, headphones, music, work, pair, sound, pairs...
	Usability	Case	case, phone, belt, fit, fits, leather, cover, plastic, screen, pocket...
		Signal	signal, anetenna, bars, phone, reception, cell, service, inside, house, unit...
Home	Service	Shipping	amazon, back, return, shipping, replacement, send, item, refund, company, order...
		Packaging	box, shipping, ordered, broken, item, packaging, shipped, damaged, opened, packed...
	Function	Power	power, cord, unit, switch, plug, turn, short, outlet, on/off, electrical...
		Cooking	cooking, cook, pan, pot, great, chicken, size, large, pasta, oven ...
	Quality	Quality 1	made, quality, product, good, design, cheap, sturdy, plastic, poor, heavy...
		Quality 2	quality, disappointed, satisfied, beat, fragile, thin, broken, break, sturdy, hand...
Outdoor	Service	Shipping	arrived, shipping, product, fast, received, item, time, delivery, shipped, ordered...
		Replacement	service, customer, amazon, back, return, called, send, replacement, refund, returned...
	Function	DVD	dvd, workouts, video, dvds, videos, fun, play, back, time, move...
		Pocket	pocket, bag, pack, carry, small, pockets, fit, large, great, hold...
	Other	Rating	stars, 5, give, star, 4, rating, reason, product, review, 3...
		Gift	bought, gift, christmas, loves, son, husband, loved, birthday, year, present...
Electronics	Service	Shipping	amazon, shipping, received, arrived, item, product, return, days, order, time...
		Support	support, customer, tech, service, call, phone, called, problem, email, back...
	Function	Portability	port, ports, plug, switch, work, works, connect, plugged, drive, connection...
		Recording	recording, camcorder, video, camera, recorder, digital, record, sound, light, tape...
	Usability	Installation	easy, manual, simple, set, user, instructions, read, setup, install, friendly...
		Mouse	mouse, logitech, trackball, buttons, hand, wheel, wireless, scroll, optical, ball...

TABLE II
THE 5 CATEGORIES USED IN EXPERIMENTS

general category	number of sub-categories under	number of topics	number of reviews	number of reviews with at least 5 votes
Watches	10	30	68,356	9,737
Cellphones	33	30	78,930	18,542
Home	1214	100	991,784	219,310
Outdoor	1880	100	510,991	72,796
Electronics	674	100	1,241,778	354,301

B. *Qualitative Results*

1) Extracted Aspects from the Generative Model: Typical topics for all of five categories are given in Table I. For each category, we show 3 high-level aspects along with their corresponding low-level aspects. The top 10 topic words for each low-level aspect are also listed. For better presentation, we also assign an aspect label for each low-level aspect.

Service and Function aspects are listed for all categories. Interestingly, different categories share many service aspects. For example, shipping is an important aspect for all categories. The topics on shipping of both categories share many topic words, e.g., amazon, shipping, received, return, etc. This indicates that people talk about shipping experience a lot when shopping online. However, an analysis later shows that service/shipping/delivery experience in reviews at Amazon.com is not very helpful to customers. Warranty, Support, and Replacement are three other commonly covered aspects across categories.

In contract, Functionality-related low-level aspects are quite different across categories. Hardly do two categories share many function aspects. For example, Accuracy only appears in Watches while Mouse function is only applicable to Electronics. Similarly, it is hard to imagine that Pocket can be an function of Cellphone or Electronics. Although people usually think that there are many common aspects between Cellphone and Electronics, we observe that each category has its unique aspects. For example, Electronics has *Portability*, *Recording*, and *OS*. while Cellphone has *Plan* and *Signal*.

We also present some examples in other aspects, including Price, Quality, Usability and Others. The top words of these aspects are all aligned well with our commonsense. It is worth noting that the rating aspect is shared by all five categories and a review is usually helpful once it addressed this aspect. Because it is not the main focus of this paper, we leave more detailed analysis and improvement as future work.

2) Aspect-Based Helpfulness Estimation: The results above show that we can extract meaningful aspects from reviews. We, however, care more about how the helpfulness differs on these aspects.

Tables III and IV show example reviews from high-level aspects with polarities of predicted helpfulness scores. A positive review example is drawn from top 1 percentile based on predicted helpfulness score while a negative review example from below 50 percentile. Example reviews are selected from Electronics and Outdoor categories. Two reviews, a more helpful one and a less helpful one, for each of the high-level aspects are listed. Only aspect-related text is presented for readability. As hypothesized, helpful reviews tend to be informative about the products, or discussing about details and user experience of the product, while the less helpful reviews are usually simple and convey little information about the product.

Although the example reviews are selected from specific categories, we can see the text is usually not category specific. Brand and Functionality are two exceptions, also aligning well with our commonsense. The Brands are usually quite different in different categories of products and some Functions are even specific for products.

TABLE III
SAMPLE REVIEWS FROM ELECTRONICS.

Aspect	Review	Polarity
Brand	... i think i had all 6 problems with this product , using 3 different phones, apple 3gs , blackberry 8320 , and now a samsung android phone that simply i can not connect.	positive (helpful)
	... these are as good, if not better than the standard ear buds that come with apple products ...	negative (unhelpful)
Functionality	... i also found a problem with one of the line-in recording features where unit is supposed to start recording only when sound is detected. ...	positive
	... Recording is easy to do and good quality as well.	negative
Price	... but the picture quality is just as good as on an expensive cable and the price is much better. ...	positive
	... very reasonably priced. ...	negative
Usability	... according to the instruction, plugging the camera in or out at the wrong time can crash the operating system. ...	positive
	... I did buy this unit, installed it and it worked flawlessly since. ...	negative

C. Quantitative Results

In order to quantitatively evaluate the proposed model, we compute the correlation coefficients between the predicted helpfulness score and that comes with the dataset. The baseline for performance comparison is the first semantic analysis to review helpfulness [5], which is a single-layer fusion of the 4 features, namely STR, UGR, LIWC, and INQUIRER. We also propose a fusion of all 5 features, which are the 4 features above plus newly introduced aspect (ASP) feature, denoted as $Fusion_{all}$. In other words, $Fusion_{all}$ is this work without using the 2-layer regressor. We also evaluate the models trained by using each feature individually.

1) On all reviews: The correlation coefficients of regression results are given in Table V. Each row corresponds to the model trained by a feature or a fusion of features, while each column corresponds to one general product category.

TABLE IV
SAMPLE REVIEWS FROM OUTDOOR

Aspect	Review	Polarity
Appearance	... the units are well made with a good white painted surface on the bracket. ...	positive (helpful)
	... my teak looks beautiful. ...	negative (unhelpful)
Quality	... I've been looking for a high quality , low-cost rain suit and was very happy to finally locate one on amazon and the material seems strong. ...	positive
	... he was impressed by the quality and feel of the watch. ...	negative
Service	... the shipping was ultra-fast , even though the vendor , googles and glasses didn't have my correct address. ...	positive
	... mine came from sunglass express, fulfilled by amazon , with free shipping ...	negative
Other	... i wish i had listened to the other people who complained about this problem. ...	positive
	...i haven't read other people's reviews lately , part of me wonders if there is something i'm not doing that they need that my own goggles don't ? ...	negative

Like in previous research [5], [15], INQUIRER is the best feature, leading in 3 of the all 5 categories, with the correlation coefficients of 0.506 on Cellphone, 0.366 on Home and 0.419 on Outdoor, respectively. UGR, LIWC and ASP are also very strong features and perform very closely to the INQUIRER feature. The UGR and ASP achieve the best performance on Watches and Electronics, with the correlation coefficients of 0.425 on Watches and 0.406 on Electronics, respectively. STR is not as good as other features but the structure information is complementary to other features.

The baseline shows better performance than using any individual feature on all 5 categories, with the highest correlation coefficient of 0.497 for Outdoor. $Fusion_{all}$, by adding the ASP feature on top of the baseline, increases the correlation coefficients on all 5 categories. Therefore, considering aspect coverage can improve helpfulness prediction.

The proposed model, in the end, achieves the best performance in all product categories, with the highest correlation coefficient of 0.576 for Cellphones. The effectiveness of ensembling regressors on different aspects has been validated (comparing the result of this work with the result of $Fusion_{all}$).

2) On High-Confidence Reviews: In the results above, the proposed approach outperforms the baseline systems. However the highest correlation coefficient 0.576 makes it less exciting. A reasonable explanation is that the helpfulness scores from reviews of only a few helpfulness votes are noise rather than signals. To validate this speculation, we evaluate the trained model on the reviews of at least 30 helpfulness votes. For convenience, those reviews are called *high-confidence* reviews. We use the same 10-fold cross validation setting in previous experiment, but only keep those reviews with 30+ votes when

TABLE V
CORRELATION COEFFICIENTS (THE HIGHER THE BETTER). ALL CORRELATIONS ARE HIGHLY SIGNIFICANT, WITH $p < 0.001$.

	Watches	Cellphones	Home	Outdoor	Electronics
STR	0.276	0.349	0.222	0.277	0.338
UGR	0.425	0.466	0.309	0.412	0.355
LIWC	0.378	0.464	0.331	0.382	0.400
INQUIRER	0.403	0.506	0.366	0.419	0.405
ASP	0.406	0.437	0.283	0.385	0.406
Baseline	0.488	0.539	0.432	0.497	0.484
$Fusion_{all}$	0.493	0.550	0.436	0.501	0.491
This work	0.518	0.576	0.475	0.527	0.526

testing. The threshold 30 is set to guarantee at least 500 test reviews for each product category. The numbers of test reviews for each category are listed in Table VI.

TABLE VI
THE NUMBER OF HIGH-CONFIDENCE REVIEWS

general category	number of reviews with at least 30 votes
Watches	648
Cellphones	1,576
Home	25,906
Outdoor	7,344
Electronics	44,413

Table VII shows the correlation coefficients for higher-confident reviews. Similar to Table V, rows correspond to models trained with different features while columns correspond to product categories. The correlation coefficients increase significantly in all cells of Table VII, compared to those in Table V. This verifies our explanation that low-vote reviews are noisy, echoing the discovery in [5] that the "X of Y approach" may not be a good estimation to helpfulness. After removing the noise, helpfulness scores can be accurately predicted. In particular, the correlation coefficient reaches 0.701 for Cellphones when using our approach. Our approach shows 7.27% improvement on average than the baseline, in terms of correlation coefficient.

TABLE VII
CORRELATION COEFFICIENTS (THE HIGHER THE BETTER) FOR HIGH-CONFIDENT REVIEWS (30+ VOTES). ALL CORRELATIONS ARE HIGHLY SIGNIFICANT, WITH $p < 0.001$.

	Watches	Cellphones	Home	Outdoor	Electronics
STR	0.363	0.444	0.251	0.312	0.383
UGR	0.438	0.575	0.306	0.420	0.390
LIWC	0.473	0.549	0.397	0.428	0.460
INQUIRER	0.493	0.598	0.431	0.491	0.479
ASP	0.447	0.616	0.307	0.407	0.487
Baseline	0.533	0.665	0.487	0.530	0.560
$Fusion_{all}$	0.540	0.687	0.491	0.544	0.567
This work	0.544	0.701	0.556	0.575	0.615

D. Aspect Bias of Helpful Reviews

It is very promising that the ASP feature is competitive compared with UGR, LIWC and INQUIRER. In single-feature prediction, ASP is the best feature for Electronics and the second best for Watches. Understanding the reviews is in deed very important to predicting helpful reviews.

We observe several aspects that consumers pay a lot of attention to by calculating the correlation coefficients between each aspect and the helpfulness score. According to the correlation coefficients, customers usually feel that reviews about product functionality are helpful regardless of the category, e.g., image quality for cameras, accessories for electronics, etc. However, the specificity of functionality aspects makes this group of features less transferable.

Reviews about usability and quality are two other types of reviews that customers usually feel helpful. Compared with functionality aspects, aspects about usability and quality are more general and transferable. But certain categories still have strong biases on these aspects, e.g. the usability for outdoor products and the quality for watches.

Rating is another low-level aspect common for all 5 categories. We find that it performs strongly regardless of the product category, indicating that customers like to read reviews by other customers and how other consumers react to such reviews.

There are also certain aspects that people cannot easily appreciate from reviews, such as service. For example, Amazon.com's shipping and return services are what customers already know about. Hence, reviews about them do not increase customers' knowledge and are less helpful. As another example, the signal reception of T-mobile or AT&T is unlikely to catch consumers' eyeballs because of its geographical dependency. Such a review by a New Yorker is not helpful to a person in San Francisco.

We also notice that reviews covering more aspects tend to be more helpful than reviews covering only a few aspects. Our explanation is that people like to research on different aspects of products before purchasing. This observation also aligns well with the fact that comprehensive reviews usually receive the most helpful votes at Amazon.com. It might be a strong motivation behind reading reviews and the reason why online shopping is so attractive.

V. CONCLUSION

In this paper, we propose an aspect-based approach to review helpfulness prediction. Our hypothesis is that a helpful review will cover many aspects of a product. A topic modeling based approach is introduced to extract aspects. Our generative model assumes that related products share similar topics but with different distributions on them. Then, in a two-layer fashion, we build a regression model which first predicts the helpfulness scores on all aspects and then ensembles them into final helpfulness score. Experimental results show that our approach can increase the performance by over 7% in terms of correlation coefficient. Further analysis shows that customers appreciate reviews covering many aspects, but coverage on certain aspects does not increase the knowledge for readers and hence does not contribute to helpfulness.

REFERENCES

[1] S. M. Mudambi and D. Schuff, "What makes a helpful online review? a study of customer reviews on amazon.com," *MIS Quarterly*, pp. 185–200, 2010.

[2] M. P. O'Mahony and B. Smyth, "Using readability tests to predict helpful product reviews," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, ser. RIAO '10, 2010, pp. 164–167.

[3] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and predicting the helpfulness of online reviews," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM '08, 2008, pp. 443–452.

[4] L. Martin and P. Pu, "Prediction of helpful reviews using emotions extraction," in *Proceedings of 28th AAAI Conference on Artificial Intelligence*, ser. AAAI'14, 2014, pp. 1551–1557.

[5] Y. Yang, Y. Yan, M. Qiu, and F. S. Bao, "Semantic analysis and helpfulness prediction of text for online product reviews," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-2015)*, July 2015, pp. 38–44.

[6] S. Moghaddam and M. Ester, "Aspect-based opinion mining from online reviews," in *Tutorial at SIGIR Conference*, ser. SIGIR'12', 2012.

[7] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM'11, New York, NY, USA, 2011, pp. 815–824.

[8] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang, "Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS)," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 193–202.

[9] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, August 2014, pp. 1–8.

[10] I. Titov and R. Mcdonald, "A joint model of text and aspect ratings for sentiment summarization," in *ACL'2008*, 2008, pp. 308–316.

[11] R. Tadano, K. Shimada, and T. Endo, "Multi-aspects review summarization based on identification of important opinions and their similarity," in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, November 2010, pp. 685–692.

[12] S. Moghaddam and M. Ester, "The FLDA model for aspect-based opinion mining: addressing the cold start problem," *Proceedings of the 22nd international conference on World Wide Web*, pp. 909–918, 2013.

[13] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ser. ECIR'11. Springer-Verlag, 2011, pp. 338–349.

[14] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proceedings of the 7th ACM Conference on Recommender Systems*, ser. RecSys '13. ACM, 2013, pp. 165–172.

[15] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06, 2006, pp. 423–430.

[16] I. Titov and R. McDonald, "Modeling Online Reviews with Multi-grain Topic Models," in *WWW'2008*, 2008, pp. 111–120.

[17] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ser. ACL'12, 2012, pp. 339–348.

[18] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: A case study on amazon.com helpfulness votes," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09, 2009, pp. 141–150.

[19] R. Krestel and N. Dokoohaki, "Diversifying product review rankings: Getting the full picture," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, ser. WI-IAT'11, 2011, pp. 138–145.

[20] D. Agarwal, B.-C. Chen, and B. Pang, "Personalized recommendation of user comments via factor models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11, 2011, pp. 571–582.

[21] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '10, 2010, pp. 56–65.

[22] W. Xiong and D. Litman, "Automatically predicting peer-review helpfulness," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ser. HLT'11, 2011, pp. 502–507.

[23] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic inquiry and word count: Liwc," 2007.

[24] P. J. Stone, R. F. Bales, J. Z. Namenwirth, and D. M. Ogilvie, "The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information," in *Behavioral Science*, 1962, pp. 484–498.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.