计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(12):21−31 [doi: 10.15888/j.cnki.csa.009342]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

利用多维评分进行在线评论的有用性预测①

吴健健, 李文畅, 时宏伟

(四川大学 计算机学院, 成都 610065)
通信作者: 时宏伟, E-mail: shihw001@126.com

摘　要: 在线评论的有用性预测任务在当前的电子商务领域中发挥着重要的作用, 该任务的目标是判断在线评论的有用性, 进而重点展示对未来消费者更有帮助的评论, 提高消费者获取信息的效率. 在本文中, 我们重点关注近年来在各大在线平台兴起的一种新的评分系统——多维评分系统, 尝试研究用户在该系统中给出的方面评分对在线评论有用性的影响. 本文提出了一个综合考虑了评论文本、用户总体评分和方面评分 3 种元素及其交互的多层次神经网络模型 HORA 来完成有用性预测任务. 通过在两个真实世界的数据集上进行的实验结果表明, 与当前的基线模型相比, HORA 在 *MAE* 和 *RMSE* 两个指标上展示了更好的结果, 同时在实验中也表现出了良好的鲁棒性, 表明了方面评分对用户的在线评论有用性感知的重要意义.

关键词: 有用性预测; 多维评分; 方面评分; 注意力机制; 层次网络; 文本分析; 自然语言处理

引用格式: 吴健健,李文畅,时宏伟.利用多维评分进行在线评论的有用性预测.计算机系统应用,2023,32(12):21–31. http://www.c-s-a.org.cn/1003-3254/9342.html

Helpfulness Prediction of Online Reviews Using Multidimensional Ratings

WU Jian-Jian, LI Wen-Chang, SHI Hong-Wei

(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Helpfulness prediction task of online reviews is significant in the contemporary e-commerce environment. It aims to evaluate the helpfulness of online reviews and then highlight the reviews more helpful to future consumers, thereby improving the consumers' efficiency in obtaining information. This study concentrates on the new multidimensional scoring system emerging on various online platforms in recent years, and tries to study the influence of aspect ratings given by users in the system on the helpfulness of online reviews. To accomplish the helpfulness prediction task, it puts forward a multi-level neural network model HORA that considers all three components of review texts, overall ratings, and aspect ratings, as well as their interconnections. The experimental results on two real-world datasets show that HORA outperforms the present baseline models in terms of *MAE* and *RMSE* and exhibits good robustness. This indicates the significance of aspect ratings for the helpfulness awareness of users' online reviews.

Key words: helpfulness prediction; multidimensional ratings; aspect ratings; attention mechanism; hierarchical network; text analysis; natural language processing (NLP)

1　引言

随着电子商务的普及和成功, 在线评论已经成为消费者、企业和组织跨期决策过程的一个有价值的参考来源. 消费者喜欢在做购买决策之前先浏览评论以收集相关信息, 无需像在线下市场中接触到实物, 其就可以直接通过浏览大量评论以获得足够的产品质量信息, 这些都能够对消费者的行为产生直接的影响. 产品的销售会受到产品评论和特定产品类别的相关因素的

① 收稿时间: 2023-06-16; 修改时间: 2023-07-19; 采用时间: 2023-08-08; csa 在线出版时间: 2023-10-20
CNKI 网络首发时间: 2023-10-23

影响[1], 而消费者对于产品的第一印象和他人的好评或者差评也都直接影响着消费者的购买行为, 因此在线评论已成为消费者在购物前识别商品质量的一个重要标志, 在很大程度上决定了消费者的意愿[2]. 但随着评论数量的不断增加, 定位有用的信息变得具有挑战性. 虽然电子商务平台收集用户对产品评论的有用性投票意见, 但实际上, 在不太受欢迎的产品中, 投票数据很少, 甚至缺失. 而提出的有用性预测旨在从投票数据中学习, 以识别高质量的评论并向客户推荐这些评论以帮助他们作出购买决策. 因此, 理解和预测在线评论的有用性行为将不仅可以大大节约消费者的决策时间, 提高购买决策正确性, 还对商家如何改进在线评论机制提供一定的辅助指导.

其实, 有用性预测在自然语言处理方面已得到了广泛的研究. 先前的研究表明, 有用性作为在线评论质量的内在度量, 受到内容和上下文特征[3,4]的影响. 内容特征包括直接来自评论文本 (如评论内容、总体评分)[5]的信息, 而对上下文特征的研究主要关注评论文本以外的信息, 如用户信息和产品特征[6]. 本研究中我们关注前者, 因为相较用户个人数据而言, 用户提供的产品评论却是网站中更容易访问到的数据 (图 1). 在内容特征中, 有用性预测通常使用评论文本和总体评分. 此外, 研究表明, 这两个特征的交互作用可以进一步提高有用性预测任务的性能[7]. 尽管如此, 现有的研究主要集中在基于上下文信息生成内容表示, 并使用总体评分来优化内容的语义, 从而做出在线评论的有用性预测, 而这在很大程度上忽略了方面评分以及方面评分与句子之间的语义关联. 同时, 仅使用总体评分和文本内容进行建模容易出现情感不一致的问题, 而方面评分的出现可以对这种不一致性进行适当的修正, 如图 2 所示.

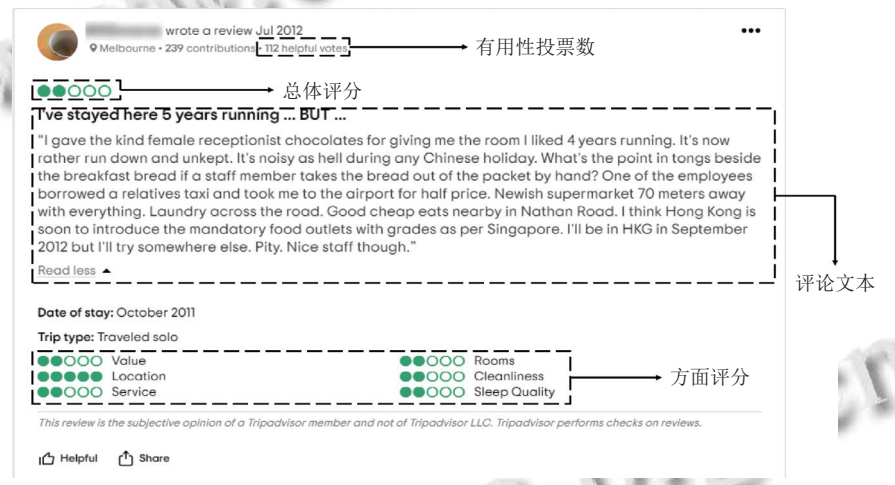


图 1 来自 TripAdvisor 的一个评论示例

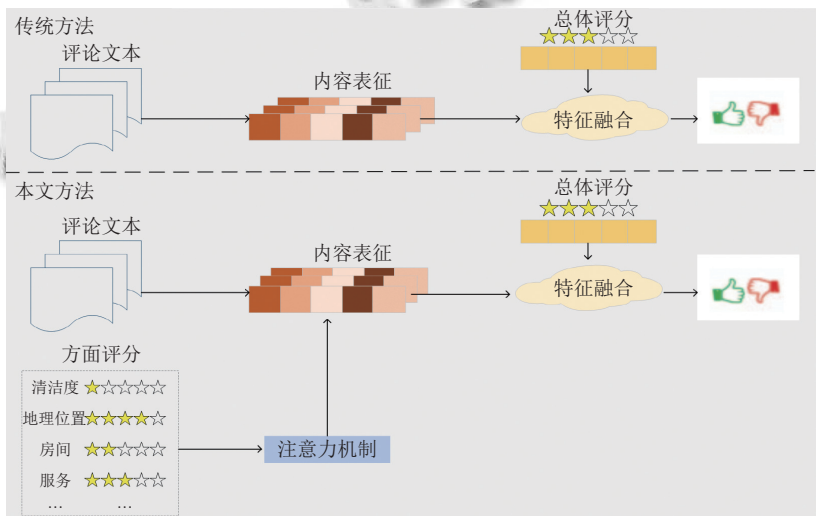


图 2 方法对比

最近, 多维评分系统[8]已经应用于多个在线平台, 如 TripAdvisor、Niche 和 WebMD, 它们要求用户除了给出评论文本和相应的总体评分外, 还对产品的具体方面进行评分 (以下简称“方面评分”)(图 1), 在用户决策的过程中, 方面评分、评论内容和总体评分在多个粒度层次上提供产品评估. 方面评分反映了用户对产品更具体的属性感受; 评论内容包括用户的意见和情绪; 总体评分则反映了用户对产品的整体满意程度. 理解有用性预测行为的挑战是如何联合执行各类评分和语义表示, 以学习一个更全面的预测行为模型. 由于评分过程反映了消费者自然地学习和概念化各个方面, 并在局部概念和整体评分之间建立关系, 因此有效的预测行为分析方法应该建立在人类认知的基础上. 而在这类系统中, 用户的评论行为就通常遵循一种从细到粗的模式. 例如, 总体评分可以看作是对多个方面的意见的粗粒度综合[9,10]. 在这项工作中, 我们假设建模这样的模式可以提高在线评论的感知有用性, 因为方面评分包含更详细的情绪, 并有能力缓解评论文本和总体评分[11]之间的不一致性问题. 具体来说, 我们提出了一个基于注意力机制的多层次网络, 利用总体评分、评论文本和方面评分来执行有用性预测任务 (图 2). 基于两个真实世界的数据集, 实验结果表明, HORA 在预测在线评论的有用性方面表现良好, 同时也展示了方面评分对在线评论有用性的预测影响. 我们的主要贡献如下.

(1) 我们提出了 HORA, 一个基于注意力机制的层次网络 (hierarchical network), 利用总体评分 (overall rating)、评论文本 (review text) 和方面评分 (aspect ratings) 来准确和稳健地执行在线评论的有用性预测任务.

(2) 据我们所知, HORA 是第 1 个将来自多维评分系统的方面评分合并到有用性预测任务的模型.

(3) 我们在两个真实数据集上进行了实验, 以验证该方法的有效性. 结果表明, HORA 的表现显著优于其他神经网络模型. 除此之外, 我们还证明了模型的稳健性.

2 相关工作

在以往的文献工作中已提出有多种特征会影响在线评论的有用性, 而这些特征主要集中在两个方面: 基于内容的特征和基于上下文的特征.

2.1 内容特征

内容特征包括评论文本, 总体评分和方面评分[4], 经典的在线评论有用性预测方法包括寻找新的手工特征, 以此来提高模型的表现. 例如基于情感[12], 方面[13], 论点[14]等特征都相继被研究学者发现其能够帮助预测在线评论的有用性. Ghose 等人[15]和 Korfiatis 等人[16]探索了评论文本的多个方面, 如主观性水平、可读性的各种衡量标准和拼写错误的程度, 以确定基于文本的重要特征, 最后发现评论中的主观性、信息性、可读性和语言正确性会影响感知有用性. 与倾向于仅包含主观或仅包含客观信息的评论相比, 混合了客观和高度主观句子的评论对感知有用性的影响更大.

另外还有一些研究人员认为, 星级评分和有用性预测之间存在联系, 因此评论文本和总体评分在以往的研究中也被广泛应用. Otterbacher 等人[17]认为评论中存在积极偏见, 即积极星级评分的评论被认为更有帮助; Martin 等人 [12]在预测在线评论的有用性时, 将从评论文本中提取的情感词汇融入到他们的特征中. Chen 等人[18]提出了一种利用字符级和字级信息的卷积网络, 取得了显著的效果. 通过测试总体评分的离散性是否对评论的有用性有影响, Lee 等人[19]引入了一个新的变量来衡量在线评论的效用. Lee 等人[20]发现, 随着时间的推移, 产品的平均评论评分是在线评论有用性预测的一个重要特征. Mukherjee 等人[21]总结了来自总体评分和其他特征的一致性和语义特征, 提供了更多可解释的结果. Zhou 等人[3]调查了总体评分和文本情绪的互动影响, 他们发现负面评价被认为比正面评价更有帮助. 除了从评论文本和总体评分中提取特征外, 一些研究还试图构建它们之间的交互作用. Qu 等人[22]使用 CNN 将总体评分和评论文本结合起来, 提高了评论有用性预测任务的整体性能. Fan 等人[23]提出了一种多任务神经网络学习架构, 认为它提高了基于满足总体评分作为辅助任务的有用性评论预测的准确性. 除此之外, Yang 等人[13]提出, 在线评论的有用性预测受到评论中提到的一些细粒度方面的影响. 他们训练了一个基于方面的提取模型来衡量方面的覆盖率, 并将性能提高了 7%. Sun 等人[24]使用了方面的数量和方面的平均长度来帮助识别有用的评论.

2.2 上下文特征

除了从评论文本和评分中获得特征之外, 诸如用

户信息和产品信息等上下文特征也被用于研究其对用户感知有用性的影响. Bilal 等人[25]断言对于每个用户来说, 历史上有用的投票、过去评论的数量、评论的平均长度都是评论有用性预测的关键特征. 根据 Qu 等人[26]的说法, 评论的有用性会受到产品信息的影响, 而用户是否意识到相关的产品信息将会影响他们的购买选择. Fan 等人[27]发现除了评论内容本身外, 评论的有用性预测还受目标产品的标题、品牌、类别和描述的影响. Li 等人[28]通过 4 种度量找到了有用的评论, 包括可信度、可读性、保密性和 LWC 特性. Hong 等人[29–31]试图从产品特性中提取信息, 构建模型并最终获得了良好的性能.

3 本文方法

3.1 模型框架

Yang 等人[32]设计了一个层次注意网络, 利用语境注意机制来识别重要的词汇和句子. 此外, Wu 等人[6]的研究表明, 对文档结构进行分层建模可以帮助生成更有效的内容表示. Du 等人[7]将评分与评论内容分开编码, 自适应地调整内容表示的潜在方面的评分信息量. 然而, 这些方法忽略了明确的方面评分, 这为理解用户对产品的每个属性的情绪提供了重要的信息. 因此, 为了更好地探索评论内容中句子之间的深层关联, 并在学习评论内容表征时更多地关注与方面相关的词汇, 我们提出了一种新的多维度层次注意网络 (简称 HORA), 如图 3 所示.

该方法首先用 BiGRU 对单词的上下文信息进行编码, 并为每个单词生成一个上下文状态表示. 然后采用单词级注意力机制来计算每个单词的语义重要性, 并通过将单词的上下文状态聚合起来得到句子表示, 之后再利用 BiGRU 对句子之间的依赖关系进行建模, 为每个句子生成一个上下文状态表示. 基于此, 再通过一个句子级注意力机制来计算每个句子的权重, 并将句子状态的加权和作为评论内容表示. 最后, 将内容表示与总体评分连接起来, 得到最终的特征向量, 并根据所得到的特征向量预测评论的有用性.

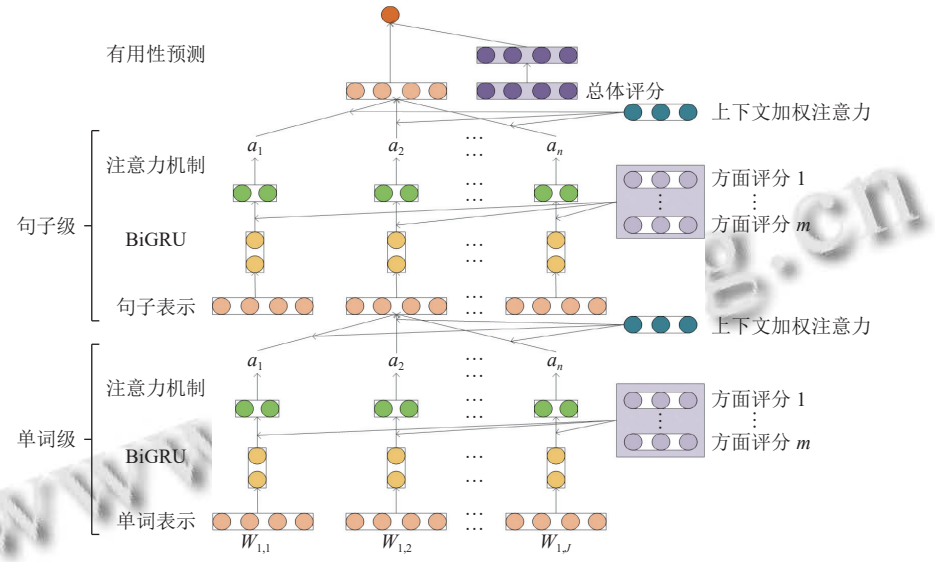


图 3 模型架构

3.2 问题公式化

我们的目标是使用评论文本、总体评分和方面评分来执行在线评论的有用性预测任务. 我们的语料库是评论的集合, 假设一个评论内容由 L 个句子组成, 每个句子都是一个单词序列. 如果我们以第 i 个句子 s_i 为例, 并假设它包含 J 个单词 $\{w_{i1},\cdots,w_{ij},\cdots,w_{iJ}\}$, 且对应的总体评分为 r, 第 k 个方面对应的方面评分为 a_k.

1) 单词级 BiGRU 层: 类似于 LSTM 可以解决长期记忆和反向传播中的梯度问题, GRU 也是一种递归神经网络. 然而, 对于非双向神经网络结构, 状态必须从前到后输出, 而在在线评论有用性预测研究任务中当前时刻的输出可以与前一时刻的状态和即将到来的时刻的状态相关联. 由于双向 GRU 网络更利于层次特征的提取, 因此我们选择了它. 首先, 我们使用 GloVe

预训练词向量$\{w_{i1},\cdots,w_{ij},\cdots,w_{iJ}\}$, 并得到单词向量的序列$\{x_{i1},\cdots,x_{ij},\cdots,x_{iJ}\}$, 作为 BiGRU 层用于提取文本深度特征的输入向量. 根据 BiGRU 神经网络模型图, 它包括前向和反向的 GRU. 第i个句子的第t个单词的词向量为x_{it}, 同时, 为了将方面评分信息整合到单词的语义重要性计算中, 首先, 通过方面嵌入矩阵E_{aspect}, 将每个方面评分a_k编码为一个向量e_k:

$$x_{it}=W_e w_{it} \tag{1}$$

$$e_k=E^k_{\text{aspect}}a_k \tag{2}$$

因为一个单词的重要性高度依赖于上下文, 所以我们需要对每个单词的上下文信息进行编码. 根据 BiGRU 神经网络模型图, 可以把 BiGRU 模型看做由前向 GRU 和反向 GRU 两部分组成, 第i个句子的第t个单词的词向量x_{it}通过 BiGRU 层进行语义编码, 由于反向网络在每个时间步长中获得的序列表示包括未来的编码信息, 最后, 将每个单词的上下文状态h_{it}编码为正向网络$\overrightarrow{h_{it}}$和后向网络$\overleftarrow{h_{it}}$得到的相应序列表示的连接.

$$\overrightarrow{h_{it}}=\overrightarrow{GRU}(x_{it}) \tag{3}$$

$$\overleftarrow{h_{it}}=\overleftarrow{GRU}(x_{it}) \tag{4}$$

$$h_{it}=\left[\overrightarrow{h_{it}},\overleftarrow{h_{it}}\right] \tag{5}$$

2) 单词级注意力机制层: 以第i个句子s_i的第t个单词的重要性计算为例. 为了将上下文信息和方面评分嵌入映射到同一语义空间中, 现将其上下文状态h_{it}和方面嵌入e_k分别输入到全连接网络中. 然后将这两个全连接网络的输出相加, 得到集成表示u_{it}, 其同时包含上下文信息和方面信息. 上标k表示这组参数属于第k个方面, 其中W_{wc}、W_{wa}、b_w为单词级注意力机制层的参数. 接着, 通过集成表示u_{it}与全局语义向量u^k_w之间的相似度来度量单词之间的语义权重α_{it}, 之后采用 Softmax 函数对语义权重进行归一化处理. 全局语义向量u^k_w可以看作是查询在潜在语义空间中“单词的重要性”的向量表示, 该值在训练过程中被随机初始化并学习. 对每个单词w_{it}重复这些操作, 然后确定每个单词的权重α_{it}. 最后, 通过对单词的加权上下文状态进行聚合, 得到基于方面a_k的句子s_i的连续向量表示S^k_i.

$$u_{it}=\tanh\left(W^k_{wc}h_{it}+W^k_{wa}e_k+b^k_w\right) \tag{6}$$

$$\alpha_{it}=\frac{\exp\left(u^{\mathrm{T}}_{it}u^k_w\right)}{\sum_{t=1}^{J}\exp\left(u^{\mathrm{T}}_{it}u^k_w\right)} \tag{7}$$

$$S^k_i=\sum_{t=1}^{J}\alpha_{it}h_{it} \tag{8}$$

现在, 我们已为每个方面生成一个句子表示, 为了得到集成所有方面信息的句子表示, 首先, 通过将S^k_i输入到一个全连接网络当中, 使得根据不同的方面评分, 将不同的句子表示映射到相同的语义空间中; 然后, 通过计算每个变换表示的t^k_i与高级语义向量x_w的相似度, 生成每个表示的权值β^k_i, 接着用 Softmax 函数进行归一化. 其中, 高级语义向量x_w可以解释为“这个方面对这个句子的情绪有多少影响”, 和u^k_w类似, 它是随机初始化的, 并直接从训练数据集中学习. 最终的句子向量表示S_i是由基于不同方面评分的表示的加权和决定的, 其中m是评论中方面评分的数量:

$$t^k_i=\tanh\left(W_{wt}S^k_i+b_{wt}\right) \tag{9}$$

$$\beta^k_i=\frac{\exp\left(\left(t^k_i\right)^{\mathrm{T}}x_w\right)}{\sum_{k=1}^{m}\exp\left(\left(t^k_i\right)^{\mathrm{T}}x_w\right)} \tag{10}$$

$$S_i=\sum_{k=1}^{m}\beta^k_i S^k_i \tag{11}$$

3) 句子级 BiGRU 层: 与单词级 BiGRU 层一样, 句子级 BiGRU 层对句子之间的依赖关系进行编码, 旨在获得句子的上下文表示. 从第 1 个句子到最后一个句子, 在每个时间步长中, 前向 GRU 网络根据当前输入的句子序列和之前的句子序列计算句子的上下文状态, 从而记忆历史语义信息. 反向 GRU 网络计算从最后一句到第 1 个句子的状态, 以编码未来的上下文信息. 在每个时间步长中, 反向 GRU 网络产生的当前状态依赖于当前的目标句子和后面的句子序列. 最后, 将前向网络和后向网络在相应的时间步长上连接起来, 得到每个句子S_i的上下文表示h_i.

$$\overrightarrow{h_i}=\overrightarrow{GRU}(S_i),\ i\in[1,L] \tag{12}$$

$$\overleftarrow{h_i}=\overleftarrow{GRU}(S_i),\ i\in[1,L] \tag{13}$$

$$h_i=\left[\overrightarrow{h_i},\overleftarrow{h_i}\right] \tag{14}$$

4) 句子级注意力机制层: 句子级注意力机制层的

目的是根据其上下文表征和方面评分来计算每个句子的语义权重, 计算过程类似于单词级注意力机制层. 首先, 将句子S_i的上下文表示h_i和方面嵌入e_k分别输入到全连接网络中, 将这两种表示转换到相同的语义空间, 其中W_{sc}和W_{sa}为全连接网络的参数. 接下来, 将这两个转换后的表示求和, 生成集成后的表示u_i, 其中上标k表示该参数用于第k个方面. 接着, 通过计算集成表示u_i与全局上下文向量u_s^k之间的相似性来确定句子S_i的权值α_i, 在句子级注意力机制中, 全局上下文向量u_s^k表示“句子的重要性”. 最后, 将句子上下文表示的加权和视为基于方面a_k的评论表示S^k.

$$u_i = \tanh\left(W_{sc}^k h_i + W_{sa}^k e_k + b_s^k\right) \tag{15}$$

$$\alpha_i = \frac{\exp\left(u_i^{\mathrm{T}} u_s^k\right)}{\sum_{i=1}^{L}\exp\left(u_i^{\mathrm{T}} u_s^k\right)} \tag{16}$$

$$S^k = \sum_{i=1}^{L}\alpha_i u_i \tag{17}$$

通过以上内容, 将基于每个方面生成一个评论内容表示, 为了获得最终的内容表示, 我们需计算每个评论表示的权重. 每个评论表示S^k首先通过一个全连接网络被映射到相同的语义空间, 然后由转换后的表示t^k与高级语义向量x_s之间的相似性确定权重β^k, 高级语义向量x_s表示“这个方面对整个评论的情绪有多大影响”, 最终的评论表示S由各方面表示的加权和得到.

$$t^k = \tanh\left(W_{st} S^k + b_{st}\right) \tag{18}$$

$$\beta^k = \frac{\exp\left(\left(t^k\right)^{\mathrm{T}} x_s\right)}{\sum_{k=1}^{m}\exp\left(\left(t^k\right)^{\mathrm{T}} x_s\right)} \tag{19}$$

$$S = \sum_{k=1}^{m}\beta^k t^k \tag{20}$$

5) 有用性预测层: 为了探讨方面评分与总体评分以及评论文本对于在线评论有用性预测的影响, 所提出的模型将综合了方面评分的内容表示和总体评分纳入了预测层. 具体地说, 首先通过全连接网络将总体评分 r 映射到与内容表示相同的潜在空间, 然后得到转换的总体评分表示e_r, 将评论表示S和变换后的总体评分表示e_r连接为最终的特征向量. 最后, 将合并后的特征向量输入到一个全连接网络, 计算在线评论的有用性. 其中S为句子级注意机制获得的句子表示, W, b为全连接网络的参数:

$$e_r = E_r \cdot r \tag{21}$$

$$\hat{y} = W(S \oplus e_r) + \mathrm{b} \tag{22}$$

最后通过减少黄金标签的分布和我们的模型在N个训练样本中的投影分布之间的均方误差来训练模型 HORA. 其中y_j是真实的有用性标签, $\hat{y}_j$是预测的有用性标签, N为训练集的样本数.

$$L = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 \tag{23}$$

4 实验与分析

4.1 实验数据集

为了验证我们的预测模型 HORA 的性能, 我们从旅游网站 TripAdvisor 和健康 IT 平台 WebMD 构建了两个数据集. TripAdvisor 网站提供与旅游相关的内容的评论, 如旅游景点、酒店和餐馆. 我们收集了 239 876 条酒店评论以及相应的总体评分和有用性投票数. 除了总体评分外, 每个评论包括以下 6 个方面的评分: {清洁度、地理位置、房间、服务、睡眠质量、价格}, 范围分为 1−5. WebMD 作为美国互联网医疗健康信息服务平台, 提供患者对各种药物的广泛评论. 我们收集了 249 470 篇药物评论, 每条评论都包括评论文本、有用性投票数、总体评分和 3 个方面评分: {易用性, 有效性, 满意度}, 取值区间仍为 1−5. 在 WebMD 数据集中, 总体评分是方面评分的组合. 数据集的统计数据汇总见表 1.

参照 Mauro 等人[33]的做法, 我们使用式 (24) 和式 (25) 来将有用性投票数规范化到区间 (0−1):

$$Helpfulness = f(|votes|) \tag{24}$$

$$f(x) = \frac{\log(x+1)}{1+\log(x+1)} \tag{25}$$

我们将数据集随机分成 3 个部分: 80% 用于训练, 10% 用于验证, 10% 用于测试. 为了评价该模型的有效性, 我们采用了两个指标, 即平均绝对误差 (MAE)[34]和均方根误差 ($RMSE$)[35]来衡量预测的有用性标签和真实有用性标签之间的差异, 其定义如下:

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{26}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k-\hat{y}_k)^2} \tag{27}$$

其中, N为样本数, y_k和$\hat{y}_k$分别为第k条评论的真实标签和预测标签. *MAE* 是预测标签值和真实标签值之间绝对差异的平均值, 而 *RMSE* 是预测标签值和真实标签值之间平方差的平均值的平方根. 因为误差在平均之前需要进行平方, 所以 *RMSE* 的值要比 *MAE* 偏大.

4.2 实验参数设置

我们在 PyTorch 框架中实现了模型, 并利用 GloVe 对单词嵌入进行了预训练. 为了加快训练速度, 我们限制一个评论文档最多有 40 个句子, 每一个句子不超过 50 个单词. 所有隐藏层的维度都设置为 50; 批量大小设置为 128. 辍学率为 0.1. 当训练时, 我们使用优化器 SGD 来更新参数, 并根据经验设置初始学习率为 0.01. 在优化过程中, 首先计算预测层的梯度, 并更新相应的参数. 基于预测层的梯度, 利用链式规则确定句子级注意力机制层的梯度, 并更新相应的参数. 然后依次对单词级注意力机制和 BiGRU 层执行类似的操作. 因此, 模型的所有参数都将被更新. 优化过程在目标函数收敛后完成. 最后, 我们根据验证集上的性能选择最佳参数, 并对测试集上的参数进行评估. 为了实验的严谨性, 我们将所有模型运行了 5 次, 最后结果见表 2.

表 1 数据集统计汇总

数据集	分类	评论数	总体评分 (平均)	方面评分 (平均)	投票数 (平均)
TripAdvisor	训练集	191 902	4.038	(4.247, 4.468, 3.995, 4.181, 4.121, 3.959)	0.692
	验证集	23 987	4.125	(4.359, 4.501, 4.109, 4.293, 4.205, 3.966)	0.613
	测试集	23 987	4.119	(4.337, 4.464, 4.126, 4.281, 4.212, 4.019)	0.612
WebMD	训练集	199 582	3.571	(4.034, 3.536, 3.141)	6.791
	验证集	24 944	3.528	(3.994, 3.488, 3.102)	6.639
	测试集	24 944	3.534	(3.929, 3.542, 3.133)	6.487

表 2 各模型在两个数据集上的结果对比

模型	特征			TripAdvisor		WebMD	
	评论文本	总体评分	方面评分	*MAE*	*RMSE*	*MAE*	*RMSE*
CNN	√	√	√	0.137	0.310	0.122	0.303
Bi-LSTM	√	√	√	0.115	0.308	0.102	0.292
EG-CNN	√	—	—	0.078	0.254	0.067	0.247
MTNL	√	√	—	0.094	0.264	0.076	0.257
CM1	√	√	—	0.102	0.282	0.089	0.274
CM2	√	√	—	0.069	0.244	0.062	0.227
ECRI	√	√	—	0.058	0.224	0.054	0.219
HORA	√	—	—	0.069	0.259	0.065	0.245
	√	√	—	0.053	0.233	0.051	0.228
	√	—	√	0.046	0.218	0.041	0.214
	√	√	√	0.033	0.198	0.032	0.192

4.3 基线模型

为了系统地评估我们提出的模型, 我们从之前的研究中选择了一系列具有代表性的基线方法, 以验证 HORA 模型的性能.

CNN[36]: 对一个句子进行卷积操作, 提取相邻特征, 然后通过池化层得到固定大小的表示.

Bi-LSTM: 对单词的上下文信息进行编码, 同时捕获过去和未来的信息, 得到单词序列的完整表示.

EG-CNN[37]: 普通 CNN 架构的一种变体, 在卷积操作之前执行字符编码和词编码.

CM1[22]: 普通 CNN 架构的一种变体, 其中将原始评分值和学习的内容表示连接在一起.

CM2[22]: 普通 CNN 架构的一种变体, 其中评分向量和单词嵌入被连接来学习内容表示.

ECRI[7]: 利用内容编码器和评分增强器将评分与评论内容分开编码, 自适应地调整内容表示的潜在方

面的评分信息量.

MTNL[23]: 多任务学习的普通 CNN 架构的变体, 学习到的内容表示作为共享特征来预测评论的有用性和原始星级评分.

4.4 与基线模型对比结果

实验结果如表 2 所示, 可将其分为两部分:使用了全部文本和评分信息的基线模型, 以及使用了部分文本和评分信息的基线模型.

从第 1 部分开始, 我们可以看到, CNN 和 Bi-LSTM 的性能相对而言比较差, 即使它使用了评论文本, 总体评分和方面评分. 而与 CNN 和 Bi-LSTM 相比, 我们的模型 HORA 无论是在 TripAdvisor 数据集还是 WebMD 数据集上, 性能都有极大的提高, 均提高约 9%−11%. 从第 2 部分开始, 首先, 我们观察到 EG-CNN 仅用到评论文本对问题进行建模, 而和同等条件下的 HORA 相比, HORA 在 TripAdvisor 数据集上的 *MAE* 和 WebMD 数据集上的 *MAE*, *RMSE* 的效果明显优于 EG-CNN, 虽然 TripAdvisor 数据集上的 *RMSE* 效果略差. 当应用更复杂的神经网络时, 比如 MTNL, CM1, CM2, ECRI, 在同等条件下, 对于 TripAdvisor 数据集, 我们的模型在 *MAE* 和 *RMSE* 方面分别比之前的基线模型高出 1%−5% 和 1%−4%. 对于 WebMD 数据集, 我们的模型在 *MAE* 和 *RMSE* 方面分别比之前的基线模型高出 0.3%−4% 和 1%−4%. 对比表明, 我们的模型架构可以生成更有效的评论内容表示以提高在线评论有用性的预测性能.

4.5 多维评分的影响

我们已经展示了 HORA 的结构对在线评论预测任务的有效性. 为了显示多维评分对于在线评论有用性的作用, 我们还比较了 HORA 在 3 种特征 (即评论文本、总体评分和方面评分) 的不同组合下的表现. 具体来说, 我们在保持其他参数设置不变的同时改变输入来执行不同维度的特征融合.

评论文本、总体评分和各个方面的评分是 HORA 中所用到的 3 种信息. 我们在表 2 下方展示了评论文本、总体评分和方面评分对在线评论有用性预测的影响. 从表中, 我们可以观察到: (1) 与只使用评论的文本特征相比, 加入总体评分可以在两个数据集中分别实现 1.6%、2.6%、1.4% 和 1.7% 的改进, 这表明该模型对总体评分进行编码, 可以获得更丰富的文档表示, 对有用性预测帮助更大. (2) 与仅用总体评分信息的模型相比, 考虑方面评分的模型在性能提升上更加明显, 可以在两个数据集中分别获得 2.3%、4.1% 和 2.4%、3.1% 的性能提升, 这表明方面评分有助于预测在线评论的有用性. (3) 在综合考虑了评论文本、总体评分和方面评分以后, 与其他基线模型相比, HORA 具有更小的平均绝对误差 (*MAE*) 和均方根误差 (*RMSE*), 即我们的模型获得了最好的性能. 这说明充分利用用户留下来的信息数据, 进行多维度信息交互可以提高在线评论有用性的预测性能.

观察结果表明, 我们的模型以一种更有效的方式整合了评论文本, 总体评分, 方面评分三者的信息, 最终改进了评论有用性的预测性能.

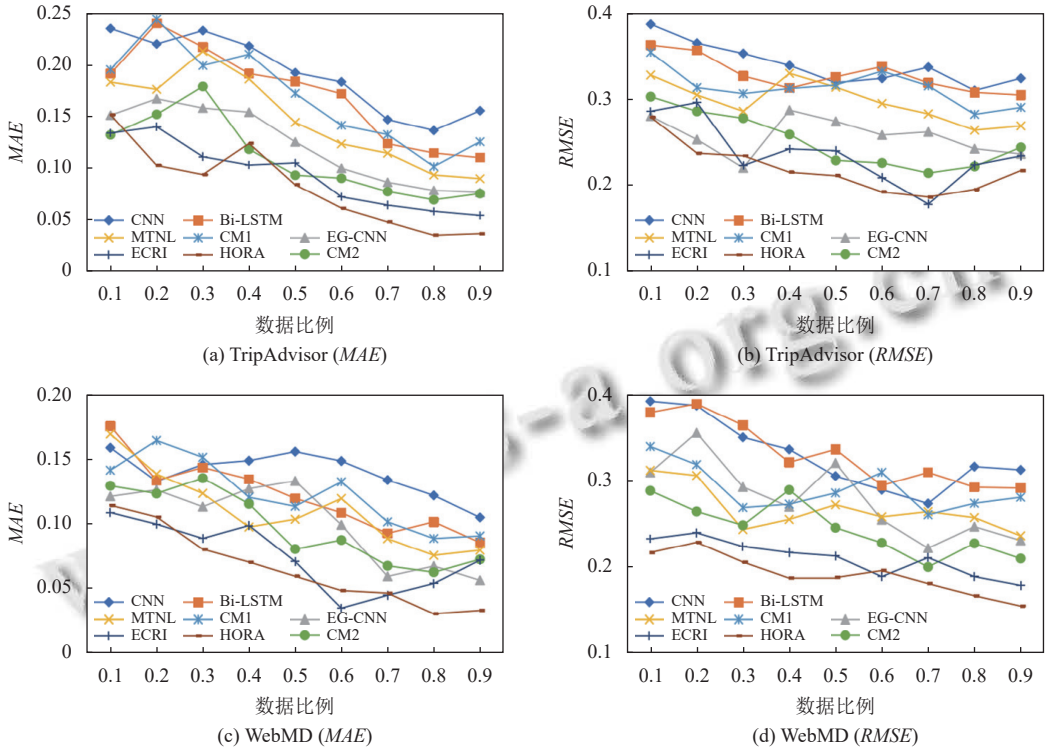
4.6 稳健性实验

为了显示 HORA 的稳健性, 我们还进行了敏感性分析[38]. 具体来说, 我们将比率从 0.1 改变到 0.9, 并检查当我们的方法与其他方法进行比较时, 性能是如何变化的. 如图 4 所示, 当训练数据的比例增加时, 两种评价指标的值都呈下降趋势, 即性能都有随着训练数据比例的增加而增加的总体趋势, 特别是当该比值等于和低于 0.7 时. 结果表明, 当比值大于 0.7 时, 大多数模型的性能都趋于稳定. 总的来说, 我们的模型比这些基线模型表现得更好. 对于一些特定的模型, 如 ECRI, 与我们在 TripAdvisor 数据集上的模型性能相当, 但它们在 WebMD 数据集上实现的性能较低. 当训练数据的比例高于 0.2 时, HORA 在 *MAE* 和 *RMSE* 两个评估指标上始终优于其他基线, 尽管大多数模型的性能曲线是混合的. 总的来说, 我们的方法比这些基线表现得更好, 而且也更稳健. 这说明在数据处理层面上, HORA 模型在面临各类噪声情况下, 如虚假恶意评论, 数据缺失等问题时, 仍然能够较好地处理这些干扰, 提供合理的评论分析和情感判断以此来作出准确的有用性预测.

4.7 理论解释

从众理论[19]认为消费者会根据大众的评价而做出购买决策. 一般消费者在确认自己的需求后, 便会自动地进入到信息搜索的过程, 而有经济头脑的消费者并不会因为最小努力原则而单纯从总体评分去做出购买决策, 因为评论星级的离散程度过大, 即评论者的总体星级水平远低于或高于平均星级水平时, 该条评论的可信度就会下降. 归因理论[39]认为评论者此时可能受到了除了产品因素以外的其他因素的刺激, 例如心情

或者恶意发布虚假信息等, 因此大部分在线购买者都会充分查看已购买者留下来的信息数据, 比如除了总体评分以外, 还有产品的各个方面的评分, 以此来帮助自己作出正确的购买决策, 从而减少错误的购买成本.



(a) TripAdvisor (*MAE*) (b) TripAdvisor (*RMSE*)

(c) WebMD (*MAE*) (d) WebMD (*RMSE*)

图 4 所有模型在不同训练数据比例下的性能表现

5 总结与展望

预测在线评论的有用性和推荐有用的在线产品评论可以帮助用户做出明智的购买决策. 之前的方法主要集中于使用评论文本和总体评分等信息预测评论的有用性. 虽然考虑了评分, 但方面评分和其他潜在的可用性数据没有考虑. 本文提出了 HORA, 一种用于评论有用性预测的深度神经网络模型, 它可以学习评分增强的内容表示. 与之前只使用文本内容和总体评分信息的工作相比, HORA 采用多层次注意力机制, 利用多维度评分信息, 尤其是方面评分来统一文本效价和评分效价之间的一致性. 在两个真实数据集上进行的大量实验表明了 HORA 在学习文本特征和利用评分信息方面的前景.

虽然我们已经证明了 HORA 的有效性, 但本研究仍有其局限性, 有待于未来的进一步改进. 首先, 虚假评论检测是一个非常重要和具有挑战性的问题. 为了避免虚假评论的影响, 我们可以通过添加数据预处理层来扩展该框架, 该层可以自动识别基于内容的特征, 并从用户的行为中去判断. 其次, 我们的方法侧重于整合总体评分, 方面评分和评论内容来预测有用性, 在未来, 可以将用户, 产品等更多相关的信息集成到统一的框架中. 最后, 当前对于在线评论有用性预测这一研究领域, 研究人员并没有建立一个统一的数据集, 未来将需要一个共同的论坛来讨论这一领域的长期愿景.

参考文献

1 冯进展, 蔡淑琴. 融合信息增益和梯度下降算法的在线评论有用程度预测模型. 计算机科学, 2020, 47(10): 69–74. [doi: 10.11896/jsjkx.190700034]

2 陈远高, 应梦茜, 毕然, 等. 管理者回复对在线评论与有用性关系的调节效应: 基于 TripAdvisor 的实证研究. 管理工程学报, 2021, 35(5): 110–116.

3 Zhou SS, Guo B. The interactive effect of review rating and text sentiment on review helpfulness. Proceedings of the 16th International Conference on Electronic Commerce and Web Technologies. Valencia: Springer, 2015. 100–111. [doi: 10.1007/978-3-319-27729-5_8]

4 Diaz GO, Ng V. Modeling and prediction of online product

review helpfulness: A survey. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 698–708. [doi: 10.18653/v1/P18-1065]

5 Karimi S, Wang F. Online review helpfulness: Impact of reviewer profile image. Decision Support Systems, 2017, 96: 39–48. [doi: 10.1016/j.dss.2017.02.001]

6 Wu Z, Dai X Y, Yin C Y, *et al*. Improving review representations with user attention and product attention for sentiment classification. Proceedings of the 2018 AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. [doi: 10.1609/aaai.v32i1.12054]

7 Du JH, Rong J, Wang H, *et al*. Helpfulness prediction for online reviews with explicit content-rating interaction. Proceedings of the 20th International Conference on Web Information Systems Engineering. Hong Kong: Springer, 2019. 795–809.

8 Schneider C, Weinmann M, Mohr PNC, *et al*. When the stars shine too bright: The influence of multidimensional ratings on online consumer ratings. Management Science, 2021, 67(6): 3871–3898. [doi: 10.1287/mnsc.2020.3654]

9 Bu JH, Ren L, Zheng S, *et al*. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 2069–2079. [doi: 10.18653/v1/2021.naacl-main.167]

10 Fei H, Li JY, Ren YF, *et al*. Making decision like human: Joint aspect category sentiment analysis and rating prediction with fine-to-coarse reasoning. Proceedings of the 2022 ACM Web Conference. Lyon: ACM, 2022. 3042–3051. [doi: 10.1145/3485447.3512024]

11 Hazarika B, Chen K, Razi M. Are numeric ratings true representations of reviews? A study of inconsistency between reviews and ratings. International Journal of Business Information Systems, 2021, 38(1): 85–106. [doi: 10.1504/IJBIS.2021.118637]

12 Martin L, Pu P. Prediction of helpful reviews using emotions extraction. Proceedings of the 2014 AAAI Conference on Artificial Intelligence. Québec City: AAAI, 2014. [doi: 10.1609/aaai.v28i1.8937]

13 Yang YF, Chen C, Bao FS. Aspect-based helpfulness prediction for online product reviews. Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence (ICTAI). San Jose: IEEE, 2016. 836–843. [doi: 10.1109/ICTAI.2016.0130]

14 Liu HJ, Gao Y, Lv P, *et al*. Using argument-based features to predict and analyse review helpfulness. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1358–1363.

15 Ghose A, Ipeirotis P G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(10): 1498–1512. [doi: 10.1109/TKDE.2010.188]

16 Korfiatis N, García-Bariocanal E, Sánchez-Alonso S. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. Electronic Commerce Research and Applications, 2012, 11(3): 205–217. [doi: 10.1016/j.elerap.2011.10.003]

17 Otterbacher J. 'Helpfulness' in online communities: A measure of message quality. Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems. Boston: ACM, 2009. 955–964. [doi: 10.1145/1518701.1518848]

18 Chen C, Yang YF, Zhou J, *et al*. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans: ACL, 2018. 602–607. [doi: 10.18653/v1/N18-2095]

19 Lee S, Lee S, Baek H. Does the dispersion of online review ratings affect review helpfulness? Computers in Human Behavior, 2021, 117: 106670.

20 Lee S, Choeh JY. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Systems with Applications, 2014, 41(6): 3041–3046. [doi: 10.1016/j.eswa.2013.10.034]

21 Mukherjee S, Popat K, Weikum G. Exploring latent semantic factors to find useful product reviews. Proceedings of the 2017 SIAM International Conference on Data Mining. Houston: SDM, 2017. 480–488.

22 Qu XS, Li XP, Rose JR. Review helpfulness assessment based on convolutional neural network. arXiv:1808.09016, 2018.

23 Fan M, Feng Y, Sun MM, *et al*. Multi-task neural learning architecture for end-to-end identification of helpful reviews. Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona: IEEE, 2018. 343–350. [doi: 10.1109/ASONAM.2018.8508623]

24 Sun XY, Han MX, Feng J. Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. Decision Support Systems, 2019, 124: 113099. [doi: 10.1016/j.dss.2019.113099]

25 Bilal M, Marjani M, Lali MI, *et al*. Profiling users' behavior, and identifying important features of review "helpfulness". IEEE Access, 2020, 8: 77227–77244. [doi: 10.1109/ACCESS.2020.2989463]

26 Qu XS, Li XP, Farkas C, *et al*. An attention model of customer expectation to improve review helpfulness prediction. Proceedings of the 42nd European Conference on Advances in Information Retrieval. Lisbon: Springer, 2020. 836–851.

27 Fan M, Feng C, Guo L, *et al*. Product-aware helpfulness prediction of online reviews. Proceedings of the 2019 World Wide Web Conference. San Francisco: ACM, 2019. 2715–2721.

28 Li ST, Pham TT, Chuang HC. Do reviewers' words affect predicting their helpfulness ratings? Locating helpful reviewers by linguistics styles. Information & Management, 2019, 56(1): 28–38. [doi: 10.1016/j.im.2018.06.002]

29 Hong Y, Lu J, Yao JM, *et al*. What reviews are satisfactory: Novel features for automatic helpfulness voting. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland: ACM, 2012. 495–504. [doi: 10.1145/2348283.2348351]

30 Kim SM, Pantel P, Chklovski T, *et al*. Automatically assessing review helpfulness. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney: ACM, 2006. 423–430.

31 Liu JJ, Cao YB, Lin CY, *et al*. Low-quality product review detection in opinion summarization. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague: ACL, 2007. 334–342.

32 Yang ZC, Yang DY, Dyer C, *et al*. Hierarchical attention networks for document classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 1480–1489.

33 Mauro N, Ardissono L, Petrone G. User and item-aware estimation of review helpfulness. Information Processing & Management, 2021, 58(1): 102434.

34 Jin ZP, Li QD, Zeng DD, *et al*. Jointly modeling review content and aspect ratings for review rating prediction. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa: ACM, 2016. 893–896. [doi: 10.1145/2911451.2914692]

35 Chen HM, Sun MS, Tu CC, *et al*. Neural sentiment classification with user and product attention. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 1650–1659.

36 Chen YH. Convolutional neural network for sentence classification [Master's thesis]. Waterloo: University of Waterloo, 2015.

37 Chen C, Qiu MH, Yang YF, *et al*. Review helpfulness prediction with embedding-gated CNN. arXiv:1808.09896, 2018.

38 Zhang Z, Wei X, Zheng XL, *et al*. Predicting product adoption intentions: An integrated behavioral model-inspired multiview learning approach. Information & Management, 2021, 58(7): 103484.

39 苗蕊, 徐健. 评分不一致性对在线评论有用性的影响——归因理论的视角. 中国管理科学, 2018, 26(5): 178–186. [doi: 10.16381/j.cnki.issn1003-207x.2018.05.018]

(校对责编: 孙君艳)