Multi-perspective Coherent Reasoning for Helpfulness Prediction of Multimodal Reviews

Junhao Liu[1,2,3]* **Zhen Hai**[3] **Min Yang**[1]† **Lidong Bing**[3]
[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]DAMO Academy, Alibaba Group
{jh.liu, min.yang}@siat.ac.cn
{zhen.hai, l.bing}@alibaba-inc.com

Abstract

As more and more product reviews are posted in both text and images, Multimodal Review Analysis (MRA) becomes an attractive research topic. Among the existing review analysis tasks, helpfulness prediction on review text has become predominant due to its importance for e-commerce platforms and online shops, i.e. helping customers quickly acquire useful product information. This paper proposes a new task Multimodal Review Helpfulness Prediction (MRHP) aiming to analyze the review helpfulness from text and visual modalities. Meanwhile, a novel Multi-perspective Coherent Reasoning method (MCR) is proposed to solve the MRHP task, which conducts joint reasoning over texts and images from both the product and the review, and aggregates the signals to predict the review helpfulness. Concretely, we first propose a product-review coherent reasoning module to measure the intra- and inter-modal coherence between the target product and the review. In addition, we also devise an intra-review coherent reasoning module to identify the coherence between the text content and images of the review, which is a piece of strong evidence for review helpfulness prediction. To evaluate the effectiveness of MCR, we present two newly collected multimodal review datasets as benchmark evaluation resources for the MRHP task. Experimental results show that our MCR method can lead to a performance increase of up to 8.5% as compared to the best performing text-only model. The source code and datasets can be obtained from https://github.com/jhliu17/MCR.

1 Introduction

Product reviews are essential information sources for consumers to acquire useful information and make purchase decisions. Many e-commerce sites such as Amazon.com offer reviewing functions that encourage consumers to share their opinions and experiences. However, the user-generated reviews vary a lot in their qualities, and we are continuously bombarded with ever-growing, noise information. Therefore, it is critical to examine the quality of reviews and present consumers with useful reviews.

Motivated by the demand of gleaning insights from such valuable data, review helpfulness prediction has gained increasing interest from both academia and industry communities. Earlier review helpfulness prediction methods rely on a wide range of handcrafted features, such as semantic features (Yang et al., 2015), lexical features (Martin and Pu, 2014), and argument based features (Liu et al., 2017), to train a classifier. The success of these methods generally relies heavily on feature engineering which is labor-intensive and highlights the weakness of conventional machine learning methods. In recent years, deep neural networks such as CNN (Chen et al., 2018, 2019) and LSTM (Fan et al., 2019) have become dominant in the literature due to their powerful performance for helpfulness prediction by learning text representation automatically. Note that these existing works on review helpfulness prediction mainly focus on the pure textual data.

As multimodal data become increasingly popular in online reviews, Multimodal Review Analysis (MRA) has become a valuable research direction. In this paper, we propose the Multimodal Review Helpfulness Prediction (MRHP) task which aims at exploring multimodal clues that often convey comprehensive information for review helpfulness prediction. In particular, for the multimodal reviews, the helpfulness of reviews is not only determined by the textual content but rather the combined expression (e.g., coherence) of multimodality data (e.g., texts and images). Taking the reviews in Table 1

*This work was conducted when Junhao Liu was an intern at DAMO Academy, Alibaba Group.

†Min Yang is the corresponding author.

as an example, we cannot identify the helpfulness score of *Review 3* solely from the text content until reading the attached images that are totally irrelevant to the product "*Teflon Pans*". The reviews that have incoherent text content and images tend to be unhelpful, even be malicious reviews. In contrast, a helpful review (e.g., *Review 2*) should contain not only concise and informative textual content but also coherent text content and images.

In this paper, we explore both text and images in product reviews to improve the performance of review helpfulness prediction. We design a novel Multi-perspective Coherent Reasoning method (denoted as MCR) to tackle the MRHP task. Concretely, we propose a product-review coherent reasoning module to effectively capture the intra- and inter-modal coherence between the target product and the review. In addition, we also devise an intra-review coherent reasoning module to capture the coherence between the text content and images of the review, which is a piece of strong evidence for review helpfulness prediction. Finally, we formulate the helpfulness prediction as a ranking problem and employ a pairwise ranking objective to optimize the whole model.

We summarize our main contributions as follows. (1) To the best of our knowledge, this is the first attempt to explore both text and images in reviews for helpfulness prediction, which is defined as the MRHP task. (2) We propose a multi-perspective coherent reasoning method for the MRHP task to conduct joint reasoning over texts and images from both the product and the review, and aggregate the signals to predict the helpfulness of multimodal reviews. (3) We present two newly-collected multimodal review datasets for helpfulness prediction of multimodal reviews. To facilitate research in this area, we will release the datasets and source code proposed in this paper, which would push forward the research in this field. (4) Extensive experiments on two collected datasets demonstrate that our MCR method significantly outperforms other methods.

2 Related Work

Most conventional approaches on review helpfulness prediction focus solely on the text of reviews, which can be generally divided into two categories based on the way of extracting predictive features: machine learning based methods with hand-crafted features (Kim et al., 2006; Krishnamoorthy, 2015)

Product Information Teflon Pans 1 Set of 3 pcs 1042-Non-stick Set of 3 
Review 1 (Helpfulness Score: 2) Overall, it is quite satisfactory. Thanks to the seller. 
Review 2 (Helpfulness Score: 4) For that price, it is more than satisfactory, even though there are a few scratches in *the pan* and the small frying pan, *the package is very neat*, *the frying pan* has been used as if it's a little burnt, it looks like it can't stand the heat, but overall I like it. 
Review 3 (Helpfulness Score: 0) Recommend for the price. Yes, the package is neat but the pan has scratched. It is unfortunate for the delivery. I ordered 4 items in this shop. but the postage has to pay double and quite very expensive. 

Table 1: Example of multimodal reviews under the same product "*Teflon Pan*". Review 1: The brief review text is insufficient to predict its helpfulness to the corresponding product, while the images provide a rich semantic supplement. Review 2: A helpful review with a good coherence between text and images. Review 3: An irrelevant image is attached to the review.

and deep learning based methods (Chen et al., 2019; Fan et al., 2018; Chen et al., 2018). The machine learning based methods employ domain-specific knowledge to extract a variety of hand-crafted features, such as structure features (Kim et al., 2006), lexical features (Krishnamoorthy, 2015), emotional features (Martin and Pu, 2014), and argument features (Liu et al., 2017), from the textural reviews, which are then fed into conventional classifiers such as SVM (Kim et al., 2006) for helpfulness prediction. These methods rely heavily on feature engineering, which is time-consuming and labor intensive. Motivated by the remarkable progress of deep neural networks, several recent studies attempt to automatically learn deep features from textual reviews with deep neural networks. Chen et al. (2019) employs a CNN model to capture the

multi-granularity (character-level, word-level, and topic-level) features for helpfulness prediction. Fan et al. (2018) proposes a multi-task neural learning model to identify helpful reviews, in which the primary task is helpfulness prediction and the auxiliary task is star rating prediction.

Subsequently, several works have been proposed to explore not only the reviews but also the users and target products for helpfulness prediction of reviews. Fan et al. (2019) argued that the helpfulness of a review should be aware of the meta-data (e.g., title, brand, category, description) of the target product besides the textual content of the review itself. To this end, a deep neural architecture was proposed to capture the intrinsic relationship between the meta-data of a product and its numerous reviews. Qu et al. (2020) proposed to leverage the reviews, the users, and items together for helpfulness prediction of reviews and devised a category-aware graph neural networks with one shared and many item-specific graph convolutions to learn the common features and each item's specific criterion for helpfulness prediction.

Different from the above methods, we take full advantage of the text content and images of reviews by proposing a novel hierarchical coherent reasoning method to learn the coherence between text content and images in a review and the coherence between the target product and the review.

3 Methodology

The overall architecture of our MCR method is illustrated in Figure 1. Our multi-perspective coherent reasoning consists of two perspectives of coherence: (i) the intra- and inter-modal coherence between a review and the target product and (ii) the intra-review coherence between the text content and images in the review. In the following sections, we will provide the problem definition of review helpfulness prediction and introduce each component of our MCR model in detail.

3.1 Problem Definition

As mentioned by Diaz and Ng (2018), we formulate the multimodal review helpfulness prediction problem as a ranking task. Specifically, given a product item P_i consisting of product related information p_i and an associated review set $R_i = \{r_{i,1}, \cdots, r_{i,N}\}$, where N is the number of reviews for p_i. Each review has a scalar label $s_{i,j} \in \{0, \cdots, S\}$ indicating the helpfulness score of the review $r_{i,j}$. The ground-truth ranking of R_i is the descending sort order determined by the helpfulness scores. The goal of review helpfulness prediction is to predict helpfulness scores for R_i which can rank the set of reviews R_i into the ground-truth result. The predicted helpfulness score $\hat{s}_{i,j}$ for the review $r_{i,j}$ is defined as follows:

$$\hat{s}_{i,j} = f(p_i, r_{i,j}), \tag{1}$$

where f is the helpfulness prediction function taking a product-review pair $\langle p_i, r_{i,j} \rangle$ as input. In multimodal review helpfulness prediction task, the product p_i consists of associated description T_p and pictures I_p, while review $r_{i,j}$ consists of user-posted text T_r and images I_r.

3.2 Feature Representation

Given a text (T_p or T_r) consisting of l_T text tokens $\{w_1, \cdots, w_{l_T}\}$ and an image set (I_p or I_r), we adopt a convolutional neural network to learn the contextualized text representation. Meanwhile, we use a self-attention mechanism on image region features to obtain the image representations. To prevent conceptual confusion, we use the subscripts p and r to indicate variables that are related to the product and the review, respectively.

Text Representation Inspired by the great success of convolutional neural network (CNN) in natural language processing (Kim, 2014; Dai et al., 2018), we also apply CNN to learn the text representation. First, we convert each token w_i in a review into an embedding vector $\mathbf{w}_i \in \mathbb{R}^d$ via an embedding layer. Then, we pass the learned word embeddings to a one-dimensional CNN so as to extract multi-gram representations. Specifically, the k-gram CNN transforms the token embedding vectors $\mathbf{w}_i$ into k-gram representations $\mathbf{H}^k$:

$$\mathbf{H}^k = \mathbf{CNN}^k(\{\mathbf{w}_1, \cdots, \mathbf{w}_{l_T}\}), \tag{2}$$

where $k \in \{1, \cdots, k_{max}\}$ represents the kernel size. k_{max} represents the maximum kernel size. $\mathbf{H}^k \in \mathbb{R}^{l_T \times d_T}$ is the k-gram representation. All the k-gram representations are stacked to form the final text representation, denoted as $\mathbf{H} = [\mathbf{H}^1, \cdots, \mathbf{H}^{k_{max}}]$. Here, we use $\mathbf{H}_p$ and $\mathbf{H}_r$ to represent the representations of text content of the product and the review, respectively.

Image Representation We use pre-trained Faster R-CNN to extract the region of interest (RoI) pooling features (Anderson et al., 2018) for the

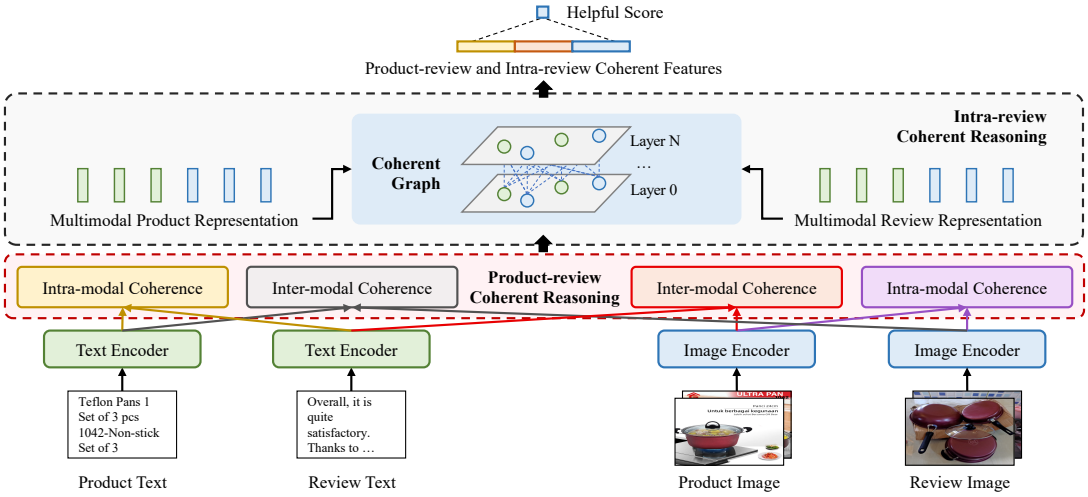

Figure 1: Model overview of our MCR method, which consists of two primary coherent reasoning components: product-review coherent reasoning and intra-review coherent reasoning.

review and product images, obtaining the fine-grained object-aware representations. All the RoI features $\mathbf{v}_i$ extracted from image sets I_p and I_r are then encoded by a self-attention module (Vaswani et al., 2017), resulting in a d_I-dimensional semantic space with non-local understanding:

$$\mathbf{V} = \mathbf{SelfAttn}(\{\mathbf{v}_1, \cdots, \mathbf{v}_{l_I}\}), \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{l_I \times d_I}$ represents the visual semantic representation and l_I is the number of extracted RoI features. Here, we use $\mathbf{V}_p$ and $\mathbf{V}_r$ to represent the product and review image features, respectively.

3.3 Product-Review Coherent Reasoning

The helpfulness of a review should be fully aware of the product besides the review itself. In this paper, we propose a *product-review coherent reasoning module* to effectively capture the intra- and inter-modal coherence between the target product and the review.

Intra-modal Coherence We propose the intra-modal coherent reasoning to measure two kinds of intra-modal coherence: (i) the semantic alignments between the product text and the review text, and (ii) the semantic alignments between product images and review images. The cosine similarity is utilized to derive the intra-modal coherence matrix. For text representations $\mathbf{H}_p^i$ and $\mathbf{H}_r^j$, we compute the corresponding coherence matrix as follow:

$$\begin{aligned} \mathbf{S}_{i,j}^{\mathbf{H}} &= \mathbf{cosine}(\mathbf{H}_p^i, \mathbf{H}_r^j), \\ \forall i, j &\in \{1, \ldots, k_{max}\}, \end{aligned} \tag{4}$$

where $\mathbf{S}_{i,j}^{\mathbf{H}}$ has the shape of $\mathbb{R}^{l_{T_p} \times l_{T_r}}$, l_{T_p} and l_{T_r} indicate the text length of the product and the review, respectively. All the coherence matrices are stacked to form the whole coherence features $\mathbf{S}^{\mathbf{H}}$. Without loss of generality, we also compute the image coherence matrix between $\mathbf{V}_p$ and $\mathbf{V}_r$ via cosine similarity. In this way, we obtain the image coherence matrix $\mathbf{S}^{\mathbf{V}}$ with the shape of $\mathbb{R}^{l_{I_p} \times l_{I_r}}$, where l_{I_p} and l_{I_r} indicate the number of RoI features of the product and review images, respectively.

Subsequently, the text and image coherence matrix (i.e., $\mathbf{S}^{\mathbf{H}}$ and $\mathbf{S}^{\mathbf{V}}$) are passed to a CNN, and the top-K values in each feature map are selected as the pooling features:

$$\mathbf{o}_{intraM} = \mathbf{TopK}(\mathbf{CNN}([\mathbf{S}^{\mathbf{H}}, \mathbf{S}^{\mathbf{V}}])), \tag{5}$$

where $\mathbf{o}_{intraM} \in \mathbb{R}^{K*M}$ is the intra-modal coherent reasoning features. M is the number of filters used in the CNN module.

Inter-modal Coherence The intra-modal coherence ignores the cross-modal relationship between the product and the review. In order to mitigate this problem, we propose the inter-modal coherent reasoning to capture two kinds of inter-modal coherence: (i) the coherence between the review text and the product images, and (ii) the coherence between the review images and the product text. Since the text representation $\mathbf{H}$ and the image representation $\mathbf{V}$ lie in two different semantic spaces, we first project them into a d_c-dimensional common latent space by:

$$\mathbf{F}^{\mathbf{H}} = \mathbf{Tanh}(\mathbf{W}_1\mathbf{H} + \mathbf{b}_1), \tag{6}$$

$$\mathbf{F}^{\mathbf{V}} = \mathbf{Tanh}(\mathbf{W}_2\mathbf{V} + \mathbf{b}_2), \tag{7}$$

where $\mathbf{F}^{\mathbf{H}} \in \mathbb{R}^{l_T \times d_c}$ and $\mathbf{F}^{\mathbf{V}} \in \mathbb{R}^{l_I \times d_c}$ are text and image representations in the common latent space, respectively.

Taking the coherence of review image and product text as an example, our inter-modal coherent reasoning aligns the features in review images $\mathbf{F}_r^{\mathbf{V}}$ based on the product text $\mathbf{F}_p^{\mathbf{H}}$. Specifically, we define the review images as the query $\mathbf{Q}_r = \mathbf{W}_Q \mathbf{F}_r^{\mathbf{V}}$ and the product text as the key $\mathbf{K}_p = \mathbf{W}_K \mathbf{F}_p^{\mathbf{H}}$, where $\mathbf{W_Q}, \mathbf{W}_K \in \mathbb{R}^{d_c \times d_c}$ are learnable parameter matrices. Hence, the inter-modal relationship $\mathbf{I}_r^{\mathbf{V}}$ can be formulated as follows:

$$\mathbf{M}_r = \mathbf{softmax}(\mathbf{Q}_r \mathbf{K}_p^T), \tag{8}$$

$$\mathbf{I}_r^{\mathbf{V}} = \mathbf{F}_r^{\mathbf{V}} + \mathbf{M}_r \mathbf{F}_p^{\mathbf{H}}, \tag{9}$$

where $\mathbf{M}_r \in \mathbb{R}^{l_I \times l_T}$ is the query attended mask. A mean-pooling operation is then conducted to get an aggregated vector of the inter-modal coherence features between the review images and the product text: $\tilde{\mathbf{I}}_r^{\mathbf{V}}$:

$$\tilde{\mathbf{I}}_r^{\mathbf{V}} = \mathbf{Mean}(\mathbf{I}_r^{\mathbf{V}}) \in \mathbb{R}^{d_c}. \tag{10}$$

Following Equations 8-10, the same procedure is employed to learn the coherence features $\tilde{\mathbf{I}}_r^{\mathbf{H}}$ between the review text and the product images. Finally, we concatenate $\tilde{\mathbf{I}}_r^{\mathbf{V}}$ and $\tilde{\mathbf{I}}_r^{\mathbf{H}}$ to form the final inter-modal coherence features $\mathbf{o}_{interM}$:

$$\mathbf{o}_{interM} = [\tilde{\mathbf{I}}_r^{\mathbf{V}}, \tilde{\mathbf{I}}_r^{\mathbf{H}}], \tag{11}$$

where $[\cdot]$ denotes the concatenate operation.

3.4 Intra-review Coherent Reasoning

Generally, consumers usually express their opinions in textual reviews and post images as a kind of evidence to support their opinions. To capture the coherence between the text content and images of the review, we should grasp sufficient relational and logical information between them. To this end, we devise an intra-review coherent reasoning module to learn the coherence between the text content and images of the review, which performs message propagation among semantic nodes of a review evidence graph and then obtains an intra-review coherence score of the multimodal review.

Specifically, we construct a review evidence graph G_r by taking each feature (each row) of $\mathbf{F}_r^{\mathbf{H}}$ and $\mathbf{F}_r^{\mathbf{V}}$ as a semantic node, and connects all node pairs with edges, resulting in a fully-connected review evidence graph with $l_T + l_I$ nodes. In a similar manner, we can construct a product evidence graph G_p with $l_T + l_I$ nodes from $\mathbf{F}_p^{\mathbf{H}}$ and $\mathbf{F}_p^{\mathbf{V}}$. The hidden states of nodes at layer t are denoted as $\mathbf{G}_r^t = \{\mathbf{g}_{r,1}^t, \ldots, \mathbf{g}_{r,n}^t\}$ and $\mathbf{G}_p^t = \{\mathbf{g}_{p,1}^t, \ldots, \mathbf{g}_{p,n}^t\}$ for the review and product evidence graphs respectively, where $n = l_T + l_I$ and t denotes the number of hops for graph reasoning. We compute the edge weights of semantic node pairs with an adjacency matrix that can be automatically learned through training. Taking the review evidence graph G_r as an example, we initialize the i-th semantic node at the first layer with $\mathbf{g}_i^0 = [\mathbf{F}_{r,i}^{\mathbf{H}}, \mathbf{F}_{r,i}^{\mathbf{V}}],\ i \in \{1, \cdots, l_T + l_I\}$. Then, the adjacency matrix $\mathbf{A}^t$ representing edge weights at layer t is computed as follows:

$$\tilde{\mathbf{A}}_{i,j}^t = \mathbf{MLP}^{t-1}([\mathbf{g}_{r,i}^{t-1}, \mathbf{g}_{r,j}^{t-1}]), \tag{12}$$

$$\mathbf{A}^t = \mathbf{softmax}(\tilde{\mathbf{A}}^t), \tag{13}$$

where $\mathbf{MLP}^{t-1}$ is an MLP at layer $t-1$. $\tilde{\mathbf{A}}_{i,j}^t$ represents semantic coefficients between a node i with its neighbor $j \in \mathcal{N}_i$. Softmax operation is used to normalize semantic coefficients $\tilde{\mathbf{A}}^t$. Then, we can obtain the reasoning features at layer t by:

$$\mathbf{g}_{r,i}^t = \sum_{j \in \mathcal{N}_i} \mathbf{A}_{i,j}^t \mathbf{g}_{r,j}^{t-1}. \tag{14}$$

By stacking L graph reasoning layers, the semantic nodes can perform coherence relation reasoning by passing messages with each other. We use $\mathbf{g}_{r,n}^L$ and $\mathbf{g}_{p,n}^L$ to denote the final reasoning hidden states of the review and product evidence graphs. Subsequently, to obtain the product-related intra-review coherent reasoning features, we adopt an attention mechanism to filter the features that are irrelevant to the product:

$$\mathbf{p} = \mathbf{Mean}(\mathbf{h}_{p,*}^L), \tag{15}$$

$$\tilde{\alpha}_i = \mathbf{MLP}([\mathbf{p}, \mathbf{g}_{r,i}^L]), \tag{16}$$

where a mean pooling operation is employed to derive the product coherent graph embedding $\mathbf{p}$. $\mathbf{MLP}$ is an attention layer to calculate the product-related features and output the attention weight $\tilde{\alpha}_i$ for the i-th node. After normalizing the attention weight with a softmax function, we use a linear combination to aggregate the intra-review coherent reasoning results $\mathbf{o}_{IRC}$:

$$\alpha = \mathbf{softmax}(\tilde{\alpha}), \tag{17}$$

$$\mathbf{o}_{IRC} = \sum_i \alpha_i \mathbf{g}_{r,i}^L. \tag{18}$$

3.5 Review Helpfulness Prediction

We concatenate the intra-modal product-review coherence features $\mathbf{o}_{intraM}$, the inter-modal product-review coherence features $\mathbf{o}_{interM}$, and the intra-review coherence features $\mathbf{o}_{IRC}$ to form the final multi-perspective coherence features $\mathbf{o}_{final} = [\mathbf{o}_{intraM}, \mathbf{o}_{interM}, \mathbf{o}_{IRC}]$. The final helpfulness prediction layer feeds $\mathbf{o}_{final}$ into a linear layer to calculate a ranking score:

$$f(p_i, r_{i,j}) = \mathbf{W}_r \mathbf{o}_{final} + \mathbf{b}_r, \quad (19)$$

where $\mathbf{W}_r$ and $\mathbf{b}_r$ denote the projection parameter and bias term. p_i represents information of the i-th product and $r_{i,j}$ is the j-th review for p_i.

The standard pairwise ranking loss is adopted to train our model:

$$\mathcal{L} = \sum_i \mathbf{max}(0, \beta - f(p_i, r^+) + f(p_i, r^-)) \quad (20)$$

where $r^+, r^- \in R_i$ are an arbitrary pair of reviews for p_i where r^+ has a higher helpfulness score than r^-. β is a scaling factor that magnifies the difference between the score and the margin. Since our MCR model is fully differentiable, it can be trained by gradient descent in an end-to-end manner.

4 Experimental Setup

4.1 Datasets

To the best of our knowledge, there is no benchmark dataset for the Multimodal Review Helpfulness Prediction task (MRHP). Hence, we construct two benchmark datasets (Lazada-MRHP and Amazon-MRHP) from popular e-commerce platforms to evaluate our method.

Lazada-MRHP in Indonesian Lazada.com is a popular platform in Southeast Asia, which is in the Indonesian language. We construct the Lazada-MRHP dataset by crawling the product information (title, description, and images) and user-generated reviews (text content and images) from Lazada. To make sure that the user feedback of helpfulness voting is reliable, we strictly extract the reviews which were published spanning from 2018 to 2019. We focus on three product categories, including *Clothing, Shoes & Jewelry* (CS&J), *Electronics* (Elec.), and *Home & Kitchen* (H&K).

Amazon-MRHP in English The Amazon review dataset (Ni et al., 2019) was collected from Amazon.com, containing meta-data of products and customer reviews from 1996 to 2018. We extract the product information and associated reviews published from 2016 to 2018. Since there are no review images in the original Amazon dataset, we crawl the images for each product and review from the Amazon.com platform. Similar to Lazada-MRHP, the products and reviews also belong to three categories: *Clothing, Shoes & Jewelry* (CS&J), *Electronics* (Elec.), and *Home & Kitchen* (H&K).

Dataset	Category	Instance Number (#P/#R) Train+Dev	Test
Lazada	CS&J	8,245/130,232	2,062/32,274
	Elec.	4,811/52,393	1,204/12,661
	H&K	3,675/46,602	920/12,551
Amazon	CS&J	15,903/348,766	3,966/87,492
	Elec.	13,205/324,907	3,327/79,570
	H&K	18,186/462,225	4,529/111,193

Table 2: Statistics of the two datasets. #P and #R represent the number of products and reviews, respectively.

Learning from user-feedback in review helpfulness prediction has been revealed effective in (Fan et al., 2019; Chen et al., 2019). Specifically, the helpfulness voting received by each review can be treated as the pseudo label indicating the helpfulness level of the review. Following the same data processing as in (Fan et al., 2019), we filter the reviews that received 0 votes in that they are under an unknown user feedback state. Based on the votes received by a review, we leverage a logarithmic interval to categorize reviews into five helpfulness levels. Specifically, we map the number of votes into five intervals (i.e., [1,2), [2, 4), [4, 8), [8, 16), [16, ∞)) based on an exponential with base 2. The five intervals correspond to five helpfulness scores $s_{i,j} \in \{0, 1, 2, 3, 4\}$, where the higher the score, the more helpful the review. Finally, the statistics of the two datasets are shown by Table 2. For both Lazada-MRHP and Amazon-MRHP, we utilize 20% of the training set per category as the validation data.

4.2 Implementation Details

For a fair comparison, we adopt the same data processing for all baselines. We use the ICU tokenizer[1] and NLTK toolkit (Loper and Bird, 2002) to separate text data in Lazada-MRHP and Amazon-MRHP, respectively. Each image is extracted as RoI features with 2048 dimensions. For the net-

[1] http://site.icu-project.org

Type	Method	Clothing			Electronics			Home		
		MAP	N@3	N@5	MAP	N@3	N@5	MAP	N@3	N@5
Text-only	BiMPM	60.0	52.4	57.7	74.4	67.3	72.2	70.6	64.7	69.1
	EG-CNN	60.4	51.7	57.5	73.5	66.3	70.8	70.7	63.4	68.5
	Conv-KNRM	62.1	54.3	59.9	74.1	67.1	71.9	71.4	65.7	70.5
	PRHNet	62.1	54.9	59.9	74.3	67.0	72.2	71.6	65.2	70.0
Multi-modal	SSE-Cross	66.1	59.7	64.8	76.0	68.9	73.8	72.2	66.0	71.0
	D&R Net	66.5	60.7	65.3	76.1	69.2	74.0	72.4	66.3	71.4
	MCR (Ours)	**69.7**	**63.8**	**68.3**	**77.4**	**71.3**	**75.9**	**74.0**	**67.8**	**72.5**

Table 3: Helpfulness review prediction results on the Lazada-MRHP dataset.

Type	Method	Clothing			Electronics			Home		
		MAP	N@3	N@5	MAP	N@3	N@5	MAP	N@3	N@5
Text-only	BiMPM	57.7	41.8	46.0	52.3	40.5	44.1	56.6	43.6	47.6
	EG-CNN	56.4	40.6	44.7	51.5	39.4	42.1	55.3	42.4	46.7
	Conv-KNRM	57.2	41.2	45.6	52.6	40.5	44.2	57.4	44.5	48.4
	PRHNet	58.3	42.2	46.5	52.4	40.1	43.9	57.1	44.3	48.1
Multi-modal	SSE-Cross	65.0	56.0	59.1	53.7	43.8	47.2	60.8	51.0	54.0
	D&R Net	65.2	56.1	59.2	53.9	44.2	47.5	61.2	51.8	54.6
	MCR (Ours)	**67.0**	**58.1**	**61.1**	**56.0**	**46.5**	**49.7**	**63.2**	**54.2**	**57.3**

Table 4: Helpfulness review prediction results on the Amazon-MRHP dataset.

work configurations, we initialize the word embedding layers with the pre-trained 300D GloVE word embeddings[2] for Amazon-MRHP and the fastText multilingual word vectors[3] for Lazada-MRHP. The text n-gram kernels are set as 1, 3, and 5 with 128 hidden dimensions. For the image representations, we set the encoded size of feature d_{l_I} as 128, and the size of common latent space d_c is set to 128. We stack two graph reasoning layers (i.e., $L = 2$) where the hidden dimension of each layer is set to 128. We adopt the Adam optimizer (Kingma and Ba, 2014) to train our model, and the batch size is set to 32. The margin hyperparameter β is set to 1.

4.3 Compared Methods

We compare MCR with several state-of-the-art review helpfulness methods. First, we compare MCR with four strong methods that rely only on the text content of reviews, including the Bilateral Multi-Perspective Matching (BiMPM) model (Wang et al., 2017), Embedding-gated CNN (EG-CNN) (Chen et al., 2018), Convolutional Kernel-based Neural Ranking Model (Conv-KNRM) (Dai et al., 2018), the Product-aware Helpfulness Prediction Network (PRHNet) (Fan et al., 2019).

We are the first to leverage images in the review for helpfulness prediction of multimodal reviews, thereby we compare our MCR model with two strong multimodal reasoning techniques: SSE-Cross (Abavisani et al., 2020) that leverages stochastic shared embedding to fuse different modality representations and D&R Net (Xu et al., 2020) that adopts a decomposition and relation network to model both cross-modality contrast and semantic association.

4.4 Evaluation Metrics

In this paper, we propose a pairwise ranking loss function for review helpfulness prediction, which fully benefits from the sampling of informative negative examples. Since the output of MCR is a list of reviews ranked by their helpfulness scores, we adopt two authoritative ranking-based metrics to evaluate the model performance: Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG@N) (Järvelin and Kekäläinen, 2017). Here, the value of N is set to 3 and 5 in the experiments for NDCG@N. MAP is a widely-used measure method evaluating the general ranking performance on the whole candidate review set, while NDCG@N merely takes into account the top N reviews in the scenario that the customers only read a limited number of reviews.

[2] http://nlp.stanford.edu/data/glove.6B.zip
[3] https://fasttext.cc/docs/en/crawl-vectors.html

5 Experimental Results

5.1 Main Results

Since we adopt the pairwise ranking loss for review helpfulness prediction, we treat the product text as the query, and the associated reviews are viewed as candidates for ranking. Table 3 and Table 4 report the results of MCR and baselines on Lazada-MRHP and Amazon-MRHP, respectively. From the results, we can make the following observations. First, EG-CNN performs worse than other text-only baselines, because EG-CNN only considers the hidden features from the review text, while other text-only methods additionally utilize the product information as a helpfulness signal. Second, the multimodal baselines (SSE-Cross and D&R Net) perform significantly better than text-only baselines. This verifies that multimodal information of reviews can help the models to discover helpful reviews. Third, MCR performs even better than strong multimodal competitors. For example, on Lazada-MRHP, MAP and NDCG@3 increase by 2.9% and 3.5% respectively over the best baseline method (i.e., D&R Net). We can observe similar trends on Amzaon-MRHP. The advantage of MCR comes from its capability of capturing the product-review and intra-review coherence.

5.2 Ablation Study

To analyze the effectiveness of different components of MCR, we conduct detailed ablation studies in terms of removing intra-review coherence (denoted as w/o intra-review), removing intra-modal coherence between product and review images (denoted as w/o intra-modal-I), removing intra-modal coherence between product and review texts (denoted as w/o intra-modal-II), removing inter-modal coherence between review text and product images (denoted as w/o inter-modal-I), and removing inter-modal coherence between review images and product text (denoted as w/o inter-modal-II). The ablation test results on the *CS&J* category of Lazada and Amazon datasets are summarized in Table 5. We can observe that the intra-review coherent reasoning has the largest impact on the performance of MCR. This suggests that the images within a review are informative evidence for review helpfulness prediction. The improvements of the intra-modal and inter-modal coherent reasoning in the product-review coherent reasoning module are also significant. However, intra-modal-I and intra-modal-II have a smaller impact on MCR than the other two variants. This may be because most product images have been always beautified, and there are significant differences between the product images and the images posted by the consumers. It is no surprise that combining all components achieves the best performance on both datasets.

Dataset	Model Variant	MAP	N@3	N@5
Lazada	**MCR** (Ours)	**69.7**	**63.8**	**68.3**
	-w/o intra-review	68.4	62.0	66.9
	-w/o intra-modal-I	69.1	63.0	67.5
	-w/o intra-modal-II	69.2	63.2	67.7
	-w/o inter-modal-I	68.9	62.7	67.3
	-w/o inter-modal-II	68.9	62.5	67.2
Amazon	**MCR** (Ours)	**67.0**	**58.1**	**61.1**
	-w/o intra-review	65.9	57.0	60.1
	-w/o intra-modal-I	66.6	57.7	60.7
	-w/o intra-modal-II	66.8	57.8	60.7
	-w/o inter-modal-I	66.5	57.5	60.5
	-w/o inter-modal-II	66.4	57.5	60.4

Table 5: The ablation study on *Clothing, Shoes& Jewelry* category of Lazada-MRHP and Amazon-MRHP.

5.3 Case Study

To gain more insight into the multimodal review helpfulness prediction task, we use an exemplary case that is selected from the test set of *Home & Kitchen* category of Amazon-MRHP to empirically investigate the effectiveness of our model. Table 6 shows a product and two associated reviews with ground-truth helpfulness scores voted by consumers. These two reviews are ranked correctly by our MCR method while being wrongly ranked by strong baselines (e.g., Conv-KNRM and PRHNet). The text content of both reviews contains negative emotion words (e.g., "disappointed" and "sad") and expresses similar information "*the product size does not meet my expectation*". It is hard for text-only methods to discriminate the helpfulness of these two reviews via solely considering the text content of reviews. After analyzing the images within the reviews, we can reveal that the *Review 1* is helpful since it provides two appropriate bed images with a brought comforter as evidence that can well support his/her claim in the text content. However, *Review 2* provides an inappropriate image with the product package, which cannot well support the claim of product size. This verifies that it is essential to capture the complex semantic relationship between the images and text content within a review for helpfulness prediction.

Product Information
Bedding printed comforter set (king, grey) with 2 pillow shams - luxurious soft brushed microfiber - goose down alternative comforter


Review 1 (Helpfulness Score: 4)
Though I like the color and look, I am very disappointed in the size. The picture on amazon shows the comforter going all the way to the floor. To be sure, I ordered the king size. As you can see in the photos, I have a queen bed and the comforter still has 18" to the floor on each side. I will try to fix it with a bed skirt.


Review 2 (Helpfulness Score: 1)
This comforter is very fluffy and does have a nice feel to it, but is far too small to actually cover much more than the top of the bed. In the picture, it nearly touched the floor on both visible sides. Likewise, it was described as a printed comforter set (grey, queen) with 2 pillow shams - luxurious soft brushed microfiber - goose down alternative comforter by utopia bedding but the item itself said nothing of being a down alternative. I'm sad that this doesn't meet my expectations.

Table 6: An example product and two associated reviews. We use underlines to highlight main opinions.

6 Conclusion

Multimodal review analysis (MRA) is extremely important for helping businesses and consumers quickly acquire valuable information from user-generated reviews. This paper is the first attempt to explore the multimodal review helpfulness prediction (MRHP) task, which aims at analyzing the review helpfulness from text and images. We propose a multi-perspective coherent reasoning (MCR) method to solve MRHP task, which fully explores the product-review coherence and intra-review coherence from both textual and visual modalities. In addition, we construct two multimodal review datasets to evaluate the effectiveness of MCR, which may push forward the research in this field. Extensive experimental results demonstrate that MCR significantly outperforms baselines by comprehensively exploiting the images associated with the reviews.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (No. 61906185), Natural Science Foundation of Guangdong Province of China (No. 2019A1515011705), Youth Innovation Promotion Association of CAS China (No. 2020357), Shenzhen Science and Technology Innovation Program (Grant No. KQTD20190929172835662), Shenzhen Basic Research Foundation (No. JCYJ20200109113441941).

References

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. 2019. Multi-domain gated cnn for review helpfulness prediction. In *The World Wide Web Conference*, pages 2630–2636.

Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest 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, Volume 2 (Short Papers)*, pages 602–607.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708.

Miao Fan, Chao Feng, Lin Guo, Mingming Sun, and Ping Li. 2019. Product-aware helpfulness prediction of online reviews. In *The World Wide Web Conference*, pages 2715–2721.

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 *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 343–350. IEEE.

Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA.

Soo-Min Kim, Patrick Pantel, Timothy Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *EMNLP*, pages 423–430.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Srikumar Krishnamoorthy. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

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*, pages 1358–1363.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. *arXiv preprint cs/0205028*.

Lionel Martin and Pearl Pu. 2014. Prediction of helpful reviews using emotions extraction. In *AAAI*, pages 1551–1557.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Xiaoru Qu, Zhao Li, Jialin Wang, Zhipeng Zhang, Pengcheng Zou, Junxiao Jiang, Jiaming Huang, Rong Xiao, Ji Zhang, and Jun Gao. 2020. Category-aware graph neural networks for improving e-commerce review helpfulness prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2693–2700.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786.

Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *ACL*, pages 38–44.