文章编号:1008-1542(2021)04-0380-09

融入情感信息词向量的评论文本情感分析方法

吕妹园,张永健,张永强,孙胜娟

(河北工程大学信息与电气工程学院,河北邯郸 056107)

摘 要:为了解决分布式词表示方法因忽略词语情感信息导致情感分类准确率较低的问题,提出了一种融入情感信息加权词向量的情感分析改进方法。依据专属领域情感词典构建方法,结合词典和语义规则,将情感信息融入到 TF-IDF 算法中,利用 Word2vec 模型得到加权词向量表示方法,并运用此方法对采集到的河北省旅游景点的评论文本与对照组进行对比实验。结果表明,与基于分布式词向量表示的情感分析方法相比,采用融入情感信息加权词向量的改进方法进行情感分析,积极文本的准确率提高了 6.1%,召回率提高了 6.6%,F 值达到了 90.3%;消极评论文本的准确率提高了 6.0%,召回率提高了 7.2%,F 值达到了 89.6%。因此,融入情感信息加权词向量的情感分析改进方法可以有效提高评论文本情感分析的准确率,为用户获得更为准确的评论观点提供参考。

关键词:自然语言处理;语义规则;情感信息;TF-IDF;Word2vec;加权词向量;情感分析

中图分类号:TP391.1 **文献标识码**:A **doi**:10.7535/hbkd.2021yx04008

Sentiment analysis method of comment text based on word vector with sentiment information

LYU Meiyuan,ZHANG Yongjian,ZHANG Yongqiang,SUN Shengjuan

(School of Information and Electrical Engineering, Hebei University of Engineering, Handan, Hebei 056107, China)

Abstract: In order to solve the problem of low accuracy of sentiment classification caused by neglecting the sentiment information of words in distributed word representation method, an improved sentiment analysis method incorporating weighted word vectors of sentiment information was proposed. According to the exclusive domain sentiment dictionary, combined with the dictionary and semantic rules, the sentiment information is integrated into the TF-IDF algorithm, and the weighted word vector representation method is obtained by using word2vec model. The method is used to compare the collected comments of tourist attractions in Hebei Province with the control group. The results show that compared with the sentiment analysis method based on distributed word vector representation, the accuracy and recall rate of positive text are increased by 6.1% and 6.6%, and the F value reached 90.3%, the accuracy and recall rate of negative text are increased by 6.0% and 7.2%, and the F value reached 89.6% by using the improved method of sentiment analysis integrated with sentiment information weigh-

收稿日期:2021-03-25;修回日期:2021-06-11;责任编辑:王淑霞

基金项目:河北省创新能力提升计划项目(19456003D)

第一作者简介:吕妹园(1996—),女,山东济南人,硕士研究生,主要从事自然语言处理方面的研究。

通讯作者:张永强教授。E-mail:120030009@qq.com

吕妹园,张永健,张永强,等.融入情感信息词向量的评论文本情感分析方法[J].河北科技大学学报,2021,42(4):380-388.

LYU Meiyuan,ZHANG Yongjian,ZHANG Yongqiang, et al.Sentiment analysis method of comment text based on word vector with sentiment information[J].Journal of Hebei University of Science and Technology,2021,42(4):380-388.

ted word vector. Therefore, the improved method of sentiment analysis integrated with sentiment information weighted word vector can effectively improve the accuracy of sentiment analysis of comment text, and provide valuable reference for users to obtain more accurate comments.

Keywords: natural language processing; semantic rules; sentiment information; TF-IDF; Word2vec; weighted word vector; sentiment analysis

随着互联网的发展，越来越多的互联网用户开始在线上发表自己的观点，如淘宝、携程网等平台上用户对商品和景点的评论，情感分析技术可以让用户更便捷地获取评论的情感倾向。情感分析的主要任务是对评论语料的情感倾向性进行分析，本质上是一种文本分类[1]，即对用户的评论文本进行积极、消极的情感倾向的分类。

最早应用于情感分析的方法是基于情感词典[2-3]的方法。该方法的核心是利用情感词典遍历匹配旅客评论文本中的情感词，并根据语义规则计算评论文本的情感倾向。文献[4]—文献[5]介绍了基于情感词典的代表研究。基于情感词典的旅游文本情感分析不需要提前对文本进行标注，操作简单易于实现，但其太过于依赖情感词典且目前大多数情感词典不是专属领域情感词典，导致情感分类的准确率较低。基于机器学习情感分析方法[6-9]最早是由 PANG 等[10]提出，使用最大熵算法和 SVM 算法进行电影评论的情感分析。CHEN 等[11]针对在线旅游情感分类准确率低的问题，把情感分类任务转变成机器学习中的多分类问题，设计了基于知识图谱的关键词扩展方法，增加了短文本的特征数量，利用机器学习构建情感分类模型进行情感分析。VALDIVIA 等[12]发现 TripAdvisor 中许多用户的星级打分和评论文本的情感极性是不同的，研究了用户情绪与自动情绪检测算法之间的匹配，利用机器学习模型识别负面意见并发现了负面评价背后的原因。YU 等[13]为了对日本旅游网站 4Traval 景点的评论进行情感分析，提出了 3 组基于统计的特征选择函数和传统的 TF-IDF 方法且制定了 7 组基于规则的方法。结果证明，特征选择函数与权重结合能够较好地提高算法的整体性能。YANG 等[14]提出了以情感词典为基础，结合卷积神经网络(CNN)和基于注意力的双向门控回归单元(BiGRU)模型(SICABG)，SICABG 模型结合了情感词典和深度学习技术的优点，克服了现有产品评论情感分析模型的不足。

在基于机器学习的情感分析研究中，一般采用分布式词向量表示方式，分布式的表示方式只考虑词语的语义信息，忽视了词语的情感信息，而在情感分析研究中，一个词语所包含的情感倾向性信息非常重要。本文结合语义规则，利用情感词典将情感信息融入到 TF-IDF 算法进行加权词向量计算，然后利用 SVM 算法进行情感分析。由于同一个情感词在不同领域文本中的情感倾向是不同的，因此研究建立一种情感种子词的筛选标准，并利用 SO-PMI 算法构建专属领域情感词典，避免发生不包含情感信息的特征词影响情感分析的准确率等问题。

1　融入情感信息的加权词向量表示

1.1　Word2vec 词向量表示

在情感分析任务中，将词语表示成低维、非稀疏的向量是关键。目前，词向量表示方法主要有 one-hot 方法和分布式词向量表示方法。one-hot 方法中词向量的维数是由词典中词语的个数决定的。该方法的缺点是如果词典的词语数目过多会导致词向量的维数过大并且向量稀疏，另外该方法还忽视了词语之间的语义关联性。分布式词向量表示方法可以把词语表示成低维向量，将所有的词向量构成一个词向量空间，并通过计算词向量的距离来判断词语的语义相似度。

研究采用分布式方法中的 Word2vec 算法训练词向量，Word2vec 算法中包括 2 种词向量训练模型：CBOW 模型和 Skip-Gram 模型[15]，Skip-Gram 模型的训练准确度更好，模型如图 1 所示。

由图 1 可知，在 Skip-Gram 模型中输入中心词语 $W(t)$ 的 one-hot 编码来预测中心词的上下文词语 $W(t-k),\cdots,W(t-1),W(t+1),\cdots,W(t+k)$ 的概率模型。其中 $W(t)$ 表示当前句子中位置为 t 的词语，k 表示与 $W(t)$ 相邻的上下文的窗口。

1.2 传统 TF-IDF 特征权重算法

TF-IDF 算法是文本分类中常用的特征权重的计算方法，该方法考虑了词语在文档中的分布情况，可以衡量词语对文本分类的重要度。

传统的 TF-IDF 公式如式(1)所示：

$$W_{ij}=tf_{ij}\times\log\left(\frac{N}{M_j}\right)。\tag{1}$$

式中：W_{ij} 表示评论文本 T_i 中词语 h_{ij} 的权重值；tf_{ij} 表示词语 h_{ij} 在评论文本中的词频；N 表示评论文本数量；M_j 表示所有评论文本中出现词语 h_{ij} 的评论文本的数量。将式(1)归一化得到式(2)：

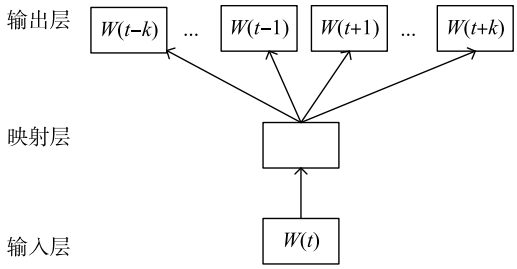

图 1 Skip-Gram 模型

Fig.1 Skip-Gram model

$$W_{ij}=\frac{tf_{ij}\times\log\left(\frac{N}{M_j}\right)}{\sqrt{\sum_{h_{ij}\in T_i}\left(tf_{ij}\times\log\left(\frac{N}{M_j}\right)\right)^2}}。\tag{2}$$

式中：h_{ij} 表示评论文本 T_i 中的第 i 个特征词。

1.3 融入情感信息的加权词向量

通过将评论文本与情感词典、程度副词词典相匹配，并结合语义规则将情感信息融入到特征权重的计算过程中。

情感词在不同的修饰词修饰下对文本情感倾向的贡献是不同的，情感词的修饰规则如下。规则 1：由程度副词修饰情感词时，句中存在关系(adv,STW)，则 $S_i=D_{i+m}\times S_i$。规则 2：否定词修饰情感词时，句中存在关系(negative,STW)，如“不高兴”，情感词“高兴”被否定词“不”修饰后由积极情感倾向变成了消极情感倾向，因此 $S_i=-1\times S_i$。规则 3：情感词由否定词和程度副词共同修饰，句中存在 2 种关系：一种为(negative,adv,STW)，如“不太满意”，此时情感词的情感倾向不改变，但情感词对文本的情感倾向贡献会被减弱，并参考文献[3]得到 $S_i=0.5\times D_{i+m}\times S_i$；一种为(adv,negative,STW)，如“太不满意”，此时情感词的情感倾向被改变，但情感词“满意”对文本的消极情感倾向的贡献程度由程度词决定，因此，$S_i=-1\times D_{i+m}\times S_i$。式中：$S_i$ 为情感词的情感极性值；D_{i+m} 为程度副词的程度极值；STW 表示情感词；negative 表示否定词；adv 表示程度副词，因此，融入情感信息的词语权重计算分 4 种情况。

第 1 种　词语 h_{ij} 为非情感词

此种情况下，词语 h_{ij} 的权重计算公式如式(3)所示：

$$W_{ij}=\frac{tf_{ij}\times\log\left(\frac{N}{M_j}\right)}{\sqrt{\sum_{h_{ij}\in T_i}\left(tf_{ij}\times\log\left(\frac{N}{M_j}\right)\right)^2}}。\tag{3}$$

第 2 种　词语 h_{ij} 为情感词且无修饰词修饰

此种情况下，词语 h_{ij} 的权重计算公式如式(4)所示：

$$W_{ij}=\frac{\left|tf_{ij}\times\log\left(\frac{N}{M_j}\right)\times S_j\right|}{\sqrt{\sum_{h_{ij}\in T_i}\left(tf_{ij}\times\log\left(\frac{N}{M_j}\right)\times S_j\right)^2}}。\tag{4}$$

式中：S_j 为情感词 h_{ij} 的情感极性值。

第 3 种　词语 h_{ij} 为情感词且满足修饰规则(adv,STW)，(negative,STW)，(adv,negative,STW)

对于此种情况，蔺璜等[16]提出程度副词的粘着性与定位性强，做状语时不仅不可前移也不能后置，只能紧靠在谓语周围，程度副词与情感词的距离不超过 3 个词。因此，当单词 h_{ij} 是情感词且情感词周围有程度副词和否定词修饰时，则判断词语 h_{ij} 前后距离为 3 的 6 个词语是否为程度副词，并将非程度副词的程度值设为 1。因此，词语 h_{ij} 的权重计算如式(5)所示：

$$W_{ij}=\frac{\left|tf_{ij}\times\log\left(\frac{N}{M_j}\right)\times S_j\times\prod_{-3\leqslant m\leqslant 3}D_{j+m}\right|}{\sqrt{\sum_{h_{ij}\in T_i}\left(tf_{ij}\times\log\left(\frac{N}{M_j}\right)\times S_j\times\prod_{-3\leqslant m\leqslant 3}D_{j+m}\right)^2}}。\tag{5}$$

式中：m 表示与词语 h_{ij} 的间隔距离，范围在[－3，3]之间；D_{j+m} 表示距离单词 h_{ij} 为 m 的词语的程度值。

第 4 种　词语 h_{ij} 为情感词且满足修饰规则(negative，adv，STW)

此种情况下，词语 h_{ij} 的权重计算如式(6)所示：

$$W_{ij}=\frac{\left|tf_{ij}\times\log\left(\frac{N}{M_j}\right)\times S_j\times 0.5\times\prod_{-3\leqslant m\leqslant 3}D_{j+m}\right|}{\sqrt{\sum_{h_{ij}\in T_i}\left(tf_{ij}\times\log\left(\frac{N}{M_j}\right)\times S_j\times 0.5\times\prod_{-3\leqslant m\leqslant 3}D_{j+m}\right)^2}}。\tag{6}$$

设 $\boldsymbol{h}_{ij}$ 为使用 Word2vec 训练得词语 h_{ij} 的词向量，则融入情感信息词语的加权词向量 $\boldsymbol{a}_{ij}$ 表示如式(7)所示。

$$\boldsymbol{a}_{ij}=\boldsymbol{h}_{ij}\cdot W_{ij}。\tag{7}$$

设语料中第 i 条评论文本为 $T_i=\{h_{i1},\cdots,h_{ij},\cdots,h_{ik}\}$，则文本 T_i 的向量表示如式(8)所示：

$$\boldsymbol{t}_i=\sum_{j=1}^{k}\boldsymbol{h}_{ij}\cdot W_{ij}。\tag{8}$$

将向量 $\boldsymbol{t}_i$ 作为特征输入到 SVM(支持向量机)中，训练可得到情感分析的分类模型。

2　专属领域情感词典的构建及特征提取改进方法

2.1　情感词典的构建

在计算词语权重时需要使用情感词典，中文文本语义博大精深，同一个情感词在不同领域文本中出现时，对文本的情感倾向贡献是不同的，如，“股票跌了”和“票价跌了”，前一个句子中“跌”的情感倾向是消极的，后一个句子中“跌”的情感倾向是积极的，因此构建专属领域情感词典是必须性的[17]。因此在进行情感分析之前依据词向量构建了一个专属领域情感词典[18-19]。

2.1.1　情感种子词典的构建

从携程网站上爬取到的 15 000 条河北省旅游景点的评论文本，使用 jieba 工具分词后得到的评论文本词集(TSet)，与知网情感词典(HowNet[20])取交集，得到一个情感词集(TSSet＝$\{s_{ij}\}$，s_{ij} 指情感倾向为 i 的 j 个情感词语)，利用 Word2vec 模型将情感词集的词变换为词向量($\boldsymbol{s}_{ki}$)，为了使情感种子词有较好的聚类效果，构建了一个基于余弦相似度的种子词集筛选标准，如式(9)和式(10)所示。

$$\mathrm{ADIS}(\boldsymbol{s}_{ki})=\frac{1}{n}\sum_{j=1}^{i}Dis\ (\boldsymbol{s}_{ki},\boldsymbol{s}_{kj})=\frac{1}{n}\sum_{j=1}^{i}\frac{\boldsymbol{s}_{ki}\cdot\boldsymbol{s}_{kj}}{\|\boldsymbol{s}_{ki}\|\times\|\boldsymbol{s}_{kj}\|}。\tag{9}$$

式中：$\boldsymbol{s}_{ki}$ 和 $\boldsymbol{s}_{kj}$ 表示情感倾向为 k 的 2 个不同的词语的词向量；$\mathrm{ADIS}(\boldsymbol{s}_{ki})$ 表示情感倾向为 k 的第 i 个情感词的平均距离。

$$\mathrm{SThreshold}_k=\frac{1}{n}\sum_{i=1}^{n}\mathrm{ADIS}(\boldsymbol{s}_{ki}),\tag{10}$$

式中：$\mathrm{SThreshold}_k$ 表示情感倾向为 k 的情感词的距离阈值。

当 $\mathrm{ADIS}(\boldsymbol{s}_{ki})>\mathrm{SThreshold}_k$ 时，将词语 s_{ki} 存入种子情感词典(SSDic)中，并标注其情感倾向为 k。

2.1.2　专属领域情感词典的构建

判断词语情感倾向的方法有 SO-PMI 算法(点互信息算法)和语义相似度算法。本文使用文献[21]改进后的 SO-PMI 算法计算词集(TSet)的词语与种子情感词典(SSDic)中词的 SO-PMI 值，以得到词集(TSet)中词语的情感倾向和情感极值。将 SO-PMI 值大于零的词语及该词语的 SO-PMI 值作为情感词的情感极值存入积极情感词典中，将 SO-PMI 值小于零的词语及该词语的 SO-PMI 值作为情感词的情感极值存入消极情感词典中，得到专属领域情感词典。

2.2 改进的特征提取方法

2.2.1 语义规则分析

句子可以分为单句和复杂句。单句指直观地表达对景点情感的句子，如“景点很美还会来!”，而复杂句是由多个单句以一定的逻辑结构结合在一起，如“城墙不错其他就一般了，古城内环境不好，卫生状况差，为什么不能搞得好一点呢?”，句中积极和消极的评论交织在一起，面临这种情况，需要从句子本身出发，弄清其逻辑结构，分析句子中对情感倾向有较大贡献的部分以及贡献较小或没有贡献的部分。将复杂句(C)表示为单句的集合，即 $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$ 。将 sc_i 设置为单句 c_i 对旅游评论文本的情感倾向贡献值，若 sc_i 为零时，单句 c_i 对文本的情感倾向无贡献，因此在进行文本情感分析时应忽略单句 c_i 。

1)总结词情感规则

若评论文本这样描述“一个四面环水的小村落，感觉还是不错的，但毕竟是有人居住的地方，所以不要四处乱转。总结一下：家距离景点近的人可以去玩，里面挺好的。”这条评论文本的最后一句含有总结词“总结一下”，这表明该句为总结句，在一段文本中总结句起到点明中心的作用，则该评论文本的情感重心落在总结句上。因此，若复杂句 C 包含的单句 c_i 中出现总结词，则 $sc_k,sc_{k+1},\cdots,sc_{i-1}=0;sc_i,sc_{i+1},\cdots,sc_n=1$。基于此，构建了总结词词典，其部分内容如表1所示。

表1 总结词词典

Tab.1 Dictionary of summary words

总结词
总之、总而言之、总结一下、反正、整体来说、综上所述、简而言之

2)转折词情感规则

除了总结词之外，转折词也会改变文本的情感重心，若复杂句 C 中的单句 c_i 包含“虽然”“尽管”则单句 c_i 对复杂句 C 的情感倾向无贡献，即 $sc_k,sc_{k+1},\cdots,sc_{i-1}=1;sc_i,sc_{i+1},\cdots,sc_j=0$，因此该类转折词其标注为一类转折词。若复杂句 C 中的单句 c_i 包含“然而”等转折词，复杂句 C 的情感重心落在单句 c_i 之后，因此将该类转折词标注为二类转折词，则 $sc_k,sc_{k+1},\cdots,sc_{i-1}=0;sc_i,sc_{i+1},\cdots,sc_j=1$。基于此，构建了转折词词典，部分内容如表2所示。

表2 转折词词典

Tab.2 Dictionary of turning words

一类转折词	二类转折词
尽管、虽然、即便、就是、几乎、不管、不足的是、只不过、只是有点、也就	但是、可是、重要的是、然而、可、但

2.2.2 改进特征提取

对于情感分类的研究，若忽略文本中一些词对情感极性大小的贡献进行无差别特征提取[22]，势必影响情感分类的准确性，增加实验工作量。本文针对复杂句式，通过对语义规则进行分析，改进了特征提取。判断评论文本中是否存在总结词。若存在，则直接提取包含总结词句子的特征词；若不存在，则判断句子中是否存在转折词。若存在转折词，则继续判断此转折词是一类词还是二类词：若是一类词，则忽略该句；若是二类词则提取句子中的特征词。若评论文本中不存在总结词和转折词，则直接提取全句的特征词。提取流程如图2所示。

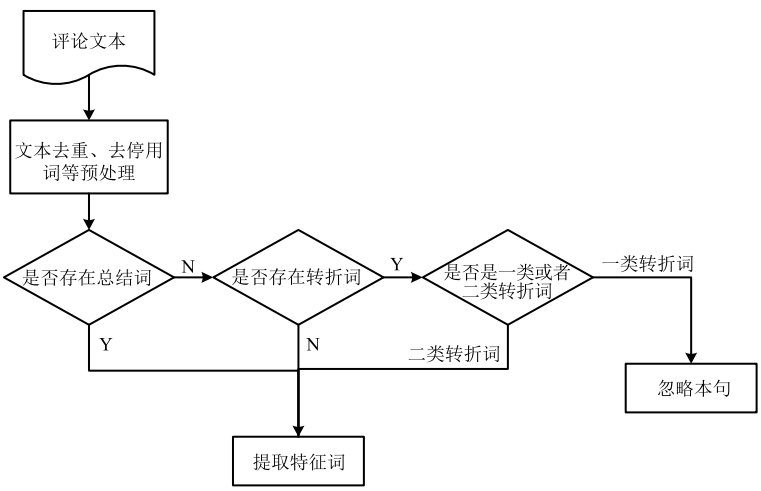

图2 特征词提取流程图

Fig.2 Feature word extraction flowchart

3 实验验证

实验硬件环境是ThinkPadE445，CPU主频2.5 GHz，内存16 GB；软件环境是Windows 10操作系统，开发工具是PyCharm 2018.2.2，开发语言是Python，分词工具是jieba，分类算法使用SVM(支持向量机)算法。

3.1　程度副词与停用词词典的处理

1）程度副词预处理。使用的程度副词来自 HowNet 词典。依据陈羽等[23]对程度副词的研究，“透顶”等词语是形容词，因此本文参考此标准删除程度词典中的此类词。

2）程度量化值的设定。根据张宗洁[24]对程度副词的研究，将程度副词按照修饰情感词的强度分为极高、高、中、低 4 类。文献[25]利用 MMTD 算法和真值程度函数计算出了程度副词的真值程度，本文参考文献[25]计算程度值的方法以及文献[26]—文献[29]为程度副词设置了程度量化值（表中用 D 表示）：1.9，1.5，1.1，0.7。另外，否定词能颠覆评论文本的情感倾向类[21]，将否定词也存入到程度词词典中，量化值设为−1。程度词词典部分内容如表 3 所示。

表 3　程度词词典

Tab.3　Dictionary of degree adverbs

ADV	D 值
最	1.9
颇为	1.5
相对	1.1
略为	0.7
不	−1

3）停用词词典处理。停用词在文本中不会传递任何信息。去除停用词词典中所含有的转折词词典和总结词词典中的词，构建适用于评论文本情感分析的停用词词典。

3.2　数据获取与数据预处理

本文以旅游网站的游客评论文本作为情感分析数据，对提出的改进方法进行实验，验证方法的有效性。

1）数据获取　从携程网上爬取赵州桥、广府古城、承德避暑山庄等河北省 30 个景点的游客评论文本数据。

2）数据清洗　分析后发现，获取到的游客评论文本中有一些是无用评论，评论文本不包含任何信息，还有一些评论文本是游客对网站服务质量的评论，以及一些重复的文本，这些数据会影响最终情感分析结果的准确性，因此手动删除此类数据。最终获取得到了 15 000 条数据。

3）数据标注　对上述经数据清洗后的携程网用户的评论数据进行人工情感倾向标注，为了标注的准确性，参考游客对景点的星级评价，将星级评价为 4 星、5 星并且评论文本明显具有积极倾向的文本标注为积极评论文本，将星级评价为 0 星和 1 星且评论文本具有明显消极倾向的文本标注为消极评论文本，最终获取得到了 10 000 条数据标注过的游客评论文本。

4）文本分词　本文使用的分词工具是 jieba，在分词前为了使分词结果更适用于本文的游客评论情感分析研究，将情感词典、程度副词词典以及转折词词典导入 jieba 词库中。

3.3　旅游专属领域词典的构建

将分词后的携程网上的游客评论文本按照语义规则分析进行种子情感词的构建，最终得到 89 个积极倾向的种子情感词和 82 个消极性倾向种子情感词，然后将种子情感词存入种子情感词典（SSDic）。

利用词典 SSDic 和专属领域情感词典方法构建旅游专属领域的情感词典（STW）。STW 词典的部分内容如表 4 所示。

表 4　STW 词典

Tab.4　Dictionary of STW

STW	S 值
闹心	−3.2877
郁闷	−3.0299
值得一去	3.4056
不虚此行	4.4074

3.4　实验评估指标

以准确率、召回率、F 值作为评价指标，评价实验方法的有效性。

准确率是指被正确分类的评论文本数量占总评论文本数量的比值，如式(11)所示：

$$P=\frac{Q_{i\text{right}}}{Q_{i\text{right}}+Q_{i\text{wrong}}}, \tag{11}$$

式中：P 为准确率；$Q_{i\text{right}}$ 是属于情感倾向类别 C_i 被正确分类的文本数量；$Q_{i\text{wrong}}$ 是属于情感倾向类别 C_i 被错误分类的文本数量。

召回率是指属于某情感倾向的文本 C_i 被正确分类的文本数量与评论文本中真正属于情感倾向 C_i 评论文本数量的比值，如式(12)所示：

$$R=\frac{Q_{i\text{right}}}{Q_{i\text{all}}}。 \tag{12}$$

式中：R 表示召回率；$Q_{i\text{all}}$ 表示实际评论文本中属于情感倾向类别 C_i 的文本数量。

F 值是准确率和召回率的调和平均值，计算公式如式(13)所示：

$$F=\frac{2\times P\times R}{P+R}。 \tag{13}$$

3.5 结果与分析

将旅客评论文本的加权词向量作为特征向量，并使用SVM算法对本文采集到的数据进行情感分析，为了测试本文所提方法的有效性，设置了4组对照实验：第1组 基于情感词典方法，利用HowNet词典和语义规则计算游客评论文本的情感倾向；第2组 利用Word2vec词向量表示方法和机器学习中SVM算法进行游客评论文本的情感分类；第3组 利用HowNet词典和文本提出的融入情感信息的加权词向量表示方法和机器学习中SVM算法进行游客评论文本的情感分类；第4组 使用本文提出的构建专属领域情感词典方法，构建旅游专属领域情感词典(STW)，结合文本提出的融入情感信息的加权词向量表示方法以及机器学习中SVM算法进行游客评论文本的情感分类，实验结果如表5所示。

表5 4组实验结果对比

Tab.5 Comparison of four groups of experimental results

实验类别	情感倾向类别	P值	R值	F值
第1组	积极	0.732	0.722	0.726
	消极	0.723	0.701	0.711
第2组	积极	0.843	0.836	0.839
	消极	0.837	0.823	0.830
第3组	积极	0.891	0.889	0.890
	消极	0.882	0.871	0.876
第4组	积极	0.904	0.902	0.903
	消极	0.897	0.895	0.896

由表5及实验分析可知：

1)第4组实验比第1组实验的准确率要高，其中积极评论文本的准确率提高了17.2%，召回率提高了18%，F值提高了17.7%；消极评论文本的准确率提高了17.4%，召回率提高了19.4%，F值提高了18.5%，因此，与基于情感词典的方法相比，使用本文提出的方法进行情感分析准确率更高，克服了过于依赖情感词典的缺点。

2) 第4组比第2组实验的准确率要高，其中积极评论文本的准确率提高了6.1%，召回率提高了6.6%，F值提高了6.4%；消极评论文本的准确率提高了6.0%，召回率提高了7.2%，F值提高了6.6%。提出的方法在进行词向量表示时考虑了词语的情感信息，提高了情感分析的准确率。

3) 第4组比第3组实验的准确率要高，其中积极评论文本的准确率提高了1.3%，召回率提高了1.3%，F值提高了1.3%；消极评论文本的准确率提高了1.5%，召回率提高了2.4%，F值提高了2.0%。实验表明，利用建立的专属领域情感词典方法结合本文提出的融入情感信息词向量情感分析方法比利用公开的情感词典HowNet结合本文提出的融入情感信息词向量情感分析方法更有效，可以提高积极和消极文本的准确率、召回率和F值，因此本文建立的专属领域情感词典是有必要的。

4 结 语

本文提出了一种融入情感信息加权词向量的情感分析方法，用以评论文本的情感倾向。对爬取的河北省游客的评论文本进行情感分析实验。结果显示，与传统的分布式词向量表示的情感分析方法相比，使用提出的改进方法进行情感分析，积极文本的准确率提高了6.1%，召回率提高了6.6%，F值提高了6.4%；消极评论文本的准确率提高了6.0%，召回率提高了7.2%，F值提高了6.6%。这表明使用提出的融入情感信息加权词向量的情感分析方法可以有效提高情感分析的准确度。

但是，本研究尚存在一些不足，所提方法无法对未登录词进行识别，在进行词向量表示时没有考虑到未登录词对文本情感倾向的贡献。未来将就未登录词的识别算法进行研究，利用专属领域情感词典构建方法，将包含情感信息的未登录词加入到情感词典中，以此获得未登录词融入情感信息的词向量表示，进而提升文本库情感分析的准确性。

参考文献/References：

[1] KHAN F H,BASHIR S,QAMAR U.TOM:Twitter opinion mining framework using hybrid classification scheme[J].Decision Support Systems,2014,57:245-257.

[2] 吴杰胜，陆奎.基于多部情感词典和规则集的中文微博情感分析研究[J].计算机应用与软件，2019，36(9)：93-99.
WU Jiesheng，LU Kui.Chinese weibo sentiment analysis based on multiple sentiment lexicons and rule sets[J].Computer Applications and Software，2019，36(9)：93-99.

[3] 万岩，杜振中.融合情感词典和语义规则的微博评论细粒度情感分析[J].情报探索，2020(11)：34-41.
WAN Yan，DU Zhenzhong.Fine-grained sentiment analysis of microblog comments based on fusion of sentiment lexicon and semantic rules[J].Information Research，2020(11)：34-41.

[4] 涂海丽，唐晓波.基于在线评论的游客情感分析模型构建[J].现代情报，2016，36(4)：70-77.
TU Haili，TANG Xiaobo.Tourist sentiment analysis model building based on online reviews[J].Modern Information，2016，36(4)：70-77.

[5] ZHANG S X，WEI Z L，WANG Y，et al.Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary[J].Future Generation Computer Systems，2018，81：395-403.

[6] 胡梦雅，樊重俊，朱玥.基于机器学习的微博评论情感分析[J].信息与电脑(理论版)，2020，32(12)：71-73.
HU Mengya，FAN Chongjun，ZHU Yue.Emotional analysis of Weibo comments based on machine learning[J].China Computer & Communication，2020，32(12)：71-73.

[7] KUMAR S，GAHALAWAT M，ROY P P，et al.Exploring impact of age and gender on sentiment analysis using machine learning[J].Electronics，2020，9(2)：374.

[8] ALOQAILY A，ALHASSAN M，SALAH K，et al.Sentiment analysis for Arabic tweets datasets：Lexicon-based and machine learning approaches[J].Journal of Theoretical and Applied Information Technology，2014.doi：10.1504/IJSNM.2015.072280.

[9] YASIN S，ULLAH K，NAWAZ S，et al.Dual language sentiment analysis model for YouTube videos ranking based on machine learning techniques[J].Pakistan Journal of Engineering and Technology，2020，3(2)：213-218.

[10] PANG B，LEE L，VAITHYANATHAN S.Thumbs up? sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10.USA：Association for Computational Linguistics，2020：79-86.

[11] CHEN W，XU Z Y，ZHENG X Y，et al.Research on sentiment classification of online travel review text[J].Applied Sciences，2020.doi：10.3390/app10155275.

[12] VALDVIA A，VICTORIA LUZON M，HERRERA F.Sentiment analysis in tripadvisor[J].IEEE Intelligent Systems，2017，32(4)：72-77.

[13] YU C M，ZHU X Y，FENG B L，et al.Sentiment analysis of Japanese tourism online reviews[J].Journal of Data and Information Science，2019，4(1)：89-113.

[14] YANG L，LI Y，WANG J，et al.Sentiment analysis for E-Commerce product reviews in Chinese based on sentiment lexicon and deep learning[J].IEEE Access，2020，8：23522-23530.

[15] MILOLOV T，SUTSKEVER I，CHENK，et al.Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2.Red Hook，NY，USA：Curran Associates Inc.2013：3000-3009.

[16] 蔺璜，郭姝慧.程度副词的特点范围与分类[J].山西大学学报(哲学社会科学版)，2003，26(2)：71-74.
LIN Huang，GUO Shuhui.On the characteristics，range and classification of adverbs of degree[J].Journal of Shanxi University(Philosophy & SociaL Science)，2003，26(2)：71-74.

[17] 严仲培，陆文星，束柬，等.面向旅游在线评论情感词典构建方法[J].计算机应用研究，2019，36(6)：1660-1664.
YAN Zhongpei，LU Wenxing，SHU Jian，et al.Construction method of sentiment lexicon for online travel reviews[J].Application Research of Computers，2019，36(6)：1660-1664.

[18] 박상민，나철원，최민성，et al.Knu korean sentiment lexicon：Bi-LSTM-based method for building a korean sentiment lexicon[J].Journal of Intelligence and Information Systems，2018，24(4)：219-240.

[19] 赵天锐，刘晨阳.基于深度学习的韩国语影评情感词典构建[J].信息技术与信息化，2021(1)：250-253.
ZHAO Tianrui，LIU Chenyang.A deep learning approach to the sentiment dictionary of korean film critics[J].Information Technology & Informatization，2021(1)：250-253.

[20] 韦婷婷，陈伟生，胡勇军，等.基于句法规则和 HowNet 的商品评论细粒度观点分析[J].中文信息学报，2020，34(3)：88-98.
WEI Tingting，CHEN Weisheng，HU Yongjun，et al.Fine-grained opinion analysis of product reviews based on syntactic rules and HowNet[J].Journal of Chinese Information Processing，2020，34(3)：88-98.

[21] 李凯.基于词典与改进信息增益的微博情感分析[D].淮南:安徽理工大学,2019.
LI Kai.Weibo Sentiment Analysis Based on Dictionary and Improved Information Gain[D].Huainan:Anhui University of Science and Technology,2019.

[22] XU G X,MENG Y T,QIU X Y,et al.Sentiment analysis of comment texts based on BiLSTM[J].IEEE Access,2019,7:51522-51532.

[23] 陈羽,徐素萍.论程度副词在书面语和口语内部的层级差异[J].文化创新比较研究,2019,3(22):92-96.
CHEN Yu,XU Suping.On the hierarchy difference between written and spoken adverbs of degree[J].Comparative Study of Cultural Innovation,2019,3(22):92-96.

[24] 张宗洁.中英文程度副词的等级数量含意对比[J].黄山学院学报,2018,20(2):52-56.
ZHANG Zongjie.A comparative study of scalar of Chinese and English degree adverbs[J].Journal of Huangshan University,2018,20(2):52-56.

[25] 何霞,杜国平,宗慧.基于中介真值程度度量的模糊语义翻译研究[J].南京邮电大学学报(自然科学版),2020,40(6):71-77.
HE Xia,DU Guoping,ZONG Hui.Research on fuzzy semantic translation based on intermediate truth degree measurement[J].Journal of Nanjing University of Posts and Telecommunications(Natural Science),2020,40(6):71-77.

[26] 敦欣卉.张云秋,杨铠西.基于微博的细粒度情感分析[J].数据分析与知识发现,2017,1(7):61-72.
GUO Xinhui,ZHANG Yunqiu,YANG Kaixi.Fine-grained sentiment analysis based on weibo[J].Data Analysis and Knowledge Discovery,2017(7):61-72.

[27] 李勇泉,李蕊,阮文奇.大型节庆活动微博用户情感态势的时空规律——以故宫上元灯会为例[J].华侨大学学报(哲学社会科学版),2019(6):27-38.
LI Yongquan,LI Rui,RUAN Wenqi.Temporal and spatial law of microblog user's emotional state in large-scale festival activities:Taking the Lantern Festival in the Forbidden City as an example[J].Journal of Huaqiao University (Philosophy & Social Sciences),2019(6):27-38.

[28] 樊振,过弋,张振豪,等.基于词典和弱标注信息的电影评论情感分析[J].计算机应用,2018,38(11):3084-3088.
FAN Zhen,GUO Yi,ZHANG Zhenhao,et al.Sentiment analysis of movie reviews based on dictionary and weak tagging information[J].Journal of Computer Applications,2018,38(11):3084-3088.

[29] 张青,韩立新,勾智楠.基于词向量和变分自动编码器的短文本主题模型[J].河北工业科技,2018,35(6):441-447.
ZHANG Qing,HAN Lixin,GOU Zhinan.Short text topic model based on word vector and variational autoencoder[J].Hebei Journal of Industrial Science and Technology,2018,35(6):441-447.