[文章编号] 1003－4684(2026)01-0132-06

基于新能源汽车在线评论的文本挖掘与情感分析

郑 列，兰海宇

（湖北工业大学理学院，湖北 武汉 430068）

[摘 要] 爬取汽车之家、易车网、爱卡汽车网的新能源汽车用户评论，对评论数据进行数据清洗、文本分词、去停用词等预处理后，采用词云图和语义网络模型对新能源汽车在线评论的文本特征进行可视化；运用 SnowNLP 对已标注的评论进行训练，在测试集上准确率达到 94.2%；进一步对未标注的外观、空间、内饰、续航、动力、操控、舒适性、性价比、智能化九个维度进行情感分析；利用 LDA 主题模型分别对外观、空间的正向评论和内饰、舒适性的负向评论进行分析，以期为新能源汽车厂商改进产品性能和提升服务质量提供参考。

[关键词] 新能源汽车；在线评论；文本挖掘；SnowNLP 情感分析；LDA 模型

[中图分类号] TP391.1；O212.8 [文献标识码] A

为进一步落实“绿色”发展理念，实现国家“双碳”目标，扩大内需，推动我国经济高质量发展，国家出台了相关政策来支持新能源汽车的发展，新能源汽车市场环境也随之得到优化，新能源汽车市场消费活力亦被激活[1-3]。目前市场上已经存在相当规模的新能源汽车消费者，产生了海量的在线评论数据。在线评论中含有大量有效信息，能够体现消费者对商品的情感倾向，同时也是消费者进行购买决策的参考信息[4-5]。通过对消费者的在线评论进行情感分析[6-8]，可以得到消费者偏好、购买缘由等，从而使商家有针对性地提升商品性能和服务质量，吸引更多消费者。因此，本文基于在线评论分析新能源汽车用户的满意度，探究新能源汽车用户在不同评论维度下的情感倾向，为提升和改善新能源汽车产品性能提供参考。

1 数据获取与处理

本文选择汽车之家、易车网、爱卡汽车网作为数据源，利用网络爬虫技术获取新能源汽车用户在线评论并对其进行数据预处理。

1.1 数据获取

本文利用八爪鱼采集器对新能源汽车在线评论进行采集，选取每个网站的新能源汽车热销榜前十的车型作为研究样本，采集的原始数据包括用户名、评论时间、汽车系列名、购车时间、购车地点、购车价格以及具体评论内容等。最终获得在线评论文本共 215 153 条，其中最满意维度的评论 23 441 条，最不满意维度的评论 23 441 条，外观、空间、内饰、续航、舒适性、动力、操控、性价比及智能化维度的评论共 168 271 条。

1.2 数据预处理

数据预处理是分析文本信息的重要步骤，删除重复、无意义的文本信息，有助于提高文本分析的精确度，保证文本信息的质量，为后续的文本分析奠定基础。对新能源汽车原始评论进行文本去重、压缩去词、短句删除，处理后得到有效评论共 213 203 条。为了减小模型复杂度、提高训练效率，构建停用词库过滤掉出现频率高但并无明确意义的词语，比如副词、语气助词等；评论文本中含有新能源汽车行业的专有名词，故引入自定义词库，进一步提高分词的准确性。

2 新能源汽车在线评论的文本特征

2.1 基于词频统计的文本特征可视化分析

高频词能在一定程度上体现消费者对新能源汽车各个维度的关注情况。表 1 展示了前 20 个高频词汇及其频数。评论数据中“外观”“设计”“颜值”等高频词展示消费者对新能源汽车外观颜值的关注；“空间”“内饰”“座椅”“后排”“噪音”“隔音”反映了消费者比较关注新能源汽车内部的舒适性和美观度；“动力”“续航”“充电”“油耗”“公里”等词说明消费者重视新能源汽车在动力续航、充电设施建设方面的

[收稿日期] 2023－12－17

[基金项目] 教育部人文社会科学研究规划基金项目(17YJA790098)

[第一作者] 郑 列(1963－)，男，湖北英山人，湖北工业大学教授，研究方向为应用数学。

[通信作者] 兰海宇(2000－)，女，湖北谷城人，湖北工业大学硕士研究生，研究方向为数据挖掘与机器学习。

问题；同时，价格、提车时间也是消费者重点考虑的因素。

表 1　高频词汇统计表

词语	频数	词语	频数	词语	频数
外观	11 086	座椅	5683	功能	3655
空间	10 708	配置	5604	颜值	3629
动力	8495	油耗	5382	后排	3507
续航	8273	价格	4486	提车	3357
内饰	7307	公里	4236	隔音	3277
设计	6883	整体	4075	噪音	3261
充电	6782	用车	3853		

绘制用户最满意和最不满意评论的词云图，可以直观地展示出新能源汽车的优缺点。从图 1 可以看出，“空间”“外观”“动力”“设计”“续航”“配置”“充电”“颜值”等词出现的频率最高，从而可以看出新能源汽车用户对于外观颜值、空间设计、充电续航等方面满意度较高；“价格”“成本”等特征词说明用户对新能源汽车的性价比比较满意。



图 1　“最满意”维度评论词云图



图 2　“最不满意”维度评论词云图

新能源汽车用户最不满意的词频统计中，“内饰”“续航”“座椅”“充电”“提车”“时间”等词出现的频率较高，可以看出用户对于新能源汽车的内饰设计、座椅质量、隔音效果、提车时间等方面不太满意，对于充电续航方面仍然存在担忧；其次，“后排”“空间”“底盘”等词反映了新能源汽车后排空间不足以及底盘太低的问题；“味道”“异味”等词反映了新车存在异味；“噪音”“声音”“异响”“隔音效果”等词说明了新能源汽车存在噪音问题；“价格”“优惠”等词说明新能源汽车用户对于购车价格以及优惠力度还有很大期待。

2.2　基于语义网络的用户情绪分析

通过词云图分析能够直观地看到用户对新能源汽车的关注重点，但很难看出这些关注点之间的内在联系。而语义网络分析可以看出高频词之间的联系，挖掘出潜在的价值信息[9-10]。本文利用 ROSTCM6 软件绘制最满意和最不满意评论的语义网络图。图 3 展示了在线评论中最满意维度评论的语义网络，可以看出“外观”“动力”“设计”“续航”“配置”等词周围的连线比较密集，是重要节点；与“外观”密切联系的特征词有“好看”“漂亮”“颜值”“吸引”“时尚”“造型”“整体”“大气”，说明新能源汽车外观设计整体较好，受到消费者喜欢；“设计”还与“时尚”“造型”“内饰”“座椅”相连，说明消费者关注新能源汽车整体造型时尚度的同时，对内饰及座椅等细节有一定要求；“动力”与“加速”“体验”“操控”“加速”之间的网络关系，说明消费者对于新能源汽车的动力比较满意，认可其在提速方面的性能。

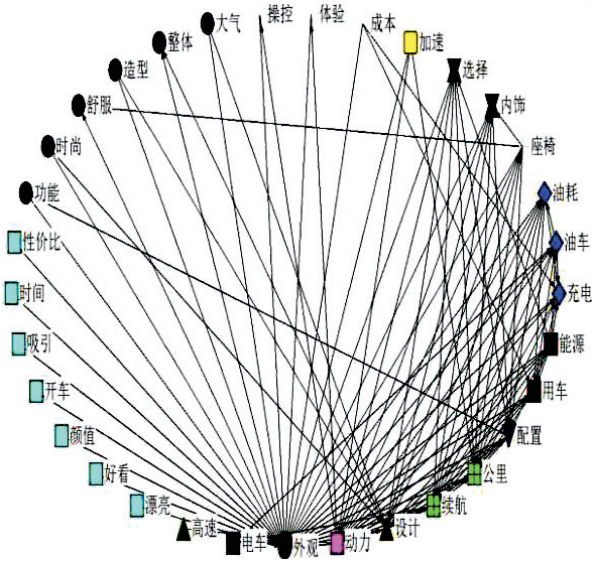

图 3　“最满意”评论的语义网络图

从最不满意评论的语义网络图（图 4）可以看出，“内饰”“座椅”“时间”“充电”等词周围的连线比较密集，说明消费者对这几个方面不满意；与“内饰”密切联系的特征词有“颜色”“风格”“用料”“材质”“简单”“做工”“塑料”，说明新能源汽车内饰用料简单、做工粗糙、颜色选择少、风格单调；“通风”“异味”“味道”之间的连线说明新能源汽车通风效果不好，内饰用料不好，导致新车存在异味；“座椅”与“后排”“功能”“开车”“调节”“舒适性”等词之间的网络关系，说明新能源汽车座椅偏硬，调节不便，后排座椅空间小，无法满足乘车人的舒适性要求；“噪音”“声音”“隔音”“噪声”之间的连线说明新能源汽车在隔音和降噪方面还有待提高；“时间”与“用车”“等待”

“销售”“优惠”之间的连线，反映消费者在销售服务、提车时间、优惠力度等方面的不满。

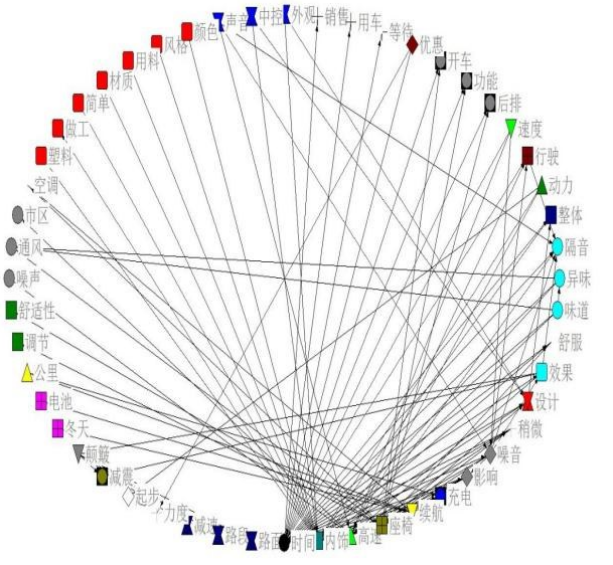

图 4 “最不满意”评论的语义网络图

3 新能源汽车在线评论的情感分类与预测

3.1 基于 SnowNLP 的情感分析

SnowNLP 情感分析[11]的基本原理如下：假设情感分析的分类有正面评价 c_1 和负面评价 c_2 两类，且每一条评论中有 $w_1,w_2,\cdots,w_n$ 共 n 个相互独立的文本空间，通过朴素贝叶斯公式分别计算出属于正面评价 $P(c_1 \mid w_1,w_2,\cdots,w_n)$ 和负面评价 $P(c_2 \mid w_1,w_2,\cdots,w_n)$ 的条件概率，计算得到概率值的范围在 0 到 1 之间，当概率值越接近 1 表示正向情感，当概率值越接近 0 表示负向情感。其计算公式可表示为：

$$P(c_1 \mid w_1,w_2,\cdots,w_n)=\frac{P(w_1,w_2,\cdots,w_n \mid c_1)}{P(w_1,w_2,\cdots,w_n)}=\frac{P(w_1,w_2,\cdots,w_n \mid c_1)}{P(w_1,w_2,\cdots,w_n \mid c_1)P(c_1)+P(w_1,w_2,\cdots,w_n \mid c_2)P(c_2)}$$

直接运用 SnowNLP 进行情感分析，发现准确率仅为 0.513，这说明在实际使用的过程中需要训练适用于新能源汽车的语料库。将“最满意”维度的评论标注为正向评论，“最不满意”维度的评论则视为负向评论，并按照 7∶3 的比例划分训练集和测试集。运用 SnowNLP 对标注好的评论进行模型训练，并用测试集对模型进行验证，分析该情感分类器的效果。加入正向情感文本和负向情感文本训练后，SnowNLP 情感分析的准确率为 0.673。观察数据后发现，负向情感文本存在类似“提车时间没多久，目前暂无不满意”等无意义的中性评论，所以对正向和负向的语料库数据进一步整理和丰富，最终情感分析准确率达到 0.942。

3.2 基于 SnowNLP 的评论情感预测

对未标注的外观、空间、内饰、续航、舒适性、动力、操控、性价比、智能化九个维度的评论文本进行向量化训练，再运用 SnowNLP 进行情感预测，得到每条评论的情感倾向。

3.2.1 不同维度评论的情感预测 SnowNLP 对九个维度的情感预测结果由图 5 所示。其中，外观和空间维度的正向情感评论数量最多，且正向情感占比达到 94%以上，可以看出消费者对于新能源汽车的外观和空间评价较高。新能源汽车往往采用流线型车身、动感十足的前脸设计、创新的座椅布局和储物设计，使得新能源汽车在外观和空间方面获得消费者青睐。续航和性价比的正向评论占比达到 88%以上，动力和智能化的正向评论占比紧随其后，说明消费者对于新能源汽车的续航、性价比、动力和智能化方面也比较满意。内饰、舒适性和操控维度的负向评论数量较多，其中舒适性维度的负向评论占比达到 70%。目前，新能源汽车减震效果不佳，并且在高速行驶时产生的风噪和轮胎噪音，一定程度上会影响乘坐舒适度。

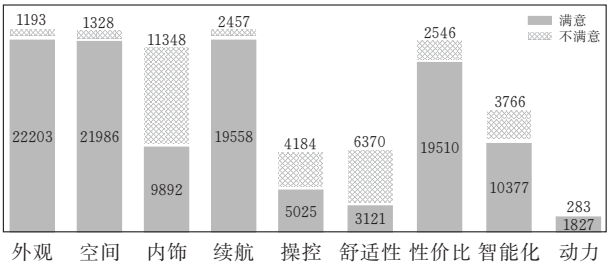

图 5 不同维度评论的情感预测结果

3.2.2 不同价位车型评论的情感预测 不同价位的新能源汽车在外观、空间、内饰、续航、舒适性、动力、操控、性价比、智能化维度上的正向情感占比情况如下。

通过对比数据，可以看出用户对于 5 万元以下的新能源汽车情感较为极端，由于价格原因，牺牲了一定程度的驾驶性能和乘车体验，导致消费者在空间、内饰、续航、舒适性、智能化、动力维度均对 5 万元以下的新能源汽车最为不满；同样也是因为价格的优惠，其在性价比方面令消费者最满意。30 万元以上的新能源汽车在内饰、舒适性、动力维度获得好评，但并没有达到消费者对其在外观和操控方面的预期，缺乏性价比。在空间、续航、智能化方面正向情感率最高的均为 10～20 万元的新能源汽车。而 5～10 万元、20～30 万元的新能源汽车在各个方面都表现得中规中矩。在内饰、空间和智能化方面，高价位新能源汽车的正向评论占比相对较高；在性价

比方面，低价位新能源汽车的正向评论占比相对较高；在舒适性和动力方面，价格越高，正向评论占比也越高；在操控方面，反而是价格越低，操控起来越顺手。高价位车型通常配备更先进的技术，空间上更为宽敞，内饰也更加精致，从而为消费者带来更为优质的体验。相比之下，低价位车型往往只提供基础功能，缺乏智能化功能，因此在智能化上不尽如人意。

表 2　不同价位车型评论的情感预测结果　%

价位	外观	空间	内饰	续航	操控	舒适性	性价比	智能化	动力
5 万以下	95.7	91.4	35.9	78.3	57.1	24.7	92.4	60.4	71.7
5－10 万	94.5	92.8	44.6	87.3	55.2	27.5	88.6	62.4	82.9
10－20 万	95.4	94.9	43.3	91.5	54.1	32.9	90.2	76.9	88.1
20－30 万	94.3	94.3	47.7	89.7	52.6	34.6	87.0	72.8	88.4
30 万以上	93.3	93.7	48.8	87.5	51.9	46.7	84.3	71.1	89.4

4　基于 LDA 主题模型的新能源汽车性能分析

由上文分析可知，新能源汽车用户对于外观、空间这两个维度的表现最为认可，但是对于新能源汽车在内饰和舒适性这两个方面普遍不太满意。为了进一步获取用户对于新能源汽车在各维度的关注点，本文针对外观、空间的正向评论和内饰、舒适性的负向评论分别进行 LDA 主题模型分析。

4.1　LDA 主题模型

LDA 是由 Blei 等[12]于 2003 年提出，是一种基于贝叶斯思想的无监督的聚类算法，可以用来识别大规模文档集或语料库中潜在的主题并对其进行分类。通过构建“文档－主题－词”的三层贝叶斯概率模型，可以获得每个主题下词语的分布概率以及每个文档对应主题的概率，根据这些概率就可以对文档进行主题分类[13]。

4.2　外观维度正向评论的 LDA 主题分析

由用户对新能源汽车外观维度评论的主题－Coherence 变化情况（图 6）可知，当主题数为 3 时，主题模型效果最佳。因此，对外观维度的正向评论提取 3 个主题，每个主题对应的前 10 个特征词如表 3 所示。

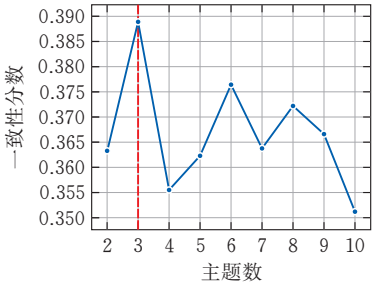

图 6　外观维度正向评论的主题数与一致性分数的关系

由表 3 可以看出，主题 1 反映了新能源汽车小巧精致，车漆颜色丰富、做工精细；主题 2 反映了新能源汽车造型独特，辨识度极高，侧面轮廓充满动感，LED 灯光增添了科技感；主题 3 反映了新能源汽车颜值高，线条设计流畅，整体造型融合了时尚、大气、运动的风格。由此可见，用户对于新能源汽车外观的满意点为颜色丰富、造型独特、设计精巧。

表 3　外观维度正向评论的主题分析

主题	外观特征词
1	好看 可爱 白色 颜色 做工 黑色 颜值 车漆 小巧 灰色
2	贯穿 造型 前脸 车灯 侧面 线条 晚上 辨识 霸气 LED
3	时尚 线条 大气 造型 整体 设计 颜值 流畅 运动 风格

4.3　空间维度正向评论的 LDA 主题分析

对空间维度的正向评论提取 2 个主题，每个主题提取 10 个特征词。由表 4 可以看出，主题 1 反映了新能源汽车具有宽敞舒适的乘坐环境，且可以满足上下班、接送孩子的代步需求；主题 2 反映了新能源汽车的空间设计也很出色，不仅座位空间很大，乘坐起来非常舒适，而且后备箱的容积也足够装载东西。由此可见，用户对于新能源汽车空间维度的满意点为宽敞舒适、储物充足。

表 4　空间维度正向评论的主题分析

主题	空间特征词
1	孩子 上下班 宽敞 家里 够用 舒服 代步 开车 后排 小孩
2	后排 储物 东西 第三排 容积 后备箱 设计 乘坐 车内 很大

4.4　内饰维度负向评论的 LDA 主题分析

对内饰维度的负向评论提取 3 个主题，每个主题对应的前 10 个特征词如表 5 所示。由表 5 可以看出，主题 1 反映了新能源汽车内饰设计简约，但整体做工欠佳，中控质感较差，颜色搭配不够美观，用料材质也不够精细，存在异味，给人整体感觉不够高端；主题 2 反映了新能源汽车座椅的颜色选择有限并且不耐脏，空调效果和通风效果都不够理想，加之座椅皮质不好，导致车内的味道不太好，让用户感到不适；主题 3 反映了车内的功能配置也不尽如人意，

如旋转屏幕在操作时不够方便，车机系统的控制面板也不够直观，导航系统的屏幕较小，给驾驶者查看地图带来一定的困难。整体来说，车内的配置相对较低，大屏显示效果也不尽如人意。由此可见，用户对于新能源汽车内饰维度的评价为质感欠佳、配置落后。

表5 内饰维度负向评论的主题分析

主题	内饰特征词
1	设计 用料 简约 做工 整体 材质 中控 质感 颜色 异味
2	座椅 皮质 颜色 空调 效果 通风 耐脏 加热 味道 后排
3	功能 语言 旋转 屏幕 控制 导航 英寸 车机 配置 大屏

4.5 舒适性维度负向评论的LDA主题分析

对舒适性维度的负向评论进行分析，主题1反映了新能源汽车座椅舒适性不足，体现在其包裹性不够好，支撑性能不够理想，调节范围有限；主题2反映了新能源汽车的空调系统在舒适性方面还有很大的提升空间，主要体现在空调系统的控制效果不理想，过滤功能欠佳，加热出风口的效果也不尽如人意；主题3反映了新能源汽车的隔音效果不理想，尤其在高速行驶时，胎噪和风噪都较为明显，给驾驶者和乘客带来不舒适的感受；同时，车辆在颠簸路段的减震效果也不够理想，使得车内人员感受到明显的震动和颠簸，影响了行驶的舒适度。

表6 舒适性维度负向评论的主题分析

主题	舒适性特征词
1	包裹 后排 空间 舒适 乘坐 座椅 支撑 调节 前排 控制
2	空调 功能 控制 温度 过滤 加热 出风口 效果 自动 设计
3	噪音 隔音 效果 颠簸 减震 安静 胎噪 行驶 车内 高速

5 结论

本文爬取汽车之家、易车网、爱卡汽车网站上新能源汽车用户在线评论数据共215 153条，对其进行文本特征分析、SnowNLP情感分析和LDA主题模型分析，得出以下结论：

利用语义网络分析用户情绪，发现消费者对于新能源汽车的外观设计、动力续航等方面普遍表示满意，而对于内饰、座椅、提车时间、充电等方面存在较多不满；消费者关注新能源汽车颜值的同时，还要求有良好的驾驶体验感和空间舒适性。通过SnowNLP情感分析可知，消费者对于不同价位的新能源汽车给出了不同的评价情感和关注点，对于低价位的新能源汽车重点关注其外观和性价比，而对于高价位的新能源汽车则更加追求功能和配置，关注其动力续航；从不同的评论维度来看，新能源汽车用户对于外观和空间这两个维度的表现最为满意，但对内饰和舒适性这两个方面却普遍不太满意。通过LDA主题模型分析，发现消费者认为新能源汽车的外观亮点在于色彩丰富、造型独特、设计精巧；在空间维度上，消费者则满意于车内空间的宽敞与舒适以及储物空间的充足；然而，对于内饰维度，消费者评价其质感欠佳，配置相对落后；在舒适性方面，消费者认为新能源汽车的座椅舒适性不足，隔音效果不理想，空调系统的控制效果难以令人满意，减震效果亦需进一步改进。

[参 考 文 献]

[1] 孙婉晴.新时期新能源汽车行业现状与战略规划研究[J].科技经济市场，2022(09)：13-15.

[2] 李旭东，何寿奎，戴庆春，等.双碳背景下多元政策对新能源汽车消费者购买意愿影响研究[J].重庆理工大学学报(自然科学)，2023，37(11)：362-371.

[3] 张松颖，高翠，龚宇鹏，等."碳中和"背景下新能源汽车发展现状及策略研究[J].时代汽车，2023(21)：89-91.

[4] 陈小燕，张晴.在线评论情感强度对消费者购买意愿的影响研究[J].中国商论，2023(20)：82-85.

[5] 朱振中，单明辉，房晓芸."陈述"还是"修辞"：在线评论语言风格对消费者购买意愿的影响[J].财经论丛，2023(11)：92-102.

[6] 刘旺达.基于在线评论的XC酒店顾客满意度评价与提升策略研究[D].西安：西安理工大学，2023.

[7] 张成元，刘云珂，赵炳清，等.基于在线文本情感分析的"唐文化"目的地形象挖掘研究[J].计量经济学报，2023，3(02)：387-407.

[8] 张金纯.基于在线评论多属性情感分析的IPA手机改进研究[D].秦皇岛：燕山大学，2023.

[9] 史宝磊.北京冬奥会期间自媒体发布内容的语义网络分析[D].石家庄：河北经贸大学，2023.

[10] 陈翔，黄璐，倪兴兴，等.基于动态语义网络分析的主题演化路径识别研究[J].情报学报，2021，40(05)：500-512.

[11] 贺海玉.大数据下的网络评论情感分析研究与实现[J].电脑知识与技术，2023，19(18)：64-66.

[12] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.

[13] 丁申宇.基于BERT-LDA的在线评论细粒度情感分析[D].兰州：兰州财经大学，2022.

Text Mining and Sentiment Analysis Based on Online Reviews of New Energy Vehicles

ZHENG Lie, LAN Haiyu

(*School of Science, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: This paper crawled the online reviews of new energy vehicles from Autohome, Yiche.com and Aika Auto. The review data is preprocessed by data cleaning, text segmentation, and stop words removal. Word cloud map and semantic network model were used to visualize the text features of online reviews of new energy vehicles. SnowNLP was used to train the model on the annotated reviews, and the final accuracy on the test set reached 94.2%. The sentiment analysis was carried out on nine dimensions of unlabeled appearance, space, interior, endurance, power, handling, comfort, cost performance, and intelligence. Finally, the LDA topic model analysis is carried out for the positive comments of appearance and space and the negative comments of interior and comfort, expecting to provide reference suggestions to improve product performance and service quality.

Keywords: new energy vehicles; online review; text mining; SnowNLP; LDA

[责任编校: 闫　品]

(上接第 107 页)

Establishment of a TaqMan qPCR Method for Quantitative Detection of Oncolytic Virus OH2

LI Le, YANG Jingru, WANG Xinya, Xia Wen, WANG Yang, HU Han, LIU Binlei

(*School of Life Science and Health Engineering, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: A real-time quantitative PCR (qPCR) assay based on a TaqMan probe was developed. The established TaqMan qPCR method was validated for amplification linearity, sensitivity, lower limit of quantification (LLOQ), and specificity. Additionally, the extraction recovery rate of the method was verified. The results demonstrated that the method exhibited a good linear amplification relationship with an $R^2 > 0.980$. The limit of detection (LOD) was 4 copies/μL, and the LLOQ was 50 copies/μL. The TaqMan qPCR method also showed excellent specificity. The establishment of this quantitative detection method lays a foundation for the biodistribution study of oncolytic virus OH2 in clinical trial samples.

Keywords: TaqMan qPCR; oncolytic virus OH2; genome copy number; method validation

[责任编校: 张　众]