山西大学学报(自然科学版)49(1):15—28,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2025090

基于方面词感知的多模态细粒度情感分析

郝雅卉[1],谢珺[1*],郝戍峰[2],孙颖[1],连浩毅[1],杨文秀[1]
(1.太原理工大学 电子信息工程学院,山西 晋中 030600;
2.太原理工大学 人工智能学院,山西 晋中 030600)

摘 要:针对现有多模态细粒度情感分析模型存在的文本和视觉特征提取不充分、在模态融合过程中忽略方面词引导作用的问题。提出了一种基于方面词感知的多模态细粒度情感分析模型。首先,设计语义对齐模块获取图像中方面词线索,增强方面词感知能力;其次,构建面向方面词的句法依赖图和情感句法图注意力网络,多角度挖掘文本内部与方面词有关的复杂依赖关系;再次,设计多特征编码模块,获取丰富的视觉特征线索;最后,设计对偶交叉注意力机制获取图文模态双向交互信息,基于门控语义图卷积网络和方面掩码机制进一步提高方面词特征质量,引入注意力机制进行方面词感知的模态融合。研究结果表明,在Twitter-2015和Twitter-2017公开数据集上的准确率相比基线模型分别平均提高3.00和2.78个百分点,Macro-*F*1值分别平均提高3.50和2.53个百分点,有效提升了多模态细粒度情感分析性能。

关键词:句法依赖图;多特征编码;情感句法图注意力网络;门控语义图卷积网络
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2026)01-0015-14

Multimodal Fine-grained Sentiment Analysis Based on Aspect Perception

HAO Yahui[1], XIE Jun[1*], HAO Shufeng[2], SUN Ying[1], LIAN Haoyi[1], YANG Wenxiu[1]
(1. *College of Electronic Information Engineering, Taiyuan University of Technology, Jinzhong* 030600*, China;*
2. *College of Artificial Intelligence, Taiyuan University of Technology, Shanxi, Jinzhong* 030600*, China*)

Abstract: The existing multimodal fine-grained sentiment analysis models have problems of insufficient extraction of textual and visual features and neglect of guiding role of aspect during modality fusion. To address these issues, an aspect percept multimodal fine-grained sentiment analysis model was proposed. Firstly, a semantic alignment module was designed to capture aspect clues in images, enhancing aspect awareness. Secondly, an aspect-oriented syntactic dependency graph and an emotion syntactic graph attention network were constructed to explore complex dependency relationships related to aspect in the text from multiple perspectives. Thirdly, a multi-feature encoding module was developed to extract rich visual feature clues. Finally, a dual cross-attention mechanism was introduced to obtain bidirectional interaction information between text and image modalities. The quality of aspect features was further improved using a gated semantic graph convolutional network and an aspect masking mechanism, while an attention mechanism was employed for aspect percept modality fusion. The experimental results demonstrated that, compared to baseline models, the proposed model achieved average accuracy improvements of 3.00 and 2.78 percentage points on the Twitter-2015 and Twitter-2017 public datasets, respectively, along with average Macro-*F*1 score increases of 3.50 and 2.53 percentage points. These

收稿日期:2025-05-14;**修回日期**:2025-09-10
基金项目:虚拟现实技术与系统全国重点实验室(北京航空航天大学)开放课题基金(VRLAB2022C11);山西省重点研发计划项目(202102020101004);山西省回国留学人员科研教研资助项目(2024-61)
作者简介:郝雅卉(1999—),女,山西吕梁人,硕士研究生,研究方向为自然语言处理、情感计算。E-mail:2101675508@qq.com
* **通信作者**:谢珺(XIE Jun),E-mail: xiejun@tyut.edu.cn
引文格式:郝雅卉,谢珺,郝戍峰,等.基于方面词感知的多模态细粒度情感分析[J].山西大学学报(自然科学版),2026,49(1):15-28. DOI:10.13451/j.sxu.ns.2025090.

findings confirmed the model's effectiveness in enhancing the performance of multimodal fine-grained sentiment analysis.
Key words: syntactic dependency graph; multi-feature encoding; emotion syntax graph attention network; gated semantic graph convolutional network

0 引言

社交媒体平台中的海量数据往往包含了用户对被评论对象的观点和态度,情感分析利用所得到的数据信息深度挖掘用户潜在情绪表达,帮助决策者更好地做出适应发展趋势的决定,在舆情检测和商品推荐领域有着广泛应用[1]。实际上,从社交媒体平台获取的数据具有多种模态,同时,用户在同一条评论中可能涉及多个讨论对象。因此获取多模态数据中针对某一给定方面词的情感态度成为了当前的研究重点,即开展多模态细粒度情感分析任务研究。

随着深度学习技术的不断精进,多模态细粒度情感分析任务已取得了不错的成果。Yu等[2]结合长短时记忆网络和残差网络(Residual Network, ResNet)对文本和图像进行表征,同时借助注意力机制捕捉文本和视觉特征中指向方面词的有效信息。Yu等[3]则是通过快速区域卷积网络(Fast Region-based Convolutional Network,Faster R-CNN)获取图像区域特征,并从不同粒度层面对方面词与图像区域进行匹配,减少视觉特征中无关噪声的干扰,提高模型预测能力。此外,为了探索不同模态之间的复杂交互关系和融合方式,Wang等[4]从全局和局部两个角度建模文本特征与视觉特征之间的全局-局部情感语义关联,并采用双线性池化和简单拼接方式获取最后的情感预测特征。尽管现有模型已取得不少成果,但是依然存在以下问题:

1)文本信息提取不足。语言表达往往遵循一定的句法逻辑,词汇本身也有一定的情感色彩,但是现有方法并没有充分利用这些信息增强特征表达能力。

2)视觉信息提取不足。图像具有直接反映强烈情感与观点的特性,能够为模型提供更加直接的情感线索,但是现有模型大多使用编码器获取视觉抽象特征,没有从更多维度丰富视觉特征来源。

3)未能充分利用方面词引导模态融合。现有方法大多利用方面词与单模态交互对齐,忽略了从多模态层面上利用方面词进行引导融合,导致情感预测特征中方面词信息不足。

针对以上问题,本文提出了一种基于方面词感知的多模态细粒度情感分析模型(Multimodal Fine-Grained Sentiment Analysis Based on Aspect Perception,AP-MFG),主要贡献为以下四点:

1)为了提高方面词感知能力,设计语义对齐模块,获取图像中与方面词相关的重要信息,并作为额外线索融入方面词表征中;

2)在文本侧构建面向方面词的句法依赖图,建立以方面词为根节点的图结构,在此基础上设计情感句法图注意力网络,从语义、句法和情感三方面建模方面词与文本上下文之间的复杂依赖关系;

3)在图像侧设计多特征编码模块,获取丰富的图像空间信息和情感信息,为进一步提高视觉特征表达能力,结合匹配损失函数和语义对齐模块结果引导视觉特征向方面词对齐;

4)利用对偶交叉注意力机制双向建模图文特征交互,设计门控语义图卷积网络,将图文上下文信息聚合在方面词节点附近,引入方面掩码机制提取方面词显著特征,增强方面词引导性。

1 相关工作

1.1 多模态特征表示

随着深度学习在自然语言处理各个领域的成熟应用,双向长短时记忆网络(Bi-directional Long Short-Term Memory,Bi-LSTM)模型和双向编码器表征(Bidirectional Encoder Representation from Transformers,BERT)等预训练语言模型在文本序列语义分析领域大放光彩[5]。在此基础之上,一些细粒度文本分析工作将句法知识加入,获取文本内部逻辑。如万宇杰等[6]利

用图卷积神经网络构建邻接矩阵来建模节点之间的依存关系,有效挖掘文本内在联系。Huang等[7]则是利用图注意力网络为每个节点的邻居节点动态分配不同的权重,自适应地聚合邻居节点特征。为进一步提高句法依赖图质量,Wang等[8]重塑和修剪普通依赖树,使模型专注于方面词和潜在观点词之间的连接。谢珺等[9]则是结合情感常识知识对句法依赖图进行增强,同时构建降噪句法图,提高了文本细粒度情感分析模型性能。但上述方法忽略了图像在情感分析任务中的重要作用。

为了在细粒度层面从文本情感分析过渡到多模态情感分析,Khan等[10]利用ResNet和Transformer模型将图像转换成文本序列,将其通过BERT网络层与文本和方面词进行融合。此外,为了过滤图像中与方面词无关的信息,一些图像-方面词匹配方法被提出。如Wan等[11]将图像经过ResNet152网络提取视觉特征表示,并使用交叉注意力机制增强方面词与视觉特征的交互。Zhao等[12]通过定义不同粒度的噪声指标来衡量每个训练样本中图像的噪声程度,并设计不同课程去噪策略减少噪声图像的影响。而Zhao等[13]则是利用SentiBank工具[14]得到图像对应的形容词名词对表示,进而设计知识增强框架提高视觉注意力和模型情感预测能力。上述工作通过对齐方面词与图像特征进一步提高了模型预测性能,但是未能从多方面充分挖掘视觉特征信息以及文本内部与方面词有关的复杂依赖关系。

1.2 多模态特征表示

为了充分交互并有效整合图文模态数据,实现对信息的更全面、准确的理解。Wang等[15]引入两个交互记忆网络监督文本和视觉特征生成,利用BERT网络层捕捉文本和视觉模态之间的连接,最后使用“[CLS]”标记的最终隐藏状态作为联合情感表示。但是上述方法均忽略了图文模态之间的对齐差异,针对该问题,Yu等[16]模型结合方面感知和多模态融合Transformer层分别捕捉图文模态内方面词关键特征和不同模态之间的互补信息,构建辅助损失对齐图文模态,并利用自注意力机制进一步挖掘多模态特征依赖关系。Wang等[17]则是设计递归注意力机制,逐步优化视觉特征表示与方面感知文本特征的对齐效果,同时结合每一步的损失函数监督模型学习,在一定步长后输出最终结果。此外,部分研究人员使用门控机制选择性融合图文特征,如Wang等[18]提出一种自适应跨模态注意力融合架构,将图文特征跨模态交互之后获取增强文本表示,接着计算门向量再次选择性融合视觉区域特征,并通过线性插值创建新样本,增强多模态表示的鲁棒性。Zhang等[19]根据图像与句子的相关性权重动态控制有效图像信息输入,使模型更加关注图文相关特征,之后使用多重注意力机制交互融合图文特征。而Wang等[4]则采用全局-局部协同融合方式挖掘图文特征全局语义关联和局部语义对齐,通过双线性池化和拼接融合方式获取最后的情感预测特征。上述模型虽然取得一定进展,但是忽略了方面词在模态融合过程中的引导作用。导致情感预测特征中方面词信息不足,模态融合不充分。

2 模型概述

AP-MFG模型结构如图1所示,主要包括以下五部分:方面词特征表示、方面感知文本特征表示、方面感知视觉特征表示、多模态融合和情感预测。

在方面词特征表示部分,借助DeepSentiBank获取图像辅助信息,利用语义对齐模块增强方面词感知能力;在方面感知文本特征表示部分构建面向方面词的句法依赖图,在此基础上设计情感句法图注意力网络,建模与方面词有关的文本复杂依赖关系,利用多头交叉注意力机制再次强化方面词与文本特征之间的交互;在方面感知视觉特征表示部分设计多特征编码模块,获取丰富视觉信息,并通过多头交叉注意力机制获取方面感知的视觉特征;在多模态融合部分引入对偶交叉注意力机制获取不同模态之间的双向交互作用,联合门控语义图卷积网络和方面掩码机制获取方面词显著特征,最后通过注意力机制获取方面引导的图文融合特征;在情感预测部分采用Softmax函数激活之前得到的情感特征表示,计算给定方面词的情感标签,同时结合匹配损失、加入正则化的交

叉熵损失函数调整模型参数,优化模型性能。

2.1 方面词特征表示

DeepSentiBank是Chen等[20]提出的一种包含2 048组形容词名词对(Adjectives and Nouns,ANPs)的概念检测器,对于输入的图像数据,DeepSentiBank可以利用深度神经网络获取ANPs与图像匹配程度。考虑到名词可以从另一个角度提供方面词线索,模型通过语义对齐模块学习名词中的方面词相关特征,并作为额外线索融入方面词特征表示中。为了避免关键信息丢失,模型选取前几组匹配度较高的ANPs作为图像辅助信息参与后续计算,具体过程如下。

2.1.1 语义对齐

鉴于BERT预训练模型在文本序列编码方面取得的不错成果,使用BERT获取方面词特征表示H_{a}和第i个名词特征表示H_{N}^{i},并利用三线性相似度算法得出二者语义匹配度,最后基于相似度得分ϕ^{i}对名词进行整合:

$$\phi^{i}=\left[H_{\mathrm{a}}/\!/H_{\mathrm{N}}^{i}/\!/H_{\mathrm{a}}\otimes H_{\mathrm{N}}^{i}\right]W_{\mathrm{a}}^{\mathrm{T}}+b_{\mathrm{a}},\tag{1}$$

$$H_{\mathrm{N}}^{\prime}=\sum\phi^{i}H_{\mathrm{N}}^{i},\tag{2}$$

其中$W_{\mathrm{a}}^{\mathrm{T}}$和$b_{\mathrm{a}}$为权重和偏差,$/\!/$为拼接运算。

2.1.2 方面词编码

将名词表示$H_{\mathrm{N}}^{\prime}$作为方面词额外线索融入方面词特征表示中:

$$H_{\mathrm{A}}=H_{\mathrm{a}}+\eta_{\mathrm{N}}H_{\mathrm{N}}^{\prime},\tag{3}$$

其中η_{N}为可调参数。

2.2 方面感知文本特征表示

为了挖掘文本中与方面词相关的复杂依赖

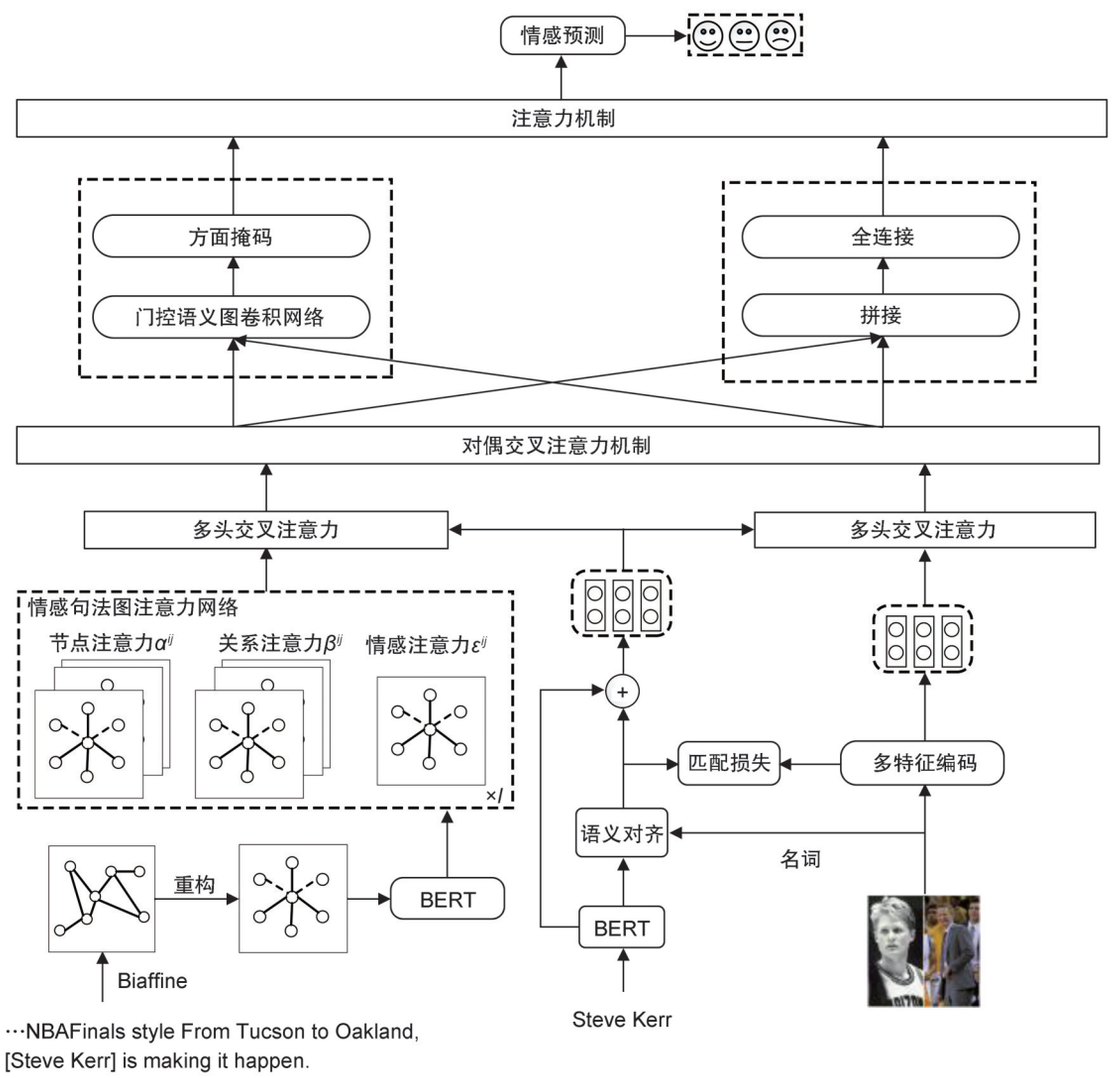

图1 AP-MFG模型结构

Fig. 1 Structure of AP-MFG model

关系,本文构建了面向方面词的句法依赖图,并设计情感句法图注意力网络,聚焦文本内与方面词相关的句法和情感特征信息,最后通过多头交叉注意力机制建模方面词与文本特征的全局交互作用。

2.2.1 构建面向方面词的句法依赖图

Biaffine句法解析器[21]能够通过双仿射变换敏锐感知文本句法结构,拥有较强的关系建模和结构捕捉能力。因此,模型利用Biaffine获取普通句法依赖图,并在此基础上,通过重建根节点以及设置虚拟句法关系,构建方面词与文本中其他单词之间的句法关联。构建规则为:将方面词设置为根节点,如果单词与方面词之间存在直接句法关联,则保留依赖关系,否则,利用节点之间的距离ℓ设置一个ℓ:con虚拟句法关系。具体过程如表1所示。

表1 面向方面词的句法依赖图算法过程

Table 1 Algorithm process of aspect-oriented syntactic dependency graph

算法1 面向方面词的句法依赖图

输入:方面词$A=\{w_1^a,w_2^a,\cdots,w_t^a\}$,文本$S=\{w_1,w_2,\cdots,w_n\}$,句法依赖图$T$,依赖关系$r$

输出:面向方面词的句法依赖图$\hat{T}$

构$\hat{T}$根节点R
for $i=1$ to t do
  for $j=1$ to n do
    if $w_j \xrightarrow{r_{ji}} w_i^a$ then
      $w_j \xrightarrow{r_{ji}} R$
    else if $w_j \xleftarrow{r_{ij}} w_i^a$ then
      $w_j \xleftarrow{r_{ij}} R$
    else
      ℓ=distance(i,j)
      $w_j \xrightarrow{\ell:\mathrm{con}} R$
    end if
  end for
end for
return $\hat{T}$

这里的方面词A为文本S的子序列,t为方面词长度,n为文本长度,r_{ij}是节点i到j的依赖关系。

2.2.2 情感句法图注意力网络

构建好图结构之后,利用BERT对句法依赖图中的节点进行编码,得到特征向量$H=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$,在此基础上使用三种类型的注意力权重捕捉不同依赖关系。

节点注意力权重:基于节点语义特征和可学习权重构建,通过多次计算的方法从不同方面关注节点关联性。

$$\theta_{kl}^{ij}=\mathrm{LeakyReLU}\left(\boldsymbol{s}_{kl}^{\mathrm{T}}\left[\boldsymbol{W}_{kl}^{s}\boldsymbol{h}_i^l /\!/ \boldsymbol{W}_{kl}^{s}\boldsymbol{h}_j^l\right]\right), \tag{4}$$

$$\alpha_l^{ij}=\frac{1}{K}\sum_{k\in K}\frac{\exp\left(\theta_{kl}^{ij}\right)}{\sum_{j\in N_i}\exp\left(\theta_{kl}^{ij}\right)}, \tag{5}$$

其中$\boldsymbol{s}_{kl}$为注意力权重,θ_{kl}^{ij}为节点i和节点j在第l层网络中第k次计算得到的相关权重系数,//是拼接运算,$\boldsymbol{W}_{kl}^{s}$为变换矩阵,α_l^{ij}为第l层的节点注意力权重,N_i为邻居节点集合。

关系注意力权重:考虑到不同句法依赖关系对节点的贡献度不同,为依赖关系r_{ij}设置权重:

$$\gamma_{\mu l}^{ij}=\sigma\left(\mathrm{ReLU}\left(r_{ij}\boldsymbol{W}_{\mu l1}+\boldsymbol{b}_{\mu l1}\right)\boldsymbol{W}_{\mu l2}+\boldsymbol{b}_{\mu l2}\right), \tag{6}$$

$$\beta_l^{ij}=\frac{1}{\Gamma}\sum_{\mu\in\Gamma}\frac{\exp\left(\gamma_{\mu l}^{ij}\right)}{\sum_{j\in N_i}\exp\left(\gamma_{\mu l}^{ij}\right)}, \tag{7}$$

其中$\boldsymbol{W}_{\mu l1}$,$\boldsymbol{W}_{\mu l2}$为权重,$\boldsymbol{b}_{\mu l1}$,$\boldsymbol{b}_{\mu l2}$为偏置,σ是非线性激活函数,β_l^{ij}为第l层的关系注意力权重,Γ是计算关系注意力权重的次数。

情感注意力权重:通过SenticNet7[22]获取每个单词的情感得分,作为额外权重对句法依赖图进行情感增强,计算方法如公式(8)所示:

$$\varepsilon_l^{ij}=\frac{\mathrm{SenticNet7}(w_i)+\mathrm{SenticNet7}(w_j)}{2}, \tag{8}$$

其中ε_l^{ij}为第l层的情感注意力权重。

结合上述三个注意力权重对文本联合编码,获取更新后文本特征$\boldsymbol{H}_{\mathrm{S}}=\{\boldsymbol{h}_1^{\mathrm{S}},\boldsymbol{h}_2^{\mathrm{S}},\cdots,\boldsymbol{h}_n^{\mathrm{S}}\}$,具体过程如公式(9)所示:

$$\boldsymbol{h}_i^{\mathrm{S}(l+1)}=\sigma\left(\sum_{j\in N_i}\left(\alpha_l^{ij}\otimes\beta_l^{ij}\otimes\varepsilon_l^{ij}\right)\boldsymbol{W}_{\mathrm{S}}\boldsymbol{h}_j^l\right), \tag{9}$$

其中$\otimes$为逐元素相乘操作,$\boldsymbol{W}_{\mathrm{S}}$是权重。

2.2.3 文本特征表示

引入多头交叉注意力机制[23]建模方面词与文本之间的交互作用,获取方面感知文本特征,计算方法如公式(10)所示:

$$\boldsymbol{H}_{\mathrm{A\text{-}S}}=\mathrm{MHCA}(\boldsymbol{H}_{\mathrm{A}},\boldsymbol{H}_{\mathrm{S}}), \tag{10}$$

其中MHCA(•)为多头交叉注意力机制运算。

2.3 方面感知文本特征表示

ResNet152网络利用步长为2的7×7卷积对图像进行下采样,从而减少模型计算量,接着在每个阶段使用不同数量的残差块单元堆叠,构建深层卷积神经网络。由于每个阶段的第一个残差块单元会对输入特征进行下采样操作,所以不同阶段的输出特征尺度不同。现有工作大多直接使用ResNet152网络的最后输出作为视觉特征,却忽略了不同尺度特征中包含的图像信息的重要性[24]。此外,由于DeepSentiBank提取到的形容词中包含了图像传达的情感信息,能够为模型提供更直接的情感线索,因此本文基于ResNet152网络和形容词信息设计多特征编码模块,并通过多头交叉注意力机制获取方面感知视觉特征表示。

2.3.1 多特征编码

结构如图2所示,为了增强视觉特征图中的有效信息表达,本文将ResNet152网络每个阶段的结果输入卷积块注意力模块[25],从通道和空间两个维度进行视觉特征增强:首先利用通道注意力机制对图像特征图的通道维度进行加权,突出重要通道的特征;其次,利用空间注意力机制对图像的空间位置进行加权,使模型能够聚焦于图像中的关键区域。

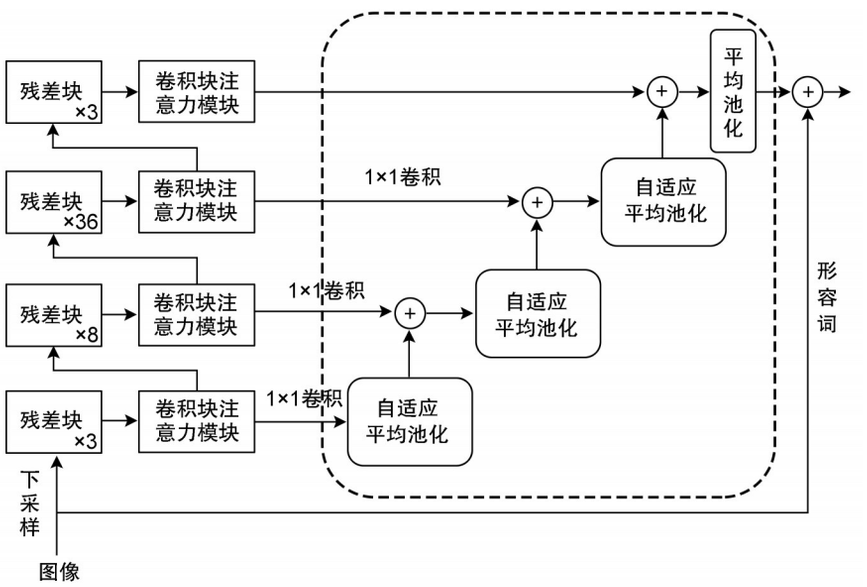

图2 多特征编码模块

Fig. 2 Multi-feature encoding module

以第一阶段提取的尺度特征f_1为例,计算过程如公式(11)至(13)所示:

$$\mathrm{M_c}(\boldsymbol{f}_1)=\sigma\Big(\mathrm{MLP}\big(\mathrm{Avg}(\boldsymbol{f}_1)\big)+\mathrm{MLP}\big(\mathrm{Max}(\boldsymbol{f}_1)\big)\Big), \tag{11}$$

$$\boldsymbol{V}_{\mathrm{c1}}=\mathrm{M_c}(\boldsymbol{f}_1)\otimes\boldsymbol{f}_1, \tag{12}$$

$$\boldsymbol{V}_{\mathrm{s1}}=\sigma\Big(\mathrm{Conv}_{7\times7}\big[\mathrm{Avg}(\boldsymbol{V}_{\mathrm{c1}})/\!/\mathrm{Max}(\boldsymbol{V}_{\mathrm{c1}})\big]\Big)\otimes\boldsymbol{V}_{\mathrm{c1}}, \tag{13}$$

其中Avg(·)为平均池化,Max(·)为最大池化操作,Conv(·)为卷积,σ为非线性激活函数,$\otimes$为逐元素相乘,$\mathrm{M_c}(\cdot)$为通道注意力机制,$\boldsymbol{V}_{\mathrm{c1}}$是通道注意力加权后的结果,$\boldsymbol{V}_{\mathrm{s1}}$为增强后的第一阶段尺度特征。

对每个阶段提取到的尺度特征重复上述步骤,并将增强后的四个多尺度特征通过点卷积调整特征通道数量,得到向量$\boldsymbol{V}_{F1}$,$\boldsymbol{V}_{F2}$,$\boldsymbol{V}_{F3}$,$\boldsymbol{V}_{F4}$,最后以自下而上逐层相加的方式进行融合,并将结果与形容词进行整合,作为多特征编码模块的输出:

$$\boldsymbol{V}_{\mathrm{F}}=\mathrm{Avg}\Big(\mathrm{AD}\big(\mathrm{AD}\big(\mathrm{AD}(\boldsymbol{V}_{F1})+\boldsymbol{V}_{F2}\big)+\boldsymbol{V}_{F3}\big)+\boldsymbol{V}_{F4}\Big), \tag{14}$$

$$\boldsymbol{H}_{\mathrm{V}}=\boldsymbol{V}_{\mathrm{F}}\boldsymbol{W}_{\mathrm{avg}}+\eta_{\mathrm{Adj}}\sum\phi^{i}\boldsymbol{H}^{i}_{\mathrm{Adj}}, \tag{15}$$

其中AD(·)表示自适应平均池化,Avg(·)为平均池化,$\boldsymbol{W}_{\mathrm{avg}}$是变换矩阵,$\boldsymbol{H}^{i}_{\mathrm{Adj}}$是第$i$个形容词经过BERT编码之后的结果,这里的形容词与语义对齐模块的名词一一对应,ϕ^{i}为语义对齐模块得到的相似度得分,η_{Adj}是可调参数。

2.3.2 视觉特征表示

为了获取方面词与视觉特征的全局交互作用,将方面词特征表示H_A和H_V共同输入多头交叉注意力机制中,得到方面感知视觉特征表示:

$$H_{A\text{-}V}=\mathrm{MHCA}(H_A,H_V)。\tag{16}$$

2.4 多模态融合

2.4.1 对偶交叉注意力机制

首先通过自注意力机制学习两个序列内部的语义依赖关系,其次采用两个多头交叉注意力机制获取文本引导的视觉特征和视觉引导的文本特征,使模型能够双向学习图文模态之间的互补信息,如公式(17)至(20)所示:

$$H_{AS}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{A\text{-}S}\boldsymbol{K}_{A\text{-}S}^{\mathrm{T}}}{\sqrt{d_{A\text{-}S}}}\right)V_{A\text{-}S},\tag{17}$$

$$H_{AV}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{A\text{-}V}\boldsymbol{K}_{A\text{-}V}^{\mathrm{T}}}{\sqrt{d_{A\text{-}V}}}\right)V_{A\text{-}V},\tag{18}$$

$$H_{SV}=\mathrm{MHCA}(H_{AS},H_{AV}),\tag{19}$$

$$H_{VS}=\mathrm{MHCA}(H_{AV},H_{AS}),\tag{20}$$

其中$\boldsymbol{Q}_{A\text{-}S}$,$\boldsymbol{K}_{A\text{-}S}$,$V_{A\text{-}S}$是$\boldsymbol{H}_{A\text{-}S}$生成的查询向量、键向量和值向量,$\boldsymbol{Q}_{A\text{-}V}$,$\boldsymbol{K}_{A\text{-}V}$,$V_{A\text{-}V}$是$H_{A\text{-}V}$生成的查询向量、键向量和值向量,$\sqrt{d_{A\text{-}S}}$,$\sqrt{d_{A\text{-}V}}$为缩放因子。

2.4.2 门控语义图卷积网络

由于不同模态所提供的情感信息对于最终的情感预测任务有不一样的贡献程度,因此,采用门控融合方式动态调整不同特征信息的重要性:

$$g=\sigma\left(\boldsymbol{W}_g\left[H_{SV}/\!/H_{VS}\right]+\boldsymbol{b}_g\right),\tag{21}$$

$$H_g=g\otimes H_{SV}+(1-g)\otimes H_{VS},\tag{22}$$

其中$\boldsymbol{W}_g$为可学习权重,$\boldsymbol{b}_g$为偏置,//为拼接操作,σ是激活函数,$\otimes$为逐元素乘法。

为了更好地将图文有用信息传递至方面词节点附近,模型利用自注意力机制自适应生成H_g内部序列的注意力矩阵$\boldsymbol{G}$,并与H_g输入图卷积网络,不断更新节点特征表示:

$$\boldsymbol{G}=\mathrm{Softmax}\left(\frac{H_g\boldsymbol{W}_G^{Q}\times\left(H_g\boldsymbol{W}_G^{K}\right)^{\mathrm{T}}}{\sqrt{d_g}}\right),\tag{23}$$

$$H_G=H_g^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}G\tilde{D}^{-\frac{1}{2}}H_g^{(l)}\boldsymbol{W}\right),\tag{24}$$

其中σ为激活函数,$\boldsymbol{W}_G^{Q}$,$\boldsymbol{W}_G^{K}$,$\boldsymbol{W}$为权重矩阵,$\sqrt{d_g}$为缩放因子,$H_g^{(l)}$为第l层的门控特征。

2.4.3 方面掩码

为了获取融合图文特征信息的方面词显著特征,模型针对门控语义图卷积网络输出中的非方面词向量采用零掩码机制进行遮盖,方面词向量则保持不变,如公式(25)所示:

$$\boldsymbol{R}_G^{\mathrm{mask}}=\{0,\cdots,\boldsymbol{r}_\tau,\cdots,\boldsymbol{r}_{\tau+t-1},\cdots,0\},\tag{25}$$

其中$\boldsymbol{r}_\tau$代表H_G中方面词的第一个单词特征向量表示,t是方面词长度。

2.4.4 模态融合

将对偶交叉注意力网络的输出通过简单拼接操作和全连接层进行融合,得到多模态特征向量$H_F=\{\boldsymbol{h}_1^F,\boldsymbol{h}_2^F,\cdots,\boldsymbol{h}_n^F\}$,引入注意力机制[26]获取方面引导的情感预测特征H_P:

$$H_F=\mathrm{ReLU}\left(\left[H_{SV}/\!/H_{VS}\right]\boldsymbol{W}_F+\boldsymbol{b}_F\right),\tag{26}$$

$$\psi_i=\sum_{m=1}^{n}\left(\boldsymbol{h}_i^F\right)^{\mathrm{T}}\boldsymbol{r}_m=\sum_{m=\tau}^{\tau+t-1}\left(\boldsymbol{h}_i^F\right)^{\mathrm{T}}\boldsymbol{r}_m,\tag{27}$$

$$\varphi_i=\frac{\exp(\psi_i)}{\sum_{m=1}^{n}\exp(\psi_m)},\tag{28}$$

$$H_P=\sum_{m=1}^{n}\varphi_m\boldsymbol{h}_m^F,\tag{29}$$

其中$\boldsymbol{W}_F$和$\boldsymbol{b}_F$是权重和偏置,//为拼接操作。

2.5 情感预测

首先,为了预测方面词情绪类别,将情感预测特征H_P送入Softmax层获取情绪极性:

$$p=\mathrm{Softmax}\left(\boldsymbol{W}_{\mathrm{pre}}H_P+\boldsymbol{b}_{\mathrm{pre}}\right),\tag{30}$$

其中$\boldsymbol{W}_{\mathrm{pre}}$和$\boldsymbol{b}_{\mathrm{pre}}$为可学习的权重矩阵和偏置项。

其次,利用均方误差构建匹配损失函数,通过最小化均方误差来调整参数,使模型学习到更能指向方面词的图像内容:

$$L_m=\frac{1}{M}\sum\left(H_N'-H_V\right)^2,\tag{31}$$

其中M为样本总数。

模型使用加入正则化的交叉熵损失函数计算出预测标签的误差值L_λ,以此缓解模型参数过多带来的过拟合问题。最终的损失函数结合L_m和L_λ,既从图像表征上减小误差,也能从最终的预测得分上优化模型:

$$L_\lambda=-y_i\log p_i+\lambda\left\|\rho\right\|^2,\tag{32}$$

$$L_s=(1-\eta)L_\lambda+\eta L_m,\tag{33}$$

其中η是一个可调整参数,y_i为真实标签,ρ是

可学习参数,λ表示L2正则化项的系数。

3 实验结果和分析

3.1 实验设置

为验证本文模型有效性,在两个公开数据集进行实验对比,分别是来自推特的公开数据集Twitter-2015和Twitter-2017[2]。Twitter-2015包含2014—2015年的推文数据,Twitter-2017包含2016—2017年的推文数据,数据集的详细信息如表2所示。

表2 数据集

Table 2 Information of datasets

	Twitter-2015			Twitter-2017		
	积极	中性	消极	积极	中性	消极
训练集	928	1 883	368	1 508	1 638	416
验证集	303	670	149	515	517	114
测试集	317	607	113	493	573	168

实验采用的评价标准主要包含准确率Accuracy(Acc)与Macro-$F1$值。本文用的GPU设备是NVIDIA GeForce RTX 3090 Ti,通过Pytorch框架实现。优化器使用Nadam,参数η_N在两个数据集上分别设置为0.5和0.2,参数η_{Adj}在两个数据集上分别设置为0.6和0.3,η在两个数据集上分别设置为0.2和0.3,其他参数设置如表3所示。

表3 参数设置

Table 3 Parameter settings

参数	Twitter-2015	Twitter-2017
最大迭代次数	8	9
学习率	2×10^{-5}	2×10^{-5}
Dropout	0.9	0.9
Batch size	16	16
句子最大输入长度	64	64
方面词最大输入长度	32	32
隐藏层维度	768	768

3.2 对比实验

为了验证模型的有效性与泛化性,将模型与其他基线模型在Twitter-2015数据集和Twitter-2017数据集上进行对比,结果如表4所示。

MIMN[27](Multi-interactive Memory Network)模型基于循环门控机制构建多跳网络,加强图文模态内部关联性。

ESAFN[2](Entity-sensitive Attention and Fusion Network)模型将文本分为左、右上下文信息,利用方面词分别交互建模,使用门控机制过滤图像噪声。

SaliencyBERT[17]模型设计了递归注意力机制,反复对齐图文模态信息。

TomBERT[28](Target-oriented Multimodal BERT)模型设计目标注意力机制对齐目标方面和图片,最后通过自注意力机制进行模态融合。

GLFFCA+BERT[4](Global-local Features Fusion with Co-attention + BERT)模型构建基于协同注意力的模态融合网络,从全局和局部的角度探索模态融合方式。

EF-CapTrBERT[10]模型基于Transformer获取图像标题,缩小图文语义差距。

HIMT[16](Hierarchical Interactive Multimodal Transformer)模型构建分层交互模块获取图文交互信息,同时通过辅助重建模块提高模态表征、预测质量。

FITE[29](Face-sensitive Image-to-emotional-text)模型提取视觉内容中的面部表情特征,进一步对齐文本方面词,最后利用门控机制进行融合。

KEF-TomBERT[13](A Knowledge-enhanced Framework-Target-oriented Multimodal BERT)构建知识增强框架获取图像关键信息,提高模型学习视觉特征的能力。

AMIFN[30](Aspect-guided Multi-view Interactions and Fusion Network)模型基于图卷积网络构建句法依赖关系,并设置图像门控机制过滤无关噪声。

从表4中可以看出,AP-MFG模型在四个评价指标上优于其他基线模型,说明了它在多模态细粒度情感分析任务上的优越性。

进一步分析发现,对比最优基线模型AMIFN,AP-MFG在Twitter-2015数据集上的准确率提升了1.22个百分点,这主要是因为模型构建的句法依赖图能够直接建立文本上下文与方面词之间的关联边,使得情感句法图注意力网络从语义、句法和情感三方面捕捉文本上下文与方面词的复杂依赖关系,帮助模型学习到更加丰富、准确的特征。

表4 算法在数据集Twitter-2015和Twitter-2017上两种评价指标的实验结果

Table 4 The experimental results of the algorithms on the datasets Twitter-15 and Twitter-17

%

模型	Twitter-2015		Twitter-2017	
	Acc	Macro-$F1$	Acc	Macro-$F1$
MIMN[27]	71.84	65.69	65.88	62.99
ESAFN[2]	73.38	67.37	67.83	64.22
SaliencyBERT[17]	77.03	72.36	69.69	67.19
TomBERT[28]	77.15	71.75	70.34	68.03
GLFFCA+BERT[4]	77.72	74.21	71.15	69.45
EF-CapTrBERT[10]	78.01	73.25	69.77	68.42
HIMT[16]	78.14	73.68	71.14	69.16
FITE[29]	78.49	73.90	70.90	68.70
KEF-TomBERT[13]	78.68	73.75	72.12	69.96
AMIFN[30]	78.69	75.50	72.29	70.21
AP-MFG	**79.91**	**75.65**	**72.89**	**70.36**

此外,对比次优模型KEF-TomBERT,在Twitter-2015数据集上的准确率提升了1.23个百分点,这是因为模型在视觉特征编码部分通过卷积块注意力模块进行特征增强,同时结合多尺度特征信息和形容词多角度提供视觉线索,此外利用语义对齐模块输出和匹配损失函数从细粒层面引导视觉特征向方面词对齐,减少了图像中与方面词无关信息的干扰。

特别的,对比HIMT模型,在Twitter-2015数据集上的准确率提升了1.77个百分点,这是因为模型在利用对偶交叉注意力机制实现图文信息双向互补的基础上,通过门控语义图卷积网络和方面掩码机制生成了语义丰富的方面词节点信息,并引入注意力机制实现方面词引导的模态融合,使模型能够学习到更加专注于方面词的情感预测特征。

现有ANPs检测器主要分为SentiBank和DeepSentiBank两种,表5展示了使用不同的ANPs检测器对模型性能的影响。可以看出,采用DeepSentiBank获取的形容词名词对能够更加准确地提取到与图像内容相匹配的信息,进而显著提高模型预测准确度。

3.3 消融实验

为了验证AP-MFG模型中各组件的效果,在其他模型框架及参数保持不变的前提下,设计了消融实验,结果如表6所示。

表5 不同形容词名词对检测器对比实验

Table 5 Comparison experiments of different ANPs detectors

%

	Twitter-2015		Twitter-2017	
	Acc	Macro-$F1$	Acc	Macro-$F1$
SentiBank	79.14	75.02	71.19	69.75
DeepSentiBank	**79.91**	**75.65**	**72.89**	**70.36**

表6 消融实验结果

Table 6 Results of ablation experiment

%

模型	Twitter-2015		Twitter-2017	
	Acc	Macro-$F1$	Acc	Macro-$F1$
w/o S	78.56	74.25	72.41	69.07
w/o R	77.16	73.11	69.92	67.81
w/o E	78.64	73.42	70.84	68.98
w/o M	79.78	75.38	72.74	69.85
w/o A	79.15	75.02	72.32	69.72
w/o D	79.74	75.22	72.72	69.62
w/o G	79.72	74.92	72.71	69.55
AP-MFG	**79.91**	**75.65**	**72.89**	**70.36**

移除语义对齐模块(w/o S):直接使用方面词的BERT编码结果和多特征编码结果进行后续实验,从表6可以看出,在Twitter-2015数据集上相比完整模型的准确率和Macro-$F1$值分别下降1.35和1.40个百分点,这是因为模型在语义对齐模块获取的名词特征既能够增强方面词感知能力,也可以有效指导视觉特征与方面词对齐,提升模型预测性能。

移除句法依赖图重建步骤(w/o R):直接把Biaffine得到的句法依赖图输入到情感句法图注意力网络。从表6可以看出,完整模型在Twitter-2015数据集上相比该模型的准确率和Macro-$F1$值分别提升2.75和2.54个百分点。说明本模型针对方面词构建的句法图能够更全面地关注到方面词和句子中其他单词之间的直接关系,过滤掉与方面词无关的特征的干扰。

移除情感句法图注意力网络(w/o E):直接使用图注意力网络建模文本依赖关系,从表6可以看出,在两个数据集上的Macro-$F1$指标相比于完整模型分别下降了2.23和1.38个百分点,说明引入关系和情感注意力权重能从不同角度挖掘方面词与文本上下文之间的内在关联,促进模型更好地学习文本特征。

移除多尺度特征编码模块中的多尺度特征提取步骤(w/o M):直接使用ResNet152网络和形容词对视觉特征进行编码。从表6可以看出,在两个数据集上的评价指标与完整模型相比都有所下降,这是因为从图像中提取的多尺度信息能够兼顾图像的局部细节和全局特征,帮助模型更全面地捕捉图像数据里的方面词线索,提升模型情感预测性能。

移除多尺度特征编码模块中融合形容词的步骤(w/o A):直接将多尺度融合后的平均池化结果作为视觉特征,从表6可以看出,相比于完整模型在两个数据集上的评价指标都有所下降,这是因为形容词能够提供更加强烈的图像情感线索,进而增强模型情感感知能力。

移除对偶交叉注意力机制(w/o D):直接把方面感知的文本和视觉特征输入到门控语义图卷积网络中,从表6可以看出,相比完整模型在两个数据集上的Macro-$F1$指标分别下降了0.72和0.74个百分点,说明利用对偶交叉注意力机制双向建模图文模态交互信息的方法可以在一定程度上挖掘模态间深层交互信息。

移除门控语义图卷积网络(w/o G):将对偶交叉注意力机制输出的图文特征直接拼接后进行预测,从表6可以看出,在两个数据集上的Macro-$F1$指标分别下降了0.73和0.81个百分点,这是因为模型基于门控机制和自注意力机制构建的门控语义图卷积网络可以有效聚合图文模态信息中方面词情感语义信息,使得方面掩码机制获取的方面词特征质量提升,促进图文模态信息更好地向方面词集中。

3.4 参数实验

由上述实验得知,加入ANPs可以增强图像对模型的情感预测效果。为验证ANPs数量对情感预测结果的影响程度,本文将ANPs数量从1设置到9进行实验,结果如图3所示。

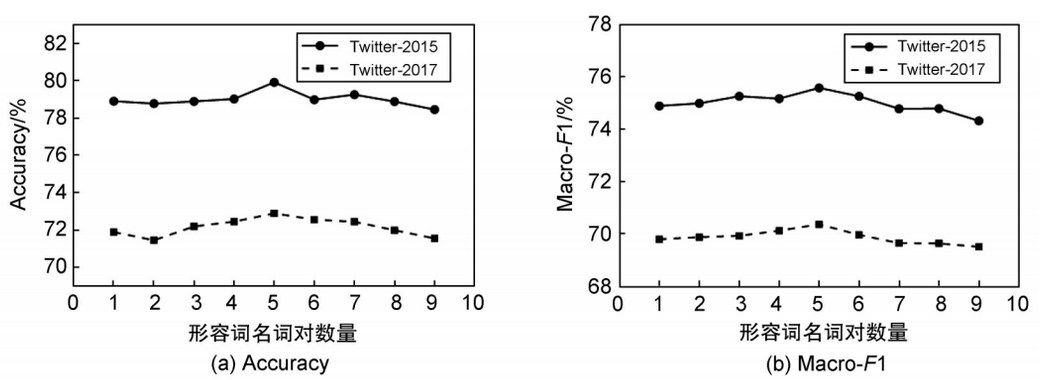

图3 形容词名词对数量对预测性能的影响

Fig. 3 The influence of the quantity of ANPs on prediction performance

两个数据集上,模型取得最好效果时对应的形容词名词对数量均为5,这是因为过多考虑形容词名词对会引入一些与方面词无关的图像噪声,但是较少的形容词名词对无法最大程度上融入图像情感信息并过滤图像噪声。

由以上实验得出,加入情感句法图注意力网络能够提高最终的文本特征表达能力。为了验证情感句法图注意力网络的节点注意力计算次数、关系注意力计算次数和图注意力层数对最终预测性能的影响程度,在两个数据集上将节点注意力计算次数设置为1到8进行测试,将关系注意力计算次数设置为1到9进行测试,将图注意力层数设置为1到5进行测试。

由图4和图5可知,当节点注意力计算次数设置为4、关系注意力计算次数设置为5时,模型在两个数据集上性能最好。这是因为模型每次计算权重时都可以关注到特征在不同方面之间的关系,多次计算注意力权重能够挖掘到更有效的关联信息,提高模型预测性能,但次数过多会带来冗余噪声,影响模型预测结果。

由图6可知,当情感句法图注意力层数为2时,模型在两个数据集上效果最好,层数的过少或者过多,性能反而下降。这是因为当层数过少时,不足以学习关系特征;层数过多可能会出现信息冗余的情况,不仅增加了计算量,还可能导致模型在决策时受到重复信息的干

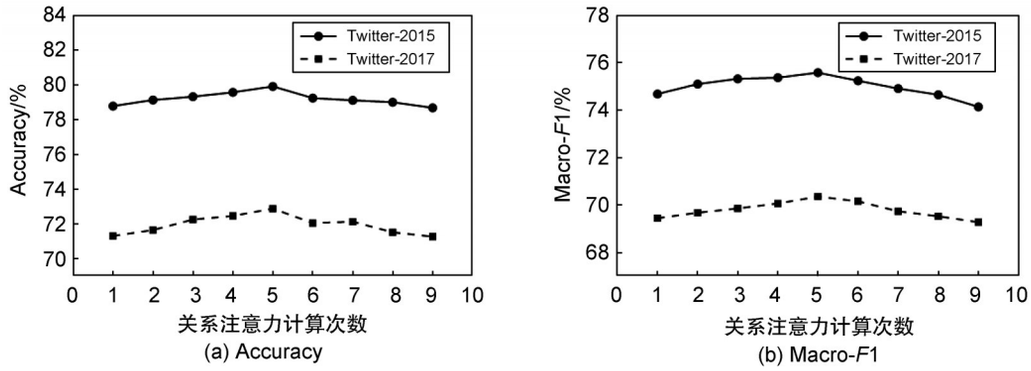

图4 节点注意力计算次数对预测性能的影响

Fig. 4 The influence of node attention calculation times on prediction performance

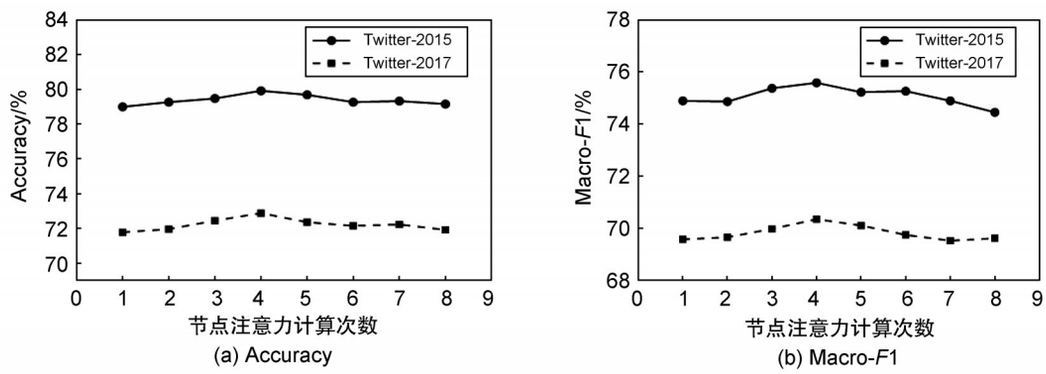

图5 关系注意力计算次数对预测性能的影响

Fig. 5 The influence of relational attention calculation times on prediction performance

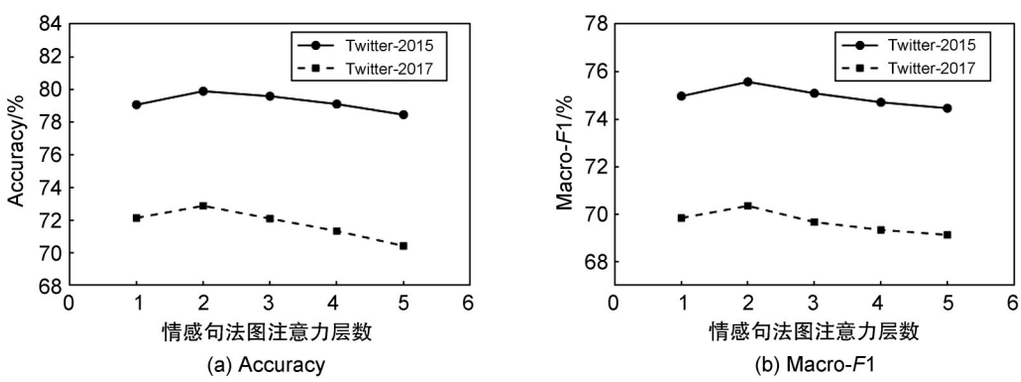

图6 情感句法图注意力层数对预测性能的影响

Fig. 6 The influence of attention levels of emotion syntactic graph on prediction performance

扰,从而影响性能。

消融实验证明了门控语义图卷积网络可以使模型学习到图文模态中方面词的重要信息。为了探索门控语义图卷积网络层数对模型性能的影响,将图卷积网络层数设置为从1到5,并在Twitter-2015和Twitter-2017两个数据集上进行实验,结果如图7所示。

由图7可以看出,当层数设置为1时,模型效果最佳,取值大于1时,模型性能下降,这说明1层图卷积网络足以学习到方面词显著特征,过多的层数会引入冗余噪声,同时随着网络层数的增加,模型的参数数量增多,复杂度也相应提高,这使得模型在训练数据上可能表现得很好,但在测试数据上泛化能力变差,出现过拟合现象,从而导致性能下降。

4 结论

本文提出了一种基于方面词感知的多模态

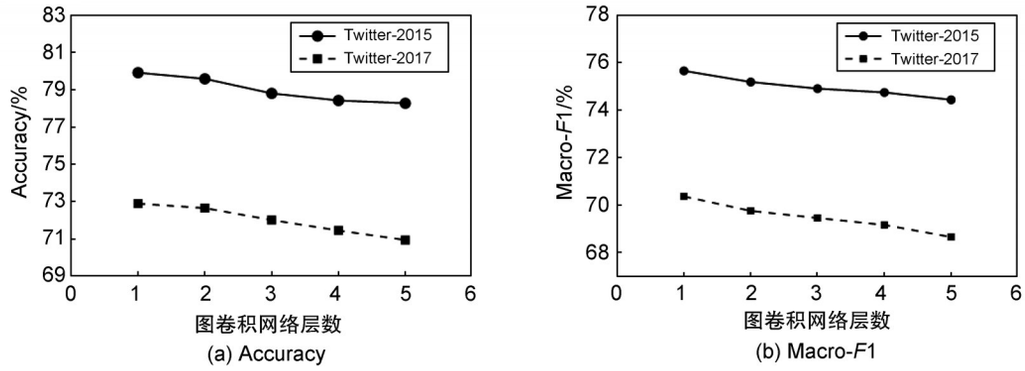

图7 门控语义图卷网络层数对模型性能的影响

Fig. 7 The impact of the numbers of network layers on the performance of the gated semantic graph volume

细粒度情感分析模型AP-MFG,有效解决了现有多模态细粒度情感分析模型存在的文本和视觉特征提取不充分、在模态融合过程中忽略方面词引导作用的问题,具体表现在以下三个方面:通过在文本侧构建面向方面词的句法依赖图和情感句法图注意力网络,充分挖掘文本内部专注于方面词的复杂依赖关系;通过设计多特征编码和语义对齐模块,获取指向方面词的关键视觉特征;通过在模态融合阶段设计对偶交叉注意力机制和门控语义图卷积网络,获取语义丰富的方面词线索,强化方面词引导作用。为验证模型的有效性,我们在Twitter-2015和Twitter-2017这两个公开数据集上进行了实验。结果显示,相较于多个基线模型,该模型在准确率和Macro-*F*1值这两项关键指标上表现更为出色,充分证明了模型的有效性。

当前社交媒体评论的表达方式发生一定变化,为了更好地提高模型泛化能力,下一步将爬取最新的外文评论数据进行标注,并将研究范畴拓展至包含视频模态的数据,在有效提取视频关键信息以及充分融合多模态数据基础上,展开多模态方面级情感分析。

参考文献:

[1] ZHAO H, YANG M Y, BAI X Y, *et al*. A Survey on Multimodal Aspect-based Sentiment Analysis[J]. *IEEE Access*, 2024, **12**: 12039-12052. DOI: 10.1109/ccdc58219.2023.10326793.

[2] YU J F, JIANG J, XIA R. Entity-sensitive Attention and Fusion Network for Entity-level Multimodal Sentiment Classification[J]. *IEEE/ACM Trans Audio Speech Lang Process*, 2019, **28**: 429-439. DOI: 10.1109/TASLP.2019.2957872.

[3] 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. Vienna, Austria: IJCAI, 2022: 4482-4488.. DOI: 10.24963/ijcai.2022/622.

[4] WANG S J, CAI G Y, LV G R. Aspect-level Multimodal Sentiment Analysis Based on Co-attention Fusion[J]. *Int J Data Sci Anal*, 2025, **20**(2): 903-916. DOI: 10.1007/s41060-023-00497-3.

[5] ZHU C, DING Q. Aspect-based Sentiment Analysis *via* Dual Residual Networks with Sentiment Knowledge[J]. *J Supercomput*, 2024, **81**(1): 131. DOI: 10.1007/s11227-024-06546-3.

[6] 万宇杰, 陈羽中. 一种用于方面级情感分析的知识增强双图卷积网络[J]. 小型微型计算机系统, 2024, **45**(1): 37-44. DOI: 10.20009/j.cnki.21-1106/TP.2022-0389.
WAN Y J, CHEN Y Z. Knowledge-enhanced Bi-graph Convolutional Network for Aspect-level Sentiment Analysis[J]. *J Chin Comput Syst*, 2024, **45**(1): 37-44. DOI: 10.20009/j.cnki.21-1106/TP.2022-0389.

[7] HUANG B X, CARLEY K. Syntax-aware Aspect Level Sentiment Classification with Graph Attention Networks [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: ACL, 2019: 5468-5476. DOI: 10.18653/v1/d19-1549.

[8] WANG K, SHEN W Z, YANG Y Y, *et al*. Relational Graph Attention Network for Aspect-based Sentiment Analysis[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 3229-3238. DOI: 10.18653/v1/2020.acl-main.295.

[9] 谢珺, 高婧, 续欣莹, 等. 基于知识增强的双Transformer

网络的方面级情感分析模型[J]. 数据分析与知识发现, 2024, **8**(11): 47−58. DOI: 10.12677/csa.2022.1212291.
XIE J, GAO J, XU X Y, *et al*. Aspect-based Sentiment Analysis Model of Dual-transformer Network Based on Knowledge Enhancement[J]. *Data Anal Knowl Discov*, 2024, **8**(11): 47−58. DOI: 10.12677/csa.2022.1212291.

[10] 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. DOI: 10.1145/3474085.3475692.

[11] WAN Y J, CHEN Y Z, LIN J L, *et al*. A Knowledge-augmented Heterogeneous Graph Convolutional Network for Aspect-level Multimodal Sentiment Analysis [J]. *Comput Speech Lang*, 2024, **85**: 101587. DOI: 10.1016/j.csl.2023.101587.

[12] ZHAO F, LI C H, WU Z, *et al*. M2DF: Multi-grained Multi-curriculum Denoising Framework for Multimodal Aspect-based Sentiment Analysis[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2023: 9057−9070. DOI: 10.18653/v1/2023.emnlp-main.561.

[13] ZHAO F, WU Z, LONG S Y, *et al*. Learning from Adjective-noun Pairs: A Knowledge-enhanced Framework for Target-oriented Multimodal Sentiment Classification[C]//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022: 6784−6794. DOI: 10.18653/v1/2023.findings-emnlp.403.

[14] BORTH D, CHEN T, JI R R, *et al*. SentiBank: Large-scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content[C]//Proceedings of the 21st ACM International Conference on Multimedia. ACM, 2013: 459−460. DOI: 10.1145/2502081.2502268.

[15] WANG Z, LIU Y, YANG J N. BERT-based Multimodal Aspect-level Sentiment Analysis for Social Media[C]// Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition. New York: ACM, 2022: 187−192. DOI: 10.1145/3573942.3573971.

[16] YU J F, CHEN K, XIA R. Hierarchical Interactive Multimodal Transformer for Aspect-based Multimodal Sentiment Analysis[J]. *IEEE Trans Affect Comput*, 2023, **14** (3): 1966−1978. DOI: 10.1109/TAFFC.2022.3171091.

[17] WANG J W, LIU Z, SHENG V, *et al*. SaliencyBERT: Recurrent Attention Network for Target-oriented Multimodal Sentiment Classification[M]//Pattern Recognition and Computer Vision. Cham: Springer International Publishing, 2021: 3−15. DOI: 10.1007/978-3-030-88010-1_1.

[18] WANG Z Y, GUO J J. Self-adaptive Attention Fusion for Multimodal Aspect-based Sentiment Analysis[J]. *Math Biosci Eng*, 2024, **21**(1): 1305−1320. DOI: 10.3934/mbe.2024056.

[19] ZHANG T Z, ZHOU G, LU J C, *et al*. Text-image Semantic Relevance Identification for Aspect-based Multimodal Sentiment Analysis[J]. *PeerJ Comput Sci*, 2024, **10**: e1904. DOI: 10.7717/peerj-cs.1904.

[20] CHEN T, BORTH D, DARRELL T, *et al*. Deepsentibank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks[EB/OL]. (2014−01−01) [2025−12−11]. https://doi.org/10.48550/arXiv.1410.8586.

[21] DOZAT T, MANNING C D. Deep Biaffine Attention for Neural Dependency Parsing[EB/OL]. (2016−01−01) [2025−12−11]. https://doi.org/10.48550/arxiv. 1611.01734.

[22] CAMBRIA E, LIU Q, DECHERCHI S, *et al*. Senticnet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis[C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2022: 3829−3839. DOI: 10.20944/preprints202001.0163.v1.

[23] TSAI Y H, BAI S J, LIANG P P, *et al*. Multimodal Transformer for Unaligned Multimodal Language Sequences[J]. *Proc Conf Assoc Comput Linguist Meet*, 2019, **2019**: 6558−6569. DOI: 10.18653/v1/p19-1656.

[24] 张英俊, 甘望阳, 谢斌红, 等. 融合多尺度特征与注意力的小样本目标检测[J]. 小型微型计算机系统, 2025, **46**(3): 689−696. DOI: 10.20009/j.cnki.21-1106/TP.2023-0509.
ZHANG Y J, GAN W Y, XIE B H, *et al*. Few-shot Object Detection Integrating Multi-scale Feature and Attention[J]. *J Chin Comput Syst*, 2025, **46**(3): 689−696. DOI: 10.20009/j.cnki.21-1106/TP.2023-0509.

[25] WOO S, PARK J, LEE J Y, *et al*. CBAM: Convolutional Block Attention Module[C]//Proceedings of the European Conference on Computer Vision. ECCV, Cham: Springer, 2018: 3−19. DOI: 10.7717/peerjcs.2100/fig-6.

[26] PONTIKI M, GALANIS D, PAVLOPOULOS J, *et al*. SemEval-2014 Task 4: Aspect Based Sentiment Analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Stroudsburg, PA, USA: ACL, 2014: 27−35. DOI: 10.3115/v1/s14-2004.

[27] XU N, MAO W J, CHEN G D. Multi-interactive Memory Network for Aspect Based Multimodal Senti-

ment Analysis[J]. *Proc AAAI Conf Artif Intell*, 2019, **33** (1): 371−378. DOI: 10.1609/aaai.v33i01.3301371.

[28] YU J F, JIANG J. Adapting BERT for Target-oriented Multimodal Sentiment Classification[C]//Proceedings of the Twenty−Eighth International Joint Conference on Artificial Intelligence. Macau, China: Morgan Kaufmann, 2019: 5408−5414. DOI: 10.24963/ijcai.2019/751.

[29] YANG H, ZHAO Y Y, QIN B. Face-sensitive Image-to-emotional-text Cross-modal Translation for Multimodal Aspect-based Sentiment Analysis[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2022: 3324−3335. DOI: 10.18653/v1/2022.emnlp-main.219.

[30] YANG J, XU M Y, XIAO Y L, *et al*. AMIFN: Aspect-guided Multi-view Interactions and Fusion Network for Multimodal Aspect-based Sentiment Analysis[J]. *Neurocomputing*, 2024, **573**: 127222. DOI: 10.1016/j. neucom.2023.127222.