基于字符级卷积神经网络的民宿顾客意见挖掘

张 振，杨 有，罗 凌，余 平
（重庆师范大学 计算机与信息科学学院，重庆 401331）

摘 要：在线评论蕴含着丰富的顾客意见信息，顾客意见挖掘可为民宿业的良性发展提供帮助。本文以携程重庆民宿板块作为实验数据源，利用追评替代原始评论，使用弱监督的方式自动扩充数据集，针对在线民宿评论隐含特征的意见挖掘，提出可视化的 LDA 主题聚类和字符级卷积神经网络情感分析方法。实验表明，使用弱监督预训练进行预分类的方法在情感分类精度上提升了 2%；所提出的情感分析方法在 F 值上比字符级递归神经网络情感分类模型提高了 1.0%；并且可视化的 LDA 主题聚类更能直观反映顾客意见，表现为顾客对民宿设施和服务的关注度较高、顾客对民宿设施和餐饮方面意见较强烈。

关键词：弱监督预训练；主题聚类；情感分析；在线评论；字符级；
中图分类号：TP391

Character-level Convolutional Neural Networks for Homestay Inn Customer Opinion Mining

ZHANG Zhen, YANG You, LUO Ling, YU Ping
(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

【Abstract】Online reviews contain a wealth of customer feedback, and customer insights can help the development of the property industry. Taking Chongqing Homestay Inn of Ctrip as data source of experiment, and using the review to replace the original comment, using the weak supervision method to automatically expand the data set, and for the opinion mining of the hidden features of the online Homestay Inn comment, the visualized LDA topic clustering and Character-level Convolution Neural Network Sentiment Analysis method are proposed. Experiments show that the method of pre-classification using weak supervisory pre-training improves the accuracy of sentiment classification by 2%,;and the proposed sentiment analysis method is 1.0% higher than the Character-level Recurrent Reural Network Model in F-value; and the visualized LDA topic cluster can more directly reflect customer opinions, which is mainly expressed by facility and service, and customers have strong opinions on facility and food.

【Keywords】weak supervision pre-train; topic cluster; sentiment analysis; online review; character-level;

1 概述

顾客意见挖掘是对顾客需求和意见的分析，对顾客评论进行分析有利于民宿服务的改进和迭代。民宿研究的内涵和外延由于一直由于没有统一标准而受到质疑[1]，2017 年 10 月 1 日起实施的《旅游民宿基本要求与评价》定义了民宿的评价标准，使得后续民宿顾客意见挖掘可以利用标准中的指标对民宿进行评价。由于民宿服务的无形性，民宿的在线评论比其他种类信息来源的影响更大，因此，借助于顾客意见挖掘改进服务质量，是快速积累竞争优势的关键[2]。

目前主流的顾客意见挖掘方式有两种，一是针对结构化数据分析，即基于结构化数据，诸如调查问卷、利克特量表、语义差别量表等，来获得可感知的、有效的属性。孙凯[3]以满意度理论、IPA 分析模型、SERVQUAL 量表模型、新公共服务理论为基础，通过文献的梳理及实际的调研成果，设计了乡村旅游公共卫生服务质量评价量表对乡村旅游公共卫生服务质量进行测评。王晓红等[4]将 1998-2017 年中文社会科学引文索引(CSSCI)数据库收录的与隐性知识研究相关的 865 篇文献作为研究对象，采用文献计量方法，对隐性知识研究的文献特征和研究热点进行分析。步会敏等[5]通过调查问卷的方式将 SERVQUAL 量表用于旅游景区服务质量的研究，并以鼓浪屿为例使各旅游景区和政府管理部门能够更确切了解景区的发展状况和景区服务质量差距，制定旅游景区服务质量优化提升方案。二是针对非结构化数据分析，即通过自然语言处理(Natural Language Processing，简称 NLP)技术、可视化技术来分析数据自身的特点。在评论网站，论坛，博客和社交媒体中可以获得大量表达意见的文本，并在情感分析系统的帮助下，这种非结构化信息可以自动转换为结构化数据，即可以捕捉到表达关于产品、服务、品牌、政治或人们可以表达意见的其他主题等[6,7]。聂卉等[8]以大众点评网餐饮业的在线评论为研究数据，使用词向量结合依存句法分析进行领域特征词典构建和用户观点抽取，实现在线评论的特征主题凝聚及用户观点的情感计算，并对自身和竞争对手进行对比分析。

针对于非结构化的意见挖掘，根据文本嵌入粒度的不同可以分为字符嵌入和词嵌入两种。目前

基金项目：重庆师范大学科研项目(YKC18025)；重庆市研究生教育教学改革研究项目(No.YJG20163009、YJG152001)；重庆市 2015 年高等学校教学改革研究(重点项目)（No.152017）。
作者简介：张振（1993-），硕士研究生，主要研究方向为数据挖掘。杨有（通讯作者），男，博士，副教授，研究兴趣包括数字图像处理和协同过滤推荐。
收稿日期： **修回日期：** **E-mail:** 630117639@qq.com

的意见挖掘技术主要考虑词或词的组合上面[9]，对于更加细粒度的模型尝试较少。研究表明，卷积神经网络(Convolutional Neural Network，简称 CNN)对于图像分类以像素作为输入在从原始信号中抽取信息的方面非常有用,从 2009 年的 ImageNet 开始已经有大量的图像识别任务开始使用 CNN，但在自然语言处理领域，CNN 的应用才刚刚开始。2011 年，Collobert 等人首次提出使用卷积神经网络建模句子。2014 年,Kim[10]提出了 Text-CNN 的模型架构,并比较 CNN、SVM 等算法在电影评论 MR、多种商品用户评论、CR 等上面的效果，证实了 CNN 算法优于其他算法，成为文本卷积神经网络的经典架构。2015 年，Yin 等[11]发现卷积神经网络在自然语言处理领域有较好的应用效果，可以用于文本情感分析。随着近几年的深度学习技术的全面发展,Zhao 等[12]使用门限层次（Gated Hierarchical）CNN，并使用金字塔状的卷积操作自适应，得到句子的向量表示。张越等[13]为了解决情感分析中监督训练样本不足的问题,利用大规模弱监督数据来训练卷积神经网络，证实了引入弱监督数据参与训练,有效增强了卷积神经网络学习情感语义的能力,从而提升了模型的准确性。刘敬学等[14]提出了一种基于字符级嵌入的卷积神经网络(CNN)和长短时记忆网络(LSTM)相结合的神经网络模型进行短文本的分类，实验证实了字符输入进行文本分类的效果更好。

民宿评论主题鲜明、特征向量稀疏、上下文主题独立，有时甚至会出现许多非常规的语句和异常字符，如拼写错误和表情符号等，传统方法由于提取的特征稀疏，识别精度不能满足现实需求，另外由于低配置机器问题也会对识别有影响。本文的主要贡献在于：1，首先对顾客民宿的评论文本进行切割，得到较短的具有主题语义的文本，然后针对超短文本的情感分析提出自己的一套解决办法，即以高频字符表为短文本向量化的字符级卷积神经网络情感分析模型（Character-level Convolutional Neural Networks for Sentiment Analysis，简称 C-CNN-SA）；2，提高了顾客意见挖掘的精确性和合理性，首先以《旅游民宿基本要求与评价》中的指标体系为基础，使用 TF-IDF 对评论关键词进行特征提取，结合 LDA 主题聚类模型对主题属性词进行主题聚类并对评价标准进行扩充，以使用以追评作为顾客最终评论的选用策略，将用户打分为 5 分的标注为积极类，打分为 1 分的标注为消极类，使用弱监督预训练的方式即使用朴素贝叶斯分类后的高置信度文本来构建训练集[13]。使用以上方式可以将不可计算的非结构化文本数据转化为可以量化的结构化数据，实现了端对端（End-to-End）的处理方式，具有实际的应用价值。

2 相关研究

主要的研究目是针对于在线民宿评论上下文主题独立和评论不一致的问题，提出利用无监督聚类和有监督分类相结合的方式来进行顾客意见挖掘，研究在不同主题下的民宿意见情感分布情况，本文涉及的相关研究包括评论民宿主题聚类和主题情感分析。

2.1 民宿主题聚类

针对民宿评论主题性强的特点，民宿主题聚类试图将顾客评论样本按照词的形式划分为若干个通常是不相交的词集，每个词集称为一个“簇”（cluster），使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大，通过聚类可以寻找民宿评论中的论点分布。通过对民宿评价进行主题聚类可以了解顾客的需求。

目前在意见挖掘中主要使用两种主题聚类方法，一种是基于距离的主题聚类的方法。李晓英等[16]通过 K-means 聚类算法演绎不同层次下样本数据的归类分布情况进行主题聚类，可解释性强，可以将高维度词向量转化为低纬度进行可视化，增加模型的可解释性。朱晓霞等[17]提出一种基于主题-情感挖掘模型的无监督情感分类方法,通过将语义角色标注、TF-IDF 和 K-means 聚类方法相结合的方法进行主题提取。房孟春等[18]通过对在线文本评论实施词语 K-means 聚类，获取隐藏其中的消费者最为关注的民宿信誉评价指标：设施设备、服务、娱乐文化因素、卫生舒适、位置、价格、餐饮等信誉评价指标。K-means 算法使用的时候需要指定 K 值，遵循贪心算法原理，试图找到使均方误差准则函数最小的簇，具体的的算法流程如下。

步骤 1，从 n 个数据对象任意选择 k 个对象作为初始聚类中心。

步骤 2，根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的类。

步骤 3，重新计算每个所获新聚类的聚类中心（该聚类中所有对象的距离均值），并求得最小均方误差，以损失函数 $MSE=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i-Y_i\right)^2$ 进行参数调优。

步骤 4，迭代步骤 2 和步骤 3，直到每个聚类不再发生变化为止。

另一种是基于概率的主题聚类方法。典型代表就是 Beli 等在 2003 年提出的隐含狄利克雷分布（Latent Dirichlet Allocation，简称 LDA）主题模型， LDA 也称为三层贝叶斯概率模型，包含词、

主题和文档三层结构，假设有i篇文档d，总共有j个关键字w，其中有k个主题z，它们之间的关系如公式1所示。

$$p(w_j \mid d_i)=\sum_{k=1}^{k} p(w_j \mid z_k)p(z_k \mid d_i) \qquad (1)$$

在民宿评论主题聚类中，LDA 主题聚类能对隐含的主题进行提取，效果要好于词频等统计归纳主题的方式。李莉等[19]针对保险网站客服聊天记录作为语料应用 LDA 建模方法获取交互式文本主题。曾子明等人使用 LDA 与爬虫软件提取演化特征中的主题特征。LDA 是一种无监督的文档主题生成模型，认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程，这些主题被集合中的所有文档所共享，每个文档有一个特定的主题比例。对应的结构如下图1所示，M表示评价数，N表示不同的主题数，其中文档的中的词频$w_{t,n}$是一个已知的统计量，它依赖于对这个话题的指派$z_{t,n}$，以及话题所对应的词频β_k；同时，话题指派$z_{t,n}$依赖于话题分布θ，θ依赖于Dirichlet分布参数α，话题词频则依赖于参数η，大矩形表示从狄利克雷分布中为每个文档d反复抽取主体分布θ_d，小矩形表示从主体分布中迭代产生文档d的词$\{w_1, w_2, w_3, \ldots, w_n\}$。

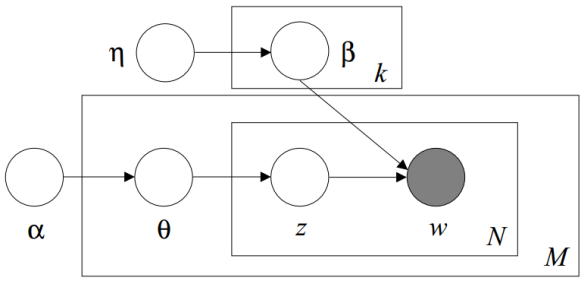

图1　LDA概率模型图

当给定一个评论集合D，包含M个评价数和N个不同的主题词数，每条评论d包含一个词序列$\{w_1, w_2, w_3, w_4 \ldots, w_N\}$，在评论集合$D$对应的LDA模型中，假设主题数目固定为$k$，则一个文档$d$的产生可以表示为以下两个步骤。

步骤1，从Dirichlet分布$p(\theta \mid \alpha)$中随机选择一个k维的向量θ_d，表示文档d中的主题混合比例。

步骤2，根据主题比例对文档d中的每个词均进行反复抽样，得到$p(w_n \mid \theta_d, \beta)$，其中参数$\alpha$是一个$k$维的Dirichlet的一个参数，如式2所示，其中$\Gamma(.)$是Gamma函数，α是模型参数，β是一个$K \times N$的矩阵，$\beta_{ij} = p(w_j = 1 \mid z_i = 1), i = 1, 2 \ldots, K; j = 1, 2, \ldots, N$。

$$p(\theta \mid \alpha)=\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \qquad (2)$$

2.2 情感分析

情感分析（Sentiment Analysis，简称SA）又称文本观点挖掘或意见挖掘，是需要先对主题进行挖掘，再进行对应属性的文本进行情感分类或者回归。情感分析的目前应用非常广泛，可以进行用户评论分析[20]、股票分析[21]、舆情分析[22]等。Vytautas等[21]根据投资者有可能在短时间内对新闻反应过度，利用新闻文章和博客文章的情感值预测比特币的浮动。

目前比较常用的情感分析方法主要有两种，一种是基于情感词典的方法是使用情感词典对文本进行情感计算从而判断其情感倾向及程度。周福礼等[23]利用 TF-IDF 算法和情感词典构建实现属性特征提取，为提升分类效率，设计并行朴素贝叶斯算法对评论信息进行情感分类。

另一种是基于机器学习的情感分析方法，包括以传统机器学习为代表的朴素贝叶斯(Naive Bayes,简称NB)、决策树(Decision Tree,简称DT)、支持向量机(Support Vector Machine,简称SVM)

等和深度学习为代表的 CNN 情感分析。樊振等[24]通过基于情感词典的方法计算出评论文本的情感倾向，然后利用用户评分的弱标注信息和基于词典方法的情感倾向对评论文本自动标注，最后,利用支持向量机(SVM)对评论文本进行情感分类。

神经网络模型的引入又一次的提升了文本分类的精度。基于卷积神经网络因为具有优秀的特征抽取能力，在情感分析中被广泛使用。卷积神经网络是一个典型的空间上的深度神经网络，能显著降低情感分析中人工抽取特征的难度，本文在正式实验开始前，使用标准的词级 CNN 和 RNN 作为预实验进行分析，对比分析在民宿评论的情感分类中的性能差异，使用 12000 条人工标注的二分类的民宿评价数据，以训练时间和精确度为指标，结果发现在训练速度上，由于 RNN 比 CNN 多出计算时间序列的步骤，RNN 在训练耗时上要比 CNN 高出 3.5 倍，并且在准确度和损失上略差于 CNN，这可能与民宿评论上下文主题独立以及评论字数较少有关，所以本文将选择 CNN 作为情感分析模型，测试结果如下图 2 所示。

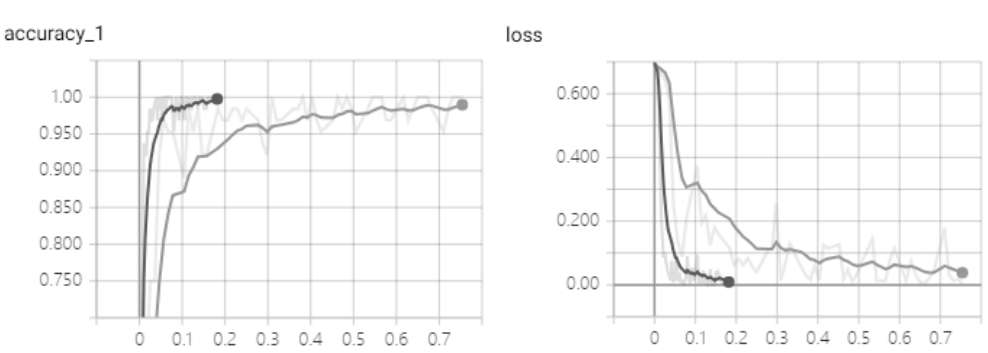

图 2　CNN 和 RNN 的对比试验（两个图横坐标都为时间（单位是小时），左图纵坐标为精确度，右图纵坐标为模型损失）

3　模型设计和算法流程

本文流程主要包括携程民宿板块评论采集、基于 TF-IDF 和 LDA 的主题属性词构建、弱监督预分类、基于字符级 CNN 民宿评论情感分析。首先采集携程重庆民宿板块的民宿评论，判断是否具有追评，并将追评作为原评论语料，结合 TF-IDF 和 LDA 对预处理后的语料进行特征聚类，并参考民宿评价标准文件,综合得出主题词和主题属性词。然后构造朴素贝叶斯弱分类器对评论文本进行预分类，并将分类后的高置信度文本作为字符级 CNN 的训练集并训练情感分析模型。最后利用主题属性词匹配主题句，使用情感分析模型进行主题下的情感极性判别和可视化，研究框架如下图 3.1 所示。

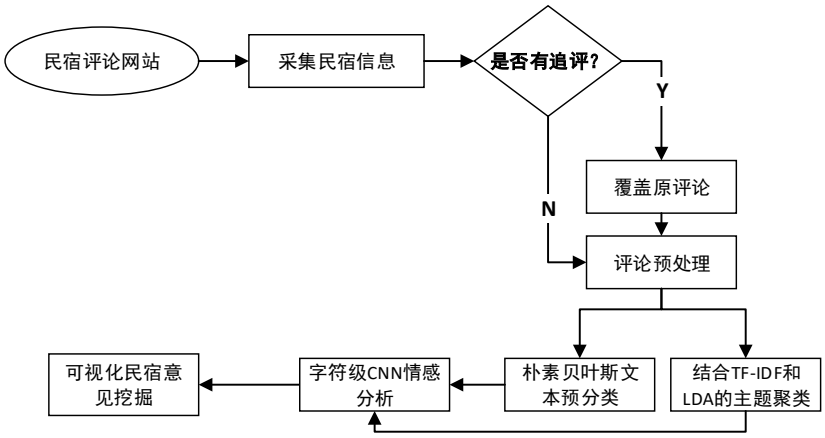

图 3　本文研究框架

3.1　TF-IDF 和 LDA 相结合的主题聚类

在民宿评论中需要对顾客讨论的话题进行判别，首先需要对文本进行向量化处理，由于传统的 TF-IDF 提取在线评论中的关键词时未考虑特征项分布出现偏差的状况，因为某个特征项如果在集合 W 中出现的次数比较多，而在其余的类别出现的少时，TF-IDF 会利用减少 IDF 的值降低该特征的权重，参考王艳东等人[25]的使用方式，假设总体文档数是 N，总词条数为 T，其中含有词条 t 的文档数是 x，而 c_t 的文档数为 y，除 c_t 外包含词条 t 的文档数为 k，TF-IDF 公式如式 3 所示。

$$TF-IDF=\frac{t}{T}\times \lg\left(\frac{N}{x+1}\right) \quad (3)$$

考虑特征项在一个类别中不同的类别间的分布情况以及特征词的位置因素对文本的区分度，因为词条出现在文档的不同位置时，对区分度的贡献大小是不一样的，所以本文采用房孟春等人提出的改进的 TF-IDF 方法来计算特征词的权重[18]，词 w 在 c_t 类中的改进 IDF 计算公式，如式 4 所示。

$$TF-IDF=\frac{t}{T}\times \lg\left(\frac{y}{x}\times k\right)=\frac{t}{T}\times \lg\left(\frac{y}{x}\right)\times k \quad (4)$$

通过改进的 TF-IDF 进行特征词向量化，利用 LDA 主题模型对主题词进行聚类可视化，根据可视化结果，按照簇内相似度高，簇间相似度低的主题选取标准，首先从《旅游民宿基本要求与评价》得出最低的主题数 k=5 来选取初始 k 值，得到初始模型，再计算各主题 t 之间的相关性，然后可视化观察聚类情况，通过增加或减少 k 的值，重新训练得到模型，再次计算主题 t 之间的相似度。重复直到得出最优的在线评论中反映用户关注点的民宿评论主题数 k 。某个词语主题的相关性，由 λ 参数来调节。如果 λ 接近 1，那么在该主题 t 下更频繁出现的词 w，跟主题 t 更相关；如果 λ 越接近 0，那么该主题 t 下更特殊、更独有（exclusive）的词 w，跟主题 t 更相关,通过调节 λ 的大小来改变领域词语 $term\_w$ 跟主题 $topic\_t$ 的相关性，主题相关性计算如式 5 所示。

$$relevance(term\_w|topic\_t)=\lambda * p(w|t)+(1-\lambda)*p(w|t)/p(w) \quad (5)$$

3.2 基于朴素贝叶斯的弱监督预分类

通过网络爬虫自动标注部分不具有追评的原评论，观察发现追评的情感往往和原始情感不一致，导致数据集数量比较少，所以使用弱分类器预先对文本进行分类，选择置信度高的文本作为后续分类器的训练文本。预分类时，我们使用的标签通常为 0 和 1，分别代表消极和积极，当朴素贝叶斯按照概率输出的时候，值域通常在[0，1]之间，通常一个类别的后验概率大于 0.5 即可分类成功，本文使用输出概率值大于 0.9 作为置信度高的积极文本，输出概率小于 0.1 的作为置信度高的消极文本，具体计算如式 6 所示。

$$p(B_j|A_k)=\frac{p(A_k|B_j)p(B_j)}{P(A_k)}\propto p(A_k|B_j)p(B_j) \quad (6)$$

其中：k 为特征数即评论的关键词数，j 为类别数。

3.3 基于 C-CNN-SA 的民宿评论情感分析

本文将字符级的文本当做原始信号，按照字符进行去重，并按照字符频率进行降序排列建立字符表，如{字符：ID}的形式，通过查询字符表中的的位置 ID 的方式将评论向量化，并且构建一维卷积核的卷积神经网络去进行特征提取，这种方式无需考虑语言的语法或语义结构，并且比词级的维度更低，消除了分词环节带来的误差，使得分类速度更快、精确度更高，例如将“阳台很大”分解为“阳”、“台”、“很”、“大”的这种字符级形式，同时对字符的位置进行标注。

参考图像处理中的像素级别处理方案，假设字典的大小为 n，通过建立字符表的方式，利用字符的 ID 将评论进行向量化，然后导入卷积神经网络进行处理。在输入层利用 Embdding 层将一个句子所有字符（$c_0,c_1,c_2,...,c_n$ 为单独的字，不进行分词处理）的字符向量进行拼接成一个句子矩阵，使用 pad 长度为 200 来覆盖 99%的文本长度，采用常用的“pre”的策略，即在文本长度不够的情况下，在前面填充 0；并对 Embdding 层的字符权重进行设置为训练更新，然后使用 Convolution1D 进行特征提取，通过一层全局最大池化层采样和两层全连接层，最后使用 softmax 层得出不同情感等级的概率分布，以概率值作为情感极性，如式 7 所示，模型结构及参数如下图 4 所示。

$$p(c)=\frac{\exp\left[\hat{y}_c\right]}{\sum_{i=1}^{c}\exp\left[\hat{y}_t\right]} \quad (7)$$

其中：p_c 是情感极性 c 的概率，c 是情感极性。

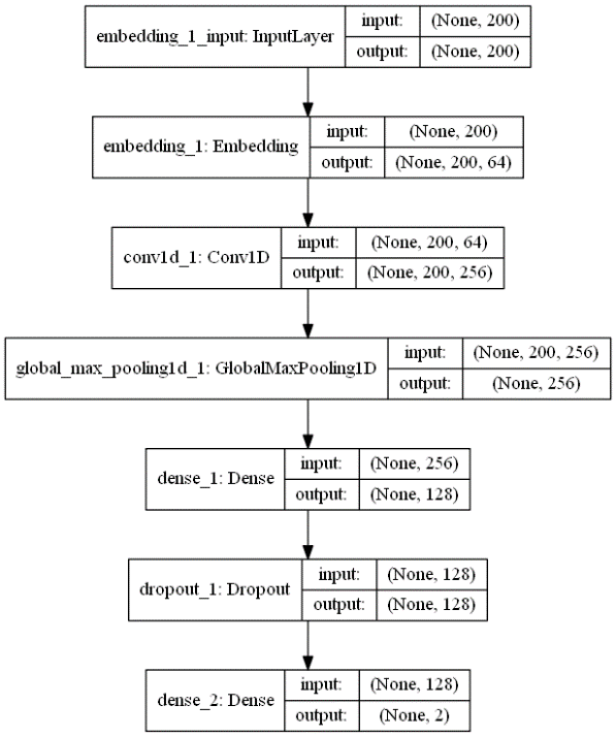

图 4　C-CNN-SA 的结构图

4　实验过程与结果分析

4.1　在线民宿评论采集与预处理

构建网络爬虫采集携程重庆民宿板块，采集 2016 年 7 月 26 至 2018 年 7 月 26 日之间所有携程重庆板块全部民宿评论，构建的主题属性词为 100，评论条数少于 100 的将影响主题提取，所以本次数据只选取评论用户数大于 100 的民宿评论和打分，最后整理出的符合条件的语料条数共有 81810 条，含有无标记的 10000 条追评。在建立民宿字典之后，利用哈工大开源 LTP 词性标注功能将标点符号利用换行符进行替代，将评论中的主题句进行分解，比如说“老板热情，房间干净整洁，而且客栈在景区内”分解为“老板热情”、“房间干净整洁”、“而且客栈在景区内”三个主题评价。

4.2　主题聚类

利用 TF-IDF 对文本进行特征提取和向量化之后，使用 pyLDAvis 对民宿评论进行可视化的主题聚类。左侧的圆圈代表了不同的主题，圆圈之间的距离是每个主题之间的相似度，能够帮助我们推算在线民宿的顾客意见数目。在选定某一个主题后，右侧面板会相应地显示出跟这个主题最近的词汇，通过总结这些词汇表达的意义，我们可以归纳出该主题的意思。参考民宿标准文件，利用实验通过 K=6 为基准，依次升高 K 值的方法进行对于主题属性词的选择，如下图 5 所示，当主题数 K=8 时各主题交叉较少，分布均匀，效果最好，选择第八个主题，内部包含的主题词有“周边环境”、“电梯”、“床上用品”、“花园”、“桌子”、“马路”等主题词，通过查看“独立”一词后接“卫生间”，通过主题词归纳之后，得出主题 8 包含的主题有，“环境”、“设施”两个主题，以下 7 个主题归纳也是同样的方式进行归纳主题，实现主题对评价的最大覆盖。

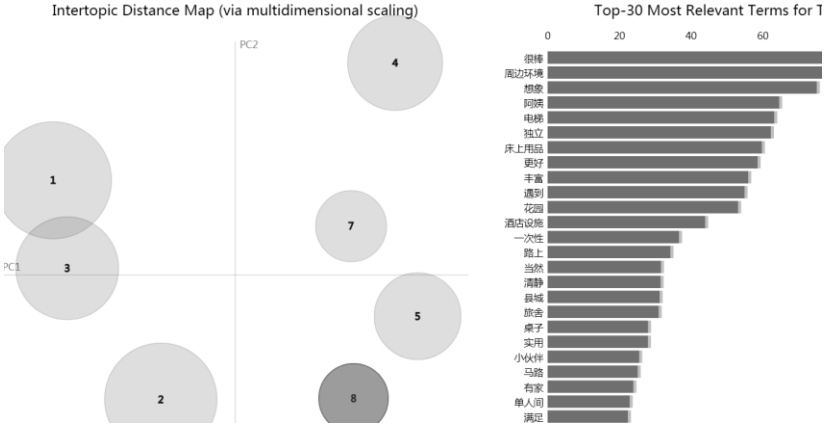

图 5　LDA 民宿主题可视化

借助于民宿规范文件和可视化聚类辅助构建民宿主题词典，构建后的的民宿主题和主题属性词如下表 1 所示。

表 1　主题属性词集

序号	主题	特征项
1	环境	环境 景点 周围 江景 声音 街 空气 夜景 小区 风景 周边 景区 景色 马路
2	价格	价格 单价 价钱 性价比 价位 房价 划算 不值
3	特色	特色 隔音 风格 设计 布置 装潢 装修 建筑 格调 结构
4	设施	设施 条件 设备 卫生间 阳台 电话 热水 硬件 电梯 马桶 大厅 空调 洗手间 电梯 摆设 被子 床 枕头 用品 桌子
5	餐饮	餐饮 食材 烧烤 咖啡 水果 餐 饭 特产 早餐 美食 味道 小吃 菜 宵夜 饭馆
6	交通	交通 地理 车站 车程 停车 位置 码头 离 中心 机场 路程 地段 海拔 火车站 汽车站
7	服务	服务 店家 态度 服务员 前台 工作人员 掌柜 老板
8	体验	体验 整体 感觉

通过属性词匹配的方式找出分句后对应的评价条数，我们对对应主题的评价条数进行统计，发现民宿评论中，顾客意见中对设施、服务、环境、交通、餐饮、特色、价格、体验的关注度依次减弱，其中对价格和体验的评论数较少，具体如下图 6 所示。

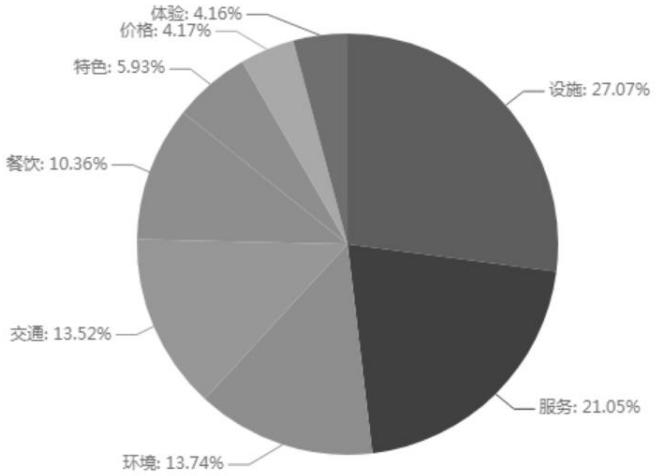

图 6　各主题评论占比图

4.3 主题意见挖掘

在 CNN 的最后一层使用 softmax 进行概率输出，可以直接输出情感倾向。利用横坐标表示对应的主题评价情感倾向，每条评论的情感评分值落在[0,1]之间，横坐标的步长设定在 0.01，越靠近 1（右部分）代表积极情感越强，越靠近 0（左部分）代表消极情感越强，概率输出在中间的位置的时候，可以认为情感为中性，情感值为 0，纵坐标表示对应的主题评论条数，通过汇总统计可得出每个主题下的情感趋势分布，可以做出了每个主题下的顾客评论情感极性图，一共 8 个主题情感趋势图，如图 7 所示，从主题的民宿顾客情感分析结果来看，可以通过分析得出以下结论。

（1）　重庆民宿的“服务”、“交通”、“体验”、“环境”、“价格”等主题评价较高，图像整体向右部分明显倾斜，这与重庆将服务业作为发展战略和“山水之城”的地理位置是分不开的，重庆交通便利，依山而建，近年来旅游指数逐年攀升，吸引大批外来游客来重庆游玩，对民宿的体验感觉新奇，从消费价格来说，重庆处于西南地区，消费较东部地区略低，价格实惠受到顾客的好评。

（2）　但是对“餐饮”、“特色”、“设施”的情感分析来看，顾客的意见比较强烈，这可能与重庆地区以吃辣为主，来重庆游玩的以外地游客居多，可能对饮食不习惯造成的，一般民宿的位置靠近景点居多，考虑成本问题，在设施上投入较少，顾客对其的意见较大，后期可以通过与景点合作来更新设施。

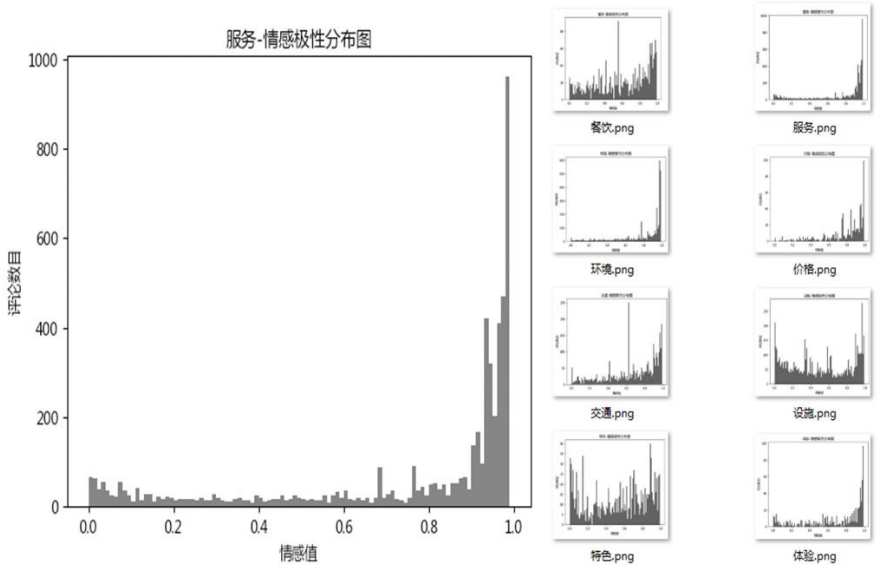

图 7　主题-情感极性分布图

情感极性图只能表示单一主题下的顾客意见倾向，我们进一步分析，按照步长为 0.2 进行情感可视化，对比多个主题下的顾客意见倾向，横坐标表示评价主题，纵坐标表示情感占比，可以同时对比多个主题的顾客意见，图中显示的情况和单一情感极性一致，顾客对民宿“设施”、“餐饮”的意见比较大，满意度较低，以后可以据此进行针对性的改善，以此来提高民宿的整体满意度，具体如图 8 所示。

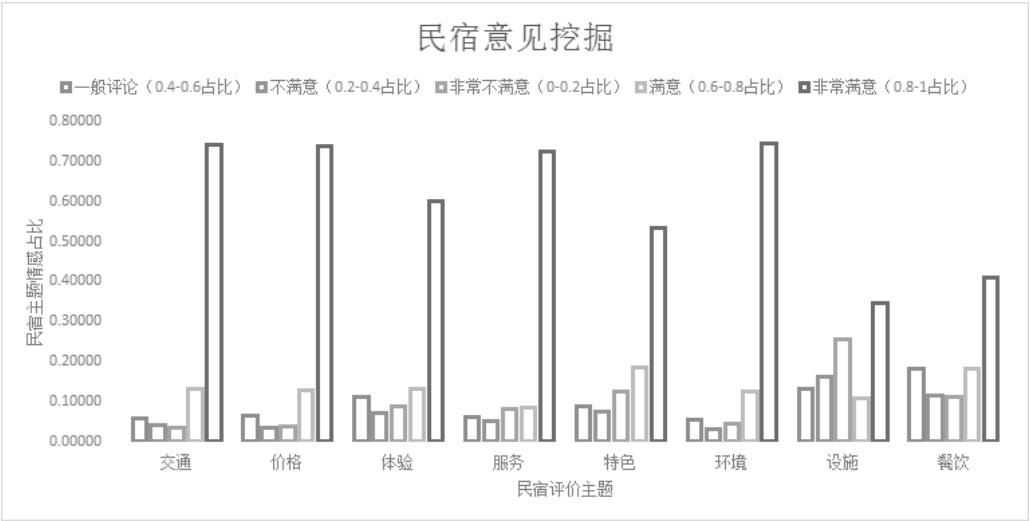

图 8　各主题下的顾客意见可视化

4.4 模型评价

为验证本文模型的有效性，同时进行 10 次实验，使用平均测试集准确度、平均精确度、平均召回率和平均 F 值作为评价指标，训练集使用 36000 经过弱训练器过滤和人工挑选的文本，测试集选用人工标记的 12000 条评论，使用词级决策树（DT）、词级朴素贝叶斯(NB)、词级支持向量机(SVM)和字符级 RNN(LSTM)四种算法和是否使用弱监督预训练的方式进行对比实验，C-CNN-SA 表示本文模型，CNN-W 表示不使用弱分类器预分类的字符级 CNN，CNN-N 表示使用标准的词级 CNN，CNN-S 表示使用去停用词后的词级 CNN，C-RNN 表示使用字符级的 LSTM，测评结果如下表所示。

表 2 模型评测数据表

分类器名称	测试集准确度	Precision(avg)	Recall(avg)	F1-score(avg)	测试条数
NB	0.9681	0.96	0.96	0.96	12000
SVM	0.9604	0.96	0.96	0.96	12000
DT	0.9504	0.95	0.97	0.96	12000
CNN-W	0.9671	0.97	0.97	0.97	12000
CNN-N	0.9738	0.98	0.98	0.98	12000

CNN-S	0.9718	0.98	0.98	0.98	12000
C-RNN	0.98	0.98	0.98	0.98	12000
C-CNN-SA	0.9875	0.99	0.99	0.99	12000

从实验结果可以看出，加入预处理步骤后，测试集的精确度提升了 2%。在情感分类上，本文利用改进模型对比传统的词级模型，在分类准确率上有一定的提升；在短文本情感分类下，字符级的粒度准确率高于词级，可能是由于预料较短的原因，使用停用词过滤可能会丧失文本信息导致分类性能下降。将字符级的文本当做原始的输入信号，直接使用一维的卷积神经网络进行特征提取，在短文本的情况下，可以无需考虑语言的单词层面的意义（包括语言的语法和语义），这种方式使得情感分析的工程得以简化。

5 结束语

从大量带有噪声和虚假的评论数据中挖掘隐藏在这些个性化评论中的情感和用户需求，将有助于企业组织和用户个人的决策行为。本文从数据驱动的角度出发，一方面证实了大数据技术来分析顾客在特定主题下的意见挖掘的可行性，挖掘出了顾客在各个主题下的满意度情况，结果可为民宿经营者和监管者提供建议，另一方面通过改进意见挖掘算法，针对民宿语料较少的问题，提出适合于民宿评论的可视化主题抽取和弱监督预训练的情感分析算法，实现在线民宿评论的隐含特征主题抽取和情感分析，并通过携程民宿评论数据的实验验证了本文方法的有效性，本文提出的顾客意见挖掘方案具有很强的通用性，对消费者、经营者和监督者也有一定的实际利用价值。但是由于本文选取的数据只选了一个在线平台，可能会造成民宿研究的局限性，未来研究将从以下两个方面开展：一是扩大数据规模，加入多平台数据；二是加入时间维度，研究民宿顾客意见在时间维度上的变化情况，以便于实时进行针对性的改进。

参考文献：

[1] 姚瑶.中国共享民宿的制度规制路径探析[J].行政管理改革,2018(10):47-51.

[2] Xiaoyue Cong. UGC QUALITY EVALUATION BASED ON METALEARNING AND CONTENT FEATURE ANALYSIS[A]. IEEE Beijing Section、中国人工智能学会（Chinese Association for Artificial Intelligence）.Proceedings of 2016 5th IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC 2016)[C].IEEE Beijing Section、中国人工智能学会（Chinese Association for Artificial Intelligence）:IEEE BEIJING SECTION(跨国电气电子工程师学会北京分会),2016:5.

[3] 孙凯. 基于游客感知的乡村旅游公共卫生服务质量评价研究[D].兰州: 西北大学,2018.

[4] 王晓红,任晓菲.基于CSSCI的我国隐性知识研究的文献计量分析[J].管理学报,2018,15(12):1854-1861.

[5] 步会敏,魏敏,林娜.基于SERVQUAL模型的旅游景区服务质量问题研究——以鼓浪屿为例[J].中国农业资源与区划,2018,39(09):190-198.

[6] Wang Hongwei, Song Yuan, Du Zhanqi, et al. Evaluation of Service Quality for Express Industry Through Sentiment Analysis of Online Reviews [J]. Journal of Beijing University of Technology, 2017, 43(3): 402-412.

[7] Zhao Zhibin, Liu Huan, Yao Lan, et al. Research on Dimensional Mining and Sentiment Analysis for Chinese Product Comments[J]. Journal of Frontiers of Computer Science and Technology, 2018, 12(3): 341-349.

[8] 聂卉，李通，何欢，刘梦圆，首欢容.基于在线评论的商业竞争情报自动获取[J].情报杂志,2018,37(10):167-173+188.

[9] 李枫林,柯佳.基于深度学习的文本表示方法[J].情报科学,2019(01):156-164.

[10] Kim Y.Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the Association for Computational Linguistics(ACL),2014,pp.655-665.

[11] Yin W, He X, Meek X. Semantic Parsing for Single-Relation Question Answering[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.2014,2: 643-648.

[12] Zhao H, Lu Z, Poupart P. Self-adaptive hierarchical sentence model[C]// International Conference on Artificial Intelligence,2015:4069-4076.

[13] 张越,夏鸿斌.基于弱监督预训练CNN模型的情感分析方法[J].计算机工程与应用,2018,54(13):27-33.

[14] 刘敬学，孟凡荣，周勇，刘兵.字符级卷积神经网络短文本分类算法[J/OL].计算机工程与应用:1-11[2018-12-22].http://kns.cnki.net/kcms/detail/11.2127.TP.20180913.0626.006.html.

[15] 王翠翠，高慧.含追加的在线评论有用性感知影响因素研究--基于眼动实验[J].现代情报,2018,38(12):70-77+90.

[16] 李晓英,周大涛.基于K-means聚类的调查问卷动态赋权统计方法[J].统计与决策,2018(23):80-83.

[17] 朱晓霞，宋嘉欣，孟建芳.基于主题—情感挖掘模型的微博评论情感分类研究[J/OL].情报理论与实践:1-11[2018-12-22].http://kns.cnki.net/kcms/detail/11.1762.G3.20181219.1124.006.html.

[18] 房孟春,曲颖.基于文本评论的在线民宿信誉评价指标关注度研究[J].地域研究与开发,2018,37(05):123-127.

[19] 李莉，林雨蓝，姚瑞波.基于LDA模型的交互式文本主题挖掘研究——以客服聊天记录为例[J].情报科学,2018,36(10):64-70.

[20] Xue Bai. The interaction design research tourism APP application under UGC mode[A]. Research Institute of Management Science and Industrial Engineering.Proceedings of 2017 International Conference on Computing,Communications and Automation(I3CA 2017)[C].Research Institute of Management Science and Industrial Engineering:计算机科学与电子技术国际学会(Computer Science and Electronic Technology International Society),2017:5.

[21] Vytautas Karalevicius,Niels Degrande,Jochen De Weerdt. Using sentiment analysis to predict interday Bitcoin price movements[J]. The Journal of Risk Finance,2018,19(1).

[22] 曾子明,万品玉.融合演化特征的公共安全事件微博情感分析[J].情报科学,2018,36(12):3-8+51.

[23] 周福礼，侯建，布朝辉，杜建辉.基于Staay多情感等级的汽车消费者行为偏好研究[J/OL].工业工程与管理:1-12[2018-12-22].https://doi.org/10.19495/j.cnki.1007-5429.2019.01.014.

[24] 樊振，过弋，张振豪，韩美琪.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.

[25] 王艳东,付小康,李萌萌.一种基于共词网络的社交媒体数据主题挖掘方法[J].武汉大学学报(信息科学版),2018,43(12):2287-2294.