Operations Research and Fuzziology 运筹与模糊学, 2022, 12(2), 157-168
Published Online May 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.122015

Hans汉斯

基于深度学习的旅游评论有效一致性综合评价

李君涛[1]，欧阳智[2*]，杜逆索[2]

[1]贵州大学数学与统计学院，贵州 贵阳
[2]贵州大学贵州省大数据产业发展应用研究院，贵州 贵阳

收稿日期：2022年3月16日；录用日期：2022年4月28日；发布日期：2022年5月5日

摘 要

旅游平台上的旅游在线评论是消费者出行旅游的重要参考，然而评分不一致与无效评论问题会导致对旅游目的地的评价失真，干扰消费者出行决策。针对旅游目的地综合评价的问题，首先通过基于负样本生成的深度学习BERT二分类模型等方式清理无效评论数据，然后基于改进的BERT粗粒度情感得分与细粒度多维特征匹配的评分计算方法构建旅游目的地的综合评价模型。结果表明模型在实验数据集上取得了较低的误差，并且在考虑数据有效性和指标多样性的前提下能较好地拟合平台得分数据。因此，在保证训练数据客观的准确的前提下，该方法具有一定实用性和泛化性。

关键词

旅游评论，综合评价，深度学习，有效性，一致性

Research on Comprehensive Evaluation of Travel Review in Effectiveness and Consistency Based on Deep Learning

Juntao Li[1], Zhi Ouyang[2*], Nisuo Du[2]

[1]School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
[2]Guizhou Big Data Academy, Guizhou University, Guiyang Guizhou

Received: Mar. 16th, 2022; accepted: Apr. 28th, 2022; published: May 5th, 2022

*通讯作者。

文章引用: 李君涛, 欧阳智, 杜逆索. 基于深度学习的旅游评论有效一致性综合评价[J]. 运筹与模糊学, 2022, 12(2): 157-168. DOI: 10.12677/orf.2022.122015

Abstract

Online reviews on tourism platforms are important references for consumers' decision-making in travelling. However, the inconsistent ratings and invalid reviews may lead to distortion of the evaluation on travel destinations and interfere with consumers' travel decisions. For the comprehensive evaluation of tourist destinations, invalid comments are firstly cleaned up by using a BERT binary classification model based on negative samples generation. And then a comprehensive evaluation model of tourist destinations is constructed based on the score calculated by the BERT coarse-grained sentiment level and the fine-grained multi-dimensional feature matching score. The results show that the model has achieved extremely low errors on the experimental data set, and it can fit well the platform score data under the premise of the data validity and indicator diversity. Therefore, when the training data are sufficiently objective and accurate, this method has significant practicability and generalization.

Keywords

Tourism Review, Comprehensive Evaluation, Deep Learning, Effectiveness, Consistency

1. 引言

随着旅游行业的快速发展，旅游支出在中国消费市场所占比重日益增大。伴随着移动互联网在日常生活中越发深入，越来越多的消费者习惯于根据平台上的消费者评论决定旅游目的地。在此过程中，评论的有效性以及旅游目的地的综合考量成为了平台管理者和消费者密切关注的问题。失真的综合评价不但影响消费者的消费体验，也影响了旅游平台的信誉。

基于消费者评论的综合评价依赖于评价的有效性和一致性。有效性是指获取客观评价得分所依赖的输入信息是否存在无效、混乱和冗余以及是否受到非理性因素的影响。现有的理论表明，评论的有效性对于产品的综合评价至关重要，而评论有效性受多种因素影响，如消费者特征、产品特征、评论特征等[1]。目前在线评论的研究，为了评价的有效性主要是通过人工去除的方式，例如王乾等[2]通过人工清洗，删除无意义评论与矛盾性评论，这些方法虽然简单直接，但是对于大规模的在线评论数据，人工清除的方式不仅效率低，鉴定困难，也容易误删除有效信息，致使最终的综合评价产生偏差。

另一方面，一致性是指相同或相似度高的文本评论所对应的得分趋于一致。例如评论“这个酒店很不错”理应得到 4~5 分的评分，但消费者会因为非客观等因素的影响给出 3 分及以下的评分。现有研究在考虑不一致的矛盾评价时往往选择赋予更低的权重，以通过忽略文本信息来获取评分的一致性，这使得文本数据和评分数据的一致性问题没有得到很好解决。同时现有研究在构建评价体系时倾向于使用语义情感等一维特征[3]，而如何将产品的多维特征合理地融入评价体系仍较少考虑。

综上，本研究将针对评论的有效性和一致性，基于深度学习中的 BERT (Bidirectional Encoder Representations from Transformers)模型[4]，通过粗粒度情感得分构建基于网评文本的产品综合评价模型，以此实现旅游产品在线综合评价。在仅给出消费者文本评论的情况下，解决旅游目的地的自动化评分问题。

本文的结构安排如下：第一部分提出研究问题并介绍本文研究的目的和意义；第二部分阐述了国内外研究现状以及本文的研究创新点；第三部分介绍实验数据集并提出了数据集的预处理方法；第四部分是模型的主体结构，包括各模块数据的分析、处理和评价方法；第五部分，介绍了模型运行所需的实验环境，分析及评价模型的运行结果；最后总结了本文的研究结论以及未来展望。

2. 相关工作

消费者评论是一类重要的用户生成内容[5]，其包含情感描述等数据特征，可用于社会科学中研究消费者行为和动机，包括如何制定定价策略[6]，消费者认知与商品类型的关系[7]，消费者特征差异[8]等等。

近年来，基于消费者评论的产品综合评价方法大致可分为两类。一类是基于分词和词频统计的传统机器学习方法，例如，常怀文等[9]提出的基于 TF-IDF (Term Frequency-inverse Document Frequency)的 LDA (Local-density Approximation)模型，该模型通过 TF-IDF 算法将文本评论进行分类量化，后将量化后的文本评论与评分数据一同采用线性加权得到市场综合评价指标体系；Mohammadamir 等[3]提出的基于单词信息增益算法的 SVM (Support Vector Machine)分类模型，将词频特征和评分作为 SVM 分类模型的输入，得到产品的综合评价。这一类方法都是将文本数据分词后将关键词或情感词的词频作为模型的输入，结合回归、SVM 等传统机器学习模型，得到最终评价。这种方法的局限在于并未充分利用文本数据的有效信息，对语句情感的鉴定过于笼统。

另一类是基于自然语言处理(NLP)深度学习模型的方法。例如，王乾等[2]提出的基于 LSTM-AE 的商品综合评分模型，该模型使用 word2vec 模型生成词向量，并结合 LSTM-AE 网络提取评论的隐含特征，通过 SVM 对隐含特征进行情感分类获得文本评分，将其与数值评分加权求和最终得到综合评分。Cao 等[10]提出的基于语义提取和生成的综合评价模型，该模型将生成的语义与评论语义提取器提取的语义进行比较，将最终生成的语义特征输入评分回归器以预测整体评分。这一类方法使用深度神经网络的词嵌入模块，大大增加了语句特征的语义丰富度，在后续的评分模块中取得了较好的效果。然而，针对结合多指标的深度网络模型，现有研究在进行综合评价时对评分数据具有依赖性，同时由于消费者主观因素的影响，存在评分标准不一致的问题；而对于仅考虑文本情感得分的评价模型，其忽略了评论中不同指标的信息，使得评价模型不够客观全面。

另一方面，评论的有效性对于产品的综合评价至关重要。在给出消费者评论的情况下，殷国鹏等[11]发现评论长度与评论的有用性为正相关关系，郝媛媛等[12]发现评论内容的正负情感对评论的有用性存在显著正向影响；基于海量增长的消费者评论数据，毛郁欣[13]等提出了基于语义特征的评论有用性评价模型，以过滤出有价值的内容。马超等[14]通过利用机器学习方法，对旅游在线评论文本和图片进行分析，评估图片对于评论感知有效性的影响。在信息更充足的情况下，安静等[15]发现消费者的认知特征和行为特征均对在线评论有效性更是有着显著正向影响。然而在实际应用中，大部分基于消费者评论的模型在数据处理阶段都只使用了简单的剔除重复数据等操作，忽视了数据的有效性问题。

针对评论有效性问题以及结合多指标评价中对评分数据的依赖问题，本文提出了基于网评文本的深度学习产品在线综合评价模型。本文的主要贡献为：

1) 在数据预处理中，通过提出基于负样本生成的 BERT 分类模型提高评论的有效性。

2) 在初步建立评论与评分映射模型中，通过引入 Focal-Loss 损失函数解决中性评论和高情感特征评论样本损失计算不平衡的问题。

3) 在建立综合评价中，通过结合改进的 BERT 模型和下游细粒度匹配的方式，使模型能建立文本数据到评分数据的对应关系，从而有效处理对评分数据的依赖以及评分不一致问题，并且以自建关键词库考虑了评论数据的多维特征。

因此，本文模型通过引入深度学习的方法，来实现基于网评文本的产品在线有效一致综合评价，最终获得对旅游目的地多方面客观的评分，进而为平台对产品的评价定位和消费者的消费参考提供建议。

3. 基于评论有效性的数据预处理

在数据预处理阶段，主要通过逐级剔除的方式以得到有效评论数据。除了用相似度阈值剔除重复评论、通过语义丰富性和关键字匹配去除广告评论和“短义”评论等传统针对文本模型的预处理方式以外，针对可能存在的混乱评论，本文引入了深度学习方法，提出了基于负样本生成的 BERT 二分类模型。

3.1. 数据集介绍

本文的任务目标在于构建仅依赖文本评论的多指标产品综合评价模型，通过基于相关评论对景区和酒店进行综合评价，并结合网站得分数据验证模型的合理性。因此，本文所用实验数据通过数据采集方法获取，并将实验数据分为 2 类。第 1 类是“艺龙”旅游平台上获取的网评文本数据和得分数据，通过爬虫技术，对网站内 2016 年 1 月 1 日至 2020 年 12 月 31 日的评论数据和各项指标评分进行抓取。其中文本数据包含了 59106 条游客对 50 个景区的旅游评论以及 25,225 条游客对 50 家酒店的文本评价，得分数据包含了各个景区和酒店的服务、位置、设施、卫生、性价比各项指标评分。该类数据将运用于整体模型的验证以及测试。第 2 类数据同样通过爬虫技术，对“去哪儿”，“马蜂窝”等其他网站内 2018 年 1 月 1 日至 2020 年 12 月 31 日的评论数据和对应评分数据进行抓取，数据包含 66,464 条带有 1~5 分评价标签的评论数据和旅游目的地的得分数据。该类数据将用于 BERT 模型的训练，以及得分参数的调整。在 BERT 模型训练时，将该类数据集打乱顺序，按照 8:1:1 的比例，分别划分训练集、验证集和测试集。

3.2. 基于负样本生成的 BERT 分类模型

首先，对于旅游评论重复数据的甄别，主要采用基于 doc2bow [16]和 TF-IDF 模型[17]的文本相似度计算方法。这里，对于文本相似度高于阈值(0.95)的评论，认为其是重复的，为无效评论。接着，基于语义词数量的无义评论筛选，将评论不重复的字数总和小于等于 3 的评论进行了删除。然后，基于文本匹配的数据分类，通过关键字匹配的方式将广告评论进行剔除，完成基础的数据预处理。

虽然进行了上述的基础预处理，但是在数据中仍不难发现有一些评论是文字乱码的无意义状态。为了解决此类的无效评论，本文提出了基于负样本生成和 BERT 分类模型来剔除此类评论。在文本数据的初始化表示时 BERT 预训练模型能生成信息量丰富的字向量表示，对文本的信息表示要显著优于 CNN (Convolutional Neural Network)、LSTM (Long Short Term Memory)等深度学习模型，即 BERT 分类模型通过预训练模型结合下游微调的方式就能实现很好地文本分类任务，故本文使用 BERT 分类模型来剔除此类评论，负样本剔除模型过程如图 1 所示。

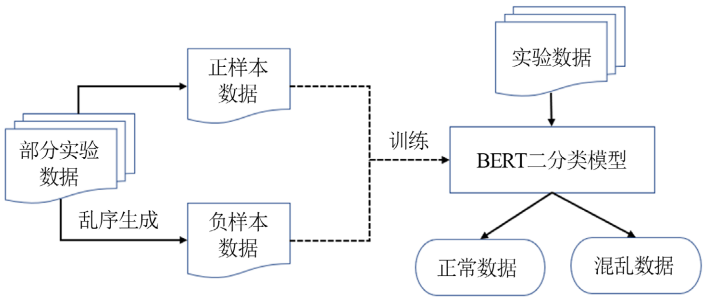

Figure 1. BERT classification model based on negative sample generation
图 1. 基于负样本生成的 BERT 分类模型

通过从训练数据中的景区和酒店评论中分别选取 4000 条正常语序的评论合并作为正面样本，将正面样本中的每条评论随机打乱顺序得到文字混乱的负面样本；随后对正面样本和负面样本分别标注各自的标签，将正面样本和负面样本合并后打乱，按 8:1:1 的比例生成训练集、验证集和测试集；接着将分好的数据输入 BERT 二分类模型中，训练得到最终用于甄别混乱评论的深度网络模型。最后将实验数据输入至上述训练好的深度网络模型中，得到最终的有效评论数据。通过上述模型，剔除的部分混乱数据如表 1 所示。

从表中可以看到，模型可以很准确地鉴别混乱评论，但对一些流行的“语气词”如“信耐”、“粉开心”等不具有鉴别能力，也将此类评论划归为混乱文本。不过由于此类评论数量不多且语义大多不丰富，因而不做进一步甄别，直接将此类数据剔除。

Table 1. Partially messed up data
表 1. 部分混乱数据

很便捷，值得信耐，非常满意
黑客；if 哪里呢；碧痕 JJ
哈哈哈拆 1 寂寞和红米哦哦
书宿舍呵呵呵属蛇呵呵呵 GG 额个
好过不哥特路口红 hold 路克拉默旅途
玩的粉开森……粉刺激……
诺特磕头哦啦了咯问喝了口虐了
还欧克，个人爱好啦啦啦啦
KKK 哦家吃饭：10K 了啦 KKK 咯弄 low 图我哦的五蕴皆空咯五
可口可乐了看看可口可乐了了拉拉
……

4. 主体模型与实验流程

针对基于消费者评论的产品综合评价问题，本文提出了结合 BERT 和下游规则匹配的产品综合评分模型，总体技术路线图如图 2 所示。

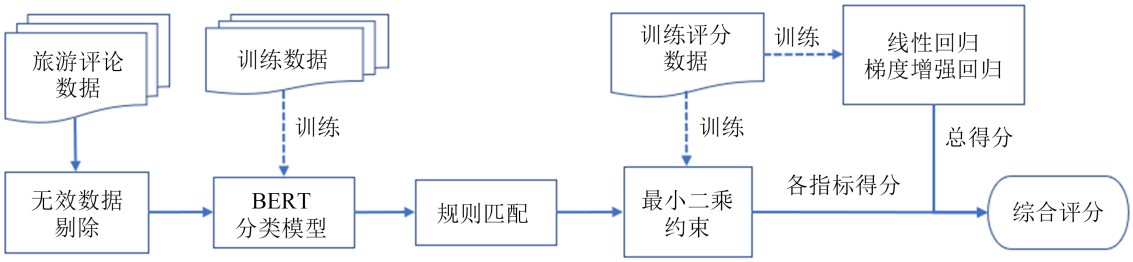

Figure 2. Comprehensive evaluation technology roadmap
图 2. 综合评价技术路线图

首先，针对于初始的旅游评论数据，采用相似度得分、规则匹配的方式对可能存在的各种重复、语义缺乏、广告等评论数据进行剔除。对于文本评论中的混乱数据，提出基于负样本生成和 BERT 分类模型的数据剔除方法。数据预处理之后，将文本数据输入至训练好的 BERT 情感分类模型中，得到各文本数据对应的基础情感得分。然后根据关键字匹配的方式得到各文本数据不同指标下的未调整得分，并通过最小二乘约束的方式得到各文本数据不同指标下的调整得分，最后采用回归的方式得到各旅游目的地的综合得分评价。

4.1. 基于改进的 BERT 模型的情感粗粒度得分

在综合评分模型中，首先将文本数据输入至改进的 BRET 分类模型中，模型结构如图 3 所示。

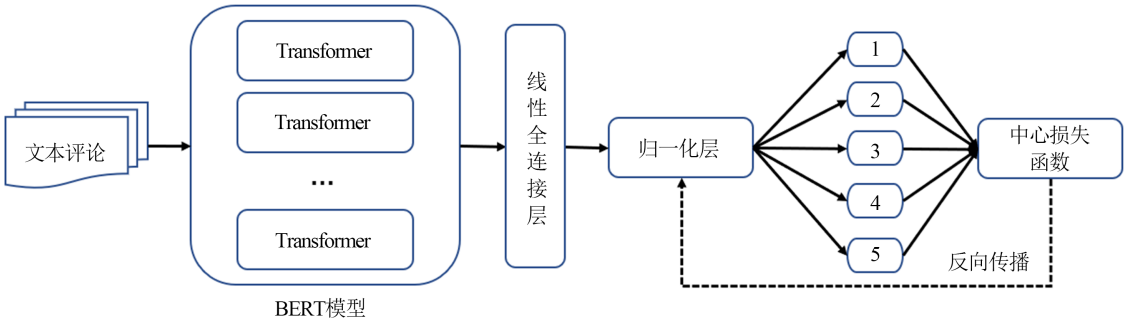

Figure 3. Improved BERT scoring model
图 3. 改进的 BERT 评分模型

文本评论通过 BERT 预训练模型编码为向量矩阵，随后通过两层线性全连接层和 SoftMax 层转化为各评分的预测概率，最后使用 Focal-Loss 损失函数计算损失并进行反向传播。通过对模型的训练，得到各评论的基础情感得分。与大部分提出的情感二分类模型不同，在基础得分阶段，模型将情感进行了五个等级的划分。由于情感程度分类任务划分越精细，任务难度则越高。因此，鉴于主观因素、边界模糊等因素，本文提出了结合粗粒度评价标准的 BERT 模型。具体而言，对于旅游目的地，不同的人关注点不同，感受不同，评分也大不相同；对于评论而言，游客的评论尺度也不同，比如“这地方挺不错的”这句话，有些人在评论完后会给出 5 分的评价，而有些人则会给出 4 分甚至 3 分。

针对这个问题，本文从训练过程和评价指标两个角度考虑处理此问题，在训练方面，对于评论数据，由于 1 分和 5 分这样的高评分对应的评论特征较为明显，模型能很轻易地预测准确，而对于 2 分至 4 分的评论数据，特征相对模糊，模型则较难训练，因而针对这种简单样本和困难样本混合的情况，本文引用了图像中目标检测领域的 focal loss 函数[18]，函数如式(1)所示，通过平衡简单和困难样本的损失权重，在预测阶段得到更高的准确率。

$$Focal\ Loss=\begin{cases}-\alpha(1-p)^{\gamma}\log(p), & \text{if } y=1\\ -(1-\alpha)p^{\gamma}\log(1-p), & \text{if } y=0\end{cases} \tag{1}$$

其中，α,γ 为超参数；y 为当前类别的标签标志，是当前类别取 1，否则取 0；p 为对应分类标签 y 的预测概率。

在评价指标方面，对于此类无固定评分标志的程度分类问题，提出了粗粒度准确率、召回率和 F1 值，作为模型好坏的评价标准，认为对于评分为 s 的评论，若模型能大概率将其预测至 $[s-1,s+1]$ 之间，便认为模型是可行的。

粗粒度精确率：

$$\widetilde{precision}=\frac{\widetilde{AP}}{\widetilde{AP}+\widetilde{FP}} \tag{2}$$

粗粒度召回率：

$$\widetilde{recall}=\frac{\widetilde{TP}}{\widetilde{TP}+\widetilde{FN}} \tag{3}$$

粗粒度 F1 值：

$$\widetilde{F1}=\frac{2*\widetilde{precision}*\widetilde{recall}}{\widetilde{precision}+\widetilde{recall}} \tag{4}$$

粗粒度 F-Score：

$$\widetilde{F_{score}}=\frac{1}{n}\sum_{i-1}^{n}\widetilde{F1}_i \tag{5}$$

其中，$\widetilde{AP}$：预测结果为正，样本近似为正；$\widetilde{TP}$：样本为正，预测结果近似为正；$\widetilde{FP}$：样本为负，预测结果近似为正；$\widetilde{FN}$：样本为正，预测结果近似为负。

改进的 BERT 分类模型训练完成后，输入测试集，得到的测试数据的混淆矩阵 M 如下所示：

$$M=\begin{bmatrix}173 & 11 & 25 & 5 & 1\\ 9 & 214 & 33 & 1 & 1\\ 8 & 18 & 1002 & 158 & 57\\ 8 & 8 & 244 & 747 & 692\\ 10 & 13 & 187 & 350 & 2722\end{bmatrix} \tag{6}$$

混淆矩阵行表示实际标签，列表示预测标签。从混淆矩阵可以计算出，使用 Focal loss 函数的 F-score 能达到 73.63%，相较于使用传统方式的交叉熵损失函数提升了 3.73%。对比不同损失下的粗粒度 F1 值的结果如图 4 所示：

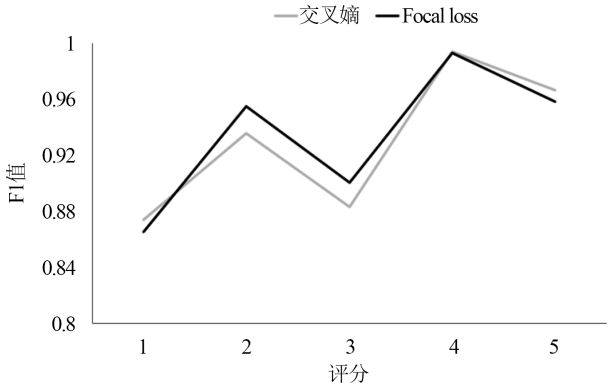

Figure 4. Coarse-grained F1 values for each score under different loss functions
图 4. 不同损失函数下各评分的粗粒度 F1 值

从图 4 中可以看出对于 1 分和 5 分这样的易分类样本，使用 Focal loss 损失函数的模型粗粒度 F1 值虽然略低于使用交叉熵损失函数的传统模型，但在中间评分这样的困难分类样本上取得了更好的效果。即模型的泛化性有所提升，而不是仅仅针对易分类样本。并且结合在 F-score 作为评价方式时的表现情况，认为在此文本分类问题上，使用 Focal loss 函数要优于交叉熵损失函数。从总体来看，在使用 Focal loss 函数的情况下模型预测结果 1 分的粗粒度 F1 值能达到 86%左右，而其他评分的粗粒度 F1 值均能达到 90%以上，故认为用该模型分类后的预测评分可以用于后续模型。

4.2. 基于 BERT 分类结果的关键字匹配

在下游微调阶段，针对服务、位置等不同指标的得分，建立了不同的敏感词与之对应，敏感词库可

以根据平台和产品类型做相应的添加和修改，增加了模型的可拓展性。通过匹配敏感词，得到各指标得分。

上述基于粗粒度的评价标准，认为改进的 BERT 分类模型的训练结果是有效的，将实验数据中的网评文本数据输入至训练好的 BERT 模型中，得到每条评论的基本得分。之后基于关键字匹配的方式得到服务、位置等各项指标的未调整得分。匹配规则如下：首先对于“服务”、“位置”、“设施”、“卫生”、“性价比”这五个方面，分别设立其对应的敏感词(包括名词和形容词)，敏感词对应褒义和贬义两部分。以服务为例，设定的部分高得分和低分敏感词如表 2 所示。

Table 2. Sensitive words related to services
表 2. 服务相关敏感词

名词：
服务　体验　感受　感觉　工作人员　前台　态度
形容词(褒义)
宾至如归　笑容可掬　精益求精　体贴入微　关怀备至 非常到位　特别舒服　无微不至　不厌其烦　很用心 很到位　太好了　超级好　百分百　特别好　非常好 特别棒　有耐心　高质量　很舒服　很不错　热情　很高 满意　贴心　满分　周到　极好　很好　超好　细心 仔细　很棒
形容词(贬义)
差　低　不　没有　黑　冷淡

随后对于每条评论中的每一个句子，先将“但是”、“不过”、“然而”等转折词替换成逗号；之后，将评论按标点符号分开成独立的句子；对于每一个句子，若句子中同时出现了敏感词的某个名词和某一个形容词，则该评论在对应指标下的基本得分加上或减去一个常数，获得每条评论各项指标下的一致评分，最后将同一旅游目的地中所有评论的各项指标平均，得到各个景区或酒店各项指标未调整的得分。

4.3. 基于最小二乘法的参数调整

由于旅游平台本身对于不同指标有着权重的设定需求，需要对不同指标赋予不同的权重。例如对景区而言，平台认为更应该看重位置这样的自然属性等要求。故有必要对得到的各个景区或酒店各项指标未调整的得分做进一步调整。结合训练数据集得分数据中的景区和酒店的服务、位置等各项指标评分，运用最小二乘法对各项指标评分进行调整。设 $k_i\left(i\in[1,5]\right)$ 为各项指标的调整系数，i取 1~5 分别对应“服务”、“位置”、“设施”、“卫生”、“性价比”这五个方面。

目标函数：

$$\min f=\sum_{j=1}^{m}\left(k_i\cdot init_{ij}-label_{ij}\right)^2 \tag{6}$$

其中 m 为景区或酒店总数；$init_{ij}$ 为关键字匹配模型得到的各项指标未调整的得分；$label_{ij}$ 为训练数据给出的旅游目的地各项指标评分。

解目标函数得：

$$k_i=\frac{\sum_{j=1}^{m}label_{ij}\times init_{ij}}{\sum_{j=1}^{m}init_{ij}^2} \tag{7}$$

将数据代入上述公式，得景区及酒店各项指标调整参数如表 3 所示。

Table 3. Adjustment parameters of various indicators of tourist destination and regression coefficient of total score
表 3. 景区及酒店的各项指标调整参数及总得分回归系数

指标	景区参数	酒店参数	景区系数	酒店系数
服务	0.884	0.954	0.3	0.208
位置	1.038	0.971	0.1	0.183
设施	0.992	0.987	0.15	0.171
卫生	1.041	0.999	0.3	0.289
性价比	0.981	0.853	0.15	0.158

4.4. 基于回归模型的总得分计算

在总得分计算中，本文采用的是多元线性回归的方式。根据训练数据中的得分数据可以发现，总得分与各项指标之间并非简单的平均值的关系。通过分析发现景区数据是单纯的线性关系，普通线性回归 R^2 达到了 1.0，即在线性回归模型中各指标的预测值完全拟合了总得分数据。从表 3 景区系数可以看出总得分相当于各项指标得分按 6:2:3:6:3 的权重加权比例得来。而酒店综合得分并不是完全线性的各指标加权组合的形式，从表 3 酒店系数可以看出各指标系数在 0.2 上下波动，最终通过回归的方式得到了基于文本输入的景区及酒店各项指标评分和总得分。

5. 实验与结果分析

5.1. 实验环境

模型代码的运行环境为 python3.6，深度网络框架为 torch1.4.0，预训练模型参数为 bert-base。在超参数设置方面，参考了 Sun 等人[19]在 BERT 上的文本分类的经验，如下设置超参数：学习率 lr = 2e−5，衰变因子 $\xi = 0.95$，批次大小 bsz = 64。对于 Focal loss 损失函数，超参数 $\alpha = 0.25$，$\gamma = 2$。为了分析 BERT 模型的损失变化，进行了多次实验，其结果如图 5 所示，当 BERT 模型在训练第 4 轮(epoch)时，在验证集上的损失开始上升。此外，训练遵循早停(early stopping)原则，当模型的损失在验证集上不再下降，就视为模型在验证集上已经收敛，可以停止训练。这能够有效地避免过拟合(Overfitting)问题，保证模型的泛化能力以及在测试集上的表现。因此将训练轮次设为 3 即可。

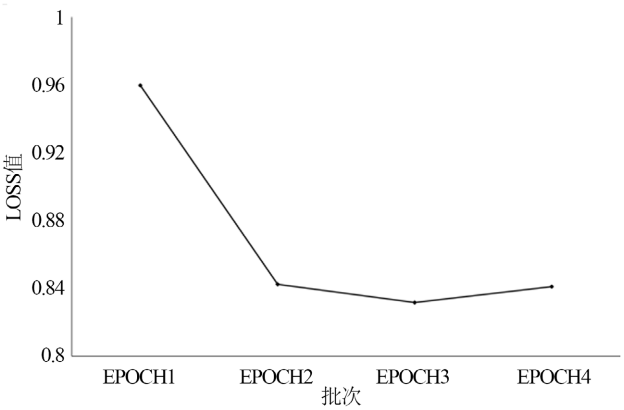

Figure 5. Average loss change graph of BERT model on validation set
图 5. BERT 模型在验证集上的平均损失变化图

5.2. 模型评价

基于上述实验流程得到景区及酒店各项指标评分和总得分，酒店部分结果如表 4 所示。

Table 4. Hotel index score and total score
表 4. 酒店各项指标评分和总评分

酒店名称	服务得分		位置得分		设施得分		卫生得分		性价比得分		总得分	
	预测	实际	预测	实际	预测	实际	预测	实际	预测	实际	预测	实际
H01	4.8	4.8	5	4.8	4.8	4.7	4.8	4.8	4.1	4.0	4.8	4.8
H02	4.9	4.7	5	4.8	4.8	4.6	4.9	4.7	4.2	4.0	4.9	4.7
H03	5	4.8	5	4.8	4.9	4.8	4.9	4.9	4.2	4.0	4.9	4.8
H04	4.9	4.9	5	4.9	4.9	4.9	4.9	4.9	4.2	4.0	4.9	4.9
H05	4.8	4.7	4.9	4.8	4.8	4.7	4.8	4.8	4.1	4.0	4.8	4.7
…												
H50	4.3	4.6	4.5	4.6	4.4	4.4	4.4	4.6	3.8	4.1	4.4	4.5

从表 4 可以看出，模型得到的基于酒店网络评论的各指标评分和总评分和实际值相差很小，在保证训练数据客观准确的情况下，模型结果能较准确地反映酒店的真实评价。

结合实验数据集中的得分数据，采用均方误差(Mean Squared Error, MSE)的方式对模型进行评价，酒店各指标及总均方误差如图 6 所示。首先，为了验证无效评论对最终的综合得分的影响，将原始数据代入综合评价模型，与去除无效评论的结果相比，从图 6 中可以发现，在景区数据集和酒店数据集上，去除无效评论后对结果都具有正向影响，获得的模型综合评分能更贴近平台数据的特点，方差值更低。另一方面，从图 6 可以看到，对于景区数据，景区总得分的拟合效果最好，而对于设施得分的拟合效果是最差为 0.114。各项指标的均方误差在 0.09 上下浮动，即模型对景区各个指标的评分平均有 0.3 左右的误差。但单就总得分而言，评分误差能缩小到 0.2。而对酒店数据而言，综合得分误差仅为 0.018，即模型能较好地拟合平台酒店综合得分数据的特点。

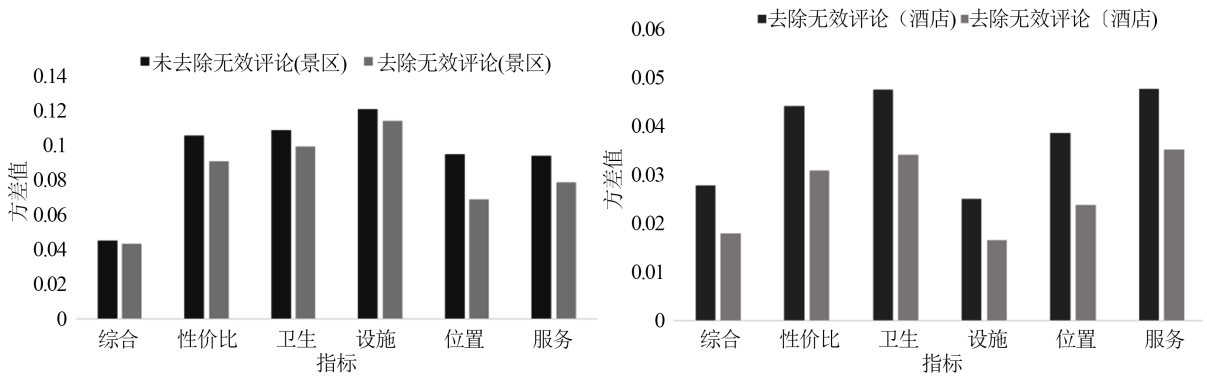

Figure 6. Variance comparison chart of tourist destination data with or without invalid comments
图 6. 景区和酒店有无效评论方差对比图

从结果对比来看，模型对酒店数据的拟合效果优于景区数据。原因可能有两方面，一是景区评论有较多的长文本，上游 BERT 分类模型由于长文本语义的丰富无法精确地捕捉情感语义，因此在生成每条评论的基础得分时产生了偏差。二是各项指标所包含内容的界定，对于酒店数据而言，服务、设施、卫

生等都有较为明确的界定，但对于景区评论而言，指标界定就显得相对模糊。比如“游乐园”本身可以是设施得分的一部分，尽管在评论时有关设施的敏感词并未被提及，但在评价时它会被一部分游客默认为高的设施得分。

需要指出的是，这里测试采用的评价标准是基于平台提供的综合得分数据集。由于缺乏客观的第三方评分，平台综合得分的历史记录对于验证模型的有效性是一个可靠的代理。本文模型的目的是通过改进的 BERT 粗粒度得分，使得相同、相似、同语义评论能获得相同的评分，解决过去评论得分的不一致问题。因此，通过从以上的分析，在提供的训练数据客观合理的前提下，结合模型的预处理方法和下游特征匹配，本文模型能够在拟合综合评分规则体系的同时，保障数据的有效性、评分的一致性和指标的多样性。

6. 总结与展望

为实现基于网评文本的产品在线综合评分，解决评分过程中存在的数据有效性、评分一致性和指标多样性问题，本文将深度学习方法引入到在线评论分析应用上，提出了基于 BERT 模型和规则匹配构建了有效一致的产品综合评价模型。首先，提出了基于负样本生成的 BERT 二分类模型来去除混乱评论，并从多个角度考虑了评论数据的有效性。在综合评分模型中，提出了粗粒度评价标准用于鉴定改进的 BERT 模型的效果，通过引入 Focal loss 函数增强中间评分的预测准确率。然后，在下游微调阶段，针对服务、位置等不同指标的得分，建立了不同的敏感词与之对应，敏感词库可以根据平台和产品类型做相应的添加和修改，增加了模型的可拓展性。实验结果表明，模型在考虑评分一致性、数据有效性和指标多样性的前提下能较好地拟合平台得分数据。因此，在实际中，本模型具有一定的实用性，训练好之后可以为平台直接即时计算出产品对应的评价综合评分。同时，除计算产品评价综合评分外，该模型还可以用于评论异常检测、文本评论监控等方面。

在未来工作中，针对预处理过程中分类模型并不能有效地鉴别“信耐”、“粉开心”等语气表义词的问题，后续将尝试改进训练数据集，使模型达到更好地鉴别混乱文本的效果。针对模型结果在景区和酒店数据集上表现出的差异，未来将继续探索如何更好地抽取长文本的隐含语义，以生成更加客观准确的旅游目的地评分。

基金项目

贵州省科学技术厅重大科技计划项目(黔科合重大专项字[2018]3002)；贵州大学培育项目(贵大培育[2020]41 号)。

参考文献

[1] 吴佳炫, 李胜利. 在线评论有用性的影响因素实证研究——以“去哪儿网”评论数据为例[J]. 文献与数据学报, 2021, 3(2): 65-76.

[2] 王乾, 傅魁. 基于 LSTM-AE 神经网络的商品评价综合评分计算方法研究[J]. 北京邮电大学学报, 2018, 20(4): 19-27.

[3] Kavousi, M. and Saadatmand, S. (2018) Estimating the Rating of Reviewers Based on the Text. In: Nagabhushan, P., Guru, D., Shekar, B. and Kumar, Y., Eds., *Data Analytics and Learning*, Springer, Singapore, 257-267.

[4] Jacob, D., Chang, M., Lee, K., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Annual Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, 2-7 June 2019, 4171-4186.

[5] 徐勇, 武雅利, 李东勤. 用户生成内容研究进展综述[J]. 现代情报, 2018, 38(11): 130-135+144.

[6] 孙燕红, 赵骞, 王子涵. 基于消费者评论的网络预售定价策略研究[J]. 中国管理科学, 2020, 28(11): 184-191.

[7] 刁雅静, 何有世, 王念新. 商品类型对消费者评论认知的影响: 基于眼动实验[J]. 管理科学, 2017, 30(5): 3-16.

[8] 张海彬. 多媒体消费者评论的信任与传播的性别差异研究[J]. 学术论坛, 2016, 38(2): 123-127.

[9] 常怀文, 陈浩文, 张宜. 消费者视角下电商发展前景的挖掘与分析——以亚马逊市场为例[J]. 特区经济, 2020(5): 89-94.

[10] Cao, R., Zhang, X. and Wang, H. (2020) A Review Semantics Based Model for Rating Prediction. *IEEE Access*, **8**, 4714-4723. https://doi.org/10.1109/ACCESS.2019.2962075

[11] 殷国鹏, 刘雯雯, 祝珊. 网络社区在线评论有用性影响模型研究——基于信息采纳与社会网络视角[J]. 图书情报工作, 2012, 56(16): 140-147.

[12] 郝媛媛, 叶强, 李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报, 2010, 13(8): 78-88+96.

[13] 毛郁欣, 朱旭东. 面向 B2C 电商网站的消费者评论有用性评价模型研究[J]. 现代情报, 2019, 39(8): 120-131.

[14] 马超, 李纲, 陈思菁, 等. 基于多模态数据语义融合的旅游在线评论有用性识别研究[J]. 情报学报, 2020, 39(2): 199-207.

[15] 安静, 郑荣, 杨明中. 消费者个体特征对在线评论有效性的影响研究[J]. 现代情报, 2017, 37(1): 106-111.

[16] Řehůřek, R. and Sojka, P. (2010) Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC* 2010 *Workshop on New Challenges for NLP Frameworks*, Valletta, 22 May 2010, 45-50.

[17] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-864.

[18] Lin, T., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324

[19] Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019) How to Fine-Tune BERT for Text Classification? 18*th China National Conference on Computational Linguistics* (*CCL* 2019), Kunming, 18-20 October 2019, 194-206. https://doi.org/10.1007/978-3-030-32381-3_16