计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：基于多模态融合的情感分析算法研究综述
作者：郭续，买日旦·吾守尔，古兰拜尔·吐尔洪
网络首发日期：2023-08-09
引用格式：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/OL]．计算机工程与应用．
https://link.cnki.net/urlid/11.2127.TP.20230809.1432.004

网络首发时间：2023-08-09 15:56:26
网络首发地址：https://link.cnki.net/urlid/11.2127.TP.20230809.1432.004

基于多模态融合的情感分析算法研究综述

郭续， 买日旦·吾守尔， 古兰拜尔·吐尔洪

新疆大学 信息科学与工程学院，新疆维吾尔自治区，乌鲁木齐 830046

摘 要：情感分析是一项新兴技术，其旨在探索人们对实体的态度，可应用于各种领域和场景，例如产品评价分析、舆情分析、心理健康分析和风险评估。传统的情感分析模型主要关注文本内容，然而一些特殊的表达形式，如讽刺和夸张，则很难通过文本检测出来。随着技术的不断进步，人们现在可以通过音频、图像和视频等多种渠道来表达自己的观点和感受，因此情感分析正向多模态转变，这也为情感分析带来了新的机遇。多模态情感分析除了包含文本信息外，还包含丰富的视觉和听觉信息，利用融合分析可以更准确地推断隐含的情感极性(积极、中性、消极)。多模态情感分析面临的主要挑战是跨模态情感信息的整合，因此，本文重点介绍了不同融合方法的框架和特点，并对近几年流行的融合算法进行了阐述，同时对目前小样本场景下的多模态情感分析进行了讨论，此外，还介绍了多模态情感分析的发展现状、常用数据集、特征提取算法、应用领域和存在的挑战。期望这一综述能够帮助研究人员了解多模态情感分析领域的研究现状，并从中得到启发，开发出更加有效的模型。

关键词：多模态；情感分析；模态融合
文献标志码:A 中图分类号:TP391 doi：10.3778/j.issn.1002-8331.2305-0439

Survey of Sentiment Analysis Algorithms Based on Multimodal Fusion

GUO Xu, Mairidan Wushouer, Gulanbaier Tuerhong

School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

Abstract：Sentiment analysis is an emerging technology that aims to explore people's attitudes toward entities and can be applied to various domains and scenarios, such as product evaluation analysis, public opinion analysis, mental health analysis and risk assessment. Traditional sentiment analysis models focus on text content, yet some special forms of expression, such as sarcasm and hyperbole, are difficult to detect through text. As technology continues to advance, people can now express their opinions and feelings through multiple channels such as audio, images and videos, so sentiment analysis is shifting to multimodality, which brings new opportunities for sentiment analysis. Multimodal sentiment analysis contains rich visual and auditory information in addition to textual information, and the implied sentiment polarity (positive, neutral, negative) can be inferred more accurately using fusion analysis. The main challenge of multimodal sentiment analysis is the integration of cross-modal sentiment information, therefore, this paper focuses on the framework and characteristics of different fusion methods and describes the popular fusion algorithms in recent years, and discusses the current multimodal sentiment analysis in small sample scenarios, in addition to the current development status, common datasets, feature extraction algorithms, application areas and challenges. It is expected that this review will help researchers understand the current state of research in the field of multimodal sentiment analysis and be inspired to develop more effective models.

Key words：multimodal; emotional analysis; modal fusion

情绪是人们对特定话题、人或实体所持有的情感体验和反应。情感分析可以识别多模态包含的情感信号，例如愤怒、喜悦、悲伤等。这项技术在商业和政治等领域中具有广泛的应用，比如商家若能了解消费者对其品牌或产品的评价，便能够找到改进其产品和服务的方向，从而提高客户满意度；政治家可以通过

基金项目：国家自然科学基金(61961039)。

作者简介：郭续(1998－)，男，硕士研究生,CCF 学生会员，主要研究方向为多模态情感分析、机器学习和深度学习；买日旦 吾守尔(1984－)，通信作者，男，博士，副教授，CCF 会员，主要研究方向为自然语言处理、语音合成、机器学习等，E-mail: mardan@xju.edu.cn；古兰拜尔 吐尔洪(1985－)，女，博士，副教授，CCF 会员，主要研究方向为网络安全和多模态机器学习等。

民意调查和对社交媒体上的情感进行分析，了解选民的态度和偏好，以制定更符合公众利益的政策。

早期的情感分析主要聚焦于文本数据，如今，在许多行业中，基于文本的情感分析已经成为了一种常见的解决方案。它被广泛应用于电影票房业绩预测[1]、股市业绩预测[2]、选举结果预测[3]等领域，但依靠文本数据是不能完全提取人类表达的所有情感，例如：模型对文本中“优秀”一词的分析通常是积极的，但是如果加上夸张或带有讽刺的表情，就很可能变成消极的情绪，这时候多模态情感分析就被提出来解决这个问题。2015 年发布的一项多模态情感分析调查报告显示，多模态系统始终比最好的单模态系统更准确[4]。

近年来，在社交多媒体平台(YouTube, Bilibili)上发布的视频博客(vlog)或口头评论，包含了用户情感的表达，例如描述用户谈论产品或电影的视频，视频不仅提供文本信息，还提供丰富的视觉和音频信息。图 1 给出了一个典型的多模态情感分析系统的总体框架，该框架可分为单模态数据处理和多模态数据融合两部分。首先，特征提取器分别应用于文本、视觉和声学数据提取特征；然后将提取的特征信息转化到融合模型中进行情感预测，这两个组件对整个模型的性能都很重要。仅进行单模态情感分析会导致对模态内相互作用的理解不足，从而降低多模态系统的性能，此外，低效率的多模态融合也会使各模态之间的相互作用得不到充分利用，从而影响多模态系统的稳定性和性能提升。因此，为了提高多模态系统的性能和稳定性，需要进行充分的模态间交互分析和高效的多模态融合。

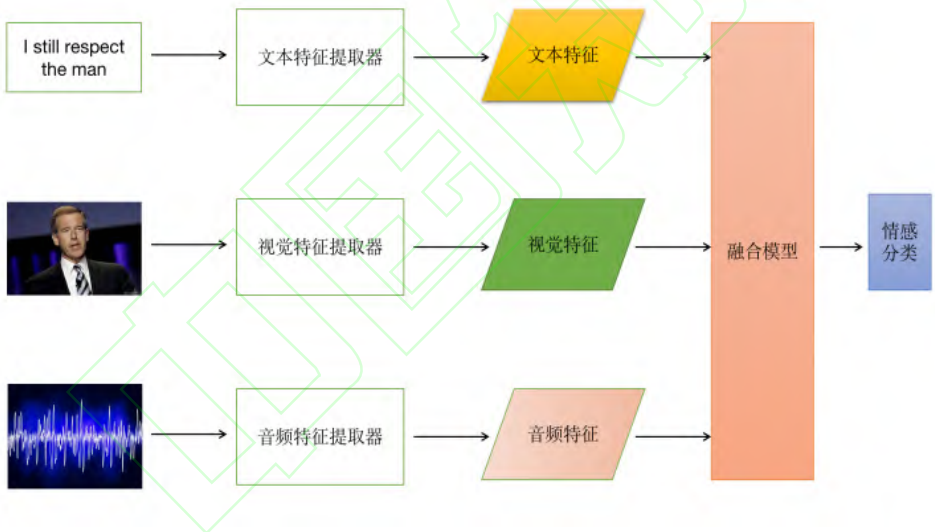

图 1 多模态情感分析系统总体框架

Fig.1 General framework of multimodal sentiment analysis system

随着对多模态情感分析的研究越来越多，了解该领域最新的研究方法变得尤为重要。在 2017 年，Poria 等人[5]就提出了从单模态到多模态融合的情感计算综述，作者阐述了情感分析的一些基本特征提取方法和模型框架，并对多模态情感分析的潜在性能提出了一些改进方法。同年，Soleymani 等人[6]定义了情感和多模态情感分析的问题，并总结了多模态情感分析在不同领域的最新进展。虽然这些综述论文对当时的多模态发展进行了全面的概述，但在过去的几年里，有许多丰富的数据集和先进的模型被提出，例如，该领域最流行的两个数据集 CMU-MOSI[7]和 CMU-MOSEI[8]在这些早期的论文中并未提及，也很少提及基于注意力机制（Attention）的模型。2019 年，Huddar 等人[9]阐述了多模态情感分析中存在的问题和挑战，回顾了近年来使用多模态情感分析的一些计算方法。2021 年，Gkoumas 等人[10]利用 CMU-MOSI 和 CMU-MOSEI 对 11 个最先进的模型进行了详细的实验分析，作者发现，使用注意力机制的模型通常能获得更好的结果。Chandrasekaran 等人[11]研究了多模态情感分析在社交媒体上的应用，并提出了大量的方法。基于以上综述论文都没有从融合方法的角度对该领域中存在的模型进行详细的描述，且近几年少量标注数据的少样本场

景更为常见，因此，本文的工作详细分析了现有模型的融合方式，建立了五种融合方法的分类框架，并详细介绍了近年来表现较好的模型算法，此外，对少样本场景下的多模态情感分析进行了讨论。

1 多模态情感分析数据集

基于多模态的情感数据集大多都来源于社交媒体发布的视频或根据自己的需求建立数据集。表 1 总结了当前多模态情感分析领域的常用数据集，并给出了数据集的名称、数据集中所包含的语言、数据来源、数据所含模态（T 代表文本，V 代表图像，A 代表音频）、数据集标注的情感标签以及提供了对该数据集的访问，其中一些数据集可以直接从网站下载，一些数据集可以通过电子邮件联系作者。

表 1 多模态情感分析数据集

Table 1 Multimodal sentiment analysis dataset

数据集名称	语言	来源	所含模态	情感标签	下载地址
CMU-MOSI[7]	英文	YouTube	T+V+A	[-3,+3]	http://immortal.multicomp.cs.cmu.edu/raw datasets/processed data/
CMU-MOSEI[8]	英文	YouTube	T+V+A	[-3,+3]	http://immortal.multicomp.cs.cmu.edu/raw datasets/processed data/
CMU-MOSEAS[12]	西班牙语、德语、葡萄牙语、法语	YouTube	T+V+A	[-3,+3]	https://bit.ly/2Svbg9f
ICT-MMMO[13]	英文	YouTube, ExpoTV	T+V+A	[-2,+2]	http://immortal.multicomp.cs.cmu.edu/raw datasets/processed data/
CH-SIMS[14]	中文	电影，电视剧，综艺	T+V+A	[-2,+2]	https://github.com/thuiar/MMSA
MOUD[15]	西班牙语	YouTube	T+V+A	[-1,+1]	http://immortal.multicomp.cs.cmu.edu/raw datasets/processed data/
IEMOCAP[16]	英文	南加州大学	T+V+A	[1,10]	http://sail.usc.edu/iemocap/
YouTube[17]	英文	YouTube	T+V+A	[-1,+1]	http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/

CMU-MOSI 数据集[7]：该数据集是第一个用于在线视频中情绪和主观性分析的意见级标注语料库，数据集包含 93 个关于电影评论的视频博客，共包含 3702 个视频片段，其中 2199 个片段为评论部分，拥有从-3 到+3 的情绪标签。

CMU-MOSEI 数据集[8]：该数据集包括来自 1000 个不同演讲者(57%男性，43%女性)的 23453 个注释视频剪辑，情感标注从-3 到+3。

CMU-MOSEAS 数据集[12]：该数据集共包含 4000 个独白视频，其中每种语言各 1000 个，涵盖了 1645 个说话者，并且使用 20 个标签来标注句子，其中包括情感、主观性和属性等，情感标签的取值范围在-3 到+3 之间。

ICT-MMMO 数据集[13]：该数据集包括来自 YouTube 和 ExpoTV 中的 370 个关于电影评论的真实影评视频。视频的长度在 1-3 分钟之间。情绪标签包括强烈消极、弱消极、中性、弱积极和强烈积极 5 种。

CH-SIMS 数据集[14]：该数据集收集了 60 个原创视频，共 2281 个视频片段，每个视频片段都有多模态注释和三个单模态注释，情绪被分为消极、弱消极、中性、弱积极和积极 5 种。

MOUD 数据集[15]：该数据集由随机选择的 80 个视频组成，从每个视频中人工选择一个 30 秒的意见片段，平均分成 6 个话语，共得到 498 个话语数据集，每个话语被标记-1 到+1 之间。

IEMOCAP 数据集[16]：该数据集由大约 12 小时的视听数据组成，10 位演员（5 位男演员和 5 位女演员）按照选定的情感剧本表演，也即兴创作了一些假想的

场景，旨在诱发特定类型的情绪，包括快乐、愤怒、悲伤、沮丧和中立状态。情感标签共有 10 个，包括愤怒、快乐、悲伤、中立等。

YouTube 数据集[17]：该数据集是在三模态情感分析任务中使用的第一个数据集，数据集共包含 47 个（13 个积极、22 个中性和 12 个消极）视频序列，它们由不同年龄和不同种族背景的人通过英语来讲述关于产品的观点，每个视频都为 30s，并被切分成 3 到 11 个话语，并赋予消极、中性和积极的情感极性。

2 特征提取

单模态特征提取是多模态情感分析系统的重要组成部分。在本节中，将分别介绍文本、音频和视觉的特征提取方法，并列出了一些使用本文所提到的特征提取方法的模型，如表 2 所示。

表 2 模型所使用的特征提取方法

Table 2 The feature extraction method used by the model

模型	文本	视觉	音频
SVM[15]	Bag-of-Words	CERT	OpenEAR
MKL[18]	Word2vec	CLM-Z	openSMILE
SAL-CNN[19]	Word2vec	CLM-Z	openSMILE
TFN[20]	GloVe	Facet	COVAREP
LMF[21]	GloVe	Facet	COVAREP
MulT[22]	GloVe	Facet	COVAREP
BC-LSTM[23]	Text-CNN	3D-CNN	openSMILE
MMMU-BA[24]	GloVe	Facet	COVAREP
CHFusion[25]	CNN	3D-CNN	openSMILE
Gated mechanism for attention[26]	CNN	3D-CNN	openSMILE
ICDN[27]	GloVe	Facet	COVAREP

2.1 文本特征提取

在文本特征提取中，常用的技术包括词袋模型(Bag-of-Words)、词频和逆向文档频率(TF-IDF)[28]、N-grams 和词嵌入等，这些技术都是为了将文本转化为机器学习算法可以处理的数值特征。Mikolov 等人[29]在 2013 年提出的 Word2Vec 是最常用的词嵌入方法。

最近的研究使用 GloVe[30]来提取文本特征，GloVe 添加了基于统计的信息，这使得 GloVe 可以同时使用语料库的全局信息和本地上下文特征。卷积神经网络（Convolutional Neural Network，CNN）最初是为图像处理任务设计的，近年来它在文本特征提取中也被广泛应用。在文本分类任务中，经过 CNN 处理后的文本特征通常被馈送到全连接层进行分类，由于 CNN 的并行性质，它在文本处理中具有快速和高效的计算能力。Poria 等人[31]的工作证明了基于卷积神经网络的方法在提取文本特征方面非常有效，该方法为每个文本构建包含重要特征的特征向量，该向量集合能够表示整个文本的特征。此外，像 BERT[32]这样的大型预训练模型也经常被使用，BERT 是一种基于 Transformer[33]架构的预训练模型，它在大量无标注文本数据上进行预训练，然后在有标注数据上进行微调以适应各种任务。相比于传统的文本特征提取方法，BERT 能够更好地处理语境、多义词和文本中的复杂关系。Munikar 等人[34]使用 BERT 模型在一个大规模情感分类数据集上进行预训练，对其进行微调以实现更细粒度的情感分类。Araci 等人[35]提出一种基于 BERT 的 FinBERT 语言模型来处理金融领域的任务，该模型能够对金融领域的语言和术语进行理解，并能够在不同的金融任务中进行微调。表 3 总结了文本特征提取常用技术的优点和不足之处。

表 3 文本特征提取技术

Table 3 Text feature extraction techniques

技术	优点	不足
Word2Vec	能学习词向量，可以保留词语的语义信息	计算量较大，需要大量的训练数据
GloVe	能够学习到词向量，对大规模语料库适用	计算量较大，需要大量的训练数据
CNN	自动学习特征表示，可以捕捉局部信息	需要大量的训练数据，对文本长度敏感
BERT	能够学习到双向上下文信息，表现优异	计算量大，需要大量的训练数据

2.2 音频特征提取

尽管深度神经网络在计算机视觉中被广泛应用于自动特征提取，但其在语音领域的应用仍然存在一些挑战。这是因为音频信号具有时间序列性质，需要对其进行序列建模，同时需要考虑语音信号的时变性质。因此，对于音频信号的处理需要采用一些特殊的技术，如 CNN、长短期记忆网络（Long Short-Term Memory Network，LSTM），双向 LSTM[36][37]和注意力机制等。目前，大多数多模态情感分析模型都使用开

源库 OpenEAR[38]、openSMILE[39]、LibROSA[40]、COVAREP[41]等来提取声学特征。其中，OpenEAR 是一个基于 MATLAB 的开源库，用于情感识别、情感表达和情感辨别任务。OpenEAR 自动计算一组声学特征，包括韵律、能量、声音概率、频谱和倒谱特征，并使用 z 标准对说话人进行标准化。openSMILE 是一个用于音频特征提取的开源库，提取频率为 30Hz，带有滑动窗口。具体来说，openSMILE 提取的特征由几个低级描述符（LLD）组成，比如 MFCC、音调和声音强度以及它们的统计函数。LibROSA 是 Python 中常用的音频分析库，可以用于提取 22050Hz 的声学特征，包括 MFCC、色度频率 CQT、STFT、调频谱、功率谱、熵等。COVAREP 声学分析框架可以用于提取声学特征，包括 12 个 MFCC、声门源参数、峰值斜率参数、最大色散商（MDQ）和 Liljencrants-Fant（LF）。有论文[42][43]指出基于音频的情感分析对于人机交互非常重要。目前，可以提取的语音相关特征分为局部特征和全局特征两组。Ayadi 等人[44]更倾向于使用全局特征，因为它们比其他特征具有更多的优势。表 4 总结了音频特征提取常用技术的优点及不足。

表 4 音频特征提取开源库

Table 4 Audio feature extraction open source library

技术	优点	不足
OpenEAR	提供了多种语音情感和语音行为特征	不支持跨平台操作，对于非专业用户，使用上较为困难
openSMILE	提供了多种语音情感和语音行为特征	处理音频时需要一些预处理步骤，对于非专业用户，使用上较为困难
LibROSA	提供了多种音频信号处理和分析功能	无法处理非线性的变换，计算速度较慢
COVAREP	提供了多种声音基本特征和高级特征	对于非专业用户，使用上较为困难

2.3 视觉特征提取

视觉特征提取技术是计算机视觉领域中的一项重要技术，其目的是从图像或视频中提取出具有代表性的特征，以便于后续的图像处理和分析。常用的视觉特征提取技术包括尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）[45]，加速稳健特征（Speeded-Up Robust Features，SURF），方向梯度直方图（Histogram of Oriented Gradients，HOG），CNN，生成对抗网络（Generative Adversarial Network，GAN），区域卷积神经网络（Region-based Convolutional Neural Network，RCNN）。其中，CNN 可以自动地从图像中学习特征表示，通过多层卷积层和池化层的组合能够提取出具有高度语义信息的特征表示。Tran 等人[46]提出了一种 3D 卷积神经网络（3D-CNN）学习时空特征的方法，可用于动作识别、相同动作判断、动态场景识别等不同任务。GAN 是一种生成模型，通过训练生成器和判别器两个模型的对抗学习，可以生成逼真的图像，并从中提取出具有代表性的特征。RCNN 是一种基于区域的 CNN，可以在图像中提取出具有代表性的目标区域，并对这些区域进行分类和检测。此外，还有一些特征提取工具，计算机表情识别工具箱 CERT[47]提供了多种视觉特征，包括面部表情、姿态、手势等，支持多种语言和平台。MultiComp OpenFace 2.0 工具包[48]可以提取 68 个面部标志、17 个面部动作单元、头部方向和眼睛凝视，其中 Facet 库能提取一组视觉特征，包括面部特征跟踪、头部姿势、HOG 特征等。表 5 总结了常用的视觉特征提取技术的优点及不足。

表 5 视觉特征提取技术

Table 5 Visual feature extraction techniques

技术	优点	不足
CNN	自动学习特征表示，准确率较高	计算量大，需要大量的训练数据
GAN	可以生成逼真的图像，具有良好的可扩展性	计算量大，训练不稳定
RCNN	可以准确地检测图像中的目标	计算量大，速度较慢，对目标尺寸变化敏感
CERT	提供了多种视觉特征，支持多种语言和平台	对于非专业用户，使用上较为困难
Open-Face 2.0	提供了多种视觉特征和情感分析功能，易于使用	需要较高的计算性能，对硬件要求较高
Facet	提供了多种面部表情识别功能和情感分析功能	对于非专业用户，使用上较为困难

3 多模态情感分析

在多模态情感分析中，模态融合的效果会直接影响结果的准确性。本章先根据不同的模态融合方式进行归纳总结，接着讨论了表现较好的模态融合算法，然后针对近几年在多模态领域中小样本场景进行了简单的介绍和讨论，最后对不同算法的性能进行了分析讨论。

3.1 基于融合方法的多模态情感分析

利用有效的方法融合来自不同模态的特征信息是多模态情感分析的主要挑战。在本节中，我们将根据其融合方法分为 5 大类，并详细描述了每个方法和模型的优点及不足。

3.1.1 特征级融合

特征级融合是指将来自不同模态（如文本、图像、音频等）的特征进行组合，形成一个统一的特征向量，以进行情感分析任务。在融合过程开始前，将不同模态(例如音频、图像和文本)中提取的特征转换为相同的格式，将其进行简单的拼接，这种方式广泛出现在多模态情感任务中[49]。表 6 总结了基于特征级融合不同方法的优点及不足。

表 6 特征级融合方法

Table 6 Feature-level fusion methods

方法	优点	不足
HMM	模型具有很好的可解释性；对时间序列建模有很好的能力；利用 HMM 的概率计算解决多个模态之间的不匹配问题	模型的性能很大程度上依赖于参数；概率计算和解码算法较复杂
SVM	适用于数据量较少的情况；选择不同的核函数适应不同类型的数据，并具有较好的泛化能力	对于噪声和异常值比较敏感，可能会导致模型过度拟合
MKL	可以对特征权重进行自动学习，有效处理不同特征子集之间的差异	对于大规模数据集，训练时间较长

Morency 等人[17]首次提出了三模态情感分析任务，并采用自动提取多模态特征的方法进行情感分析。他们通过自动识别话语文本中的情感线索，生成语篇特征，从视频序列中自动提取视觉特征和音频特征。在提取每个模态的特征后，将它们串联在一起，并使用三模态 HMM 分类器来学习输入信号的隐藏结构。Pérez-Rosas 等人[15]提出了一种基于话语级别的情感分析方法，并构建了首个在话语层面进行情感分析的 MOUD 数据集。该方法通过构建词汇表和使用简单的加权图特征作为文本情感特征，使用 OpenEAR 进行语音特征提取，使用 CERT 进行面部特征提取，随后将其进行特征级融合并输入到支持向量机（Support Vector Machine，SVM）分类器中以得到情感极性。S. Park 等人[50]则采用支持向量机(SVMs)进行分类，支持向量回归进行回归实验，并使用径向基函数核作为预测模型。

Poria 等人[18]在提取三模态特征后，使用两种不同的特征选择器来减少特征数量。一种是基于循环相关的特征子集选择(CFS)，另一种是基于主成分分析(PCA)。该特征选择方法除了提高了模型的处理时间外，对实验结果也有一定的改善。随后将处理后的特征向量串联起来，利用多核学习(MKL)算法对分类器进行训练。第二年，作者[51]在其工作的基础上提出了卷积递归多核学习(CRMKL)模型，具体使用卷积 RNN 进行视觉情感检测，进一步改善了实验结果。

3.1.2 决策级融合

决策级融合首先对每个模态进行情感分析，然后将单模态情感决策纳入最终决策的不同机制。这种方法的优点是，每个模式都可以使用其最适合的分类器来学习其特征。但是，由于每个模态都建立了独立的分类器，模态间的交互往往不能有效地建模，且分类器的学习过程会变得繁琐且耗时[5]。表 7 总结了决策级融合不同方法的优点及不足。

表 7 决策级融合方法

Table 7 Decision-level fusion methods

方法	优点	不足
平均	简单易懂，容易实现	忽略了每个模型的权重差异
选择-加性	考虑每个模态的权重差异，灵活性高	参数需要人为设定，需要进行调优
多数投票	能在不同程度噪声的模态中获得更可靠的结果	无法考虑不同结果之间的相对置信度

Nojavanasghari 等人[52]针对这三种模式（视觉，听觉和文本）中的每一种训练了一个单峰分类器，然后对单个单峰分类器的置信度进行平均，从而做出最终预测。Yu 等人[53]利用预训练好的词向量训练 CNN 进行文本情感分析，使用 DNN 和广义缺失进行视觉情感分析，采用平均策略和权重融合最终的结果。Hussain 等人[54]提出了基于加权多数投票技术的混合融合方法来完成情感分类。

由于一些数据集中注释的内容较少，一些人物的身份特征会对结果产生影响，Wang 等人[55]提出了 Select-Additive Learning 方法，该方法通过选择性地学习个体特定的特征和共享特征，改善了多模态情感分析中的跨个体泛化性能。随后，又在其基础上提出了一个 SAL-CNN 模型[19]，在对 CNN 模型进行充分训练后，利用 SAL 提高模型的通用性和预测情绪。

3.1.3 基于张量融合

基于张量的方法主要是通过计算单模态句子表示的张量积来得到多模态句子表示。这需要首先将输入表示转换为高维张量，然后将其映射回低维输出向量空间，这是一种典型的非串联特征融合方法。张量的强大之处在于，它能够捕捉跨越时间、特征维度和多种形式的重要高阶相互作用[56]。然而，这种方法的缺点是计算复杂度呈指数增长，并且在不同模态之间缺乏细粒度的字级交互。表 8 总结了基于张量融合不同方法的优点及不足。

表 8 基于张量融合方法

Table 8 Based on tensor fusion method

方法	优点	不足
TFN	学习端到端的模态内和模态间动力学；张量融合层使用嵌入层的三个输出向量的三倍笛卡尔积	得到的表示具有非常大的维数，有大量的参数
LFM	类似于 TFN，但增加了一个额外的低秩因子，以减少计算内存；能在大范围的低阶设置中稳健地执行并且在训练和推断中更有效	局部相互作用的建模被忽略了
T2FN	基于张量秩最小化的正则化方法	不完美数据增加张量秩

Zadeh 等人[20]提出了一种张量融合网络(TFN)模型，通过使用张量表示和张量操作，TFN 模型能够更好地捕捉多模态数据中的动态特性和模态间的相互关系。该模型分为三个部分：模态嵌入子网络、张量融合层和情感子网络。模态嵌入子网络使用带有遗忘门的 LSTM 网络来学习与时间相关的语言表示，然后将这些表示连接到一个全连接网络来获得语言嵌入。对于声学和视觉特征，分别使用 FACET 和 COVAREP 提取特征，然后经过平均池化处理后连接到深度神经网络以获得嵌入。在张量融合层中，对嵌入层的三个输出向量使用三阶笛卡尔积，充分结合了张量融合中的单峰、双峰和三峰相互作用。将得到的多模态张量传递给一个完全连通的深度神经网络——情感推理子网络，从而得到预测结果。良好的多模态时间序列通常展现出时间和模态之间的相关性，这种相关性可以通过低秩张量表示来捕捉[57]。为了解决将输入转换为张量时计算复杂度呈指数级增加的问题，Liu 等人[21]提出了一种低秩多模态融合(Low-rank Multimodal Fusion, LMF)方法，该方法使用低秩张量进行多模态融合，大大降低了计算复杂度。该模型与 TFN 相似，但将权值分解为低秩因子，减少了参数的数量。实验结果表明，该模型在大范围的低秩情况下具有良好的鲁棒性，在训练和推理方面比其他张量表示方法更有效。由于存在不完美的模态、缺失的条目以及受噪声干扰的情况，这可能会破坏这些自然相关性并导致高阶张量表示的出现，为了解决此类问题，Paul 等人[58]提出了一种基于张量秩最小化正则化方法的时间张量融合网络(T2FN)模型。该模型可以学习多模态数据中真实关联和潜在结构的张量表示，并对其秩进行有效的归一化，T2FN 在 TFN 的基础上添加了时间分量，增强了捕获高阶张量表示的能力，从而提高了预测性能，此外，T2FN 还能够适应不完整数据，反映了其鲁棒性。

3.1.4 基于上下文融合

之前的方法通常将每个话语视为一个独立的实体，忽略了视频中话语之间的依赖性[59]。基于语境的融合则考虑了语境中其他话语与目标话语之间的联系，从而获得了更好的融合效果，基于循环神经网络的模型通常用于整合上下文信息。表 9 总结了基于上下文融合方法的优点及不足。

表 9 基于上下文融合方法

Table 9 Contextual fusion based approach

方法	优点	不足
CHFusion	层次融合结构使每一对模态相互作用并组合成一个三模态矢量，捕捉了模态之间的相互关系；每一层使用 RNN 提取语境感知的话语特征	在 MOSEI 数据集上进行实验时，负面类别的性能较差
BC-LSTM	提出了一种基于 lstm 的语境话语级特征提取框架；该模型保留了话语的顺序，使连续话语能够共享信息	每个话语的重要性和它对每个情态的具体贡献是不被考虑的

Majumder 等人[25]提出了一种感知上下文的分层融合(CHFusion)模型，如图 2 所示，该模型使用分层结构不断融合多模态信息，并在每层融合后更新上下文信息。首先，通过不同的特征提取器得到三种模态话语的单模态特征，利用 GRU 提取语境感知的篇章特征，将包含上下文信息的单峰特征通过全连接层成对组合，融合后形成双峰特征向量，与单峰情况一样，GRU 也用于感知上下文。最后，通过全连接层将三个双模态融合向量组合成一个三模态融合向量，并使用 GRU 来传递上下文信息，模型的输出由 softmax 层生成。

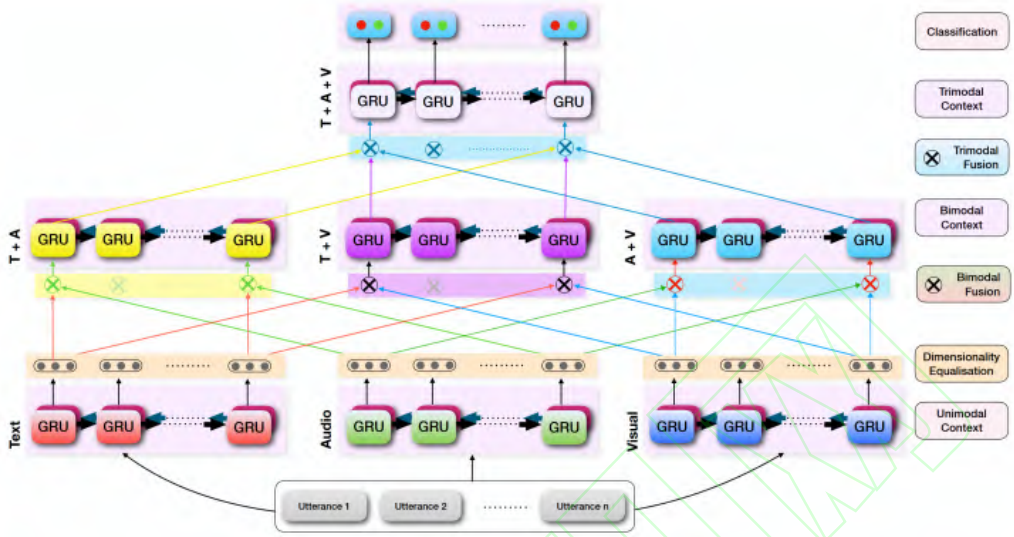

图 2 感知上下文的分层融合模型（CHFusion）

Fig.2 Hierarchical Fusion Model for Perceptual Context (CHFusion)

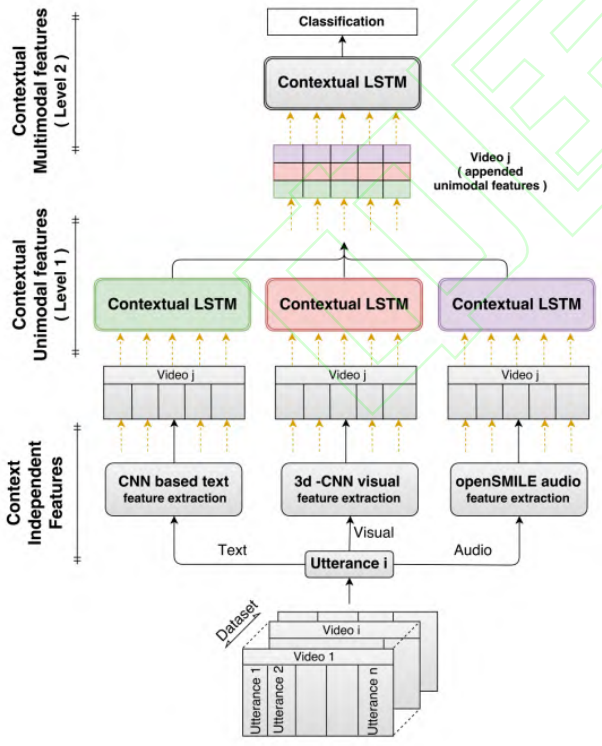

图 3 基于上下文相关的多模态话语特征的层次结构

Fig.3 A hierarchy of multimodal discourse features based on contextual relevance

Poria 等人[23]提出了 BC-LSTM 模型来捕获相同视频环境下话语的语境信息，图 3 为其结构框图。该模型首先使用不同的特征提取器提取不包含上下文信息的单峰特征，在提取单模态特征时，使用 text-CNN 提取文本特征，openSMILE 提取音频特征，3D-CNN 提取视觉特征，然后将这些特征输入到 LSTM 网络中。

为了更好地捕获上下文信息，作者将常规的 LSTM 替换为双向 LSTM，这样一个话语就可以从前面和后面的话语中获得信息。最后，将得到的包含上下文信息的单峰特征串联起来，并传递给类似的独立 BC-LSTM 进行训练，最后输出情感分类结果。

3.1.5 基于注意力机制

基于注意力机制的融合方法通过计算每个模态的注意力分数来进行加权融合。注意力分数反映了每个模态在情感分类中的重要性。这些分数是从每个模态的特征中计算出来的，并通过一个归一化函数进行标准化。在计算完注意力分数之后，每个模态的特征向量会乘以其对应的注意力分数，然后将它们相加以得到融合特征向量。最后，融合特征向量被送入分类器进行情感分类。基于注意力机制的融合方法有多种变体，如多头注意力机制、自注意力机制等。表 10 总结了基于注意力机制不同方法的优点及不足。

表 10 基于注意力机制融合方法

Table 10 Attention-based mechanism fusion method

方法	优点	不足
MulT	跨模态注意模块通过直接关注	当部分语言信息缺失时，

	其他模态中的低级特征来融合多模态信息；可以直接应用于非对齐的多模态流	模型缺乏对目标情态的持续关注和泛化，整体性能略低
ICDN	对传统 Transformer 的编码器进行了改进，使其能够接收多模态信息的输入;改进自监督单模态标签生成模块(ULGM)，利用单模态进行多任务学习，对多模态融合进行补充和推广	Mapping Transformers 使用后，仍然需要包含原始模态丰富信息的特征来进一步减少模态差异
UniMSE	使用一个统一的多模态知识共享框架来解决 MSA 和 ERC 任务	通用标签的生成只考虑文本形态，而没有考虑声学和视觉形态
CM-BERT	利用文本和音频模态的相互作用来微调预先训练好的 BERT 模型；使用掩蔽多模态注意动态调整词的权重	无法从未对齐的多模态数据中学习更好的表示
SPECTRA	适用于广泛的语音文本任务；能够挖掘对话中的语境信息来丰富话语表征	依赖于具有明确的词级语音-文本对齐标注的大规模会话语料库；只涉及语音和文本形式，缺乏图像等更多的模态

Tsai 等人[22]提出了一种名为 MulT（Multimodal Transformer）的多模态模型。该模型利用定向的双向跨模态注意力机制来实现不同时间步长的多模态序列之间的交互，并潜在地将信息从一种模态传递到另一种模态。为了捕捉时间信息，模型还引入位置嵌入（Positional Embedding），将时间信息融入序列中。此外，还引入了跨模态 Transformer 模块使不同模态之间可以通过跨模态注意力相互交互，每个模态都与其他两个模态进行交互，接收低级别的跨模态信息，并持续更新序列表示。最后，使用自注意力机制对具有相同目标模态的跨模态 Transformer 进行时间信息聚合，将输入信号连接到全连接层以进行情感或情绪的预测。受 Transformer 结构和 MulT[22]在多模态应用的启发，Zhang 等人[27]提出了一种集成一致性与差异网络 (integrated Consistency and Difference Networks, ICDN)的方法。该模型由多个映射注意模块组成，以每个模态的低级特征为基础进行更深层次的模态融合。首先，ICDN 使用映射转换器(MT)将剩余两种模式的低级特征映射到第三种模式，以弥补该模式缺失部分造成的损失。与 MulT 不同，MT 模块放弃了解码器，并改进了编码器，以使用自我注意技术获得更丰富的模态相关信息。其次，Transformer 被用来提取模态特征，改善模态之间的长期依赖关系和对上下文信息的注意。最后，对 SELF-MM[60]中的自监督方法进行改进，获取单模态情感标签，在多任务学习指导下使多模态特征最终融合。Kumar 等人[26]提出了一种改进多模态情感分析的方法。利用自我注意捕捉长期语境和门控机制选择性地学习交叉参与特征。当单模态信息不足以判断情绪时，门控函数强调交叉交互作用;当单模态信息足以预测情绪时，门控函数对交叉模态信息的权重较低。现有的方法大多将情感与情绪分开研究，没有充分挖掘二者之间的互补知识，基于此，Hu 等人[61]提出了一个多模态情感知识共享框架(UniMSE)，该框架将多模态情感分析(Multimodal Sentiment Analysis，MSA)和对话中的情绪识别(Emotion Recognition in Conversations，ERC)任务从特征、标签和模型中统一起来，在句法和语义层面进行模态融合，并在模态和样本之间引入对比学习，以更好地捕捉情感和情绪之间的差异和一致性。BERT 是一种有效的预训练语言表示模型，以往的研究大多只对文本数据进行了微调，但并没有引入多模态信息来学习更好的表示，Yang 等人[62]提出了跨模态的 BERT(Cross-Modal BERT，CM-BERT)，该模型依赖于文本和音频模态的相互作用来微调预先训练好的 BERT 模型。掩蔽多模态注意作为 CM-BERT 的核心单元，通过结合文本和音频的模态信息动态调整词的权重。由于现有的语音文本预训练方法未能挖掘对话中的语境信息来丰富话语表征，Yu 等人[63]提出了语音文本对话理解的显式跨模态对齐预训练模型(SPECTRA)，图 4 为其结构框图。具体来说，考虑语音模态的时间性，设计了一种新的时间位置预测任务来捕获语音-文本对齐。该预训练任务的目的是预测每个文本单词在相应语音波形中的开始和结束时间。此外，为了解语音对话的特点，将文本对话预训练的响应选择任务推广到语音文本对话预训练场景。

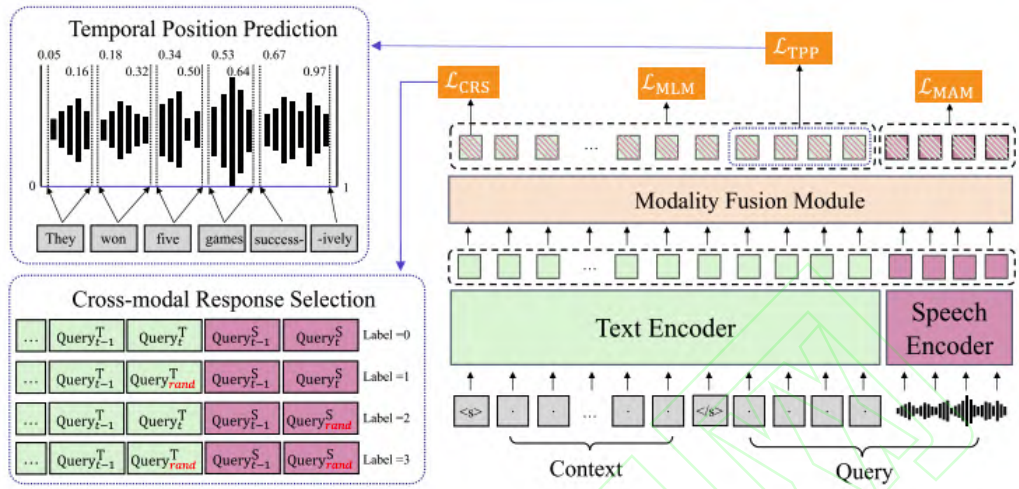

图 4 SPECTRA 整体框架

Fig.4 SPECTRA overall framework

3.2 多模态融合相关算法

由于多模态技术的发展，也相继出现了很多相关算法。在模态融合技术早期，大多数研究者使用基于机器学习的方法训练模态的分类器，比如 SVM，Logistic 回归以及 K-近邻分类器等，还有一些使用基于规则的方法，如平均和多数投票等。由于深度学习的发展，研究者将深度学习的方法引入到模态融合中，近年也提出了一些表现较好的算法。本小节将对当前表现较好的模型算法进行详细的讨论。

受 TFN[20]和 LMF[21]的启发，Barezi 等人[64]提出了一种基于模态的冗余减少融合(MRRF)模型，以理解和调节各模态在多模态推理任务中的相对贡献，模型如图 5 所示。MRRF 使用 Tucker[65][66]张量分解，它允许每种模态有不同的压缩率，使模型能够适应不同模态之间有用信息量的变化。该压缩方法在信息损失最小的情况下降低了模型复杂度和参数数量，可以作为正则化器来避免过拟合。具体来说，该方法首先从输入模态 D 形成一个外积张量，通过张量 W 将其投射到特征向量 H，并将其作为输入传递给神经网络，由于可训练投影张量 W 代表了大量的参数，为了减少这个参数，使用 Tucker 的张量分解。公式如下：

$$W = R \times_1 W_1 \times_2 W_2 \times_3 W_3 \tag{1}$$

$$H=W\ D \tag{2}$$

其中算子 X_k 是张量 R 的 k 模积。

由于话语之间存在相互依赖和联系，Jiang 等人[67]提出了基于注意的双向门控循环单元(AT-BiGRU)模型，整体框架如图 6 所示。该模态考虑了话语之间的语境关系，并采用一种新的框架-信息增强融合 AT-BiGRU (IEFAT-BiGRU)来融合每个话语的多模态特征来进行多模态情感分析，IEFAT-BiGRU 的架构如图 7 所示。具体来说，三种模态的数据通过 CT-BiLSTM 得到了文本、音频和视觉的上下文相关的单模态特征，并且它们都具有相同的维数，随后计算多模态信息增强融合特征 Z，为了考虑周围话语对目标话语的影响，利用 AT-BiGRU 模型来放大与目标话语高度相关的语境信息，随后送入 softmax 进行情感分类。通过实验可以得出，该框架在 CMU-MOSI 和 IEMOCAP 两种不同的基准数据集上分别获得了 82.85%和 65.77%的精度。多模态信息增强融合特征 Z 计算公式如下：

$$Z=n * Z_a +p * Z_v +m * Z_t \tag{3}$$

这里的 Z_a 代表音频特征，Z_v 代表视觉特征，Z_t 代表文本特征，在实验中，Z_a，Z_v，Z_t 的值被设置为[1,10]之间的整数。

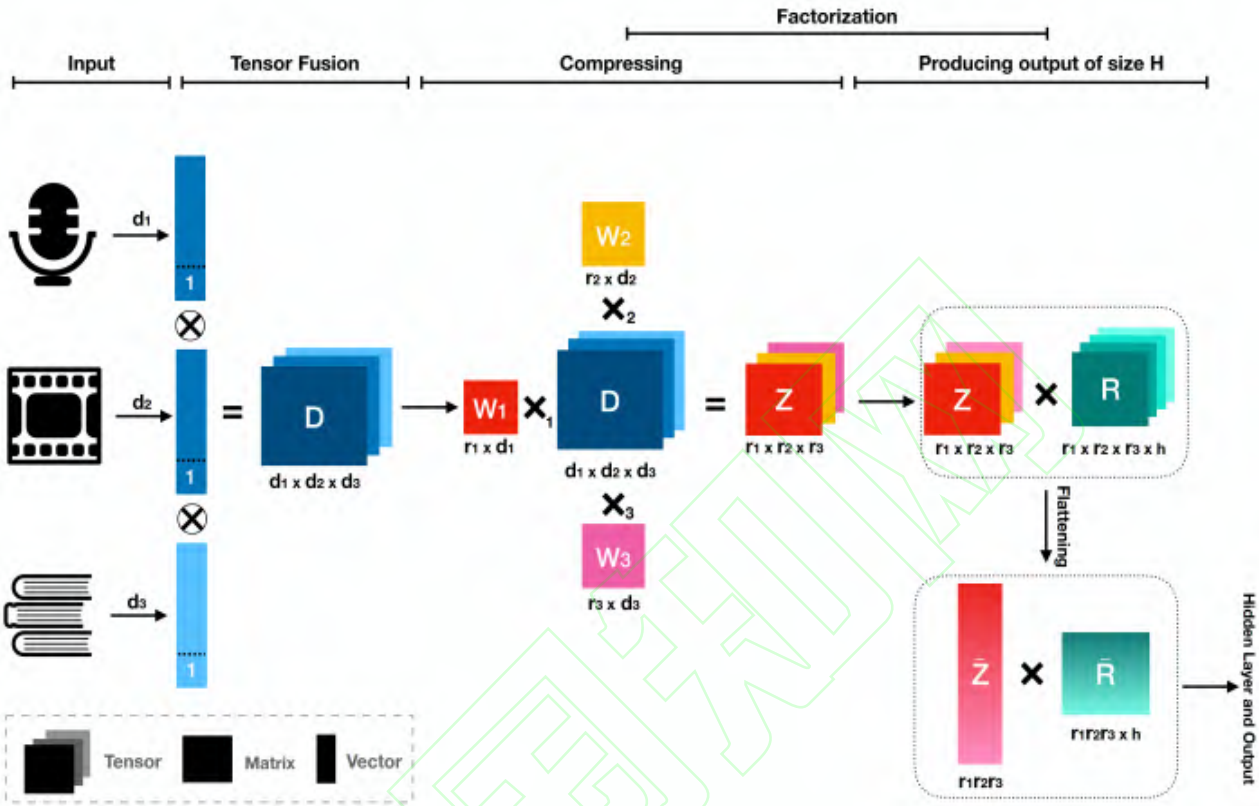

图 5 基于模态的冗余减少多模态融合图

Fig.5 Mode-based redundancy reduction multimodal fusion diagram

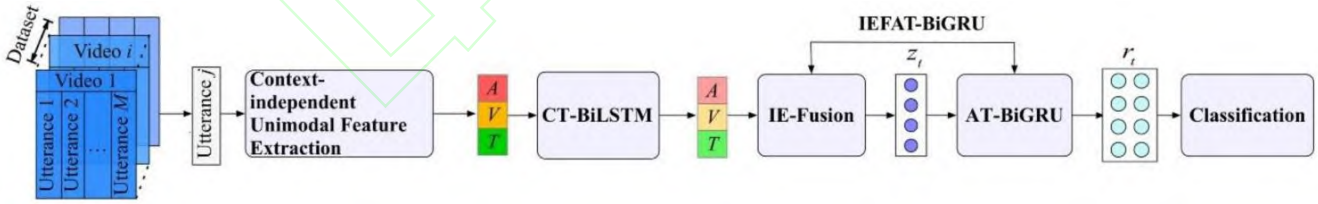

图 6 AT-BiGRU 整体框架

Fig.6 AT-BiGRU overall framework

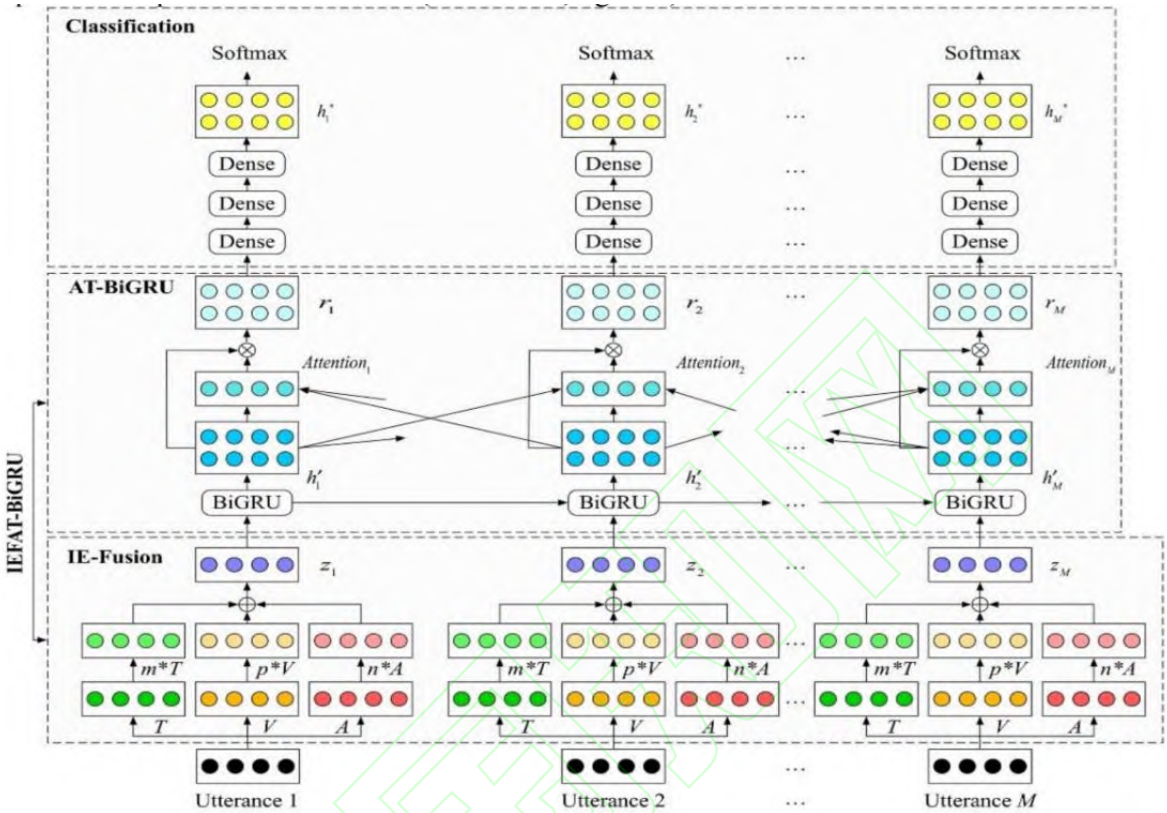

图 7 IEFAT-BiGRU 架构

Fig.7 IEFAT-BiGRU Architecture

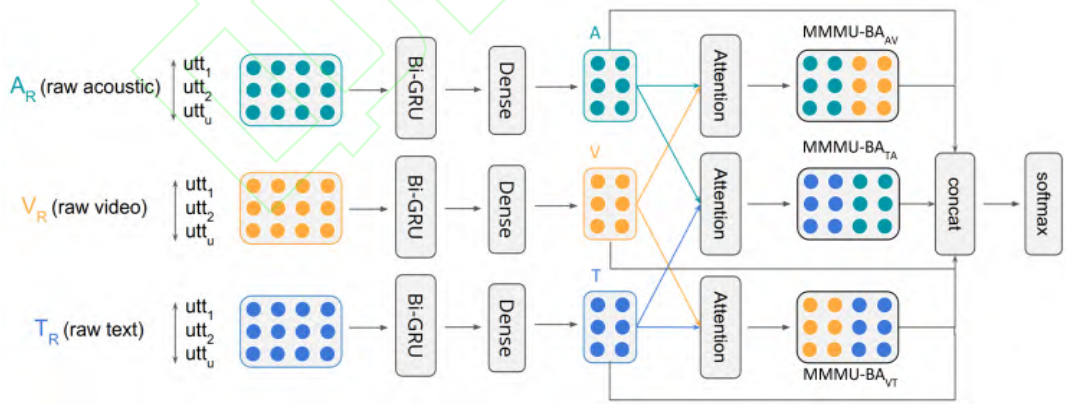

图 8 多模态多话语-双模态注意框架

Fig.8 Multimodal Multidiscourse - Bimodal Attention Framework

Ghosal 等人[24]提出了一种基于循环神经网络的多模态多话语-双模态注意框架（MMMU-BA），图 8 为其结构框图。首先将文本，视觉和音频输入三个独立的双向门控循环单元（Bi-GRU），通过一个全连接层对输出应用多模态注意，接着采用了一种双模态注意框架，其中注意函数被应用到两两形式表征（T+V，T+A，V+A），最后，将两两注意的输出与表示形式连接并传递到 softmax 层进行分类。具体来说，在两两模态的表示计算上，先计算一对匹配矩阵M_1，M_2，接着使用 softmax 函数计算双模态注意矩阵M_1和M_2每个话语的概率分数 $N_1(i,j)$，$N_2(i,j)$，其实是计算了语境话语的注意权重，随后在多模态多话语注意矩阵

上应用软注意力来计算模态上的注意表示 O_1， O_2，接着在每个模态和其他模态的多模态话语特征表征之间计算一个乘法门控函数 A_1， A_2，最后将注意矩阵 A_1 和 A_2 串联起来得到 V 和 T 之间的 $MMMU-BA_{VT}$，以同样的方法计算 $MMMU-BA_{TA}$ 和 $MMMU-BA_{AV}$， $MMMU-BA_{VT}$ 的计算方式如下。

$$M_1 = V \cdot T^T \quad \& \quad M_2 = T.V^T \tag{4}$$

$$N_1(i,j) = \frac{e^{M_1(i,j)}}{\sum_{k=1}^{u} e^{M_1(i,k)}} \quad \text{for} \quad i,j = 1,\ldots,u \tag{5}$$

$$N_2(i,j) = \frac{e^{M_2(i,j)}}{\sum_{k=1}^{u} e^{M_2(i,k)}} \quad \text{for} \quad i,j = 1,..,u \tag{6}$$

$$O_1 = N_1 \cdot T \quad \& \quad O_2 = N_2 \cdot V \tag{7}$$

$$A_1 = O_1 \odot V \quad \& \quad A_2 = O_2 \odot T \tag{8}$$

$$MMMU-BA_{VT} = \text{concat}\left[A_1, A_2\right] \tag{9}$$

3.3 基于小样本场景下的多模态情感分析

近年来，对于某些应用程序来说，收集足够的训练样本通常很昂贵或很困难，许多深度学习模型在使用大量标记数据进行训练时表现良好，但有限的数据可能会降低深度学习模型的能力。小样本学习的核心思想是创建合成任务来模拟只给出少量标记数据的场景，并利用元学习等算法对模型进行训练以避免过拟合。MAML[68]和 Reptile[69]是基于梯度的方法，它们将元学习器设计成一个优化器，可以在给定新例子的几个优化步骤内学习更新模型参数。另一种方法使用度量学习方法优化输入数据的特征嵌入，如 ProtoNet[70]、关系型网络[71]、匹配网络[72]和 DeepEMD[73]。一些研究使用基于图的方法解决小样本问题，对于每个任务，它们将实例设置为节点，并将实例之间的关系设置为边。然后，基于图的方法 GNN、TPN[74]和 DPGN[75]通过递归聚合和转换相邻节点来细化节点表示。MetaOptNet[76]提倡使用线性分类器，它可以优化为凸学习问题，而不是最近邻方法。LEO[77]利用编码器-解码器架构来挖掘潜在的生成表示，并在极低数据状态下预测高维参数。

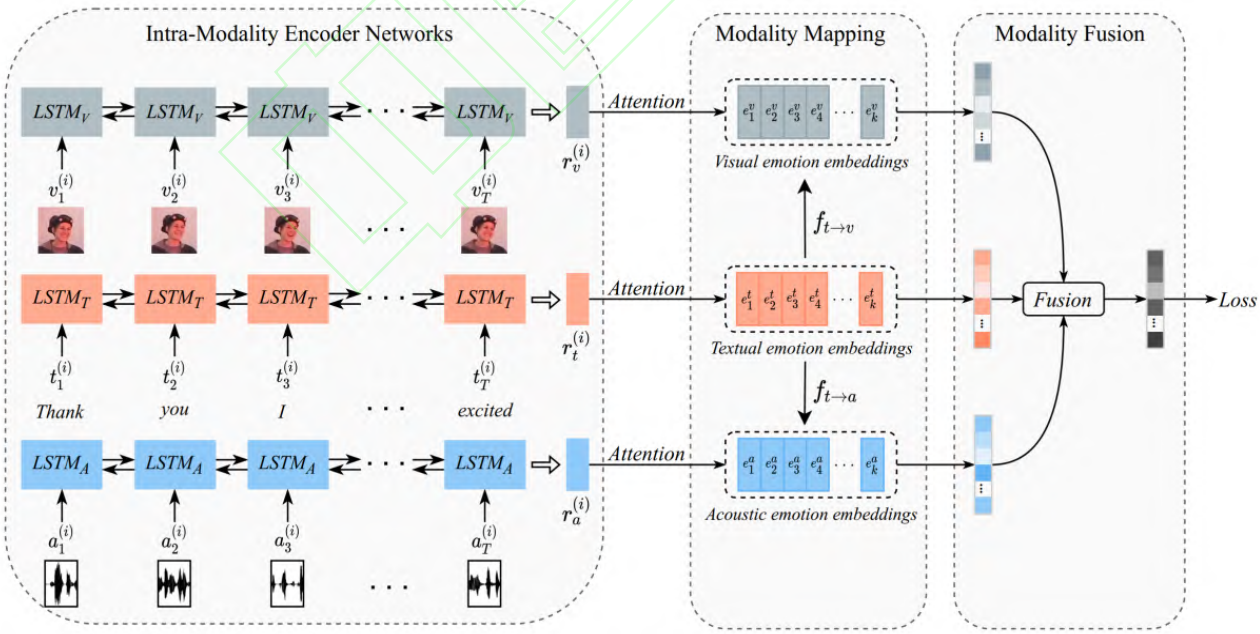

图 9 基于情感嵌入的模态可转移模型

Fig.9 A modal transferable model based on emotional embedding

由于先前的模型不能很好地处理低资源情绪，特别是看不见的情绪，因此，Dai 等人[78]提出了一个带有情感嵌入的模态可转移模型来解决上述问题，该模型如图 9 所示。首先使用预先训练好的词嵌入来表示

文本数据的情感类别。然后，学习两个映射函数，将这些嵌入转换为视觉和听觉空间。对于每个模式，该模型计算输入序列和目标情绪之间的表示距离，并根据距离进行预测。这样模型可以直接适应任何模态中看不到的情绪，因为我们有它们预先训练好的嵌入和模态映射功能。实验表明，在零样本和小样本场景下，该模型优于先前的基线。具体来说，该模型有三部分组成：模态内编码器网络，模态映射模块和多模态融合模块。在模态内编码器网络中，每一种模态都使用双向长短期记忆作为编码器来处理序列并得到向量表示，在模态映射中，使用情感词嵌入来将情感的语义信息注入到模型中，对于文本模态，使用预先训练好的 GloVe 嵌入 K 个情感词，对于另外两种模式，由于没有现成的预训练好的情感嵌入，该模型学习了两个映射函数，将文本空间的向量投射到声学 E_a 和视觉空间 E_v，在模态融合阶段，计算每个模态的序列表示和情绪嵌入之间的相似度得分 $s_t^{(i)}$，$s_a^{(i)}$，$s_v^{(i)}$，并融合模态进行加权求和所有向量。计算公式如下：

$$E_a = f_{t\to a}(E_t) \in \mathbb{R}^{K\times d_a} \tag{10}$$

$$E_v = f_{t\to v}(E_t) \in \mathbb{R}^{K\times d_v}. \tag{11}$$

$$s_t^{(i)} = E_t r_t^{(i)}, s_a^{(i)} = E_a r_a^{(i)}, s_v^{(i)} = E_v r_v^{(i)} \tag{12}$$

$$s^{(i)} = \text{Sigmoid}\left(w_t s_t^{(i)} + w_a s_a^{(i)} + w_v s_v^{(i)}\right) \tag{13}$$

由于文本提示忽略了来自其他形式的信息，Yang 等人[79]提出了多模态概率融合提示用于小样本场景下的图文情感检测，为多模态情感检测提供了多样化的线索，模型的结构如图 10 所示。作者首先设计了一个统一的多模态提示，以减少不同模态提示的差异，为了提高模型的稳健型，在每个输入中利用了多个不同的提示。具体来说，首先分别为不同的模式设计提示，对于文本模态，在 LM-BF[80]的推动下，使用预先训练的 T5 模型[81]，该模型能够为数据集自动生成多个文本模板，接着，对生成的模板进行排序，并选择 top-Nt 模板作为候选文本提示，对于图像模态，为了缓解不同模态之间的差距，通过 ClipCap[82]生成图像的文本描述 C，并将其作为图像提示符，接着利用 NF-ResNet[83]提取原始图像表示并将其投影到文本特征空间中，接着，设计多个多模式提示 $\mathcal{P}_m$，最后将多模态分类任务看作完形填空问题，接着在给出多个提示的情况下进行多模态情感检测，最后，基于概率融合不同 n 个多模态提示中获得标签 l 的融合分布，计算公式如下所示。

$$C = \text{Clip}\,Cap(I) \tag{14}$$

$$V = W_i\,\text{Pool}(\text{ResNet}(I)) + b_i \tag{15}$$

$$\tilde{V} = \text{reshape}(V) = \left[v^1,\ldots,v^j,\ldots,v^{N_i}\right], v^j \in \mathbb{R}^{d_t} \tag{16}$$

$$\mathcal{P}_m = [s]\tilde{V}\ \text{is}\ C[/s]T\text{Itwas[mask].[/s]}, \tag{17}$$

$$p\left(\hat{l}\middle|\left\{\mathcal{P}_m^j\right\}_{j=1}^n\right) \propto \frac{\prod_{j=1}^n p\left(\hat{l}\middle|\,\mathcal{P}_m^j\right)}{p(\hat{l})^{n-1}} \tag{18}$$

其中 I 代表图像，N_i 是超参数，表示多模态提示符中初始图像表示的槽数，d_t 表示文本嵌入预训练语言模型的维数。T 为原始文本序列，“It was”为文本提示符。

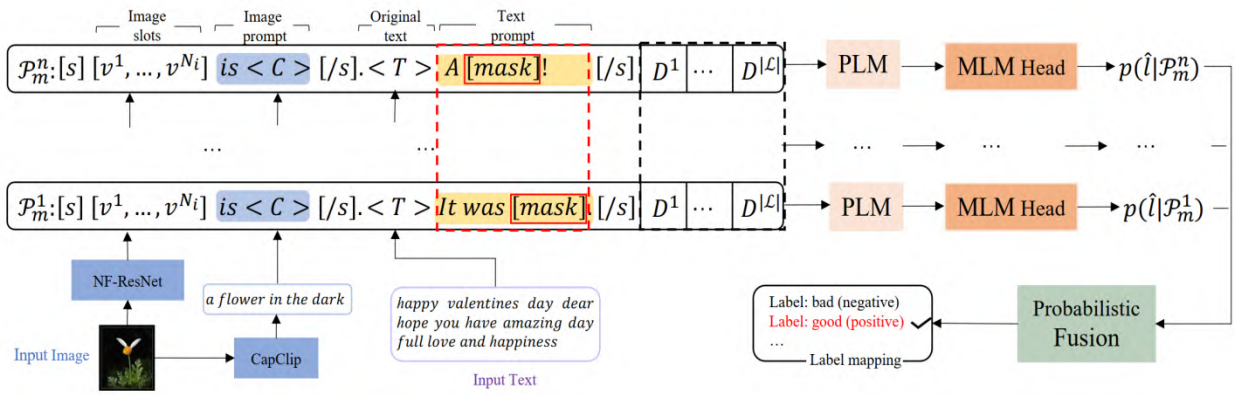

图 10 多模态概率融合提示（MultiPoint）模型

Fig.10 Multimodal Probabilistic Fusion Prompts (MultiPoint) model

在研究多模态情感分析时，小样本场景与数据量充足场景中存在着一些共性及差异性。首先，在共性方面主要体现在特征提取、迁移学习及模型选择和调优上。在小样本场景和数据量充足场景下，都需要设计有效的特征提取方法来捕捉多模态数据中的情感信息，这对于数据量较少的小样本场景变得更为重要；其次，在小样本场景和数据量充足场景下都使用迁移学习作为一种常用策略，其能够帮助提高模型的性能和泛化能力；选择适合的模型架构和算法，并对模型进行合适的调优对小样本场景和数据量充足情况下都起着至关重要的作用。小样本场景与数据量充足场景下同样也存在着一些差异性，结果如表 11 所示。这些共性和差异性揭示了在小样本场景和数据量充足场景下研究多模态情感分析时的关键考虑因素。研究者应根据具体情况选择适当的方法和策略来克服数据量限制，并提高情感分析的性能和泛化能力。

表 11　小样本场景和数据量充足场景下的差异性

Table 11　Differences between Few-shot scenarios and scenarios with sufficient data volume

特点	小样本场景	数据量充足场景
数据量	数据量较少	数据量较大
数据质量	数据质量较低	数据量质量较高
特征提取	根据实际情况可能需要手动设计特征提取方法	可以使用深度学习进行特征提取
模型训练	模型训练通常需要更小的批量大小	可以使用更大的批量大小进行训练
迁移学习	迁移学习更为关键	迁移学习可以用于加速训练
模型评估	评估指标需要适应小样本情况	可以使用常规的评估指标
算法选择	可以选择较简单的算法	可以选择更复杂的算法
泛化性能	泛化性能可能较差	有更好的泛化性能

3.4　不同算法实验分析对比

本小节将前文中对视频信息进行情感分析所提到的算法进行对比研究，对比结果如表 12 所示。以下表中的评价指标都为 Accuracy，表中的模态信息 A，V，T 分别代表 Audio，Video，Text。

通过表 12 可以看出，在进行单模态情感分析时，文本要高于视觉和音频，此外，在进行双模态情感分析时，T+V 和 T+A 相比与 V+A 有更好的表现，说明在进行情感分析时，文本信息是非常重要的线索。基于上下文融合的方法要高于基于张量融合的方法，说明对话语之间的上下文相关性建模可以改善分类，同时，IEFAT-BIGRU 在同种类别中准确率最高，证明了引入 AT-BiGRU 模型可以进一步放大与目标话语高度相关的语境信息。可以看出，基于注意力机制的融合方法在所提出的融合方法中表现出了优异的结果，其中 MMMU-BA 使用双模态组合(即双模态注意框架)的注意计算比单模态组合的自我注意计算更有效。Gated mechanism for attention 使用注意力机制以及门控单元选择性地学习不同模式之间的噪声鲁棒交互和利用视频中出现的长期上下文依赖关系并使用自我注意对提高性能是有帮助的。CM-BERT 和 SPECTRA 都使用了预训练语言模型来作为文本编码器，并结合音频数据进行多模态融合，在 MOSI 数据集上表现出了非常优异的准确率，说明在文本特征提取中，使用预训练语言模型能够提取出更加丰富的情感信息，同时也说明文本和音频形态之间的互动可以提供更加全面的信息，捕捉更多的情感特征。SPECTRA 在以往的文本-音频预训练语言模型的基础上考虑了对话中的语境信息，并在 MOSI 数据集上进一步提高了准确率，说明语境信息对于提高准确率是有帮助的。UniMSE 将多模态情感分析(MSA)和对话中的情绪识别(ERC)任务从特征、标签和模型中统一起来并表现出了较为优异的准确率，说明在多模态情感分析中，捕捉情感和情绪之间的差异和一致性对其性能是有提升的。通过上述分析，使用两种或两种以上模态进行情感分析时能得到最佳的效果，并且在多种模态进行交互时，考虑上下文和语境信息，使用预训练语言模型来作为文本编码器以及利用注意力机制对提升性能是非常重要的。

4　多模态情感分析的应用

通过自动化的情感分析，可以以低成本和高效率的方式获取客户的情感反馈，有助于改善产品的设计和提高用户满意度。图 11 列出了多模态情感分析的各种应用。

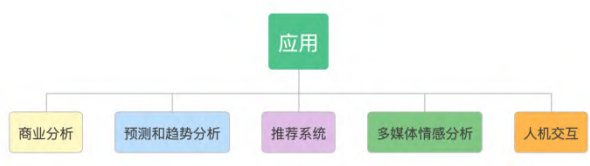

图 11　多模态情感分析应用

Fig.11　Multimodal Sentiment Analysis Applications

表 12 MOSI 数据集上不同算法 Accuracy 比较

Table 12 Comparison of different algorithms Accuracy on MOSI dataset

融合方法	LFM	CHFusion	BC-LSTM	IEFAT-BIGRU	MMMU-BA	MulT	Gated mechanism for attention	CM-BERT	UniMSE	SPECTRA
融合策略	基于张量	基于上下文	基于上下文	基于上下文	基于注意力机制	基于注意力机制	基于注意力机制	基于注意力机制	基于注意力机制	基于注意力机制
T	-	-	78.1	79.65	-	-	-	-	-	-
V	-	-	55.8	61.18	-	-	-	-	-	-
A	-	-	60.3	62.45	-	-	-	-	-	-
T+V	-	79.3	80.2	81.78	81.51	-	-	-	-	-
T+A	-	79.1	79.3	81.25	80.58	-	-	84.5	-	87.5
V+A	-	58.8	62.1	62.23	65.16	-	-	-	-	-
T+V+A	76.4	80.0	80.3	82.85	82.31	83.0	83.91	-	86.9	-

4.1 商业分析

多模态情感分析在商业智能领域有着广泛的应用，尤其在分析顾客对产品或品牌的评价方面最为典型。这些研究不仅为产品的生产者提供了参考意见，还让消费者能够准确评估所购商品的质量，做出更理性的购买决策[84]。比如，企业可以利用多模态情感分析了解客户的需求并对产品不断改进，同时也能制定创新的营销策略[85]。

4.2 预测和趋势分析

情感分析可以被用来追踪民意，并且可以帮助预测市场的一些情景。举例来说，通过分析电影评论，可以预测电影的票房表现。Apala 等人[86]使用 Weka 中的 KMeans 聚类工具，对 Twitter、YouTube 和 IMDB 电影数据库中的电影进行票房预测。

4.3 推荐系统

许多应用程序会根据用户的历史搜索体验来提供相应的推荐服务。例如，淘宝使用推荐系统在主页上向客户推荐相关产品，YouTube 使用自动播放的方式推荐相关视频。Dang 等人[87]提出将情感分析引入推荐系统可以显著提高推荐质量，特别是在数据稀疏的情况下。

4.4 多媒体情感分析

多媒体情感分析是情感分析领域的一个新兴分支。Ellis 等人[88]构建了一个多模态情感分析系统，可以自动分析广播视频新闻并生成电视节目摘要，多模态情感分析技术还可以用于识别具有政治说服力的内容。

4.5 人机交互

人机交互领域通过分析用户的语音、面部表情、手势等多种模态的情感，可以提高人机交互的效率和准确性。例如，在智能客服领域，可以使用多模态情感分析来更好地理解用户的需求和情感状态，从而提供更加个性化和优质的服务。另外，在虚拟现实和增强现实等领域，多模态情感分析也可以提高用户体验和交互效果。

5 多模态情感分析的挑战

在进行多模态情感分析时，需要确定哪个模式的信息更具有分量，并且需要减少异质输入数据之间的噪声。为了达到这个目的，需要设计更好的融合方法和融合模型，以提高多模态情感分析的准确性和效率。除此之外，在多模态情感分析领域还有一些其他的挑战，本节将介绍这些挑战，如图 12 所示。

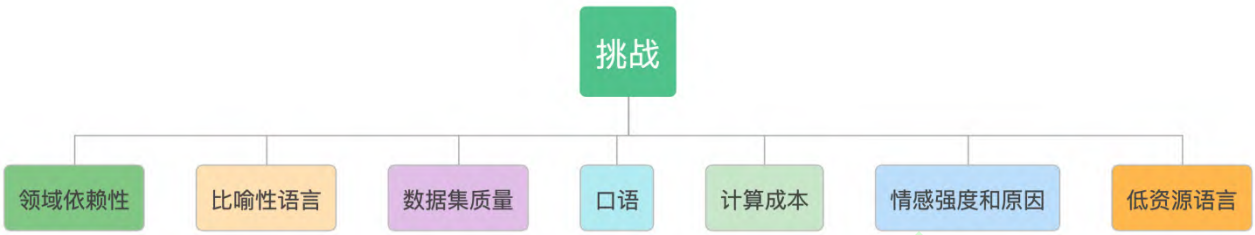

图 12 多模态情感分析挑战

Fig.12 Multimodal Sentiment Analysis Challenges

5.1 领域依赖性

将一个在特定领域训练好的模型迁移到其他不同领域时，通常会出现性能显著下降的情况。例如，采用一个在产品评论中培训过的情感分析模型来分析微博中蕴含的情感。针对这一问题，基于提示学习的方法是一个值得研究的方向。Mao 等人[89]对基于提示学习的情感分析进行了系统的实证研究，以考察 PLMs 对情感计算的偏向程度。

5.2 比喻性语言

在多模态情感分析领域，歧义、反讽和隐喻等比喻性语言的分析一直是一项具有挑战性的任务，例如，表面上赞扬一个产品的评论可能是为了表达消极的情绪，但传统的情感分析方法却认为它们是积极的。人们已经提出了一些方法来检测文本中的反语[90][91]，但是这个问题还没有完全解决。许多因素会影响反讽的理解，如语气、情境、背景信息以及流行语等，而这些幽默因素是机器难以理解的。Poria 等人[92]的研究表明，将声音和面部表情结合到多模态情感分析中，可以提高识别讽刺评论的成功率。

5.3 数据集质量

社交媒体是一个十分丰富的数据仓库，为我们的研究提供了大量的数据。然而，所收集的数据集在质量和领域上各不相同，有些数据仅供互联网上特定人群或特定领域使用。为了提高数据集的质量和可用性，Grosman 等人[93]提出了一种新的基于 web 的文本标注工具 ERAS。该工具不仅实现了主流标注系统的主要功能，而且还集成了随机文档选择、重新标注阶段、预热标注等一系列机制来提高标注过程和标注数据集本身的质量。

5.4 口语

口语是一个具有挑战性的任务，因为口语充满了口音、语速变化等语音信号的差异，并且常常带有语气和音调，甚至包含口头禅、缩略语、方言等非正式的语言形式，例如“我觉得...嗯...是的...好”，这些特点都使得口语情感分析变得更加困难。Zhang 等人[94]提出了一种深度强化学习机制来选择有效的情感相关词，并对每个模态去除无效词。

5.5 计算成本

为了获得更高的精度和更好的结果，需要增加数据集的大小，并设计更高效的模型[84]。一方面，训练具有巨大语料库的模型需要高端 GPU 设备，另一方面，高复杂度的模型很难适应某些特定场景。传统的 SVM 和 NB 模型计算量不大，但其结果并不理想，相反，现在流行的神经网络和注意力模型在计算上很昂贵。Han 等人[95]提出了一种结合分层注意力机制和前馈神经网络来检测抑郁个体的新型编码器，它使用的训练参数比传统编码器更少。

5.6 情感强度和原因

同一个情感可能具有不同的强度水平，举例来说，我们可以观察以下几句话：“我好喜欢今天的天气。”、“我认为这部电影真好看。”和“我喜欢现在的生活。”这三句话都被标记为表达了“快乐”情绪。但在这三种表述中，快乐的强度是不同的，这样的细微差别对于深入理解文本的情感内容非常重要。考虑到情绪产生的原因可以提高对文本或语音中正确情绪的准确性。举个例子：“我太高兴了！因为下雨了！”在这种情况下，系统能够检测到他快乐的原因是下雨。通过关注情绪产生的原因，可以更深入地理解情感表达背后的动机和情境。

5.7 低资源语言

低资源语言指的是语言资源稀缺的语言类型，为了克服资源匮乏的问题，可以采用半监督学习、无监督学习和迁移学习等方法从零开始构建语言资源[96]，这些方法可以利用少量标注数据和大量未标注数据来提高模型性能。

6 总结和未来工作

本文综述了近年来多模态情感分析领域的研究进展，讨论了该领域中最流行的数据集和特征提取方法，并重点分析了基于不同融合方法的框架和特点，对主流的模型算法进行了详细说明以及对当前小样本场景中多模态情感分析所用的方法进行了讨论，并对不同方法的实验结果进行了分析讨论，最后介绍了多模态情感分析的应用和现有方法面临的挑战。该文献综述表明，多模态情感分析利用互补信息渠道进行情感分析，通常优于单模态方法。它还有潜力增强目前受益于单模态情感分析的其他工具，如实体识别和主观性分析。希望这一综述将鼓励这一跨学科领域的进一步发展。

在未来的工作中，一个值得探索的领域是理解对话中的情绪。在谈话中，一个人表达的情绪会影响到其他人。相关研究表明，话语语境有助于理解人类语言，如果多模态系统能够模拟人类的情感依赖，则多模态情感研究将取得重大进展。利用多模态和多语言特征进行情感分析，可以进一步支持跨学科研究，这些研究可以影响从商业到政治、教育到医疗保健领域的技术进步。

参考文献：

[1] ASUR S, HUBERMAN B A. Predicting the future with social media[C]//2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. IEEE, 2010, 1: 492-499.

[2] BOLLEN J, MAO H, ZENG X. Twitter mood predicts the stock market[J]. Journal of computational science, 2011, 2(1): 1-8.

[3] TUMASJAN A, SPRENGER T, SANDNER P, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment[C]//Proceedings of the international AAAI conference on web and social media. 2010, 4(1): 178-185.

[4] D'MELLO S K, KORY J. A review and meta-analysis of multimodal affect detection systems[J]. ACM computing surveys (CSUR), 2015, 47(3): 1-36.

[5] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information fusion, 2017, 37: 98-125.

[6] SOLEYMANI M, GARCIA D, JOU B, et al. A survey of multimodal sentiment analysis[J]. Image and Vision Computing, 2017, 65: 3-14.

[7] ZADEH A, ZELLERS R, PINCUS E, et al. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. arXiv preprint arXiv:1606.06259, 2016.

[8] ZADEH A A B, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2236-2246.

[9] HUDDAR M G, SANNAKKI S S, RAJPUROHIT V S. A survey of computational approaches and challenges in multimodal sentiment analysis[J]. Int. J. Comput. Sci. Eng, 2019, 7(1): 876-883.

[10] GKOUMAS D, LI Q, LIOMA C, ET AL. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis[J]. Information Fusion, 2021, 66: 184-197.

[11] CHANDRASEKARAN G, NGUYEN T N, HEMANTH D J. Multimodal sentimental analysis for social media appli- cations: A comprehensive review[J]. Wiley Interdis- ciplinary Reviews: Data Mining and Knowledge Discovery, 2021, 11(5): e1415.

[12] ZADEH A, CAO Y S, HESSNER S, et al. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. NIH Public Access, 2020, 2020: 1801.

[13] WÖLLMER M, WENINGER F, KNAUP T, et al. Youtube movie reviews: Sentiment analysis in an audio-visual context[J]. IEEE Intelligent Systems, 2013, 28(3): 46-53.

[14] 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.

[15] PÉREZ-ROSAS V, MIHALCEA R, MORENCY L P. Utterance-level multimodal sentiment analysis[C]//Pro ceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 973-982.

[16] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42: 335-359.

[17] MORENCY L P, MIHALCEA R, DOSHI P. Towards multimodal sentiment analysis: Harvesting opinions from the web[C]//Proceedings of the 13th international conference on multimodal interfaces. 2011: 169-176.

[18] PORIA S, CAMBRIA E, GELBUKH A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]// Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 2539-2544.

[19] WANG H, MEGHAWAT A, MORENCY L P, et al. Select-additive learning: Improving generalization in multimodal sentiment analysis[C]//2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017: 949-954.

[20] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.

[21] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv preprint arXiv:1806.00064, 2018.

[22] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]// Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, 2019: 6558.

[23] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context- dependent sentiment analysis in user-generated videos[C]// Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). 2017: 873-883.

[24] GHOSAL D, AKHTAR M S, CHAUHAN D, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C]//proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 3454-3466.

[25] MAJUMDER N, HAZARIKA D, GELBUKH A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-based systems, 2018, 161: 124-133.

[26] KUMAR A, VEPA J. Gated mechanism for attention based multi modal sentiment analysis[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 4477-4481.

[27] ZHANG Q, SHI L, LIU P, et al. ICDN: integrating consistency and difference networks by transformer for multimodal sentiment analysis[J]. Applied Intelligence, 2022: 1-14.

[28] SALTON G , BUCKLEY C . Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5):513-523.

[29] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[30] 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.

[31] PORIA S, CAMBRIA E, HOWARD N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content[J]. Neurocomputing, 2016, 174: 50-59.

[32] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[33] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[34] MUNIKAR M, SHAKYA S, SHRESTHA A. Fine-grained sentiment classification using BERT[C]//2019 Artificial Intelligence for Transforming Business and Society (AITB). IEEE, 2019, 1: 1-5.

[35] ARACI D. Finbert: Financial sentiment analysis with pre-trained language models[J]. arXiv preprint arXiv:1908.10063, 2019.

[36] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Bidirectional LSTM networks for improved phoneme classification and recognition[C]//Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II 15. Springer Berlin Heidelberg, 2005: 799-804.

[37] EYBEN F, WÖLLMER M, GRAVES A, et al. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues[J]. Journal on Multimodal User Interfaces, 2010, 3: 7-19.

[38] EYBEN F, WÖLLMER M, SCHULLER B. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit[C]//2009 3rd international conference on affective computing and intelligent interaction and workshops. IEEE, 2009: 1-6.

[39] 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.

[40] MCFEE B, RAFFEL C, LIANG D, et al. librosa: Audio and music signal analysis in python[C]//Proceedings of the 14th python in science conference. 2015, 8: 18-25.

[41] DEGOTTEX G, KANE J, DRUGMAN T, et al. COVAREP—A collaborative voice analysis repository for speech technologies[C]//2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2014: 960-964.

[42] TAJADURA-JIMÉNEZ A, VÄSTFJÄLL D. Auditory- induced emotion: A neglected channel for communication in human-computer interaction[J]. Affect and emotion in human-computer interaction: From theory to applications, 2008: 63-74.

[43] VOGT T, ANDRÉ E, WAGNER J. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation[J]. Affect and Emotion in Human-Computer Interaction: From Theory to Applications, 2008: 75-91.

[44] EL AYADI M, KAMEL M S, KARRAY F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern recognition, 2011, 44(3): 572-587.

[45] LOWE D G . Distinctive Image Features from Scale- Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[46] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[47] LITTLEWORT G, WHITEHILL J, WU T, et al. The computer expression recognition toolbox (CERT)[C]//2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). IEEE, 2011: 298-305.

[48] BALTRUSAITIS T, ZADEH A, LIM Y C, et al. Openface 2.0: Facial behavior analysis toolkit[C]//2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018: 59-66.

[49] 孙影影,贾振堂,朱昊宇.多模态深度学习综述[J].计算机工程与应用,2020,56(21):1-10.

SUN YINGYING, JIA ZHENTANG, ZHU HAOYU. Survey of multimodal deep learning[J]. Computer Engineering and Applications,2020,56(21):1-10.

[50] PARK S, SHIM H S, CHATTERJEE M, et al. Multimodal analysis and prediction of persuasiveness in online social multimedia[J]. ACM Transactions on Interactive Intelligent Systems (TiiS), 2016, 6(3): 1-25.

[51] PORIA S, CHATURVEDI I, CAMBRIA E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016: 439-448.

[52] NOJAVANASGHARI B, GOPINATH D, KOUSHIK J, et al. Deep multimodal fusion for persuasiveness prediction[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 284-288.

[53] YU Y, LIN H, MENG J, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks[J]. Algorithms, 2016, 9(2): 41.

[54] HUSSAIN M S, CALVO R A, AGHAEI POUR P. Hybrid fusion approach for detecting affects from multichannel physiology[C]//Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4. Springer Berlin Heidelberg, 2011: 568-577.

[55] WANG H, MEGHAWAT A, MORENCY L P, et al. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis[J]. arXiv preprint arXiv:1609.05244, 2016.

[56] KOSSAIFI J, LIPTON Z C, KOLBEINSSON A, et al. Tensor regression networks[J]. The Journal of Machine Learning Research, 2020, 21(1): 4862-4882.

[57] YANG X, YUMER E, ASENTE P, et al. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5315-5324.

[58] LIANG P P, LIU Z, TSAI Y H H, et al. Learning representations from imperfect time series data via tensor rank regularization[J]. arXiv preprint arXiv:1907.01011, 2019.

[59] 胡新荣,陈志恒,刘军平,彭涛,叶鹏,朱强.基于多模态表示学习的情感分析框架[J].计算机科学, 2022, 49(S2): 631-636.

HU XINRONG, CHEN ZHIHENG, LIU JUNPING, PENG TAO, YE PENG, ZHU QIANG. Sentiment Analysis Framework Based on Multimodal Representation Learning[J]. Computer Science,2022,49(S2):631-636.

[60] YU W, XU H, YUAN Z, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 10790-10797.

[61] HU G, LIN T E, ZHAO Y, et al. Unimse: Towards unified multimodal sentiment analysis and emotion recognition[J]. arXiv preprint arXiv:2211.11256, 2022.

[62] YANG K, XU H, GAO K. Cm-bert: Cross-modal bert for text-audio sentiment analysis[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 521-528.

[63] YU T, GAO H, LIN T E, et al. Speech-Text Dialog Pre-training for Spoken Dialog Understanding with Explicit Cross-Modal Alignment[J]. arXiv preprint arXiv:2305. 11579, 2023.

[64] BAREZI E J, FUNG P. Modality-based factorization for multimodal fusion[J]. arXiv preprint arXiv:1811.12624, 2018.

[65] TUCKER L R. Some mathematical notes on three-mode factor analysis[J]. Psychometrika, 1966, 31(3): 279-311.

[66] HITCHCOCK F L. The expression of a tensor or a polyadic as a sum of products[J]. Journal of Mathematics and Physics, 1927, 6(1-4): 164-189.

[67] JIANG D, ZOU D, DENG Z, et al. Contextual multimodal sentiment analysis with information enhancement[J]. Journal of Physics: Conference Series, 2020, 1453(1): 012159 (5pp).

[68] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// International conference on machine learning. PMLR, 2017: 1126-1135.

[69] NICHOL A, ACHIAM J, SCHULMAN J. On first-order meta-learning algorithms[J]. arXiv preprint arXiv:1803.02999, 2018.

[70] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[J]. Advances in neural information processing systems, 2017, 30.

[71] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: Relation network for few-shot learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1199-1208.

[72] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[J]. Advances in neural information processing systems, 2016, 29.

[73] ZHANG C, CAI Y, LIN G, et al. Deepemd: Differentiable earth mover's distance for few-shot learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[74] LIU Y, LEE J, PARK M, et al. Learning to propagate labels: Transductive propagation network for few-shot learning[J]. arXiv preprint arXiv:1805.10002, 2018.

[75] YANG L, LI L, ZHANG Z, et al. Dpgn: Distribution propagation graph network for few-shot learning[C]//Pro- ceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 13390-13399.

[76] LEE K, MAJI S, RAVICHANDRAN A, et al. Meta-learning with differentiable convex optimization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10657-10665.

[77] RUSU A A, RAO D, SYGNOWSKI J, et al. Meta-learning with latent embedding optimization[J]. arXiv preprint arXiv:1807.05960, 2018.

[78] DAI W, LIU Z, YU T, et al. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition[J]. arXiv preprint arXiv:2009.09629, 2020.

[79] YANG X, FENG S, WANG D, et al. Few-shot Multimodal Sentiment Analysis based on Multimodal Probabilistic Fusion Prompts[J]. arXiv preprint arXiv:2211.06607, 2022.

[80] GAO T, FISCH A, CHEN D. Making pre-trained language models better few-shot learners[J]. arXiv preprint arXiv: 2012.15723, 2020.

[81] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[82] MOKADY R, HERTZ A, BERMANO A H. Clipcap: Clip prefix for image captioning[J]. arXiv preprint arXiv:2111. 09734, 2021.

[83] BROCK A, DE S, SMITH S L. Characterizing signal propagation to close the performance gap in unnormalized resnets[J]. arXiv preprint arXiv:2101.08692, 2021.

[84] WANKHADE M, RAO A C S, KULKARNI C. A survey on sentiment analysis methods, applications, and challenges[J]. Artificial Intelligence Review, 2022, 55(7): 5731-5780.

[85] MADHU S. An approach to analyze suicidal tendency in blogs and tweets using Sentiment Analysis[J]. Int. J. Sci. Res. Comput. Sci. Eng, 2018, 6(4): 34-36.

[86] APALA K R , JOSE M , MOTNAM S , et al. Prediction of movies box office performance using social media[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining. IEEE, 2013.

[87] DANG C N, MORENO-GARCÍ A M N, PRIETA F D. An approach to integrating sentiment analysis into recommender systems[J]. Sensors, 2021, 21(16): 5666.

[88] ELLIS J G, JOU B, CHANG S F. Why we watch the news: a dataset for exploring sentiment in broadcast video news[C]//Proceedings of the 16th international conference on multimodal interaction. 2014: 104-111.

[89] MAO R, LIU Q, HE K, et al. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection[J]. IEEE Transactions on Affective Computing, 2022.

[90] CASTRO S , HAZARIKA D , V PÉ REZ-ROSAS, et al.Towards multimodal sarcasm detection (an obviously perfect paper), in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4619-4629.

[91] LIU B, ZHANG L. A survey of opinion mining and sentiment analysis[M]//Mining text data. Springer, Boston, MA, 2012: 415-463.

[92] PORIA S, HUSSAIN A, CAMBRIA E, et al. Combining textual clues with audio-visual information for multimodal sentiment analysis[J]. Multimodal sentiment analysis, 2018: 153-178.

[93] GROSMAN J S , FURTADO P , RODRIGUES A , et al. Eras: Improving the quality control in the annotation process for Natural Language Processing tasks[J]. Informa- tion systems, 2020(Nov.):93.

[94] ZHANG D, LI S, ZHU Q, et al. Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 148-156.

[95] HAN S, MAO R, CAMBRIA E. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings[J]. arXiv preprint arXiv:2209.07494, 2022.

[96] BIRJALI M, KASRI M, BENI-HSSANE A. A com- prehensive survey on sentiment analysis: Approaches, challenges and trends[J]. Knowledge-Based Systems, 2021, 226: 107134.