基于 YOLOv8 的多模态方面类别情感分析模型的设计与实现

张晓薇 [1]
ZHANG Xiaowei

摘 要 多模态基于方面的情感分析（multimodal aspect-based sentiment analysis, MABSA）作为一种更加全面的情感分析方式，旨在从不同模态中提取方面并识别其情感。现有的研究大多以文本的语义特征分析为主，粗略地将整个图像特征与文本特征进行拼接或者基于注意力机制对特征进行融合，以产生相应的特征表示。这将导致图像中的细粒度元素未能被充分挖掘与利用，同时还会向模型引入无关干扰信息，从而降低模型的准确率。为了解决上述问题，文章提出了一种基于 YOLOv8 的多模态方面类别情感分析模型（multimodal aspect-category sentiment analysis model based on YOLOv8, YL-MACSA）来检测多模态数据集中与方面相关的语义和情感信息。实验结果表明，该模型在跨模态细粒度融合方面取得了良好的效果。

关键词 YOLOv8；多模态；方面类别；情感分析

doi：10.3969/j.issn.1672-9528.2026.04.009

0 引言

随着移动设备和社交网络的蓬勃发展，越来越多的用户倾向于在社交平台上把图片和文字放在一起来表达自己的情感观点，这使得多模态情感分析成为近年来备受关注的研究课题 [1]。在服务业中，尤其是酒店餐饮行业，聚焦用户体验对企业至关重要，而实现这一目标的有效途径，便是借助多模态方面级情感分析技术，深入挖掘用户提供的包含图文信息的评论与反馈。这种分析不仅能捕捉用户对整体体验的评价，更能精准定位到具体方面，如酒店的房间设施、餐饮的菜品口味、服务人员的态度等，从而为企业调整提供明确依据。

一般来说，方面级情感分析的主要研究方向涉及识别各个方面的情感元素，包含对方面术语、方面类别、意见术语以及情感极性的抽取 [2]。例如，在文本“the pizza is deliciou”中，方面术语为 pizza，方面类别为 food，意见术语为 delicious，情感极性为 positive。对于多模态数据集而言，方面级情感分析模型需要从多种不同模态中挖掘丰富的细粒度信息，如何将这些信息对齐并整合为一个统一的多模态表示，是目前面临的主要问题 [3]。从上述的例子中可以看出，“方面类别”相对于“方面术语”而言，是一个笼统的概念，这种较为笼统且预先定义的方面可以更好地协调两种模态之间的详细信息。相比之下，“方面术语”有时是不存在的，因此本文选择“方面类别”来连接这两种形态，并进一步得到每个方面类别下对应的情感极性。

对于多模态数据，模型可通过额外模态来补充文本中用户未明确提及的相关信息。如图 1 所示，用户在文本“close to the night market and beach”中明确表达了对酒店“位置”方面的积极态度，文本“very good, very clean”没有明确说明用户想要表达的方面，但结合图 1 可知，这里指“房间”和“公共设施”两个方面。由此可见，对于这种在线评论中常见的模糊表述，图像能够有效帮助模型更好地挖掘方面信息以及对应的情感极性。

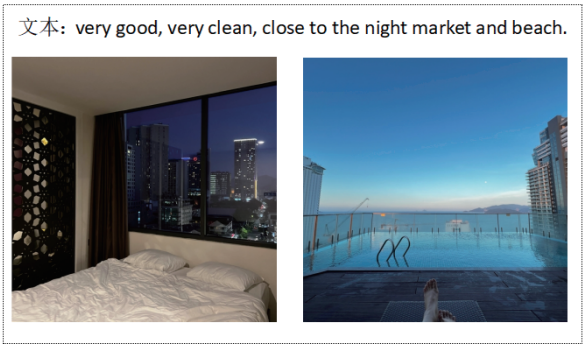

图 1 酒店评论示例

在多模态方面及情感分析领域，Xu 等 [4] 首先在 2019 年手动标注了一个多模态基于方面的情感分类数据集（Multi-ZOL Dataset），该数据集收集了大量关于不同品牌手机的多模态评论，并提出了一种多模态交互式记忆网络（MIMN），该模型采用双通道记忆网络分别建模文本和图像数据，并利用注意力机制捕获面向方面类别的文本和视觉表示，实现模态内与模态间的信息交互。同年，Yu 等 [5] 提出的 ESAFN 模型整合了注意力机制、门控机制和双线性交互，用以捕捉实

1. 太原学院智能与信息工程系 山西太原 030032
[基金项目]2024 年太原学院校级项目（24TYQN15）

体级多模态情感分析中，模态内部及模态之间的动态关系。

此外，一些研究人员试图将基于 Transformer 的方法应用于该任务。Yu 等 [6] 在 2019 年提出了一种基于 BERT 的多模态框架，该框架利用两个基于 BERT 的模块从文本和图像中提取模态内特征，以及第三个基于 BERT 的模块通过整合文本和视觉信息来提取跨模态特征。2021 年，Khan 等 [7] 提出的 EF-CaTrBERT 模型利用 transformer 生成图像描述，并构建辅助句子进行多模态方面情感分析。2024 年，Yang 等 [3] 提出 MGAM 模型，该模型通过构建辅助句子和异构图，学习模态间的交互关系。

综上所述，与基于纯文本的方面情感分析相比，多模态方面级情感分析侧重于从文本和视觉等不同模态信息中捕捉细粒度情感特征，这些模态的联合使用可以补充相关信息，增强情感表达。因此，本文提出一种基于 YOLOv8 的多模态方面类别情感分析模型（YL-MACSA），该模型能学习文本元素与视觉元素内部及跨模态的交互作用，提高在多模态数据中模型的方面提取与情感分析能力。

1 本文模型

对于每个给定的文本图像对，都有一个文本 $S=\{w_1,w_2,\cdots,w_L\}$ 和一个图像列表 $G=\{G_1,G_2,\cdots,G_K\}$。预先定义了一个方面类别集 $C=\{C^1,C^2,\cdots,C^N\}$。这里的 L、K 和 N 分别表示句子中的单词数量、图像数量和预定义方面类别数量。给定一个输入对 (S, G) 和一个特定的方面类别 C^n，组成一个多模态输入样本 (C^n, S, i)，目标是确定与该方面相关的情感标签。方面情感极性分为 0 表示不相关，1 表示消极，2 表示中性，3 表示积极。

1.1 图像处理

在图像处理阶段，模型应用 YOLOv8 算法对给定的一组图像 $G=\{G_1, G_2,\cdots,G_K\}$ 进行目标检测，得到每张图像 G_k 的一组感兴趣区域（region of interest, RoI）$R^k=\{R^k_1,R^k_2,\cdots,R^k_J\}$，$J$ 表示第 k 张图的 RoI 个数。接着，使用 ResNet-152 模型得到每张图像 G_k 的方面类别 A_k 和每个感兴趣区域 R^k_j 的方面类别 A_j^k，其中 $A_k\in\{0,1\}^N$、$A_j^k\in\{0,1,...,n\}$。用公式表示为：

$$R^k=\mathrm{YOLOv8}(G_k), k\in[1,K] \tag{1}$$

$$A_k=\mathrm{ResNet}(G_k) \tag{2}$$

$$A_j^k=\mathrm{ResNet}(R_j^k) \tag{3}$$

1.2 特征提取

特征提取阶段由视觉特征提取与文本特征提取两部分组成。在视觉特征提取阶段，针对给定的图像 G_k 和感兴趣区域 R^k_j，使用 ResNet-152 网络分别对图像和感兴趣区域提取视觉特征。对于图像 G_k，直接获取其网络最后一层卷积层的输出结果作为基础视觉特征；对于感兴趣区域 R^k_j，则在提取卷积层特征后，进一步执行空间维度的均值池化操作，以压缩空间信息并保留关键特征。最后，通过线性映射层将上述两类视觉特征投影至与文本特征一致的特征空间，得到每张图像对应的视觉特征$\boldsymbol{v}^{G_k}\in\mathbb{R}^{49\times d}$与每个感兴趣区域对应的视觉特征$\boldsymbol{v}^{R_j^k}\in\mathbb{R}^d$，其中 d 表示文本特征维度。

在文本特征提取阶段，考虑到 RoBERTa 模型在自然语言处理领域的优异性能，本研究选用 xlm-Roberta-large 模型进行文本特征提取。此外，鉴于构建辅助句子在前人研究中取得的良好效果 [3]，本文在文本特征提取阶段将图像方面类型信息融入模型输入。具体而言，模型构建的辅助句子包括四个部分：方面类别 C^n、上下文 S、图像类别 A^G 和 RoI 类别 A^R，将其连接起来形成模型的输入序列：X=[CLS] C^n [SEP] S [SEP]A^G [SEP] A^R[SEP]。将该输入序列送入 xlm-Roberta-large 模型后，取 [CLS] 令牌对应的最终隐藏状态，作为文本的特征表示，用公式表示为：

$$\boldsymbol{H}_T=\mathrm{Roberta}(X) \tag{4}$$

$$\boldsymbol{H}_{[\mathrm{CLS}]}=\boldsymbol{H}_T^0 \tag{5}$$

1.3 细粒度融合阶段

在细粒度融合阶段，模型应用跨模态注意力机制（cross-modal attention mechanism）来建模上下文与图像之间的交互关系，具体公式为：

$$\boldsymbol{H}_{G_k}=\mathrm{CM}-\mathrm{Attention}(\boldsymbol{H}_T,\boldsymbol{v}^{G_k},\boldsymbol{v}^{G_k}) \tag{6}$$

式中：输入序列 $\boldsymbol{H}_T$ 为跨模态注意力机制的查询向量；图像特征$\boldsymbol{v}^{G_k}$为跨模态注意力机制的值向量和键向量；$\boldsymbol{H}_{G_k}\in\mathbb{R}^{L_\mathrm{seq}\times d}$表示基于图像 G_k 生成的上下文表征，接着，模型获取第一个维度向量$\boldsymbol{H}^0_{G_k}$，此向量是指基于单张图像生成的上下文表征的首个维度向量，最终，模型将一组图像 G 与文本的跨模态注意力表征拼接起来，生成统一的图像基于上下文表征$\boldsymbol{H}_G$。

$$\boldsymbol{H}_G=\{\boldsymbol{H}^0_{G_1},\boldsymbol{H}^0_{G_2},\ldots,\boldsymbol{H}^0_{G_K}\} \tag{7}$$

式中：$\boldsymbol{H}_G\in\mathbb{R}^{K\times d}$表示文本与 K 张图像经注意力机制融合后的特征向量。此外，模型也应用跨模态注意力机制来建模上下文与 RoI 之间的交互关系。

$$\boldsymbol{H}_{R^k}=\mathrm{CM}-\mathrm{Attention}(\boldsymbol{H}_T,\boldsymbol{v}^{R^k},\boldsymbol{v}^{R^k}) \tag{8}$$

式中：$\boldsymbol{v}^{R^k}\in\mathbb{R}^{J\times d}$表示第 k 个图像的感兴趣区域的特征；$\boldsymbol{H}_{R^k}\in\mathbb{R}^{L_\mathrm{seq}\times d}$表示基于 RoI 图像生成的上下文表征。接着，模型获取第一个标记的表征（即$\boldsymbol{H}^0_{R^K}$），该表征是将文本表征与 RoI 图像表征融合后在所有预定义方面类别上的核心关联信息。最终，将所有这些表征拼接起来，生成统一的 RoI 基于上下文表征。

$$\boldsymbol{H}_R=\{\boldsymbol{H}^0_{R^1},\boldsymbol{H}^0_{R^2},\ldots,\boldsymbol{H}^0_{R^K}\} \tag{9}$$

式中：$\boldsymbol{H}_R \in \mathbb{R}^{K\times d}$表示文本表征与 K 张 RoI 图像特征经注意力机制融合后的特征向量。

在最终阶段，将文本特征 $\boldsymbol{H}_{[CLS]}$、$\boldsymbol{H}_G$ 和 $\boldsymbol{H}_R$ 进行拼接，形成统一的多模态表征 $\boldsymbol{H}^A$。表征被传递至另一个多模态注意力层，以捕捉不同模态之间的依赖关系与交互作用。

$$\boldsymbol{H}^A = \text{concat}(\boldsymbol{H}_{[CLS]}, \boldsymbol{H}_G, \boldsymbol{H}_R) \tag{10}$$

$$\boldsymbol{H} = \text{MM}-\text{Attention}(\boldsymbol{H}^A, \boldsymbol{H}^A, \boldsymbol{H}^A),\ \boldsymbol{H}^A \in \mathbb{R}^{(1+2k)\times d} \tag{11}$$

在获取多模态自注意力输出 $\boldsymbol{H}$ 的首个令牌 $\boldsymbol{H}^0$，作为最终隐藏状态后，将该状态传入线性函数进行维度转换，随后通过 softmax 函数完成情感分类操作，用公式表示为：

$$P(y) = \text{softmax}(\boldsymbol{W}^T\boldsymbol{H}^0 + \boldsymbol{b}) \tag{12}$$

式中：$\boldsymbol{W}^T$ 和 $\boldsymbol{b}$ 为可学习参数。

为优化模型框架中的所有可学习参数，本研究采用最小化标准交叉熵损失函数的方式实现参数更新，用公式表示为：

$$\text{loss} = -\frac{1}{D}\frac{1}{N}\sum_{i=1}^{D}\sum_{j=1}^{N}\text{logp}(y^{i,j}) \tag{13}$$

式中：D 是样本数量；N 是预定义方面的数量。

2 实验及结果分析

为保障文本与图像数据的真实性及可靠性，本研究从提供酒店在线预订服务的大型在线旅游平台中，收集并标注了 2 000 条含图像的英文评论数据。数据筛选过程中，限定每条评论最多包含 6 张图像，且评论文本长度不超过 128 。结合前人研究成果 [8] 与本数据集特征，本研究选用“位置、食物、房间、设施、服务、公共区域”作为方面类别标准。这六类方面类别能尽可能地覆盖酒店领域用户的核心关注点，且能独立应用于文本数据或图像数据的标注任务。

本研究对单一样本（文本 - 图像对）的标注采用三阶段递进式设计，具体包括文本标注、图像标注与文本 - 图像联合标注。在文本标注阶段，需将文本情感对应至六个预定义方面类别，并标注各方面类别下的情感极性。其中，情感极性划分为 4 类，具体定义为：0 代表“不相关”，1 代表“消极”，2 代表“中性”，3 代表“积极”。在图像标注阶段，需要在图像中检测感兴趣区域（RoI），并将图像及 RoI 标注为预定义的方面类别。需说明的是，酒店领域视觉内容的情感传达具有较强模糊性，准确判断并标注图像及 RoI 的情感标签存在挑战，为避免标注者间因主观判断差异产生误导性信息，本研究未对图像及 RoI 进行情感标签标注。

在文本 - 图像联合标注阶段，核心任务是为图像标注阶段识别出的方面类别补充情感标签。若某方面类别在文本标注阶段已获得对应的情感标签，则直接沿用该标签，若某方面类别仅在图像标注阶段新出现，则以文本模态的语义信息为依据，为其标注情感标签。数据集的完整标注流程如图 2 所示。

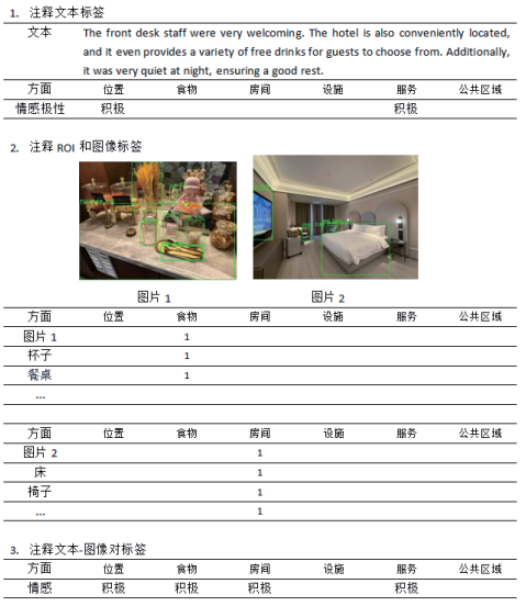

图 2 数据集的注释过程

数据集的详细信息如表 1 所示，该数据集一共包含 2 000 条标注好的带图像的英文用户评论数据。按照 8:1:1 的比例划分为训练集、验证集以及测试集。从情感极性来看，绝大多数情感标签呈现积极倾向，存在明显的失衡现象。

表 1 多模态方面类别数据集

数据集	评论	图像	ROI	积极	中性	消极
训练集	1 600	3 029	4 965	3 608	753	455
验证集	200	375	530	465	104	56
测试集	200	356	595	428	99	64

图 3 展示了多模态数据集中不同方面所占比例，从图中分析可知，收集到的数据集存在方面类别失衡问题。各方面类别中所占比例前三分别是：房间、公共区域与服务，三方面类别合计占比超 70%。

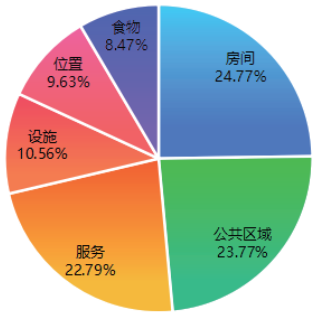

图 3 多模态各方面类别占比

接下来，介绍模型使用的评估指标，在各类分类任务中，精确率、召回率及 F_1 值是常用的评估标准。然而，根据上述分析可知，数据集存在类别分布与情感极性失衡的问题，平

均宏观 F_1 分数能平等对待每个类别，避免多数类对指标的过度影响，更客观反映模型对少数类的预测性能，因此模型选择该指标作为评估模型性能的核心指标，同时将平均宏观精确率与平均宏观召回率作为补充参考指标。其计算公式为：

$$\text{Precision}_M = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i} \tag{14}$$

$$\text{Recall}_M = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i} \tag{15}$$

$$F_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2\text{TP}_i}{2\text{TP}_i+\text{FN}_i+\text{FP}_i} \tag{16}$$

式中：对于多类别分类任务中的第 i 个类别，TP_i 表示该类别被正确预测为自身的样本数量；FP_i 表示其他类别被错误预测为该类别的样本数量；FN_i 表示第 i 类实际存在但被错误预测为其他类别的样本数量。上述评估指标通过先类别内计算、再类别间平均的宏观策略，适配了数据集的失衡特性，为模型性能评估提供了更可靠的量化依据。

由表 2 所示的实验结果可知，与仅以文本数据为输入的基准模型相比，引入额外图像信息的多模态模型在各项评估指标上均表现更优。这一结果充分表明，图像模态能够提供文本模态之外的互补性信息，此类信息可作为文本语义的有效补充，为多模态方面级情感分析任务构建更丰富的语义理解基础，进而缓解单一文本模态在情感表达细节捕捉上的局限性。

表 2 多模态方面类别数据集上的实验结果

形式	模型	精确率 /%	召回率 /%	F_1 值 /%
文本	roberta	70.14	67.42	67.89
文本－图像对	YL-MACSA	71.27	68.67	69.59

各方面的实验结果如表 3 所示，不同方面的情感分析性能存在差异，“房间”方面的各项指标相对领先，而“设施”方面的性能有待进一步提升，这也反映出模型在不同方面的情感特征捕捉与分析能力有所不同。

表 3 不同方面的实验结果

方面	精确率 /%	召回率 /%	F_1 值 /%
位置	73.37	71.70	72.40
食物	73.72	69.89	71.66
房间	77.20	76.61	76.85
设施	64.09	58.17	60.29
服务	73.83	74.47	74.02
公共区域	65.41	61.16	62.31

3 结语

本文设计并实现了基于 YOLOv8 的多模态方面类别情感分析模型（YL-MACSA），该模型通过 YOLOv8 提取图像中的感兴趣区域（RoI），结合 ResNet-152 获取图像与 RoI 的类别信息，再将视觉语义融入 XLM-RoBERTa-large 的文本特征提取过程，最终通过跨模态注意力机制，实现文本特征、图像特征及 RoI 特征的细粒度融合，形成统一的多模态表征以完成情感分类任务。模型验证了图像细粒度特征对多模态情感分析任务的增益作用，也证明了“方面类别”作为跨模态对齐桥梁的有效性，为多模态情感分析的细粒度融合提供了可行思路。

参考文献：

[1] 汤浩杰 . 基于社交网络和签到数据的兴趣点推荐系统设计与实现 [D]. 北京 : 北京邮电大学 ,2021.

[2]ZHANG W X, LI X, DENG Y, et al. A survey on aspect-based sentiment analysis: tasks, methods, and challenges[J]. IEEE transactions on knowledge and data engineering, 2023, 35(11): 11019-11038.

[3]YANG H, SI Z M, ZHAO Y Y, et al. MACSA: a multimodal aspect-category sentiment analysis dataset with multimodal fine-grained aligned annotations[J]. Multimedia tools and applications, 2024, 83: 81279.

[4]XU N, MAO W J, CHEN G D. Multi-interactive memory network for aspect based multimodal sentiment analysis[EB/OL].(2019-07-17)[2026-04-02].https://www.semanticscholar.org/paper/Multi-Interactive-Memory-Network-for-Aspect-Based-Xu-Mao/594971a09fe6ee479e862379eb4fa97c3272aeed.

[5]YU J F, JIANG J, XIA R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification[J].IEEE/ACM transactions on audio, speech, and language processing, 2019,28: 429-439.

[6]YU J F, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. POD Publisher: SPIE, 2019: 5408-5414.

[7]KHAN Z, FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation[EB/OL]. (2021-08-05)[2026-04-02].https://doi.org/10.48550/arXiv.2108.01682.

[8]NGUYEN Q H, NGUYEN M V T, NGUYEN K V. New benchmark dataset and fine-grained cross-modal fusion framework for vietnamese multimodal aspect-category sentiment analysis[EB/OL]. (2024-05-01)[2026-04-02]. https://doi.org/10.48550/arXiv.2405.00543.

【作者简介】

张晓薇（1998—），女，山西太原人，硕士，助教，研究方向：自然语言处理。

（收稿日期：2025-09-08 修回日期：2026-04-09）