基于目标对齐和语义过滤的多模态情感分析

欧阳梦妮,樊小超,帕力旦·吐尔逊
(新疆师范大学 计算机科学技术学院,新疆 乌鲁木齐 830054)

摘 要:近年来许多研究工作利用注意力机制捕捉意见目标相应的视觉表征进行情感预测,但这种方法在细粒度意见目标对齐方面效果并不理想。为此,提出一种基于目标对齐和语义过滤的多模态情感分析方法。首先,引入目标识别方法 Deepface 获取图像的粗粒度意见目标,并使用映射方法,将粗粒度意见目标映射到细粒度意见目标,实现模态内的目标对齐。其次,利用 Deepface 获取粗粒度意见目标的情绪词并将其和视觉表征融合,使模型更准确地理解和表示意见目标的情感倾向。最后,引入图文匹配模型 CLIP 来评估图像与意见目标之间的语义关联性,从而过滤多余的视觉模态数据噪声。实验表明,提出的意见目标对齐和语义过滤能更好地利用视觉模态信息,提高情感预测的准确性。

关键词:方面级情感分析;目标对齐;语义过滤;噪声;多模态
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2024)10-0171-07
doi:10.20165/j.cnki.ISSN1673-629X.2024.0209

Multimodal Sentiment Analysis Based on Target Alignment and Semantic Filtering

OUYANG Meng-ni, FAN Xiao-chao, Palidan Turson
(School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China)

Abstract: In recent years, many studies have utilized attention mechanisms to capture visual representations corresponding to opinion targets for sentiment prediction, but such methods are not ideal for fine-grained opinion target alignment. To address this, a multimodal sentiment analysis method based on target alignment and semantic filtering is proposed. First, the target recognition method Deepface is introduced to obtain coarse-grained opinion targets from images, and a mapping method is used to map these coarse-grained opinion targets to fine-grained opinion targets, achieving intra-modal target alignment. Second, emotion words associated with coarse-grained opinion targets obtained by Deepface are fused with visual representations, enabling the model to more accurately understand and represent the emotional tendencies of opinion targets. Finally, the text-image matching model CLIP is introduced to evaluate the semantic correlation between images and opinion targets, thereby filtering out redundant visual modal data noise. Experiments demonstrate that the proposed opinion target alignment and semantic filtering can better utilize visual modal information and improve the accuracy of sentiment prediction.

Key words: aspect-based sentiment analysis; target alignment; semantic filtering; noise; multimodal

0 引 言

多模态情感分析是传统的基于文本情感分析的一个扩展领域,除了文本特征,还将考虑语音、视觉等其他模态特征[1]。方面级情感分析旨在识别文本句子中特定意见目标的情感极性。由于不同模态的内容往往密切相关,多模态方面级情感分析利用多种模态信息能够帮助模型更好地分析用户对不同方面的情感[2]。多模态方面级情感分析在各个领域都有着广泛的应用,例如识别客户意见、优化推荐系统[3]等。多模态方面级情感分析不再局限于传统的单一模态任务,而是整合了多种模态特征[4],更贴近人类理解世界的过程,因此吸引了越来越多的关注。

多模态情感分析过程中涉及到文本、音频、视频等多种模态特征,它们存在不同步或不一致的现象,这种现象可能会对模型性能产生负面影响[5]。多模态对齐方法的引入,能够有效地缓解该类问题的出现。

收稿日期:2023-11-15 **修回日期**:2024-03-15
基金项目:新疆维吾尔自治区自然科学基金项目(2022D01A99);国家自然科学基金项目(62066044,62167008);新疆师范大学 2022 年度青年拔尖人才项目(XJNUQB2022-23)
作者简介:欧阳梦妮(1998-),女,硕士研究生,研究方向为自然语言处理;通信作者:帕力旦·吐尔逊(1970-),女,副教授,博士,研究方向为自然语言处理。

Baltrusaitis 等[6]将多模态对齐分为显式对齐和隐式对齐两种方式。显式对齐通过计算人工定义或数据中自动学习的模态特征之间的相似度来对齐不同模态特征。隐式对齐不依赖于数据中的对齐标签,而是在模型训练期间自动学习如何潜在地对齐数据中的不同模态特征。通常情况下,隐式对齐方法利用注意力机制来自动学习多种模态特征之间的对应关系,无需手动设计特征对齐规则。相比之下,隐式对齐方法为多模态情感分析任务提供了一种灵活且通用的模态对齐方式[7],被广泛应用于方面级多模态情感分析领域。

Yu 等人[8]采用注意力机制来实现意见目标和图像之间的对齐,从而得到意见目标敏感的视觉表征。然而,该方法忽略了文本与图像之间的粒度差异。通常情况下,图像中出现的意见目标是粗粒度对象,如图 1 中的“white man”,而句子中对应的意见目标通常是一个细粒度的实体,如“white man”的名字“Koke”。意见目标粒度的不一致可能导致注意力机制无法捕获到对应的视觉表征。从场景图像中提取粗粒度和细粒度的意见目标能够使模型更好地学习意见目标与图像特征之间的对齐关系[9]。然而,不同的意见目标表达相同情感可能采用不同的视觉表征。例如图 1 中的人物均表现出正向的情感,但是笑容的图像表征却并不相同。视觉表征的多样性不可避免地导致了其稀疏性,这使得很难学习视觉表征和情感标签之间的精确映射关系。此外,图文模态之间的语义关系不一致,将会使图片模态的信息成为噪声,从而导致模型整体性能下降。

图 1 目标对齐示例

基于以上问题,该文提出了一种基于目标对齐和语义过滤的多模态情感分析模型(multimodal sentiment analysis based on target alignment and semantic filtering,TASF)。为了更好地对齐意见目标和视觉表征,该文引入了目标识别方法 Deepface,通过 Deepface 获取图像的粗粒度意见目标,并使用映射方法,将粗粒度意见目标映射到细粒度意见目标,实现模态内的目标对齐。此外,为了减少视觉表征多样性对模型性能的影响,TASF 模型利用 Deepface 获取粗粒度意见目标的情绪词并将其和视觉表征相融合,使模型能更准确地理解和表示意见目标的情感倾向。由于文本语义特征在多模态情感分析中占有主导地位[10],为了减少多余视觉特征对模型的性能影响,该文引入了图文匹配模型(Contrastive Language-Image Pre-training,CLIP)来评估图像与意见目标之间的语义关联性,从而过滤多余的视觉模态的数据噪声。主要贡献如下:

●针对意见目标和视觉模态目标粒度不一致问题,该文采用 Deepface 获取图像的粗粒度意见目标,并将其映射到细粒度意见目标,从而实现模态内的目标对齐。此外,使用 Deepface 提取了粗粒度意见目标的情绪词汇,并将其与视觉表示融合,以增强视觉表示中的情感信息,降低了视觉多样性对模型性能的影响。

●针对不同模态间语义关系不一致问题,该文采用 CLIP 来评估图片和意见目标间的关联性,若图片和意见目标关联性较弱,则通过关系权重弱化图片特征,以便减少和任务无关的数据噪声,而更依赖文本语义特征进行多模态情感分析。

●该文提出的基于目标对齐和语义过滤的多模态情感分析模型 TASF 在公开数据集 Twitter-2015 和 Twitter-2017 上进行了实验。实验结果表明,目标对齐、语义过滤以及视觉表征的情感强化能够有效提升多模态方面级情感分析的性能。

1 相关工作

方面级情感分析已成为近年来自然语言处理领域的研究热点之一[11]。多模态对齐是目前多模态情感分析所要面对的主要挑战之一。接下来,该文将从模态显式对齐和隐式对齐两个方面对多模态情感分析的研究现状进行综述。

显式对齐依赖人工定义或从数据中学习的不同模态特征之间的相似度。蔡宇扬等人[12]通过将各个模态映射到公共空间实现模态对齐。Khan 等人[13]使用字幕 Transformer 将图片模态信息转化为文本辅助句子,最后与文本模态融合实现对齐。Hazarika 等人[14]通过分布对齐学习共享子空间中的不同模态表示的共性。

隐式对齐在深度学习模型训练期间对数据进行潜在的对齐。Zhang 等[15]通过两个基于内存网络的模块用于捕获内部模态的特征,实现文本与意见目标的对齐以及图像与意见目标的对齐。Yang 等[16]通过视觉处理器分别提取出图像和文本数据的特征表示,利用多视角注意力机制实现图像与文本对齐。Huddar 等[17]通过双向 LSTM 对文本和图像数据进行上下文编码,并使用注意力机制实现文本和图像对齐。

综上所述,意见目标对齐能够使模型更好地理解和分析不同模态中的情感信息和意见目标之间的关联性,有助于提升模型的整体性能。

2　方　法

该文提出的基于目标对齐和语义过滤的多模态方面级情感分析方法(TASF),如图 2 所示。TASF 包括四个部分:文本语义提取模块、意见目标对齐与情感增强模块、语义过滤与融合模块、情感分析模块。文本语义提取模块采用 BERT 提取意见目标和文本内容的语义特征,并采用了 Transformer 获取二者的语义关联性特征。意见目标对齐与情感增强模块采用 Deepface 提取图片的粗粒度意见目标及其情感信息,通过映射关系实现粗细粒度的意见目标对齐。对齐后的意见目标与视觉表征进行融合,从而捕获视觉模态情感表征。通过映射关系,获取与细粒度意见目标最相关的粗粒度意见目标情感信息,从而强化了视觉表征中的情感信息。语义过滤与融合模块采用 CLIP 检测意见目标与图片模态特征的语义相关性,减少视觉特征的噪声,并与文本语义特征进行融合。最后,通过情感分析模块预测方面的情感倾向性。

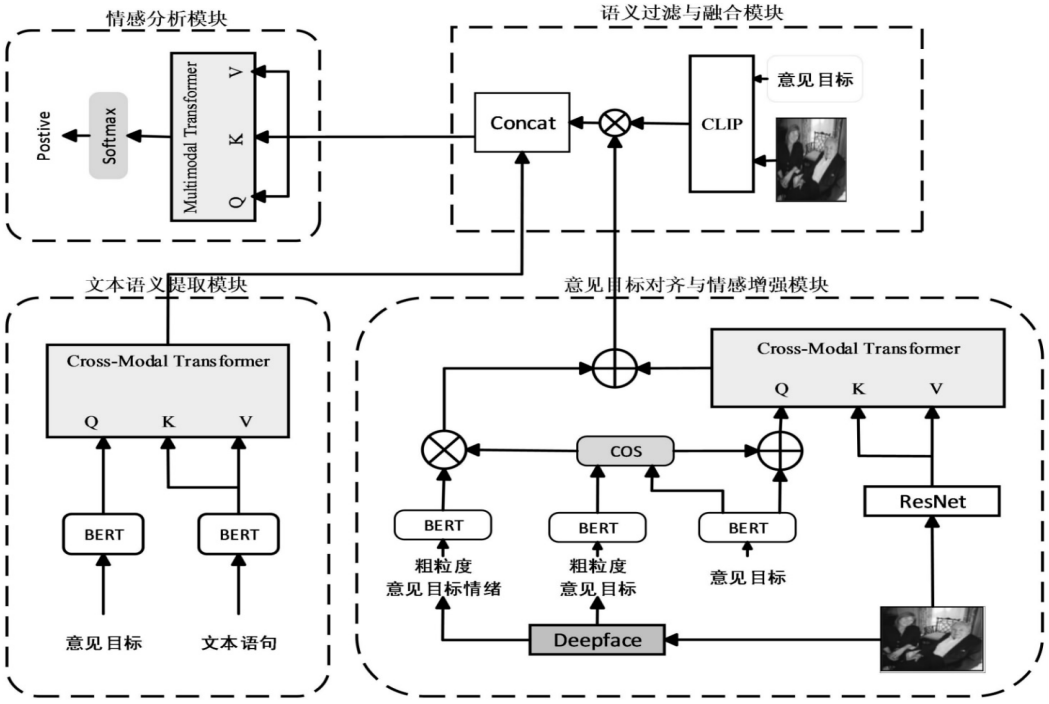

图 2　基于目标对齐和语义过滤的多模态情感分析框架

2.1　文本语义提取模块

该文的目标是对于给定的多模态数据集 D,C 为样本,$C \in D, C=(S,V,T)$,其中 $S=(x_1,x_2,\cdots,x_n)$ 为上下文语句,V 为文本对应的图片,$T=(t_1,t_2,\cdots,t_m)$ 为细粒度意见目标,$T \subseteq S$,训练多模态方面级情感预测模型,从而正确地预测情感值 $\hat{y}$,$\hat{y} \in$(正向,负向,中性)。

由于 BERT[18] 可以从大型语料库中使用预先训练的模型参数生成上下文感知的单词表示,并且具有学习任意两个输入文本之间对齐的能力,因此该文采用 BERT 提取文本语义特征。为了将 BERT 用于方面级文本情感特征提取,将意见目标 T 和上下文文本 S 作为 BERT 模型的输入,得到上下文文本表征 $H_s \in R^{n\times d}$,细粒度意见目标表征 $H_T \in R^{m\times d}$,其中 n 和 m 表示句子长度和意见目标长度,d 为维度。最后,通过跨模态 Transformer 模型[19] 获取 H_T 和 H_S 的深层语义特征 $H_{T\to S}$。

2.2　意见目标对齐与情感增强模块

对于多模态方面级情感分析任务,文本和图像的意见目标通常存在粒度差异。文本中通常给出了细粒度的意见目标,而模型自动获取的图像中的意见目标通常为粗粒度。粗细粒度意见目标对齐是为了确保模型能够正确地将文本中的细粒度目标与图像中的粗粒度目标关联起来,以便更好地捕获图片模态情感信息。此外,粗粒度意见目标及其情感信息具有关联性,因此粗粒度意见目标的情感信息可以辅助细粒度意见目标的情感识别。

鉴于该文专注于研究图片模态中包含人物的多模态方面级情感分析,数据集中的意见目标多为人物,而 Deepface 专注于面部识别和面部特征分析,同时具备评估人物情感状态的能力。因此,该文采用 Deepface 和 BERT 提取粗粒度的意见目标表征 H_{cg} 以及与其对

应的情感表征 H_e，其中 $H_{cg}^i \in R^{l\times d}$，$H_e \in R^{l\times d}$，$l$ 为粗粒度意见目标的个数，d 为维度。

为了计算多个粗粒度的意见目标和细粒度意见目标的关联性，该文采用余弦相似度作为衡量标准，计算了 H_T 和 H_{cg}^i 的相关性并取其最大值 α^m，然后由 α^m 获取与细粒度意见目标最相关的粗粒度意见目标 $\tilde{H}_{cg}$ 及其情感 $\tilde{H}_e$，计算公式如下：

$$\tilde{H}_{cg} = \alpha^m H_{cg}^m \tag{1}$$

$$\tilde{H}_e = \alpha^m H_e^m \tag{2}$$

为了更好地对齐粗细粒度意见目标，该文计算了粗细粒度混合意见目标表征 H_{FT}，计算公式如下：

$$H_{FT} = H_T + \boldsymbol{W}_{cg}\tilde{H}_{cg} \tag{3}$$

其中，$\boldsymbol{W}_{cg}$ 为可训练矩阵，用于调节 $\tilde{H}_{cg}$ 重要性。对于与文本关联的图片 V，采用图像识别模型 ResNet-152[20] 获取最后一个卷积层的输出作为图像表征，计算公式如下：

$$\mathrm{ResNet}(V) = \{r_j \mid r_j \in R_{2048}, j = 1,2,\cdots,49\} \tag{4}$$

计算过程中，模型将原始图像分割成 49 个区域，每个区域用 2 048 维的向量 r_j 表示。接下来，采用线性函数将视觉特征投影到文本特征空间，计算公式如下：

$$H_V = \boldsymbol{W}_V \mathrm{ResNet}(V) \tag{5}$$

其中，$\boldsymbol{W}_V \in R^{d\times 2048}$ 是可训练矩阵。最后，通过跨模态 Transformer 模型获取 H_V 和 H_{FT} 的深层语义表征 $H_{T\to V}$。

直觉上，H_{cg}^m 与细粒度意见目标相似度最高，与 $\tilde{H}_{cg}$ 对应的情感信息 $\tilde{H}_e$ 更可能有助于细粒度意见目标的情感倾向性计算。因此，为了更好地利用粗粒度的意见目标的情感特征，将其与 $H_{T\to V}$ 进行融合，计算公式如下：

$$H_E = H_{T\to V} + \boldsymbol{W}_{eg}\tilde{H}_e \tag{6}$$

其中，$\boldsymbol{W}_{eg}$ 为可训练矩阵，H_E 是意见目标对齐与情感增强后的图文语义表征。

2.3 语义过滤与融合模块

在图文多模态情感分析中，由于文字可以明确表达情感和观点，因此文本通常被认为是主要的情感信息来源。而图像信息通常被视为一种辅助信息，用于从视觉角度补充文本信息，增强对文本情感的理解。然而，由于图文不匹配、融合策略不佳等原因，图片信息可能会引入噪声或不相关的信息，进而影响模型性能。该文采用 CLIP 方法[21] 检测图片与文本之间的相关性。CLIP 使用 Transformer 作为文本编码器，使用 ResNet 和 Vision Transformer(ViT) 作为图像编码器。通过文本编码器和图像编码器分别获取细粒度意见目标和图像的模态表示并计算二者的余弦相似度 R_t。接着，将得到过滤后的图片模态的视觉表示，计算公式如下：

$$H_r = R_t H_E \tag{7}$$

最后，将 $H_{T\to S}$ 和 H_r 拼接在一起，通过 Transformer 模型得到多模态输出表示 H。

2.4 情感分析模块

该文将多模态信息的最终表示 H 作为 softmax 的输入，从而计算多模态方面级情感倾向性，计算公式如下：

$$p(\hat{y} \mid H) = \mathrm{softmax}(\boldsymbol{W}_M^{\mathrm{T}} H) \tag{8}$$

其中，$\boldsymbol{W}_M \in R^{d\times 3}$ 为权重矩阵。

为了衡量模型的预测概率分布 $\hat{y}$ 与真实标签 y 之间的差异，采用了交叉熵损失函数：

$$\mathrm{loss} = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \tag{9}$$

其中，k 为样本数，y 为真实标签，$\hat{y}_i$ 为预测标签。

3 实 验

本节首先介绍数据集和评价指标，其次对比分析了基线模型与文中模型 TASF 的性能差异，然后通过消融实验，验证了模型各部分的有效性，最后通过实例分析进一步解释了模型性能提升的原因。

表 1 数据集详细信息

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Pos	928	303	317	1 508	515	493
Neg	368	149	113	416	144	168
Neu	1 883	670	607	1 638	517	573
Total	3 179	1 122	1 037	3 562	1 176	1 234
AT	1.348	1.336	1.354	1.410	1.439	1.450
Words	9 023	4 238	3 919	6 027	2 922	3 013
AL	16.72	16.74	17.05	16.21	16.37	16.38
Pictures	2 101	727	674	1 745	577	587

3.1 实验数据与评价指标

该文使用由 Yu 等[8] 标注了意见目标情感标签的公共数据集 Twitter-2015 和 Twitter-2017。两个数据集均为图文构成的多模态数据集，每条推文均标注了方面术语以及情感标签，情感极性为正(Pos)、负(Neg)、中(Neu)。数据集被划分为训练集(Train)、验证集(Dev)和测试集(Test)，数据集的详细信息如表 1 所示，其中 AT 为平均方面术语个数，Words 为包含词汇数量，AL 为平均长度。为了和基线模型保持一致，采用了准确率 ACC 和 Macro-F1 值作为评估模型的

评价指标。

3.2　实现细节

所有模型实现采用 Tensorflow 框架。Bert 采用 uncased_L-12_H-768_A-12 模型,学习率为 2e-5,丢失率为 0.1。每个阶段的 batch,size,epoch 分别为 16,100。每一轮结束后,模型在验证集上进行测试。句子输入和意见目标输入的最大长度分别设置为 64 和 32,隐藏维度和注意头数设置为 768 和 12。

3.3　实验结果分析

为了验证提出的基于目标对齐和语义过滤的多模态方面级情感分析 TASF 的性能,对比了以下基线方法:

AE-LSTM[22]:采用基于注意力的长短期记忆对文本的情感进行分类。

BERT[18]:通过微调预训练 BERT 模型对文本进行情感分类

ESAFN[23]:使用注意机制捕获与目标相关的上下文信息来确定上下文表征。通过一个文本融合层来聚合上下文表征以及目标表征。以目标为导向的视觉注意机制,提取与目标密切相关的重要视觉块,通过门控机制消除视觉语境带来的噪声,最后通过多模态表征进行情感分类。

EF-Net[24]:使用多头自注意网络来提取上下文的特征,使用胶囊网络进行视觉模态的特征提取,使用多头注意网络进行多模态特征融合,最后对多模态情感进行分类。

EF - CaTrBERT[13]:通过训练好的字幕 Transformer 将图片映射到文本空间,将目标的标记与字幕 Transformer 预测的图像描述的标记连接,再与文本语句按句子对分类模式结合进行多模态情感分析。

MSFNet[25]:通过交互式注意力机制使方面表示与单模态表示交互,并通过残差连接保留更多方面的表示。提出特征平滑策略和多通道注意力交互网络,以实现不同模态间的深度交互。

AMCGC+BERT[26]:通过同时建模方面指向的模态内上下文语义关联和跨模态的细粒度对齐来提升情感分析性能。

SaliencyBERT[7]:提出了一种多模态 BERT 体系结构捕获模态内和模态间的动态,并逐步优化面向目标的文本特征和视觉特征的对齐,进行情感分类。

TomBERT[8]:通过目标注意机制自动学习意见目标和图像之间的对齐,接着堆叠了一组自注意层,自动捕获它们的模态间相互作用,最后进行多模态情感分析。

主要实验结果如表 2 所示。可以观察到:(1)在基于文本的方法中,BERT 模型的表现优于 AE-LSTM 模型以及其他一些多模态情感分析模型。这一结果证实了 BERT 作为一个有效的预训练模型,能够提供丰富的语义特征,从而提升模型性能;(2)TASF 和 TomBERT 的表现优于大多数多模态模型,主要原因是采用跨模态的多头注意力能够学习更具有鲁棒性的表示;(3)相比于 AMCGC+BERT 模型,文中模型在 Twitter-2015 数据集上展现出更优越的性能。尽管 AMCGC+BERT 模型也致力于细粒度的意见目标对齐,但它未能充分考虑视觉和文本信息之间的相关性。文中模型在 Twitter-2017 数据集上的表现略逊于 AMCGC+BERT,但这一差异主要是由于两个数据集本身的特性所致。相较于 Twitter-2015,Twitter-2017 数据集所使用的图片样本较少,这一事实也反映出文中模型在处理具有更多视觉信息数据集时具有更好的性能表现;(4)TASF * 是五个 TASF 模型的各个评价指标的标准方差,这五个 TASF 模型是在不同随机种子设置下训练,可以看到各个评价指标的标准方差都较低,这表明 TASF 框架训练出的模型性能稳定;(5)TASF 在两个数据集上取得了具有竞争力的结果,文中模型的准确性和 Macro-F1 值在 Twitter-2015 数据集和 Twitter-2017 数据集上高于基线模型,这表明,关注模态内与模态间对齐能使视觉注意精准捕捉到意见目标的视觉表示。

表 2　结果对比

Model	Twitter-2015		Twitter-2017	
	ACC	Macro-F1	ACC	Macro-F1
AE-LSTM	70.30	63.43	61.67	57.97
BERT	74.15	68.86	68.15	65.23
ESAFN	73.38	67.37	67.83	64.22
EF-Net	73.65	67.90	67.77	65.32
MSFNet	74.46	—	69.67	—
EF-CaTrBERT	78.01	73.25	69.77	68.42
AMCGC+BERT	77.73	72.48	**71.15**	**69.69**
SaliencyBERT	77.03	72.36	69.69	67.19
TomBert	77.15	71.15	70.50	68.04
TASF	**78.05**	**73.27**	70.99	69.13
TASF *	0.38	0.67	0.34	0.56

3.4　消融实验

为了验证 TASF 的两个主要组成部分:模态内意见目标对齐与情感增强模块(AE),模态间语义过滤与融合模块(FF)的重要性,进行了一系列消融实验。根据表 3 可以观察到:(1)TASF w/o AE 在两个数据集上的性能都有所下降,验证了利用粗细粒度意见目标融合对提高视觉注意能力的合理性,说明 AE 可以通过粗细粒度意见目标对齐提高情绪预测能力;(2)

TASF w/o FF 在两个数据集上的性能都有所下降,这也证明了 FF 的有效性,过滤模态间的数据噪声能有效提升模型性能;(3)TASF w/o FF 比 TASF w/o AE 性能高,说明 AE 比 FF 更有效。这是具有可解释性的,因为数据中视觉模态的数据噪声相对于整个数据集来说较少,因此对于模型性能的影响相对有限;(4)该文采用 matplotlib 库实现 Twitter-2015 测试集的可视化。

表 3 消融实验

Model	Twitter-2015		Twitter-2017	
	ACC	Macro-F1	ACC	Macro-F1
TASFw/oAE&FF	76.13	70.26	68.11	66.83
TASF w/o AE	77.18	71.24	69.87	68.37
TASF w/o FF	77.71	72.36	70.19	68.74
TASF	**78.05**	**73.27**	**70.99**	**69.13**

如图 3 所示,可以看到 TASF 学习的多模态输出表示明显比 TASF w/o AE 学习的更可分离。这说明 AE 确实可以降低情绪预测的难度,粗细粒度意见目标对齐可以帮助意见目标捕获相应的视觉表示。

(a) TASF w/o AE&FF (b) TASF

图 3 TASF w/o AE&FF 与 TASF 的样本分布情况

3.5 案例分析

为了进一步证明 TASF 的有效性,在表 4 中,通过 TASF、TASF w/o AE&FF、TASF w/o AE 和 TASF w/o FF 对三个例子的预测结果可以明显地观察到:在示例(a)中,TASF w/o AE&FF 和 TASF w/o AE 没能准确预测“Lady Gaga”的情感,这主要是因为这两个方法不能实现图片模态的粒度对齐,捕捉到意见目标的视觉表征,从而获取准确的图片模态情感信息。在示例(b)中,图片模态并不能为情感分析提供有用的情感信息,对 TASF w/o AE&FF 和 TASF w/o FF 方法来说,图片模态信息反而成为噪声,因此错误预测“E3 festiva”的情感。结果表明,文中方法 TASF 表现良好,可以通过控制图像信息的融合,通过意见目标粒度对齐实现视觉表征的精准捕捉,以及通过视觉表征和情感标签之间的精确映射函数获得所有正确的意见目标的情感。

表 4 基于目标对齐和语义过滤的多模态情感分析示例

例子	(a)		(b)	
视觉模态				
文本模态	According to WWD, Mark Ronson is hoping to have Lady Gaga's new album released by year-end.		devolver loses $ 100 k on banned E3 festival	
TASF w/o AE&FF	(Lady Gaga, Pos) (WWD, Neu)	× √	(E3 festiva, Neg)	×
TASF w/o AE	(Lady Gaga, Pos) (WWD, Neu)	× √	(E3 festiva, Neu)	√
TASF w/o FF	(Lady Gaga, Neu) (WWD, Neu)	√ √	(E3 festiva, Neg)	×
TASF	(Lady Gaga, Neu) (WWD, Neu)	√ √	(E3 festiva, Neu)	√

4 结束语

该文提出了基于目标对齐和语义过滤的多模态情感分析(TASF)。在 Deepface 的帮助下,设计了模态内意见目标对齐与情感增强模块,在 CLIP 的帮助下,设计了模态间语义过滤与融合模块,以提高 TASF 任务的视觉注意能力和情绪预测能力。大量的实验结果表明,该模型比其他最先进的方法具有更好的性能。进一步的分析也验证了该模型的优越性。

参考文献:

[1] GANDHI A, ADHVARYU K, PORIA S, et al. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information Fusion, 2023, 91: 424-444.

[2] ZHANG W, 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, 2022, 35: 11019-11038.

[3] 余 鹏,刘星雨,程 颢,等. 在线课程推荐系统综述[J]. 计算机工程与应用,2023,59(22):1-14.

[4] 郭 续,买日旦·吾守尔,古兰拜尔·吐尔洪. 基于多模态融合的情感分析算法研究综述[J]. 计算机工程与应用,

2024,60(2):1-18.
[5] 陈国伟,张鹏洲,王　婷,等. 多模态情感分析综述[J]. 中国传媒大学学报:自然科学版,2022,29(2):70-78.
[6] BALTRUŠAITIS T,AHUJA C,MORENCY L P. Multimodal machine learning:a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(2):423-443.
[7] WANG J,LIU Z,SHENG V,et al. Saliencybert:recurrent attention network for target-oriented multimodal sentiment classification[C]//Pattern recognition and computer vision:4th Chinese conference. Beijing:Springer,2021:3-15.
[8] YU J,JIANG J. Adapting bert for target-oriented multimodal sentiment classification[C]//Proceedings of the twenty-eighth international joint conference on artificial intelligence. Macao:IJCAI,2019:5408-5414.
[9] ZHAO F,LI C,WU Z,et al. Learning from different text-image pairs:a relation-enhanced graph convolutional network for multimodal ner[C]//Proceedings of the 30th ACM international conference on multimedia. Lisboa:ACM,2022:3983-3992.
[10] ZENG J,ZHOU J,LIU T. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities[C]//Proceedings of the 2022 conference on empirical methods in natural language processing. Abu Dhabi:EMNLP,2022:2924-2934.
[11] WANG X,LU A,LIU J,et al. Intelligent interaction model for battleship control based on the fusion of target intention and operator emotion[J]. Computers & Electrical Engineering,2021,92:107196.
[12] 蔡宇扬,蒙祖强. 基于模态信息交互的多模态情感分析[J]. 计算机应用研究,2023,40(9):2603-2608.
[13] KHAN Z,FU Yun. Exploiting BERT for multimodal target sentiment classification through input space translation[C]//Proceedings of the 29th ACM international conference on multimedia. [s. l.]:ACM,2021:3034-3042.
[14] HAZARIKA D,ZIMMERMANN R,PORIA S. Misa:modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proceedings of the 28th ACM international conference on multimedia. Seattle:ACM,2020:1122-1131.
[15] ZHANG Z,WANG Z,LI X,et al. ModalNet:an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network[J]. World Wide Web,2021,24:1957-1974.
[16] YANG X,FENG S,WANG D,et al. Image-text multimodal emotion classification via multi-view attentional network[J]. IEEE Transactions on Multimedia,2020,23:4014-4026.
[17] HUDDAR M G,SANNAKKI S S. RAJPUROHIT V S. Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM[J]. Multimedia Tools and Applications,2021,80:13059-13076.
[18] DEVLIN J,CHANG M W,LEE K,et al. Bert:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of naacL-HLT. Minneapolis:NAACL-HLT,2019:4171-4186.
[19] TSAI Y H H,BAI Shaojie,LIANG P P,et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of annual meeting of the association for computational linguistics. Florence:ACL,2019:6558-6569.
[20] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:770-778.
[21] CHERTI M,BEAUMONT R,WIGHTMAN R,et al. Reproducible scaling laws for contrastive language-image learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver:IEEE,2023:2818-2829.
[22] WANG Y,HUANG M,ZHU X,et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. Austin:EMNLP,2016:606-615.
[23] YU J,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.
[24] GU D,WANG J,CAI S,et al. Targeted aspect-based multimodal sentiment analysis:an attention capsule extraction and multi-head fusion network[J]. IEEE Access,2021,9:157329-157336.
[25] XIANG Yan,CAI Yunjia,GUO Junjun. MSFNet:modality smoothing fusion network for multimodal aspect-based sentiment analysis[J]. Frontiers in Physics,2023,11:1187503.
[26] 王顺杰,蔡国永,吕光瑞,等. 方面级多模态协同注意图卷积情感分析模型[J]. 中国图象图形学报,2023,28(12):3838-3854.