计算机科学
COMPUTER SCIENCE

# 基于注意力机制的多模态在线评论有用性预测研究

张逸安, 杨颖, 任刚, 王刚

引用本文

张逸安, 杨颖, 任刚, 王刚.基于注意力机制的多模态在线评论有用性预测研究[J]. 计算机科学, 2023, 50(8): 37-44.

ZHANG Yian, YANG Ying, REN Gang, WANG Gang. Study on Multimodal Online Reviews Helpfulness Prediction Based on Attention Mechanism [J]. Computer Science, 2023, 50(8): 37-44.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600204

# 基于注意力机制的多模态在线评论有用性预测研究

张逸安[1]　杨　颖[2]　任　刚[2]　王　刚[2]
1 南京大学信息管理学院　南京 210023
2 合肥工业大学管理学院　合肥 230009
(yianzh1004@163.com)

**摘　要**　在电子商务时代,在线评论被视为一类重要的商品评价,深刻影响着消费者的决策过程。但是指数级增长的评论数量和非结构化的评论数据给评论有用性预测模型的特征选择和精确度提升带来了挑战。此外,目前的研究主要集中于浅层特征和评论文本的特征提取,往往忽略了评论照片所包含的图像信息,同时评论文本、照片、浅层特征这些多模态的信息需要应用多模态融合方法进行信息的提炼融合。基于此,文中将评论照片和评论文本作为影响在线评论有用性的潜在特征,并根据KAM知识采纳理论设计浅层特征集合。对于3种模态的数据,提出了一种基于协同注意力机制的三模态评论有用性预测模型(TMCAM),用于实现跨模态信息的交互和融合。实验结果检验了TMCAM模型的优越性能,证明了图像和文本信息的互补能够达到比单一模态信息更好的效果;浅层特征能够辅助预测评论有用性;相比简单的模态特征拼接,利用协同注意力机制进行跨模态信息交互有助于提升对评论有用性的感知。

**关键词**:评论有用性;协同注意力机制;多模态融合;自然语言处理;深度学习
**中图法分类号**　TP391.1

# Study on Multimodal Online Reviews Helpfulness Prediction Based on Attention Mechanism

ZHANG Yian[1], YANG Ying[2], REN Gang[2] and WANG Gang[2]
1 School of Information Management, Nanjing University, Nanjing 210023, China
2 School of Management, Hefei University of Technology, Hefei 230009, China

**Abstract**　In the e-commerce era, online reviews are regarded as important product evaluations, which profoundly influence consumers' decision-making process. However, the exponentially increasing number of reviews and unstructured review data pose challenges to feature selection and accuracy improvement of review helpfulness prediction. In addition, current research mainly focuses on shallow features and feature extraction of review texts, the image information contained in review photos is often ignored. Besides, multi-modal information such as review text, photos, and shallow features needs to be refined and fused by applying multi-modal fusion methods. Based on these, this paper regards review photos and review text as a latent feature affecting the helpfulness of online reviews, and designs a shallow feature set according to the KAM knowledge adoption theory. For the data of three modalities, a deep prediction model, i.e., three-modal review helpfulness prediction based on co-attention mechanism (TMCAM) is proposed, which can achieve the interaction and fusion of cross-modal information. The superior performance of the TMCAM model is tested through experiments, and it is proved that the complementation of image and text information can achieve better results than single modal information. Besides, shallow features can help predict the reviews helpfulness. Moreover, compared with simple modal features splicing, using collaborative attention mechanism for cross-modal information interaction helps to improve the perception of reviews helpfulness.

**Keywords**　Review helpfulness, Co-attention mechanism, Multimodal fusion, Natural language processing, Deep learning

## 1　引言

随着社交媒体和电子商务的发展,在线评论渗透到网络消费的各个方面,被视为一类重要的商品评价,蕴含丰富的信息和待发掘的商业价值。消费者通过阅览评论信息,来减少购买过程中的不确定性和风险,从而提升购买体验,帮助其做出决策[1]。根据2016年BrightLocal的调查显示,有90%的消费者会阅读至少10条在线评论以形成对该商品的评价。

到稿日期:2022-06-22　返修日期:2022-11-04
基金项目:国家自然科学基金(72071062,71471054,72071061)
This work was supported by the National Natural Science Foundation of China(72071062,71471054,72071061).
通信作者:王刚(wgedison@hfut.edu.cn)

然而，真实的购买体验会比较复杂，评论中隐含的非对称信息并没有完全呈现给消费者。同时，指数级增长的评论数量、参差不齐的评论质量会使得在线评论的利用程度降低[2-3]。这与消费者希望花费较短的时间做出满意的决策相悖，因此需要设计一个科学有效的衡量在线评论有用性的机制，帮助消费者筛选出真正有价值的评论。

目前，大量研究者在计量经济学或传统机器学习领域通过人工提取与在线评论相关的浅层特征，并使用各类回归模型[4-5]、支持向量机[3,6]等方法来预测在线评论的有用性。然而，单纯用浅层特征难以完整表征在线评论。相比以上方法，使用深度学习网络来进行文本特征学习已经被证明是有效的，并且能够取代传统特定任务的特征工程[7]。Saumya 等[8]使用了预训练的 Word2Vec 与 Glove 进行文本词嵌入，并使用两层的卷积神经网络计算在线评论有用性得分。Fan 等[9]使用 Bi-LSTM 处理在线评论信息。Xu 等[10]引入了 BERT 预训练模型作为特征提取器。

一系列研究证明了深度学习方法在评论的特征表示上有更本质的刻画，但仍然存在以下不足：1)浅层特征不应被深层特征完全取代，部分浅层特征可以作为一种特殊的视角或模态去表征评论信息；2)评论照片作为在线评论的重要组成部分，对消费者感知评论有用性产生了不可忽视的影响，但在以往的研究中并没有得到重视；3)在线评论中存在多模态的数据，它们并不是独立对在线评论有用性产生影响[11]，这些多模态数据之间的交互作用如何建模、是否对评论有用性的预测产生影响仍然存疑。

综上所述，本文从深度学习的视角出发，提出了一种基于协同注意力机制的三模态评论有用性预测模型(TMCAM)，将在线评论的文本、照片和浅层特征作为输入信息，预测在线评论的有用性。该模型结合了图像、文本预训练模型和 KAM 知识采纳理论提取特征集合，并使用所构建的协同注意力机制实现多模态信息的融合。本文通过与其他深度学习模型的对比，检验了 TMCAM 模型的优越性能。通过消融实验，探讨了图像信息是否对人们感知在线评论存在潜在贡献；验证了浅层特征相对于深层特征的必要性；并对多模态特征融合的效用进行了考察。

## 2 相关工作

### 2.1 在线评论有用性预测

在线评论有用性的具体含义是什么？Mudambi 等[4]认为一条有用的在线评论能够促使消费者做出购买的决策。Pan 等[12]和 Li 等[13]把在线评论的有用性定义为，消费者认为在线评论能够促使其产生购买决策或衡量其作出判断的程度。消费者对评论有用性的感知取决于消费者对评论来源和评论内容的主观态度。当潜在消费者阅读评论时，通过点击评论下方的“投票”按钮表达对评论的感知价值。因此，在实际研究中，在线评论的有用性由评论的帮助性投票总数或者帮助性投票占所有投票总数的比例来衡量[14-15]。

在线评论有用性的研究方法主要可以分为计量经济学方法、传统机器学习方法和深度学习方法。计量经济学与机器学习方法集中于挖掘在线评论的浅层特征与评论有用性的关系，例如评论评级[2,4,15]、评论易读性特征[3,16]、评论者相关特征[1,17]。但人工提取评论文本的特征存在许多困难，不能很好处理评论文本这些非结构化数据。

随着深度学习的兴起，评论文本的特征学习有了更强大的工具。Du 等[18]提出，相比浅层次的外围特征，深度学习模型能够从评论文本中学习更丰富的语义特征。越来越多的研究者开始使用深度学习模型对评论文本提取评论特征。但在线评论并不只有文本数据，还有评论照片信息。以往在线评论的照片由于技术上的挑战，其信息价值并没有被很好地理解。Ma 等[11]意识到了评论照片可能存在的价值，发现同时考虑评论文本和图像信息会产生更好的预测结果。然而，该研究是假设评论文本和照片独立对在线评论有用性产生影响，没有考虑跨模态信息的交互作用。在线评论有用性的研究中存在多模态的数据，有必要采用多模态融合的技术方法。这种基于多模态信息和特征融合的研究思路如图 1 所示。

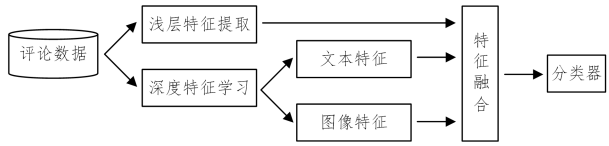



图 1 基于多模态信息和特征融合的研究思路图

Fig. 1 Research idea based on multimodal information and feature fusion

### 2.2 多模态融合

不同的存在形式或信息来源均可被称为一种模态，由两种或两种以上模态组成的数据被称为多模态数据。多模态融合指对多模态数据进行统筹处理，融合或提炼跨模态数据，从而为下游任务提供更全面充足的信息。在多模态融合任务中，不管是文本还是图像信息都被深度学习模型表示为实值向量，这使得跨多个模态执行信息处理成为可能[19]。

多模态融合方法可以分为与模型无关的方法和基于模型的方法。前者不依赖于具体的深度学习模型，融合过程简易且容易造成信息损失。而基于模型的方法具有更好的性能和稳定性，其中协同注意力机制(Co-Attention)为多模态信息融合提出了一种出色的解决方案。传统的注意力机制通常只处理单模态的信息，通过查询向量(Query)和键向量(Key)计算注意力分布来实现对输入信息(Value)的加权。而协同注意力机制是对不同模态的信息序列进行计算[20]。具体来说，查询向量和键向量在协同注意力机制中分别代表了不同的模态信息，如文本、图像的信息融合，通过计算文本特征和图像特征的相似度，来得到对各自模态信息的注意力分布，实现信息提炼。多模态融合和注意力机制的典型应用领域有视觉问答、图文匹配等，典型的模型有 HieCoAtt[21]，SAN[22]，DAN[23]，SAA[24]等。多模态融合的概念适用于在线评论有用性预测，能够为不同模态评论信息的交互融合提供思路。

## 3 基于协同注意力机制的三模态评论有用性预测

### 3.1 整体模块框架

本文对研究问题进行了定义。假设在评论有用性预测任务中，有 $N$ 条在线评论，其中第 $i(i=1,2,\cdots,N)$ 条评论包含一段评论文本 $Review_i$、第一张评论照片 $Image_i$ 和浅层特征

集 $\boldsymbol{S}_i$，预测出的在线评论有用性为 $H_i$。其中 $\boldsymbol{S}_i=\{\boldsymbol{s}_{i1},\boldsymbol{s}_{i2},\cdots,\boldsymbol{s}_{iL}\}$，$\boldsymbol{s}_{ij}$ 为第 $i$ 条评论的第 $j$ 个浅层特征，$L$ 为特征个数。评论文本和照片经过特征提取后，$\boldsymbol{R}_i=\{\boldsymbol{w}_{i1},\boldsymbol{w}_{i2},\cdots,\boldsymbol{w}_{iT}\}$，$\boldsymbol{w}_{ij}$ 为第 $i$ 条评论的第 $j$ 个字符的特征表示，$T$ 为评论文本嵌入长度；$\boldsymbol{I}_i=\{\boldsymbol{v}_{i1},\boldsymbol{v}_{i2},\cdots,\boldsymbol{v}_{iZ}\}$，$\boldsymbol{v}_{ij}$ 为第 $i$ 条评论照片的第 $j$ 个图像区域的特征表示，$Z$ 为区域数量。

结合以上定义，本文构建了一种基于协同注意力机制的三模态评论有用性预测模型（Three-Modal Review Helpfulness Prediction Based on Co-Attention Mechanism，TMCAM）。整体模型由输入模块、特征表示模块、协同注意力网络模块和输出模块组成，如图 2 所示。为简洁起见，在接下来的公式描述中忽略模型中的偏差项 $b$。

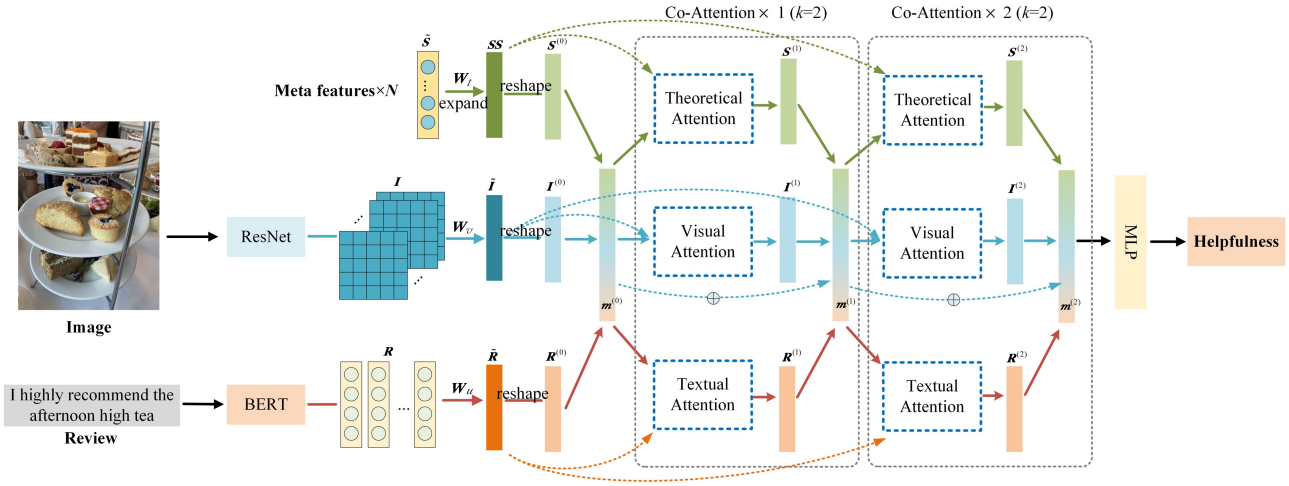



图 2 基于协同注意力机制的三模态评论有用性预测模型 TMCAM

Fig. 2 TMCAM model

### 3.2 特征表示模块

#### 3.2.1 文本、图像特征表示

（1）评论文本特征表示

TMCAM 使用 BERT 预训练模型提取评论文本特征，BERT 能够处理不同长短的评论文本，准确挖掘长距离文本的语义信息。本文设置文本最大长度为 $T=512$。

$$\boldsymbol{R}_i=BERT_{\text{Pre}}(Review_i) \tag{1}$$

对于每段评论文本 $Review_i$，使用 BERT 进行深度特征提取后表示为 $R_i=\{\boldsymbol{w}_{i1},\boldsymbol{w}_{i2},\cdots,\boldsymbol{w}_{iT}\}\in\mathbb{R}^{T\times d_w}$。$R_i$ 会经过一个全连接层 $\boldsymbol{W}_u$ 实现维度对齐，生成 $\widetilde{\boldsymbol{R}}_i=\{\widetilde{\boldsymbol{w}}_{i1},\widetilde{\boldsymbol{w}}_{i2},\cdots,\widetilde{\boldsymbol{w}}_{iT}\}\in\mathbb{R}^{T\times d}$。

$$\widetilde{R}_i=\text{LeakyRelu}(W_uR_i) \tag{2}$$

（2）评论图像特征表示

ResNet 网络应用于图像数据，通过引入残差结构解决了梯度消失和梯度爆炸的问题。TMCAM 模型采用 ResNet 进行图像特征的提取，输出取 ResNet 网络最后一个平均池化层（Average Pooling）前的结果。

$$\boldsymbol{I}_i=ResNet_{\text{Pre}}(Image_i) \tag{3}$$

对于每张评论照片 $Image_i$，使用 ResNet 进行深度特征提取后表示为 $\boldsymbol{I}_i=\{\boldsymbol{v}_{i1},\boldsymbol{v}_{i2},\cdots,\boldsymbol{v}_{iZ}\}\in\mathbb{R}^{Z\times d_v}$，$\boldsymbol{I}_i$ 是一个三维的张量，其中 $Z=14\times14$。对维度 $Z$ 进行铺平操作后，$\boldsymbol{I}_i$ 同样经过一个全连接层 $\boldsymbol{W}_v$，生成 $\widetilde{\boldsymbol{I}}_i=\{\widetilde{\boldsymbol{v}}_{i1},\widetilde{\boldsymbol{v}}_{i2},\cdots,\widetilde{\boldsymbol{v}}_{iZ}\}\in\mathbb{R}^{Z\times d}$。

$$\widetilde{\boldsymbol{I}}_i=\text{LeakyRelu}(\boldsymbol{W}_v\boldsymbol{I}_i) \tag{4}$$

#### 3.2.2 浅层特征提取

在已有研究中，有许多基于浅层特征的用于解释在线评论有用性的理论。KAM 知识采纳理论于 2003 年由 Sussman 等[25]提出，并受到了广泛认可。该理论认为信息接受者对信息有用性的感知是由信息本身的质量和信息来源的可信度决定的[25]。TMCAM 模型依据 KAM 理论进行浅层特征的提取，从评论信息质量和评论信息来源两个维度，共 11 个特征来定义在线评论有用性，如表 1 所列。

表 1 基于 KAM 理论提取的浅层特征集合

Table 1 Shallow feature set extracted based on KAM theory

| 特征类别 | 特征 | 特征说明 |
|---|---|---|
| 评论信息来源 | $s_{i1}$ | 评论者粉丝数 |
| | $s_{i2}$ | 评论者发布评论数 |
| | $s_{i3}$ | 专家标签 |
| | $s_{i4}$ | 评论者发布照片数 |
| 评论信息质量 | $s_{i5}$ | 评论评级 |
| | $s_{i6}$ | 评论长度 |
| | $s_{i7}$ | Flesch-Kincaid Reading Ease score(FKRE) |
| | $s_{i8}$ | Automated Readability Index(ARI) |
| | $s_{i9}$ | Simple Measure of Gobbledygook(SMOG) |
| | $s_{i10}$ | Gunning Fog Index(GFI) |
| | $s_{i11}$ | Dale-Chall Score(DCS) |

评论信息来源维度包含 4 个特征：评论者粉丝数、评论者发布评论数、专家标签和评论者发布照片数。评论信息质量维度包含 7 个特征：评论评级、评论长度以及 5 个易读性指标，即 FKPE，ARI，SMOG，GFI，DCS。11 个特征共同构成了 $\boldsymbol{S}_i$，随后将 $\boldsymbol{S}_i$ 输入两层的浅层感知机 $\boldsymbol{W}_t^1$ 和 $\boldsymbol{W}_t^2$，初步对浅层特征进行信息学习，$\widetilde{\boldsymbol{S}}_i=\{\widetilde{s}_{i1},\widetilde{s}_{i2},\cdots,\widetilde{s}_{iL}\}\in\mathbb{R}^d$。

$$\widetilde{\boldsymbol{S}}_i=\text{LeakyRelu}(\boldsymbol{W}_t^2(\text{LeakyRelu}(\boldsymbol{W}_t^1\boldsymbol{S}_i))) \tag{5}$$

### 3.3 协同注意力网络模块

如图 2 所示，TMCAM 模型通过多个步骤同时实现对文本、图像和浅层特征的注意力机制，并且通过记忆变量 *memory* 实现跨模态的信息交互，如图 3 所示。

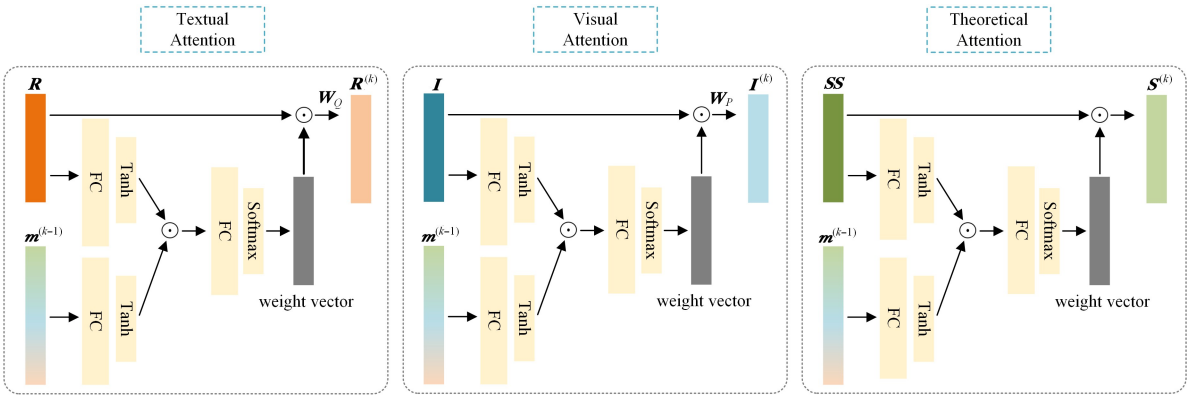



图 3 TMCAM 的三重注意力架构图

Fig. 3 Triple attention architecture diagram of TMCAM

### 3.3.1 文本注意力机制

文本注意力机制的目的是通过关注评论文本的重要部分来提炼评论文本信息。在协同注意力网络迭代的第 $k$ 次，$\widetilde{\boldsymbol{R}}_i=\{\widetilde{\boldsymbol{w}}_{i1},\widetilde{\boldsymbol{w}}_{i2},\cdots,\widetilde{\boldsymbol{w}}_{iT}\}$ 经过文本注意力机制生成优化向量 $\boldsymbol{R}_i^{(k)}$。

$$\boldsymbol{C}_{w,ij}^{(k)}=\mathrm{Tanh}(\boldsymbol{W}_w^{(k)}\widetilde{\boldsymbol{w}}_{ij})\odot\mathrm{Tanh}(\boldsymbol{W}_{w,m}^{(k)}\boldsymbol{m}_i^{(k-1)}) \tag{6}$$

$$\boldsymbol{\alpha}_{w,ij}^{(k)}=\mathrm{softmax}(\boldsymbol{W}_{w,C}^{(k)}\boldsymbol{C}_{w,ij}^{(k)}) \tag{7}$$

$$\boldsymbol{R}_i^{(k)}=\mathrm{Tanh}(\boldsymbol{Q}^{(k)}\sum_{j=1}^{T}\boldsymbol{\alpha}_{w,ij}^{(k)}\widetilde{\boldsymbol{w}}_{ij}) \tag{8}$$

其中，$\boldsymbol{m}_i^{(k-1)}$ 是记忆向量，由第 $k-1$ 次迭代时编码生成，具体生成步骤在 3.3.4 节详细介绍；$\odot$ 代表对应位置元素相乘，$\boldsymbol{C}_{w,ij}^{(k)}$ 是第 $k$ 次文本注意力机制生成的评论文本融合相似度；$\boldsymbol{\alpha}_{w,ij}^{(k)}$ 为概率向量。其中 $\boldsymbol{W}_w^{(k)}$，$\boldsymbol{W}_{w,m}^{(k)}$，$\boldsymbol{W}_{w,C}^{(k)}$，$\boldsymbol{Q}^{(k)}$ 均为可学习的网络参数，$\boldsymbol{W}_{w,C}^{(k)}$ 作为线性层进行融合相似度的加权求和，$\boldsymbol{Q}^{(k)}$ 实现与其余注意力模块信息的维度兼容。

### 3.3.2 图像注意力机制

图像注意力机制的结构与文本注意力机制大体类似。将评论照片特征与记忆向量 $\boldsymbol{m}^{(k-1)}$ 进行交互以实现特征提炼。在协同注意力网络的第 $k$ 次迭代，$\widetilde{\boldsymbol{I}}_i=\{\widetilde{\boldsymbol{v}}_{i1},\widetilde{\boldsymbol{v}}_{i2},\cdots,\widetilde{\boldsymbol{v}}_{iZ}\}$ 经过图像注意力机制生成优化向量 $\boldsymbol{I}_i^{(k)}$。

$$\boldsymbol{C}_{v,ij}^{(k)}=\mathrm{Tanh}(\boldsymbol{W}_v^{(k)}\widetilde{\boldsymbol{v}}_{ij})\odot\mathrm{Tanh}(\boldsymbol{W}_{v,m}^{(k)}\boldsymbol{m}_i^{(k-1)}) \tag{9}$$

$$\boldsymbol{\alpha}_{v,ij}^{(k)}=\mathrm{softmax}(\boldsymbol{W}_{v,C}^{(k)}\boldsymbol{C}_{v,ij}^{(k)}) \tag{10}$$

$$\boldsymbol{I}_i^{(k)}=\mathrm{Tanh}(\boldsymbol{P}^{(k)}\sum_{j=1}^{Z}\boldsymbol{\alpha}_{v,n}^{(k)}\widetilde{\boldsymbol{v}}_{ij}) \tag{11}$$

### 3.3.3 浅层特征注意力机制

浅层特征注意力机制旨在实现浅层特征与深度学习模型的融合。与图像和文本深度特征不同，$\widetilde{\boldsymbol{S}}_i$ 为一维向量，因此将其重复扩展至一个二维张量，即用一个元素相同的向量来替代原来的一个浅层特征。

$$\boldsymbol{SS}_i=\mathrm{expand\_as}(\widetilde{\boldsymbol{S}}_i) \tag{12}$$

$\boldsymbol{SS}_i=\{ss_{i1},ss_{i2},\cdots,ss_{iL}\}\in\mathbb{R}^{d\times d}$，$\boldsymbol{SS}_i$ 通过与记忆向量 $\boldsymbol{m}^{(k-1)}$ 进行交互生成优化向量 $\boldsymbol{S}_i^{(k)}$。

$$C_{s,ij}^{(k)}=\mathrm{Tanh}(\boldsymbol{W}_s^{(k)}ss_{ij})\odot\mathrm{Tanh}(\boldsymbol{W}_{s,m}^{(k)}\boldsymbol{m}_i^{(k-1)}) \tag{13}$$

$$\alpha_{s,ij}^{(k)}=\mathrm{softmax}(\boldsymbol{W}_{s,C}^{(k)}\boldsymbol{C}_{s,ij}^{(k)}) \tag{14}$$

$$\boldsymbol{S}_i^{(k)}=\mathrm{permute}(\sum_{ss}\boldsymbol{\alpha}_{s,ij}^{(k)}\boldsymbol{ss}_{ij}) \tag{15}$$

在生成 $\boldsymbol{S}_i^{(k)}$ 时，与评论文本和图像不同，选择对每个 $ss_{ij}$ 进行加权求和，原因是 $\boldsymbol{SS}_i$ 是通过重复扩展而来的，维度 $j$ 代表的是不同视角的浅层特征，每个 $ss_{ij}$ 代表的是同一个浅层特征的不同含义扩展。因此，这样求和保留了浅层特征的视角全面性，permute(·)是维度交换函数。

### 3.3.4 记忆向量与网络迭代

在 TMCAM 模型中，为了实现跨模态的信息交互，支持三重注意力机制的信息融合，本文构造了一个记忆向量 $\boldsymbol{m}^{(k)}$，$\boldsymbol{m}^{(k)}$ 的更新模式为：

$$\boldsymbol{m}_i^{(k)}=\boldsymbol{m}_i^{(k-1)}+[\boldsymbol{R}_i^{(k)},\boldsymbol{I}_i^{(k)},\boldsymbol{S}_i^{(k)}] \tag{16}$$

考虑到浅层特征与深度特征在数据结构上可能存在的差异性，对 3 个模态的特征进行拼接。记忆向量 $\boldsymbol{S}_i^{(k)}$ 作为一种跨模态联合表征，在注意力机制中扮演着 *Query* 的角色，指导着各个模态信息的注意力。需要注意的是，TMCAM 的协同注意力网络能够进行迭代，以不断优化和提炼记忆变量，$k$ 代表迭代次数。当 $k=1$ 时，$\boldsymbol{m}_i^{(0)}$ 的初始化方式为：

$$\boldsymbol{m}_i^{(0)}=[\boldsymbol{R}_i^{(0)},\boldsymbol{I}_i^{(0)},\boldsymbol{S}_i^{(0)}] \tag{17}$$

其中，

$$\boldsymbol{R}_i^{(0)}=\frac{1}{T}\sum_{j=1}^{T}\widetilde{\boldsymbol{w}}_{ij} \tag{18}$$

$$\boldsymbol{I}_i^{(0)}=\frac{1}{Z}\sum_{j=1}^{Z}\widetilde{\boldsymbol{v}}_{ij} \tag{19}$$

$$\boldsymbol{S}_i^{(0)}=\widetilde{\boldsymbol{S}}_i \tag{20}$$

在第一次迭代中生成记忆向量 $\boldsymbol{m}_i^{(1)}$ 后，$\boldsymbol{m}_i^{(1)}$ 进入第二次迭代，再次与 $\widetilde{\boldsymbol{R}}_i$，$\widetilde{\boldsymbol{I}}_i$ 和 $\boldsymbol{SS}_i$ 计算注意力向量，生成新的记忆向量 $\boldsymbol{m}_i^{(2)}$，$\boldsymbol{m}_i^{(K)}$ 是最后一次迭代生成的记忆向量。记忆向量的生成方式借鉴了 ResNet 的残差结构思想，将浅层的输出结果 $\boldsymbol{m}_i^{(k-1)}$ 传入 $\boldsymbol{m}_i^{(k)}$ 的计算中，以缓解网络结构过于复杂导致的梯度消失问题。同时 $\boldsymbol{m}_i^{(k-1)}$ 作为一种历史知识，在 $\boldsymbol{m}_i^{(k)}$ 的生成中被继承，增加了网络结构的稳定性和鲁棒性。

## 3.4 输出模块

本文采用多层感知机 MLP 作为分类器。MLP 是一种非线性神经网络，本文设置了 3 层的 MLP 作为分类器，并使用了不同的非线性激活函数。

$$H_i=\boldsymbol{W}_{\mathrm{MLP}}^{3}(\mathrm{Tanh}(\boldsymbol{W}_{\mathrm{MLP}}^{2}(\mathrm{LeakyRelu}(\boldsymbol{W}_{\mathrm{MLP}}^{1}\boldsymbol{m}_i^{(K)})))) \tag{21}$$

训练的目标是在训练数据中最小化损失函数，即均方误差 MSE(Mean Squared Error)，用 $(H_i,y_i)_{i=1}^{\mathrm{batchsize}}$ 表示一组训练数据，$y_i$ 是真实标签，表示为：

$$MSE(error)=\frac{1}{n}\sum_{i=1}^{n}(H_i-y_i)^2 \tag{22}$$

### 3.5 模型的流程

综上所述，TMCAM 模型的流程如算法 1 所示。

**算法 1** TMCAM 算法

输入：评论文本 $Review_i$，评论照片 $Image_i$，浅层特征集 $S_i$

输出：预测的评论有用性 $H_i$

For epoch in range(epochs)：

  For data in datasets：

    $\mathbf{R}_i = \mathrm{BERT}_{Pre}(Review_i)$，$\widetilde{\mathbf{R}}_i = \mathbf{W}_u\mathbf{R}_i$

    $\mathbf{I}_i = \mathrm{ResNet}_{Pre}(Image_i)$，$\widetilde{\mathbf{I}}_i = \mathbf{W}_v\mathbf{I}_i$

    $\widetilde{\mathbf{S}}_i = \mathbf{W}_t\mathbf{S}_i$，$\mathbf{SS}_i = \mathrm{expand\_as}(\widetilde{\mathbf{S}}_i)$

    $\mathbf{R}_i^{(0)} = \frac{1}{T}\sum_{j=1}^{T}\widetilde{\mathbf{w}}_{ij}$，$\mathbf{I}_i^{(0)} = \frac{1}{Z}\sum_{j=1}^{Z}\widetilde{\mathbf{v}}_{ij}$，$\mathbf{S}_i^{(0)} = \widetilde{\mathbf{S}}_i$

    $\mathbf{m}_i^{(0)} = [\mathbf{R}_i^{(0)}, \mathbf{I}_i^{(0)}, \mathbf{S}_i^{(0)}]$

    For k in range(1，k_steps)：

      $\mathbf{R}_i^{(k)} = \mathrm{Attention\_Text}(\widetilde{\mathbf{w}}_{ij}, \mathbf{m}_i^{(k-1)})$

      $\mathbf{I}_i^{(k)} = \mathrm{Attention\_Image}(\widetilde{\mathbf{v}}_{ij}, \mathbf{m}_i^{(k-1)})$

      $\mathbf{S}_i^{(k)} = \mathrm{Attention\_Shallow}(\mathbf{ss}_{ij}, \mathbf{m}_i^{(k-1)})$

      $\mathbf{m}_i^{(k)} = \mathbf{m}_i^{(k-1)} + [\mathbf{R}_i^{(k)}, \mathbf{I}_i^{(k)}, \mathbf{S}_i^{(k)}]$

    $H_i = \mathrm{MLP}(\mathbf{m}_i^{(K)})$

  End For

  $\mathrm{MSE(error)} = \frac{1}{n}\sum_{i=1}^{n}(H_i - y_i)^2$

  Gradient Descent

End For

## 4 实验与分析

### 4.1 数据集与预处理

本文使用 Yelp.com 的餐饮类数据来训练和测试所提出的模型，共爬取 11 532 条数据，其中有 2 318 条数据包含评论照片，经过数据审查去除重复、为空值的数据，共有 2 315 条数据。可以直接爬取到的浅层特征有：评论评级、评论者粉丝数、专家标签、评论者发布照片数、评论者发布评论数。评论长度和易读性特征由评论文本计算产生。专家标签为“Elite”或“0”，将其处理为 0-1 变量。评论者粉丝数、评论者发布照片数、评论者发布评论数 3 个特征出现了严重的长尾现象，进行对数化以保持数据分布的合理性。最后对所有的浅层特征进行归一化。

本文使用每条评论的帮助性投票的总数作为标签进行模型训练和测试。针对异常高的投票数，通过设置上限来对其进行标准化，这个上限是所有非 0 帮助性投票的平均值。这与文献[8,26]的处理方式类似，旨在确保模型的平稳学习。

### 4.2 实验参数设置

本实验采用 MSE 作为损失函数，使用 Adam 优化器，批量大小设置为 128，初始学习率为 $3\times10^{-4}$，并使用了学习率衰减算法，衰减规律如下：

$$lr_{iteration} = 0.5^{\frac{iteration}{decay\_iterations}} \times lr_0 \tag{23}$$

其中，$decay\_iterations$ 设置为 500，$lr_0$ 即为初始学习率 $3\times10^{-4}$。考虑到数据量并不巨大，对实验数据按照 7∶3 分别设置为训练集和测试集，每次训练打乱训练集顺序。在处理评论文本时，使用谷歌预训练好的权重参数。默认设置 TMCAM 的参数 $k$ 为 2，ResNet 模块选择 ResNet-18。在模型的网络层中，Dropout 技术被使用，设置 Dropout 概率为 0.5。合理的神经网络权重初始化选择同样重要，鉴于 TMCAM 模型频繁使用 Relu 家族的激活函数，因此网络权重使用 Kaiming 进行初始化。

### 4.3 实验设置

#### 4.3.1 对比实验设计

为了检验 TMCAM 模型的有效性，本实验设计了基于注意力机制的多模态模型和经典的基线模型。

(1)SAA[24]：一种问题引导的图像注意力机制，通过跨模态信息叠加、卷积的方式生成融合信息，然后用融合信息与图像信息实现注意力机制。使用 LSTM 提取文本特征，使用 ResNet 提取图像特征。

(2)HieCoAtt[21]：一种协同注意力机制，并提出了一种 3 层架构探讨文本和图像区域之间的联系。使用 LSTM 提取文本特征，使用 ResNet 提取图像特征。

(3)DAN[23]：一种协同注意力机制，其中包含了对文本和图像的注意力，通过点乘的方式融合两个注意力输出的向量。使用 Bi-LSTM 提取文本特征，使用 ResNet 提取图像特征。

(4)BR：分别使用 BERT 和 ResNet 进行文本和图像的特征提取，随后进行特征的简单拼接，最后输入 MLP 分类器。

(5)BT：一种简单的基线模型，使用预训练 BERT 进行文本特征的提取，随后通过 MLP 分类器输出结果。

#### 4.3.2 消融实验设计

(1)消融实验一：多模态信息的互补

本消融实验旨在验证多模态信息的互补作用，因此需要考察常见单模态和组合模态对在线评论有用性预测的影响。实验设置如表 2 所列。

表 2 多模态信息互补的消融实验

Table 2 Ablation experiments on multimodal information complementarity

| | 模态数据 | 处理方式 |
|---|---|---|
| ITS | 评论文本、照片、浅层特征 | TMCAM |
| IT | 评论文本、评论照片 | 特征拼接+MLP |
| Text-only | 评论文本 | BERT+MLP |
| Image-only | 评论照片 | ResNet+MLP |
| Meta-only | 浅层特征 | MLP |

ITS 表示同时输入 3 种模态数据，同时使用本文提出的 TMCAM 模型进行训练。IT 表示同时输入评论文本和评论照片数据，只进行特征拼接，并输入 MLP。Text-only，Image-only 和 Meta-only 表示只使用单模态的数据。

(2)消融实验二：跨模态融合方式

本组消融实验在保持所有三模态数据都输入的情况下，对跨模态融合的方式展开实验，如表 3 所列。

表 3 针对不同跨模态融合方式的消融实验

Table 3 Ablation experiments for different cross-modal fusion methods

| | 模态数据 | 处理方式 |
|---|---|---|
| A(ITS) | 评论文本、照片、浅层特征 | TMCAM |
| C(ITS) | | 特征拼接+MLP |

A(ITS)表示使用 TMCAM 模型进行模态数据的融合，C(ITS)表示只对评论文本、评论照片和浅层特征进行拼接。

(3)消融实验三:迭代次数 $k$

本组消融实验对 TMCAM 模型中协同注意力网络的迭代次数对评论有用性的影响展开实验,结果如表 4 所列。

表 4 参数 $k$ 的消融实验

Table 4 Ablation experiments on parameter $k$

| | 模态数据 | 处理方式 |
|---|---|---|
| K1 | 评论文本、评论照片、浅层特征 | TMCAM |
| K2 | | |
| K3 | | |
| K4 | | |

K1 代表 TMCAM 模型的协同注意力网络层迭代一次,K2 代表迭代两次,以此类推。

### 4.4 对比实验及结果分析

对比实验旨在比较本文提出的 TMCAM 模型与相似技术领域的其他模型的绩效。实验结果如图 4 所示。柱状图中,本文探讨两类指标,即$MSE_m$ 和$MSE_{5ep\text{-}100ep}$,$MSE_m$ 代表模型的最佳表现,$MSE_{5ep\text{-}100ep}$ 是第 5—100 个 epoch 上 MSE 的均值,代表模型的平均性能和收敛情况。测试集收敛曲线为模型在测试集上每次 epoch 的 MSE 数值。


(a)$MSE_m$ (b)$MSE_{5ep\text{-}100ep}$

图 4 对比实验结果

Fig. 4 Comparative experimental results

图 4 列出了对比实验结果,图 5 为对应的模型 MSE 收敛图。通过与基线方法进行对比,可知 TMCAM 模型在两个指标上都具有最优的结果,即 TMCAM 模型优于其他基线方法。其次,基于注意力机制的模型比不使用注意力机制的 BR 和 BT 表现更好。此外,BR 和 DAN 具有较小的$MSE_m$ 值,但训练后期模型过拟合严重,而 TMCAM,SAA 和 HieCoAtt 具有更好的收敛效果和平稳性。这些对比实验表明融合了评论文本、照片信息和浅层特征的 TMCAM 模型具有更好的特征融合能力,达到了最好的预测效果,能够更好地对在线评论有用性预测问题进行建模。

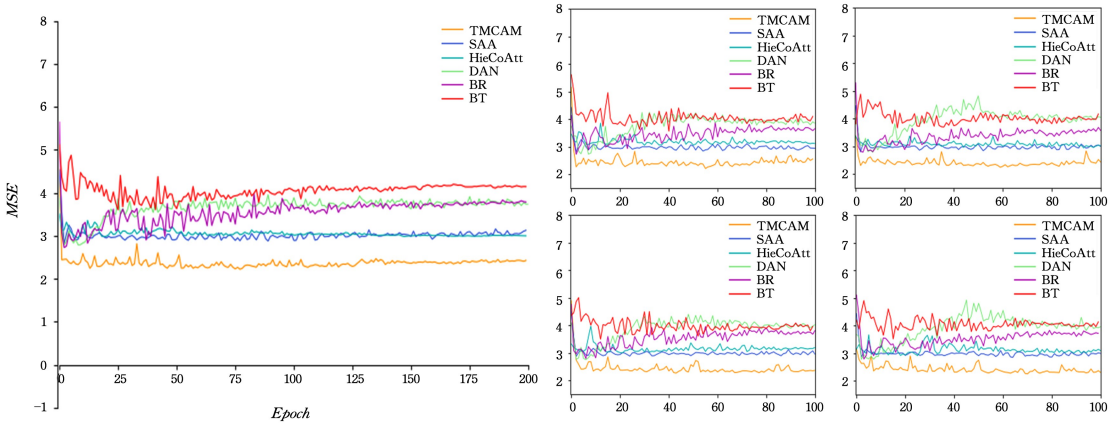



图 5 对比实验的测试集收敛曲线

Fig. 5 Convergence curve of test set for comparative experiment

### 4.5 消融实验及结果分析

#### 4.5.1 消融实验一:多模态信息的互补

消融实验一的结果如图 6 和图 7 所示。由图 6 和图 7 可知,同时输入 3 种模态信息的 ITS 具有最好的预测效果,同时输入评论文本和评论照片的 IT 模型表现次之。在单独输入单模态信息的实验中,只输入评论文本的 Text-only 具有最低的 MSE。


(a)$MSE_m$ (b)$MSE_{5ep\text{-}100ep}$

图 6 消融实验一的实验结果

Fig. 6 Results of ablation experiment 1

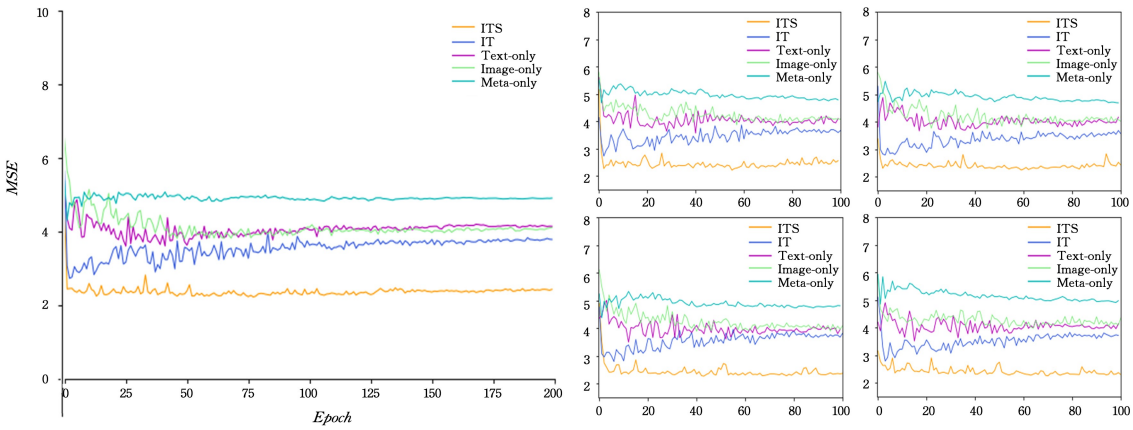


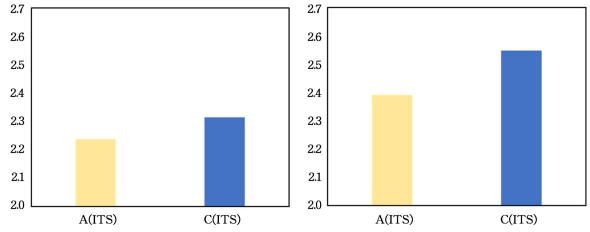

图 7 消融实验一的测试集收敛曲线

Fig. 7 Test set convergence curve of ablation experiment 1

实验结果证明了多模态信息之间的互补作用，评论照片和评论文本信息的互补能够达到更好的效果。其次，浅层特征能够辅助预测评论的有用性，其所包含的信息不能完全被评论文本、图像深层特征所取代。此外，同时考虑三模态的信息，实现跨模态信息交互能够达到最好的绩效。

4.5.2 消融实验二：跨模态融合方式

消融实验二是针对跨模态融合方式的不同进行对比。实验结果如图 8、图 9 所示。由图 8 和图 9 可知，使用 TMCAM 模型，即 A(ITS)，具有更小的 $MSE_{\min}$ 和 $MSE_{5ep\text{-}100ep}$，比 C(ITS)具有更好的跨模态特征融合能力，说明对评论文本、照片和浅层特征使用协同注意力机制能够有效提取和融合多模态的信息。C(ITS)将特征拼接作为特征融合的方式，在精度和收敛效果上都差于A(ITS)。实验结果证明，本文设计的协同注意力网络以及特征提取方式具有一定的科学性和合理性，符合在线评论有用性预测的内在逻辑。




(a) $MSE_m$ (b) $MSE_{5ep\text{-}100ep}$

图 8 消融实验二的实验结果

Fig. 8 Results of ablation experiment 2

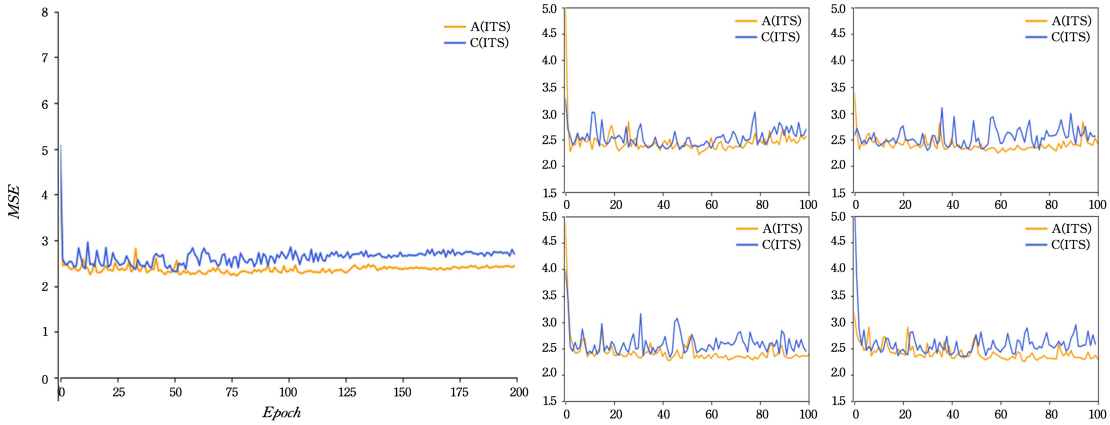


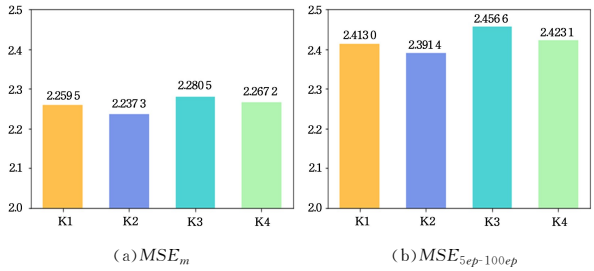

图 9 消融实验二的测试集收敛曲线

Fig. 9 Test set convergence curve of ablation experiment 2

4.5.3 消融实验三：迭代次数 $k$

本组实验围绕 TMCAM 模型的协同注意力网络迭代次数展开对比，实验结果如图 10 所示。

由图 10 可知，不同迭代次数的 TMCAM 模型都获得了较好的表现，这说明所设计的基于协同注意力机制的深度预测模型能够很好地预测在线评论的有用性，即便是最浅层的模型(即 $k=1$)，也能发挥较好的特征融合能力。虽然提升效果微弱，但也能够发现当迭代次数为 2 时，预测效果有微弱的提升。




(a) $MSE_m$ (b) $MSE_{5ep\text{-}100ep}$

图 10 消融实验三的实验结果

Fig. 10 Results of ablation experiment 3

**结束语** 为了更好地预测在线评论的有用性，本文提出了一种基于协同注意力机制的三模态评论有用性预测模型(TMCAM)，它利用 BERT 和 ResNet 预训练模型提取评论文本和评论照片的深度特征，基于 KAM 知识采纳理论构建浅层特征集合，3 种模态的特征通过一种可迭代的协同注意力机制来实现多模态特征的交互和提炼，最后输入多层感知机进行评论有用性的预测。对比实验结果证明了 TMCAM 较基线模型具有更好的预测效果；消融实验揭示了评论文本、照片和浅层特征的结合能够达到最好的预测效果，TMCAM 模型实现的三模态信息交互具有更好的特征融合能力。

在未来的研究中，我们将在更广泛、更优质的数据上进行实验。此外，受深度学习难以解释的限制，本文没有从严格的解释性视角去展示各模态数据是如何交互的，只是从结果的角度出发给出了一些普适性的结论，希望在未来的研究中对多模态信息的交互过程进行进一步探索。

## 参 考 文 献


[1] HONG H, XU D, WANG G A, et al. Understanding the determinants of online review helpfulness: A meta-analytic investigation[J]. Decision Support Systems, 2017, 102: 1-11.

[2] KARIMI S, WANG F. Online review helpfulness: Impact of reviewer profile image[J]. Decision Support Systems, 2017, 96: 39-48.

[3] DU J, RONG J, MICHALSKA S, et al. Feature selection for helpfulness prediction of online product reviews: An empirical study[J]. PloS One, 2019, 14(12): e0226902.

[4] MUDAMBI S M, SCHUFF D. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon. com[J]. MIS Quarterly, 2010, 34(1): 185-200.

[5] SALEHAN M, KIM D J. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics[J]. Decision Support Systems, 2016, 81: 30-40.

[6] CHATTERJEE S. Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach[J]. International Journal of Hospitality Management, 2020, 85: 102356.

[7] PRIETO A, PRIETO B, ORTIGOSA E M, et al. Neural networks: An overview of early research, current frameworks and new challenges[J]. Neurocomputing, 2016, 214: 242-268.

[8] SAUMYA S, SINGH J P, DWIVEDI Y K. Predicting the helpfulness score of online reviews using convolutional neural network[J]. Soft Computing, 2020, 24(15): 10989-11005.

[9] FAN M, FENG C, GUO L, et al. Product-Aware Helpfulness Prediction of Online Reviews[C]// The World Wide Web Conference. 2019.

[10] XU S, BARBOSA S E, HONG D. Bert feature based model for predicting the helpfulness scores of online customers reviews [C]// Future of Information and Communication Conference. Cham: Springer, 2020: 270-281.

[11] MA Y, XIANG Z, DU Q, et al. Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning[J]. International Journal of Hospitality Management, 2018, 71: 120-131.

[12] PAN Y, ZHANG J Q. Born unequal: a study of the helpfulness of user-generated product reviews[J]. Journal of retailing, 2011, 87(4): 598-612.

[13] LI M, HUANG L, TAN C H, et al. Helpfulness of online product reviews as seen by consumers: Source and content features [J]. International Journal of Electronic Commerce, 2013, 17(4): 101-136.

[14] LIU A X, LI Y, XU S X. Assessing the Unacquainted: Inferred Reviewer Personality and Review Helpfulness[J]. MIS Quarterly, 2021, 45(3): 1113-1148.

[15] REN G, HONG T. Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews[J]. Information Processing & Management, 2019, 56(4): 1425-1438.

[16] MALIK M S I, HUSSAIN A. An analysis of review content and reviewer variables that contribute to review helpfulness[J]. Information Processing & Management, 2018, 54(1): 88-104.

[17] LEE S, CHOEH J Y. The determinants of helpfulness of online reviews[J]. Behaviour & Information Technology, 2016, 35(10/11/12): 853-863.

[18] DU J, RONG J, WANG H, et al. Neighbor-aware review helpfulness prediction [J]. Decision Support Systems, 2021, 148: 113581.

[19] LI H. Deep learning for natural language processing: advantages and challenges[J]. National Science Review, 2018, 5(1): 24-26.

[20] BRAUWERS G, FRASINCAR F. A General Survey on Attention Mechanisms in Deep Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 3279-3298.

[21] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2016: 289-297.

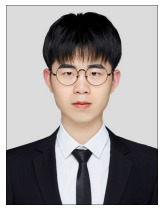
[22] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 21-29.

[23] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 299-307.

[24] KAZEMI V, ELQURSH A. Show, ask, attend, and answer: A strong baseline for visual question answering[J]. arXiv: 1704.03162, 2017.

[25] SUSSMAN S W, SIEGAL W S. Informational influence in organizations: An integrated approach to knowledge adoption[J]. Information Systems Research, 2003, 14(1): 47-65.

[26] SAUMYA S, SINGH J P, BAABDULLAH A M, et al. Ranking online consumer reviews[J]. Electronic Commerce Research and Applications, 2018, 29: 78-89.

**ZHANG Yian**, born in 1999, postgraduate. His main research interests include deep learning and user information behavior.

**WANG Gang**, born in 1980, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include information systems and machine learning.

(责任编辑：喻藜)