**引用格式**：陈国伟，张鹏洲，王婷，叶前坤.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(02)：70-78.
**文章编号**：1673-4793(2022)02-0070-09

# 多模态情感分析综述

陈国伟，张鹏洲*，王婷*，叶前坤

（中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

**摘要**：情感分析是人工智能领域中重要的研究课题，在人工客服情绪安抚、抑郁症患者状态分类、刑侦辅助心理研究等多个方面都有着广泛使用。结合多个模态进行情感分析可以有效改善单模态情感分析的局限性，本文围绕多模态情感分析介绍了模态数据表示、融合方法、模型使用等方面，梳理概括相关研究内容。最后，总结了多模态情感分析存在的问题，对多模态情感分析的未来进行了展望和探讨。

**关键词**：多模态；情感分析；模态融合
**中图分类号**：TP391　　**文献标识码**：A

# Review on multimodal sentiment recognize

CHEN Guowei, ZHANG Pengzhou*, WANG Ting*, YE Qiankun

*(State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)*

**Abstract:** Sentiment analysis is one of the major research topics in the field of artificial intelligence, and it is widely used in emotional comfort of artificial customer service, classification of depression patients, and research of criminal investigation. Combining multiple modalities in sentiment analysis can effectively eliminate the limitations of single-modal. This paper introduces data representation, fusion methods, and model improvement around multi-modal sentiment analysis, and summarizes related research content. Finally, it summarizes the existing problems of multi-modal sentiment analysis, and discusses the future prospects of multi-modal sentiment analysis.

**Keywords**: multi-modal; sentiment recognize; modal fusion

## 1　引言

情感是一种心理状态，通常会导致人们的行为方式和计算理性相冲突，是人类等高等生物区分于计算机的显著属性[1]，人类的大脑具有意识的维度，著名的情感学家Scherer[2]将情感定义为组成过程，Scherer指出，情绪在适应生物体生命中频繁发生和典型模式的重大事件方面发挥着重要作用，情绪范围很难界定，而愤怒、喜悦、恐惧、悲伤等功利主义情绪相对频繁出现。情感在人工智能领域有着重要的研究价值。情感分析，又称情感计算、意见挖掘，最早起源于Picard提出的“情感计算”概念[3]，Picard指出，情绪在人类的思维、推断和决策中发挥着重要作用，情感计算机可以通过识别人类情感来提高决策能力。目前对基本情感暂未有严格的定义，在

**基金项目**：国家重点研发计划（2020AAA0108700）

**作者简介（*为通讯作者）**：陈国伟（1984-），男，助理研究员，主要从事涉华舆情分析研究。Email：cuc_chenguowei@cuc.edu.cn；张鹏洲（1968-），男，教授，主要从事多模态情感分析研究，Email：zhangpengzhou@cuc.edu.cn；王婷（1999-），女，硕士研究生，主要从事语音情感识别研究。Email：wangting226@163.com

情感分析的研究中，大部分学者使用六种基本情感：悲伤、高兴、恐惧、厌恶、惊讶、愤怒。

情感分析的研究对象主要是情感交互时发生变化的外部行为，如文本、语音参数、人脸表情、肢体动作、生理参数。模态是某件事发生或经历的方式，许多人将模态和感官模态相联系，感官模态代表我们主要的交流和感觉渠道，如视觉或触觉。[4]传统的情感分析主要根据文本、语音、面部表情等模态中的一种模态进行分析。在情感分析的发展过程中，早期会对文本进行基于情感词典的情感分析，根据情感词的强度对文本进行分数统计，国内外都有着丰富的情感词典，如SetiWordNet[5]、知网Hownet情感词典、台湾大学NTUSD情感词典，早期对语音进行情感分析最常用的方法是HMM模型，如Lin等人在2005年结合HMM和SVM模型对语音情感进行分类[6]。早期的表情识别通常是基于几何特征和纹理特征的，如李悦等人采用多种表情特征分类进行人脸表情识别[7]。在机器学习的发展下，一些常见的机器学习算法如SVM、贝叶斯分类器、KNN等都被广泛运用于各个模态的情感分析中，如Bhakre等人使用朴素贝叶斯实现音频的四种情绪状态分类[8]，Mehmood等人使用SVM和KNN对脑电信号进行情感分类[9]。在神经网络开始兴起后，卷积神经网络、长短期记忆网络等神经网络在各个模态的情感分析中也得到良好运用，如Wang等人使用CNN-LSTM对文本进行维度情感分析[10]。近年兴起的Attention机制、迁移学习、预训练模型也被广泛运用于单模态情感分析中，如Munikar等人使用预训练的Bert模型对其进行微调，实现细粒度的情感分类[11]。利用单模态进行情感分析具有一定的局限性，人类表达情感时会通过声音、内容、表情、肢体语言等多种方式联合表达情感。多模态来自多个异构源，如图1所示，相对于单模态情感分析，利用多种模态可以更加准确的捕捉情感信息，部分研究者致力于结合多个模态进行情感分析，利用表征学习、模态对齐、模态融合等方法，有效改善了利用单模态进行情感分析的局限性。

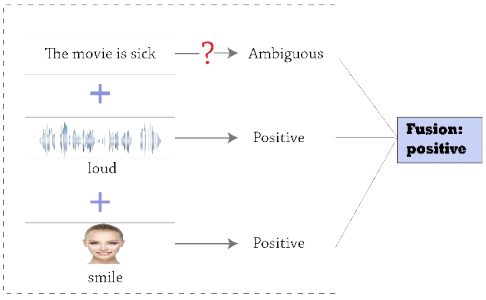


**图1 不同模态的互补性**

在多模态情感分析的研究历程中，许多会议和竞赛推动着多模态情感分析的进步，如2011年起在德国慕尼黑工业大学开始举办的多模态情感识别的竞赛（AVEC）；2016和2017年的多模态情感识别挑战（MEC2016、MEC2017），使用中文自然音频-视觉数据库（CHEAVD）作为挑战数据集，促进了汉语多模态情感分析的研究；科大讯飞在2020年的IFLYTEK A.I.开发者大赛中的多模态情感分析赛道。在情感分析的研究历程中，多模态情感分析起步较晚，但仍有许多研究学者和机构致力于多模态情感分析的研究，国内外许多顶级会议如NLPCC(CCF International Conference on Natural Language Processing and Chinese Computing)、ACL(The Association for Computational Linguistics)、INTERSPPECH(Conference of the International Speech Communication Association)、ICASSP(IEEE International Conference on Acoustics, Speech and SP)、EMNLP(Conference on Empirical Methods in Natural Language Processing)、NAACL(The North American chapter of the association for computational linguistics)、AAAI(Association for the Advance of Artificial Intelligence)上收录了许多围绕多模态对齐、融合、情感识别模型展开研究的论文。

## 2　情感表达模型描述

人类的情感是复杂繁琐的认知过程，很难对人类情感进行简单的概括，现阶段的情感模型大多分为两种，分别是离散情感模型和维度情感模型。

### 2.1　离散情感模型

离散情感模型将情感分为独立的类别，著名的心理学家Ekman[12]等人总结了六种基本情绪：快乐、悲伤、愤怒、恐惧、惊讶、厌恶。且这六种基本情感可以组合派生出其他复合情绪。Roseman[13]等人通过评价因素对情感进行评估，给出17种基本情绪。由于情绪的复杂性，很难精确的对其进行模拟，在实际使用场景中，针对离散情感模型的分类模型比较常见。

### 2.2　维度情感模型

相较于离散情感模型，维度情感模型更加具有普适性，可以有效的对情绪强度进行描述。Russell[14]等人提出了基于愉悦度和激励度两个维度进行情感模拟的二维情感模型，采用环状结构对情感进行描述（见图2）。在维度情感模型中，认同度最高的PAD模型是基于愉悦度（Pleasure）、唤醒度（Arousal）、支配度（Dominance）三

个维度的模型[15],PAD三维情感模型可以有效解释人类的情感,模拟情绪的相似和对立性(见图3)。

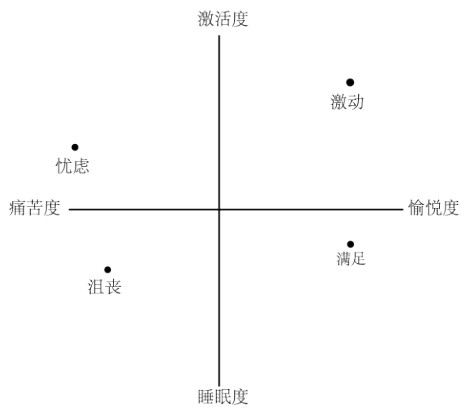


图2 二维情感模型

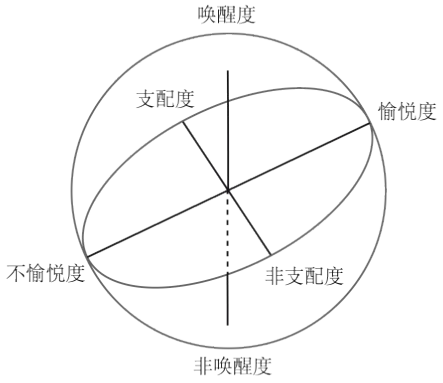


图3 PAD情感模型

## 3 基于多模态的情感分析方法论

如图4所示,多模态情感识别对所采用的数据集中的不同模态进行预处理,提取特征后采用不同的模态融合方式,输入模型对融合的模态信息进行情感识别,分为对句子级别和对话级别进行多模态情感分析,训练模型后为测试样本匹配情感标签,从而评估预测结果。

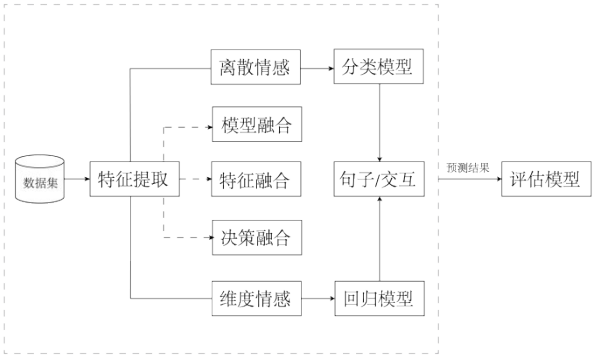


图4 多模态情感分析方法论

### 3.1 单模态数据表示

(1)文本特征表示

对于文本的特征表示,主要是将文本转化为可供机器识别的语言,通常有两种表示方法,分别是One-hot Representation(离散表示)和Distribution Representation(分布式表示)。常用的文本特征工具是词向量模型,词向量将文本进行向量表示,维度是自己事先定义的,相似的词会有相似的向量表示。常用的词向量模型如Google发布的Word2vec模型[16],主要依赖Skip-grams或CBOW两个模型,分别通过中心词预测附近词,以及通过附近词预测中心词,如文献[17]中使用Word2vec模型进行文本模态的特征提取。GloVe[18]词向量使用共现矩阵考虑了全局信息,ELMo词向量[19]能够随着语言环境的变化捕捉词语中和语境相关的含义。2018年Google提出Bert预训练模型[20],许多学者使用大规模语料进行预训练,学习语义关系后进行下游任务词向量的输入,如文献[21]分别使用GloVe词向量和Bert模型来进行文本特征的表示并比较其性能。

(2)语音特征表示

声学特征涵盖丰富的信息,通过对声学特征进行分析可以获取其传递的情感信息,对分类器进行情感识别有着显著影响。最常使用的声学特征有梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient,简称MFCC)、能量/幅度特征、线性预测倒谱系数(Linear Predictor Cepstral Coeffcients,简称LPCC)等。基于Python的Librosa工具可以对语音进行时频处理、提取多种语音特征、绘制声音各类相关图像(如频谱图),Schuller团队在2015年开发了OpenSmile工具[22],可以对语音进行预处理和特征提取,对帧能量、帧强度、自相关函数、幅度谱加权等多种特征进行提取。如文献[23]中使用OpenSmile提取MFCC和LPCC等特征作为语言的情感识别特征。

(3)图像/视频特征表示

人脸信息是根据五官等不断变化的,含有丰富的情感信息。人脸图像/视频特征提取主要基于几何特征和纹理特征。几何特征根据五官的位置、大小、比例关系等使用一组矢量对人脸进行表示。纹理特征主要有SIFT、局部二值模型(Local Binary Patterns,简称LBP)、Gabor小波系数、HOG等,如文献[24]中提取LBP和Gabor特征作为人脸信息特征;文献[25]使用SIFT作为表情识别的特征。对于动态图像序列,

光流法反映了动态帧中灰度的变化，可以反映基于人脸肌肉的运动。Python 的 OpenCV 和 Dlib 库常用于人脸特征关键点识别。如文献[26]中使用 OpenCV 来检测人脸。Brandon 等人在 2016年提出的 OpenFace[27]工具也可以提取面部特征，获取低维表示，用于表情分析。如文献[28]中使用 OpenFace 工具来进行视频模态的处理，用于过滤无关信息，提取面部特征。

除此之外，神经网络也被广泛运用于特征提取中，如 Cambria 等人使用深度卷积神经网络提取文本和视觉特征[29]，Wang 等人使用神经网络进行人脸面部特征提取[30]。

### 3.2　多模态融合方法

多模态数据从不同角度（文本、语音、视频等）对对象进行描述，涵盖比单模态信息更加丰富的信息量，不同的模态信息在内容上可以互补。在进行多模态情感分析任务时，要明确如何融合不同模态的特征信息，保证模态的语义完整性，实现不同模态之间的良好融合，不同的融合方式也会影响任务结果。根据模态融合的方式的不同，可以分为早期基于特征的融合，中期基于模型的融合，晚期基于决策的融合。

（1）早期特征融合

如图 5 所示，特征级融合是早期在特征提取后的浅层融合，将多个模态进行特征的直接连接，即浅层的拼接、相加、加权求和。在进行深度学习之前，往往会使用特征工程来提取模态特征。特征融合要将不同模态的多种特征整合到一个公共空间，由于各个模态的差异性，往往涵盖大量的冗余信息，会采取降维方法来消除冗余信息，通常采用主成分分析（Principal Component Analysis，简称 PCA）等方式。

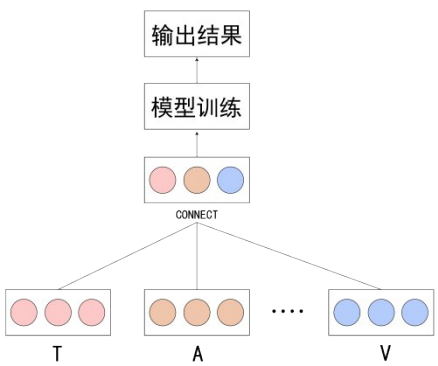


图 5　早期特征融合

文献[31]认为基于特征的模型学习低等级语音信号情绪特征的能力有限，提出了一种多模态双重递归编码模型，对文本和音频序列的双模态信息进行编码，将文本和音频进行特征融合后分类，将准确率提高到了 71.8%，有效解决了错误预测为中立的情况。文献[32]采用卷积神经网络提取文本的浅层特征，最后使用包含 100个神经元的全连接层连接文本表征，采用 OpenSMILE 工具中的 is13compare1config 文件，提取共计 6373个语音特征，进行标准化后采用全连接层降维至 100 层，采取 3D-CNN 提取面部表情和视觉特征，利用含有 100个神经元的全连接层提取视频特征。在经过上述处理后，各个模态具有相同的维度，文献中将单个模态通过简单的线性连接形成维度为 300 的多模态映射。文献[33]提出一种分层的多模态情感分析层次融合网络（Hierarchical Feature Fusion Network，简称 HFFN），涵盖了局部融合模块和全局融合模块，通过滑动窗口探索局部的跨模态融合，有效降低了计算复杂度，通过 ABS-LSTM 网络探索全局多模态向量，引入记忆细胞的双向残差连接和隐藏状态，使用注意力机制整合两个层次下的融合机制。实验结果证明，HFFN 能够有效提高准确率，三模态下的融合机制表现最好。

（2）中期模型融合

如图 6 所示，基于模型的融合是将不同的模态数据共同输入网络，基于模型的中间层进行融合，模型融合的好处是可以选择融合的位置，也可以实现模态间的交互性，基于模型的融合通常使用多核学习、神经网络、图像模型等方法。

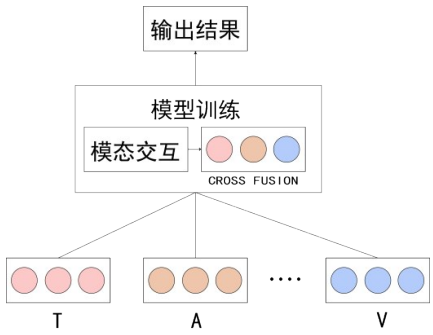


图 6　中期模型融合

文献[34]介绍了 Multiple Kernel Learning(MKL）算法，多核学习可以更好的融合异构数据，将多模态特征输入到核空间，能够获得比单核更好的性能。图像模型也是常见的模型融合方法，可以很好的利用不

同模态的时间信息,但多核学习和图像模型在多模态情感分析中的使用较少,神经网络是多模态情感分析中最常使用的模型融合算法。文献[35]使用Bert和VQ-Wav2Vec预训练模型提取特征,使用浅层的特征融合将提取出的特征向量直接连接起来;使用CoAttention机制,进行语音和文本间的模态交换,具体的方法是,一个模态的Key-Value是通过另一个模态的Query来进行计算的。文献比较了不同的融合机制如何影响模型的性能,结果显示,纯文本的效果比纯语音要好,浅层的特征融合比单模态效果要好,而在网络固定的情况下,Co-Attentional模型的表现更好,更多的交互使得Co-Attentional机制的适应性更好。文献[36]使用BiGRU学习文本,使用VGG网络学习图片,提出了使用视觉信息对齐文档的VistaNet网络,将视觉模态整合到文本信息。文献[37]提出了基于耦合平移的融合网络,使用Transformer的解码部分,使用并行融合策略,在公共数据集上达到了先进的性能。

(3)后期决策融合

如图7所示,决策级融合是在后期各个单模态分别训练完后,将各个模态的结果进行决策打分,即对每个模态的预测结果进行集成。在某些模态数据缺失时,决策级融合也能具有良好表现,且来自不同模态的数据可以分别运用合适的分类器进行训练,不同模态间的错误不会互相影响。决策级融合常见的融合机制有加权、投票、集成学习、规则融合等。

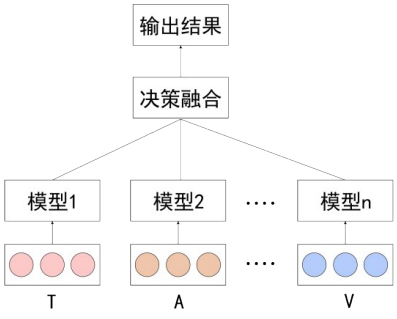


图7 后期决策融合

文献[38]认为组合多个模态信息有助于解决模糊信息,开发并比较了特征融合和决策融合,具体的方法是使用PCA进行降维,去除冗余特征后,进行简单的未加权拼接和使用训练单独模态所学习到的权值实现特征融合,使用分类器利用权重进行决策融合。结果 表明,特征融合、决策融合两种融合方式都能取得比单模态更好的性能;PCA可以改善直接连接的特征融合性能;不使用PCA进行降维的视频、文本双模态在决策融合中表现更好;而在三种融合方式中,学习模态权重,不使用PCA的视频、文本双模态在所有实验中表现最佳,可以达到76%的准确率。文献[39]提出了一种量子认知来驱动决策的多模态决策级融合策略。具体的方法是,通过将话语分为积极与消极情绪判断的量子叠加状态在一个具有正算子值测度的复值希尔伯特空间上,将单模态分类器建模为复值空间上的互不相容的可观察量,从训练数据中估计复值希尔伯特空间和单模态可观察量,然后从学习到的单模态可观察量中建立测试话语的最终多模态情感状态。文献[40]指出,在微表情识别中运用语音作为辅助信息,可以有效提高模型准确率,其提出了包含数据级和决策级融合的T1MF方法,使用张量融合网络生成文本、音频和文本、视频的嵌入,在决策融合层面,根据后验概率输出单个分类器的得分矩阵,引入软融合进行决策从而获得新的预测标签。

### 3.3 多模态对齐方法

多模态对齐是寻找两个或两个以上模态之间的对应关系,在多模态融合的过程中,存在文本、音频、视频不同步的现象,采用多模态对齐方法可以有效解决这种问题,现在的主流做法是基于时间序列的。文献[41]指出,语音和文本在时间上存在固有的共存关系,对齐对多模态学习是有益的,文献将语音经过语音识别(ASR)识别成文字,利用注意力网络学习语音和文本在时间域上的一致性,计算语音编码器和文本编码器隐藏状态间的权重,将语音特征和文本特征在词的层次上结合,实现语音模态和文本模态的对齐,实验结果证明,该方法优于直接连接的方法,体现了文本和语音学习对齐的优势。

### 3.4 多模态训练模型

在多模态情感分析的研究历程中,对模型的改进也是研究重点,目前已经有许多优秀的成果,主要可以分为句子级多模态情感分析和对话级多模态情感分析。

(1)基于句子级的多模态情感分析

由于对话间交互过程建模难度大,多模态情感分析的研究大部分都是基于独立的话语,许多学者致力于判断孤立句子的情感极性,基于句子级别使用先进的方法如注意力机制、迁移学习、预训练模型等对模型进行改进。文献[42]采用并行的交叉和自注意力机制来模拟模态之间的交互关系。文献[43]指出,不

是所有单模态的贡献度都相同，文献提出了一个轻量级掩码层M3，在训练过程中对主模态文本进行掩盖，提高弱模态的贡献，结果证明，多模态掩码能有效提高模型准确率。

（2）基于对话级的多模态情感分析

情感分析的适用场景多半是基于真实对话场景下的，对话之间会产生交互关系，而当说话人进行交谈时，说话人自身和说话人之间的情感也存在着依赖关系，捕捉这种交互性也可以有效提高情感识别的准确率。在多模态情感分析的研究历程中，许多学者开始着眼于对话间的交互关系，如文献[32]指出，说话人的自我影响与情绪惯性有关，情绪可以在一个时刻延续到另一个时刻，文献中使用GRUs对说话人的历史语句进行建模，基于改进记忆网络的协同记忆网络（CMN）模型，通过注意力的跳跃捕获说话者之间的依赖关系。文献[44]提出了DialogueRNN模型，对双方说话者和全局进行三方建模，使用三个GRU进行存储，采用注意力机制，实现了更好的上下文表征。而文献[45]提出了DialogueGCN模型，通过图神经网络为对话者之间的依存关系建模上下文，有效改善了上下文理解和依存关系。

## 4 多模态情感分析相关数据集

如表1所示，在多模态情感分析的研究中，科研人员们创建了许多不同模态和类型的数据集以供研究。在双模态数据集中，较常使用的有从Yelp.com评论网站收集评论构建的Yelp数据集[46]、手机评论相关信息的Multi-ZOL数据集、YouTube上电影评论视频为主的CMU-MOSI数据集[47]；在三模态数据集中，使用较多的有从YouTube搜集构建的YouTube数据集、由10位演员情感互动交流12h视听数据的IEMOCAP数据集[48]、从美剧《老友记》中剪辑出的聊天对话片段等MELD数据集。

双模态数据集中的模态并不总是相同的，一般是文本、图像和语音的两两组合。其中Yelp数据集是由233569张图像和44305条文本组成的；Multi-ZOL数据也是由图像和文本组成，共有5288条数据，每条数据至少有一条文本和一个图像；CMU-MOSI数据集是从YouTube上收集的93个有关电影评论的视频，共89人2~5 min的电影评论；CHEAVD2.0[49]则由视频和音频数据组成；SEMAINE[50]由150人参与录制的959段对话数据组成，包含视频和音频数据；DEAP数据集则由32名受试者观看视频的脑电信号和面部表情的视频数据组成。

三模态数据集中的模态多是由文本、图像、音频和视频组合而成。其中较为出名的数据集IEMOCAP由10位演员情感互动交流的大约12个小时视听数据组成，有音频、文本和视频数据；YouTube数据集评论视频数据由47人对产品的评论视频组成，有文本、视频、音频数据；CH-SIMS数据集[51]由2281个视频片段组成，且视频片段中只有说话者的面部，有文本、图像、音频数据；ICT-MMMO数据集由370个影评视频组成，有文本、图像、音频数据；MELD数据集[52]由1433个聊天片段组成，有文本、视频、音频数据。

**表1 多模态相关数据集**

| | 名称 | 模态 | 标签 | 资源 |
|---|---|---|---|---|
| 双模态情感数据集 | Yelp | 图像、文本 | 五类 | https://www.yelp.com/dataset/challenge |
| | Multi-ZOL | 图像、文本 | 十类 | https://github.com/xunan0812/MIMN |
| | CMU-MOSI | 图像、文本 | 六类 | https://www.amir-zadeh.com/dataset |
| | CHEAVD2.0 | 视频、音频 | 八类 | http://www.chineseldc.org/emotion.html |
| | SEMAINE | 视频、音频 | 八类 | http://semaine-db. eu |
| | DEAP | 脑电、视觉 | 九类 | http://www.eecs.qmul.ac.uk/mmv/datasets/deap/ |
| 三模态情感数据集 | IEMOCAP | 文本、视频、音频 | 九类 | http://sail.usc. edu/iemocap/ |
| | YouTube | 文本、视频、音频 | 三类 | 邮件获取:stratou@ ict.usc.edu |
| | CH-SIMS | 文本、图像、音频 | 三类 | https://github.com/thuiar/MMSA |
| | ICT-MMMO | 文本、图像、音频 | 三类 | https://github.com/A2Zadeh/CMU-MultimodalSDK |
| | MELD | 文本、视频、音频 | 七类 | https://affective-meld. Github.io/ |

## 5 结束语

本文对多模态情感分析的研究进展和主要研究过程进行了梳理,阐述了多模态的特征提取、融合方式和模型改进等方面的现状,介绍了多模态的相关数据集资源。在多模态情感分析领域仍有许多待解决的问题,其面临的挑战如下:

1)不同模态之间的特征可靠性不完全一样,目前大部分研究表明文本模态的可靠性较强,且不同模态之间存在着依赖关系。

2)模态连接后容易产生高维灾难,增加了计算复杂度。且融合模型的时候很难利用模态间的互补性,由于模态采样率、噪音类型、强度等因素的不同,在同一时刻模态的密集程度不同,不同模态信息很难做到完全对齐。

3)目前大部分多模态情感分析都采用了文本、视频和音频,脑电、生理信号等模态的数据集缺少,考虑其他更多的模态,可以给多模态情感分析领域带来更多可能性。

4)情感是复杂的决策过程,人类的决策在某些情况下是高度非理性的,大部分研究多半考虑了不同模态之间融合,如何搭配不同的模态模拟复杂决策过程也尤为重要。

5)在对话时通常存在许多个谈话者,考虑到在真实世界的对话场景中,人的情感中通常存在着交互,时序信息显得尤为重要,在不同的对话场景中相同的话语也会存在不同的含义,对情感交互性进行研究,可以有效提高模型在实际场景中的泛化能力。

### 参考文献(References):


[1] Cabanac M. What is emotion?[J]. Behavioural processes, 2002, 60(2): 69-83.

[2] Scherer K R. Emotion as a process: function, origin and regulation[J]. Social Science Information, 1982, 21(4-5): 555-570.

[3] Picard R W. Affective computing[M]. Cambridge, Mass: MIT press, 2000.

[4] 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.

[5] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 384-394.

[6] Lin Y L, Wei G. Speech emotion recognition based on HMM and SVM[C]. International Conference on Machine Learning and Cybernetics, 2005, 8: 4898-4901.

[7] 李悦, 黄永明, 章国宝, 刘海彬.基于角度差和散度均值特征的人脸表情识别[J].中南大学学报(自然科学版), 2013, 44(S2):250-253.

[8] Bhakre S K, Bang A. Emotion recognition on the basis of audio signal using Naive Bayes classifier[C]. International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016: 2363-2367.

[9] Mehmood R M, Lee H J. Emotion classification of EEG brain signal using SVM and KNN[C]. IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2015: 1-5.

[10] Wang J, Yu L C, Lai K R, et al. Dimensional sentiment analysis using a regional CNN-LSTM model[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016: 225-230.

[11] Munikar M, Shakya S, Shrestha A. Fine-grained sentiment classification using bert [C]. Artificial Intelligence for Transforming Business and Society (AITB), 2019, 1: 1-5.

[12] Ekman P. Expression and the nature of emotion[J]. Approaches to Emotion, 1984, 3(19): 344.

[13] Roseman I J, Spindel M S, Jose P E. Appraisals of emotion-eliciting events: Testing a theory of discrete emotions [J]. Journal of Personality and Social Psychology, 1990, 59 (5): 899.

[14] Russell J A. A circumplex model of affect[J]. Journal of personality and social psychology, 1980, 39(6): 1161.

[15] Russell J A, Mehrabian A. Evidence for a three-factor theory of emotions [J]. Journal of Research in Personality, 1977, 11(3): 273-294.

[16] Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.′ s negative-sampling word-embedding method [DB/OL]. arXiv preprint arXiv:1402.3722, 2014.

[17] Yang H J, Lee G S, Kim S H. End-to-end learning for multimodal emotion recognition in video with adaptive loss [J]. IEEE MultiMedia, 2021, 28(2): 59-66.

[18] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.

[19] Peters E M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations[C]. Proceedings of the 2018 Conference of the North American Chapter of the Associa-

tion for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018: 2227-2237.

[20] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: 4171-4186.

[21] Chen F, Sun Z, Ouyang D, et al. Learning What and When to Drop: Adaptive Multimodal and Contextual Dynamics for Emotion Recognition in Conversation [C]. Proceedings of the 29th ACM International Conference on Multimedia, 2021: 1064-1073.

[22] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor [C]. Proceedings of the 18th ACM international conference on Multimedia, 2010: 1459-1462.

[23] Yang K, Wang C, Gu Y, et al. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition [J]. IEEE Transactions on Affective Computing(Early Access), 2021.

[24] Sánchez-Lozano E, Lopez-Otero P, Docio-Fernandez L, et al. Audiovisual three-level fusion for continuous estimation of russell′s emotion circumplex [C]. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013: 31-40.

[25] Zhong L, Liu Q, Yang P, et al. Learning multiscale active facial patches for expression analysis [J]. IEEE transactions on cybernetics, 2014, 45(8): 1499-1510.

[26] Liu J, Chen S, Wang L, et al. Multimodal emotion recognition with capsule graph convolutional based representation fusion [C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 6339-6343.

[27] Baltrušaitis T, Robinson P, Morency L P. Openface: an open source facial behavior analysis toolkit [C]. IEEE Winter Conference on Applications of Computer Vision (WACV), 2016: 1-10.

[28] Sun X, Huang J, Zheng S, et al. Personality Assessment based on Multimodal Attention Network Learning with Category-based Mean Square Error [J]. IEEE Transactions on Image Processing, 2022, 31: 2162-2174.

[29] Cambria E, Hazarika D, Poria S, et al. Benchmarking multimodal sentiment analysis [C]. International Conference on Computational Linguistics and Intelligent Text Processing, Springer, Cham, 2017: 166-179.

[30] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6897-6906.

[31] Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text [C]. IEEE Spoken Language Technology Workshop (SLT), 2018: 112-118.

[32] Hazarika D, Poria S, Zadeh A, et al. Conversational memory network for emotion recognition in dyadic dialogue videos [C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018: 2122-2132.

[33] Mai S, Hu H, Xing S. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing [C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 481-492.

[34] Gehler P, Nowozin S. On feature combination for multiclass object classification [C]. IEEE 12th International Conference on Computer Vision, 2009: 221-228.

[35] Siriwardhana S, Reis A, Weerasekera R, et al. Jointly Fine-Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition [C]. Proc Interspeech, 2020, 3755-3759.

[36] Truong Q T, Lauw H W. Vistanet: Visual aspect attention network for multimodal sentiment analysis [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 305-312.

[37] Tang J, Li K, Jin X, et al. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network [C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 5301-5311.

[38] Williams J, Comanescu R, Radu O, et al. Dnn multimodal fusion techniques for predicting video sentiment [C]. Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), 2018: 64-72.

[39] Gkoumas D, Li Q, Dehdashti S, et al. Quantum cognitively motivated decision fusion for video sentiment analysis [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 827-835.

[40] Sun J, Yin H, Tian Y, et al. Two-level multimodal fusion for sentiment analysis in public security [J]. Security and Communication Networks, 2021.

[41] Xu H, Zhang H, Han K, et al. Learning alignment for multimodal emotion recognition from speech [C]. Proc Inter-

speech 2019: 3569-3573.

[42] Sun L, Liu B, Tao J, et al. Multimodal cross-and self-attention network for speech emotion recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 4275-4279.

[43] Georgiou E, Paraskevopoulos G, Potamianos A. M3: Multi-Modal Masking applied to sentiment analysis[C]. Proc Interspeech 2021: 2876-2880.

[44] Majumder N, Poria S, Hazarika D, et al. Dialoguernn: an attentive rnn for emotion detection in conversations[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 6818-6825.

[45] Ghosal D, Majumder N, Poria S, et al. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation[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) 2020: 154-164

[46] Asghar N. Yelp dataset challenge: review rating prediction [DB/OL]. arXiv preprint arXiv:1605.05362, 2016.

[47] Zadeh A, Zellers R, Pincus E, et al. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. IEEE Intelligent Systems, 2016, 31 (6): 82-88.

[48] Busso C, Bulut M, Lee C C, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335-359.

[49] Li Y, Tao J, Chao L, et al. CHEAVD: a Chinese natural emotional audio-visual database[J]. Journal of Ambient Intelligence and Humanized Computing, 2017, 8(6): 913-924.

[50] McKeown G, Valstar M, Cowie R, et al. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE transactions on affective computing, 2011, 3(1): 5-17.

[51] Yu W, Xu H, Meng F, et al. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3718-3727.

[52] Poria S, Hazarika D, Majumder N, et al. Meld: a multimodal multi-party dataset for emotion recognition in conversations[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 527-536.

**编辑:**王谦,王雨田