doi:10.3772/j.issn.2095-915x.2023.02.004

开放科学（资源服务）标识码（OSID）

面向用户评论的店面画像构建研究
——以美团网为例

曾金[1,2,3] 黄新杰[3] 黄廷海[3]

1. 武汉大学信息管理学院 武汉 430072;
2. 湖北经济学院信息管理学院 武汉 430205;
3. 湖北经济学院大数据与数字经济研究院 武汉 430205

摘要：[目的 / 意义] 从美团餐饮用户评论数据中构建店面画像评价维度和店面画像得分，助力门店提升数字化和智能化管理水平，实现用户价值深度挖掘及门店最大化收益。[方法 / 过程] 首先抽取关键词进行 BERT 词向量表示，通过高斯混合维度聚类、维度识别及维度词挖掘来构建店面画像评价维度，然后构建领域情感词典和句法分析获得维度词得分，最后为构建的店面画像进行打分。[局限] 不足之处在于本次实验仅以武汉地区的美食自助餐评论为主，其构建的行业适用性范围有限。[结果 / 结论] 构建维度分别是：味道、环境、服务、菜品、价格，画像维度得分：5.0、2.13、2.1、1.31、1.0。实验结果证实，该方法能够有效构建店面画像，且构建的各维度得分与语料库得分具有一致性，实验表明该构建方法能取得良好效果。

关键词：用户评论；数字画像；情感词典；画像构建

中图分类号：G35 TP391

Research on the Construction of Storefront Portrait Oriented to User Comments: Taking of Meituan as an Example

ZENG Jin[1,2,3] HUANG Xinjie[3] HUANG Tinghai[2,3]

1. School of Information Management, Wuhan University, Wuhan, HuBei, 430072, China;
2. School of Information Management, Hubei University of Economics, Wuhan 430072, HuBei, China;
3. Institute of Big Data and Digital Economy, Hubei University of Economics, Wuhan 430072, HuBei, China

基金项目 国家社会科学基金项目“突发公共卫生事件用户画像构建与舆情演化机制研究”（21BTQ045）。

作者简介 曾金（1982-），博士，副教授，研究方向为文本挖掘、舆情分析，E-mail：jinzeng@whu.edu.cn；黄新杰（2001-），助理研究员，研究方向为文本挖掘；黄廷海（1999-），助理研究员，研究方向为文本挖掘。

引用格式 曾金，黄新杰，黄廷海 . 面向用户评论的店面画像构建研究——以美团网为例 [J]. 情报工程，2023, 9(2): 38-50.

Abstract: [Objective/Significance] We construct the store portrait evaluation dimensions and store portrait score from the Meituan catering user review data to help stores improve the level of digital and intelligent management, realize in-depth user value mining and maximize store revenue. [Methods/Process] Firstly, keywords are extracted for BERT word vector representation, and the evaluation dimension of storefront portrait is constructed through Gaussian mixture dimension clustering, dimension recognition and dimension word mining, and then the domain sentiment dictionary and syntactic analysis are constructed to obtain dimension word scores, and finally the constructed store portraits are scored. [Limitations] The disadvantage is that this experiment only focuses on the reviews of gourmet buffets in Wuhan, and the scope of industry applicability of its construction is limited. [Results/Conclusions] The construction dimensions include: taste, environment, service, dishes, and price, and the portrait dimension scores are 5.0, 2.13, 2.1, 1.31 and 1.0. Experimental results confirm that the proposed method can effectively construct the storefront portrait, and the scores of each dimension of the construction are consistent with the scores of the corpus, and the experiment shows that the construction method can achieve good results.

Keywords: User reviews; Digital portraits; Emotional dictionary; Portrait building

引言

随着互联网技术的发展和在线电商行业的兴起，传统餐饮行业正享受各种电商平台带来的红利。近年来，社会重大突发公共卫生事件不可避免给餐饮行业造成一系列重大影响和改变，特别是疫情防控期间餐饮行业外卖线上订单明显增长，数字化和智能化技术在一定程度上能帮助企业提升销量和降低运营成本[1]。互联网技术和社会环境的不断演化对于传统餐饮行业的发展和变革带来了前所未有的挑战和机遇，同时也给很多餐饮店面的盈利模式带来了全新思考。

在线电商步入智能化时代的核心主要是以大数据平台对餐饮企业进行驱动，用户的网络行为数据通过电商平台最终以“可视化”的形式展现在企业面前；餐饮企业自身借助互联网做运营和管理时在各种电商平台上留下了大量的用户评论数据，这些评论数据将直接影响用户的消费决策[2]。作为商家如何通过互联网技术将这些评论数据为自身构建清晰画像，并对画像进行有效挖掘和评估，发现自身不足从而来调整和改善产品或服务，对于后疫情时代下餐饮行业精细化运营和精准营销服务有着重要的意义，为餐饮行业数字化赋能起到一定作用。

1 相关研究

画像这一概念最早源于美术，主要用于对摄影及艺术的表达[3]。通过查阅数字画像相关文献发现，尚无发现对数字画像详细定义。最早提出关于数字画像相关研究源于交互设计之父 Alan Cooper[4]，以调研为目的通过访谈或观察等方法获得用户的相关特征数据，并不是将数字化技术对用户画像构建和生成。互联网场景下的用户画像是真实数据虚拟形象的一种表示，当前互联网环境下用户画像研究是通过运用数字化技术对用户生成内容描述用户特征的一种方法，其主要呈现过程包括数据采集、分

doi:10.3772/j.issn.2095-915x.2023.02.004

析建模和生成画像三个步骤[5]，主要针对用户进行精准营销或广告推荐。用户画像理论研究主要基于用户行为[6]和用户兴趣[7]数据作为标签来描述用户画像模型构建。数字画像概念较为宽泛，除了对用户进行画像外，其概念研究范围还包括杨美婷等[8]对乳制品进行分类来构建产品画像；任中杰等[9]针对突发事件舆情数据对用户进行情感画像；肖寒琼等[10]采集用户评论来挖掘用户的消费心理和消费偏好通过三层次体验理论模型来构建用户心理画像；黄家娥等[11]分析企业自身属性、企业竞争属性和企业客户属性三大要素来构建企业画像；李纲等[12]将文本和图像进行融合构建多模态语义分类模型来识别事件画像，叶晟之[13]基于多源数据探索“人城互动”视角下街道画像，马亚雪等[14]将城市划分为社会、信息及物理三个空间，将城市大数据对三个空间进行交互、映射，可以反映城市运行的整体状态，从而来构建城市画像。

互联网已经将人类社会带入数字化经济时代，数据的采集、生产、加工、分析和挖掘变得越来越便利。数字化技术的深度研究和应用已经推动各个行业进行数字化转型，驱动商业生态系统重构。数字画像实际上是利用数字化技术将数据转化成为信息或知识结构特征的标签化或可视化分析模型。随着不同学科的发展与交叉融合，数字画像技术及研究已逐渐渗透并应用到各个领域，如：数字政务[15]、电子商务[16]、数字图书馆[17]、智慧教育[18]、健康医疗[19]、酒店旅游[20]、科技情报[21]、犯罪司法[22]等。

综上所述，数字画像所面对的对象和应用领域非常广泛，其数据也是多源异构，在分析和构建画像的过程中运用的技术较为丰富。目前，大多对于数字画像研究存在以下问题：其一，在画像的构建过程中用到了用户行为数据或企业内部数据，这些数据很难获取，同时也涉及用户数据隐私问题。其二，画像评价维度模型及评价方法较为单一，很少从句法分析、情感强度和用户主观评分的视角融合来构建画像评价维度模型。其三，画像应用对象和领域研究不均衡，较多的学者研究的对象是基于用户或政府视角，很少针对店面画像视角进行研究。因此，针对店面像研究，发现文献 [23] 提出将用户选购行为和在线点评行为数据来构建服装行业门店画，但也仅仅只是停留在一种架构设计，并无实际验证效果；文献 [24] 将餐饮门店评论数据进行情感维度正负向的分句数量层面上刻画维度得分来构建维度画像，该方法构建门店画像粒度粗，其构建模型并无具体验证方法。本文提出的方法是将 SO-PMI 值与用户评分融合来构建门店画像，在维度得分的基础构建上更为细粒度情感强度和极性，并通过机器学习模型来量化和验证细粒度情感分类结果。

2 店面画像模型构建方法实现

2.1 研究思路

本文中用户评论主题挖掘研究思路如图 1 所示，详细步骤包括：

（1）数据预处理

对用户评论数据进行处理（中文分词、去停用词和词性标注），获得处理语料库。

（2）店面画像维度构建

①使用 TF-IDF[25] 关键词抽取方法获取关键词；②使用中文 BERT 预训练模型，得到关键词的 BERT[26] 词向量编码；③轮廓系数法（Silhouette Coefficient）[27] 确定最优主题个数；④通过高斯混合聚类算法确定维度；⑤构建种子主题词；⑥利用相似度确定各维度下最优维度词。

（3）构建领域情感词典

①确定种子词；②词性过滤；③ SO-PMI[28] 确定领域情感词典。

（4）店面画像维度得分计算

①句法分析，提取特定句法结构的词，构建四元组；②匹配维度词表与领域情感词典；③融合 SO-PMI 与用户评分；④店面画像构建。

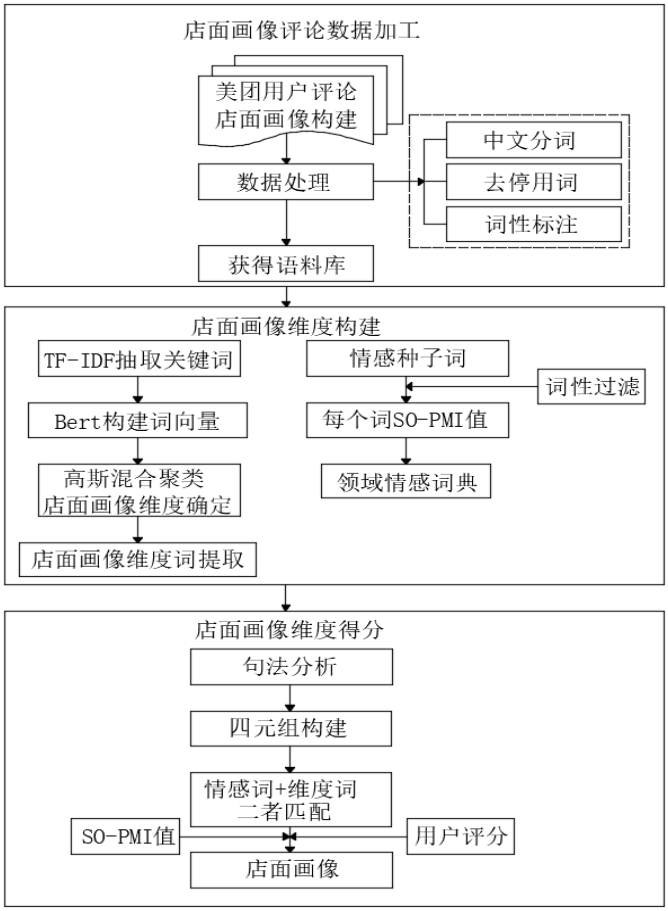

图 1　用户评论店面画像构建研究思路

2.2　店面画像维度构建

（1）店面画像关键词抽取及语义表示

在店面画像维度挖掘任务中，餐饮行业特点在于维度词和关键词重合性大，维度词通常在关键词中抽取。通过 TF-IDF 关键词抽取方法获取最优数量关键词后，将 TF-IDF 排名前 q 的词作为维度关键词集合 $KD=\{kd_1,kd_2,kd_3,\cdots,kd_q\}$。然后生成词向量模型，由于 BERT 词向量具有良好的语义表示，其独特的训练结构使其能捕捉上下文信息，故使用中文 BERT 预训

doi:10.3772/j.issn.2095-915x.2023.02.004

练模型得到关键词表征。

（2）店面画像维度确定

由于 KMeans 不适合 BERT 词向量的高维空间表示，相比之下，高斯混合是多维高斯模型概率分布的混合表示，从而拟合出任意形状的数据分布。故本文对维度关键词进行高斯混合聚类，使用轮廓系数法确定画像 s 维度个数。

高斯混合模型可以看作是由 K 个单高斯模型组合而成，是一个由 K 个子分布组成的混合分布，计算观测数据在总体分布中的概率。高斯混合模型的概率分布如公式（1）所示：

$$P(x|\theta)=\sum_{k=1}^{K}a_k\varnothing(x|\theta_k) \quad (1)$$

其中 a_k 是观测第 k 个子模型的概率，$\varnothing(x|\theta_k)$ 是第 k 个子模型的高斯分布函数，x 为观测样本。

其对数化极大似然函数如公式（2）所示：

$$\log L(\theta)=\sum_{j=1}^{N}\log(\sum_{j=1}^{K}a_k\varnothing(x|\theta_k)) \quad (2)$$

模型求解使用 EM 迭代算法，通过 Jensen 不等式极大化下界做到极大化似然函数，不断迭代直到收敛确定参数。

通过对聚类结果进行观察，存在很明显的聚类错分现象，即维度下的维度词不能代表该维度，故考虑在已确定的维度下筛选维度词，进一步提高维度分类的准确率。由于维度词数量不多，文本采取机器结合人工的方式进行筛选。筛选完毕后，对包含 m 个维度词的维度种子词集记为 $SD=\{sd_1,sd_2,sd_3,\cdots,sd_m\}$。

维度种子词机器结合人工筛选完毕后，进一步对剩下的关键词即候选维度词划分维度，将其记为 $CD=KD/SD=\{cd_1,cd_2,cd_3,\cdots,cd_k\}$，k 为候选维度词数。有以下步骤：

①计算店面画像每个维度向量，如公式（3）所示：

$$V_{dim}=\frac{V_{w1}+V_{w2}+V_{w3}+\ldots+V_{w(n_w)}}{n_w} \quad (3)$$

其中 n_w 表示对应 dimension 中维度种子词的个数，V_{w1} 表示维度种子词的 BERT 编码表示。

②对 CD 中所有词与维度做余弦相似度计算，如公式（4）所示：

$$dim[\underbrace{argmax}_{n}(similarity(dim[n],cd_w))] \quad (4)$$

③候选维度词 w 的最后划分维度为 $similarity(dim[n],cd_w)=\frac{V_{dim[n]}\cdot V_{cd_w}}{\|V_{dim[n]}\|*\|V_{cd_w}\|}$。

2.3 领域情感词典构建

（1）构建种子情感词集

本文结合通用情感词典情感词汇本体构建种子情感词集，情感词汇本体包含正面评价词语 11 229 个，中性评价词语个 5 375，负面评价词语 10 783 个，褒贬两性词汇 78 个，总的情感词 27 466 个 [29]。具体构建种子情感词集步骤如下：

①语料库匹配。使用 jieba 将语料库中文本分词，与上述情感词汇本体中情感词匹配，若出现则计入候选种子情感词集。

②情感词强度匹配。由于种子情感词需要具有较强的代表性及情感的突出性，故本文结合情感词汇本体中情感词强度进一步筛选，将候选种子情感词中强度大于 8 的作为情感种子词集并标注情感倾向。

通过上述步骤得到种子情感词集，其中正向情感词 449 个，负向情感 201 个。

（2）基于 SO-PMI 算法生成领域情感词典

点间互信息（PMI）主要用于计算词语的

语义相似度，就是统计两个词语在文本中同时出现的概率，如果概率越大，其相关性就越紧密，关联度就越高，反之关联度就越低。两个词语 word1 和 word2 的 PMI 值计算公式（5）所示：

$$PMI(word1,word2)=\log_2\left(\frac{P(word1\&word2)}{P(word1)P(word2)}\right) \quad (5)$$

其中 P(word1|word2) 表示两个词语 word1 与 word2 共同出现的概率，P(word1) 与 P(word2) 分别表示两个词语单独出现的概率，如果两个词语在数据集的某个小范围共同出现的概率越大，说明两个词语的关联度越大；反之，两个词语的关联度越小。

情感倾向点互信息法 (SO-PMI) 是将 PMI 方法引入计算词语的情感倾向中，以此达到捕捉情感词的目的。其计算公式（6）所示：

$$SO\text{–}PMI(word1)=\sum_{Pword\in Pwords}PMI(word1,Pword)-\sum_{Pword\in Nwords}PMI(word1,Nword) \quad (6)$$

其中，word1 是为确定情感极性的情感词，Pword 和 Nword 为正负情感种子词。如果最终的差值大于某一阈值时，word1 为正面情感词；反之则为负面情感词，进而可以判断情感词的情感极性，得到情感词的情感强度。

在使用 SO-PMI 方法构建美团评论的情感字典时，首先需要具备美团评论语料、种子情感词。先计算 PMI 值，接着根据情感种子词计算每个候选词的 SO-PMI 值，这里的候选词指的是不包含：‘u’助词，‘w’标点符号，‘x’非语素字符号字符，‘p’介词，‘q’量词，‘m’数词，这些词对于本文研究意义不大。最后得出美团餐饮行业领域的正、负向情感词，构建字典流程如图 2 所示：

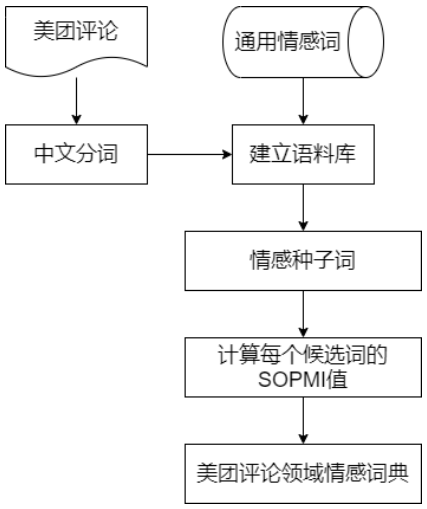

图 2　构建领域情感词典流程图

本文以与候选维度词有关系的情感种子词的 SO-PMI 值作为维度词的一种维度得分。基于 PMI 的情感倾向点互信息法主要是依据统计思想来度量词与词之间的相关性，进而确定某个词的情感极性及强度，该方法对于大多数的语料是通用的。这种方法构建情感词典需事先指定情感种子词集，适用于正、负两类情感类别词典的构建。所以，本文使用 SO-PMI 算法提高美团评论正、负向情感词的准确率，构建美团领域情感词典。

2.4　门店评分

由于每条评论中的用户评分具有重要意义，是用户对该产品的主观感受的反映，故本文提取每条评论的用户评分作为评论中维度词的另一维度得分，如一条评论中包含维度词“好吃”，且该条用户评分为 4，故将“好吃”对应的维度“味道”在该评论的维度得分赋值为 4。

2.5　门店画像构建

（1）门店画像维度词表提取及维度匹配

doi:10.3772/j.issn.2095-915x.2023.02.004

经过上述的门店画像维度提取方法得到语料的维度与其对应的维度词作为维度来构建门店画像维度词表。维度匹配主要基于维度词表，从每一句文本中找到它的维度词与维度词表匹配，得到该评论文本隶属的维度，一条评论文本可以同时隶属多个维度，随后根据句法依存关系确定该维度词对应的情感词、情感词对应的否定词、依存关系，得到句法分析的四元数组（维度词、情感词、否定词、依存关系），然后对其中情感词进一步匹配情感词典得到情感词倾向，结合否定词个数得到最后该句文本的情感倾向，进而对每条评论文本有二元数组（维度，情感强度）。

在文本中，往往用“。”来间隔意义不同的分句，在做词的依存分析时，应该分开考虑这些分句。由于每条语料由这些分句组成，所以本文先计算每个分句的二元数组，进一步再得出由分句组成的整句的二元数组，再进一步得到全文的维度得分。本文借助哈工大语言技术平台 LTP 完成。

步骤如下：

①将每个整句以“。”分隔得到每个整句的分句列表 $\{S_1, S_2, ...S_{n\_sentences}\}$。

②对每个分句进行句法分析得到分句的维度词与情感词，找到含有 SBV（主谓关系）、VOB（动宾关系）、ATT（定中关系）的维度词与情感词。三种句法关系示例如下：

主谓关系 (SBV)：名词作为主语，若该词汇为情感词，判断谓语词汇的情感倾向，如“红烧肉很不错”；若该词汇不为情感词，直接找与谓语有直接关系的所有词汇，判断它们的情感倾向。

动宾关系 (VOB)：名词作为宾语，动词则为谓语，如“喜欢这个环境”，判断谓语的情感倾向。

定中关系 (ATT)：名词作为中心语，前面为修饰语，如“不错的环境”，则直接判断该修饰语的情感倾向。

③确定否定词。从前面在句子里找出的情感词，定位与该情感词存在状中关系（ADV）、动宾关系（VOB），关系的词汇，判断其中是否存在否定词汇以及否定词的个数（奇数个否定词为否定，偶数个否定词为肯定）。如“这菜 不是 不太 好吃”。

④确定候选四元数组集合。对全文的每个分句进行上述步骤得到四元数组（维度词、情感词、否定词、依存关系），得到句法分析候选四元数组集合 CD=$\{cd_1,cd_2,\cdots,cd_n\}$，其中 cd_i 表示某一个四元数组，n 为通过句法分析得出的四元数组总数。

⑤对分句依存关系分析匹配维度词表。将每个分句得到的四元数组中的维度词在维度词表中匹配，同时将四元数组中的情感词在领域情感词典中匹配，若维度词出现在维度词表且情感词出现在特定领域情感词典中，则保留该四元数组，否则将该四元数组标注为“no topic”。得到最终四元数组集合 FD=$\{fd_1, fd_2,\cdots, fd_m\}$，其中 fd_i 为④中候选四元组集合元素或“no topic”。

⑥将⑤中的最终四元组集合中每个四元组 $fd_1,\cdots,fd_s$ 中的维度词替换成所在维度，如（口味，重，[‘有点’]，SBV）变为（味道，重，[‘有点’]，SBV）。

（2）维度得分计算

由于在（1）的句法分析中找到了与维度有

直接关系的所有情感词，且这些情感词可表现用户对该维度词所在的维度的看法，故本文以情感得分作为维度得分，又由于情感得分有内在与外在属性，即其评论本身固有的情感值与用户主观情感值，所以本文的情感得分从两方面度量：①通过 PMI 方式构建的领域情感词典中的 SO-PMI 值作为固有情感值；②用户的评论评分作为主观情感值。随后结合否定词表确定每个句子中每个维度的两方面情感得分，并结合两方面的维度得分作为最终的维度得分。具体步骤如下：

1）构建维度 SO-PMI、维度 - 用户评分二元数组。遍历（1）中得到的最终四元组集合 FD，提取其中每个四元组的情感词。第一方面，在领域情感词典中匹配得到情感词的 SO-PMI 值，随后在否定词表中匹配否定词个数，若出现奇数个否定词则记与候选情感倾向相反，对应情感得分也相反，反之相同，如公式（7）、（8）所示：

$emotion_1$=(–1)^n*(emotion_sopmi)---
若情感词为正向 （7）

$emotion_1$=(–1)$^{n+1}$*(emotion_sopmi)---
若情感词为负向 （8）

其中 n 为四元组中否定词个数，emotion$_s$opmi 为情感词的 SO-PMI 值。由于 SO-PMI 值与用户评分维度不统一，故本文将 SO-PMI 值统一 1-5 分，如公式（9）所示：

$$emotion\_sopmi = 3 + \frac{emotion\_sopmi - (max - min)/2}{(max - min)/2 - min} \quad (9)$$

等式右侧的 emotion_sopmi 为原来的值，max 表示 emotion_sopmi 中最大的值，min 表示 emotion_sopmi 中最小的值。算出统一后的 emotion_sopmi 得到维度 SO-PMI 二元数组。

然后，将每条评论的用户评分作为该评论中情感词的主观情感值，如公式（10）、（11）所示：

$emotion_2$=(–1)^n*(emotion_sopmi)---
若情感词为正向 （10）

$emotion_2$=(–1)$^{n+1}$*(emotion_sopmi)---
若情感词为负向 （11）

其中 n 为否定词在否定词表中索引到的个数，emotion_score 为情感词所属评论的用户评论评分。得到维度 - 用户评分二元数组。

2）计算维度的整体得分。综合 1）中二元数组，将每条评论对应维度得分相加得到整条评论不同维度得分，再将每条评论的不同维度得分按照维度相加得到总体的维度 SO-PMI 得分与维度 - 用户评分得分。具体公式如（12）所示：

$$score_i = \sum_{k=1}^{n} \sum_{m=1}^{C_k} comment_s core_{m_i} \quad (12)$$

其中 C_k 为第 K 条评论中二元数组个数，$comment_s core_{m_i}$ 为第 m 个二元数组中对应维度为 i 的情感值，情感值从 SO-PMI 与用户评分两方面计算。利用上述公式分别计算出两方面的总体维度得分。

3）综合维度 SO-PMI 与维度 - 用户评分。本文采用平均方式综合得到全文最终的维度得分，得到各维度得分占比且划分到 1-5 分。

3 实验

3.1 实验环境和数据准备

实验环境 i7 2.6GHz，8 核 CPU、16G 内存

doi:10.3772/j.issn.2095-915x.2023.02.004

及 64 位 Windows10 操作系统，开发工具 Jupyter Notebook 平台。所有数据通过 Python 网络爬虫在美团电子商务网上进行抓取，处理后得到的数据样例如表 1 所示。

表 1　数据样例表

评论	评分
味道不错，虽然是半条鱼，但份量很足，就是偏咸，可能是口味差异的原因。西红柿牛腩煲满分。…	4.0
第一次来这儿吃，沙拉很好吃，鱼也不错，那个牛腩嘛就一般了，鱼上得很慢，快吃完了才上的，整体还...	3.0

3.2　参数设置及维度聚类结果

使用中文 BERT 预训练模型对关键词集合（KD）中关键词进行词向量表示，使用轮廓系数法确定最优 K 值，得到图 3 轮廓系数法确定最优维度数。

同时结合上图结果分析，选取最优维度 k 值为 5，故本文将维度数设置为 5。

通过词向量语义聚类及迭代计算，发现前 300 个维度词较为合适，每个维度下的种子维度词如表 2 所示：

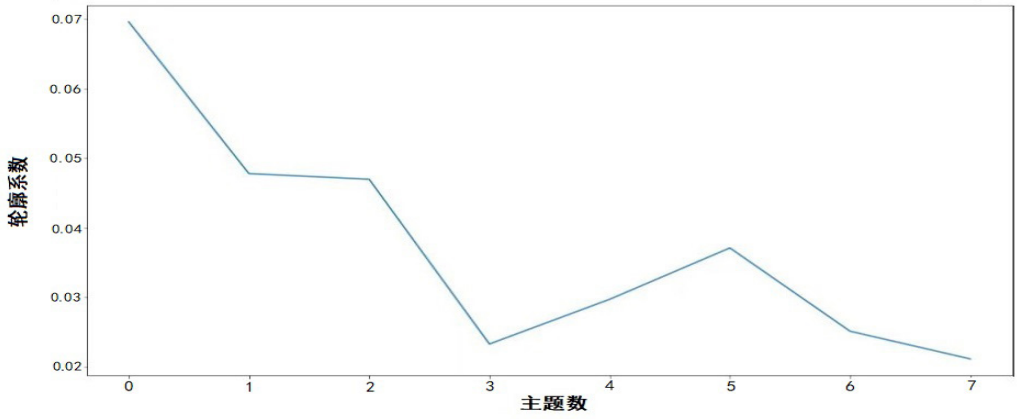

图 3　轮廓系数法确定最优维度数

表 2　种子维度词表

维度	种子维度词	种子维度词数量
味道	好吃、好喝、棒棒、超赞、喜欢、新鲜、口感、入味、辣味、尝尝、辣得、口味、过瘾...	71
服务	热情、到位、周到、超好、贴心、态度、好评、不行、态度、满意、失望、打包、推荐、改进、还好...	46
价格	免费、优惠、实惠、性价比、经济、买单、团购、划算、购买、活动、消费、涨价、价钱、价位、支持、购买...	51
环境	位置、地方、学校、光谷、装修、店面、改进、空调、桌子、位子、选择、聚会、估计、同学、朋友、好找...	74
菜品	食材、甜点、饮料、海鲜、烧烤、火锅、榴莲、披萨、烤肉、酱料、鱼肉、份量、套餐...	58
总数		300

3.3　领域情感词典效果评估

通过多次试验，发现词向量维度选取 300 时效果较好，表 5 是 Random Forest、Light GBM、Bi-LSTM 三种分类模型对分类效果进

行评估来检验领域情感词典的构建结果，选择80% 作为训练集，20% 作为验证集评估，样本输入为每个情感词对应词向量，输出为情感极性，得到准确率、精准率、召回率和 F1 值如表 3 所示。

ROC 曲线是衡量机器学习分类效果的重要工具，其线下面积即 AUC 值越大说明模型效果越好。结合表 3 和图 4 看出模型分类效果较好，反映了领域词典构建的有效性，ROC 曲线如图 4 所示。

表 3　分类指标评估表

序号	分类器	Acc准确率（%）	P精确率（%）	R召回率（%）	F1值（%）
1	Random Forest	84.51	84.88	99.41	91.57
2	Light GBM	85.01	84.96	1.00	91.86
3	Bi-LSTM	83.39	86.87	94.69	90.61

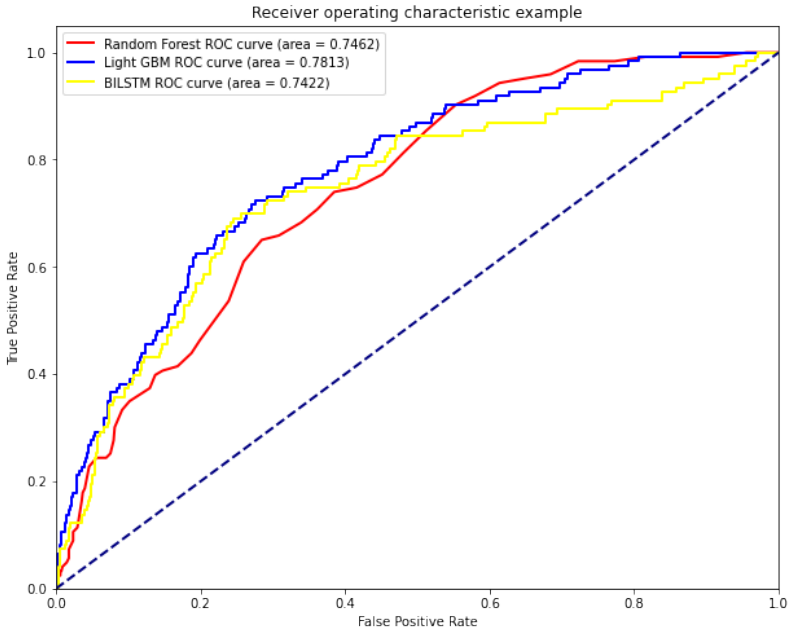

图 4　ROC 曲线图

3.4　句法分析与四元组构建

本文使用 Python 编程语言中的 <sep> 字符串将每条评论隔开，借助哈工大 LTP 平台对每条评论做句法分析，找出其中具有 SBV、VOB 与 ATT 依存关系的结构，构建四元数组（维度词，情感词，否定词列表，依存关系），部分句法分析结果如表 4 所示：

表 4　句法分析样例表

维度词	情感词	否定词列表	依存关系
份量	很足	[‘是’]	SBV
汤野菌	不太	[‘喜欢’]	ATT
…	…	…	…

3.5　门店画像结果展示

下表分别为基于用户评分与 SO-PMI 计算

doi:10.3772/j.issn.2095-915x.2023.02.004

出的维度得分，各维度占比得分使用公式（13）映射到 1-5 分之间。

$$dim\_score_i = 3 + \frac{D_i - (D_{min} + D_{max})/2}{D_{max} - (D_{min} + D_{max})/2} * 2 \quad (13)$$

其中，dim_score_i 表示第 i 个维度映射后的得分，D_i 表示维度集的第 i 个维度，D_{min} 表示维度集中最终维度得分最小的维度，D_{max} 表示维度集中最终维度得分最大的维度。如表 5 所示。

店面画像维度映射后得分雷达图和饼图，结果的可视化如图 5 和图 6 所示。

表 5　维度得分表

维度	维度-用户评分	维度-SOPMI	最终维度得分	维度映射后的得分
味道	592 108.00	500 453.99	546 280.99	5.00
环境	263 346.0	193 525.56	228 435.78	2.13
服务	268 659.0	181 541.26	225 100.13	2.10
菜品	160 869.0	113 088.77	136 978.88	1.31
价格	128 669.0	77 523.05	103 096.02	1.00

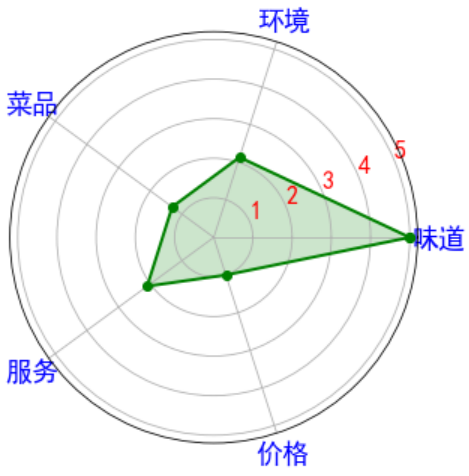

图 5　维度映射后的得分雷达图

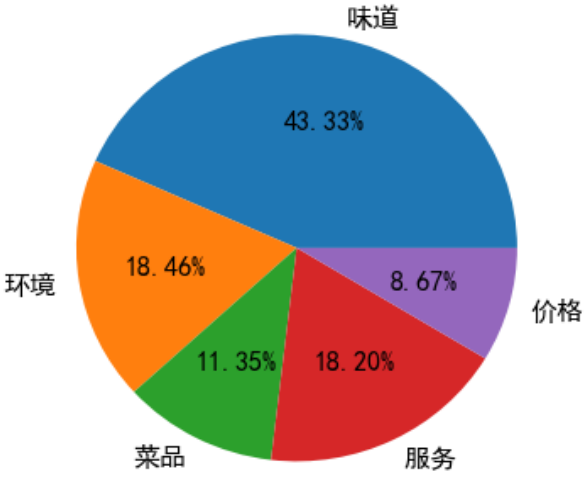

图 6　维度得分占比饼图

3.6　实验结果分析

（1）通过聚类结果图 3 和表 2 得出，本文构建店面画像 5 个维度分别是：味道、环境、服务、菜品、价格。并获取 5 个维度下的 300 个维度词，发现所构建的维度与维度词表具有较高的相似性，说明该方法能够有效识别维度和维度词，为店面画像维度定义和维度词识别提供了一种较为科学的构建思路和方法。

（2）表 3 和图 4 得出，本文提出的基于 SO-PMI 构建领域情感词典方法具有较强的准确性和可靠性。三种分类模型的 F1 值评估指标均在 90% 以上，说明情感词典在情感词的正负情感（强度和极性）识别具有较强的区分度，证明了所构建餐饮领域情感词典的有效性。

（3）表 4 得出，本文通过哈工大 LTP 平台能有效的通过句法分析得出维度词与情感词的关系，并能有效的提取否定词和准确识别情感词对应维度的依存关系，正确的判断维度词的

情感倾向及强度。

（4）从表 5、图 5 和图 6 可以看到，“味道”维度得分最高，“价格”维度得分最低，维度得分结果与维度重要性占比结果具有一定的一致性，进一步证明了本文方法的有效性。

（5）通过数据统计，菜品、环境、味道、服务、价格维度中的维度词共计出现 44 927、70 741、156 100、36 522、74 002 次。通过句法分析找到并计算各维度词对应的情感值得分与用户评分得分，可以得到：在 SO-PMI 情感值得分方面，每个维度平均得分为 2.51、2.73、3.20、2.12、2.45；在用户评分得分方面，每个维度平均得分为 3.57、3.72、3.79、3.52、3.63；与图 6 结果各维度占比排序一致。

（6）用户的线上评论和评分非常主观，这也和本文所构建的画像维度通过 SO-PMI 和句法分析融合后计算出来的情感得分非常接近，这也进一步验证模型效果很不错。即便用户不评分，构建的模型也能够给出和用户接近的评分，足以证明本文所构建店面画像具有一定的科学性和实用价值。

4 结束语

本文以美团自助餐文本评论数据为例，提出一种餐饮店面画像构建方法，通过对词语的客观句法结构与用户的主观评分两方面构建维度得分，分别使用 SO-PMI 与句法分析方法，融合客观的情感强度、极性与主观的词语评分进主题维度词中，对于维度词的得分，该方法有双向的纠正作用，即能解决评论中存在不清晰或不合适表达时，还可以通过用户评分进行纠正，识别存在不真实用户评分时通过对评论中客观情感词计算纠正，为构建的店面画像进行真实合理打分，最终实验结果证明了门店画像维度构建的有效性和可靠性，由于该方法主要用到用户评论与用户评分，故能适用于大部分美食类评价平台。在后继研究工作中，本文会从用户评论数据中对事件或菜品进行抽取，来识别用户关注的热门菜品、健康饮食和餐饮中的热点话题。上述问题的探索和研究，对后疫情时代数字化和智能化发展具有非常重要的意义。

参 考 文 献

[1] 邓桦 . 新冠肺炎疫情对全球商业环境的影响分析 [J]. 竞争情报 , 2020, 16(5): 50-57.

[2] 吴应良 , 黄媛 , 王选飞 . 在线中文用户评论研究综述 : 基于情感计算的视角 [J]. 情报科学 , 2017, 35(6): 159-163, 170.

[3] 陈奕良 . 基于街景图片大数据的街道数字画像方法研究 [D]. 南京 : 东南大学 , 2020.

[4] Cooper A, Robert Reimann R, Cronin D. About Face 3: The Essentials of Interaction Design[M]. New Jersey: Wiley Publishing Inc., 2007: 19-22.

[5] 刘海鸥 , 孙晶晶 , 苏妍嫄 , 等 . 国内外用户画像研究综述 [J]. 情报理论与实践 , 2018, 41(11): 155-160.

[6] Adomavicius G, Tuzhilin A. Using data mining methods to build customer profiles[J]. Computer, 2001, 34(2): 74-82.

[7] Godoy D, Amandi A. User profiling for web page filtering[J]. IEEE Internet computing, 2005, 9(4): 56-64.

[8] 肖扬 . 基于产品画像的汽车推荐研究 [D]. 大连 : 大连外国语大学 , 2022.

[9] 任中杰 , 张鹏 , 兰月新 , 等 . 面向突发事件的网络用户画像情感分析——以天津 “8·12” 事故为例 [J]. 情报杂志 , 2019, 38(11): 126-133.

[10] 肖寒琼 , 张馨遇 , 肖宇晗 , 等 . 基于方面词的用户

doi:10.3772/j.issn.2095-915x.2023.02.004

消费心理画像方法 [J/OL]. 数据分析与知识发现 :1-15

[11] 黄家娥 , 李静 , 胡潜 . 基于企业画像的行业信息精准服务研究 [J]. 情报科学 , 2022, 40(2): 99-104, 112.

[12] 李纲 , 张霁 , 毛进 . 面向突发事件画像的社交媒体图像分类研究 [J]. 数据分析与知识发现 , 2022, 6(Z1): 67-79.

[13] 叶晟之 . 基于多源大数据的城市街道数字画像研究与验证 [D]. 南京 : 东南大学 , 2021.

[14] 马亚雪 , 李纲 , 谢辉 , 等 . 数字空间视角下的城市数据画像理论思考 [J]. 情报学报 , 2019, 38(1): 58-67.

[15] 王志刚 , 邱长波 . 基于主题的政务微博评论用户画像研究 [J]. 情报杂志 , 2022, 41(3): 159-165.

[16] 李佳慧 , 赵刚 . 基于大数据的电子商务用户画像构建研究 [J]. 电子商务 , 2019(1): 46-49.

[17] 陈慧香 , 邵波 . 国外图书馆领域用户画像的研究现状及启示 [J]. 图书馆学研究 , 2017(20): 16-20.

[18] 尹婷婷 , 龚思怡 , 曾宪玉 . 基于用户画像技术的教育资源个性化推荐服务研究 [J]. 数字图书馆论坛 , 2019(11): 29-35.

[19] 马费成 , 周利琴 . 面向智慧健康的知识管理与服务 [J]. 中国图书馆学报 , 2018, 44(5): 4-19.

[20] 单晓红 , 张晓月 , 刘晓燕 . 基于在线评论的用户画像研究——以携程酒店为例 [J]. 情报理论与实践 , 2018, 41(4): 99-104, 149.

[21] 赵辉 , 化柏林 , 何鸿魏 . 科技情报用户画像标签生成与推荐 [J]. 情报学报 , 2020, 39(11): 1214-1222.

[22] 张凯琳 . 论犯罪心理画像技术在侦查中的应用 [D]. 杭州 : 浙江大学 , 2018.

[23] 谢慧志 . 基于用户选购及在线点评行为的门店画像研究 [D]. 广州 : 华南理工大学 , 2016.

[24] 蒋志强 . 基于点评餐饮评论数据的情感分析 [D]. 苏州 : 苏州大学 , 2019.

[25] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[26] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26.

[27] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational and applied mathematics, 1987, 20: 53-65.

[28] Turney P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[J]. arXiv preprint cs/0212032, 2002.

[29] 徐琳宏 , 林鸿飞 , 潘宇 , 等 . 情感词汇本体的构造 [J]. 情报学报 , 2008, 27(2): 180-185.