基于BERTopic的新能源汽车评价分析

——以比亚迪为例

赵成兵，汪 瑶，储立峥

（安徽建筑大学数理学院，合肥 230601）

摘要：随着新能源汽车在全球市场的快速发展，用户对产品的关注已从单一性能转向多维需求。以比亚迪七款热门车型为研究对象，基于汽车之家和懂车帝收集的27 283条用户评价，采用DeepSeekV2模型提取短文本并利用BERTopic模型进行主题建模，识别出纯电性能、刀片电池、噪声控制等核心用户关注点。研究结果表明，用户对续航能力与电池技术持积极评价，但对胎噪、风噪及新车异味等舒适性问题反馈较为负面。研究的主要贡献包括：通过DeepSeekV2模型对长文本拆分，提升主题建模对非结构化数据的处理能力；结合BM25加权的c-TF-IDF算法和MMR优化技术，在语义嵌入与层次聚类的基础上，揭示用户多维关注点及其内在关联，为新能源汽车产品设计与市场策略提供数据支持。

关键词：BERTopic；比亚迪；新能源汽车；主题建模

中图分类号：F49；U270.38 **文献标志码：**A **文章编号：**1671－1807(2025)14－0083－07

新能源汽车的快速发展在全球范围内引发了广泛关注，尤其在“双碳”发展理念的指导下，中国社会和经济发展逐步面向低碳能源转型，新能源汽车逐渐成为未来汽车行业发展的必然趋势[1]。近年来，消费者对新能源汽车的需求已从基础功能逐渐转向续航能力、智能化与舒适性等多维度特性。与此同时，大量用户评论数据的涌现为企业了解市场需求、优化产品策略提供重要资源，但其非结构化特性也增加了分析难度。

国内外学者针对新能源汽车用户评论的研究逐渐增多。在国际研究中，Reimers和Gurevych[2]提出基于深度语义嵌入技术的消费者行为分析方法，有效提高了文本分析的精度与效率；国内方面，张永安和周怡园[3]结合时间序列与主题建模技术，揭示了新能源汽车从性能导向到用户体验导向的转变。尽管如此，现有研究在精准挖掘主题间的关联性与动态演化方面仍有待提升。

以比亚迪旗下七款新能源汽车为研究对象，结合用户评论数据，利用BERTopic（BERT-based topic modeling）模型进行主题建模与分析。研究的核心目标是通过深入分析比亚迪新能源汽车领域的成功经验，探讨其科技成果如何推动企业在激烈的市场竞争中取得领先地位，同时为同行业企业提供借鉴，帮助它们在产品设计、消费者需求理解以及市场拓展方面获得更高的市场份额。

1 数据与方法

技术路线如图1所示，包括文本数据爬虫、DeepSeekV2大模型提取、BERTopic建模三个部分。

1.1 数据来源

数据来源包括懂车帝和汽车之家，分别涵盖27 283条评论，涉及比亚迪旗下王朝系列与海洋系列七款热门车型。数据采集主要集中在2021年1月至2024年8月的用户口碑，以保证数据的时效性。如表1所示。

1.2 LLM提取观点

在新能源汽车用户评论中，由于评论文本具有明显的口语化和情绪化表达，词语使用往往不够精确，传统的基于词频或关键词匹配的方法难以准确解析文本的潜在语义。因此，基于大语言模型（large language model，LLM）模型的信息提取方法，

收稿日期：2025-03-10

基金项目：安徽省高校省级自然科学基金重点项目（KJ2021A0631，2024AH050257）；安徽省高校省级人文社会科学基金（2023AH040035）

作者简介：赵成兵（1970—），男，安徽庐江人，博士，教授，研究方向为几何分析与动态系统；通信作者汪瑶（1998—），女，安徽潜山人，硕士研究生，研究方向为经济模型优化；储立铮（2000—），男，安徽岳西人，硕士研究生，研究方向为应用统计。

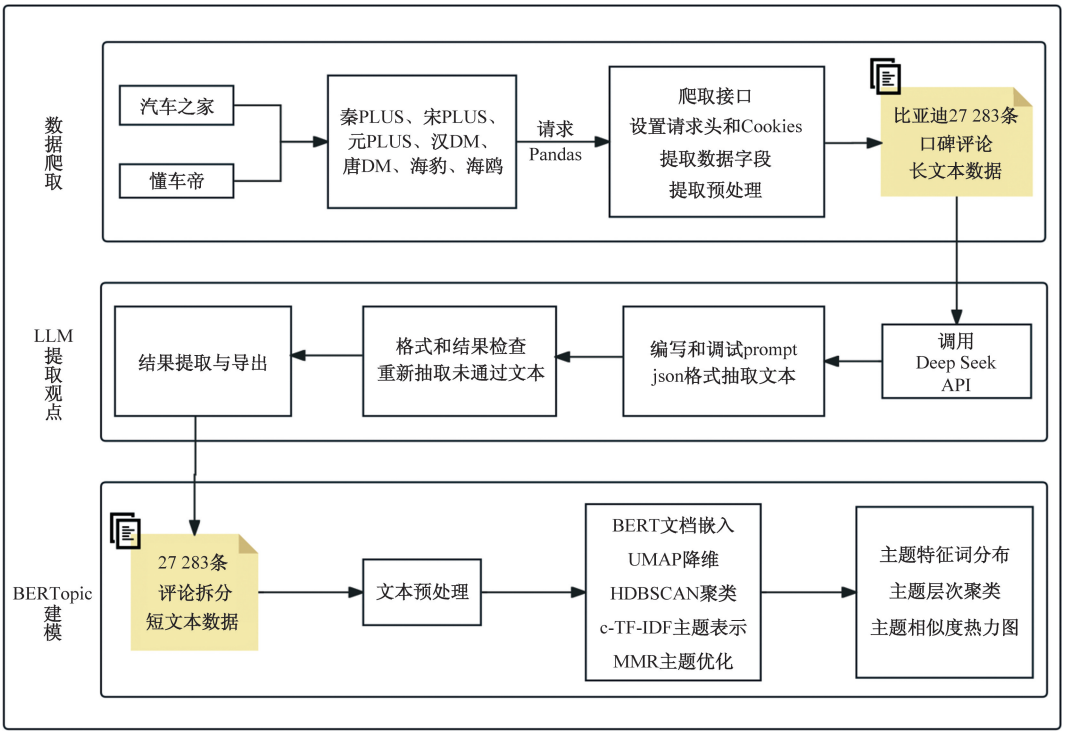

图 1　技术路线

表 1　数据来源

车型	评论/条		
	汽车之家	懂车帝	总计
宋 PLUS	3 992	2 957	6 949
秦 PLUS	3 886	2 318	6 204
汉 DM	3 982	1 807	5 789
唐 DM	1 274	1 770	3 044
元 PLUS	1 656	730	2 386
海豹	1 740	379	2 119
海鸥	606	186	792
总计	17 136	10 147	27 283

旨在通过清洗和标准化操作去除无关信息，对用户评论进行拆分和观点提取，以确保文本的语义完整性和一致性。

首先，利用正则表达式清洗文本内容，通过公式 $L(x) \leqslant T_{\min}$ 或 $L(x) \geqslant T_{\max}$ 检测文本长度是否超出阈值，并剔除异常条目。同时，文本长度与向量维度的检测也确保嵌入处理的稳定性。随后，通过大语言模型生成语义嵌入向量，捕捉文本的潜在语义，并结合余弦距离公式实现相似性聚类，从而获得高语义相关度的子句集合。

$$\text{Similarity}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\boldsymbol{s}_i \boldsymbol{s}_j}{\|\boldsymbol{s}_i\| \|\boldsymbol{s}_j\|} \tag{1}$$

式中：$\boldsymbol{s}_i$ 和 $\boldsymbol{s}_j$ 分别为两段句子的嵌入向量；Similarity $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ 是两者的余弦相似度。通过这种方式，模型确保切分后的短文本具有高语义清晰度。此外，为确保观点提取的分布与原始评论分布的一致性，采用 KL 散度（Kullback-Leibler divergence）公式计算。

$$D_{KL}(P_{\text{extracted}} \parallel P_{\text{original}}) = \sum_x P_{\text{extracted}}(x) \ln \frac{P_{\text{extracted}}(x)}{P_{\text{original}}(x)} \tag{2}$$

式中：$P_{\text{extracted}}(x)$ 为提取的概率分布，是对于每个事件 x 提取的概率；$P_{\text{original}}(x)$ 为原始概率分布，是对于每个事件 x 原始的概率。这些变量代表了两个不同的概率分布，KL 散度用于衡量这两个分布之间的差异。

多层次的质量控制机制优化结果，包括利用加权损失函数平衡各项质量指标。这套基于深度语言模型的信息提取流程有效提升了文本数据的结构化程度和语义解读能力，为后续主题建模奠定了坚实基础。如表 2 所示。

1.3　BERTopic 模型

BERTopic[4] 是一种结合 BERT（bidirectional encoder representations from transformers）嵌入、降维和聚类算法的主题建模技术，相比于传统的 LDA[5]（latent dirichlet allocation）、CTM（correlated topic model）等主题模型，BERTopic 的优势在于弥合了基于密度聚类和基于中心采样之间的不兼容问题[6]，不需要人工确定主题数量。

表 2　文本数据观点提取示例

车型	元 PLUS
原始评论	以前开的冒险家,一个月 1 200～1 400 元的油钱,跑得比较多,家里还有一个油车,就有换购电车的想法,本地就一个 4s 店,正好有我要的配置颜色的现车,提车的上午卖掉冒险家,下午就提了 510 km 超越,到现超过 600 km,续航还是比较准,不暴力驾驶,应该要超过 510 km,内饰是我比较烦躁的地方,不过看了一周多,也比较习惯了,储物格太少,车门储物太小,整体还是比较满意。还有,我买了接近 3 000 元延保,~~*,论坛看了几个同学也买了,不知道能不能买,有同学来说说么? 我纯上下班代步,周边 200 km 的地方可以去,价格 15 元/km 的纯电车,纵观所有车型,不是外观不让我满意就是品牌让我不放心,不想买了车,过几年车商没了。哦对了,空间是我很满意的地方,4 400 mm 多的车身,比我以前冒险家 4 600 mm 的车身后排都要大,满足了,大概就这些吧
观点句	正好有我要的配置颜色的现车
	续航还是比较准,不暴力驾驶,应该要超过 510 km 续航
	内饰是我比较烦躁的地方,不过看了一周比较习惯了,储物格太少,车门储物太小
	价格 15 元/km 左右的纯电车,纵观所有车型,不是外观不让我满意就是品牌让我不放心
	空间是我很满意的地方,4 400 mm 多的车身,比我以前冒险家 4 600 mm 的车身后排都要大,满足了

BERT 嵌入是指使用 BERT 模型[7]生成的文本表示。BERT 模型是一种基于深度学习的语言表示模型,它利用 Transformer 的编码器部分构造了一个双向多层的架构,这使其能够在词向量表示中保留更丰富的语义信息[8]。通过预训练比亚迪新能源汽车口碑数据,多个 Transformer 双向编码器用于编码文本字符和缩略语,利用注意力计算词汇与其他所有词汇之间的关系和重要程度,以获取词汇之间的相互关系和内部结构,从而对样本句子进行编码。计算式为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (3)$$

式中:$\boldsymbol{Q}$ 为查询向量矩阵;$\boldsymbol{K}$ 为查询向量矩阵;$\boldsymbol{V}$ 为内容向量矩阵;$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 为用于计算输入字向量之间的关系;d_k 为编码器的输入向量矩阵维度[9]。

为了更好地理解和处理这些高维数据,模型通过均匀流形近似与投影(uniform manifold approximation and projection,UMAP)算法对高维语义向量进行降维,将其投射到低维空间,可以处理更高维度的数据集,还具有较强的可解释性。在运行聚类算法之前,使用 UMAP 对数据集进行降维操作,可以极大提高聚类精度和时间等方面的聚类性能[10]。使用基于密度的空间聚类与噪声的层次聚类算法(hierarchical density-based spatial clustering of applications with noise,HDBSCAN)对降维后的向量进行聚类,将相似的文本聚为同一主题。在每个主题的关键词提取阶段,引入最佳匹配 25(best matching 25,BM25)加权机制的聚类的词频-逆文档频率(clustered term frequency-inverse document frequency,c-TF-IDF)算法,计算每个主题中词的重要性权重,并提取能够代表主题的关键词。

c-TF-IDF 是 BERTopic 中的核心算法[11],用于对主题级别的词频进行权重计算。然而,传统的 c-TF-IDF算法中,词频是线性增长的,这可能导致高频词对权重的放大,影响主题的关键词提取效果。同时,对于包含大量词的长文档,可能会产生权重偏差。因此,引入 BM25 的加权机制,对 c-TF-IDF 进行优化[12]。BM25 加权的公式为

$$W_{x,c} = \frac{(k_1+1)f(x,c)}{f(x,c)+k_1\left(1-b+b\dfrac{|D_c|}{\text{avgdl}}\right)}\ln\frac{N+1}{n_x+1} \quad (4)$$

式中:$W_{x,c}$为词 x 在主题 c 中的重要性权重;$f(x,c)$ 为词 x 在主题 c 中的词频;$|D_c|$ 为主题 c 的文档总长度;avgdl 为语料库中文档的平均长度;N 为语料库的总文档数;n_x 为包含词 x 的文档数;k_1、b 为 BM25 的调节参数,分别控制词频的非线性加权和文档长度归一化。[13]

最终,在降维阶段,UMAP 的邻近样本点数量设置为 100,嵌入数据的降维空间维度设置为 192,最小距离参数设置为 0.000 135,以平衡数据的全局结构和局部特性,从而保留语义嵌入的丰富信息。随后,在聚类阶段,采用 HDBSCAN 算法,最小聚类规模设置为 120,核心点邻居数量设置为 500,并使用欧几里得距离作为度量方式,以保证聚类结果的稳定性和主题的可解释性。在文本特征向量化阶段,使用 Count Vectorizer 对文本进行分词和特征数值化,结合 BM25 权重优化后的 c-TF-IDF 算法对文本特征加权,从而突出关键主题词汇并抑制高频噪声词。使用一致性评分可对 BM25 加权的效果进行对比检验,该方法基于语料库中文档关键词的共现关系,通过 gensim 库中的 Coherence Model 计算。对于每个主题,提取关键词集合,结合语料库中的词频信息,计算主题关键词之间的共现概率。

$$C=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\ln\frac{P(w_i,w_j)+\varepsilon}{P(w_i)P(w_j)} \tag{5}$$

式中：C 为主题一致性得分；N 为主题中关键词的数量；$P(w_i,w_j)$ 为关键词 w_i 和 w_j 的共现概率；$P(w_i)$、$P(w_j)$ 为关键词的边际概率；ε 为平滑参数，用于避免对数值为零的问题。

此外，为了进一步优化主题建模的表现，采用最大边际相关性（maximal marginal relevance，MMR）方法提高关键词的多样性，确保每个主题保留高质量关键词。MMR 的选择过程可以通过式 6 来表示。

$$\mathrm{MMR}(R)=\arg\max_{D_i\in\mathrm{Candidates}\setminus R}\left[\lambda\mathrm{Sim}(D_i,Q)-(1-\lambda)\max_{D_j\in R}\mathrm{Sim}(D_i,D_j)\right] \tag{6}$$

式中：R 为当前已选结果的集合；$\mathrm{Sim}(D_i,Q)$ 是为候选文档 D_i 与查询 Q 的相似度；$\mathrm{Sim}(D_i,D_j)$ 为候选文档 D_i 与已选文档 D_j 的相似度；λ 为一个控制参数，范围为 [0,1]，用来平衡查询相关性和多样性[12]。

2　结果与讨论

在主题特征词分布中，用户关注的需求呈现多维度的趋势，涉及车辆性能、舒适性、环保性和实用性等方面。特别是在内饰气味、乘坐空间和能耗表现等主题上，关键词权重突出，表明消费者对环保、车内空气质量、空间舒适性和充电便利性有较高的需求。

从图 2 主题特征词分布和关键词权重中用户对新能源汽车需求的主要方向。内饰气味（主题 4）是用户最关注的主题，关键词“异味”（1.118 2）和“气味”（0.765 2）表明新车内饰气味，尤其是刺鼻异味和环保材料的选择成为核心问题。乘坐空间（主题 5）反映对“后排的空间”（0.550 6）和“成年人”

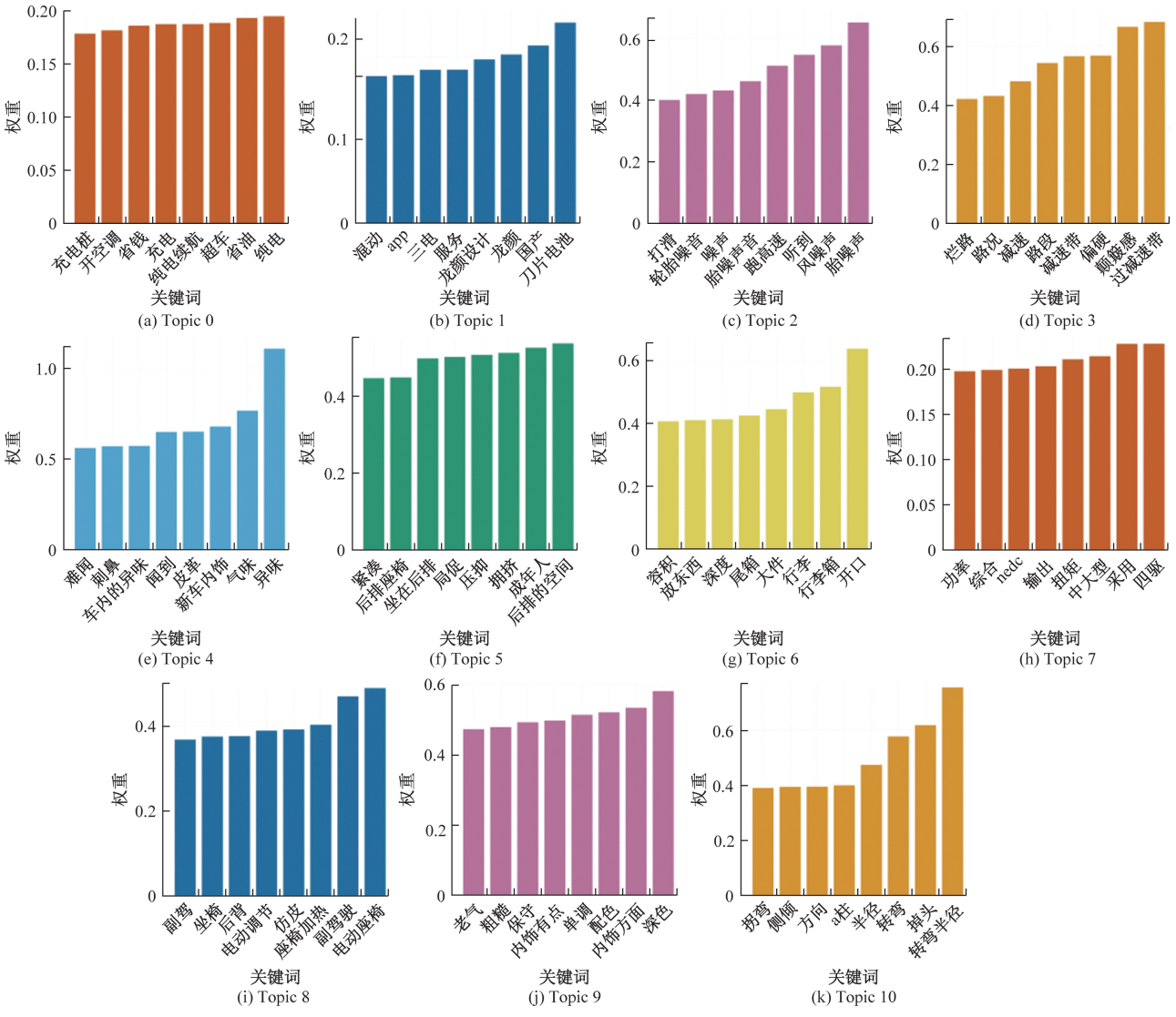

图 2　主题特征词分布

(0.535 1)的需求,显示出用户对紧凑型车型后排空间不足的关注。能耗表现(主题 0)聚焦在“充电”(0.205 5)和“纯电”(0.203 8)关键词上,用户希望提升充电便捷性和续航精准性,同时强调动力表现。行驶舒适性(主题 3)关注通过减速带时的“颠簸感”(权重 0.666 7)和避震性能,用户对车辆通过复杂路段时的减震表现提出了更高要求。最后,噪声控制(主题 2)中的“风噪”(权重 0.727 7)和“胎噪声”(权重 0.625 2)是重点,高速行驶时的噪声对驾驶体验的影响显著。这些数据为企业优化内饰材料、空间设计、能耗管理、避震系统和车身降噪提供清晰方向。

图 3 主题层次聚类结果显示,不同主题之间的关联性形成明显的层级结构,其中一些主题表现较高的相似性,而另一些主题则具有较强的独立性。在高度相似的主题中,储物空间与乘坐空间(主题 6 和主题 5)紧密相关,关键词如“行李箱”“尾箱”“后排空间”等反映了用户对车辆内部空间的实用性和舒适性的共同关注。同时,噪声与行驶体验(主题 2 和主题 3)也具有较高的相似性,用户对“胎噪声”“风噪声”等行驶噪声以及“减速带”“颠簸感”等复杂路况的适应性需求体现了对整体行驶舒适性的高度重视。在中等相似的主题中,视觉设计与内饰气味(主题 9 和主题 4)显示出用户对车辆内饰的整体体验需求,既要求视觉上的美观性,也关注嗅觉上的舒适性,减少异味干扰。而能耗与动力性能(主题 0 和主题 7)则结合了用户对车辆实际续航能力和动力输出稳定性的综合考量。独立性较高的主题包括座椅与配置(主题 8)和品牌与技术(主题 1)。座椅配置主要集中在“电动座椅”“座椅加热”

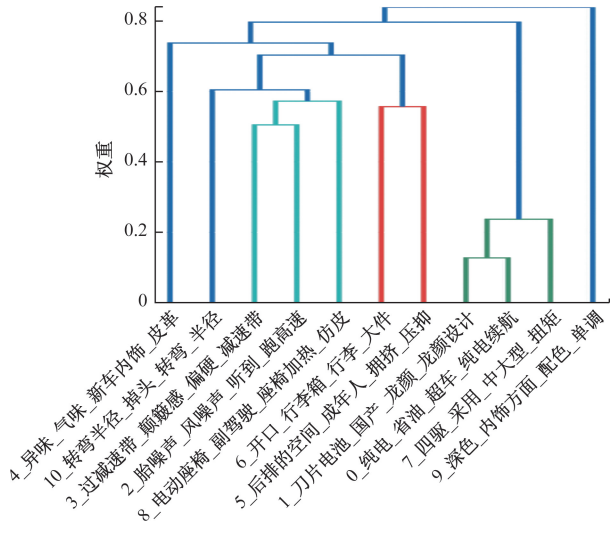

图 3　主题层次聚类

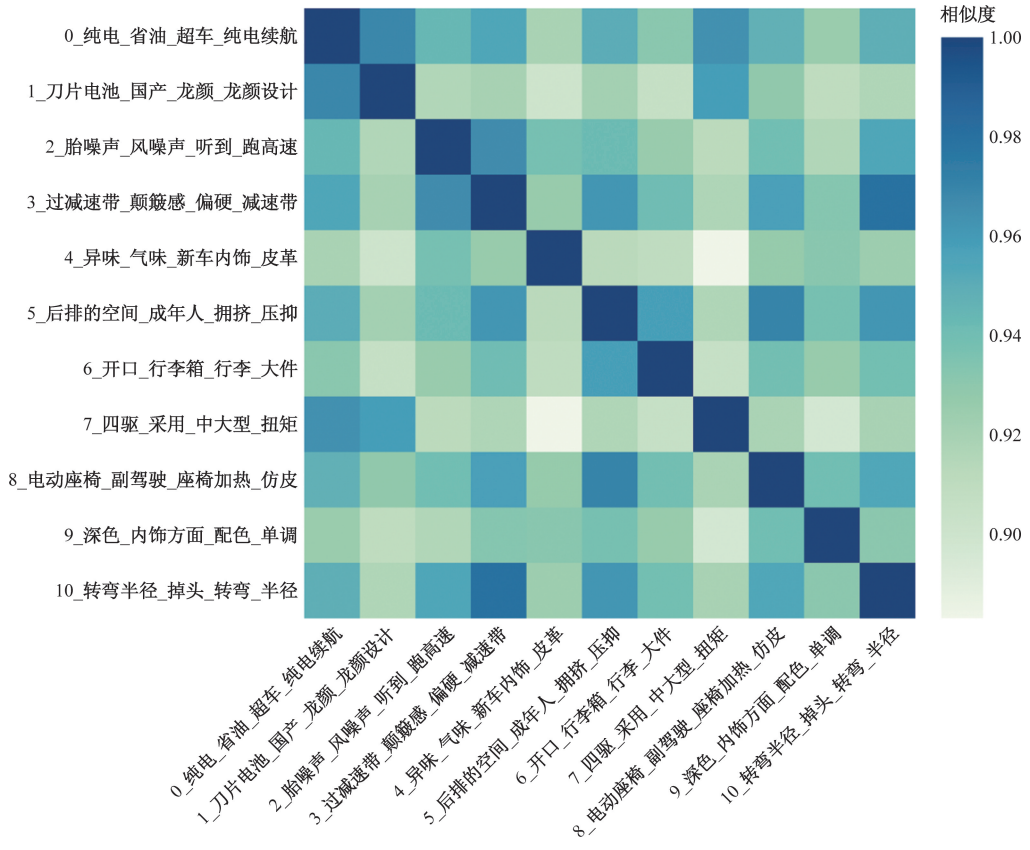

图 4　主题相似度热力

等功能需求上，体现了用户对座椅智能化和舒适度的关注，具有较强的独立性。品牌与技术则聚焦于“刀片电池”“龙颜设计”等技术创新和品牌形象，反映出用户对新能源汽车差异化竞争力的期待，与其他主题的关联性较弱。整体来看，层次聚类进一步揭示了用户关注点之间的结构性关系，为企业在产品优化和市场策略方面提供了具体方向。

通过主题相似度热力图（图4）可以看出，新能源汽车用户评论中的各主题之间存在不同程度的相似性。图中颜色越深表示主题之间的相似性越高，例如，储物空间（主题6）与乘坐空间（主题5）的相似性较高，反映了用户对车辆内部空间整体设计的共同关注。此外，内饰气味（主题4）与视觉设计（主题9）存在一定关联性，表明用户对内饰的综合体验包括嗅觉和视觉方面的统一需求。相比之下，品牌技术（主题1）和行驶噪声（主题2）等主题的相似性较低，体现了这些主题在用户关注点中的独立性。整体来看，相似度热力图揭示了主题间的联系和差异，为深入理解用户需求提供了定量化依据。

3 结论与展望

基于BERTopic主题建模方法，对新能源汽车用户评论数据进行深入分析，结合深度语义嵌入、UMAP降维、HDBSCAN聚类以及引入BM25加权的c-TF-IDF算法，有效提取了用户关注的11个主题。结果显示，用户对内饰气味、乘坐空间、能耗表现、行驶舒适性和噪声控制五大方面关注度最高。其中，BM25加权优化了c-TF-IDF的关键词提取效果，使得主题更具解释性；同时，通过MMR算法提高关键词多样性，确保每个主题关键词的代表性更全面。

对于新能源汽车行业而言，借助该模型，不仅可以更全面地挖掘用户需求，还能够持续监测用户关注点的动态变化。这一方法为企业提供精确调整产品设计、提升用户体验的策略，尤其是在提升内饰环保性、优化空间布局、增强续航表现与驾驶舒适性等方面。比亚迪作为行业领军企业，凭借其在新能源汽车领域的技术积累和创新，已经在这些方面取得显著成绩。通过其在内饰材料环保性、车内空间优化以及动力系统提升方面的成功实践，比亚迪为行业提供宝贵的经验和有效的技术路径。同时，行业内的其他企业可以借鉴比亚迪在品牌宣传和技术创新方面的经验，强化其技术竞争力，从而在市场中获得更强的竞争优势。

总体而言，该模型为行业提供行之有效的用户需求识别方法论，并通过比亚迪的发展经验为其他企业提供值得借鉴的策略和技术支持。未来，随着更多用户需求识别方法论的应用，新能源汽车行业将在产品创新与市场竞争力提升方面获得更强的动力。

参考文献

[1] 韩顺杰，于渲铎，李东奇，等. 基于改进量子粒子群算法的新能源汽车换电站优化布局[J]. 科学技术与工程，2024，24(27)：11720-11725.

[2] REIMERS N，GUREVYCH I. Sentence-BERT：sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong：USAACL，2019：3982-3992.

[3] 张永安，周怡园. 新能源汽车补贴政策工具挖掘及量化评价[J]. 中国人口·资源与环境，2017，27(10)：188-197.

[4] 余博，管超，戴淑庚. 人民币国际化、汇率波动与双边贸易——基于“一带一路”国家面板门槛模型的分析[J]. 统计与信息论坛，2020，35(7)：57-65.

[5] GROOTEN D. BERTopic：Neural topic modeling with a class-based TF-IDF，procedure[J]. arXiv，2022，3：05794.

[6] LI Q，GUO Y，TIAN C. Effect of material crack flaws on dynamic fracture behavior[J]. Science Technology and Engineering，2016，18(28)：1-5.

[7] DEVLIN J，CHANG W，LEE K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[C]// In Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Minneapolis：Google AI Language，2018：4171-4186.

[8] 陈晨，石赫，徐悦，等. 基于BERT-BiLSTM的油田安全生产隐患文本分类[J]. 科学技术与工程，2024，24(29)：12650-12657.

[9] 郝宽公，董兵，吴悦，等. 基于BERT-Bi-LSTM-CRF模型的机场类中文航行通告要素实体识别[J]. 科学技术与工程，2024，24(10)：4182-4188.

[10] 陈健飞，卜凡亮，王一帆. 基于CoSENT和改进K-Means的冒犯性评论文本主题识别[J]. 科学技术与工程，2024，24(31)：13442-13449.

[11] 李雄，刘允才. 视觉机制研究对机器视觉的启发示例[J]. 中国图象图形学报，2013(2)：152-156.

[12] ROMANE，YU J. A topic modeling comparison between LDA，NMF，Top2Vec and BERTopic to demystify twitter posts[J]. Frontiers in Sociology，2022，67(5)：886498.

[13] ZIEGLER C N, MC NEE S. Improving recommendation lists through topic diversification[C]// Proceedings of the 14th International Conference on WorldWide Web. Chiba, Japan: Keio University, 2005: 22-32.

Analysis of New Energy Vehicle Evaluations Based on BERTopic: A Case Study of BYD

ZHAO Chengbing, WANG Yao, CHU Lizheng

(School of Mathematics and Physics, Anhui Jianzhu University, Hefei 230601, China)

Abstract: Taking seven popular BYD models as research objects, 27 283 long-text reviews from AutoHome and DongCheDi were collected. Effective short texts were extracted using the DeepSeekV2 model, followed by BERTopic modeling to identify key consumer focus areas, including pure electric performance, blade battery and noise control. The experimental results reveal that consumers give positive feedback on battery technology and range capability but express concerns about comfort-related issues such as tire noise, wind noise and interior odors. The contributions of this study include proposing the use of the DeepSeekV2 model to split long texts, enhancing the adaptability of BERTopic in data analysis, and combining semantic embedding and hierarchical clustering techniques to uncover consumers' multidimensional concerns and their structural relationships.

Keywords: BERTopic; BYD; new energy vehicles; topic modeling