计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(8):42−48 [doi: 10.15888/j.cnki.csa.006488]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于餐饮业网络评论的消费者情感极性分析①

杨博文

(南京财经大学 经济学院 统计系, 南京 210023)
通讯作者: 杨博文, E-mail: yangbowen0717@163.com

摘　要: 首先, 根据餐饮业网络评论文本对消费者情感极性进行预测, 建立了 Lasso-Logistic 和 Lasso-PCA 两个预测模型. 相比之下, Lasso-PCA 模型整合了更多的变量信息, 对文本的情感极性具有更好的预测效果; 但是 Lasso-PCA 模型对变量的解释能力较弱, 尤其在解释变量维度较高的情况下, Lasso-PCA 模型很难分析出解释变量对被解释变量的影响. 其次, 对 Lasso-Logistic 模型的变量选择结果进一步分析发现, 特色菜、服务态度和环境以及“美中不足”之处是影响消费者情感极性的显著因素.

关键词: 网络评论文本; 情感极性; Lasso 算法; Logistic 回归; 稀疏主成分回归

引用格式: 杨博文.基于餐饮业网络评论的消费者情感极性分析.计算机系统应用,2018,27(8):42–48. http://www.c-s-a.org.cn/1003-3254/6488.html

Analysis of Consumer Sentiment Polarity Based on Chinese Online Review of Catering Industry

YANG Bo-Wen

(Department of Statistics, School of Economics, Nanjing University of Finances and Economics, Nanjing 210023, China)

Abstract: First, to predict consumer sentiment polarity based on Chinese online review of catering industry, this study establishes Lasso-Logistic and Lasso-PCA models. By comparison, Lasso-PCA model is more accurate by integrating more information of variables. However, Lasso-PCA model has weaker explanatory power especially in the scenario of high dimensional data. Second, using the variable selection results of Lasso-Logistic model, we find that specialties, service attitude, and the external environment, as well as “a fly in the ointment” are the significant factors affecting the consumer's emotional polarity directly.

Key words: text mining; sentiment polarity; Lasso algorithm; Logistic regression; sparse principal component regression

1　引言

1.1　研究背景

在电子商务蓬勃发展的信息化时代, 越来越多的互联网用户在线评价自己的消费, 这些文本的评论信息作为消费者亲身体验的反馈, 涵盖了大量的有用信息. 一方面以往消费者对产品的评价可以帮助潜在消费者事前对产品有所了解, 便于消费者根据自身需要做出消费决策; 另一方面也可以作为反馈信息帮助商家了解消费者的购买意愿、跟踪商品的售后服务等, 进而不断改进、提高自身竞争力.

消费者情感极性分析[1−3](Sentiment Polarity Analysis) 是文本分析的一大分支, 一般可以分为积极 (Positive) 和消极 (Negative) 两类, 只有准确地把握了消费者的情感极性才能做好客户的维护、挖掘潜在客户、弥补欠缺进而提升自身的市场竞争能力. 本文旨在运用高维数据变量选择方法[4]关注两方面的核心内容, 一方面寻求较好的消费者情感极性预测模型; 另一方面, 以往的研究重在分类预测, 而对评论背后隐含的商业价值很少深入探究, 本文希望借助 Lasso 算法的变量选择优势挖掘出影响消费者情感极性的关键因素.

① 收稿时间: 2017-12-21; 修改时间: 2018-01-11; 采用时间: 2018-01-19; csa 在线出版时间: 2018-07-28

1.2 研究现状

从国内外研究现状来看, 目前对文本数据的分析主要涉及提取文本特征、文本特征关联分析、文本内容识别, 以及文本情感极性分析等方面. 其中提取文本特征是对文本信息进一步分析的基础, 所谓特征提取就是根据评论文本的分词结果, 选择对文本具有代表性的关键词. 特征选取主要有两种不同的思路, 一种是构造评估函数法[5,6], 另一种是在事先挑选的初始种子集为起点, 对候选特征集合采用不断迭代的方法确定最终的特征集合[7].

以往对特征提取和文本情感极性的分析, 大都以词频和语义分析为主. Hu 用形容词作为观点词分析英文评论的情感极性, 借助 WordNet 将要判断情感倾向的词条与给定情感倾向的同义词或反义词词网相匹配, 词条的情感倾向与同义词具有相同的情感倾向, 与反义词具有相反的情感倾向[8]. Turney 以形容词和副词为分析对象, 运用 PMI 方法分别计算给定词与“excellent”、“poor”的点互信息 (PMI), 两者相减, 若为正值则情感极性为正, 反之为负. 近年来, 部分学者在对词的分析上做了进一步延伸, 如根据词条在不同文本类别间分布不均的情况, 提出了对特征项加权的方法判断情感极性[9]. 随着大数据时代的到来, 相关的机器学习方法在情感极性分析中也越来越受欢迎[10–12]. Pang 等根据事先既定的有关形容词的积极词料集和消极词料集, 分别运用朴素贝叶斯 (Naive Bayes)、最大熵 (Maximun Entropy) 和支持向量机 (Support Vector Machines) 三种方法进行文本的情感极性分析并在不同的情况下进行了对比[13]. 王健等基于主题概率模型 (LDA) 实现了文本分类, 并取得较好的分类效果[14].

1.3 研究思路

以上研究对文本情感极性的预测, 主要有两种思路, 第一种是由特征词或特征项的情感极性加权进行预测; 第二种是运用机器学习方法对文本的情感极性进行预测, 主要包括支持向量机、朴素贝叶斯、最大熵等方法等. 除此之外, 鉴于 L1-正则项对高维数据良好的惩罚特性, Lasso 稀疏模型已经被成功的应用于文本分类领域[15–19]. 鉴于此, 本文运用 Lasso-Logistic 和 Lasso-PCA 模型[20–22]对餐饮业文本评论的情感极性进行分析. 一方面, 作为对比找到更好的分类模型; 另一方面, 笔者借助 Lasso-Logistic 较好的模型解释能力对影响消费者情感极性的关键因素深入分析. 相比于 Lasso-Logistic 模型, 目前鲜有对 Lasso-PCA 模型的应用文献, 基于稀疏数据的主成分模型在解决数据稀疏性的同时, 保留了较多的变量信息, 但该方法对文本的情感极性预测效果有待于探讨.

根据以上文献综述, 本文提出以下研究思路: (1) 对数据进行预处理, 包括提取评论样本、分词等. (2) 运用 TF-IDF 算法初步提取关键词. (3) 分类预测. 以消费者情感极性为被解释变量, 以高维稀疏关键词词频矩阵为解释变量, 结合 Lasso 算法, 运用带惩罚的 Logistic 和 PCA 两种方法对消费者情感极性进行预测. (4) 借助 Lasso 的变量选择结果, 运用 Logistic 模型对显著影响消费者情感的因素做进一步的分析.

2 数据来源与处理

2.1 数据来源

本文数据来源为大众点评网上某餐厅的消费者评论的文本内容和评分等级, 共 2293 条评论记录. 因变量是评论者对消费情况的总体评价的星级数据, 分为 5 个等级, 1 颗星代表最低评价, 5 颗星代表最高评价. 考虑到实际情况, 消费者一般倾向于给出较高的星级指数, 在评分为 3 的样本中大都带有消极的情绪, 如表 1 所示. 因此, 在分析过程中将 1 颗至 3 颗星的样本视为情感极性为负; 将 4 颗星和 5 颗星的样本视为情感极性为正. 这里随机抽取了 400 条积极样本和 400 条消极样本, 作为对消费者情感极性分析的总样本.

表 1 部分样本信息

评分 (level)	评论文本
1 或 2	味道太烂了, 而且老板娘一直和老顾客说话, 不太理会我们, 什么态度, 气死我了, 下次不来了
	一般般, 菜品分量太小了, 适合儿童来吃, 环境还不错, 倒是看环境看不饱啊, 也就来这么一次吧, 聚餐吃不好
3	感觉环境不错, 菜量着实有点小, 服务还是不错的, 感觉宫保鸡丁真的不如眉州东坡的.
	大份牛蛙和小份牛蛙相差三十块钱但是里面的牛蛙是一样多的毛血旺里没有毛肚黄喉里面只有血块豆皮和豆芽和一片肥牛
4 或 5	味道很正, 服务也很好, 喜欢这里的水煮鱼、牛蛙、毛血旺、生煎包…
	团购点评这两道菜个人认为是每次都会点的, 味道和别的地方的有点区别, 肉质很嫩！赞

首先提取 1000 个关键词作为初始特征集, 然后遍历每一条评论的分词结果, 分别统计特征词在每条评论中出现的频数, 以由此得到的稀疏矩阵作为解释变量. 不失一般性, 在分析过程中选用样本的 80% 作为训练集, 20% 作为测试集进行样本外预测.

2.2 数据处理

数据处理的第一阶段是利用 R 软件的加载包 jiebaR 对网络评论文本进行分词, 首先在分词的过程中去除常用停用词 (stop words, 如介词、冠词、限定词等); 同时考虑到分词结果会产生数值型的分词结果, 所以在数据的预处理过程中删除了数值型的分词结果; 最后运用该软件包提供的词频-逆向文本频率算法 (TF-IDF) 提取关键词, 作为备选特征词集合.

TF-IDF 算法是提取文本关键词常用的统计方法, 用以评估一字词对一个文本的重要程度. 其基本思想是如果一个词比较少见, 但是它在这个文档中出现多次, 那么它很可能就反映了这个文档的某方面特性, 可以作为该文档的关键词 . 该算法分为词频 (Term Frequency, TF) 和逆向文本频率 (Inverse Document Frequency, IDF) 两部分. TF 即一个词在目标文本中出现的频率, 见式 (1). IDF 是对该词代表的信息量的衡量, IDF 值的计算需要一个词料库, 由词料库中总文件数除以包含该词的文档数, 再将商取对数得到, 见式 (2). TF - IDF 值即 TF 与 IDF 的乘积, 见式 (3). 这里选用的是 R 软件 jiebaR 包自带的词料库.

$$tf(t,d)=\frac{n_{t,d}}{\sum_k n_{k,d}} \tag{1}$$

$$idf(t,D)=\log\frac{N}{|\{d\in D:\ t\in d\}+1|} \tag{2}$$

$$TF-IDF=tf(t,d)\times idf(t,D) \tag{3}$$

其中, $n_{t,d}$ 是词 t 在文档 d 中出现的频率; N 代表词料库中的文件数, $N=|D|$; $|\{d\in D: t\in d\}|$代表词料库中包含词 t 的文档数, 为避免该词不在词料库中的情况, 将 $|\{d\in D: t\in d\}|+1$ 作为分母.

运用 TF-IDF 算法可以得到对文本内容具有代表性的关键词, 根据文本的分词结果统计出 1000 个关键词的词频矩阵, 如表 2 所示. 从表中可以看出, TF-IDF 值较大有“水煮鱼”、“川菜”、“味道”和“毛血旺”、“宫保鸡丁”等名词性词汇, 以及“不错”、“好吃”和“喜欢”等带有情感极性的形容词、副词和动词.

表 2 部分关键词词频矩阵

序号	1	2	3	…	1000
关键词样本	水煮鱼	川菜	味道	…	解暑
1	1	0	0	…	0
2	0	2	0	…	0
3	0	1	0	…	0
4	0	0	0	…	0
…	…	…	…	…	…
799	0	0	1	…	0
800	0	0	0	…	0
求和	274	299	458	…	2
TF-IDF	2753.2	2611.9	2596.7		22.5

此外, 从词频的角度来看, “味道”、“不错”的频率明显高于“水煮鱼”和“川菜”; 但是从 IF-IDF 值来看, “水煮鱼”和“川菜”的值则高于“味道”、“不错”. 这是因为, “味道”和“不错”出现的频率虽然高, 但是对文本内容的代表性不够, “水煮鱼”和“川菜”则直接反映出了文本的主题, 具有更好的代表性. 同时可以看出, “水煮鱼”和“毛血旺”、“宫保鸡”具有较高的频数和 TF-IDF 值, 且“水煮鱼”高于“毛血旺”和“宫保鸡”, 说明这三个菜品在该家餐厅中比较具有特色, 尤其是“水煮鱼”, 建议作为餐厅的特色菜来打造. 同时也说明了以表 2 的关键词词频矩阵作为解释变量对文本的情感极性进行预测和分析, 既很好的将文本型数据转化为数值型数据又不失对文本内容的代表性.

3 消费者情感极性的预测模型

3.1 Lasso-Logistic 预测模型

从表 2 的关键词词频矩阵可以知道, 解释变量具有明显的高维性和稀疏性的特点. 由于关键词数目过多, 且大部分数据为 0, 为解决自变量矩阵中存在的奇异问题, 必须首先对数据进行降维, 这也是将 Lasso 算法运用到 Logistic 回归和主成分回归的根本出发点.

Lasso 算法加入的惩罚项为 L1 范数, 即参数向量中各个元素绝对值之和, 由两部分构成, 一部分为 Logistic 回归的负对数似然函数, 另一部分为 L1-正则项, Lasso 的目的是求得使 $f(\beta)$ 最小的解, 即式 (4) 所示.

$$\hat{\beta}=\arg\min_{\beta}\left\{\sum_{i=1}^{n}\log\left(1+\mathrm{e}^{-y_i x_i^{\mathrm{T}}\beta}\right)+\lambda\sum_{j}^{p}|\beta_j|\right\} \tag{4}$$

关于 λ 的选择, 这里运用的是 10 折交叉验证的方法[23−25]. 本文借用 R 软件中软件包 glmnet 来实现

Lasso 算法, 如需程序代码, 可向作者索取.

由于 Lasso 算法中 λ 的取值具有一定程度的随机性, 因此每次提取出的关键词的个数并不相同, 为了不影响预测结果, 经过几次实验发现, 在提取大于 1000 个关键词时 Lasso 的稀疏解的个数没有明显增加, 所以最终提取了 TF-IDF 值较大的前 1000 个关键词词频作为初始解释变量. 由 Lasso 算法运用 (4) 式得到稀疏解, 然后将得到的系数不为 0 的关键词提取出来, 作为最终 Logistic 回归的解释变量对消费者的情感极性进行预测. 模型预测效果见表 3 和图 1.

表 3 Lasso-Logistic 预测效果混淆矩阵

		真实值		
		1(正类)	0(负类)	合并
预测值	1(正类)	54(TP)	28(FP)	82
	0(负类)	28(FN)	50(TN)	78
	合并	82	78	160

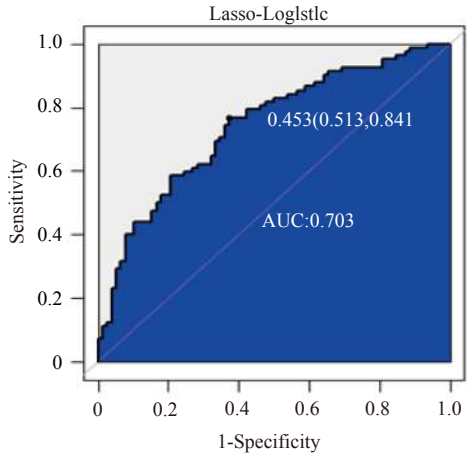

图 1 Lasso-Logistic 预测结果的 ROC 曲线

表 3 列出了在分类阈值设为 0.5 时由 Lasso-Logistic 模型得到的测试集预测结果的混淆矩阵, 根据混淆矩阵可以得到, 模型对测试集预测精度为 65%; 同时由表 3 纵向比较结果可以得出, 预测结果的敏感性 (True Positive Rate, TPR) 和特异性 (False Positive Rate, FPR) 分别为 0.66 和 0.36, 分别刻画的是正确预测为正类占真实值中正类的比例、分类器错认为正类的负实例占所有负实例的比例, 如式 (5), (6) 所示.

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

$$FPR = \frac{FP}{FP+TN} \tag{6}$$

为了更好的体现出模型的预测效果, 这里采用 ROC 曲线下面积 (AUC) 来评价模型的预测效果. 从图 1 中可以看出, 模型预测得到的 AUC 值为 0.703, Lasso-Logistic 预测方法在基于文本评论的消费者情感极性的分析上是有效的.

3.2 Lasso-PCA 预测模型

主成分分析 (Principle Components Analysis, PCA) 可以直接通过矩阵的奇异值分解 (PMD) 得到, 如式 (7) 所示. 具体来说, 是通过对原始变量进行一个基的变换, 实现变量的重新组合, 组合后得到的 p 个新的变量称为主成分, 前 r ($r<p$) 个主成分携带了原始变量 X 的主要信息. 主成分分析的优良特性使其在数据降维方面得到的广泛的应用, 然而在高维数据, 尤其是稀疏的高维数据的情况下, 传统的主成分分析的求解受到挑战. 因此, 本文借鉴文献[22]提出的 SPC 方法, 通过对 V 施加惩罚, L1-正则项, 运用 PMD(·, L1) 来实现高维稀疏矩阵的主成分分析[22].

$$\mathbf{X}=\mathbf{UDV}^{\mathrm{T}},\mathbf{U}^{\mathrm{T}}\mathbf{U}=\mathbf{I}_n,\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{I}_p,d_1\geq d_2\geq\cdots\geq d_p>0 \tag{7}$$

$$\begin{aligned}&\text{maximize}_{u_k,v_k}\, u_k^{\mathrm{T}}\mathbf{X}v_k\\ &\text{s.t. } \|v_k\|_1\leq c,\|u_k\|_2^2\leq 1,\|v_k\|_2^2\leq 1,u_k\perp u_1,\cdots,u_{k-1}\end{aligned} \tag{8}$$

R 软件提供的 PMA 软件包提供了很好的分析工具. 为了使模型具有可比性, 这里的主成分分析沿用上文中 Lasso-Logistic 预测模型抽取的测试集和训练集, 选取与 Lasso-Logistic 预测模型的变量相同数目的主成分, 将 Lasso-PCA 得到的稀疏主成分作为解释变量, 运用 Logistic 回归对消费者的情感极性进行预测, 模型预测效果如表 4 和图 2 所示.

表 4 Lasso-PCA 预测效果混淆矩阵

		真实值		
		1(正类)	0(负类)	合并
预测值	1(正类)	58	28	86
	0(负类)	24	50	74
	合并	82	78	160

同样地, 根据模型的预测结果可以得到 Lasso-PCA 对测试集预测混淆矩阵, 如表 4 所示. 根据表 4 可以得到, 模型对测试集样本预测的正确率为 67.5%, 灵敏性和特异性分别为 0.71 和 0.36. 因此, 和 Lasso-Logistic 模型相比, Lasso-PCA 模型具有更高的预测精度和灵敏性. 仍然采用 ROC 曲线来进一步评价模型的预测结果, 如图 2 所示. 本次抽样得到的 Lasso-PCA 模型的 AUC 值为 0.742, 略高于 Lasso-Logistic 模型的 AUC 值 0.703. 综合以上分析来看, Lasso-PCA 模型对

基于文本评论的消费者情感极性的预测是有效的, 并且初步判断 Lasso-PCA 模型比 Lasso-Logistic 模型具有更好的预测效果.

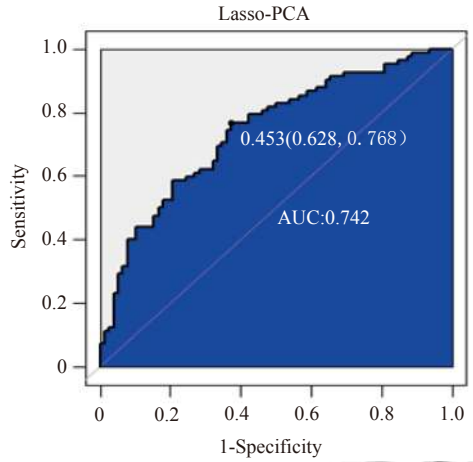

图 2 Lasso-PCA 预测结果的 ROC 曲线

3.3 两种预测模型的比较

以上关于 Lasso-Logistic 模型和 Lasso-PCA 模型的比较建立在一次抽样的基础上, 由于每次抽样都是随机的, 因此以上关于模型的比较也具有一定的随机性, 为了更好的比较两个模型的预测效果, 本文对以上研究过程重复 100 次, 分别得到 Lasso-Logistic 模型和 Lasso-PCA 模型的 100 个 AUC 值, 比较结果如图 3 所示.

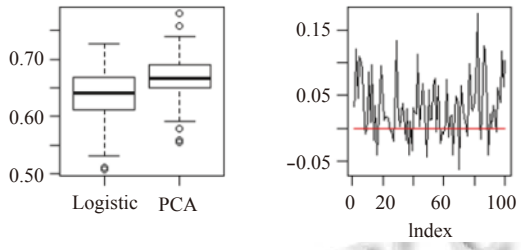

图 3 两种预测模型预测效果比较

图 3 呈现出了 100 次实验得到的预测结果的 AUC 值. 由图 3(a) 的箱线图可以看到, Lasso-PCA 模型的预测精度的平均值略高于 Lasso-Logistic 模型预测精度的平均值, 其中 Lasso-PCA 模型得到的 AUC 的均值 0.667, Lasso-Logistic 模型得到的 AUC 的均值为 0.635. 对每次抽样的预测精度求差, 由 Lasso-PCA 模型的预测精度减去 Lasso-Logistic 模型的预测精度, 两者差值如图 3(b) 所示. 二者差值虽然在 0 的两侧都有分布, 但上侧明显高于下侧且上侧的值的分布更密集, 说明 Lasso-PCA 模型的预测精度相对高于 Lasso-Logistic 模型的预测精度.

4 消费者情感极性的影响因素分析

文本评论呈现了消费者对消费行为较为细致的评价, 也是评分的根本依据, 主要由评价对象和对评价对象的情感倾向两部分组成. 从餐饮业的角度来看, 影响消费者情感极性的因素有很多, 包括味道、服务、环境、地理位置、心理预期等等. 探索这些因素是如何影响消费者评价的, 对商家提高服务质量、改善营销策略具有非常重要的意义.

Lasso-PCA 模型虽有较好的预测效果, 但模型的解释能力欠佳, 因此, 考虑到 Lasso-Logistic 模型较强的解释性, 本文借助 Lasso-Logistic 预测模型变量选择的结果进一步对影响消费者情感极性的影响因素进行分析. 由于 Lasso-Logistic 模型中由 Lasso 算法得到的稀疏解具有一定的随机性, 本文进行了两次回归以减小随机性对结果的影响. 这里主要关注回归结果中显著的变量, 结果如表 5 所示.

表 5 两个 Lasso-Logistic 回归的结果

回归一			回归二		
变量	系数	标准误	变量	系数	标准误
常数项	–0.325*	0.183	常数项	–0.397**	0.171
不错	0.615***	0.142	不错	0.557***	0.134
喜欢	0.560***	0.183	喜欢	0.580***	0.173
好吃	0.456**	0.188	好吃	0.462***	0.172
非常	1.208***	0.306	非常	1.082***	0.289
最好	2.078**	0.945	最好	1.862**	0.874
下次	1.409***	0.513	下次	1.102**	0.445
值得	1.518***	0.553	值得	1.291**	0.510
川菜馆	0.950**	0.402	川菜馆	0.836**	0.388
必点	1.892**	0.828	必点	1.640**	0.803
地道	0.961***	0.370	地道	0.829**	0.343
中规中矩	–2.053***	0.790	中规中矩	–1.876**	0.776
态度	–1.609**	0.759	态度	–1.577**	0.734
半天	–2.783**	1.255	半天	–2.810**	1.199
电梯	–2.308*	1.231	电梯	–2.434**	1.212
一次	–0.789*	0.441	一次	–0.711*	0.418
没有	–0.514**	0.204	没有	–0.537***	0.195
知道	–1.085**	0.473	知道	–0.982**	0.447
可能	–1.008**	0.462	可能	–0.905**	0.442
不会	–1.797**	0.709	不会	–1.685**	0.658
还算	–1.683**	0.673	还算	–1.491**	0.648
印象	–1.936***	0.734	印象	–1.561**	0.635
感觉	–0.459***	0.173	感觉	–0.379**	0.162
还行	–0.627*	0.368	每次	0.747**	0.348
辣味	2.009*	1.197	昏暗	–2.116*	1.187
广义似然比检验	0.000***		广义似然比检验	0.000***	

Signif. codes: '***' 0.01, '*' 0.05, '*' 0.1

表 5 呈现出了回归结果中显著变量的相关信息, 从表中可以看出, 两次回归得到的显著性变量存在很大的相似性. 两次回归都得到了 25 个显著变量, 其中有 23 个变量在两个回归结果中同时显著. 此外, 从回归系数可以看出, 所有在两次回归中同时显著的变量具有相同的正负号, 且系数大小相差不大, 说明模型具有很好的稳健性. 这些显著的特征词或特征项隐含了影响消费者情感极性的重要因素, 按照属性不同可以将其分成 6 类, 如表 6 所示.

表 6 显著的回归变量分类

属性	变量名称	影响方向
积极情感词汇	不错、喜欢、好吃、非常、最好	正向
中性情感词汇	中规中矩、还算、还行、感觉、可能	负向
消极情感词汇	没有、不会、一次	负向
体现“特色”的词汇	川菜馆、辣味、必点、地道	正向
体现“服务和环境”的词汇	态度、半天、电梯、昏暗、印象	负向
体现“潜在消费”的词汇	下次、值得、每次	正向

三类带有情感倾向的词汇和三类表示特征属性的词汇分别从不同角度体现了消费者情感极性. 从总体上来说, 带有情感色彩的词汇最能直观地表达消费者的情绪; 虽然影响餐饮业消费者情感极性的因素众多, 但是餐厅“特色”、“服务和环境”却是消费者最为关注的; 通过关注含有“下次”、“值得”和“每次”的评论, 可以有效识别潜在消费能力. 具体地, 从以下 5 个方面进行分析.

(1) 从两次回归结果中可以看出, “不错”、“喜欢”、“好吃”以及程度副词“非常”和“最好”的系数在两个回归中的系数都显著为正. 相比之下, “没有”、“不会”和“一次”这类含有负面情绪的词汇, 回归系数显著为负. 这一结果也是符合常理的, 好的评价对应高的评分; 而对于没有达到满意的消费行为, 消费者往往对不满意之处吐槽, 评分自然也低.

(2) “中规中矩”、“还算”和第一个回归中“还行”的系数显著为负, 说明评论中出现“中规中矩”、“还算”这两个词汇的消费者对消费行为更加倾向于持负面的态度, 服务中的美中不足之处很容易引起消费者的消极情绪. 同时, 这一结论对商家也具有一定的警醒作用, 商家应该对此类评论加以重视, 根据评论内容分析对应消费者的消费心理, 扑捉到自身服务的欠缺之处, 如果能够弥补美中不足之处可能就会带来意想不到的利润.

(3) “必点”的系数在两个回归中的结果都显著为正, 体现出了消费者对某个菜品的青睐; “地道”和“川菜馆”在两个回归结果中显著为正, “辣味”也在回归一中显著为正, 体现出了餐厅的独特之处. 这些都是最能体现出一个餐厅特色的词汇, 系数显著为正的回归结果说明餐厅特色菜是影响消费者评价的一个关键因素, 说明餐饮业的商家在经营过程中要有能力打造出自己的特色, 并且注重招牌菜的推广, 这在很大程度上有利于餐厅的经营, 从而提升自身的市场竞争力.

(4) “态度”、“半天”、“电梯”以及第二个回归中“昏暗”的系数显著为负, 说明服务态度和环境的好坏直接影响了消费者的心理, 强调了餐厅服务态度和外部环境特征的重要性. 现代人的消费观念不断转换, 对服务的要求也随之提高, 更是体现在方方面面. 好的服务态度和就餐环境给消费者更加舒适、放松的感觉, 直接影响消费者的情绪, 对消费者的评分起到重要作用.

(5) “下次”、“值得”和“每次”的回归系数显著为正, 体现出了顾客再次消费的潜质, 说明这类消费者对消费行为的整体评价较高, 再次消费的可能性很大. 商家为提高顾客忠诚度、改善经营状况, 要时常关注这类消费者的消费动向, 注意维护此类消费者的顾客忠诚度.

5 结论及启示

本文将 Lasso 算法运用到网络评论的文本分析中, 首先建立了 Lasso-Logistic 和 Lasso-PCA 两个模型对消费者情感极性进行预测. 由分析结果可知, 两种预测模型都取得了一定的预测效果. 根据 100 次随机抽样结果, Lasso-PCA 预测模型的 AUC 平均值达到 0.67, 而 Lasso-Logistic 预测模型的 AUC 平均值为 0.64. 相比之下, Lasso-PCA 模型整合了更多的变量信息, 对文本的情感极性具有更好的预测效果; 但是 Lasso-PCA 模型对变量的解释能力较弱, 尤其在解释变量维度较高的情况下, Lasso-PCA 模型很难分析出解释变量对被解释变量的影响. 因此, 文中第 4 节借助 Lasso-Logistic 模型分析了影响消费者情感极性的显著性因素作为补充分析. 结果表明, 餐厅特色、餐厅的服务态度和外部环境等是影响消费者情感极性的主要因素. 另外, “中规中矩”和“还算”两个特征项的系数显著为负也反映了消费者对消费行为的高标准、高期望, 即使在市场逐渐细分的大环境下, 商家要想维护顾客忠诚度以长期生存下去, 也必须根据市场要求不断完善自己.

参考文献

1 李胜宇, 高俊波, 许莉莉. 面向酒店评论的情感分析模型. 计算机系统应用, 2017, 26(1): 227–231. [doi: 10.15888/j.cnki.csa.005511]

2 Clavel C, Callejas Z. Sentiment analysis: From opinion mining to human-agent interaction. IEEE Transactions on Affective Computing, 2016, 7(1): 74–93. [doi: 10.1109/TAFFC.2015.2444846]

3 Zheng LJ, Wang HW. Sentimental polarity and strength of online cellphone reviews based on sentiment ontology. Journal of Industrial Engineering and Engineering Management, 2017, 31(2): 47–54.

4 曾津, 周建军. 高维数据变量选择方法综述. 数理统计与管理, 2017, 36(4): 678–692.

5 Weiser M. The computer for the 21st century. IEEE Pervasive Computing, 2002, 1(1): 19–25. [doi: 10.1109/MPRV.2002.993141]

6 Aizawa A. An information-theoretic perspective of TF-IDF measures. Information Processing & Management, 2003, 39(1): 45–65.

7 Turney PD. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, PA, USA. 2002. 417–424.

8 Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA. 2004. 168–177.

9 覃世安, 李法运. 文本分类中 TF-IDF 方法的改进研究. 现代图书情报技术, 2013, (10): 27–30. [doi: 10.11925/infotech.1003-3513.2013.10.05]

10 Mascolo C, Capra L, Zachariadis S, *et al*. XMIDDLE: A data-sharing middleware for mobile computing. Wireless Personal Communications, 2002, 21(1): 77–103. [doi: 10.1023/A:1015584805733]

11 Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1–47. [doi: 10.1145/505282.505283]

12 Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP. Philadelphia, PA, USA. 2002. 79–86.

13 王健, 张俊妮. 统计模型在中文文本挖掘中的应用. 数理统计与管理, 2017, 36(4): 609–619.

14 Geladi P, Kowalski BR. Partial least-squares regression: A tutorial. Analytica Chimica Acta, 1986, 185: 1–17. [doi: 10.1016/0003-2670(86)80028-9]

15 曹芳, 朱永忠. 基于多重共线性的 Lasso 方法. 江南大学学报 (自然科学版), 2012, 11(1): 87–90.

16 方匡南, 章贵军, 张惠颖. 基于 Lasso-Logistic 模型的个人信用风险预警方法. 数量经济技术经济研究, 2014, 31(2): 125–136.

17 倪新洁, 梁彪, 倪佩可. 结合 LASSO 算法与 logistic 回归模型的 P2P 信贷审批结果研究. 统计与管理, 2015, (8): 44–47.

18 吴方照, 王丙坤, 黄永峰. 基于文本和社交语境的微博数据情感分类. 清华大学学报 (自然科学版), 2014, 54(10): 1373–1376, 1383.

19 郑文斌. 基于正则化线性统计模型的文本分类研究[博士学位论文]. 杭州: 浙江大学, 2012.

20 Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics, 2006, 15(2): 265–286. [doi: 10.1198/106186006X113430]

21 Jolliffe IT. A note on the use of principal components in regression. Applied Statistics, 1982, 31(3): 300–303. [doi: 10.2307/2348005]

22 Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 2009, 10(3): 515–534. [doi: 10.1093/biostatistics/kxp008]

23 胡局新, 张功杰. 基于 K 折交叉验证的选择性集成分类算法. 科技通报, 2013, 29(12): 115–117. [doi: 10.3969/j.issn.1001-7119.2013.12.039]

24 王运生, 谢丙炎, 万方浩, 等. ROC 曲线分析在评价入侵物种分布模型中的应用. 生物多样性, 2007, 15(4): 365–372.

25 邹洪侠, 秦锋, 程泽凯, 等. 二类分类器的 ROC 曲线生成算法. 计算机技术与发展, 2009, 19(6): 109–112.