Modeling and Simulation 建模与仿真, 2025, 14(2), 543-552
Published Online February 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.142174

Hans汉斯

基于多模态的新能源汽车用户需求挖掘

陈佳佳，赵敬华

上海理工大学管理学院，上海

收稿日期：2025年1月27日；录用日期：2025年2月20日；发布日期：2025年2月28日

摘　要

本文针对新能源汽车市场快速扩张背景下，消费者在线评论中多模态信息的情感分析需求，提出了一种基于改进多头注意力机制的多模态情感分析模型。该模型通过跨模态和自注意力机制的融合，有效提升了新能源汽车在线评论中情感倾向的识别精度。实验结果表明，该模型在多个数据集上的性能优于现有方法，为新能源汽车用户需求挖掘提供了新的视角和工具。

关键词

新能源汽车，多模态，在线评论

Multimodal-Based User Demand Mining for New Energy Vehicles

Jiajia Chen, Jinghua Zhao

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 27th, 2025; accepted: Feb. 20th, 2025; published: Feb. 28th, 2025

Abstract

This paper addresses the need for sentiment analysis of multimodal consumer reviews in the rapidly expanding new energy vehicle market by proposing an improved multi-head attention mechanism-based multimodal sentiment analysis model. The model effectively enhances the recognition accuracy of emotional tendencies in online reviews by integrating cross-modality and self-attention mechanisms. Experimental results demonstrate that the model outperforms existing methods

文章引用: 陈佳佳, 赵敬华. 基于多模态的新能源汽车用户需求挖掘[J]. 建模与仿真, 2025, 14(2): 543-552.
DOI: 10.12677/mos.2025.142174

across multiple datasets, providing a new perspective and tool for mining user requirements in new energy vehicles.

Keywords

New Energy Vehicles, Multimodal, Online Review

1. 引言

随着全球能源危机的加剧和环境污染问题的日益严重，新能源汽车(NEVs)作为替代传统燃油车的重要选择，其市场规模正在迅速扩大[1]。消费者对新能源汽车的态度和反馈，尤其是通过在线评论表达的情感和意见，对于汽车制造商来说具有极高的价值[2]。在线评论中所包含的多模态信息，如文本、图像和视频，为理解消费者情感提供了丰富的素材[3]。然而，单一模态的情感分析方法难以全面捕捉用户的真实感受，多模态情感分析(MSA)因此成为新能源汽车领域的关键技术[4]。MSA 旨在从多种模态数据中提取情感信息，以更全面地理解和预测用户的情感倾向[5]。

近年来，研究者们提出了多种模型来处理多模态情感分析任务，包括基于注意力机制的方法[6]、预训练语言模型[7]以及多模态大模型[8]。任刚等人[9]提出的 ITRHP 模型利用图片和文本的多模态信息，基于图文匹配技术，通过 Faster R-CNN 和 Bi-GRU 模型分别提取图像和文本特征，并通过协同注意力机制提高特征表达的一致性，显著提升了模型的分类准确性。张换香等人[10]则聚焦于中文隐式情感分析，提出融合多模态信息的方法，通过 BiLSTM 网络挖掘各单模态内部的上下文信息，并结合多头互注意力机制捕捉与文本相关的语音和视觉特征，有效提升了隐式情感分析的准确率。李懋林等人[11]针对社交平台数据的多模态方面的情感分析，提出了基于置信度引导的提示学习(CPL)模型，该模型通过自注意力网络的置信度评估样本的分类难度，并采取不同的适应性模板提示，以引导大语言模型生成辅助情感线索，显著提高了情感分类的准确率。

然而，现有的基于多头注意力机制的模型在处理新能源汽车在线评论时仍面临挑战，尤其是在模态间信息融合和细粒度情感分类方面[12]。为了解决上述问题，本文提出了一种基于改进多头注意力机制的模型，专门针对新能源汽车在线评论进行多模态情感分析。该模型通过引入跨模态注意力和自注意力的融合策略，增强了模型对不同模态间情感信息的捕捉能力。

2. 新能源汽车用户需求挖掘模型

2.1. 基于 Bert 模型的新能源汽车属性情感分类

以往为了得到用户对新能源汽车不同属性的喜好程度通常的做法是针对不同的属性，搭建不同的分类器，得到对应的情感分析结果，但这需要构建多个模型，成本高，效率低。为解决这一问题，本文旨在基于 Bert 构建一个统一分类模型，同时对评论整体以及不同属性进行情感分析，得到关于新能源汽车全面的情感分数分布，为下一步研究用户的情感，需求趋势变化提供准确和全面的信息输入。本文对 Bert 的预训练模型进行微调，并得到基于 Bert 的每条输入评论的 embedding 表示。改进的 Bert 模型结构如图 1 所示。

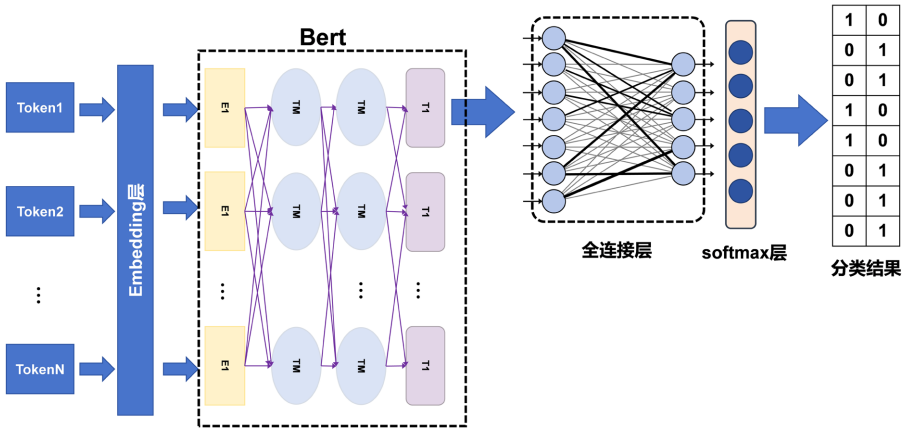

Figure 1. Schematic structure of improved Bert model
图 1. 改进 Bert 模型结构示意图

2.2. 基于多模态融合的新能源汽车属性情感分析

本文描述了使用一种新颖的多模态融合方法来分析和提取用户情感信息的过程，该方法结合了 Bert 模型和 ResNet152 模型。如图 2 展示了跨模态注意力的流程图。文本编码层使用 Bert 对内容进行编码，生成文本信息的特征向量。图像编码层采用 ResNet-152 对图像进行编码，产生尺寸为 7*7 的序列特征向量，并通过全连接层获得全连接特征向量。注意力层通过应用自注意力层聚合这两组序列特征向量，并通过取平均值来获得最终的注意力层特征向量。

注意力聚合层将注意力层特征向量与文本端的特征向量以及图像端的全连接层特征向量进行拼接。拼接后的特征向量通过一个全连接层，得到多模态聚合特征向量。分类层将两个多模态聚合特征向量相加，并应用 SoftMax 层进行分类。

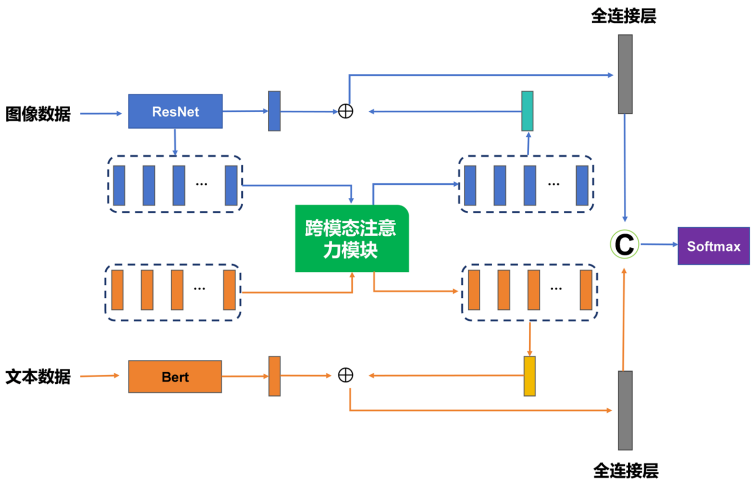

Figure 2. A multimodal sentiment analysis network architecture based on cross-modal attention
图 2. 基于跨模态注意力的多模态情感分析网络架构

该建模方法的核心是跨模态注意力机制(见图 3)，注意除了文本信息和图像信息，为适用在更宽泛的场景，本方法还提供第三种模态的输入建模方式，如视频，语音等。跨模态注意力模型首先通过特定模

态的编码器处理各自的模态数据。然后，编码后的特征分别输入到自注意力或跨注意力的多头注意力(Multi-Head Attention, MHA)模块中。在每个注意力模块的输出处，采用时间平均操作生成语句片段的全局表示。随后将得到的特征进行拼接，并通过统计池化层获取其均值和标准差。接着，将均值和标准差向量的拼接结果输入全连接层。最终，通过 Softmax 操作得到情感分类预测。

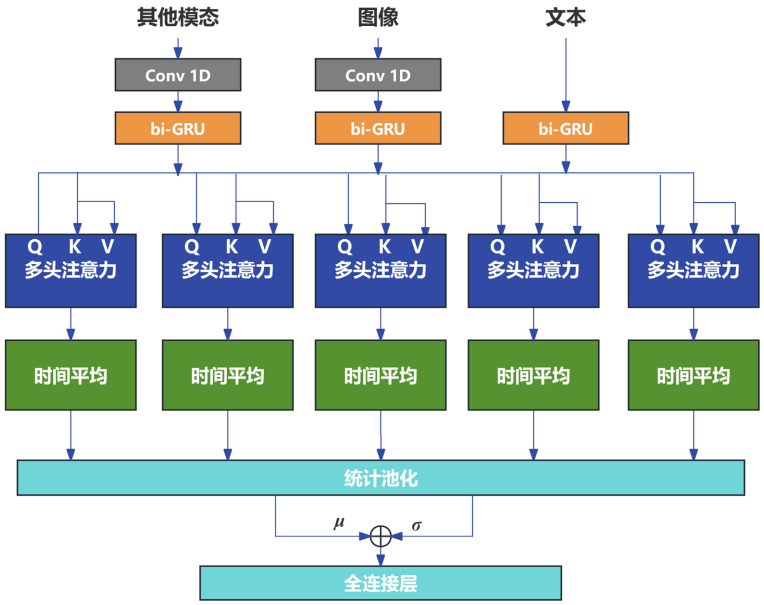

Figure 3. Architecture of a multimodal cross-attention model
图 3. 多模态交叉注意力模型的架构

设 $X_v \in R^{t_v \times d_v}$ 为与文本片段对应的视觉特征，其中 tv 是序列长度，dv 是特征维度。视觉编码器由一维卷积层和双向 GRU(门控循环单元)组成。卷积层通过找到与任务相关的模式来优化输入特征序列，其操作如下：

$$X'_v(t') = b(t') + \sum_{k=0}^{t_v-1} \left(W(t',k) * X_v(k) \right) \tag{1}$$

其中 $X'_v \in R^{t'_v \times d'_v}$ 是输出，长度为 t'_v，维度为 d'_v，$t' \in [0, t'_v - 1]$，*表示卷积运算符，W 表示与卷积层相关的权重，b 是偏置项。因此，卷积层不仅修改了序列长度，还改变了特征维度。对于序列中的每个元素，双向 GRU 层计算以下函数：

$$\begin{cases} r_t = \sigma\left(W_{ir} X'_v(t) + b_{ir} + W_{hr} h_{t-1} + b_{hr}\right), \\ z_t = \sigma\left(W_{iz} X'_v(t) + b_{iz} + W_{hz} h_{t-1} + b_{hz}\right), \\ n_t = \phi_h\left(W_{in} X'_v(t) + b_{in} + r_t \odot \left(W_{hn} h_{t-1} + b_{hn}\right)\right), \\ h_t = (1 - z_t) \odot n_t + z_t \odot h_{t-1} \end{cases} \tag{2}$$

其中和 h_{t-1} 是在时间 t 和 $t-1$ 的隐藏状态，$X'_v(t)$ 是在时间 t 的输入。r_t、z_t 和 n_t 分别是重置门、更新门和新门，W 和 b 是相应的权重和偏置，σ 和 φ_h 是 sigmoid 函数曲正切函数，$\odot$ 是哈达玛积。在 bi-GRU 的输出中，每个时间步的前向和后向隐藏状态被连接在一起，精炼的视觉特征可以表示为 $e_v \in R^{t'_v \times d''}$，其中 d'' 是 GRU 中隐藏神经元数量的两倍。

本文使用多头注意力(MHA)模块来进行自注意力和交叉注意力建模。每个 MHA 模块需要 3 个输入，分别是查询(Query, Q)、键(Key, K)和值(Value, V)，每个输入首先通过线性层被投影到 H 个不同的子空间

中，其中 H 指的是头的数量。每个子空间的投影 $h \in \{0,\cdots,H-1\}$ 可以计算为：

$$Q_h = W_h^Q e_m \tag{3}$$

$$K_h = W_h^K e_m \tag{4}$$

$$V_h = W_h^V e_m \tag{5}$$

其中 $m \in \{a,v,l\}$ 表示模态。在每个这些子空间中，对投影进行缩放点积注意力操作。对于子空间 h，注意力操作表示为：

$$Att_h(Q_h,K_h,V_h) = \text{Softmax}\left(\frac{Q_h K_h^{\mathrm{T}}}{\sqrt{d_k}}\right)V_h \tag{6}$$

其中 Att_h 和 d_k 分别表示子空间 h 中的注意力操作和特征维度。所有 H 个注意力的输出被连接起来，并通过一个线性层，以获得 MHA 模块的最终输出。

在意力模型中，相同模态对应的输入序列被用作查询(Q)、键(K)和值(V)。这有助于捕捉每个模态中的模态内交互。对于交叉注意力模型，统计池化是在跨模态序列的时间平均值的连接上进行的，而对于自注意力模型，统计池化是在所有模态的自注意力序列的时间平均值的连接上进行的。这两个模型的分类器为：

$$\hat{y} = \text{Softmax}\left(f_{\theta 2}\left(f_{\theta 1}\left(\left[\mu \| \sigma\right]\right)\right)\right) \tag{7}$$

其中，μ 和 σ 是从统计池化层的输出中获得的均值和标准差，k 代表连接操作，$f_{\theta 1}$ 和 $f_{\theta 2}$ 分别表示带有参数 $\theta 1$ 和 $\theta 2$ 的两层全连接层，$\hat{y}$ 表示情感预测的独热向量。

3. 多模态情感分析模型设计

3.1. 情感词典构建

本文通过总结已有构建方法，提出了一种新颖的可扩展的词典构建方式，其构建流程如图 4 所示：

Figure 4. Flowchart of sentiment dictionary construction
图 4. 情感词典构建流程图

第一步，对已进行清理和分词的文本库中的词语进行词性分类。本文使用 Jieba 的词性标注功能得到每个词的词性标识(如表 1)，然后根据本文的词性保留逻辑只保留相关词性的词。

Table 1. Part-of-speech tagging table
表 1. 词性含义表

标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	t	时间
nr	人名	ns	地名	nt	机构名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词

续表

m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间

第二步，统计高频词汇，即本文选择少量(约 300 个)用户常用的词作为种子词。

第三步，针对第二步筛选出来的高频词，对其进行手工标注，将其分为正向和负向两个类别，严格控制标注质量。

第四步，词典扩展中，利用词嵌入模型 word2vec 对每个词生成一个词向量。该向量可以用来计算词之间的相似度，利用此相似度，可以找到和第三步生成的高频种子词相似的词，并根据对应种子词的情感标签自动对扩展的词的情感进行分类，从而实现情感词典的扩展。

最后一步，人工再次介入对扩展词的标注进行抽样检测，控制自动生成的标签质量。

3.2. 结合多模态模型和情感词典进行情感分析

本文讨论如何将多模态情感分析模型通过微调得到针对不同属性的情感分析结果和通过构建情感词典得到关键词的情感标签结合起来从而进一步提升情感分析的结果。

首先使用多模态情感分析模型进行情感特征提取，包括(1) Bert 模型初始化，如加载预训练的 Bert 模型和相应的分词器。(2) 文本编码：将预处理后的文本通过 Bert 的分词器转换为 Bert 输入的格式，通常包括输入序列的 token ids、attention masks 等。(3) 图像编码：将预处理的图像数据输入到多模态模型中的 ResNet 图像处理分支。(4) 特征提取：将文本特征和图像特征进行融合，提取出整个句子的特征表示。(5) 情感打分：可以使用线性层(如全连接层)接在融合的特征向量后，用于计算情感得分，通常是正面情感和负面情感的概率，这里记为 Sb。

然后结合情感词典。利用预处理步骤中标记的情感词，结合情感词典对文本中的情感词进行打分。例如，对每个正向词汇赋予一个正分数，对负向词汇赋予零。然后对文本中的所有词汇情感分数进行累加，并根据句子长度进行归一化，得到整个句子的情感词典分数记为 Sd。

最后将多模态模型输出的情感概率与情感词典的打分结合，从而融合多模态模型和情感词典结果。即将多模态模型的输出与情感词典的结果作为两个独立的特征输入到一个机器学习模型中，训练一个最终的分类器来预测情感标签和对应的概率，即以多模态模型的结果 Sb 和情感词典的结果 Sd 作为输入特征，构建一个逻辑回归模型来得到最终的情感分析结果。

通过这种方法，多模态情感分析模型能够捕捉到上下文中的深层次情感信息，而情感词典则能够补充多模态模型可能忽略的显式情感表达，两者结合可以提高情感分析的整体效果。

3.3. 对比实验与模型评价

为了验证多模态模型的有效性，本文设置对比试验，除了和针对文据的 Bert 模型本身进行比较，本文还采用情感分析领域的经典模型 TextCNN 和循环神经网络(RNN)作为基线模型。并采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 分数。其中准确率是分类正确的样本数与总样本数之比。它衡量了模型整体分类的正确性，公式如下：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

其中 TP (True Positive)为真正例，模型正确预测为正类的样本数。TN (True Negative)为真反例，模型正确预测为负类的样本数。FP (False Positive)为假正例，模型错误预测为正类的样本数。FN (False Negative)为假反例，模型错误预测为负类的样本数。

4. 实验设计与分析

4.1. 数据预处理

本文选择懂车帝平台(http://www.dongchedi.com/)“口碑”模块用户评论作为主要数据源，选择了特斯拉 Model Y、极氪 001(Zeekr 001)、小米 SU7 等几款新能源车型作为研究对象，本文爬取从最早 2020 年四月直到 2024 年 10 月 1 日期间的所有评论，共 7276 条文字评论，25279 张图片评论每个车型的评论数分布如图 5 所示。

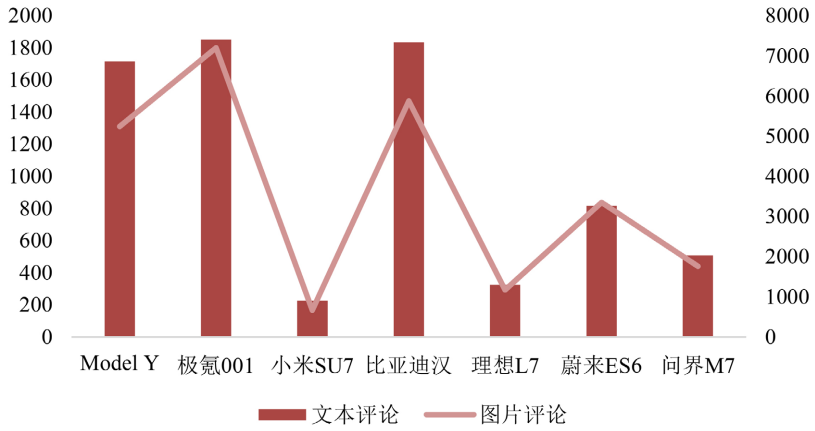

Figure 5. Distribution of reviews for new energy vehicles
图 5. 新能源汽车评论数分布

对评论数据进行数据清理和预处理的过程包括分词，去掉停用词，使用汽车专用词典等。数据处理之后每条评论由一系列 token 表示。对于图像数据，本文将其按评论分成组，依次输入到多模态情感分析模型。

4.2. 多模态情感分析实验结果

本文对处理好的评论数据分为训练集和测试集，其比例为 8:2。为了对多模态情感分析模型进行微调，各项参数设置如表 2 所示。

实验结果按照评论整体，内部空间，驾驶体验，续航能力，外观，内饰，配置，价格/性价比等属性总结如图 6 所示。实验结果表明 TextCNN 整体优于 RNN，部分由于评论通常较长，而 RNN 在捕捉长期依赖关系时易产生梯度消失的问题，导致 RNN 的表现不如基于 CNN 的 TextCNN 好。而 Bert 模型无论是在评论整体还是针对不同的属性，指标都全面超越 TextCNN，体现了注意力机制在提取文本语义方面的领先性。另一方面，通过融合文本信息和图像信息，本文提出的多模态情感分析模型在所有属性中均显著超越 Bert 模型，显示出不同模态之间的互补作用，也证明了多模态情感分析的有效性。

Table 2. Experimental parameter environment
表 2. 实验参数环境

参数	值
优化函数	Adam
学习率	0.001
Batch_size	64
embedding	256
doprout	0.5
训练轮数	300

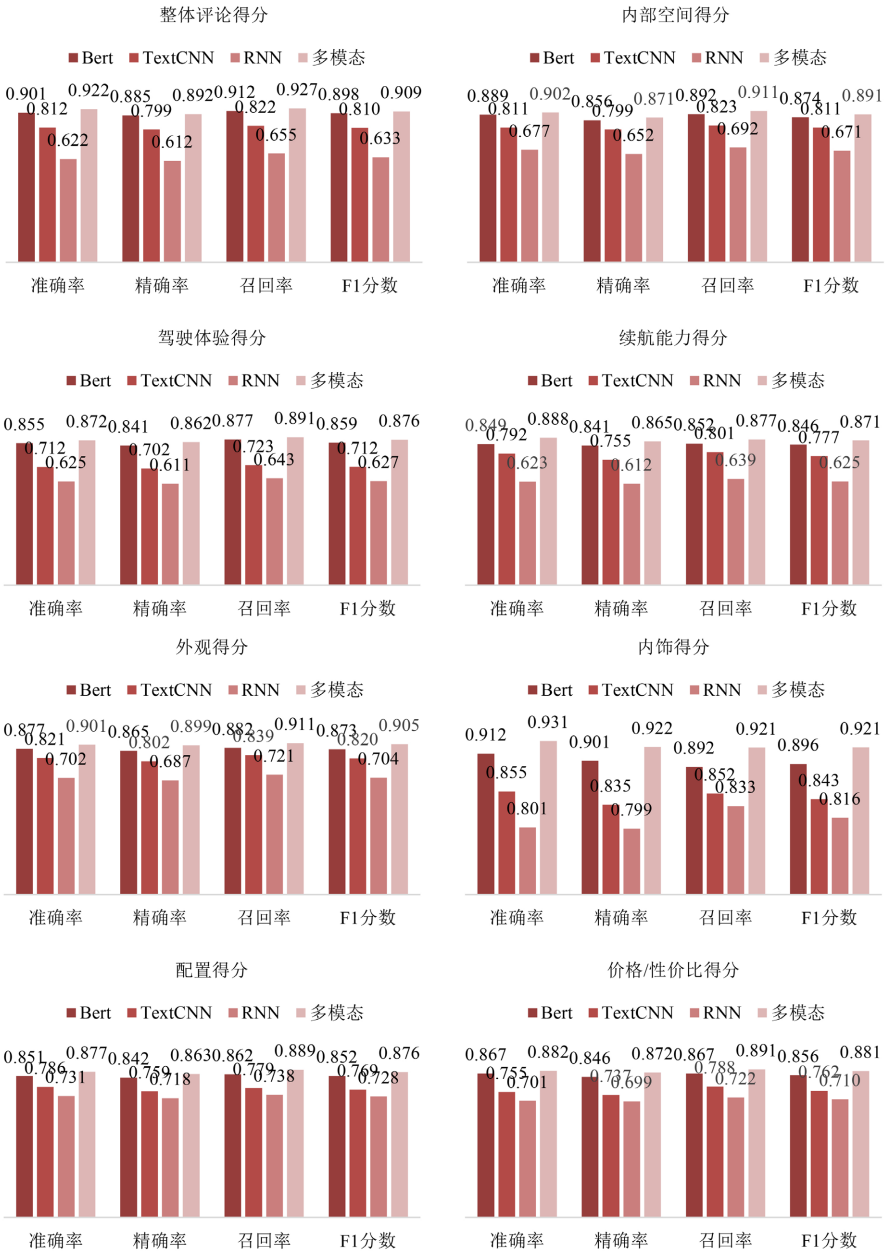

Figure 6. Experimental results of various models
图 6. 各模型实验结果

5. 结语

本文通过构建基于改进多头注意力机制的多模态情感分析模型，对新能源汽车在线评论的情感倾向进行了深入分析。研究结果表明，该模型在情感识别精度和鲁棒性方面表现优异，尤其在处理图文并茂的复杂评论场景时，展现出显著的优越性。

通过对实验结果的系统分析，本文得出以下主要结论：

(1) 多模态融合的优势：多模态情感分析模型通过整合文本和图像信息，能够从多个维度捕捉用户情感，显著提升了情感分析的准确性和全面性。与单一模态模型相比，该模型在处理复杂情感表达时更具优势，能够更精准地识别用户的真实情感倾向。

(2) 模型架构的有效性：融合跨模态注意力和自注意力机制的多头注意力模型，能够有效捕捉不同模态间的情感关联，并显著提升情感分类的精度。实验结果表明，该模型在多个新能源汽车评论数据集上均优于传统的情感分析模型(如 TextCNN 和 RNN)，验证了其在多模态情感分析任务中的有效性。

(3) 实际应用价值：多模态情感分析模型不仅在情感识别精度上取得了突破，还为新能源汽车用户需求挖掘提供了新的视角和工具。该模型能够为汽车制造商提供更精准的用户反馈，助力产品优化和市场营销策略的制定。

综上所述，本文提出的多模态情感分析模型为新能源汽车领域的用户情感分析提供了一种高效、准确的方法。未来研究将进一步探索多模态情感分析在其他领域的应用，并优化模型架构以提升其在更大规模数据集上的性能。

致　谢

在本研究工作即将完成之际，我谨向所有在研究过程中给予我支持和帮助的人表示最深切的感谢。我要特别感谢我的导师赵敬华副教授，她的悉心指导和宝贵建议对本研究的完成至关重要。同时，我也要感谢我的家人，他们对我的理解和支持是我不断前进的动力。没有他们，这份成果不可能实现。再次表达我最深的感激之情。

基金项目

国家自然科学基金资助项目(72201173)。

参考文献

[1] 蒲中敏, 张晨曦, 徐泽水. 基于在线评论的消费者偏好挖掘研究综述[J]. 中国管理科学, 2025, 33(1): 209-220.

[2] Poria, S., Hazarika, D., Majumder, N. and Mihalcea, R. (2023) Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, **14**, 108-132. https://doi.org/10.1109/taffc.2020.3038167

[3] Thandaga Jwalanaiah, S.J., Jeena Jacob, I. and Mandava, A.K. (2022) Effective Deep Learning Based Multimodal Sentiment Analysis from Unstructured Big Data. *Expert Systems*, **40**, e13096. https://doi.org/10.1111/exsy.13096

[4] Valstar, M., Pantic, M., Ambadar, Z., *et al*. (2015) A Survey into the Detection and Analysis of Facial Expression: Overview of Psychology-Based Models, Databases, Evaluation Protocols, and Applications. *IEEE Transactions on Affective Computing*, **6**, 366-387.

[5] Cambria, E., Poria, S., Bajpai, R., *et al*. (2018) Sentic Computing: A Common-Sense-Based, Cognition-Inspired Paradigm. *IEEE Computational Intelligence Magazine*, **13**, 70-80.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[7] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[8] Radford, A., Narasimhan, K., Salimans, T., *et al*. (2018) Improving Language Understanding by Generative Pre-Training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[9] 任刚, 程玲凤, 贾子垚, 王安宁. ITRHP 模型: 一种基于图文匹配的多模态评论有用性预测方法[J/OL]. 数据分析与知识发现, 1-16. http://kns.cnki.net/kcms/detail/10.1478.G2.20241218.1636.014.html, 2025-2-24.

[10] 张换香, 李梦云, 张景. 基于多模态信息融合的中文隐式情感分析[J]. 计算机工程与应用, 2025, 61(2): 179-190.

[11] 李懋林, 林嘉杰, 杨振国. 基于置信度引导提示学习的多模态方面级情感分析[J/OL]. 计算机科学, 1-12. http://kns.cnki.net/kcms/detail/50.1075.tp.20240926.1755.010.html, 2025-2-24.

[12] Wang, S. and Manning, C.D. (2018) Fast Dropout Training for Multimodal Sentiment Analysis. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 October-4 November 2018, 4877-4883.