Product-Aware Helpfulness Prediction of Online Reviews

Miao Fan, Chao Feng, Lin Guo, Mingming Sun, Ping Li
Cognitive Computing Lab (CCL), Baidu Research
{fanmiao,v_fegchao,v_guolin02,sunmingming01,liping11}@baidu.com

ABSTRACT

Helpful reviews are essential for e-commerce and review websites, as they can help customers make quick purchase decisions and merchants to increase profits. Due to a great number of online reviews with unknown helpfulness, it recently leads to promising research on building automatic mechanisms to assess review helpfulness. The mainstream methods generally extract various linguistic and embedding features solely from the text of a review as the evidence for helpfulness prediction. We, however, consider that the helpfulness of a review should be fully aware of the metadata (such as the title, the brand, the category, and the description) of its target product, besides the textual content of the review itself. Hence, in this paper we propose an end-to-end deep neural architecture directly fed by both the metadata of a product and the raw text of its reviews to acquire product-aware review representations for helpfulness prediction. The learned representations do not require tedious labor on feature engineering and are expected to be more informative as the *target-aware* evidence to assess the helpfulness of online reviews. We also construct two large-scale datasets which are a portion of the real-world web data in Amazon and Yelp, respectively, to train and test our approach. Experiments are conducted on two different tasks: helpfulness identification and regression of online reviews, and results demonstrate that our approach can achieve state-of-the-art performance with substantial improvements.

KEYWORDS

E-commerce; online reviews; helpfulness prediction; neural networks; benchmark datasets

ACM Reference Format:
Miao Fan, Chao Feng, Lin Guo, Mingming Sun, Ping Li. 2019. Product-Aware Helpfulness Prediction of Online 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, 7 pages. https://doi.org/10.1145/3308558.3313523

1 INTRODUCTION

The e-commerce and review websites dramatically facilitate our daily lives as they provide a massive number of products and businesses available online and allow us to access to the comments made by experienced consumers. In order to find a desirable product, we prefer to browsing the online reviews of a product in addition to its descriptions, as we believe that the online reviews can provide more subjective and informative opinions from various perspectives on the product besides the objective descriptions given by its merchant. With the explosive growth of the transaction volume of the e-commerce market in recent years, it is common for an online product commented and rated by thousands of purchasers as illustrated by Figure 1. As a result, it is quite time-consuming for potential consumers to sift through all the online reviews with uneven qualities to make purchase decisions.

Therefore, helpful reviews are essential for e-commerce services, as they are able to bridge the gap between customers and merchants in a *win-win* manner, where customers tend to make quick purchase decisions once catching sight of helpful reviews and merchants can increase profits by surfacing the helpful reviews. To discover the helpful reviews, some platforms such as Amazon and Yelp have launched a module (see the "Helpful" buttons beneath each customer review shown by Figure 1) which allows users to

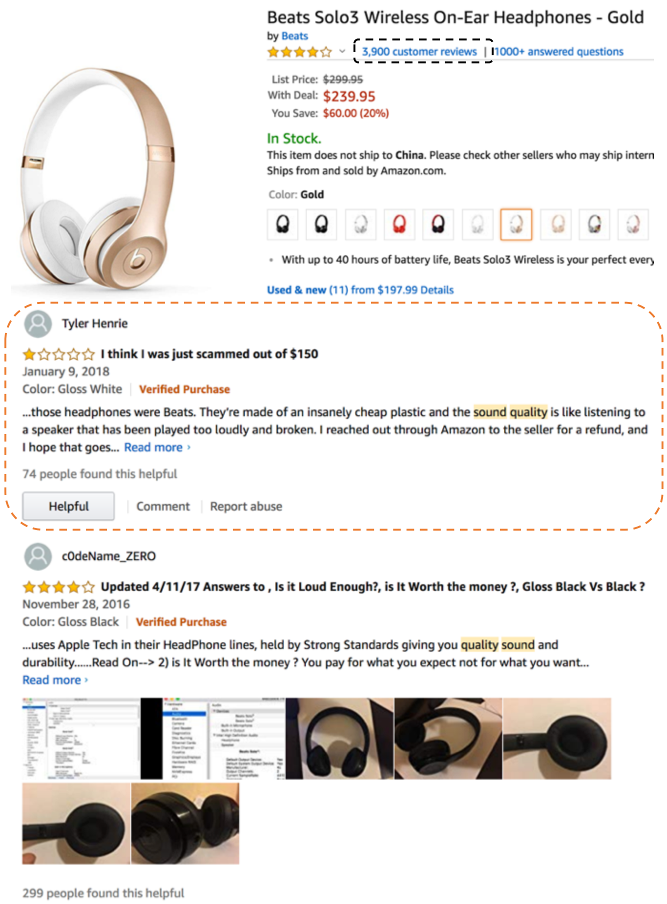

Figure 1: A screenshot of an online product sold on Amazon.com: *Beats Solo3 Wireless On-Ear Headphones - Gold* **which receives 3,900 customer reviews. However, we have no ideas on the helpfulness of most reviews.**

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313523

give feedbacks on the helpfulness of online reviews. This featured module proved by a recent study[1] increases the revenue of Amazon with an estimated 27 billion U.S. dollars annually. Although the crowd-sourcing module could help find a fraction of helpful reviews, roughly 60% online reviews in Amazon.com and Yelp.com did NOT receiving any vote of helpfulness or unhelpfulness. This phenomenon on unknown helpfulness is even more common in low-traffic items including those less popular and new arrival products.

It leads to a promising research direction on building an automatic helpfulness prediction system for online reviews, which we believe could be as useful as a product recommendation engine in e-commerce. So far as we know, a series of work on review helpfulness prediction has been proposed from two perspectives: 1) some work leverages domain-specific knowledge to extract a wide range of hand-crafted features (including structural, lexical, syntactic, emotional, semantic and argument features) from the text of reviews as the evidence for off-the-shelf learning tools; and 2) recent studies modify the convolutional neural network [9, 10] to acquire low-dimensional features from the raw text of online reviews. Generally speaking, these mainstream approaches extract various linguistic and embedding features solely from the text of a review as the evidence for helpfulness prediction.

We, however, suggest that the helpfulness of a review should be fully aware of the meta-data (e.g., title, brand, category, description) of the target product besides the textual content of the review itself. Take the online customer review (circled by the orange dashed box) shown by Figure 1 as an example. The effective features to indicate that it is a helpful review on the product: *Beats Solo3* are probably the phrase "*sound quality*" and word "*headphone*". But the same textual features are hardly considered to be *helpful* when appearing in a comment on a *Nikon Digital Camera*.

In this paper, we propose an end-to-end deep neural architecture to capture the intrinsic relationship between the meta-data of a product and its numerous comments that could be beneficial to discover the helpful reviews. Our model is directly fed by both the title of a product and the raw text of its reviews, and then acquire product-aware review representations from the supervision of helpfulness scores given by the crowd-sourcing module. The learned neural representations do not require tedious labor on feature engineering and are expected to be more informative as the product-aware evidence to assess the helpfulness of online reviews. Given the drawbacks that prior systems have been evaluated by different datasets and not built on the successes of each other, we also construct two large-scale datasets, i.e. *Amazon-9* and *Yelp-5*, which are a portion of the real-world web data in Amazon.com and Yelp.com, respectively, for the successive assessment on the helpfulness prediction of online reviews. The mainstream approaches mentioned in this paper are re-implemented in addition to our model, for fair comparison. Extensive experiments are conducted on two different application scenarios: helpfulness identification and regression of online reviews using the two benchmark datasets. Experimental results demonstrate that our model can achieve state-of-the-art performance on the two tasks with significant absolute improvements (**4.25%** AUROC in the identification of helpful reviews and **4.40%** R^2-score in the regression of helpfulness voting).

[1] https://articles.uie.com/magicbehindamazon/

2 RELATED WORK

An up-to-date and comprehensive survey on various approaches on helpfulness prediction of online reviews was recently conducted by Ocampo Diaz and Ng [17]. According to their survey, these mainstream methods generally extract various linguistic and embedding features solely from the text of an online review as the evidence for helpfulness prediction. We go further and categorize those approaches into two classes in terms of the way of acquiring supportive features for helpfulness prediction, i.e., learning with hand-crafted features (see Section 2.1) and from deep neural networks (see Section 2.2), respectively.

2.1 Learning with Hand-crafted Features

So far as we know, a series of conventional approaches on review helpfulness prediction leverage domain-specific knowledge to extract a wide range of hand-crafted features from the text of customer reviews as the evidence fed into off-the-shelf classifiers or regressors such as SVM [1, 5] or Random Forest [12, 23]. According to [13] and [26], these hand-crafted features involve:

- *Structural features* (STR) [16, 25]: The structural evidence refers to the number of tokens, the number of sentences, the average length of sentences, and even the star rating of an online review. These important features indicate the attitude of the buyers when they write down their comments.
- *Lexical features* (LEX) [8, 24]: Inspired by the idea of text classification, the bag-of-words (BOW) features are essential to helpfulness prediction of online reviews. Explicitly, we usually remove the stop words and non-frequent words, extract unigrams (UGR) and bigrams (BGR) and weight these terms by the measurement of *tf-idf* as the lexical features.
- *Syntactic features* [8]: We can also obtain the part-of-speech (POS) tag of each token in a review. The syntactic features are composed of the percentages of tokens that are nouns, verbs, adjectives, and adverbs, respectively.
- *Emotional features*: Martin and Pu [14] used the Geneva Affect Label Coder (GALC) dictionary [20] to define 36 emotion states of a review. The emotional features include the number of occurrences of each emotional state plus one additional dimension for the number of non-emotional words.
- *Semantic features*: Yang et al. [26] leveraged the General Inquirer (INQUIRER) dictionary [22] to map each word in a review into a semantic tag. This is similar to the way of obtaining emotional features. The semantic features are formulated by a vector where each entry records the number of the occurrences of each semantic tag.
- *Argument features*: Liu et al. [13] explored more intricate linguistic features such as evidence-conclusion discourse relations, also known as arguments, to study the helpfulness of an online review. To be exact, they adopted different granularities of argument features, e.g., the number of arguments, the number of words in arguments, etc.

2.2 Learning from Deep Neural Nets

The emergence of Deep Learning [11] brings in a good insight that we do not have to manually design heuristic rules to extract domain-specific features for learning tasks. To avoid the tedious

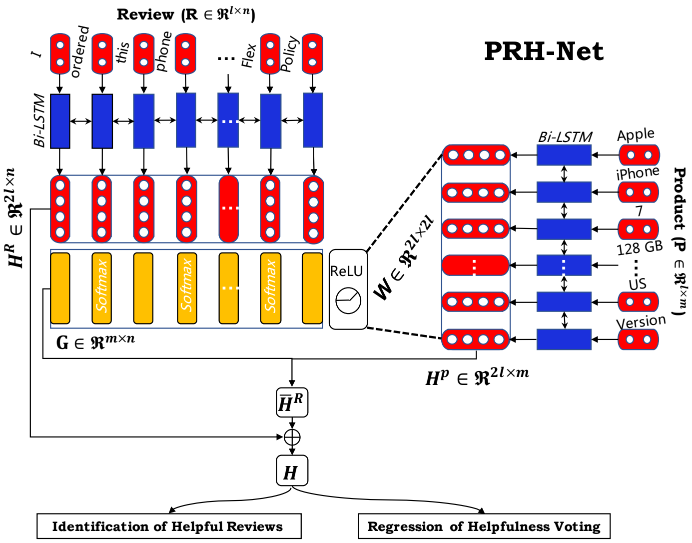

Figure 2: Framework of our neural architecture for product-aware review helpfulness prediction: PRH-Net. $\mathbf{H}^P \in \mathbb{R}^{2l \times m}$ and $\mathbf{H}^R \in \mathbb{R}^{2l \times n}$ are the contextual embeddings of the words in the review R and the product P, respectively. $\mathbf{G} \in \mathbb{R}^{m \times n}$ is the semantic matching matrix between R and P. $\overline{\mathbf{H}}^R \in \mathbb{R}^{2l \times n}$ is the product-aware review repesentation attended by G.

labor on feature engineering for helpfulness prediction of online reviews, the research community recently made several attempts:

- *Embedding-gated CNN* (EG-CNN) [2, 3]: This work adopts a convolutional neural network model [9] that is able to extract multi-granularity text features from reviews. As different words may contribute to the meaning of a review diversely, Chen et al. [2] suggest using word-level embedding-gates to control the word embeddings fed into the CNN model.
- *Multi-task Neural Learning* (MTNL) [4]: Fan et al. [4] introduce a multi-task neural learning (MTNL) paradigm for identifying helpful reviews. The main task, i.e. identifying helpful reviews, leverages the convoluted neural representations from the textual content of reviews. In order to make more accurate predictions on review helpfulness, the convoluted embeddings are also used to fit the star ratings of reviews as an auxiliary task.

Though MTNL [4] and EG-CNN [2, 3] are two newly proposed neural approaches which perform higher on review helpfulness prediction than the hand-crafted methods, these mainstream approaches generally acquire various embedding features solely from the textual content of a review as the evidence without considering the corresponding product, which we believe is inadequate for building a feasible system to discover helpful online reviews.

3 PROPOSED MODEL

In this section, we elaborate our neural model, abbr. PRH-Net, for product-aware review helpfulness prediction. PRH-Net is devised by the motivation that the helpfulness of an online review should be fully aware of the title of its target product besides the textual content of the review itself. As shown by Figure 2, it is composed of two components: 1) the local contextual embeddings of a review and 2) the product-aware distributed representations of the review. We will then explain how to model the two components.

Suppose that we have a product title P and one of its online reviews R. We use m and n to denote the number of tokens/words in the product title P and the review R, respectively. Firstly, we align each token with the embedding dictionary acquired by the word embedding approaches such as Word2Vec [15], Glove [18] or Elmo [19] to initialize the distributed representations of the product title $\mathbf{P} \in \mathbb{R}^{l \times m}$ and the review $\mathbf{R} \in \mathbb{R}^{l \times n}$.

To achieve the local contextual embeddings of the review R, we use a Bi-LSTM network [21] which takes the word embeddings of the review $\mathbf{R}$ as input:

$$\mathbf{H}^R = \text{Bi-LSTM}(\mathbf{R}). \tag{1}$$

$\mathbf{H}^R \in \mathbb{R}^{2l \times n}$ stands for the contextual embeddings where each word can obtain two hidden units with the length of $2l$ encoding both the backward and the forward contextual information of the review locally.

Similarly, we can re-fine the word embeddings of the product title $\mathbf{P}$ via another Bi-LSTM network:

$$\mathbf{H}^P = \text{Bi-LSTM}(\mathbf{P}), \tag{2}$$

and achieve the contextual embeddings of the product title $\mathbf{H}^P \in \mathbb{R}^{2l \times m}$. To make the contextual embeddings of the review fully aware of the product title, we devise a word-level matching mechanism as follows,

$$\mathbf{Q} = \text{ReLU}(\mathbf{W}^P \mathbf{H}^P + \mathbf{b}^P \otimes \mathbf{e}_P)^{\mathrm{T}} \mathbf{H}^R \tag{3}$$

where $\mathbf{W}^P \in \mathbb{R}^{2l \times 2l}$ is the weight matrix and $\mathbf{b}^P \in \mathbb{R}^{2l}$ is the bias vector for the Rectifier Linear Unit (ReLU). The outer product $\otimes$ copys the bias vector $\mathbf{b}^P$ m times to generate a $2l \times m$ matrix. Then $\mathbf{Q} \in \mathbb{R}^{m \times n}$ is the sparse matrix that holds the word-level matching information between the product title P and the review R. If we further apply the softmax function to each column of $\mathbf{Q}$, we will obtain $\mathbf{G} \in \mathbb{R}^{m \times n}$, the i-th column of which represents the normalized attention weights over all the words in product title P for the i-th word in the review R:

$$\mathbf{G} = \text{softmax}(\mathbf{Q}). \tag{4}$$

Then we can use the attention matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$ and the contextual embeddings of the product $\mathbf{H}^P \in \mathbb{R}^{2l \times m}$ to re-form the product-aware review representation $\overline{\mathbf{H}}^R \in \mathbb{R}^{2l \times n}$:

$$\overline{\mathbf{H}}^R = \mathbf{H}^P \mathbf{G}. \tag{5}$$

Driven by original motivation, we need to join the local contextual embeddings of the review ($\mathbf{H}^R$) and the product-aware distributed representations of the review ($\overline{\mathbf{H}}^R$) together for predicting its helpfulness with the feature matrix $\mathbf{H} \in \mathbb{R}^{2l \times n}$:

$$\mathbf{H} = \mathbf{H}^R + \overline{\mathbf{H}}^R. \tag{6}$$

$\mathbf{H}$ can also benefit from the idea of ResNet [6] that efficiently acquires the residual between $\mathbf{H}^R$ and $\overline{\mathbf{H}}^R$, and provides a highway to update $\mathbf{H}^R$ if the residual is tiny.

Table 1: The statistics of the Amazon-9 dataset for helpfulness prediction of online reviews in Amazon.com. #(P.): the number of products; #(R.): the number of reviews; #(R. ≥ 1v.): the number of the reviews, each receiving at least 1 vote regardless of helpfulness/unhelpfulness; #(R. ≥ 0.75h.r.): the number of the reviews, each regarded as helpful by at least 75% votes.

Category (Amazon-9)	Training Set				Test Set			
	# (P.)	# (R.)	# (R. ≥ 1v.)	# (R. ≥ 0.75h.r.)	# (P.)	# (R.)	# (R. ≥ 1v.)	# (R. ≥ 0.75h.r.)
Books	1,157,801	16,240,648	8,884,688	5,538,695	288,901	4,066,140	2,223,022	1,378,901
Clothing, Shoes & Jewelry	381,980	4,599,396	1,480,371	1,076,792	96,018	1,149,157	371,251	269,701
Electronics	223,058	6,220,960	2,597,464	1,660,149	56,117	1,577,881	658,894	422,155
Grocery & Gourmet Food	77,019	1,044,446	441,009	294,932	19,301	250,526	105,264	70,896
Health & Personal Care	118,097	2,395,311	1,094,046	686,932	29,552	581,376	267,255	166,135
Home & Kitchen	176,288	3,406,157	1,479,561	1,090,746	44,415	839,357	361,686	267,284
Movies & TV	126,719	3,614,284	2,078,390	996,024	31,686	867,957	499,574	243,851
Pet Supplies	44,926	997,365	362,499	270,017	11,232	236,222	85,979	63,946
Tools & Home Improvement	107,504	1,538,851	637,665	450,244	26,997	386,130	162,769	114,990
TOTAL	2,413,392	40,057,418	19,055,639	12,064,531	604,219	9,954,746	4,735,694	2,997,859

Table 2: The statistics of the Yelp-5 dataset for helpfulness prediction of online reviews in Yelp.com. #(P.): the number of products; #(R.): the number of reviews; #(R. ≥ 1v.): the number of the reviews, each receiving at least 1 vote regardless of helpfulness/unhelpfulness; #(R. ≥ 0.75h.r.): the number of the reviews, each regarded as helpful by at least 75% votes.

Category (Yelp-5)	Training Set				Test Set			
	# (P.)	# (R.)	# (R. ≥ 1v.)	# (R. ≥ 0.75h.r.)	# (P.)	# (R.)	# (R. ≥ 1v.)	# (R. ≥ 0.75h.r.)
Beauty & Spas	14,066	301,822	162,108	90,003	3,565	76,864	41,555	23,006
Health & Medical	12,034	173,480	103,158	66,616	3,009	43,764	25,979	16,791
Home Services	13,233	201,234	116,011	76,709	3,346	48,085	27,980	18,846
Restaurants	44,597	2,948,004	1,472,688	590,588	11,152	706,793	352,268	142,722
Shopping	22,929	355,366	212,954	100,373	5,725	87,312	51,571	24,105
TOTAL	106,859	3,979,9061	2,066,919	924,289	26,794	962,818	499,353	225,470

4 EXPERIMENTS

4.1 Benchmark Datasets

We find two well-formatted JSON resources online which contain plenty of meta-data (including titles, brands, categories, and descriptions) of products and numerous customer reviews. One is the data collection[2] of Amazon.com crawled by He and McAuley [7] up to July 2014. The other one is the dump file[3] directly provided by Yelp.com for academic purposes.

We use the product ids ("*asin*" in Amazon and "*business_id*" in Yelp) as the foreign keys to align the meta-data of products with corresponding online reviews. 80% products with reviews are randomly picked as the training set, leaving the rest as the test set. In this way, two benchmark datasets, i.e. *Amazon-9* and *Yelp-5*, are prepared, and the statistics of the two datasets are shown by Table 1 and Table 2, respectively.

In line with Table 1, *Amazon-9* covers more than 3 million products spreading over nine different categories in Amazon.com. About 50 million online reviews are included, but less than 48% (roughly 24 million reviews) of them receive at least 1 vote regardless of helpfulness/unhelpfulness by crowd-sourcing. As for *Yelp-5* shown by Table 2, it contains about 130 thousand businesses which fall into five categories in Yelp.com. The proportion of voted reviews in *Yelp-5* is also lower, i.e., roughly 52% (about 2.5 million reviews).

In this work, we regard the reviews which receive at least 1 vote, i.e. the column named after # (R. ≥ 1v.) in Table 1 and Table 2, as the experimental samples. In Amazon.com, the crowd-sourcing module for voting helpful reviews provides an "X of Y" score of helpfulness where "Y" stands for the total number of users who participate in voting and "X" denotes the number of users who think the review is helpful. Yelp.com offers more options: *useful*: X,*cool*: Y, and *funny*: Z, to the users who are willing to give feedbacks. Regardless of the difference, we generally regard the reviews which receive at least 0.75 ratio of helpfulness/usefulness, i.e. # (R. ≥ 0.75h.r.), as positive samples, leaving the others as the negative samples.

4.2 Comparison Methods

We compare our model (PRH-Net) with a wide range of prior arts mentioned in Section 2. Specifically, we re-implement the methods learning with hand-crafted features and from deep neural networks. The up-to-date neural approaches involve the embedding-gated CNN (EG-CNN) [2, 3] and the multi-task neural learning (MTNL) architecture [4] for review helpfulness prediction. The hand-crafted

[2]The data collection is available at http://jmcauley.ucsd.edu/data/amazon/links.html
[3]The dump file can be downloaded from https://www.yelp.com/dataset

Table 3: Comparison of the performance (AUROC) of mainstream approaches on identifying helpful reviews evaluated by the test sets of Amazon-9. (*italic fonts**: the best performance among the baseline approaches; bold fonts: the state-of-the-art performance of all the approaches)

Category (Amazon-9)	**Area under Receiver Operating Characteristic (AUROC)**								
	STR	LEX	GALC	INQUIRER	FUSION (SVM)	FUSION (R.F.)	EG-CNN	MTNL	PRH-Net
Books	0.595	0.572	0.610	0.620	0.594	0.601	0.625	*0.629**	**0.652** (+0.023)
Clothing, Shoes & Jewelry	0.559	0.538	0.565	*0.608**	0.587	0.557	0.590	0.592	**0.614** (+0.006)
Electronics	0.590	0.555	0.593	*0.627**	0.584	0.588	0.615	0.618	**0.644** (+0.017)
Grocery & Gourmet Food	0.540	0.526	0.566	0.618	0.537	0.556	0.613	*0.638**	**0.715** (+0.077)
Health & Personal Care	0.560	0.533	0.569	0.617	0.599	0.565	0.617	*0.624**	**0.672** (+0.048)
Home & Kitchen	0.572	0.545	0.576	0.609	0.579	0.573	0.605	*0.611**	**0.630** (+0.019)
Movies & TV	0.613	0.562	0.624	0.637	0.605	0.617	0.648	*0.652**	**0.675** (+0.023)
Pet Supplies	0.560	0.542	0.585	0.603	0.548	0.558	0.580	*0.619**	**0.679** (+0.060)
Tools & Home Improvement	0.584	0.548	0.580	0.592	0.565	0.586	0.607	*0.621**	**0.644** (+0.023)
MACRO AVERAGE	0.575	0.547	0.585	0.615	0.578	0.578	0.611	*0.623**	**0.658** (+0.035)
MICRO AVERAGE (Primary)	0.587	0.559	0.598	0.620	0.589	0.591	0.620	*0.625**	**0.651 (+0.026)**

Table 4: Comparison of the performance (AUROC) of mainstream approaches on identifying helpful reviews evaluated by the test sets of Yelp-5. (*italic fonts**: the best performance among the baseline approaches; bold fonts: the state-of-the-art performance of all the approaches)

Category (Yelp-5)	**Area under Receiver Operating Characteristic (AUROC)**								
	STR	LEX	GALC	INQUIRER	FUSION (SVM)	FUSION (R.F.)	EG-CNN	MTNL	PRH-Net
Beauty & Spas	0.512	0.500	0.527	0.570	0.521	0.541	0.571	*0.581**	**0.642** (+0.061)
Health & Medical	0.525	0.517	0.538	0.576	0.539	0.531	0.580	*0.596**	**0.665** (+0.069)
Home Services	0.528	0.528	0.562	0.584	0.535	0.538	0.563	*0.603**	**0.732** (+0.129)
Restaurants	0.559	0.516	0.552	0.582	0.569	0.554	0.581	*0.605**	**0.658** (+0.053)
Shopping	0.528	0.518	0.560	0.609	0.542	0.555	0.572	*0.619**	**0.674** (+0.055)
MACRO AVERAGE	0.530	0.516	0.548	0.584	0.541	0.544	0.573	*0.601**	**0.674** (+0.073)
MICRO AVERAGE (Primary)	0.548	0.516	0.551	0.584	0.559	0.551	0.578	*0.604**	**0.663 (+0.059)**

features include the structural features (STR) [16, 25], the lexical features (LEX) [8, 24], the emotional features (GALC) [14] and the semantic features (INQUIRER) [26]. We also add two more experiments on integrating all the hand-crafted features via the Support Vector Machines (SVM) and the Random Forest (R.F.) model for review helpfulness prediction.

4.3 Application Scenarios

Most previous studies just reported their performance on either the task of review helpfulness identification or regression. In this part, we assess the performance of PRH-Net by comparing our model with all the other approaches on both tasks.

4.3.1 Identification of Helpful Reviews. We use the data shown by the column named #(R. $\geq$ 1v.) in Table 1 and Table 2 to conduct this task. Within the data for binary classification, the reviews belonging to the column #(R. $\geq$ 0.75h.r.) are regarded as positive samples. As both the training and test sets are imbalanced, we adopt the Area under Receiver Operating Characteristic (AUROC) as the metric to evaluate the performance of all the approaches on helpful review identification. As shown by Table 3 and Table 4, MTNL [4] achieves the up-to-date performance on this classification task among the baseline approaches as it shows the best performance on 12 of 14 categories in *Amazon-9* and *Yelp-5* datasets. Our model (PRH-Net) surpasses MTNL on both datasets and obtains state-of-the-art (micro-averaged) results of 65.1% AUROC (*Amazon-9*) and 66.3% AUROC (*Yelp-5*) with absolute improvements of **2.6%** AUROC and **5.9%** AUROC, respectively.

4.3.2 Regression of Helpfulness Voting. In this task, all the approaches are required to predict the fraction of helpful votes that each review receives. We still use the data in the column named of #(R. $\geq$ 1v.) in Table 1 and Table 2 as the training and test sets. The Squared Correlation Coefficient (R^2-score) is adopted as the metric to evaluate the performance of all the approaches on helpfulness score regression. Table 5 and Table 6 show that MTNL [4] achieves the up-to-date performance on this regression task among the baselines. Our model (PRH-Net) outperforms MTNL on both datasets and obtains state-of-the-art (micro-averaged) results of 55.2% R^2-score (*Amazon-9*) and 58.2% R^2-score (*Yelp-5*) with absolute improvements of **3.5%** R^2-score and **5.3%** R^2-score, respectively.

Table 5: Comparison of the performance (R^2-score) of mainstream approaches on helpfulness voting regression evaluated by the test sets of Amazon-9. (*italic fonts: the best performance among the baseline approaches; bold fonts: the state-of-the-art performance of all the approaches)**

Category (Amazon-9)	Squared Correlation Coefficient (R^2-score)								
	STR	LEX	GALC	INQUIRER	FUSION (SVM)	FUSION (R.F.)	EG-CNN	MTNL	PRH-Net
Books	0.250	0.234	0.312	0.458	0.401	0.441	0.506	*0.549**	**0.586** (+0.037)
Clothing, Shoes & Jewelry	0.245	0.307	0.368	0.428	0.360	0.376	0.491	*0.510**	**0.567** (+0.057)
Electronics	0.223	0.291	0.333	0.450	0.416	0.357	0.394	*0.489**	**0.516** (+0.027)
Grocery & Gourmet Food	0.242	0.319	0.419	0.450	0.426	0.425	0.469	*0.506**	**0.569** (+0.063)
Health & Personal Care	0.233	0.332	0.376	0.433	0.420	0.399	0.506	*0.509**	**0.537** (+0.028)
Home & Kitchen	0.236	0.298	0.330	0.464	0.339	0.361	0.402	*0.498**	**0.513** (+0.015)
Movies & TV	0.253	0.228	0.312	0.387	0.360	0.352	0.393	*0.453**	**0.495** (+0.042)
Pet Supplies	0.237	0.237	0.285	0.420	0.301	0.339	0.400	*0.473**	**0.523** (+0.050)
Tools & Home Improvement	0.234	0.201	0.287	0.437	0.247	0.302	0.398	*0.481**	**0.503** (+0.022)
MACRO AVERAGE	0.239	0.272	0.336	0.436	0.363	0.372	0.440	*0.496**	**0.534** (+0.038)
MICRO AVERAGE (Primary)	0.243	0.258	0.325	0.445	0.385	0.399	0.463	*0.517**	**0.552 (+0.035)**

Table 6: Comparison of the performance (R^2-score) of mainstream approaches on helpfulness voting regression evaluated by the test sets of Yelp-5. (*italic fonts: the best performance among the baseline approaches; bold fonts: the state-of-the-art performance of all the approaches)**

Category (Yelp-5)	Squared Correlation Coefficient (R^2-score)								
	STR	LEX	GALC	INQUIRER	FUSION (SVM)	FUSION (R.F.)	EG-CNN	MTNL	PRH-Net
Beauty & Spas	0.283	0.384	0.511	0.537	0.349	0.418	0.550	*0.552**	**0.624** (+0.072)
Health & Medical	0.253	0.454	0.433	0.557	0.478	0.459	0.572	*0.573**	**0.635** (+0.062)
Home Services	0.264	0.452	0.416	0.554	0.492	0.481	0.570	*0.575**	**0.645** (+0.070)
Restaurants	0.256	0.338	0.383	0.501	0.356	0.407	0.510	*0.518**	**0.564** (+0.046)
Shopping	0.306	0.347	0.417	0.523	0.334	0.400	0.537	*0.540**	**0.606** (+0.066)
MACRO AVERAGE	0.272	0.395	0.432	0.534	0.402	0.433	0.548	*0.552**	**0.615** (+0.063)
MICRO AVERAGE (Primary)	0.264	0.355	0.402	0.512	0.367	0.414	0.523	*0.529**	**0.582 (+0.053)**

5 CONCLUSION AND FUTURE WORK

This paper engages in the emerging research on helpfulness prediction of online reviews. Our idea is driven by the motivations that 1) the helpfulness of an online review should be fully aware of the meta-data of its target product besides the textual content of the review itself; 2) the hand-crafted features requiring tedious labor are domain-specific and prone to high generation error; and 3) there is no widely-used benchmark dataset for constantly improving intelligent systems to precisely assess the helpfulness of online reviews.

To address the problems above, we contribute an end-to-end neural architecture which can automatically acquire product-aware review representations besides the textual embeddings of reviews as more informative evidence for review helpfulness prediction. We also construct two large-scale and real-world benchmark datasets, i.e. *Amazon-9* and *Yelp-5*, for the sake of 1) fairly conducting the performance comparison of all the approaches on review helpfulness prediction, and 2) leaving the datasets available for successive studies. Specifically, we run extensive experiments on our newly constructed datasets under the application scenarios of helpfulness identification and regression. Experimental results demonstrate that our model surpasses all the mainstream approaches and achieves state-of-the-art performance with substantial improvements.

For future work, we might consider to study the following topics related to helpfulness prediction of online reviews:

- User-specific and explainable recommendation of helpful reviews: As different users may concern about various aspects of the products online, helpful review recommendation needs to be more user-specific and self-explainable.
- Cross-domain helpfulness prediction of online reviews [3]: Given that it costs a lot on manually annotating plenty of helpful reviews in a new domain, we should explore effective approaches on transferring useful knowledge from limited labeled samples in another domain.
- Enhancing the prediction of helpful reviews with unlabeled data: As a small proportion of reviews could be heuristically regarded as helpful or unhelpful, it, therefore, becomes a promising study to automatically predict the helpfulness of online reviews based on the small amount of labeled data and a huge amount of unlabeled data.

REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology* 2, 3 (2011), 27.

[2] Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Bao. 2018. Review Helpfulness Prediction with Embedding-Gated CNN.

[3] Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. Cross-Domain Review Helpfulness Prediction Based on Convolutional Neural Networks with Auxiliary Domain Discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 2 (Short Papers)*. Association for Computational Linguistics, 602–607.

[4] Miao Fan, Yue Feng, Mingming Sun, Ping Li, Haifeng Wang, and Jianmin Wang. 2018. Multi-Task Neural Learning Architecture for End-to-End Identification of Helpful Reviews. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 343–350.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research (JMLR)* 9 (2008), 1871–1874.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[7] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*. International World Wide Web Conferences Steering Committee, 507–517.

[8] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically Assessing Review Helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 423–430.

[9] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1746–1751.

[10] Yann LeCun and Yoshua Bengio. 1998. The Handbook of Brain Theory and Neural Networks. (1998), 255–258.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[12] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[13] Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using Argument-based Features to Predict and Analyse Review Helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1369–1374.

[14] Lionel Martin and Pearl Pu. 2014. Prediction of Helpful Reviews Using Emotions Extraction. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 1551–1557.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS)*. 3111–3119.

[16] Susan M. Mudambi and David Schuff. 2010. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.Com. *MIS Quarterly* 34, 1 (March 2010), 185–200.

[17] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 698–708.

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[19] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for Computational Linguistics, 2227–2237.

[20] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[21] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing (TSP)* 45, 11 (1997), 2673–2681.

[22] Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science* 7, 4 (1962), 484–498.

[23] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences* 43, 6 (2003), 1947–1958.

[24] Wenting Xiong and Diane Litman. 2011. Automatically Predicting Peer-Review Helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*. Association for Computational Linguistics, 502–507.

[25] Wenting Xiong and Diane J Litman. 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews.. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics, 1985–1995.

[26] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 38–44.