基于自适应噪声和方面图关联学习增强多模态方面级情感分析

黄 辰[1,2,3,4],刘会杰[1,2,3,4],张 龑[1,2,3,4*],杨 超[1,2,3,4],宋建华[2,3,4,5]

(1. 湖北大学计算机学院,湖北武汉 430062;2. 智能感知系统与安全教育部重点实验室,湖北武汉 430062;3. 大数据智能分析与行业应用湖北省重点实验室,湖北武汉 430062;4. 湖北省高校人文社科重点研究基地-绩效评价信息管理研究中心,湖北武汉 430062;5. 湖北大学网络空间安全学院,湖北武汉 430062)

摘 要: 多模态方面级情感分析(Multimodal Aspect-Based Sentiment Analysis,MABSA)旨在从多模态输入数据中准确识别方面术语并判定其情感极性. 现有研究致力于融合多模态信息以提升情感分析性能. 然而,在面临多方面和多情感场景时,它们仍然面临两个关键挑战:(1)缺乏对多模态输入数据中方面术语的全面感知;(2)存在情感语义偏差,现有模型倾向于关注与特定方面术语关联性强的情感语义,而忽略了关联性较低但同样重要的情感语义. 为了克服这些问题,本文提出了一种结合自适应噪声和方面图关联学习的新型多模态方面级情感分析方法(Adaptive Noise and Aspect Graph Association Learning,ANAGAL),旨在增强多方面和多情感场景下的分析性能. 具体而言,通过专门设计的自适应噪声增强模块以补充方面信息,从而增强模型对方面术语的感知能力,并提高模型鲁棒性. 此外,设计方面图关联学习模块来关联所有方面术语,并学习与之相关的情感语义. 同时,引入额外的参数进行情感校准,使模型能够学习更多常见的情感语义偏差,从而更准确地捕捉方面术语及其对应的情感极性. 在公共数据集上的大量实验评估表明,ANAGAL在克服这些挑战方面表现优异. 与现有基线模型相比,ANAGAL在Twitter-2015和Twitter-2017数据集上将精确率分别提升了1.46个百分点和1.56个百分点,在MASAD(Multimodal Aspect Sentiment Analysis Dataset)和EmoMeta数据集上将精确率提升了2.48个百分点和1.55个百分点.

关键词: 多模态;方面级情感分析;预训练语言模型;噪声增强;方面图关联学习;图注意力网络

基金项目: 湖北省重大攻关项目(JD)(No.2023BAA018);湖北省科技计划重大科技专项(No.2024BAA008);武汉市知识创新专项项目(No.202311901251001)

中图分类号: TP391;TP399 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3397-13

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250533

Enhancing Multimodal Aspect-Based Sentiment Analysis with Adaptive Noise and Aspect Graph Association Learning

HUANG Chen[1,2,3,4], LIU Hui-jie[1,2,3,4], ZHANG Yan[1,2,3,4*], YANG Chao[1,2,3,4], SONG Jian-hua[2,3,4,5]

(1. *School of Computer Science*, *Hubei University*, *Wuhan*, *Hubei* 430062, *China*;
2. *Key Laboratory of Intelligent Sensing System and Security*, *Ministry of Education*, *Wuhan*, *Hubei* 430062, *China*;
3. *Hubei Key Laboratory of Big Data Intelligent Analysis and Application*, *Wuhan*, *Hubei* 430062, *China*;
4. *Hubei Province Project of Key Research Institute of Humanities and Social Sciences at Universities-RCIMPE*, *Wuhan*, *Hubei* 430062, *China*;
5. *School of Cyber Science and Technology*, *Hubei University*, *Wuhan*, *Hubei* 430062, *China*)

Abstract: Multimodal aspect-based sentiment analysis (MABSA) aims to accurately identify aspect terms and determine their sentiment polarity from multimodal input data. Existing studies focus on integrating multimodal information to improve sentiment analysis performance. However, they still face two critical challenges in multi-aspect and multi-sentiment scenarios: (1) a lack of comprehensive perception of aspect terms in multimodal input data; and (2) sentiment semantic bias, where current models tend to focus on sentiment semantics strongly correlated with specific aspect terms, while ignoring weakly associated yet equally important sentiment cues. To address these issues, we propose a novel multimodal aspect-based sentiment analysis method, ANAGAL (Adaptive Noise and Aspect Graph Association Learning), which integrates

收稿日期:2025-06-22;录用日期:2025-08-27;责任编辑:覃怀银
*通讯作者:张龑

adaptive noise handling and aspect-graph association learning to enhance analytical performance in scenarios involving multiple aspects and multiple sentiments. Specifically, an adaptive noise enhancement module is designed to supplement aspect information, thereby improving the model's aspect perception and robustness. In addition, an aspect graph correlation learning module is introduced to associate all aspect terms and learn related sentiment semantics. Extra parameters are further incorporated to calibrate sentiment representations, enabling the model to capture more generalized sentiment biases and better identify sentiment polarity associated with each aspect term. Extensive experimental evaluations on public datasets demonstrate that ANAGAL performs exceptionally well in addressing these challenges. Compared to existing state-of-the-art MABSA models, ANAGAL improves precision by 1.46 percentage points and 1.56 percentage points on the Twitter-2015 and Twitter-2017 datasets, and by 2.48 percentage points and 1.55 percentage points on the MASAD (Multimodal Aspect Sentiment Analysis Dataset) and EmoMeta datasets.

Key words: multimodal; aspect-based sentiment analysis; pre-trained language model; noise augmentation; aspect-graph association learning; graph attention network

Foundation Item(s): Major Project of Hubei Province (JD) (No.2023BAA018); Major Science and Technology Special Project of Hubei Science and Technology Plan (No.2024BAA008); Wuhan Knowledge Innovation Special Project (No.202311901251001)

1 引言

多模态情感分析(Multimodal Sentiment Analysis, MSA)任务旨在对整个多模态输入进行情感极分类,关注整体的情感倾向. 多模态方面级情感分析(Multimodal Aspect-Based Sentiment Analysis, MABSA)是情感计算领域中一项复杂且更具挑战性的研究. MABSA通过结合多种模态信息(如文本、图像、音频和视频等),旨在从多模态输入数据中识别所有方面术语并判定其对应的情感极性[1,2],从而实现细粒度的情感分析. 例如,对于文本模态输入"食物很美味,但服务很差",MSA的目标是判断整体情感倾向为负面,而MABSA则需要识别"食物"和"服务"两个方面术语,并判定"食物"为正面情感,"服务"为负面情感. 此外,与依赖单一文本数据的传统情感分析方法不同[3],MABSA通常包含三个关键子任务:多模态方面术语提取(Multimodal Aspect Term Extraction, MATE)、多模态方面情感分类(Multimodal Aspect Sentiment Classification, MASC)以及联合多模态方面情感分析(Joint Multimodal Aspect Sentiment Analysis, JMASA)[4,5]. 其中,MATE旨在识别和提取给定多模态数据中的所有方面术语[6]. MASC旨在同时处理方面术语提取和情感分类,实现统一分析[7]. JMASA则以端到端的方式直接联合学习方面术语及其对应的情感分类[8].

近期的研究工作广泛依赖于融合图像和文本模态信息来进行情感分析. Ju等人[9]首次提出了JMASA任务,并设计了结合图像-文本关系的联合学习方法,用于评估视觉内容对方面情感判别的贡献. 随后,Ling等人[5]进一步提出了统一的多模态编码器-解码器架构,覆盖各类预训练任务. 最近,Yang等人[10]设计了一种多任务学习框架,旨在从图像-文本对中提取方面-情感对. Zhou等人[11]提出了面向方面的网络,以缓解视觉与文本噪声对情感分析的干扰. 为提升模型在多方面多情感场景下的识别能力,Zhu等人[12]引入了方面增强与文本简化机制,以准确捕捉方面及其相应的情感.

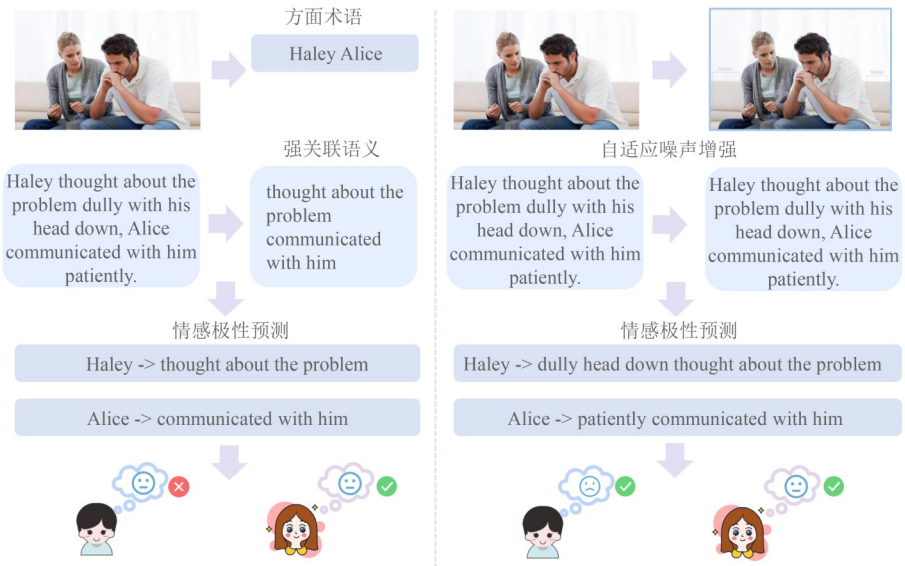
尽管已有研究在MABSA任务上取得了显著进展,但现有方法仍然面临两大局限:

(1)缺乏对多模态输入数据中方面术语的全面感知. 在现实场景中,图像和文本内容往往包含大量噪声,一些图像和文本区域可能误导模型对情感的判断,另一些则可能有助于理解. 然而,大多数先前的MABSA方法将噪声视为干扰因素,致力于在MABSA任务中抑制它. 这样可能会限制模型对方面术语的全面感知,并降低模型鲁棒性.

(2)存在情感语义偏差. 现有模型倾向于关注与特定方面术语高度关联的情感语义,忽略了关联性较低但同样重要的信息.

图1展示了现有MABSA方法和本文提出的ANAGAL(Adaptive Noise and Aspect Graph Association Learning)的直观对比,一个多方面多情感的案例,包含"Alice"和"Haley"两个方面术语,分别对应"中性"与"负面"情感. 然而,现有模型往往仅捕捉诸如"Alice手势交流"和"Haley思考问题"等强关联语义,导致错误地将"Haley"的情感判定为"中性". 而关联性较弱的情感语义,如"沉闷地""耐心地"等副词,以及"低着头"等状态描述词,虽然不直接关联特定方面术语,但蕴含情感倾向,能有效辅助模型准确地判定方面情感. 因此,忽略关联性较弱的情感语义会导致方面术语学习受限,从而影响MABSA任务的整体性能.

为了解决上述问题,本文提出了一种结合自适应噪声和方面图关联学习的新型多模态方面级情感分析方法ANAGAL. 首先,ANAGAL采用预训练语言模型对图像和文本信息进行编码. 其次,设计了一种自适应噪



(*a*) 现有的MABSA方法　　(*b*) 本文的ANAGAL方法

图1　现有的MABSA方法和本文的ANAGAL方法的直观对比

声增强模块,在训练过程中采样图像和文本信息的噪声来补充方面信息. 为了模拟真实环境,噪声采样被设置为自适应,动态采样数据量和比例,从而增强模型对方面术语的感知能力,并提高模型鲁棒性. 然后,提出了一个方面图关联学习模块,以关联所有方面术语,并学习与之相关的情感语义. 同时,引入额外的参数进行情感校准,使模型能够学习更多常见的情感语义偏差,从而更准确地捕捉方面术语及其对应的情感极性. 最后,在四个数据集上执行MATE、MASC和JMASA任务. 验证了该方法的有效性.

针对上述挑战,本研究的主要贡献如下:

(1)针对复杂的多方面多情感场景,本文提出了一种结合自适应噪声增强和方面图关联学习的新型多模态方面级情感分析方法ANAGAL. 该方法利用预训练语言模型进行模态特征编码,并通过自适应噪声采样策略补充方面信息,从而提升模型对方面术语的感知能力,并提高模型鲁棒性.

(2)设计了新的方面图关联学习模块,以关联所有方面术语,并学习与之相关的情感语义. 同时,引入额外的参数进行情感校准,使模型能够学习更多常见的情感语义偏差,从而准确捕捉方面术语对应情感极性.

(3)在四个基准数据集上进行了广泛的实验,全面分析和验证了ANAGAL的有效性、优越性和鲁棒性. 并与现有的MABSA方法进行对比,结果表明ANAGAL优于现有最佳方法.

2　相关工作

多模态方面级情感分析(MABSA)旨在从多模态输入数据中准确识别所有方面术语并判断其情感倾向,是一种细粒度情感分析任务. Lu等人[13]对该任务的定义、方法和挑战进行了综述. Zhao等人[14]从多模态情感识别角度提供理论支持. 近年来,针对MABSA的研究越来越丰富,主要可分为MATE、MASC和JMASA等三个子任务.

多模态方面术语提取(MATE)旨在从多模态数据中提取与观点相关的方面术语[15,16]. Yang等人[10]提出CMMT(Cross-Modal Multitask Transformer)以学习方面与情感感知的模态内表征. Yu等人[17]使用HIMT(Hierarchical Interactive Multimodal Transformer)对图像-文本交互建模,并提取具有语义概念的显著特征. Weng等人[18]采用全新方法MIECF(Multi-faceted Information Extraction and Cross-mixture Fusion)整合全局和局部多模态表示. Guo等人[19]则引入MGICL(Multi-Grained Interaction Contrastive Learning)用于MATE的多粒度对比学习. 尽管上述方法在方面术语提取任务中取得了一定进展,但大多忽视现实场景中的噪声信息,这可能会在一定程度上限制模型对方面术语的全面感知能力,并降低模型鲁棒性. 而本文提出的自适应噪声增强通过采样噪声数据来补充方面信息,可以有效解决这一问题.

多模态方面情感分类(MASC)旨在识别与分类多模态输入数据中的情感类别[20,21]. Yu等人[22]利用多级情感区域的视觉信息进行情感分析. 尽管取得了进展,但文本模态信息往往被忽视. Fan等人[23]开发了用于多任务学习的多模态双重原因分析框架来进行情感检测. Khan等人[24]利用BERT(Bidirectional Encoder Representations from Transformers)结合对象感知转换进行多模态目标情感分类. Jia等人[25]开发新型情感区域识别与融合网络ARFN(Affective Region and Fusion Network)以注重视觉文本的多模态协调. Yu等人[26]设计

了用于多任务学习的图像-目标匹配网络ITM(Image-Target Matching),以解决图像-文本对齐和情感分类问题.与上述方法不同,本文设计了方面图关联学习模块,以关联所有方面术语,并学习与之相关的情感语义.同时,引入额外的参数进行情感校准,使模型能够学习更多常见的情感语义偏差,以更准确地捕捉方面术语及其对应的情感极性.

联合多模态方面情感分析(JMASA)旨在联合识别多模态输入数据中的方面术语和情感极性[27,28].Ju等人[9]提出了多模态联合方法JML(Joint Multimodal Learning),同时完成方面术语提取与情感分类.Ling等人[5]设计统一的多模态编码器-解码器架构VLP-MABSA(Vision-Language Pre-training for Multimodal Aspect-Based Sentiment Analysis)用于多任务预训练.Zhou等人[11]提出了面向方面的网络AOM(Aspect-Oriented for Multimodal),用于方面相关的语义情感检测.Yang等人[10]提出了CMMT,从视觉和文本内容中联合提取图像-文本对.Zhu等人[12]提出方面增强与文本简化模型AETS(Aspect Enhancement and Text Simplification)用于准确捕捉方面及其相应情感.与上述研究不同,本文提出了一种新的多模态方面级情感分析模型,具有自适应噪声采样和方面图关联学习,可同时处理MATE、MASC和JMASA任务.

3 方法设计与分析

本节详细阐述了ANAGAL模型.ANAGAL的整体架构如图2所示,主要包含三个模块.(1)自适应噪声增强模块:获取包含自适应噪声的方面信息.(2)方面图关联学习模块:关联所有方面术语,并学习与之相关的情感语义.同时,引入额外的学习参数进行情感校准,使模型能够学习更多常见的情感语义偏差.(3)学习目标模块:同时处理方面术语提取和情感分类任务.

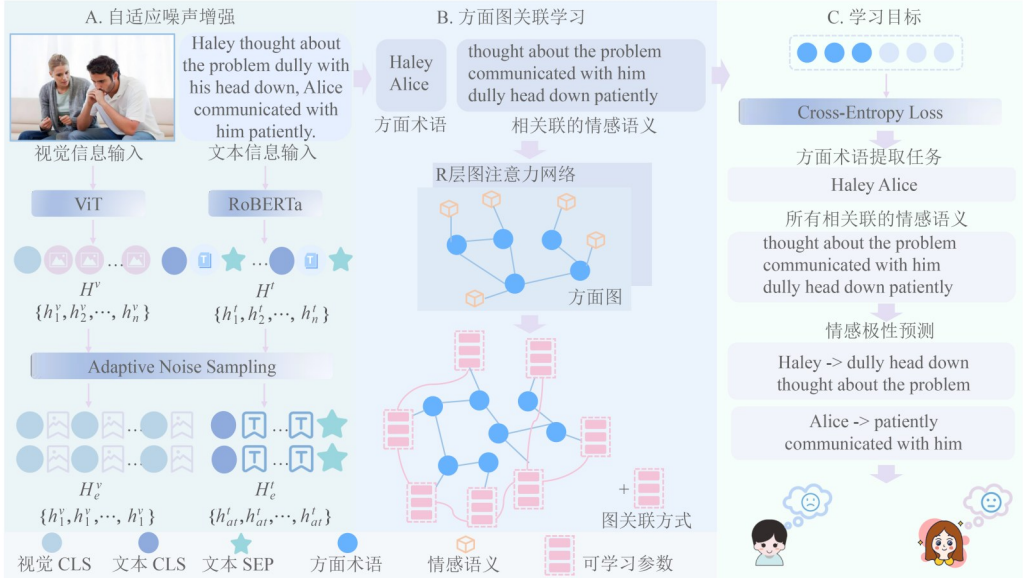


图2 ANAGAL的整体架构图

3.1 初始准备

形式上,本文专注于从图像和文本模态信息上提取所有特定方面术语,并准确识别其对应的情感极性.在多模态方面级情感分析任务中定义了三个子任务:MATE、MASC和JMASA.给定一个包含N个样本的方面级情感分析数据集$D=\{T_i,V_i,A_i,S_i\}_{i=1}^{N}$.其中每个样本都由一个包含$n$个词的文本序列$T=\{T_1,T_2,\cdots,T_n\}$和一个图像$V$组成.最终目标是从图像-文本对中识别所有方面术语$A=\{a_1,a_2,\cdots,a_m\}$及其对应的情感极性$S=\{s_1,s_2,\cdots,s_m\}$.$m$表示文本序列$T$中的方面术语数量.具体的输入输出如下:

$$\text{MATE: input}=\{T_1,T_2,\cdots,T_n,V\},$$
$$\text{output}=\{a_1,a_2,\cdots,a_m\};$$
$$\text{MASC: input}=\{T_1,T_2,\cdots,T_n,V,a_1,a_2,\cdots,a_m\},$$
$$\text{output}=\{s_1,s_2,\cdots,s_m\};$$
$$\text{JMASA: input}=\{T_1,T_2,\cdots,T_n,V\},$$
$$\text{output}=\{(a_1,s_1),(a_2,s_2),\cdots,(a_m,s_m)\}.$$

其中,对于方面术语$a_i\in\{B,P,N\}$,B表示方面术语的起始位置;P表示方面词的内部位置;N表示除方面词外的其他部分.对于情感极性$s_i\in\{\text{Pos},\text{Neu},\text{Neg}\}$,Pos表示积极;Neu表示中性;Neg表示消极.

3.2 自适应噪声增强

自适应噪声增强旨在采样噪声数据来补充方面信息,从而提升模型对方面术语的感知能力,并提高模型

鲁棒性. 遵循以往的研究工作[12],为了确保不同方法之间性能比较的一致性和准确性. 本文采用了大多数基线方法常用的编码器以及提取方式,分别使用RoBERTa (Robustly optimized BERT approach)[29] 和 ViT (Vision Transformer)[30]对文本和图像进行特征编码. 对于文本序列,在其起始和结束处分别插入特殊标记[CLS]和[SEP],以表示文本的开始和结束. 对于图像序列,仅在其起始处插入[CLS]标记,作为图像的开始标识. 本文使用RoBERTa和ViT编码器分别获得了文本隐藏状态 $H^t=\{h_1^t,h_2^t,\cdots,h_n^t\}$ 和图像隐藏状态 $H^v=\{h_1^v,h_2^v,\cdots,h_n^v\}$,计算方式如下所示:

$$H^t=\text{RoBERTa}(T_t) \tag{1}$$

$$H^v=\text{ViT}(V_v) \tag{2}$$

其中,T_t 表示文本序列;V_v 表示图像序列;$H^t\in R^{n\times d}$ 表示文本隐藏状态;$H^v\in R^{n\times d}$ 表示图像隐藏状态;n 表示词的数量;d 表示隐藏状态维度.

已有研究表明[31,32],模型可以容忍一定程度的噪声,这不仅增强了模型鲁棒性,同时还不会显著降低方面术语的感知能力. 受此启发,本文设计了一种自适应噪声采样策略,引入图像和文本中的噪声数据来尝试提升模型对方面术语的感知能力,并提高模型鲁棒性.

首先,利用spaCy工具根据词性信息从文本中提取名词,构建候选方面术语集 $\text{AT}=\{\text{at}_1,\text{at}_2,\cdots,\text{at}_k\}$. 其中 at_i 表示第 i 个候选方面术语,具有名词词性且来自文本序列 T. 然后,本文针对不同的数据集设计两种不同的采样策略:平均噪声采样和高斯噪声采样. 平均噪声采样有助于平滑噪声并减轻异常值对模型的影响,高斯噪声采样则引入随机扰动以增强模型泛化能力. 平均噪声采样的计算方法如下:

$$\mathbf{ns}_a=S_i\left(\text{RV}_{\min}+(\text{RV}_{\max}-\text{RV}_{\min})\cdot\text{RU}\right) \tag{3}$$

其中,S_i 为可学习的噪声缩放因子;$\text{RV}_{\min}$ 和 $\text{RV}_{\max}$ 分别表示随机变量区间上的最小值和最大值;RU为服从区间 $[\text{RV}_{\min},\text{RV}_{\max}]$ 上的均匀分布随机变量;$\mathbf{ns}_a$ 为最终生成的噪声输入向量. 高斯噪声采样的计算方法如下所示:

$$\mathbf{ns}_g=S_i(\alpha_{\text{em}}\cdot g+\delta_a),g\sim G(0,\boldsymbol{C}) \tag{4}$$

其中,在训练过程中自适应调整嵌入方面信息中的噪声比例,α_{em} 和 δ_a 分别表示原始噪声嵌入方面信息的标准差和均值;$G(0,\boldsymbol{C})$ 表示均值为0、协方差矩阵为 $\boldsymbol{C}$ 的高斯分布;g 是从高斯分布中采样的随机变量;$\mathbf{ns}_g$ 为最终生成的噪声输入向量. 为了确定采样向量的含噪声比例,本文采用如下自适应噪声采样:

$$H_{at}^t=\begin{cases}(1-S_i)\cdot H^t+\mathbf{ns}_m, & \text{if}\,\lambda>S_n\\ H^t, & \text{otherwise}\end{cases} \tag{5}$$

为实现自适应全局噪声采样,这里引入了随机因子 λ,并将其与嵌入向量的采样因子 S_n 进行比较. 随机因子 λ 在[0,1]区间内均匀分配,而采样因子 S_n 也在训练过程中自适应地调整到[0,1]区间内. 通过比较二者的大小,决定是否将噪声采样向量 $\mathbf{ns}_m$ 补充到文本隐藏状态 H^t 中,从而形成增强的候选方面术语文本隐藏状态 H_{at}^t.

因此,使用自适应噪声采样对文本隐藏状态进行增强,形成更加鲁棒的文本表示. 具体如下所述:

$$H_e^t=\text{NoiseSampling}(H^t) \tag{6}$$

其中,NoiseSampling(·)表示上述设计的自适应噪声采样操作;H_e^t 表示增强后的文本表示. 此外,图像隐藏状态 H^v 中[CLS]位置的向量 H_1^v 表示图像的全局信息,通过复制扩展为与增强文本表示相同的序列 $H_c^v=\{h_1^v,h_1^v,\cdots,h_1^v\}$,该序列同样使用自适应噪声采样操作进行增强,以便获得感知候选方面术语的图像表示 H_e^v,具体操作如下所示:

$$H_e^v=\text{NoiseSampling}(H_c^v) \tag{7}$$

3.3　方面图关联学习

方面图关联学习旨在关联所有方面术语,并学习与之相关的情感语义. 同时,引入额外的可学习参数进行情感校准,使模型能够学习更多常见的情感语义偏差,从而更准确地捕捉方面术语及其对应的情感极性.

首先,考虑到候选方面术语周围的情感语义(形容词、副词)是决定情感极性的关键因素,本文从位置特征与属性特征两个角度对方面术语进行建模. 候选方面术语的属性特征 $A^{b+c}\in\mathbf{R}^{d_t}$ 表示如下所示:

$$A_i^{b+c}=E_b^i\cdot H_b^i\oplus E_c^i\cdot H_c^i \tag{8}$$

其中,H_b^i 和 H_c^i 表示第 i 个候选方面术语的属性和类别名称;E_b^i 和 E_c^i 分别为其嵌入表示;$\oplus$ 表示拼接操作,属性特征维度为 d_t. 候选方面术语的位置特征 $O^{p+s}\in\mathbf{R}^{d_a}$ 可表示为如下所示:

$$O_i^{p+s}=\boldsymbol{W}_Q(\boldsymbol{W}_V\cdot o_p^i\oplus\boldsymbol{W}_Q\cdot o_s^i) \tag{9}$$

其中,o_p^i 和 o_s^i 表示候选方面术语的位置信息;$\boldsymbol{W}_Q$ 和 $\boldsymbol{W}_V$ 为可学习权重矩阵;位置特征维度为 d_a. 最终,每个候选方面术语的联合特征 h_i 可表示为如下所示:

$$h_i=[O_i^{p+s},A_i^{b+c}] \tag{10}$$

再将所有候选方面术语组成方面术语特征集合 F^v,可表示如下:

$$F^v=O_o=[h_1,h_2,\cdots,h_N] \tag{11}$$

其中,$[h_1,h_2,\cdots,h_N]$ 表示所有候选方面术语;N 表示候选方面术语数量.

其次,为了建立基于上述特征的可视图,这里定义一个具有 $2N$ 个节点的可视图 g_a,记为 $G_n^g=\{g_1,g_2,\cdots,g_{2N}\}$. 节点集合初始化为方面术语表示 $\{O_1^{p+s},A_1^{b+c},O_2^{p+s},A_2^{b+c},\cdots,O_N^{p+s},A_N^{b+c}\}$. 然后,定义边

$E_{ij}\in\mathbf{R}^{2N\times 2N}$，其中 $i,j\in\{1,2,\cdots,2N\}$，$\mathrm{Csim}(\cdot)$ 表示计算属性向量之间的余弦相似度函数，$\mathrm{GS}_{ij}^{\mathrm{IoU}}$ 表示基于位置特征的几何相似度．计算方式如下：

$$E_{ij}=\begin{cases}\mathrm{Csim}(A_i^{b+c},A_j^{b+c}), & \text{if } i\%2=j\%2\\ \mathrm{GS}_{ij}^{\mathrm{IoU}}(O_i^{p+s},O_j^{p+s}), & \text{if } i\%2=j\%2=1\\ 0, & \text{otherwise}\end{cases}\tag{12}$$

此外，将该图与 R 层自适应图注意力网络相结合，该网络能够动态注入注意力分数，且自动调整不同类型节点之间的融合程度．如下所示：

$$\beta_{ij}^{d}=\frac{\mathrm{ReLU}\left(\beta_d^T(m_i^d\cdot\boldsymbol{W}_{\mathrm{proj}}^{n1}\oplus m_j^d\cdot\boldsymbol{W}_{\mathrm{proj}}^{n2})\right)}{\sum_{j=1}^{R}\exp\left(\mathrm{ReLU}\left(\beta_d^T(m_i^d\cdot\boldsymbol{W}_{\mathrm{proj}}^{d1}\oplus m_j^d\cdot\boldsymbol{W}_{\mathrm{proj}}^{d2})\right)\right)}\tag{13}$$

$$F_k^{g'}=\frac{1}{D}\sum_{j=1}^{D}\sum_{p\in N_i}\beta_{ij}^{d}\cdot m_p^d\tag{14}$$

其中，$\boldsymbol{W}_{\mathrm{proj}}^{n1}$、$\boldsymbol{W}_{\mathrm{proj}}^{n2}$、$\boldsymbol{W}_{\mathrm{proj}}^{d1}$ 和 $\boldsymbol{W}_{\mathrm{proj}}^{d2}$ 是可学习权重矩阵；ReLU（Rectified Linear Unit）是激活函数；β_{ij}^{d} 表示节点 i 和节点 j 之间的注意力分数；$F_k^{g'}$ 表示 d 层中最终的节点融合特征；D 是注意力头数量；N_i 表示节点 i 的邻居集合．

最后，由于模型存在情感语义偏差，本文期望方面图能够学习更多常见的情感语义偏差，以更全面地感知方面术语，并捕捉其对应的情感倾向．因此，设计了一种新的图关联学习方法，以引入额外的参数进行整体的情感校准．具体而言，在每个节点上添加一组可学习参数 $P=\{p_1,p_2,\cdots,p_{2N}\}$：

$$F^s=\mathrm{FFN}\left(\mathrm{LN}(F_n^{g'}+P)+F_n^{g'}+P\right)\tag{15}$$

其中，FFN（Feed-Forward Network）和 LN（LayerNormalization）被应用于 $F_n^{g'}$ 以获得最终融合特征表示 F^s．由于关联方式有多种，将在实验部分进一步地分析以验证有效性．

3.4 学习目标

为了优化方面术语提取与情感分类任务，本文使用相同的预训练语言模型进行特征编码，并为每个任务添加统一的分类器．具体而言，两个子任务均采用标准的交叉熵损失函数进行训练．如下所示：

$$\mathrm{Loss}_a=-\frac{1}{N}\sum_{i=1}^{N}y_i^a\cdot\lg\widehat{y_i^a}\tag{16}$$

$$\mathrm{Loss}_s=-\frac{1}{N}\sum_{i=1}^{N}y_i^s\cdot\lg\widehat{y_i^s}\tag{17}$$

其中，N 是训练样本的数量，方面术语提取任务的预测结果用 $\widehat{y_i^a}$ 表示，其对应真实标签为 y_i^a；情感分类任务的预测结果用 $\widehat{y_i^s}$ 表示，真实标签为 y_i^s．

最后，为了平衡并同时优化两个独立的预测任务：方面术语提取和情感分类，定义了一个全局损失函数，它是各个任务损失函数的加权总和．定义如下：

$$L=\alpha\cdot\mathrm{Loss}_a+\beta\cdot\mathrm{Loss}_s\tag{18}$$

其中，$\alpha,\beta\in(0,1)$ 是用于平衡不同预测任务之间损失的超参数．在模型训练过程中，使用梯度方向传播来最小化全局损失函数．

4 实验与分析

在本节中，本文使用四个公开数据集针对 MATE、MASC 和 JMASA 任务进行了一系列实验，以评估所提出的 ANAGAL 模型的有效性．实验结果旨在回答以下五个研究问题：

RQ1：ANAGAL 和其他的 MABSA 模型相比，在 MATE、MASC 和 JMASA 任务上的性能如何？

RQ2：主要模块如何影响 ANAGAL 的性能？

RQ3：不同关联学习方式如何影响 ANAGAL 性能？

RQ4：参数配置如何影响 ANAGAL 的性能？

RQ5：ANAGAL 在真实数据集上的表现效果如何？

4.1 实验设置

数据集：在本研究中，利用四个公开的 MABSA 基准数据集，Twitter-2015[22]、Twitter-2017[22]、MASAD（Multimodal Aspect Sentiment Analysis Dataset）[33] 和 EmoMeta[13] 来评估模型性能．这些数据集的相关统计细节汇总在表 1 中，包含多模态数据（图像-文本对），方面术语以及情感标注，展示了多方面和多情感上的不同特征．其中 EmoMeta 为中文数据集，包含 5 000 个隐喻广告文本-图像对的中文字幕，实现了细粒度情感标注．本文将使用这些数据集进行后续实验分析．

表 1 实验数据集的统计数据

参数	Twitter-2015			Twitter-2017			MASAD			EmoMeta		
	训练集	验证集	测试集	训练集	验证集	测试集	训练集	验证集	测试集	训练集	验证集	测试集
积极	928	303	317	1 508	515	493	1 999	712	503	869	309	247
中性	1 883	670	607	1 638	517	573	1 729	420	374	1 749	728	226
消极	368	149	113	416	144	168	1 838	512	464	462	168	242
句子数量	3 502			2 910			1 840			5 000		
单个句子	2 159 (61.65%)			976 (33.54%)			1 453 (64.71%)			4 133 (82.66%)		
多方面	1 343 (38.35%)			1 934 (66.46%)			829 (35.29%)			867 (17.34%)		
多情感	1 257 (35.89%)			1 690 (58.08%)			2 273 (58.89%)			1 376 (27.52%)		

评估指标. 在MATE和JMASA任务上通过微观F1分数(Micro-F1, MF1)、精确率和召回率来评估ANAGAL的性能. 而在MASC任务上,遵循先前的研究[11],使用准确率和F1分数作为评估指标评估模型性能.

基线模型. 为了展示ANAGAL模型的有效性,本文将ANAGAL与以下四种类型的MABSA方法进行比较:

(1)文本ABSA方法. SPAN[34]是一个基于跨度的端到端片段的提取-分类框架,它直接提取从句子中提取多个观点,然后进行分类. RoBARTa[29]使用Transformer编码器进行方面级情感识别. D-GCN(Directional Graph Convolutional Networks)[35]是一种遵循序列标记范式执行任务,并基于BERT的图卷积和序列标签相结合的网络,使用适当的架构对输入之间的依赖关系进行建模. BART(Bidirectional and Auto-Regressive Transformer)[36]则是一个用于七个ABSA任务的预训练语言模型.

(2)JMASA方法. OSCGA-collapse(Object-level Semantic Contextual Graph Attention-collapse)[37]、RpBERT-collapse(Relation propagation Bidirectional Encoder Representations from Transformers-collapse)[38]和UMT-collapse(Unified Multimodal Transformer-collapse)[39]都是使用相同的视觉输入来压缩标记,并利用视觉内容中共享的视觉线索进行方面级情感分析. JML[9]是一种用于同时处理方面术语提取和情感分类的多模态联合方法. VLP-MABSA[5]是一种统一的多模态编码-解码架构,可用于所有预训练任务. CMMT[10]是一种多任务学习框架,用于从图像-文本对中提取方面-情感对进行情感分析. AOM[11]是一种面向方面的网络,用于减轻图像-文本对之间复杂的交互,并减少视觉和文本内容的噪声. Atlantis[40]是一个面向美学特征的多粒度融合网络,用于JMASA任务. DQPSA(Dual Query Prompt Sentiment Analysis)[41]是一种用于多模态方面级情感分析的新型能量模型机制. AETS[12]首次提出方面增强和文本简化来进行多方面多情感下的多模态情感分析.

(3)MATE方法. RAN(Region-aware Alignment Network)[42]首次结合对象和文本特征提取方面术语. UMT(Unified Multimodal Transformer)[43]提供了一个统一架构来解决视觉上下文偏差. OSCGA(Object-level Semantic Contextual Graph Attention)[36]是一个将视觉和文本信息联合到实体中进行预测的神经感知模型.

(4)MASC方法. ESAFN(Entity-Sensitive Attention and Fusion Network)[43]是一种针对MASC任务的注意力融合网络. TomBERT(Target-oriented multimodal Bidirectional Encoder Representations from Transformers)[22]是针对目标的情感分类方法. CapTrBERT(Caption-Transformer Bidirectional Encoder Representations from Transformers)[24]是一种直接将图像输入到空间中进行转换的双流模型.

实现细节. 本研究在Windows系统、Python版本3.10、PyTorch版本1.12.0和NVIDIA RTX 4090 GPU的环境下实现. 在训练过程中,学习率设置为2×10^{-5}, dropout设置为0.1,隐藏层大小设置为768. 对于Twitter-2015和Twitter-2017数据集,采用高斯噪声采样,而对于MASAD和EmoMeta数据集,则采用平均噪声采样.

4.2　主要结果分析(RQ1)

在本节中,充分展示了ANAGAL和当前最先进方法在JMASA、MATE和MASC任务上的对比. 实验结果表明ANAGAL模型的有效性,实现了最优效果.

(1)JMASA上的表现. JMASA分析结果如表2所示. 首先,ANAGAL优于所有纯文本模型和其他的多模态方面级情感分析方法,这主要得益于本文提出的自适应噪声增强模块和方面图关联学习模块. 其次,ANAGAL在Twitter-2015、Twitter-2017、MASAD数据集上的所有指标均达到了最高水平,精确率分别实现了1.46、1.56和2.48个百分点的显著提升. 此外,其他指标(召回率和MF1)也显示出相对于SOTA模型的改进. 这也表明了ANAGAL在增强方面术语感知能力、提高鲁棒性以及准确捕捉方面术语对应情感极性的有效性. 而在EmoMeta数据集上,精确率提升了1.55个百分点,召回率和MF1分别下降了0.53个百分点和1.08个百分点. 这可能是因为EmoMeta数据集上包含的多情感句子较少,方面图关联学习模块难以充分发挥方面情感关联作用,限制了模型性能.

(2)MATE上的表现. 如表3所示,ANAGAL实现了整体的最优性能. 与第二好的AETS模型相比,在大多数指标上均达到了当前最优水平. 例如,在Twitter-2017数据集上,准确率提高了2.82个百分点. 在EmoMeta数据集上,精确率、召回率和MF1分别提高了1.67、0.66和0.17个百分点. 但ANAGAL在Twitter-2015数据集上精确率下降了0.34个百分点,召回率下降了0.66个百分点. 这可能是由于Twitter-2015和数据集中包含多方面术语的句子比较少(只有35%左右),无法充分发挥模型的性能.

(3)MASC上的表现. 表4展示了ANAGAL在MASC任务中的优异表现. 在Twitter-2017、MASAD和EmoMeta数据集上的所有指标均优于第二的AETS模型,实现了整体的最优效果. 例如,准确率分别提升了0.55、0.50和0.41个百分点,F1分数分别提升了0.88、0.32和0.17个百分点. 这些均可以显示出ANAGAL在方面情感分类任务上的优越性.

表2 不同方法在Twitter-2015、Twitter-2017、MASAD和EmoMeta数据集上执行JMASA任务的结果 单位: %

方法	Twitter-2015			Twitter-2017			MASAD			EmoMeta		
	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1
SPAN*	53.70	53.88	53.73	59.64	61.78	60.57	55.42	57.71	56.10	40.71	45.09	47.92
RoBERTa*	61.82	65.30	63.40	65.51	66.83	66.22	58.30	56.29	57.03	50.96	53.74	53.03
D-GCN*	58.27	58.82	59.40	64.18	64.13	64.10	56.10	56.42	55.20	49.02	49.10	51.49
BART*	62.90	65.03	63.86	65.22	65.57	65.40	60.13	59.45	58.72	52.41	54.26	55.72
OSCGA-collapse*	63.11	63.70	63.24	63.52	63.51	63.55	61.75	60.20	61.42	51.28	53.01	55.80
RpBERT-collapse*	49.22	46.86	48.06	57.04	55.48	56.23	52.60	52.33	51.20	41.73	46.79	49.02
UMT-collapse*	61.03	60.42	61.60	60.78	60.03	61.75	58.42	57.30	57.90	52.03	50.06	53.19
JML	65.05	63.20	64.07	66.52	65.48	66.01	60.54	60.70	59.85	55.18	52.94	54.30
VLP-MABSA	65.12	68.30	66.57	66.92	69.18	67.98	61.14	61.20	60.90	57.39	57.13	59.11
CMMT	64.60	68.72	66.50	67.62	69.42	68.51	61.80	62.07	62.53	60.04	56.82	58.05
AOM	67.92	69.30	68.60	68.40	71.02	69.68	63.20	62.97	63.18	58.37	57.49	59.28
Atlantis	65.62	69.20	67.26	68.58	70.32	69.38	60.92	61.37	60.59	56.48	57.15	58.08
DQPSA	71.70	72.04	71.89	71.07	70.18	70.57	66.40	66.87	65.90	61.20	59.84	**62.88**
AETS	69.70	74.68	72.16	72.62	73.68	73.10	68.72	67.90	65.24	61.92	**60.73**	62.06
ANAGAL	**73.16**	**76.10**	**73.80**	**74.18**	**75.84**	**74.15**	**71.20**	**69.42**	**66.18**	**63.47**	60.20	61.79

注:最优结果以粗体显示,次优结果以下划线表示,*表示的结果来自文献[11].

表3 不同方法在Twitter-2015、Twitter-2017和EmoMeta数据集上执行MATE任务的结果 单位: %

方法	Twitter-2015			Twitter-2017			EmoMeta		
	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1
RAN*	80.52	81.53	81.02	90.70	90.68	90.01	64.02	64.90	65.21
UMT*	77.80	81.72	79.62	86.70	86.80	86.64	62.31	62.83	63.10
OSCGA*	81.70	82.11	81.93	90.24	90.70	90.37	64.27	64.01	63.82
JML	83.60	81.21	82.33	92.04	90.68	91.40	67.90	65.08	66.27
VLP-MABSA	83.60	87.91	85.73	90.74	92.60	91.67	67.92	66.17	67.28
CMMT	83.90	88.12	85.92	92.24	93.92	93.11	66.58	67.35	66.73
AOM	84.60	87.92	86.27	91.73	92.82	92.28	68.13	70.20	69.49
DQPSA	88.30	87.14	87.68	92.10	93.52	94.28	72.19	72.74	71.38
AETS	**89.40**	**92.46**	90.88	93.30	97.32	95.31	73.52	75.28	76.05
ANAGAL	89.06	91.80	**91.74**	**96.12**	**97.64**	**95.88**	**75.19**	**75.94**	**76.22**

注:最优结果以粗体显示,次优结果以下划线表示,*表示的结果来自文献[11].

4.3 消融实验分析(RQ2)

在本节中,为了评估ANAGAL模型中每个组件的重要性,主要分析以下四个变体:(1)排除视觉信息(Visual);(2)无自适应噪声采样(Adaptive Noise Sampling,ANS);(3)排除方面图关联学习(Aspect Graph Association Learning,AGAL);(4)无可学习参数(Learnable Parameters,LP). 这些修改旨在评估分析关键组件在Twitter-2015、Twitter-2017、MASAD和EmoMeta数据集上对JMASA任务的性能影响. 这些变体对模型性能的影响都记录在表5中.

(1)w/o 视觉信息: 是一种排除视觉信息,仅依赖文本数据的ANAGAL变体. 与完整方法相比,该变体在各项指标上均呈现下降,凸显了视觉信息的重要性.

(2)w/o 自适应噪声采样: 是一种不包含自适应噪声采样的ANAGAL变体. 在Twitter-2015和Twitter-2017数据集上,精确率分别下降3.86个百分点和5.10个百分点,MF1分别下降2.52个百分点和1.35个百分点,召回率也出现了下降. 表明自适应噪声对增强方面术语感知能力以及提高鲁棒性的重要性.

(3)w/o方面图关联学习: 是一种不使用方面图关联学习的ANAGAL变体. 移除方面图关联学习模块,意味着忽视模型关联的所有情感语义.在MASAD和EmoMeta数据集上,精确率分别下降5.96个百分点和3.71个百分点,召回率分别下降5.22个百分点和2.37个百分点. 这些结果表明了学习情感语义对优化模型的重要性.

(4)w/o 可学习参数: 表示可学习参数可以让模型学习更多情感语义偏差,在Twitter-2015和Twitter-2017

表4　不同方法在Twitter-2017、MASAD和EmoMeta数据集上执行MASC任务的结果　　单位:%

方法	Twitter-2017		MASAD		EmoMeta	
	准确率	F1分数	准确率	F1分数	准确率	F1分数
TomBERT*	67.82	64.20	64.31	63.78	60.42	61.03
ESAFN*	70.50	68.03	66.12	65.71	62.18	61.98
CapTrBERT*	72.30	70.20	67.24	67.10	64.73	64.25
JML	72.70	—	68.40	—	65.04	—
VLP-MABSA	73.82	71.80	68.90	68.23	65.82	66.10
CMMT	73.80	—	69.10	—	66.21	—
ITM	72.62	72.03	68.15	67.85	65.30	65.04
AOM	76.42	75.02	73.20	72.93	66.94	69.73
DQPSA	75.02	—	76.18	—	68.72	—
AETS	76.60	75.20	76.30	75.92	70.05	71.03
ANAGAL	**77.15**	**76.08**	**76.80**	**76.24**	**70.46**	**71.20**

注:最优结果以粗体显示,次优结果以下划线表示,*表示的结果来自文献[11].

表5　ANAGAL关键组件的消融实验分析　　单位:%

方法	Twitter-2015			Twitter-2017			MASAD			EmoMeta		
	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1
w/o Visual	72.80	75.42	71.35	72.32	72.90	73.46	70.53	69.40	66.04	63.04	60.12	61.70
w/o ANS	69.30	74.50	71.28	69.08	70.30	72.80	66.98	63.57	65.30	60.28	58.49	60.24
w/o AGAL	68.56	70.18	69.37	66.25	68.42	69.02	65.24	64.20	64.22	59.76	57.83	59.82
w/o LP	72.52	75.26	73.70	71.90	74.32	74.10	70.62	68.91	66.10	62.93	60.10	61.68
ANAGAL	**73.16**	**76.10**	**73.80**	**74.18**	**75.84**	**74.15**	**71.20**	**69.42**	**66.18**	**63.47**	**60.20**	**61.79**

注:最优结果以粗体显示.

上召回率分别下降了0.84个百分点和1.52个百分点.在MASAD和EmoMeta数据集上,MF1分别下降了0.08个百分点和0.11个百分点.进一步证明了图关联方式对模型识别方面术语的有效性.

值得注意的是,为验证模型性能提升主要取决于两大策略:自适应噪声采样和方面图关联学习.本文采用RoBERTa和ViT进行统一编码处理,并在训练过程中,控制参数设置不变.MATE、MASC和JMASA任务在基准数据集上的消融实验结果表明,当移除自适应噪声采样时,模型的性能显著下降,这些结果验证了本研究中开发的自适应噪声采样增强方面术语感知能力的有效性,同时提升了模型鲁棒性.此外,缺少方面图关联学习也会对模型性能产生较大影响,表明方面图关联学习可以更加准确地捕捉方面术语对应的情感极性.

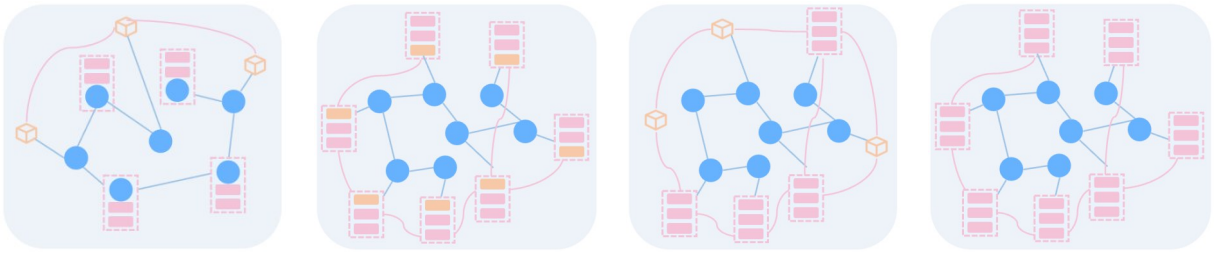
4.4　关联学习方式分析(RQ3)

如3.3节最后所述,在方面图建模过程中,可学习参数的关联方式尚未得到充分利用.因此,本文进行了对比实验以探索不同关联方式在多方面多情感场景下进行JMASA任务的有效性.如图3所示,设计了四种图关联方式.对象关联(Object Association,OA):引入为图节点提供附加参数以帮助在图学习过程中调整对象特征.属性关联(Attribute Association,AA):通过属性参数优化关联的节点.外部关联(External Association,EA):表示在现有两种类型之外重新设计一个新节点,并在模型训练过程中可调试.统一关联(Unified Association,UA):表示为所有节点提供参数,这种方式也被ANAGAL采用.

表6展示了四种图关联方式在Twitter-2015、Twitter-2017、MASAD和EmoMeta数据集上执行JMASA任务的结果.根据实验结果可以得出以下结论:统一关联在Twitter-2015和Twitter-2017数据集上的精确率和召回率达到最佳性能,而在MASAD和EmoMeta数据集上精确率和MF1达到最佳性能.表明可学习参数对模型的重要性.其次是对象关联获得第二好的结果,外部关联是唯一引入新节点的,并且与对象关联相比,同样得到了有利的结果,而属性关联则整体表现不佳.

4.5　参数敏感性分析(RQ4)

本文使用学习率、损失权重α和β、随机因子λ、可学习参数数量等各种超参数,在Twitter-2015数据集上对模型性能进行全面评估.最终的评估结果如图4所示.结果表明,当学习率为2×10^{-5}、损失权重$\alpha=0.5$、$\beta=0.5$、随机因子$\lambda=0.4$、可学习参数数量为4时,模型的性能达到最佳.因此,本文将选择性地确定最优参数配置,以进行MATE、MASC和JMASA实验.



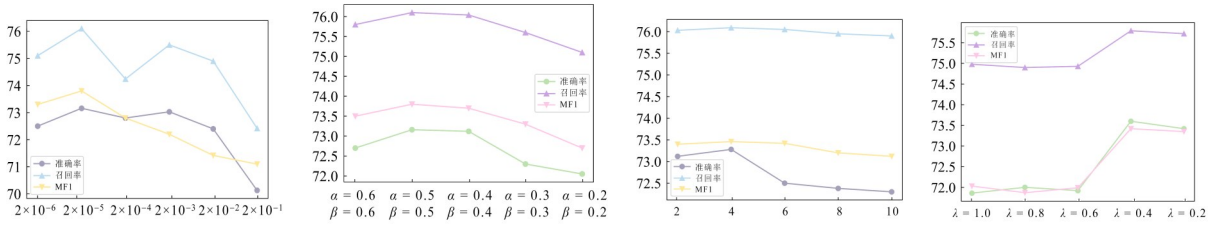
(*a*) 对象关联 (*b*) 属性关联 (*c*) 外部关联 (*d*) 统一关联

图3 不同图关联学习方式的消融实验

表6 不同图关联学习方式在Twitter-2015、Twitter-2017、MASAD和EmoMeta数据集上执行JMASA任务的结果 单位: %

方法	Twitter-2015			Twitter-2017			MASAD			EmoMeta		
	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1	精确率	召回率	MF1
OA	72.80	75.60	**73.92**	74.01	75.68	**74.36**	70.93	69.37	66.10	62.71	60.05	61.64
AA	71.92	74.13	73.02	73.62	75.12	73.40	70.58	69.01	65.28	62.48	59.27	60.78
EA	72.64	75.84	73.46	73.46	75.40	73.85	71.04	**69.57**	65.92	63.29	**60.34**	61.27
UA	**73.16**	**76.10**	73.80	**74.18**	**75.84**	74.15	**71.20**	69.42	**66.18**	**63.47**	60.20	**61.79**

注:最优结果以粗体显示.



(*a*) 学习率 (*b*) 损失权重 (*c*) 可学习参数数量 (*d*) 随机因子

图4 ANAGAL的参数敏感性分析

4.6 案例分析(RQ5)

为了进一步展示ANAGAL的有效性,本文在Twitter-2015数据集上设计了两个测试案例,并展现了不同方法的预测结果. 最终结果如图5所示,在第一个案例中,尽管VLP-MABSA和AETS可以正确地检测到Haley的中性情绪,但却错误地预测Alice的抑郁情绪为中性情绪,而ANAGAL准确地预测到每个方面术语对应的情感倾向. 在第二个案例中,VLP-MABSA错误地预测方面术语Hachiko对应的情感倾向,而AETS和ANAGAL都成功地预测每个方面术语及其对应的情感倾向. 综上所述,本文的ANAGAL模型始终表现出优越的性能,全面地感知所有方面术语,并准确地捕捉对应的情感极性,主要归结于ANAGAL中专门设计的自适应增强模块和方面图关联学习模块.

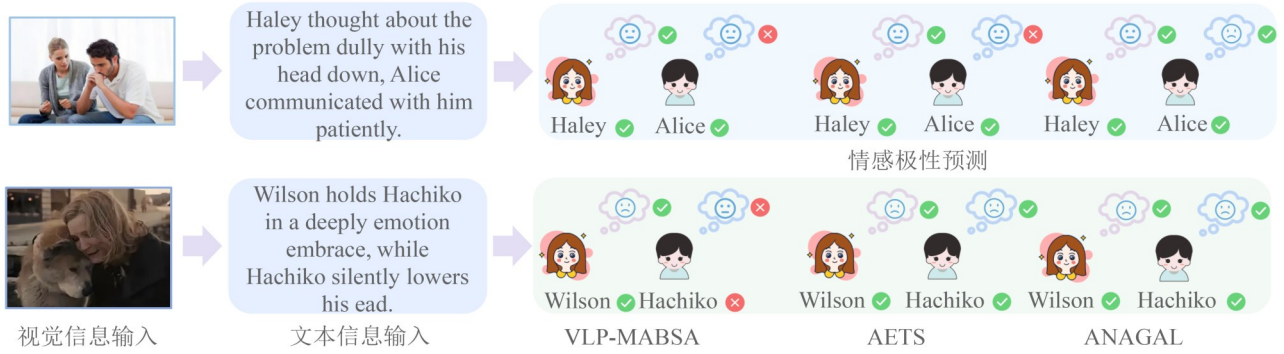


图5 VLP-MABSA、AETS和本文的ANAGAL在两个测试案例上的结果分析

5　结束语

本文介绍了ANAGAL,一种名为自适应噪声增强和方面图关联学习的新型多模态方面级情感分析方法.ANAGAL包括一个自适应噪声增强模块,用于增强模型对方面术语的全面感知能力,并提升模型鲁棒性.此外,还开发了一个方面图关联学习模块,以关联所有方面术语,并学习与之相关的情感语义.同时,引入一组额外的可学习参数进行情感校准,使模型能够学习更多常见的情感语义偏差,从而更准确地捕捉方面术语及其对应的情感极性.该方法用于分析MATE、MASC和JMASA三个子任务,在四个广泛使用的方面级情感分析数据集上的大量实验评估表明,ANAGAL在MABSA任务中优于当前最先进的方法.

尽管本文提出的ANAGAL方法在多个MABSA数据集上表现优越,但仍存在一些局限性与改进空间.首先,ANAGAL主要结合图像和文本模态,尚未覆盖视频、音频等更复杂的多模态组合,模型的模态泛化能力仍有提升空间.其次,方面图的构建依赖于预设规则与模型预测,存在一定的结构偏差风险,未来可探索端到端的方面图结构自适应学习机制.此外,在实际应用中还存在模态缺失、模态冲突等问题,本文未深入涉及此类场景.对于未来的工作,旨在继续深入探索多模态方面级情感分析任务,并在ANAGAL基础上进行.

参考文献

[1] WANG J Y, MOU L T, MA L, et al. AMSA: Adaptive multimodal learning for sentiment analysis[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(3s): 1-21.

[2] 张换香, 彭俊杰. 基于方面级情感分析的深度语义挖掘模型[J]. 电子学报, 2024, 52(7): 2307-2319.
ZHANG H X, PENG J J. A deep semantic mining model based on aspect-level sentiment analysis[J]. Acta Electronica Sinica, 2024, 52(7): 2307-2319. (in Chinese)

[3] YIN S, ZHONG G Q. TextGT: A double-view graph transformer on text for aspect-based sentiment analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 19404-19412.

[4] YANG X C, FENG S, WANG D L, et al. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt[EB/OL]. (2022-05-18) [2025-03-24]. https://arXiv.org/abs/2305.10169.

[5] LING Y, YU J F, XIA R. Vision-language pre-training for multimodal aspect-based sentiment analysis[EB/OL]. (2022-04-21)[2025-03-24]. https://arXiv.org/abs/2204.07955.

[6] HAN Z X, HU M T, BAI Y H, et al. DEQA: Descriptions enhanced question-answering framework for multimodal aspect-based sentiment analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(22): 23987-23995.

[7] WANG D, HE Y, LIANG X, et al. TMFN: A target-oriented multi-grained fusion network for end-to-end aspect-based multimodal sentiment analysis[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Paris: ELRA, 2024: 16187-16197.

[8] XIAO L W, MAO R, ZHAO S, et al. Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis[J]. IEEE Transactions on Affective Computing, 2025. DOI:10.1109/TAFFC.2025.3565506.

[9] JU X C, ZHANG D, XIAO R, et al. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 4395-4405.

[10] YANG L, NA J C, YU J F. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis[J]. Information Processing & Management, 2022, 59(5): 103038.

[11] ZHOU R, GUO W Y, LIU X M, et al. AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis[EB/OL]. (2023-05-31) [2025-03-24]. https://arXiv.org/abs/2306.01004.

[12] ZHU L L, SUN H L, GAO Q S, et al. Aspect enhancement and text simplification in multimodal aspect-based sentiment analysis for multi-aspect and multi-sentiment scenarios[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(2): 1683-1691.

[13] LU X Y, LIU Y X, ZHANG D Y, et al. EmoMeta: A multimodal dataset for fine-grained emotion classification in Chinese metaphors[C]//Companion Proceedings of the ACM on Web Conference 2025. New York: ACM, 2025: 3080-3083.

[14] ZHAO T Y, MENG L G, SONG D W. Multimodal aspect-based sentiment analysis: A survey of tasks, methods, challenges and future directions[J]. Information Fusion, 2024, 112: 102552.

[15] ZHENG C M, FENG J H, CAI Y, et al. Rethinking multimodal entity and relation extraction from a translation point of view[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.

Stroudsburg: ACL, 2023: 6810-6824.

[16] 王友卫, 刘瑞, 凤丽洲. 基于用户性格和语义-结构特征的文本评论情感分类方法[J]. 电子学报, 2024, 52(5): 1657-1669.
WANG Y W, LIU R, FENG L Z. A sentiment classification method for text comments based on user personality and semantic-structural features[J]. Acta Electronica Sinica, 2024, 52(5): 1657-1669. (in Chinese)

[17] YU J F, CHEN K, XIA R. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2023, 14(3): 1966-1978.

[18] WENG Y, CHEN L, WANG S, et al. MIECF: Multi-faceted information extraction and cross-mixture fusion for multimodal aspect-based sentiment analysis[J]. Heliyon, 2024, 10(12): e32967.

[19] GUO A B, ZHAO X, TAN Z, et al. MGICL: Multi-grained interaction contrastive learning for multimodal named entity recognition[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York: ACM, 2023: 639-648.

[20] TRUONG Q T, LAUW H W. VistaNet: Visual aspect attention network for multimodal sentiment analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 305-312.

[21] YE J J, ZHOU J, TIAN J F, et al. Sentiment-aware multimodal pre-training for multimodal sentiment analysis[J]. Knowledge-Based Systems, 2022, 258: 110021.

[22] YU J F, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C]//International Joint Conference on Artificial Intelligence (IJCAI 2020). California: IJCAI, 2020: 5407-5414.

[23] FAN R, HE T T, CHEN M H, et al. Dual causes generation assisted model for multimodal aspect-based sentiment classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(5): 9298-9312.

[24] KHAN Z, FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 3034-3042.

[25] JIA L, MA T H, RONG H, et al. Affective region recognition and fusion network for target-level multimodal sentiment classification[J]. IEEE Transactions on Emerging Topics in Computing, 2024, 12(3): 688-699.

[26] YU J F, WANG J M, XIA R, et al. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. New York: IJCAI Press, 2022: 4482-4488.

[27] YANG J, XIAO Y L, DU X. Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis[J]. Knowledge-Based Systems, 2024, 293: 111724.

[28] WU C, XIONG Q Y, YI H L, et al. Multiple-element joint detection for aspect-based sentiment analysis[J]. Knowledge-Based Systems, 2021, 223: 107073.

[29] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2025-03-24]. https://arXiv.org/abs/1907.11692.

[30] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2025-03-24]. https://arXiv.org/abs/2010.11929.

[31] XUE X J, ZHANG C X, NIU Z D, et al. Multi-level attention map network for multimodal sentiment analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(5): 5105-5118.

[32] WANG S J, CAI G Y, LV G R. Aspect-level multimodal sentiment analysis based on co-attention fusion[J]. International Journal of Data Science and Analytics, 2025, 20(2): 903-916.

[33] ZHOU J, ZHAO J B, HUANG J X, et al. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis[J]. Neurocomputing, 2021, 455: 47-58.

[34] HU M H, PENG Y X, HUANG Z, et al. Open-domain targeted sentiment analysis via span-based extraction and classification[EB/OL]. (2019-06-10) [2025-03-24]. https://doi.org/10.48550/arXiv.1906.03820.

[35] CHEN G M, TIAN Y H, SONG Y. Joint aspect extraction and sentiment analysis with directional graph convolutional networks[C]//Proceedings of the 28th International Conference on Computational Linguistics. New York: IJCAI Press, 2020: 272-279.

[36] YAN H, DAI J Q, JI T, et al. A unified generative framework for aspect-based sentiment analysis[EB/OL]. (2021-06-08)[2025-03-24]. https://arXiv.org/abs/2106.04300.

[37] WU Z W, ZHENG C M, CAI Y, et al. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1038-1046.

[38] SUN L, WANG J Q, ZHANG K, et al. RpBERT: A text-

image relation propagation-based BERT model for multi-modal NER[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13860-13868.

[39] YU J F, JIANG J, YANG L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 3342-3352.

[40] XIAO L W, WU X J, XU J J, et al. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis[J]. Information Fusion, 2024, 106: 102304.

[41] PENG T S, LI Z C, WANG P, et al. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 18869-18878.

[42] WU H Q, CHENG S L, WANG J J, et al. Multimodal aspect extraction with region-aware alignment network[M]// Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2020: 145-156.

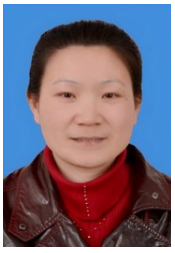
[43] YU J F, JIANG J, XIA R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 429-439.

作者简介

黄　辰　男,1983年生,福建龙岩人. 湖北大学计算机学院教授. 主要研究方向为人工智能、脑机接口.
E-mail: huang@hubu.edu.cn

杨　超　男,1982年生,湖北武汉人. 湖北大学计算机学院教授. 主要研究方向为智能计算、信息安全等.
E-mail: stevenyc@hubu.edu.cn

刘会杰　男,2000年生,湖北黄石人. 湖北大学计算机学院硕士研究生. 主要研究方向为人工智能、脑科学、情感分析.
E-mail: liuhj@stu.hubu.edu.cn



宋建华　女,1973年生,湖北襄阳人. 湖北大学网络空间安全学院教授. 主要研究方向为网络与信息安全.
E-mail: sjhhubu@126.com

张　龑　男,1974年生,湖北宜昌人. 湖北大学计算机学院教授. 主要研究领域为信息安全、大数据分析. 中国电子学会会员编号:E190197582M.
E-mail: zhangyan@hubu.edu.cn