

Gradient-Boosted Decision Tree for Listwise Context Model in Multimodal Review Helpfulness Prediction

Thong Nguyen¹, Xiaobao Wu², Xinshuai Dong³, Anh Tuan Luu^{2*},
Cong-Duy Nguyen², Zhen Hai⁴, Lidong Bing⁴

¹National University of Singapore, Singapore

²Nanyang Technological University, Singapore

³Carnegie Mellon University, USA

⁴DAMO Academy, Alibaba Group

e0998147@u.nus.edu, anhtuan.luu@ntu.edu.sg

Abstract

Multimodal Review Helpfulness Prediction (MRHP) aims to rank product reviews based on predicted helpfulness scores and has been widely applied in e-commerce via presenting customers with useful reviews. Previous studies commonly employ fully-connected neural networks (FCNNs) as the final score predictor and pairwise loss as the training objective. However, FCNNs have been shown to perform inefficient splitting for review features, making the model difficult to clearly differentiate helpful from unhelpful reviews. Furthermore, pairwise objective, which works on review pairs, may not completely capture the MRHP goal to produce the ranking for the entire review list, and possibly induces low generalization during testing. To address these issues, we propose a listwise attention network that clearly captures the MRHP ranking context and a listwise optimization objective that enhances model generalization. We further propose gradient-boosted decision tree as the score predictor to efficaciously partition product reviews' representations. Extensive experiments demonstrate that our method achieves state-of-the-art results and polished generalization performance on two large-scale MRHP benchmark datasets.

1 Introduction

E-commerce platforms, such as Amazon and Lazada, have achieved steady development. These platforms generally provide purchasers' reviews to supply justification information for new consumers and help them make decisions. Nevertheless, the quality and usefulness of reviews can vary hugely: some are helpful with coherent and informative content while others unhelpful with trivial or irrelevant information. Due to this, the Multimodal Review Helpfulness Prediction (MRHP) task is proposed. It ranks the reviews by predicting their helpfulness scores based on the textual and visual

modality of products and reviews, because helpful reviews should comprise not only precise and informative textual material, but also consistent images with text content (Liu et al., 2021; Nguyen et al., 2022). This can help consumers find helpful reviews instead of unhelpful ones, resulting in more appealing E-commerce platforms.

In MRHP, multimodal reviews naturally form ranking partitions based on user votings, where each partition exhibits distinct helpfulness feature level (Ma et al., 2021). As such, the MRHP score regressor's function is to assign scores to indicate the partition for hidden features of product reviews. However, current MRHP approaches employ fully-connected neural networks (FCNNs), which cannot fulfill the partition objective. In particular, FCNNs are ineffective in feature scaling and transformation, thus being inept at feature space splitting and failing to work efficiently in ranking problems that involve ranking partitions (Beutel et al., 2018; Qin et al., 2021). An illustration would be in Figure 1, where the helpfulness scores predicted by FCNNs do not lucidly separate helpful and unhelpful reviews. Severely, some unhelpful reviews possess logits that can even stay in the range of helpful ones, bringing about fallacious ranking.

In addition to incompetent model architectures, existing MRHP frameworks also employ suboptimal loss function: they are mostly trained on a pairwise loss to learn review preferences, which unfortunately mismatches the listwise nature of review ordering prediction. Firstly, the mismatch might empirically give rise to inefficient ranking performance (Pasumarthi et al., 2019; Pobrotyn and Białobrzeski, 2021). Second, pairwise training loss considers all pairs of review as equivalent. In consequence, the loss cannot differentiate a pair of useful and not useful reviews from a pair of moderately useful and not useful ones, which results in a model that distinguishes poorly between useful and moderately useful reviews.

*Corresponding Author

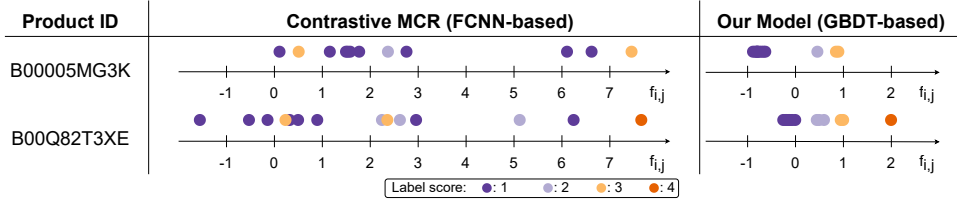


Figure 1: Examples of helpfulness scores produced by score regressors built upon neural network and gradient-boosted decision tree. We present the content of the product and review samples in Appendix E.

To address these issues, we first propose a Gradient-Boosted Decision Tree (GBDT) as the helpfulness score regressor to utilize both its huge capacity of partitioning feature space (Leboeuf et al., 2020) and differentiability compared with standard decision trees for end-to-end training. We achieve the partition capability with the split (internal) nodes of the tree implemented with non-linear single perceptron, to route review features to the specific subspace in a soft manner.

Furthermore, we develop a theoretical analysis to demonstrate that pairwise training indeed has lower model generalization than listwise approach. We proceed to propose a novel listwise training objective for the proposed MRHP architecture. We also equip our architecture with a listwise attention network that models the interaction among the reviews to capture the listwise context for the MRHP ranking task.

In sum, our contributions are four-fold:

- We propose a novel gradient-boosted decision tree score predictor for multimodal review helpfulness prediction (MRHP) to partition product review features and properly infer helpfulness score distribution.
- We propose a novel listwise attention module for the MRHP architecture that conforms to the listwise context of the MRHP task by relating reviews in the list.
- We perform theoretical study with the motivation of ameliorating the model generalization error, and accordingly propose a novel MRHP training objective which satisfies our aim.
- We conducted comprehensive experiments on two benchmark datasets and found that our approach significantly outperforms both text-only and multimodal baselines, and accomplishes state-of-the-art results for MRHP.

2 Background

In this section, we recall the Multimodal Review Helpfulness Prediction (MRHP) problem. Then, we introduce theoretical preliminaries which form the basis of our formal analysis of the ranking losses for the MRHP problem in the next section.

2.1 Problem Definition

Following (Liu et al., 2021; Han et al., 2022; Nguyen et al., 2022), we formulate MRHP as a ranking task. In detail, we consider an instance X_i to consist of a product item p_i , composed of product description T^{p_i} and images I^{p_i} , and its respective review list $R_i = \{r_{i,1}, r_{i,2}, \dots, r_{i,|R_i|}\}$. Each review $r_{i,j}$ carries user-generated text $T^{r_{i,j}}$, images $I^{r_{i,j}}$, and an integer scalar label $y_{i,j} \in \{0, 1, \dots, S\}$ denoting the helpfulness score of review $r_{i,j}$. The ground-truth result associated with X_i is the descending order determined by the helpfulness score list $Y_i = \{y_{i,1}, y_{i,2}, \dots, y_{i,|R_i|}\}$. The MRHP task is to generate helpfulness scores which match the groundtruth ranking order, formulated as follows:

$$s_{i,j} = f(p_i, r_{i,j}), \quad (1)$$

where f represents the helpfulness prediction model taking $\langle p_i, r_{i,j} \rangle$ as the input.

2.2 Analysis of Generalization Error

The analysis involves the problem of learning a deep θ -parameterized model $f^\theta : \mathcal{X} \rightarrow \mathcal{Y}$ that maps the input space \mathcal{X} to output space \mathcal{Y} and a stochastic learning algorithm \mathcal{A} to solve the optimization problem as follows:

$$f^{\theta^*} = \arg \min_{f^\theta} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}} [l(f^\theta; (\mathbf{x}, \mathbf{y}))], \quad (2)$$

where \mathbb{P} denotes the distribution of (\mathbf{x}, \mathbf{y}) , l the loss function on the basis of the difference between $\hat{\mathbf{y}} = f^\theta(\mathbf{x})$ and \mathbf{y} , and $R_{\text{true}}(f^\theta) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}} [l(f^\theta; (\mathbf{x}, \mathbf{y}))]$ is dubbed as the true risk.

Since \mathbb{P} is unknown, R_{true} is alternatively solved through optimizing a surrogate empirical risk $R_{\text{emp}}(f_{\mathcal{D}}^{\theta}) = \frac{1}{N} \sum_{i=1}^N l(f^{\theta}; (\mathbf{x}_i, \mathbf{y}_i))$, where $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ denotes a training dataset drawn from \mathbb{P} that $f_{\mathcal{D}}^{\theta}$ is trained upon.

Because the aim of deep neural model training is to produce a model f^{θ} that provides a small gap between the performance over \mathcal{D} , i.e. $R_{\text{emp}}(f_{\mathcal{D}}^{\theta})$, and over any unseen test set from \mathbb{P} , i.e. $R_{\text{true}}(f_{\mathcal{D}}^{\theta})$, the analysis defines the main focus to be the generalization error $E(f_{\mathcal{D}}^{\theta}) = R_{\text{true}}(f_{\mathcal{D}}^{\theta}) - R_{\text{emp}}(f_{\mathcal{D}}^{\theta})$, the objective to be achieving a tight bound of $E(f_{\mathcal{D}}^{\theta})$, and subsequently the foundation regarding the loss function’s Lipschitzness as:

Definition 1. (*Lipschitzness*). A loss function $l(\hat{\mathbf{y}}, \mathbf{y})$ is γ -Lipschitz with respect to $\hat{\mathbf{y}}$ if for $\gamma \geq 0$, $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^K$, we have:

$$|l(\mathbf{u}, \mathbf{y}) - l(\mathbf{v}, \mathbf{y})| \leq \gamma \|\mathbf{u} - \mathbf{v}\|, \quad (3)$$

where $\|\cdot\|$ denotes the l_1 -norm, K the dimension of the output $\hat{\mathbf{y}}$.

Given the foundation, we have the connection between the properties of loss functions and the generalization error:

Theorem 1. Consider a loss function that $0 \leq l(\hat{\mathbf{y}}, \mathbf{y}) \leq L$ that is convex and γ -Lipschitz with respect to $\hat{\mathbf{y}}$. Suppose the stochastic learning algorithm \mathcal{A} is executed for T iterations, with an annealing rate λ_t to solve problem (2). Then, the following generalization error bound holds with probability at least $1 - \delta$ (Akbari et al., 2021):

$$E(f_{\mathcal{D}}^{\theta}) = R_{\text{true}}(f_{\mathcal{D}}^{\theta}) - R_{\text{emp}}(f_{\mathcal{D}}^{\theta}) \leq L \sqrt{\frac{\log(2/\delta)}{2N}} + 2\gamma^2 \sum_{t=1}^T \lambda_t \left(2\sqrt{\frac{\log(2/\delta)}{T}} + \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{1}{N} \right). \quad (4)$$

Theorem (1) implies that by establishing a loss function \mathcal{L} with smaller values of γ and L , we can burnish the model generalization performance.

3 Methodology

In this section, we elaborate on our proposed architecture, listwise attention network, tree-based helpfulness regressor, and listwise ranking loss along with its comparison against the pairwise one from the theoretical perspective. The overall architecture is illustrated in Figure 2.

3.1 Multimodal Encoding

Our model receives product description T^{p_i} , product images I^{p_i} , review text $T^{r_{i,j}}$, and review images $I^{r_{i,j}}$ as input. We perform the encoding procedure for those inputs as follows.

Textual Encoding. For both product text T^{p_i} and review text $T^{r_{i,j}}$, we index their sequences of words into the word embeddings and forward to the respective LSTM layer to yield token-wise representations:

$$\mathbf{H}^{p_i} = \text{LSTM}^p(\mathbf{W}_{\text{emb}}(T^{p_i})), \quad (5)$$

$$\mathbf{H}^{r_{i,j}} = \text{LSTM}^r(\mathbf{W}_{\text{emb}}(T^{r_{i,j}})), \quad (6)$$

where $\mathbf{H}^{p_i} \in \mathbb{R}^{l^{p_i} \times d}$, $\mathbf{H}^{r_{i,j}} \in \mathbb{R}^{l^{r_{i,j}} \times d}$, l^{p_i} and $l^{r_{i,j}}$ denote sequence lengths of the product and review text, respectively, d the hidden dimension.

Visual Encoding. We adapt a pre-trained Faster R-CNN to extract ROI features of m objects $\{\mathbf{e}_t^{p_i}\}_{t=1}^m$ and $\{\mathbf{e}_t^{r_{i,j}}\}_{t=1}^m$ for product and review images, respectively. We then feed those object features into the self-attention module to obtain visual representations as:

$$\mathbf{V}^{p_i} = \text{SelfAttn}(\{\mathbf{e}_t^{p_i}\}_{t=1}^m), \quad (7)$$

$$\mathbf{V}^{r_{i,j}} = \text{SelfAttn}(\{\mathbf{e}_t^{r_{i,j}}\}_{t=1}^m), \quad (8)$$

where $\mathbf{V}^{p_i}, \mathbf{V}^{r_{i,j}} \in \mathbb{R}^{m \times d}$, and d denotes the hidden size.

3.2 Coherence Reasoning

We then learn intra-modal, inter-modal, and intra-entity coherence among product-review elements.

Intra-modal Coherence. There are two types of intra-modal coherence relations: (1) product text - review text and (2) product image - review image. Initially, we designate self-attention modules to capture the intra-modal interaction as:

$$\mathbf{H}_{i,j}^{\text{intraM}} = \text{SelfAttn}([\mathbf{H}^{p_i}, \mathbf{H}^{r_{i,j}}]), \quad (9)$$

$$\mathbf{V}_{i,j}^{\text{intraM}} = \text{SelfAttn}([\mathbf{V}^{p_i}, \mathbf{V}^{r_{i,j}}]). \quad (10)$$

Then, intra-modal interaction features are passed to a CNN, then condensed into hidden vectors via pooling layer:

$$\mathbf{z}_{i,j}^{\text{intraM}} = \text{Pool}(\text{CNN}([\mathbf{H}_{i,j}^{\text{intraM}}, \mathbf{V}_{i,j}^{\text{intraM}}])), \quad (11)$$

where $[\cdot]$ denotes the concatenation operator.

Inter-modal Coherence. The inter-modal coherence comprises two relation types: (1) product text (pt) - review image (ri) and (2) product image (pi) -

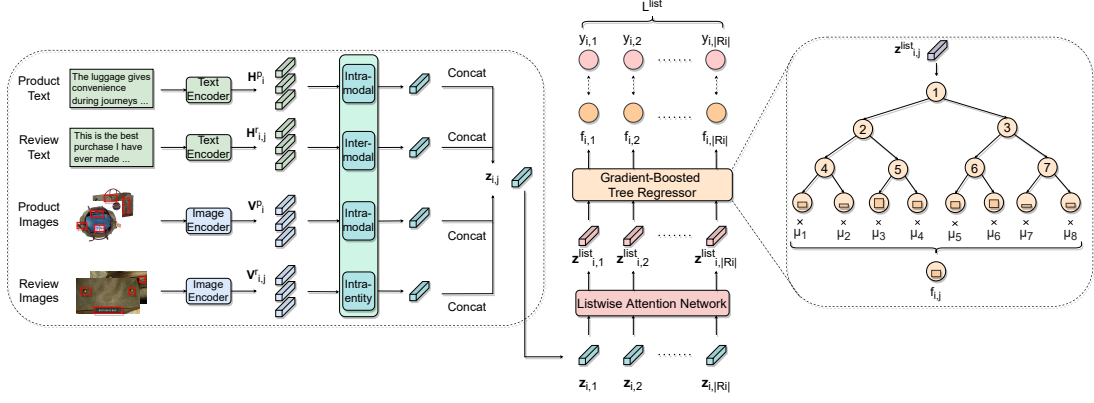


Figure 2: Illustration of our Multimodal Review Helpfulness Prediction model.

review text (rt). Similar to the intra-modal coherence, we first perform cross-modal correlation by leveraging the self-attention mechanism:

$$\mathbf{H}_{i,j}^{\text{pt-ri}} = \text{SelfAttn}([\mathbf{H}^{p_i}, \mathbf{V}^{r_{i,j}}]), \quad (12)$$

$$\mathbf{H}_{i,j}^{\text{pi-rt}} = \text{SelfAttn}([\mathbf{V}^{p_i}, \mathbf{H}^{r_{i,j}}]). \quad (13)$$

Thereafter, we pool the above features and concatenate the pooled vectors to attain the inter-modal vector:

$$\mathbf{z}_{i,j}^{\text{pt-ri}} = \text{Pool}(\mathbf{H}_{i,j}^{\text{pt-ri}}), \quad (14)$$

$$\mathbf{z}_{i,j}^{\text{pi-rt}} = \text{Pool}(\mathbf{H}_{i,j}^{\text{pi-rt}}), \quad (15)$$

$$\mathbf{z}_{i,j}^{\text{interM}} = [\mathbf{z}_{i,j}^{\text{pt-ri}}, \mathbf{z}_{i,j}^{\text{pi-rt}}]. \quad (16)$$

Intra-entity Coherence. Analogous to the inter-modal coherence, we also conduct self-attention and pooling computation, but on the (1) product text (pt) - product image (pi) and (2) review text (rt) - review image (ri) as follows:

$$\mathbf{H}_i^{\text{pt-pi}} = \text{SelfAttn}([\mathbf{H}^{p_i}, \mathbf{V}^{p_i}], \quad (17)$$

$$\mathbf{H}_{i,j}^{\text{rt-ri}} = \text{SelfAttn}([\mathbf{H}^{r_{i,j}}, \mathbf{V}^{r_{i,j}}]), \quad (18)$$

$$\mathbf{z}_i^{\text{pt-pi}} = \text{Pool}(\mathbf{H}_i^{\text{pt-pi}}), \quad (19)$$

$$\mathbf{z}_{i,j}^{\text{rt-ri}} = \text{Pool}(\mathbf{H}_{i,j}^{\text{rt-ri}}), \quad (20)$$

$$\mathbf{z}_{i,j}^{\text{intraR}} = [\mathbf{z}_i^{\text{pt-pi}}, \mathbf{z}_{i,j}^{\text{rt-ri}}]. \quad (21)$$

Eventually, the concatenation of the intra-modal, inter-modal, and intra-entity vectors becomes the result of the coherence reasoning phase:

$$\mathbf{z}_{i,j} = [\mathbf{z}_{i,j}^{\text{intraM}}, \mathbf{z}_{i,j}^{\text{interM}}, \mathbf{z}_{i,j}^{\text{intraR}}]. \quad (22)$$

3.3 Listwise Attention Network

In our proposed listwise attention network, we encode list-contextualized representations to consider relative relationship among reviews. We

achieve this by utilizing self-attention mechanism to relate list-independent product reviews' features $\{\mathbf{z}_{i,1}, \mathbf{z}_{i,2}, \dots, \mathbf{z}_{i,|R_i|}\}$ as follows:

$$\{\mathbf{z}_{i,j}^{\text{list}}\}_{j=1}^{|R_i|} = \text{SelfAttn}(\{\mathbf{z}_{i,j}\}_{j=1}^{|R_i|}), \quad (23)$$

where R_i denotes the review list associated with product p_i .

3.4 Gradient-boosted Decision Tree for Helpfulness Estimation

In this section, we delineate our gradient-boosted decision tree to predict helpfulness scores that efficaciously partition review features.

Tree Structure. We construct a d_{tree} -depth binary decision tree composed of internal nodes \mathcal{N} ($|\mathcal{N}| = 2^{d_{\text{tree}}-1} - 1$) and leaf nodes \mathcal{L} ($|\mathcal{L}| = 2^{d_{\text{tree}}}$). Our overall tree structure is depicted in Figure 2.

Score Prediction. Receiving the list-attended vectors $\{\mathbf{z}_i^{\text{list}}\}_{i=1}^N$, our decision tree performs soft partitioning through probabilistic routing for those vectors to their target leaf nodes. In such manner, each internal node n calculates the routing decision probability as:

$$p_n^{\text{left}} = \sigma(\text{Linear}(\mathbf{z}^{\text{list}})), \quad (24)$$

$$p_n^{\text{right}} = 1 - p_n^{\text{left}}, \quad (25)$$

where p_n^{left} and p_n^{right} denote the likelihood of directing the vector to the left sub-tree and right sub-tree, respectively. Thereupon, the probability of reaching leaf node l is formulated as follows:

$$\mu_l = \prod_{n \in \mathcal{P}(l)} (p_n^{\text{left}})^{\mathbb{1}^{l_n}} \cdot (p_n^{\text{right}})^{\mathbb{1}^{r_n}}, \quad (26)$$

where $\mathbb{1}^{l_n}$ denotes the indicator function of whether leaf node l belongs to the left sub-tree of the

internal node n , equivalently for $\mathbb{1}^{rn}$, and $\mathcal{P}(l)$ the node sequence path to leaf l . For example, in Figure 2, the routing probability to leaf 6 is $\mu_6 = p_1^{\text{right}} p_3^{\text{left}} p_6^{\text{right}}$.

For the score inference at leaf node l , we employ a linear layer for calculation as follows:

$$s_{l,i,j} = \text{Linear}_l(\mathbf{z}_{i,j}^{\text{list}}). \quad (27)$$

where $s_{l,i,j}$ denotes the helpfulness score generated at leaf node l . Lastly, due to the probabilistic routing approach, the final helpfulness score $f_{i,j}$ is the average of the leaf node scores weighted by the probabilities of reaching the leaves:

$$f_{i,j} = f(p_i, r_{i,j}) = \sum_{l \in \mathcal{L}} s_{l,i,j} \cdot \mu_l. \quad (28)$$

3.5 Listwise Ranking Objective

Since MRHP task aims to produce helpfulness order for a list of reviews, we propose to follow a listwise approach to compare the predicted helpfulness scores with the groundtruth.

Initially, we convert two lists of prediction scores $\{f_{i,j}\}_{j=1}^{|R_i|}$ and groundtruth labels $\{y_{i,j}\}_{j=1}^{|R_i|}$ into two probability distributions.

$$f'_{i,j} = \frac{\exp(f_{i,j})}{\sum_{t=1}^{|R_i|} \exp(f_{i,t})}, \quad y'_{i,j} = \frac{\exp(y_{i,j})}{\sum_{t=1}^{|R_i|} \exp(y_{i,t})}. \quad (29)$$

Subsequently, we conduct theoretical derivation and arrive in interesting properties of the listwise computation.

Theoretical Derivation. Our derivation demonstrates that discrimination computation of both listwise and pairwise functions (Liu et al., 2021; Han et al., 2022; Nguyen et al., 2022) satisfy the preconditions in Theorem (1).

Lemma 1. *Given listwise discrimination function on the total training set as $\mathcal{L}^{\text{list}} = -\sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \log(f'_{i,j})$, where P denotes the product set, then $\mathcal{L}^{\text{list}}$ is convex and γ^{list} -Lipschitz with respect to $f'_{i,j}$.*

Lemma 2. *Given pairwise discrimination function on the total training set as $\mathcal{L}^{\text{pair}} = \sum_{i=1}^{|P|} [-f_{i,r^+} + f_{i,r^-} + \alpha]^+$, where r^+, r^- denote two random indices in R_i and $y_{i,r^+} > y_{i,r^-}$, and $\alpha = \max_{1 \leq j \leq |R_i|} (y_{i,j}) - \min_{1 \leq j \leq |R_i|} (y_{i,j})$, then $\mathcal{L}^{\text{pair}}$ is convex and γ^{pair} -Lipschitz with respect to f_{i,r^+}, f_{i,r^-} .*

Based upon the above theoretical basis, we investigate the connection between $\mathcal{L}^{\text{list}}$ and $\mathcal{L}^{\text{pair}}$.

Theorem 2. *Let $\mathcal{L}^{\text{list}}$ and $\mathcal{L}^{\text{pair}}$ are γ^{list} -Lipschitz and γ^{pair} -Lipschitz, respectively. Then, the following inequality holds:*

$$\gamma^{\text{list}} \leq \gamma^{\text{pair}}. \quad (30)$$

Theorem 3. *Let $0 \leq \mathcal{L}^{\text{list}} \leq L^{\text{list}}$ and $0 \leq \mathcal{L}^{\text{pair}} \leq L^{\text{pair}}$. Then, the following inequality holds:*

$$L^{\text{list}} \leq L^{\text{pair}}. \quad (31)$$

We combine Theorem (1), (2), and (3), to achieve the following result.

Theorem 4. *Consider two models $f_{\mathcal{D}}^{\text{list}}$ and $f_{\mathcal{D}}^{\text{pair}}$ under common settings trained to minimize $\mathcal{L}^{\text{list}}$ and $\mathcal{L}^{\text{pair}}$, respectively, on dataset $\mathcal{D} = \{p_i, \{r_{i,j}\}_{j=1}^{|R_i|}\}_{i=1}^{|P|}$. Then, we have the following inequality:*

$$E(f_{\mathcal{D}}^{\text{list}}) \leq E(f_{\mathcal{D}}^{\text{pair}}), \quad (32)$$

where $E(f_{\mathcal{D}}) = R_{\text{true}}(f_{\mathcal{D}}) - R_{\text{emp}}(f_{\mathcal{D}})$.

As in Theorem (4), models optimized by listwise function achieve a tighter bound on the generalization error than the ones with the pairwise function, thus upholding better generalization performance. We provide proofs of all the lemmas and theorems in Appendix A. Indeed, empirical results in Section 4.6 also verify our theorems.

With such foundation, we propose to utilize listwise discrimination as the objective loss function to train our MRHP model:

$$\mathcal{L}^{\text{list}} = -\sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \log(f'_{i,j}). \quad (33)$$

4 Experiments

4.1 Datasets

For evaluation, we conduct experiments on two large-scale MRHP benchmark datasets: Lazada-MRHP and Amazon-MRHP. We present the dataset statistics in Appendix B.

Amazon-MRHP (Liu et al., 2021) includes crawled product and review content from Amazon.com, the international e-commerce brand, between 2016 and 2018. All of the product and review texts are expressed in English.

Lazada-MRHP (Liu et al., 2021) comprises product information and user-generated reviews from Lazada.com, a popular e-commerce platform in

Southeast Asia. Both product and review texts are written in Indonesian.

Both datasets are composed of 3 categories: (1) *Clothing, Shoes & Jewelry* (Clothing), (2) *Electronics* (Electronics), and (3) *Home & Kitchen* (Home). We divide the helpfulness votes of the reviews into 5 partitions, i.e. $[1, 2)$, $[2, 4)$, $[4, 8)$, $[8, 16)$, and $[16, \infty)$, corresponding to 5 helpfulness scores, i.e. $y_{i,j} \in \{0, 1, 2, 3, 4\}$.

4.2 Implementation Details

For input texts, we leverage pretrained word embeddings with fastText embedding (Bojanowski et al., 2017) and 300-dimensional GloVe word vectors (Pennington et al., 2014) for Lazada-MRHP and Amazon-MRHP datasets, respectively. Each embedded word sequence is passed into an 1-layer LSTM whose hidden dimension is 128. For input images, we extract their ROI features of 2048 dimensions and encode them into 128-dimensional vectors. Our gradient-boosted decision tree score predictor respectively exhibits a depth of 3 and 5 in Lazada-MRHP and Amazon-MRHP datasets, which are determined on the validation performance. We adopt Adam optimizer, whose batch size is 32 and learning rate $1e-3$, to train our entire architecture in the end-to-end fashion.

4.3 Baselines

We compare our approach with an encyclopedic list of baselines:

- **BiMPM** (Wang et al., 2017): a ranking model that uses 2 BiLSTM layers to encode input sentences.
- **EG-CNN** (Chen et al., 2018): a RHP baseline which leverages character-level representations and domain discriminator to improve cross-domain RHP performance.
- **Conv-KNRM** (Dai et al., 2018): a CNN-based system which uses kernel pooling on multi-level n-gram encodings to produce ranking scores.
- **PRH-Net** (Fan et al., 2019): a RHP baseline that receives product metadata and raw review text as input.
- **SSE-Cross** (Abavisani et al., 2020): a cross-modal attention-based approach to filter non-salient elements in both visual and textual input components.

- **DR-Net** (Xu et al., 2020): a combined model of decomposition and relation networks to learn cross-modal association.
- **MCR** (Liu et al., 2021): an MRHP model that infers helpfulness scores based on cross-modal attention-based encodings.
- **SANCL** (Han et al., 2022): a baseline which extracts salient multimodal entries via probe-based attention and applies contrastive learning to refine cross-modal representations.
- **Contrastive-MCR** (Nguyen et al., 2022): an MRHP approach utilizing adaptive contrastive strategy to enhance cross-modal representations and performance optimization.

4.4 Main Results

Inspired by previous works (Liu et al., 2021; Han et al., 2022; Nguyen et al., 2022), we report Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG@N), where $N = 3$ and $N = 5$. We include the performance of baseline models and our approach in Table 1 and 2.

On Amazon dataset, we consistently outperform prior methods of both textual and multimodal settings. Particularly, our architecture improves over Contrastive-MCR on MAP of 15.2 points in Clothing, NDCG@3 of 20.4 points in Electronics, and NDCG@5 of 21.0 points in Home subset. Furthermore, we accomplish a gain in MAP of 2.2 points in Clothing over PRH-Net, NDCG@3 of 16.4 points in Electronics and NDCG@5 of 11.8 points in Home category over Conv-KNRM baseline, where PRH-Net and Conv-KNRM are the best prior text-only baselines.

For Lazada dataset, which is in Indonesian, we outperform Contrastive-MCR with a significant margin of MAP of 10.4 points in Home, NDCG@5 of 11.6 points in Electronics, and NDCG@3 of 12.4 points in Clothing domain. The text-only variant of our model also gains a considerable improvement of 4.7 points of NDCG@5 in Clothing, 5.0 points of MAP in Electronics over PRH-Net, and 1.4 points of NDCG@3 in Home over Conv-KNRM model.

These outcomes demonstrate that our method is able to produce more sensible helpfulness scores to polish the review ranking process, not only being efficacious in English but also generalizing to other language as well. Over and above, it is worth pointing out in Lazada-Electronics, the textual setting of our approach even achieves higher helpfulness

Setting	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
	PRH-Net	58.3	42.2	46.5	52.4	40.1	43.9	57.1	44.3	48.1
	Our Model	60.5	51.7	52.8	59.8	56.9	57.9	63.4	59.4	60.2
Multimodal	SSE-Cross	65.0	56.0	59.1	53.7	43.8	47.2	60.8	51.0	54.0
	DR-Net	65.2	56.1	59.2	53.9	44.2	47.5	61.2	51.8	54.6
	MCR	66.4	57.3	60.2	54.4	45.0	48.1	62.6	53.5	56.6
	SANCL	67.3	58.6	61.5	56.2	47.0	49.9	63.4	54.3	57.4
	Contrastive-MCR	67.4	58.6	61.6	56.5	47.6	50.8	63.5	54.6	57.8
	Our Model	82.6	80.3	79.3	74.2	68.0	69.8	81.7	76.5	78.8

Table 1: Helpfulness review prediction results on the Amazon-MRHP dataset.

Setting	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
	PRH-Net	62.1	54.9	59.9	74.3	67.0	72.2	71.6	65.2	70.0
	Our Model	66.4	59.6	64.6	79.3	63.8	78.0	72.9	67.1	71.5
Multimodal	SSE-Cross	66.1	59.7	64.8	76.0	68.9	73.8	72.2	66.0	71.0
	DR-Net	66.5	60.7	65.3	76.1	69.2	74.0	72.4	66.3	71.4
	MCR	68.8	62.3	67.0	76.8	70.7	75.0	73.8	67.0	72.2
	SANCL	70.2	64.6	68.8	77.8	71.5	76.1	75.1	68.4	73.3
	Contrastive-MCR	70.3	64.7	69.0	78.2	72.4	76.5	75.2	68.8	73.7
	Our Model	78.5	77.1	79.0	87.9	86.7	88.1	85.6	78.8	83.1

Table 2: Helpfulness review prediction results on the Lazada-MRHP dataset.

Dataset	Model	MAP	N@3	N@5
Amazon	Our Model	81.7	76.5	78.8
	- w/ $d_{z_{i,j}}$ -8-4-2-1 NN	64.6	55.2	58.6
	- w/ $d_{z_{i,j}}$ -32-16-8-4-2-1 NN	70.6	59.8	63.8
	- w/ $d_{z_{i,j}}$ -32-32-32-32-1 NN	64.9	57.1	59.9
	- w/o $\mathcal{L}^{\text{list}}$	72.4	64.7	67.1
	- w/o LAN	64.8	55.8	59.3
Lazada	Our Model	85.6	78.8	83.1
	- w/ $d_{z_{i,j}}$ -8-4-2-1 NN	76.2	69.3	74.3
	- w/ $d_{z_{i,j}}$ -32-16-8-4-2-1 NN	78.7	71.9	77.6
	- w/ $d_{z_{i,j}}$ -32-32-32-32-1 NN	77.6	70.9	75.2
	- w/o $\mathcal{L}^{\text{list}}$	78.0	71.3	75.8
	- w/o LAN	76.5	69.9	74.4

Table 3: Ablation study on the Home category of Amazon-MRHP and Lazada-MRHP datasets.

prediction capacity than the state-of-the-art multimodal baseline, i.e. the Contrastive-MCR model.

4.5 Ablation Study

To verify the impact of our proposed (1) Gradient-boosted decision tree regressor, (2) Listwise ranking loss, and (3) Listwise attention network, we conduct ablation experiments on the Home category of the Amazon and Lazada datasets.

GBDT Regressor. In this ablation, we substitute our tree-based score predictor with various FCNNs score regressor. Specifically, we describe each substitution with a sequence of dimensions in its

fully-connected layers, and each hidden layer is furnished with a Tanh activation function.

As shown in Table 3, FCNN-based score regressors considerably hurt the MRHP performance, with a decline of NDCG@3 of 16.7 points, and MAP of 6.9 points in the Amazon and Lazada datasets, respectively. One potential explanation is that without the decision tree predictor, the model lacks the partitioning ability to segregate the features of helpful and non-helpful reviews.

Listwise Ranking Loss. As can be observed in Table 3, replacing listwise objective with the pairwise one degrades the MRHP performance substantially, with a drop of NDCG@3 of 11.8 scores in Amazon, and NDCG@5 of 7.3 scores in Lazada dataset. Based on Theorem 4 and Table 4, we postulate that removing listwise training objective impairs model generalization, revealed in the degraded MRHP testing performance.

Listwise Attention Network (LAN). We proceed to ablate our proposed listwise attention module and re-execute the model training. Results in Table 3 betray that inserting listwise attention brings about performance upgrade with 16.9 and 9.1 points of MAP in Amazon-MRHP and Lazada-

Category-Dataset	Method	Training MAP	Testing MAP	Δ_{MAP}
Electronics-Amazon	$f_D^{\theta, \text{pair}}$	89.3	68.8	20.5
	$f_D^{\theta, \text{list}}$	78.4	74.2	4.2
Electronics-Lazada	$f_D^{\theta, \text{pair}}$	91.5	70.1	21.4
	$f_D^{\theta, \text{list}}$	89.7	87.9	1.8

Table 4: Training-testing performance of our model trained with listwise and pairwise ranking losses on the Electronics category of Amazon and Lazada datasets.

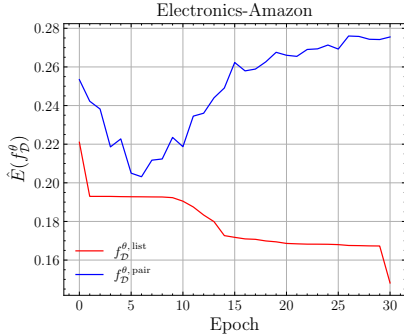


Figure 3: Generalization error curves per training epoch on the Electronics category in Amazon-MRHP dataset.

MRHP, respectively. We can attribute the improvement to the advantage of listwise attention, i.e. supplying the MRHP model with the context among product reviews to assist the model into inferring the reviews’ ranking positions more precisely.

4.6 Analysis of Generalization Error

Figures 3 and 4 illustrate the approximation of the generalization error $\hat{E}(f_D^\theta) = R_{\text{val}}(f_D^\theta) - R_{\text{train}}(f_D^\theta)$ of the model after every epoch, where R_{val} and R_{train} indicate the average loss values of the trained model f_D^θ on the validation and training sets, respectively. Procedurally, due to different scale of the loss values, we normalize them to the range $[0, 1]$. The plots demonstrate that generalization errors of our MRHP model trained with the listwise ranking loss are constantly lower than those obtained by pairwise loss training, thus exhibiting better generalization performance. Additionally, as further shown in Table 4, $f_D^{\theta, \text{list}}$ incurs a smaller training-testing performance discrepancy $\Delta_{\text{MAP}} = |\text{MAP}_{\text{training}} - \text{MAP}_{\text{testing}}|$ than $f_D^{\theta, \text{pair}}$, along with Figures 3 and 4 empirically substantiating our Theorem (4).

4.7 Case Study

In Figure 1, we present helpfulness prediction results predicted by our proposed MRHP model and Contrastive-MCR (Nguyen et al., 2022), the previous best baseline. While our model is capable of producing helpfulness scores that evidently sepa-

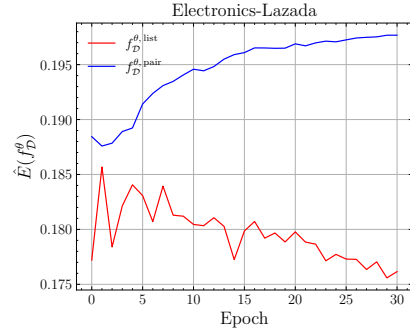


Figure 4: Generalization error curves per training epoch on the Electronics category in Lazada-MRHP dataset.

rate helpful with unhelpful product reviews, scores generated by Contrastive-MCR do mingle them. Hypothetically, our method could partition product reviews according to their encoded helpfulness features to obtain inherent separation. We provide more detailed analysis of the partitioning capability of our model and individual produced helpfulness scores in Appendix D and E.

5 Related Work

For real-world applications, existing methods are oriented towards extracting hidden features from input samples (Kim et al., 2006; Krishnamoorthy, 2015; Liu et al., 2017; Chen et al., 2018; Nguyen et al., 2021). Modern approaches have gradually taken into account additional and useful modalities, for instance meta-data (Tuan et al., 2016; Fan et al., 2019; Qu et al., 2020), images (Liu et al., 2021; Han et al., 2022), etc. They also depend on hand-crafted features, such as argument-based (Liu et al., 2017), lexical (Krishnamoorthy, 2015; Luu et al., 2015), and semantic features (Yang et al., 2015; Luu et al., 2016; Nguyen and Luu, 2022) to utilize automatic deep representation learning to train the helpfulness predictor. Some also utilize unsupervised learning techniques to polish the learned representations of input samples (Wu et al., 2020, 2023a; Nguyen and Luu, 2021; Wu et al., 2022, 2023b).

Despite performance upgrade, deep neural approaches for multimodal RHP (MRHP) problem, have been shown to still be inept at modeling partitioned and ranking data (Qin et al., 2021), which is the crucial characteristic of MRHP reviews (Ma et al., 2021). In this work, we seek to address those issues for the MRHP system with our proposed tree-based helpfulness predictor and listwise architectural framework.

6 Conclusion

In this paper, for the MRHP task, we introduce a novel framework to take advantage of the partitioned structure of product review inputs and the ranking nature of the problem. Regarding the partitioned preference, we propose a gradient-boosted decision tree to route review features towards proper helpfulness subtrees managed by decision nodes. For the ranking nature, we propose listwise attention network and listwise training objective to capture review list-contextualized context. Comprehensive analysis provides both theoretical and empirical grounding of our approach in terms of model generalization. Experiments on two large-scale MRHP datasets showcase the state-of-the-art performance of our proposed framework.

Limitations

Firstly, from the technical perspective, we have advocated the advantages of our proposed listwise loss for the MRHP task in terms of generalization capacity. Nevertheless, there are other various listwise discrimination functions that may prove beneficial for the MRHP model training, for example NeuralNDCG (Pobrotyn and Białobrzęski, 2021), ListMLE (Xia et al., 2008), etc. Moreover, despite the novelty of our proposed gradient-boosted tree in partitioning product reviews into helpful and unhelpful groups, our method does not employ prior contrastive representation learning, whose objective is also to segregate helpful and unhelpful input reviews. The contrastive technique might discriminate reviews of distinctive helpfulness features to bring further performance gain to multimodal review helpfulness prediction. At the moment, we leave the exploration of different listwise discrimination functions and contrastive learning as our prospective future research direction.

Secondly, our study can be extended to other problems which involve ranking operations. For instance, in recommendation, there is a need to rank the items according to their appropriateness to present to the customers in a rational order. Our gradient-boosted decision tree could divide items into corresponding partitions in order for us to recommend products to the customer from the highly appropriate partition to the less appropriate one. Therefore, we will discover the applicability of our proposed architecture in such promising problem domain in our future work.

Acknowledgements

This work was supported by Alibaba Innovative Research (AIR) programme with research grant AN-GC-2021-005.

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.
- Ali Akbari, Muhammad Awais, Manijeh Bashar, and Josef Kittler. 2021. How does loss function affect generalization performance of deep learning? application to human age estimation. In *International Conference on Machine Learning*, pages 141–151. PMLR.
- Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 46–54.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.
- 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.
- 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.
- Wei Han, Hui Chen, Zhen Hai, Soujanya Poria, and Lidong Bing. 2022. Sancl: Multimodal review helpfulness prediction with selective attention and natural contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5666–5677.
- Soo-Min Kim, Patrick Pantel, Timothy Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430.

- Srikumar Krishnamoorthy. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.
- Jean-Samuel Leboeuf, Frédéric LeBlanc, and Mario Marchand. 2020. Decision trees as partitioning machines to characterize their generalization properties. *Advances in Neural Information Processing Systems*, 33:18135–18145.
- 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. *arXiv preprint arXiv:1707.07279*.
- Junhao Liu, Zhen Hai, Min Yang, and Lidong Bing. 2021. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5927–5936.
- Anh Tuan Luu, Jung-jae Kim, and See Kiong Ng. 2015. Incorporating trustiness and collective synonym/contrastive evidence into taxonomy construction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1022.
- Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 403–413.
- Jiaqi Ma, Xinyang Yi, Weijing Tang, Zhe Zhao, Lichan Hong, Ed Chi, and Qiaozhu Mei. 2021. Learning-to-rank with partitioned preference: fast estimation for the plackett-luce model. In *International Conference on Artificial Intelligence and Statistics*, pages 928–936. PMLR.
- Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Advances in Neural Information Processing Systems*, 34:11974–11986.
- Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.
- Thong Nguyen, Xiaobao Wu, Anh-Tuan Luu, Cong-Duy Nguyen, Zhen Hai, and Lidong Bing. 2022. Adaptive contrastive learning on multimodal transformer for review helpfulness predictions. *arXiv preprint arXiv:2211.03524*.
- Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11103–11111.
- Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2019. Self-attentive document interaction networks for permutation equivariant ranking. *arXiv preprint arXiv:1910.09676*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Przemysław Pobrotyn and Radosław Biało-brzeski. 2021. Neuralndcg: Direct optimisation of a ranking metric via differentiable relaxation of sorting. *arXiv preprint arXiv:2102.07831*.
- Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2021. Are neural rankers still outperformed by gradient boosted decision trees?
- 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.
- Luu Anh Tuan, Siu Cheung Hui, and See Kiong Ng. 2016. Utilizing temporal information for taxonomy construction. *Transactions of the Association for Computational Linguistics*, 4:551–564.
- 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.
- Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liangming Pan, and Anh Tuan Luu. 2023a. Infocfm: A mutual information maximization perspective of cross-lingual topic modeling. *arXiv preprint arXiv:2304.03544*.
- Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023b. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*. PMLR.
- Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online. Association for Computational Linguistics.
- Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.
- 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 *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 38–44.

A Proofs

Lemma 1. Given listwise loss on the total training set as $\mathcal{L}^{\text{list}} = -\sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \log(f'_{i,j})$, where P denotes the product set, then $\mathcal{L}^{\text{list}}$ is convex and γ^{list} -Lipschitz with respect to $f'_{i,j}$.

Proof. Taking the second derivative of Equation (33), we have

$$\nabla_{f'_{i,j}}^2 \mathcal{L}^{\text{list}} = \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} \frac{y'_{i,j}}{(f'_{i,j})^2} > 0, \quad (34)$$

proving the convexity of $\mathcal{L}^{\text{list}}$.

The Lipschitz property of $\mathcal{L}^{\text{list}}$ can be derived from such property of the logarithm function, which states that

$$|\log(u) - \log(v)| = \left| \log\left(1 + \frac{u}{v} - 1\right) \right| \leq \left| \frac{u}{v} - 1 \right| = \left| \frac{1}{v}(u - v) \right| \leq \gamma |u - v|, \quad (35)$$

where the first inequality stems from $\log(1 + x) \leq x \forall x > -1$ and γ is chosen s.t. $|v| \geq \frac{1}{\gamma}$.

Let $x = \frac{u_{i,j}}{y_{i,j}}$, $z = \frac{v_{i,j}}{y_{i,j}}$. Applying the above result for $\mathcal{L}^{\text{list}}$, we obtain

$$\left| \log(u_{i,j}) - \log(v_{i,j}) \right| = \left| \log\left(\frac{u_{i,j}}{y_{i,j}}\right) - \log\left(\frac{v_{i,j}}{y_{i,j}}\right) \right| \leq \gamma \left| \frac{u_{i,j}}{y_{i,j}} - \frac{v_{i,j}}{y_{i,j}} \right|, \quad (36)$$

Multiplying both sides by $y_{i,j}$, and integrating the summation on all inequalities for $i \in \{1, 2, \dots, |P|\}$ and $j \in \{1, 2, \dots, |R_i|\}$, we achieve

$$\sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} |y_{i,j} \log(u_{i,j}) - y_{i,j} \log(v_{i,j})| \leq \gamma \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} |u_{i,j} - v_{i,j}|. \quad (37)$$

Ultimately, we obtain:

$$|\mathcal{L}^{\text{list}}(\mathbf{u}, \mathbf{y}) - \mathcal{L}^{\text{list}}(\mathbf{v}, \mathbf{y})| \leq \gamma^{\text{list}} |\mathbf{u} - \mathbf{v}|, \quad (38)$$

Where $\gamma^{\text{list}} = \gamma$. This proves the γ^{list} -Lipschitz property of $\mathcal{L}^{\text{list}}$.

Lemma 2. Given pairwise loss on the total training set as $\mathcal{L}^{\text{pair}} = \sum_{i=1}^{|P|} [-f_{i,r^+} + f_{i,r^-} + \alpha]^+$, where r^+, r^- denote two random indices in R_i and $y_{i,r^+} > y_{i,r^-}$, and $\alpha = \max_{1 \leq j \leq |R_i|} (y_{i,j}) - \min_{1 \leq j \leq |R_i|} (y_{i,j})$, then $\mathcal{L}^{\text{pair}}$ is convex and γ^{pair} -Lipschitz with respect to f_{i,r^+}, f_{i,r^-} .

Proof. Let $h_i^{\text{pair}}(\langle f_{i,r^+}, f_{i,r^-} \rangle, \mathbf{y}_i) = [-f_{i,r^+} + f_{i,r^-} + \alpha]^+$, $\mathbf{u}_i = \langle f_{i,u^+}, f_{i,u^-} \rangle$, $\mathbf{v}_i = \langle f_{i,v^+}, f_{i,v^-} \rangle$ be two inputs of h_i^{pair} . For $\theta \in [0, 1]$, we have

$$\begin{aligned} h_i^{\text{pair}}(\theta \mathbf{u}_i + (1 - \theta) \mathbf{v}_i, \mathbf{y}_i) &= h_i^{\text{pair}}(\theta \langle f_{i,u^+}, f_{i,u^-} \rangle + (1 - \theta) \langle f_{i,v^+}, f_{i,v^-} \rangle, \mathbf{y}_i) \\ &= h_i^{\text{pair}}(\langle \theta f_{i,u^+} + (1 - \theta) f_{i,v^+}, \theta f_{i,u^-} + (1 - \theta) f_{i,v^-} \rangle, \mathbf{y}_i) \\ &= [-\theta f_{i,u^+} + (1 - \theta) f_{i,v^+} + \theta f_{i,u^-} + (1 - \theta) f_{i,v^-} + \alpha]^+ \\ &= [\theta(-f_{i,u^+} + f_{i,u^-} + \alpha) + (1 - \theta)(-f_{i,v^+} + f_{i,v^-} + \alpha)]^+ \\ &\leq \theta[-f_{i,u^+} + f_{i,u^-} + \alpha]^+ + (1 - \theta)[-f_{i,v^+} + f_{i,v^-} + \alpha]^+ \\ &= \theta h_i^{\text{pair}}(\mathbf{u}_i, \mathbf{y}_i) + (1 - \theta) h_i^{\text{pair}}(\mathbf{v}_i, \mathbf{y}_i). \end{aligned} \quad (39)$$

Employing summation of the inequality on all $i \in \{1, 2, \dots, |P|\}$, we have

$$\mathcal{L}^{\text{pair}}(\theta \mathbf{u} + (1 - \theta) \mathbf{v}, \mathbf{y}) \leq \theta \sum_{i=1}^{|P|} h_i^{\text{pair}}(\mathbf{u}_i, \mathbf{y}_i) + (1 - \theta) \sum_{i=1}^{|P|} h_i^{\text{pair}}(\mathbf{v}_i, \mathbf{y}_i) = \theta \mathcal{L}^{\text{pair}}(\mathbf{u}, \mathbf{y}) + (1 - \theta) \mathcal{L}^{\text{pair}}(\mathbf{v}, \mathbf{y}), \quad (40)$$

which proves the convexity of $\mathcal{L}^{\text{pair}}$.

Regarding the Lipschitz property, we first show that h_i^{pair} holds the property:

$$|h_i^{\text{pair}}(\mathbf{u}_i, \mathbf{y}_i) - h_i^{\text{pair}}(\mathbf{v}_i, \mathbf{y}_i)| = [(-u_i^+ + u_i^- + \alpha) - (-v_i^+ + v_i^- + \alpha)]^+ = [-u_i^+ + u_i^- - v_i^- + u_i^-]^+. \quad (41)$$

Note that $y_{\min} \leq u_i^+, u_i^-, v_i^+, v_i^- \leq y_{\max}$, since we take the non-negative values in (41). Thus,

$$|h_i^{\text{pair}}(\mathbf{u}_i, \mathbf{y}_i) - h_i^{\text{pair}}(\mathbf{v}_i, \mathbf{y}_i)| \leq 2(y_{\max} - y_{\min}). \quad (42)$$

Similarly, applying the aforementioned observation, we have:

$$|\mathbf{u}_i - \mathbf{v}_i| = |u_i^+ - v_i^+| + |u_i^- - v_i^-| \geq 2(y_{\max} - y_{\min}). \quad (43)$$

Combining (42) and (43) leads to:

$$|h_i^{\text{pair}}(\mathbf{u}_i, \mathbf{y}_i) - h_i^{\text{pair}}(\mathbf{v}_i, \mathbf{y}_i)| \leq \gamma^{\text{pair}} |\mathbf{u}_i - \mathbf{v}_i|, \quad (44)$$

such that $\gamma^{\text{pair}} \geq 1$. Adopting the summation of (44) on all $i \in \{1, 2, \dots, |P|\}$, we obtain:

$$|\mathcal{L}^{\text{pair}}(\mathbf{u}, \mathbf{y}) - \mathcal{L}^{\text{pair}}(\mathbf{v}, \mathbf{y})| = \left| \sum_{i=1}^{|P|} h_i^{\text{pair}}(\mathbf{u}_i, \mathbf{y}_i) - \sum_{i=1}^{|P|} h_i^{\text{pair}}(\mathbf{v}_i, \mathbf{y}_i) \right| \leq \gamma^{\text{pair}} \sum_{i=1}^{|P|} |\mathbf{u}_i - \mathbf{v}_i| = \gamma^{\text{pair}} |\mathbf{u} - \mathbf{v}|. \quad (45)$$

The Lipschitz property of $\mathcal{L}_{\text{pair}}$ follows result (45).

Theorem 2. Let $\mathcal{L}^{\text{list}}$ and $\mathcal{L}^{\text{pair}}$ are γ^{list} -Lipschitz and γ^{pair} -Lipschitz, respectively. Then, the following inequality holds:

$$\gamma^{\text{list}} \leq \gamma^{\text{pair}}. \quad (46)$$

Proof. In order to prove Theorem (2), we first need to find the formulation of γ^{list} and γ^{pair} . We leverage the following lemma:

Lemma 3. A function \mathcal{L} is γ -Lipschitz, if γ satisfies the following condition (Akbari et al., 2021):

$$\gamma = \sup_{f_{i,j}} |\mathcal{L}'_{i,j}(f_{i,j})|. \quad (47)$$

With the foundation in mind, we take the derivative of $\mathcal{L}_{i,j}^{\text{list}}$ and $\mathcal{L}_{i,j}^{\text{pair}}$:

$$(\mathcal{L}_{i,j}^{\text{list}}(f_{i,j}))' = \left[y'_{i,j} \log \frac{\sum_{t=1}^{|R_i|} \exp(f_{i,t})}{\exp(f_{i,j})} \right]' = y'_{i,j} \left[\frac{\exp(f_{i,j})}{\sum_{t=1}^{|R_i|} \exp(f_{i,t})} - 1 \right] = -y'_{i,j} \left[\frac{\sum_{k=1, k \neq j}^{|R_i|} \exp(f_{i,k})}{\sum_{t=1}^{|R_i|} \exp(f_{i,t})} \right], \quad (48)$$

$$(\mathcal{L}_{i,j}^{\text{pair}}(f_{i,j}))' = \pm 1. \quad (49)$$

(48) and (49) imply that

$$\left| [\mathcal{L}_{i,j}^{\text{list}}(f_{i,j})]' \right| \leq y'_{i,j} \leq 1 = \left| [\mathcal{L}_{i,j}^{\text{pair}}(f_{i,j})]' \right|. \quad (50)$$

Combining equation (50) and Lemma (3), we obtain $\gamma^{\text{list}} \leq \gamma^{\text{pair}}$. ■

Theorem 3. Let $0 \leq \mathcal{L}^{\text{list}} \leq L^{\text{list}}$ and $0 \leq \mathcal{L}^{\text{pair}} \leq L^{\text{pair}}$. Then, the following inequality holds:

$$L^{\text{list}} \leq L^{\text{pair}}. \quad (51)$$

Proof. Adoption of Jensen's inequality on $\mathcal{L}_{\text{list}}$ gives:

$$\mathcal{L}_{\text{list}} = - \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \log f'_{i,j} \quad (52)$$

$$= \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \log \frac{\sum_{t=1}^{|R_i|} \exp(f_{i,t})}{\exp(f_{i,j})} \quad (53)$$

$$= \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \left(\log \sum_{t=1}^{|R_i|} \exp f_{i,t} - f_{i,j} \right) \quad (54)$$

$$= \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \left(\log \left(\frac{1}{|R_i|} \sum_{t=1}^{|R_i|} \exp(f_{i,t}) \right) - f_{i,j} + \log |R_i| \right) \quad (55)$$

$$\leq \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} \left(\frac{1}{|R_i|} \sum_{t=1}^{|R_i|} f_{i,t} - f_{i,j} + \log |R_i| \right) \quad (56)$$

$$\leq \sum_{i=1}^{|P|} \sum_{j=1}^{|R_i|} y'_{i,j} (f^{\max} - f^{\min} + \log |R_i|) \quad (57)$$

$$= |P|(f^{\max} - f^{\min}) + |P| \log |R_i|, \quad (58)$$

where $f^{\min} \leq f_{i,j} \leq f^{\max}, \forall i, j$. Now, such bounds of $f_{i,j}$ on $\mathcal{L}^{\text{pair}}$ yields:

$$\mathcal{L}^{\text{pair}} = \sum_{i=1}^{|P|} [-f_{i,r^+} + f_{i,r^-} + \alpha]^+ \leq |P|(f^{\max} - f^{\min}) + |P|(y^{\max} - y^{\min}), \quad (59)$$

where $y^{\max} = \max_{1 \leq i \leq |P|} \max_{1 \leq j \leq |R_i|} (y_{i,j})$, $y^{\min} = \min_{1 \leq i \leq |P|} \min_{1 \leq j \leq |R_i|} (y_{i,j})$. Note that Table 5 reveals that $\max |R_i| \leq 2043$. Therefore, $\log |R_i| \leq 3.31$, whereas $y^{\max} - y^{\min} = 4$, giving rise to the conclusion $\log |R_i| \leq y^{\max} - y^{\min}$. Therefore,

$$L^{\text{list}} \leq L^{\text{pair}}, \quad (60)$$

which concludes the proof of Theorem (3).

Theorem 4. Consider two models $f_{\mathcal{D}}^{\text{list}}$ and $f_{\mathcal{D}}^{\text{pair}}$ learned under common settings utilizing listwise and pairwise ranking losses, respectively, on dataset $\mathcal{D} = \{p_i, \{r_{i,j}\}_{j=1}^{|R_i|}\}_{i=1}^{|P|}$. Then, we have the following inequality:

$$E(f_{\mathcal{D}}^{\text{list}}) \leq E(f_{\mathcal{D}}^{\text{pair}}). \quad (61)$$

where $E(f_{\mathcal{D}}) = R_{\text{true}}(f_{\mathcal{D}}) - R_{\text{emp}}(f_{\mathcal{D}})$.

The inequality immediately follows from Theorems (1), (2) and (3). From Theorems (1) and (2), because T and N are constant, the second term of $\mathcal{L}^{\text{list}}$ is always smaller than that of $\mathcal{L}^{\text{pair}}$. From Theorems (1) and (3), we realize that $L^{\text{list}} \leq L^{\text{pair}}$, thus proving the smaller value of the first term of L^{list} .

B Dataset Statistics

In this section, we provide dataset statistics of the Amazon and Lazada datasets on the MRHP task. All of the numerical details are included in Table 5.

Dataset	Category	Train	Dev	Test	Max #R/P
Amazon	CS&J	12K/277K	3K/71K	4K/87K	691
	Elec.	10K/260K	3K/65K	3K/80K	836
	H&K	15K/370K	4K/93K	5K/111K	2043
Lazada	CS&J	7K/104K	2K/26K	2K/32K	540
	Elec.	4K/42K	1K/11K	1K/13K	346
	H&K	3K/37K	1K/10K	1K/13K	473

Table 5: Statistics of MRHP datasets. Max #R/P denotes the maximum number of reviews associated with each product.

C Generalization Errors of the Models trained with Listwise and Pairwise Ranking Losses

In this Appendix, we illustrate the empirical evolution of generalization errors of pairwise-trained and listwise-trained models on the remaining categories of the Amazon-MRHP and Lazada-MRHP datasets. The discovered characteristics regarding generalization in Figures 5 and 6 agree with those in Section 4.6, corroborating the intensified generalizability of our proposed listwise ranking loss.

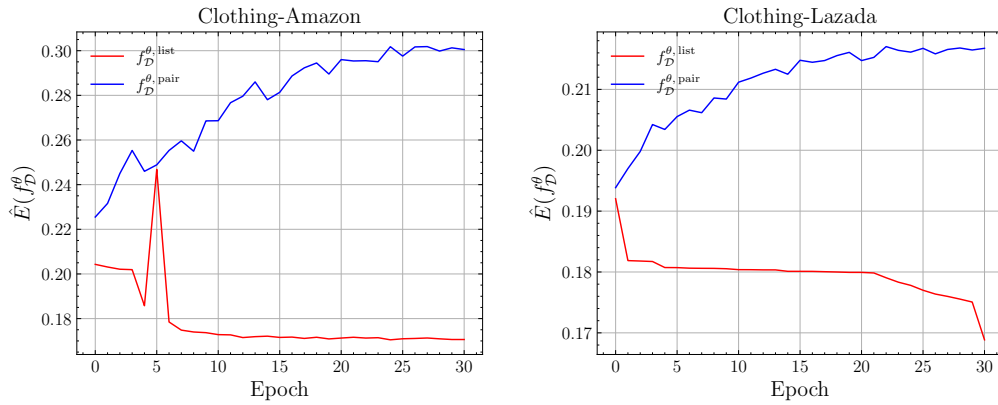


Figure 5: Generalization error curves per training epoch on the Clothing category in Amazon-MRHP and Lazada-MRHP datasets.

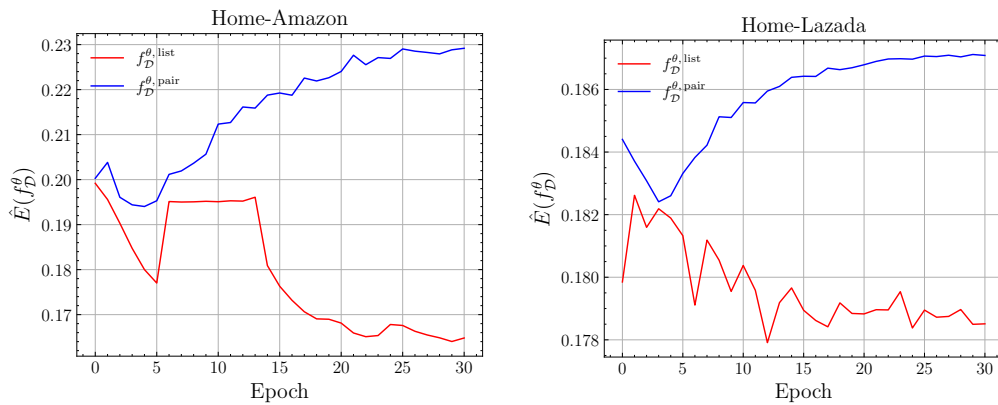


Figure 6: Generalization error curves per training epoch on the Home category in Amazon-MRHP and Lazada-MRHP datasets.

D Analysis of Partitioning Function of Gradient-Boosted Decision Tree

We examine the partitioning operation of our proposed gradient-boosted decision tree for the multimodal review helpfulness prediction. In particular, we inspect the $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_{|\mathcal{L}|}]$ probabilities, which route review features to the target leaf nodes in a soft approach. Our procedure is to gather $\boldsymbol{\mu}$ at the leaf nodes for all reviews, estimate their mean value with respect to each leaf, then plot the results on Clothing and Home of the Amazon and Lazada datasets, respectively, in Figures 7, 8, 9, 10, and 11.

From the figures, we can observe our proposed gradient-boosted decision tree's behavior of assigning high routing probabilities $\{\mu_i\}_{i=1}^{|\mathcal{L}|}$ to different partitions of leaf nodes, with the partitions varying according to the helpfulness scale of the product reviews. In consequence, we can claim that our GBDT divides the product reviews into corresponding partitions to their helpfulness degrees, thus advocating the partitioned preference of the input reviews.

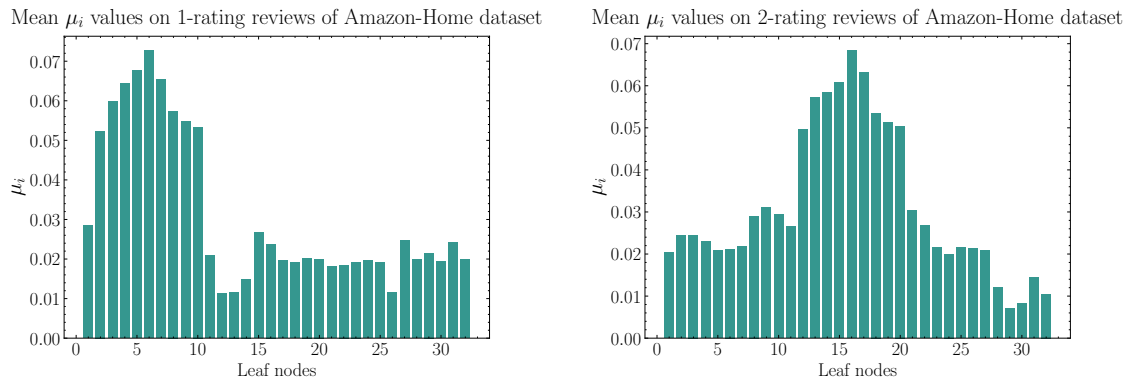


Figure 7: Mean μ_i routing probabilities at the proposed GBDT's leaves for 1-rating and 2-rating reviews in Amazon-Home dataset.



Figure 8: Mean μ_i routing probabilities at the proposed GBDT's leaves for 3-rating and 4-rating reviews in Amazon-Home dataset.

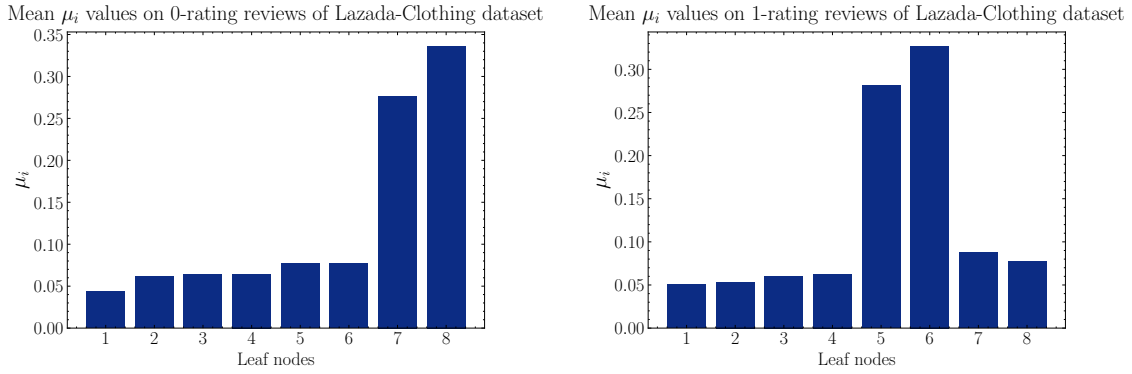


Figure 9: Mean μ_i routing probabilities at the proposed GBDT’s leaves for 0-rating and 1-rating reviews in Lazada-Clothing dataset.

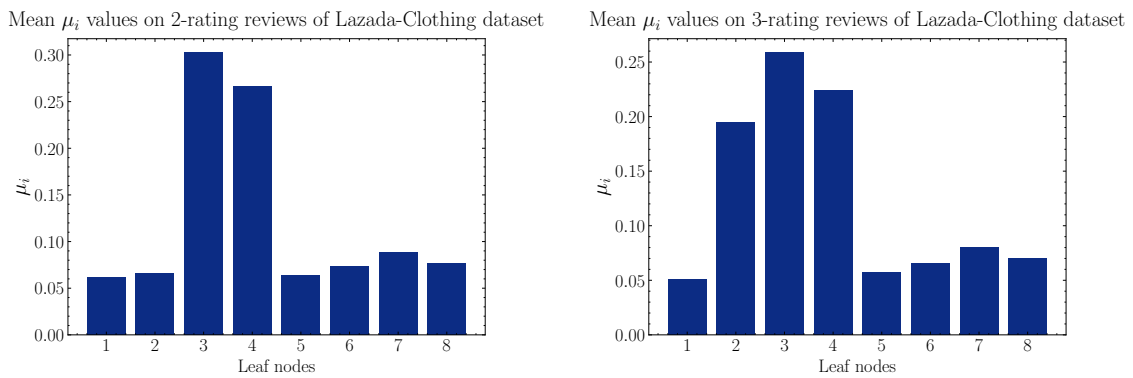


Figure 10: Mean μ_i routing probabilities at the proposed GBDT’s leaves for 2-rating and 3-rating reviews in Lazada-Clothing dataset.

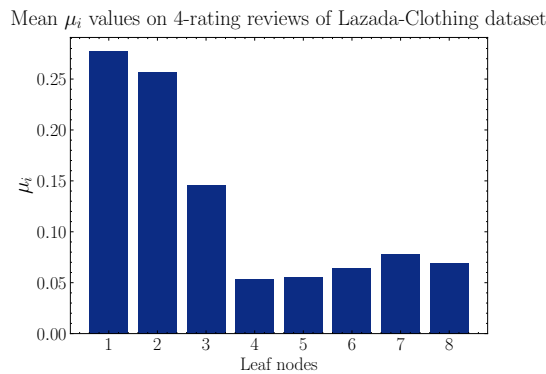


Figure 11: Mean μ_i routing probabilities at the proposed GBDT’s leaves for 4-rating reviews in Lazada-Clothing dataset.

E Examples of Product and Review Samples

We articulate product and review samples in Figure 1, comprising their textual and visual content, with the helpfulness scores generated by Contrastive-MCR (Nguyen et al., 2022), whose score predictor is FCNN-based, and our GBDT-based model.

E.1 Product B00005MG3K

Libbey Imperial 16-Piece Tumbler and Rocks Glass Set




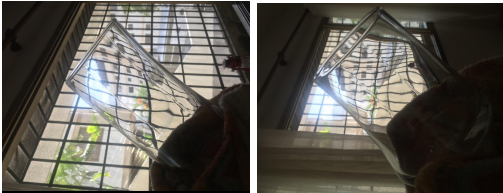
Review Information	NN-based Score	Tree-based Score
<p>Review 1 - Label: 1 These are fun, but I did learn that ice maker ice shaped like little half moon as many USA freezers have as their automatic ice maker, fit the curves of this class perfectly and will use surface water tension cohesion to slide up the glass inside to your mouth and act like a dam to block your drink believe it or not. So i have gotten used to that for personal use and know how to tilt the glass now, but when friends come, I use square tubes from an ice tray so I don't have to explain it to them or chance them spilling on themselves.</p>	1.467	-0.724
<p>Review 2 - Label: 1 If I could give less than a star I would. I am very disappointed in how low quality this product is and would not recommend buying it.</p>	1.147	-0.874
<p>Review 3 - Label: 1 Very cool & futuristic looking.</p> 	6.622	-0.964
<p>Review 4 - Label: 1 These are attractive glasses which seem a good deal more classy than the cost here would imply. They feel higher end and when you plink one with your fingernail it'll give off a fine crystal like ring. They are every bit as attractive as they look in the pictures.</p>	1.731	-0.868
<p>Review 5 - Label: 3 Mixed reviews did not deviated me from getting this set. Just the add shape is a turn on. A very well packed box arrived bubble wrap with every glass intact. The glasses are beautiful and everything I expected. One thing though, It's interesting that there is only one picture on the page. This picture shows no detail. Used to many types of glass drink-ware, the first thing I noticed is the "seams" on each glass (see pictures). This makes obvious the fact that these are mold made. This is the reason for 4 stars. Being using them for just a couple of weeks by the time I wrote this review. Will update as time goes on.</p> 	0.494	0.882

Table 6: Generated helpfulness scores on reviews 1-5 for product B00005MG3K.


Review Information	NN-based Score	Tree-based Score
Review 6 - Label: 1 I hate going through the hassle of returning things but it had to be done with this purchase.	0.044	-0.778
Review 7 - Label: 1 The short glasses are nice, but the tall ones break easily. SUPER easily. I had two of them break just by holding them. I will absolutely not be reordering this.	0.684	-0.800
Review 8 - Label: 1 I love these. We had them in a highly stylized Japanese restaurant and were psyched to find them here. Tall glasses have a "seam". No tipping or breakage yet as mentioned by other reviewers.	0.443	-0.897
Review 9 - Label: 2 It's true that the taller 18-oz glasses are delicate. If you're the kind of person who buys glassware expecting every glass to last 20 years, this set isn't for you. If you're the kind of person who enjoys form over function, I'd highly recommend them.	2.333	0.435
Review 10 - Label: 1 Quality is good. Does not hold water from the underside if you put it in the dishwasher.	6.074	-0.844
Review 11 - Label: 1 I have owned these glasses for 20-plus years. After breaking most of the tall ones, I looked around for months to find great glasses but still thought these were the best, so I bought more.	2.615	-0.923
Review 12 - Label: 3 I am soooooo disappointed in these glasses. They are thin. Of course, right after opening we put in the dishwasher and upon taking them out it looked like they were washed with sand! We could even see the fingerprints. And we have a watersoftener! In the photo I have included, this is after one dishwasher washing! 	7.529	0.836

Table 7: Generated helpfulness scores on reviews 6-12 for product B00005MG3K.

E.2 Product B00Q82T3XE

Dasein Frame Tote Top Handle Handbags Designer Satchel Leather Briefcase Shoulder Bags Purses



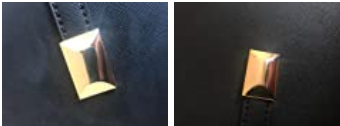
Review Information	NN-based Score	Tree-based Score
<p>Review 1 - Label: 1 I really loved this and used it to carry my laptop to and from work. I used the cross-body strap. However, the metal hardware of the strap broke after three months, and the stitching where the cross-body strap attached to the purse ripped off the same week. Love this ourselves but the handles are too short for me to wear comfortably without the cross body strap.</p>	0.281	-0.192
<p>Review 2 - Label: 1 Hello, I am Alicia and work as a researcher in the health area. Moreover, I was looking for a feminine, classical and practical bag-briefcase for my work. I would like to begin with the way you show every product. I love when I can see the inner parts and the size of the bag, not only using measures but when you show a model using the product too. Also, the selection of colour is advantageous a big picture with the tone selected. There are many models, sizes and prices. I consider that is a right price for the quality and design of the product. The products I bought have a high-quality appearance, are professional and elegant, like in the pictures! I was not in a hurry, so I was patient, and the product arrived a couple of days before the established date. The package was made thinking in the total protection of every product I bought, using air-plastic bubbles and a hard carton box. Everything was in perfect conditions. I use them for every day- work is very resistant, even in rain time I can carry many things, folders and sheet of paper, a laptop. Their capacity is remarkable. The inner part is very soft and stands the dirty. I am enjoying my bags! All the people say they are gorgeous!</p>	2.938	-0.138
<p>Review 3 - Label: 1 This purse has come apart little by little within a month of receiving it. First the thread that held on the zipper began to unravel. Then the decorative seam covering began to come off all over the purse. Yesterday I was on my way into the grocery and the handle broke as I was walking. I've only had it a few months. Poorly made.</p>	0.460	-0.226
<p>Review 4 - Label: 1 I bought this because of reviews but i am extremely disappointed... This bag leather is too hard and i don't think i will use it</p>	-0.646	-0.067
<p>Review 5 - Label: 2 There are slight scratches on the hardware otherwise great size and it's a gorgeous bag. Got it for use while I'm in a business casual environment.</p> 	5.094	-0.493

Table 8: Generated helpfulness scores on reviews 1-5 for product B00Q82T3XE.


Review Information	NN-based Score	Tree-based Score
Review 6 - Label: 1 Tight bag, has no flexibility. stiff. But I do receive a lot of compliments.	-1.794	-0.222
Review 7 - Label: 1 I love this bag!!! I use it every day at work and it has held up to months of use with no sign of wear and tear. It holds my laptop, planner, and notebooks as well as my large wallet and pencil case. It holds so much! I've gotten so many compliments on it. It feels and looks high quality.	0.819	-0.284
Review 8 - Label: 3 This bag is perfect! It doubles as somewhat of a "briefcase" for me, as it fits my iPad, planner, and files, while still accommodating my wallet and normal "purse" items. My only complaint was that Jre scratches already on the gold metal accents when I unwrapped it from the packaging. Otherwise- great deal for the price! 	0.259	0.939
Review 9 - Label: 2 I believe this the most expensive looking handbag I have ever owned. When your handbag comes in its own bag, you are on to something wonderful. I also purchased a router in the same order, and I'm serious, the handbag was better wrapped and protected. Now for a review : The handbag is stiff, but I expected that from other reviews. The only reason I didn't give a five star rating is because it is not as large as I hoped. A laptop will not fit. Only a tablet. This is a regular good size purse, so don't expect to be able to carry more than usual. I probably won't be able to use it for my intended purpose, but it is so beautiful, I don't mind.	2.695	0.462
Review 10 - Label: 1 Look is great can fit HP EliteBook 8470p (fairly bulky laptop 15 inch), but very snug. I can only fit my thin portfolio and the laptop into bag.	-0.235	-0.189
Review 11 - Label: 1 This bag is really great for my needs for work, and is cute enough for every day. Other reviews are correct that this is a very stiff-leather bag, but I am fine with that. I love the color and the bag is super adorable. I get so many compliments on this. Also, I travelled recently and this was a perfect bag to use as your "personal item" on the airplane- it zips up so you don't have to worry about things falling out and is just right for under the seat. I love the options of having handles AND the long strap. I carry an Iphone 6+ (does not fit down in the outside pocket completely but I use the middle zipper pouch for my tech), wallet, glasses, sunglasses, small makeup bag, a soapdish sized container that I use for holding charger cords (fits perfect in the inside liner pockets), and on the other side of the zipper pouch I carry an A5-sized Filofax Domino.	6.290	-0.194

Table 9: Generated helpfulness scores on reviews 6-11 for product B00Q82T3XE.


Review Information	NN-based Score	Tree-based Score
<p>Review 12 - Label: 3</p> <p>Absolutely stunning and expensive looking for the price. I just came back from shopping for a tote bag at Macy's and so I had the chance to look and feel at all the different bags both high end brand names and generic. This has a very distinguished character to it. A keeper. The size is rather big for an evening bag as long as it is not a formal one. I like that it can accommodate a tablet plus all other things we women consider must haves. The silver metal accents are just of enough amount to give it umph but not superfluous to make it look tacky. The faux ostrich material feels so real. The whole bag is very well balanced. Inside it has two zippered pockets and two open pockets for cell phone and sun glasses. Outside it has one zippered pocket by the back. I won't be using the shoulder strap too much as the handles are long enough to be carried on the shoulders.</p>	2.262	0.923
<p>Review 13 - Label: 4</p> <p>I added pictures. I hate the fact that people selling things do not give CLEAR defined pictures. This purse was well shipped. Not one scratch... and I don't think there COULD have been a scratch made in shipping. The handles and the bottom are a shiny patent leather look. The majority of the case is a faux ostrich look. It has a 'structure' to it. Not a floppy purse. There is a center divider that is soft and has a zipper to store things. One side (inside) has two pockets that do not zipper. One side (inside) has a zippered pocket. It comes with a long shoulder strap. Please see my photos. So far I really like this purse. The water bottle is a standard 16.9oz.</p> 	7.685	1.969
<p>Review 14 - Label: 2</p> <p>Love this purse! When I opened the package it seemed like it was opening a purse I had purchased for \$450.00 it was packaged so nicely!! Every little detail of the purse was covered for shipping protection. This was/is extremely impressive to me for a purse I paid less than \$40.00 for. Wow. It's roomy & has many pockets inside. And med/large purse I'd say, but I like that it's larger in length than height. It's very classic looking yet different with texturing. I always get many compliments on it. Believe me I have many purses & currently this is one of my favorites!! I have already & will continue to purchase Dasein brand handbags.</p>	2.309	0.584

Table 10: Generated helpfulness scores on reviews 12-14 for product B00Q82T3XE.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section 7
- A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.
- A3. Do the abstract and introduction summarize the paper’s main claims?
Abstract and Section 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Section 4

- B1. Did you cite the creators of artifacts you used?
Section 4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix B

C Did you run computational experiments?

Section 4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 4

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Not applicable. Left blank.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Not applicable. Left blank.

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.