MSA Review

多模态情感分析：对历史、数据集、多模态融合方法、应用、挑战和未来发展的系统综述

摘要

情感分析（SA）在人工智能（AI）和自然语言处理（NLP）领域取得了很大的发展。自动化分析用户对产品或服务的情感需求日益增长。意见越来越多地以视频形式而不仅仅是文本形式在线分享。这导致了多模态情感分析（MSA）成为一个重要的研究领域。MSA利用机器学习和深度学习在多模态特征提取、融合和情感极性检测等各个阶段的最新进展，旨在降低错误率并提高性能。本综述论文对主要分类和最新发布的多模态融合架构进行了研究。最近的MSA架构发展分为十个类别，分别是早期融合、后期融合、混合融合、模型级融合、张量融合、分层融合、双模态融合、基于注意力的融合、基于量子的融合和单词级融合。文章还比较了几种架构演变在MSA融合类别方面的优势和局限性。最后，提出了一些跨学科应用和未来研究方向。

1. 引言

自Web 2.0的引入以来，人们变得更加热衷于在网络上表达和分享自己对日常活动和全球问题的观点。社交媒体的发展也极大地促进了这些活动，为我们提供了一个透明的平台，与世界各地的人分享我们的观点。这些基于网络的电子口碑（eWOM）评论被广泛应用于商业和服务行业，让顾客发表自己的意见。因此，情感分析已经成为一个新的、令人兴奋的研究领域。情感分析也被称为舆情挖掘和情绪识别，它是情感分析的两种类型。情感分析用于提取和分析公众的情绪和观点。在研究界、学术界、政府和服务行业中，它越来越受到欢迎。识别人类情绪的过程被称为情绪识别。人们对于识别他人情绪的能力差异很大。利用技术来帮助个体识别情绪是一个相对较新的研究领域。情感计算指的是自动识别个体的情绪或情感。

这是一个新兴的研究领域，旨在使智能系统能够感知、推断和理解人类的情绪。这个跨学科领域涵盖了计算机科学、心理学、社会科学和认知科学。尽管情感分析和情绪识别是两个不同的学科，但它们被归为情感计算的范畴之下。情绪和情感在我们的日常生活中起着重要的作用。它们在以人为中心的环境中有助于决策、学习、沟通和情境意识。在过去的二十年中，人工智能研究人员一直在研究如何赋予机器认知能力，使它们能够像人类一样识别、分析和表达情绪和情感。所有这些努力都是情感计算的结果。用户提供的产品、服务和事件评论具有很大的商业价值。它们帮助其他用户做出决策，比如购买新产品，并且对于企业来说，在产品监控、促进更好的客户关系、制定更好的营销策略以及改进和创新服务方面都非常有益。消费者渴望了解在多个网络平台和社交媒体上的言论，基于这些言论，他们做出购买或使用任何产品或服务的决策。

这就是为什么情绪识别和情感分析成为一个不断增长的研究趋势。然而，自动分析大量数据并生成方面的摘要是一项极具挑战性的任务。从自然语言中识别和提取情感是一项困难的努力，它需要对语言的句法和语义规范有深入的理解。此外，意见文本通常是非正式的，包括俚语、讽刺、挖苦、缩写和表情符号，这使得分析变得更加困难。情感分析采用数据挖掘、信息检索和自然语言处理方法，从大量的文本来源中识别和提取意见。而多模态情感分析从人们的行为观察中提取其思想、感受和情绪。行为线索可以是书面写作、面部表情、语言、生理迹象和动作等形式。

情绪与人类密不可分，这就是为什么情绪理解是类似人类的人工智能的关键组成部分。一个人的情绪经常反映在他们的自然语言中。情绪识别在自然语言处理领域因其在情感分析、基于评论的系统、医疗保健和其他领域中的多种应用而备受关注。一组研究人员讨论了在新闻标题中检测情绪的想法。为了解决文本情感识别的挑战，已经建立了许多情感词典。目前，对话式或多模态情感识别在自然语言处理领域日益受到关注，因为它能够从诸如Facebook、YouTube、Reddit、Twitter等平台上的大量公开对话数据中挖掘意见。它还可以在其他行业中使用，比如医疗保健（例如作为心理健康预测工具）、教育（了解学生的沮丧情绪和学生咨询）、刑事犯罪中的欺诈检测等。在对话环境中进行情感识别也是实现情感感知互动的必要条件，需要深入理解用户的情感。为了满足这些需求，对话式情感识别系统必须既有效又可扩展。然而，由于各种研究难题，这是一个具有挑战性的课题。

机器学习现在是一个成熟的领域，涉及到从数据或经验中进行自动学习的任何活动。软件或机器通过接触数据和经验来提升特定任务性能的能力是机器学习的核心。目前，深度学习是机器学习中一个炙手可热的领域。在大数据分析的背景下，深度学习算法所获得的知识大多未被开发利用。深度学习已经在许多大数据领域中得到应用，主要是为了改善分类和建模结果。现代深度学习算法生成的改进的分类建模结果可以应用于各种应用领域[6]。卷积神经网络（CNN）是一种广泛应用于图像处理的深度学习算法。在[6]的一项研究中，对CNN近期改进的全面分析被提出。互联网上的大量数据通常可以是结构化的、半结构化的或非结构化的，并且可能来自各种数据库。结构化数据通常是以定义良好和标准化的格式高度组织的。半结构化数据不遵循传统的表格数据模型，尽管它们以某种预定格式（如电子邮件或XML数据）组织。非结构化数据没有任何预定的数据模型，尽管它们具有内部数据结构，并且通常以文本和非文本格式存在，尤其用于大数据。处理如此庞大的数据量是一项艰巨的任务。由于深度学习通过隐藏层进行复杂处理，加快了自动化过程，因此它正变得越来越受欢迎。利用深度学习可以自动化和加速情感分析过程中的各个耗时阶段，如特征选择、特征提取、学习参数、处理特征向量和生成预测。

在基于深度学习的方法中，隐藏层用于在输入层和输出层之间执行复杂处理，并充当一个黑盒子，因此中间隐藏层中的数据表示通常保持未知。在本次调查中，我们着重研究了融合多个模态（如视觉、听觉和文本特征）的不同融合技术，用于多模态情感分析。总结了最常用和最新的多模态情感分析数据集。介绍了流行数据集及其创建阶段。讨论了当前研究中最新的融合技术。还研究了基于最新的机器学习和深度学习创新的各种多模态分类算法。描述了情感分析在不同应用领域的使用以及在各种应用中的未来发展前景。还讨论了利用视频、心理信号、脑电图信号和其他信息的异构数据的利用，以及多语言数据、跨语言数据、跨领域数据和混合编码数据格式。

本文的其余部分安排如下。第2节解释了多模态情感分析的基本原理和需求。第3节总结了一些最重要的多模态情感分析数据集。第4节介绍了利用最新创新的不同融合架构的全面调查。第5节定义了多模态情感分析在各个领域的应用。第6节描述了每种融合架构的限制以及模型的优缺点。第7节讨论了多模态情感分析研究的未来方向，最后在第8节给出了结论和总结。

2. 多模态情感分析基础

在传统的情感分析系统中，只使用一种模态来确定用户对主题的积极或消极观点。多模态情感分析是传统基于文本的情感分析的一个子集，它包括文本以外的其他模态，如语音和视觉特征。使用视觉特征是因为视觉呈现可以比长篇的文字或口语更有效地解释或描述某事物，并且可以被广泛应用于准确预测相关数据的情感。对于多模态情感分析，可以使用各种双模态组合，例如语音+图像、图像+文本、语音+文本、语音+脑电图信号。使用两种模态的系统称为双模态情感分析系统。使用所有三种模态的系统称为三模态情感分析系统。

不同的模态在各种形式和大小上都是可用的，例如音频或口述的词语（例如笑声、哭声、语调），时间图像（例如微笑、凝视、面部表情）和文本数据（例如从口述词语转录的数据）或者以脑电图信号的形式存在。这些是多模态情感分析中最常研究的模态。为了进行最佳分类，只使用下面列出的特征子集。文本由积极或消极的词语组成，被称为极性词，明确定义了极性，例如词语“惊人”显示积极极性，而词语“糟糕”表示消极极性。文本被分为单词组或短语、字符N-Gram、音素N-Gram和其他文本元素。听觉特征包括笑声、停顿、语调、哭声以及句子中声音音调分布、话语速度等。微笑、皱眉、手势、姿势、凝视和其他视觉元素（如眼神接触）都是面部表情的示例[7]。

在文本指标中，单词被用于情感预测。但是，一个句子中只有少数单词对于了解情感是重要的，其余的单词要么是停用词（用于语法上构建句子），要么是与情感相关的字符n-gram，即一组字符在情感方面具有可比性。还考虑了音素n-gram，它类似于字符n-gram，但添加了音素的信息。

在听觉指标中，音频数据非常重要，用于生成转录内容。一些重要的听觉特征包括停顿。例如，话语中较多的停顿表示中性情感。音调也用于揭示主观性。高音调传达焦虑或热情，而低音调传达严肃。另一个重要的听觉特征是讲话时的强度或能量。高能量话语的极性通常通过对单个单词或短语的关注来表现出来。

在视觉指标中，视觉情感与文本数据一起用于分析情感。面部表情在情感识别中起着关键作用。微笑是明显的特征，表示积极情感。微笑是一种积极的情绪。由于最近的技术进展，摄像头现在可以检测微笑，甚至可以检测微笑的强度并给出一个评分。同样，凝视和眼神接触表示积极情感，而如果一个人目光离开则表示消极或中性情感。

因此，演讲者的面部朝向可以用来确定眼神接触或凝视。

图1展示了使用来自异构数据源的不同音频-视觉特征进行多模态融合的情感分析过程的不同阶段。首先，从互联网或网络中获取结构化、半结构化或非结构化的各种数据作为输入，然后对数据进行预处理。在预处理阶段，根据问题要求对数据进行清洗和选择，使用降维技术进行处理。然后，使用各种特征提取算法提取特征。音频特征从视频中的口述词汇中提取，而时间图像则从视频中提取。将文本数据转录为语音生成文本转录。接下来，从提取的特征中生成多模态特征向量，并使用各种分类算法对数据进行分类。然后，将分类结果发送到特定的应用程序中[7]。

2.1. 模态的重要性

在多模态情感分析中，使用不同的模态来从对话中找出情感状态。最常用的模态是文本、音频和视觉模态。每种模态都对情感的更好预测做出了贡献，文献表明与单模态系统相比，双模态和三模态系统可以改善结果。每种模态都对提高准确性有一定的重要贡献。

文本模态：文本模态是所有模态中占主导地位的模态。它在识别隐藏情感方面起着关键作用。文本情感分析取得了非常好的结果，但现在大多数具有观点的数据都是以视频形式分享，而不是文本形式。

视觉模态：视觉特征有助于更好地识别潜在的情感或意见。例如，如果有一个文本“这是一只很好的老鼠”。仅使用文本数据很难确定这是指真实的动物老鼠还是计算机鼠标。视觉在这种情况下发挥了作用，并且文本和视觉的组合形成了双模态系统，与单模态系统相比，能够产生更好的结果。

音频模态：声学特征用于从视频中生成文本数据，同时可以识别说话者的语调。所有三种模态的组合生成了更好的分析模型。在幽默、讽刺和常识检测的情况下，视觉可能会产生错误的预测，但模态的组合可以正确识别情感。

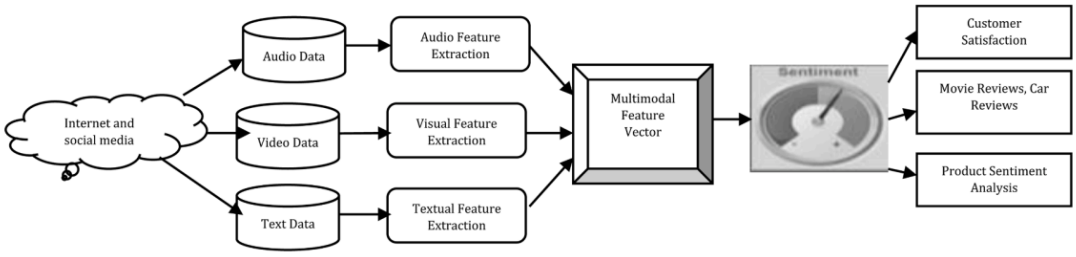

Fig 1. MSA process model [7].

3. 热门数据集

创建用于多模态情感分析的数据集涉及以下阶段。

数据获取：数据获取由两个单词组成：数据和获取。数据指的是可以结构化或非结构化的原始事实和数字。获取指的是为了实现特定目标而收集数据。术语"数据获取"指的是在数据存储、清理、预处理和其他方法中使用之前，从相关来源收集数据的行为。为此，从多个互联网视频共享平台收集视频以创建多模态数据集。使用指定的搜索短语从网络中收集视频，可以使用自动或半自动工具。通过面部检测来分析网络视频中是否存在一个人物在画面中进行独白，以确保视频是单人讲述。通常选择那些演讲者专注于摄像头的视频。根据经常搜索的主题收集视频，并限制每个频道的视频数量以增加多样性。

数据预处理：数据可能存在缺失数据、不正确或虚假值，并且可能不包含相关的特定属性。数据预处理是提高数据质量的必要步骤。MSA数据集的生成预计涵盖从20岁中期到50岁末的广泛年龄范围。数据集中的大多数说话者是英国或美国的英语本土人。数据集中还包括一小部分非母语但流利的英语说话者。一些说话者戴着眼镜。很少有带有方言或口音的说话者的视频。

数据后处理：众所周知，从口语中获取的语言信息有助于获取情感，并且是上下文解释的关键组成部分。音频数据会自动转录，以便促进听觉、视觉和文本模态之间的交互。谷歌云语音API和亚马逊转录的视频转录在这个任务中具有适当的质量。其中包括非语言线索和听觉方面，如笑声、音乐和主题。口语转录包括标点符号（例如句号、问号、感叹号），每个转录的单词都有开始和结束时间戳以及持续时间。这些元数据有助于将文本与注释（样本率差异）和其他模态对齐。

数据注释：数据注释是对图像、视频帧、音频和文本数据进行标注的过程，主要用于监督式机器学习和半监督式机器学习中的数据集训练。它帮助机器理解输入并相应地采取行动。对于每个剪辑，每个注释者将其情感状态判定为-1（消极）、0（中性）或1（积极）。使用独立的人力资源进行注释。然后，为了进行回归和多分类任务，使用平均标记结果。

表1总结了用于多模态情感分析的数据集生成中最常用和最新的进展。第一列显示数据集的名称，第二列显示数据集发布的年份。第三列显示引入数据集的研究参考文献，而第四列显示每个数据集涵盖的各种模态。第五列显示数据集中使用的评论视频数量，而第六列显示收集视频的来源平台。第七列显示数据集中男性和女性说话者的数量，接着第八列显示视频中使用的语言，而第九列列出了视频涵盖的许多主题，例如电影评论和产品评论。第十列显示了一个链接，该链接指向一个公共领域可用的数据集。

在表1中描述的较新数据集出现之前，以下两个数据集在研究目的中被广泛使用。第一个是YouTube观点数据集，由[8]生成。它是一个用于多模态分析的情感数据集。它包含了来自YouTube的47个视频，这些视频不与任何特定主题相关。其中有27位男性演讲者和20位女性演讲者。该数据集包括手动转录的文本以及自动提取的音频和视觉元素。它还包含了自动提取的话语。第二个是由[9]创建的西班牙多模态观点数据集，它是一个用于西班牙语的多模态情感分析数据集。它包括105个视频，已经对话语级别的情感极性进行了注释。长时间间隔被用于自动提取话语，大多数影片有6-8个话语。总共有550个话语在数据集中。在任何建议的数据集中都没有情感强度的注释。它们主要关注极性。此外，正如介绍中所指出的，它们更倾向于关注视频或话语分析，而不是细粒度的情感分析。

Table 1
Summary of popular datasets for multimodal sentiment analysis.

Name	Year	Reference	Modalities	No. of videos	Source	No. of speakers	Language	Topics Covered	Available At
MOSI	2016	[10]	$A+V+T$	93	YouTube	89 41-female 48-male		General Indexed by #vlog	*https://www.amir-zadeh.com/datasets*
CMU-MOSEI	2018	[11]	$A+V+T$	3228	YouTube	1000	English	250 Reviews Debate Consulting	*https://www.amir-zadeh.com/datasets*
MELD	2019	[90]	$A+V+T$	–	TV Series -Friends	Multi Speaker	English	Dialogues from TV series-Friends	*https://affective-meld.github.io.*
Memotion Analysis	2020	[12]	$V+T$	–	Reddit, Facebook, etc.	–	English	52 Hillary Trump Minion	*https://github.com/terenceylchow124/Meme-MultiModal/tree/main/data/memotion*
CH-SIMS	2020	[13]	$A+V+T$	2281			Chinese	General	*https://github.com/thuiar/MMSA.*
CMU-MOSEAS	2021	[14]	$A+V+T$	4000	YouTube	1645	Spanish Portuguese German French	General 250 topics	–
MuSe-CaR	2021	[15]	$A+V+T$	291	YouTube	70	English	Vehicle Review	*https://www.muse-challenge.org/challenge/data*
B-T4SA	2021	[16]	$V+T$	470k tweets	Tweeter	–	Other than English	General	*https://www.t4sa.it/*
FACTIFY	2022	[17]	$V+T$	50,000 tweets	Tweeter	–	English	20 Politics Governance	–
MEMOTION 2	2022	[18]	$V+T$	10,000 images	Reddit, Facebok, etc.	–	English	Politics Religion Sports	–

3.1. MOSI数据集

MOSI由[10]创建。MOSI代表多模态意见级别情感强度（Multimodal Opinion Level Sentiment Intensity）。它包括多模态观察、从口语词语和视觉手势生成的转录，以及自动生成的音频和视觉特征。它还包括在意见级别上的主观性分割、具有高一致性的情感强度注释以及词语、视觉和声学特征之间的对齐。它由93个YouTube视频组成，以#vlog作为标签进行分组。其中有89个不同的说话者，其中41位是女性，48位是男性。它是第一个在意见级别上包含主观性和情感强度注释的多模态情感分析数据集。

3.2. CMU-MOSEI数据集

CMU多模态意见情感和情绪强度（CMU-MOSEI）是在线视频中句子级情感分析和情绪识别的最大数据集[11]。在CMU-MOSEI中，您可以找到来自1000多位演讲者和250个主题的超过65小时的注释视频。每个视频片段都包括与音频同步的音素级手动转录。这些视频是从YouTube上的一个在线视频分享服务中获取的。根据统计数据，您可以从任何YouTube频道获取的视频数量限制为十个。根据统计数据，它

总共有来自3228个视频的23,453个句子。该数据集涵盖了各种各样的主题，但前三个主题分别是针对不同产品和服务的评论（16.2%）、关于各种主题的辩论（2.9%）和咨询（2.9%）。其余主题在数据集中几乎均匀地分布。

3.3. MELD数据集

这是多模态情感对话数据集（Multimodal EmotionLines Dataset），它是EmotionLines数据集[90]的扩展。它是一个多模态多方对话情感识别数据集，是情感识别中最广泛使用的数据集。MELD包括与EmotionLines相同的对话实例，但除了文本外，它还包括音频和视觉元素。MELD包含了大约1400个对话和13000个话语，来自电视剧《老友记》。它是一个多方对话的数据集，包括多个说话者。每个话语都用情感和情绪标签进行了注释，并涵盖了音频、视觉和文本模态。对话中的每个音节都被分配了这七种情绪之一：愤怒、厌恶、悲伤、喜悦、中性、惊讶和恐惧。MELD中的每个语音还有情绪注释（积极、消极或中性）。它的主要目的是训练用于情感识别的对话上下文建模。

3.4. Memotion分析数据集

这是一个用于在线meme情感分析的多模态数据集。meme通常具有幽默性，并试图与受众产生共鸣。其中许多meme希望通过在生活的不同阶段表达团结来与受众建立联系。有些meme纯粹是有趣的，而其他一些则对当前事件进行批判性的讽刺。为了收集meme数据，总共评估了52个不同的世界流行类别，如希拉里、特朗普、小黄人、宝宝教父等。使用谷歌的图像搜索服务获取meme图片。使用亚马逊的众包平台Mechanical Turk（AMT）工作者对情感类别进行注释，标记为幽默、讽刺、冒犯、动力，并量化了特定类别的效果传达强度。数据集中的整体情感跨越了四个类别（非常负面、负面、中性、正面、非常正面），收集了约10,000个样本[12]。

3.5. CH-SIMS数据集

CH-SIMS是一个中文单模态和多模态情感分析数据集，具有独立的单模态注释，包含了2281个来自真实环境的精细视频片段。它同时具有多模态和独立的单模态注释。它使研究人员能够研究模态之间的关系，或者使用独立的单模态注释进行单模态情感分析。它包含来自60个原始视频的2281个视频片段。SIMS拥有多样的角色阵容，涵盖了广泛的年龄范围，并具有出色的制作价值。[13]。

3.6. CMU-MOSEAS数据集

CMU-MOSEAS（CMU多模态意见情感、情绪和属性）数据集包括四种语言：西班牙语（全球总计超过5亿名使用者）、葡萄牙语（全球总计超过2亿名使用者）、德语（全球总计超过2亿名使用者）和法语（全球总计超过2亿名使用者）。这些语言起源于罗马或德国，来自欧洲，我们也在这里获取了大部分的视频。这些语言在非洲和加勒比地区以及北美和南美大陆的部分地区也有使用，并伴有不同的方言。然而，欧洲方言在不同的地区大多是可以理解的，虽然也有一些例外。CMUMOSEAS数据集包含了40,000个样本，涵盖了1645名演讲者和250个主题，在这四种核心语言（法语、德语、葡萄牙语和西班牙语）中是最大的数据集。CMUMOSEAS中有20个注释标签，包括情感和主观性、情绪和个性特征。随着多模态学习的进展，数据集和相应的描述符将公开提供，并定期添加新的特征描述符[14]。

3.7. MuSe-CaR数据集

MuSe-CaR数据库是一个大型的多模态（视频、音频和文本）数据集，旨在通过野外获取的数据来学习更多。多模态情感分析涉及到在产品评论（例如车辆评论）中发生的情感参与，其中情感与主题或实体相关联。专业人士、半职业人士（影响者）和业余评论者的年龄大致在20多岁到50多岁之间。其中大部分是来自英国或美国的母语英语使用者，也有少数非母语但流利的英语使用者。MuSe-CaR数据集是在野外获取的，旨在创建适当的方法并进一步了解多模态情感分析。MuSe-CaR的建设考虑了各种计算任务，包括情绪和实体识别，主要是为了提高机器对情感（即情绪）与实体及其评论组成部分之间关联的理解能力。其中包含了MuSe-CaR数据集的高质量子集，供MuSe 2020挑战赛使用，该子集包括来自291部电影和70名主持人（以及额外的20名旁白）的36小时52分钟08秒的视频数据，这些数据来自YouTube。MuSe-CaR视频的主题仅限于车辆评论。车辆制造商的数量有限，仅限于配备最新技术的高端品牌（宝马、奥迪、奔驰），以确保讨论的实体和方面（例如半自动驾驶功能）在各种视频（以及不同的制造商）中出现。大部分评价是由半专业或专业的评论者（例如YouTube的"影响者"）撰写的。在MuSeCaR中使用的所有YouTube频道都明确同意将其数据用于学术研究[15]。

3.8. B-T4SA数据集

B-T4SA是T4SA数据集的子集，包含47万个样本。每个样本都包含文本和图像数据。训练集约占数据集的80%，而验证集和测试集各占10%。B-T4SA旨在解决T4SA中重复条目、微小短语、格式错误的图形和不平衡的类别等问题[16]。

3.9. FACTIFY数据集

FACTIFY是一个多模态事实核查数据集[17]。它是世界上最大的多模态事实核查公共数据集。它涵盖了来自印度和美国的新闻的5万个数据点。FACTIFY是一组照片、文本断言、参考文本来源和图像的集合。它被分为支持、无证据和反驳三个类别。它是基于可访问性、受欢迎程度和每日发布帖子的，按日期从印度和美国的新闻Twitter账号中收集得到，其中印度的账号有Hindustan Times 1、ANI2，美国的账号有ABC3、CNN4。此外，这些Twitter账号以其公正和无偏见的立场而闻名。从每个推文中检索到推文文本和图像。该数据集总共有5万个样本，每个类别中的样本数量相等。数据集的训练-验证-测试划分比例为70:15:15。数据集中的大部分主张与政治和政府有关。根据最常见的20个实体，美国和印度的政党和领导人都在主张中被提及。

3.10. MEMOTION 2数据集

这是第一个用于迷因分类的大规模多模态数据集[18]。人类表达了广泛的情绪，包括愤怒、憎恨、悲伤、宁静、恐惧等，强度不同。Memotion 2.0是一个数据集，专注于将情绪及其强度划分为离散的标签。它还包括与迷因情感相对应的标签。Memotion 2.0在之前的版本（Memotion 1.0）的基础上进行了扩展，包括了从各种社交媒体网站收集的新的1万个迷因。它是从多个来源收集而来的。在缩小了许多感兴趣的主题（如政治、宗教和体育）之后，迷因被手动下载。数据集包括10,000张图片，按照训练-验证-测试的比例划分为8500张、1500张和1500张。每个迷因都注释了整体情感（积极、中性、消极）、情绪（幽默、讽刺、冒犯、动力）和情绪强度的尺度（0-4级）。

Table 2
Performance summary of various multimodal fusion variants with its architectural categories.

Ref.	Model Name	Year	Fusion Type	Method	Dataset	Main Objective	Accuracy/ F1 Score
[19]	Tri-modal HMM	2011	Early Fusion	HMM	YouTube	Introduces trimodal sentiment analysis Five multimodal features were identified: polarised words, smiles, gazes, pauses, and voice tone. A new YouTube dataset has been introduced.	55.3%
[20]	Text-audio-Visual	2013	Early Fusion	SVM	Spanish Multimodal Opinion Dataset	The usage of multiple modalities together boosts performance. A new Spanish Multimodal Opinion Dataset has been introduced. Check the portability of model on English dataset	75%
[21]	Proposed Multi Method	2015	Early Fusion	SVM	eNTERFACE	Novel multimodal information extraction agent	87.95%
[22]	Multimodal Model	2016	Early Fusion	SVM	POM	Persuasiveness prediction in Online Social Multimedia	70.85%
[23]	MARN	2018	Early Fusion	LSTHM	ICT-MMMO	Novel architecture to understand Human Communication Comprehension	86.3%
[24]	Audio-Visual	2013	Late Fusion	HMM, CERT	AVEC	Classifier Fusion using Kalman Filter	68.5%
[25]	Ref	2014	Late Fusion	SMO (Sequential Minimal Optimization)	YouTube	Measuring personality trait using behavioural signal processing Automatic recognition of personality trait using Vlogs Model's performance also compared with the emotional feature set which is poor	62.6%
[26]	Multi CNN	2015	Late Fusion	CNN	Real World Twitter Dataset TD1 and TD2	Sentiment Analysis of Multimodal tweets using CNN	79%
[27]	ICT-MMMO	2013	Hybrid Fusion	BLSTM, SVM	ICT-MMMO	Multimodal sentiment analysis in online review videos Hybrid Fusion by combining Early and Late Fusion	71.3%
[28]	Bimodal with Unimodal	2015	Hybrid Fusion	Deep CNN, MKL		Parallelizable decision-level data fusion method Multiple kernel learning for training classifier	86.27%
[29]	Three modalities	2016	Hybrid Fusion	ELM	YouTube	Explanation of feature extraction as well as model building Comparison using various machine learning methods	80%
[30]	HMM-BLSTM	2012	Model-Level Fusion/ Utterance-Level	HMM, BLSTM	IEMOCAP	At the utterance level flow of affective expressions Context-sensitive approaches for emotion recognition within a multimodal, hierarchical approach indicate potentially relevant patterns	72.35
[31]	Context-aware BLSTM	2013	Model-Level Fusion	BLSTM	SEMAINE	Context-sensitive LSTM-based audio-visual emotion recognition	65.2%
[32]	TFN	2017	Tensor Fusion	TFN	CMU-MOSI	End-to-end learning of intra-modality and inter-modality dynamics	77.1%
[33]	MRRF	2019	Tensor Fusion	MRFF	CMU-MOSI	Modality-based Redundancy Reduction Fusion (MRRF) modality-based tensor factorization Modality-based Redundancy Reduction Fusion in MRRF is a tensor fusion and factorization method that allows for modality-specific compression rates while also reducing parameter complexity.	77.46%
[34]	T2FN	2019	Tensor Fusion	T2FN	CMU-MOSI	Regularization method based on tensor rank minimization Random drop settings	–
[35]	MTFN—CMM	2021	Tensor Fusion	MTFN—CMM	CMU-MOSI CMU-MOSEI	Emotional fusion in multimodal data is effective Cross-modal modelling with a multi-tensor fusion network	80.9%
[36]	CHFusion	2018	Hierarchical Feature Fusion	CNN	CMU-MOSI	Hierarchical Fusion, in which the two modalities are fused first, followed by the three modalities. Context-Aware Hierarchical Fusion (CHFusion)	80%
[37]	HFNN	2019	Hierarchical Feature Fusion	BiLSTM	CMU-MOSI	'Divide, conquer, and combine' is a strategy used for multimodal fusion. For a thorough interpretation of multimodal embeddings, fusion have been done hierarchically so that both local and global interactions are taken into account. Global fusion is used to obtain an overall view of multimodal embeddings via a specifically designed ABS-LSTM. Two levels of attention mechanism are used: Regional Interdependence Attention and Global Interaction Attention	80.19%
[38]	BBFN	2021	Bimodal Fusion		CMU-MOSEI	An innovative end-to-end Bimodal Fusion network that conducts fusion (relevance increment) and separation (difference increment) on pairs modality representations is introduced. Modality-specific feature space separator and gated	86.2%

(*continued on next page*)

Table 2 (*continued*)

Ref.	Model Name	Year	Fusion Type	Method	Dataset	Main Objective	Accuracy/ F1 Score
						control mechanism is used	
[39]	BIMHA	2022	Bimodal Fusion		CH-SIMS	Bimodal Information-augmented Multi-Head Attention Inter-modal interaction and inter-bimodal interaction	82.71%
[40]	CATF-LSTM	2017	Attention Based Fusion	CATF-LSTM	CMU-MOSI	Attention-based fusion mechanism, termed AT-Fusion	81.30%
[41]	MMHA	2020	Attention Based Fusion	MMHA	MOUD, CMU-MOSI	Learn how unimodal features are related, and capture the internal structure of unimodal features Learn the connections between the various modalities, and focus on the contributing elements.	82.71%
[42]	Bi-LSTM with attention model	2021	Attention Based Fusion	Bi-LSTM	CMU-MOSI	Before fusion, a unique attention-based multimodal contextual fusion technique extracts contextual information from the utterances	80.18%
[43]	QMR	2018	Quantum Based Fusion		Flicker GI	Quantum-inspired Multimodal Sentiment Analysis fill the 'semantic gap' and model the correlations between different modalities via density matrix Quantum Interference inspired Multimodal decision Fusion (QIMF)	88.24%
[44]	QMN	2020	Quantum Based Fusion	QT+LSTM	MELD	Quantum-like multimodal network (QMN), which combines quantum theory (QT) mathematical framework with a long short-term memory (LSTM) network Dynamics of intra- and inter-utterance interaction modelling choice correlations between multiple modalities using a quantum interference-inspired decision fusion approach. To create better predictions about social impact amongst speakers, a quantum measurement-inspired strong-weak influence model was developed.	75.60%
[45]	QMF	2020	Quantum Based Fusion		CMU-MOSEI	The word interaction within a single modality and the interaction across modalities are formulated with superposition and entanglement respectively at different stages. Quantum-theoretic multimodal fusion framework	79.74%
[46]	MFN	2018	Word-level Fusion	MFN	MOUD	MFN shows a consistent trend for both classification and regression	80.4%
[47]	DFG	2018	Word-level Fusion	MFN+DFG	CMU-MOSEI	Multi-view sequential learning that consists of three main components System of LSTMs consists of multiple Longshort Term Memory (LSTM) networks, Delta-memory Attention Network	85.2%
[48]	RMFN	2018	Word-level Fusion	LSTHM	CMU-MOSI	Gated Modality mixing Network Multimodal Shifting	85.4%
[49]	RAVEN	2019	Word-level Fusion	RAVEN	CMU-MOSI	Recurrent Attended Variation Embedding Network (RAVEN) Nonverbal Sub-networks Gated Modality mixing Network Multimodal Shifting	78.0%

4. 多模态融合技术

不同模态的融合是多模态情感分析的核心。多模态融合是从各种来源接收的数据中进行过滤、提取和组合所需特征的过程。然后进一步分析这些数据，提取意见并评估态度。表2列出了许多融合过程及其解释。早期的研究[87-89]表明，任何三种模态的组合、双模态系统和三模态系统的准确性都优于单模态系统。

数据融合、特征融合和决策融合是组合或融合数据的三种方法。大部分工作采用了决策融合。在特征级融合中，通过组合独立输入特征形成联合特征向量。特征级融合混合了韵律和面部表情元素。多模态情感分析是一个相对较新的研究领域。结合输入模式可以提高分析的精确性。通过将来自多种模态的特征（如文本、音频和视觉元素）组合成通用特征向量，然后对该向量进行标准化，创建一个通用特征向量。然后，将产生的组合特征向量传输以进行分析。这些基本算法用于从音频、视频和文本等多种模态中提取冗余信息，以及视频话语之间的上下文信息。图2总结了各种融合类别的最新模型。

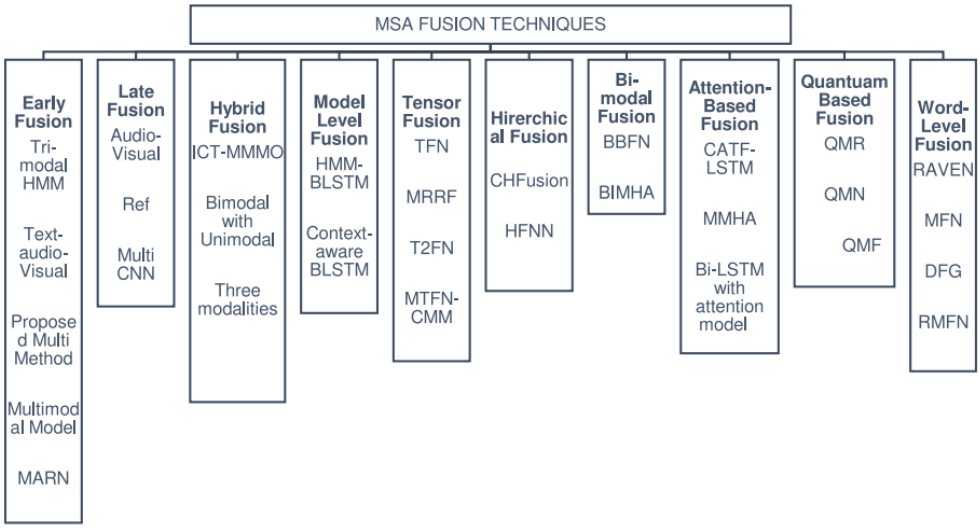

Fig. 2. Multimodal fusion models for multimodal sentiment analysis.

4.1. 早期融合-特征级融合

特征级融合（有时称为早期融合）将每种模态（文本、音频或视觉）的所有特征组合成一个特征向量，然后输入分类算法。特征级融合的好处在于它允许不同多模态特征之间的早期关联，这可能会导致更好的任务完成。应用这种策略的挑战之一是整合不同元素。这种融合方法的缺点是时间同步，因为收集到的特征属于多个模态，在许多方面可能有很大差异。因此，在融合过程之前需要将特征转换为所需的格式。在特征级别上对模态进行融合面临着将差异很大的输入特征整合在一起的挑战。这意味着在重新训练模态的分类系统时，需要同步各种输入，这是一个困难的过程。如何创建由具有不同时间尺度、度量级别和时间结构的不同模态特征组成的可接受联合特征向量是一个未解决的问题。将音频和视频特征连接成一个单一的特征向量，就像现有的使用特征级数据融合的人体情感分析器所做的那样，显然不是答案。这种融合方法无法有效地表示模态内的动态变化。它无法过滤掉来自多个模态的冲突或冗余数据。图3展示了一个早期融合的架构。

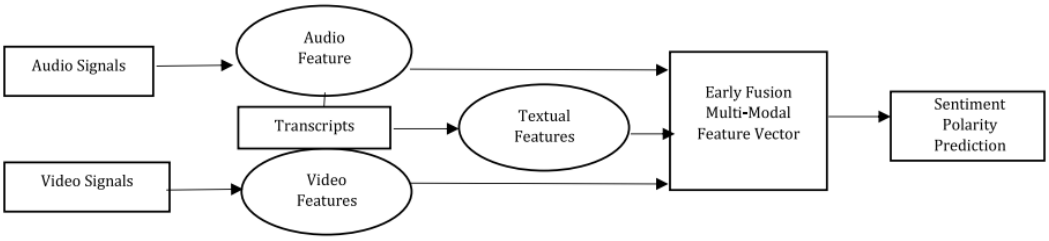

Fig. 3. Early Fusion for Multimodal Sentiment Analysis.

4.2. 晚期融合-决策级融合

在晚期融合中，首先独立处理和分类每种模态的特征。然后将分类结果融合形成最终的决策向量，进而得出情感预测。由于融合发生在分类之后，这个过程被称为晚期融合。由于早期融合存在的挑战，大多数研究人员选择决策级融合，在这种方法中，每个模态的输入分别建模和分类，然后将单模态识别的结果集成在最后。在机器学习和模式识别领域，决策级融合，也被称为分类器融合，是一个热门话题。许多研究已经证明了分类器融合相对于独立分类器的优越性，因为不同分类器之间的错误是不相关的。由于许多模态产生的决策通常具有相同的数据形式，融合来自不同模态的决策比特征级融合更容易。这种

融合过程的另一个好处是每个模态可以使用最佳分类器或模型来学习其特征。当分析任务需要不同的分类器时，决策级融合步骤中所有这些分类器的学习过程变得困难且耗时。图4展示了晚期融合的架构。

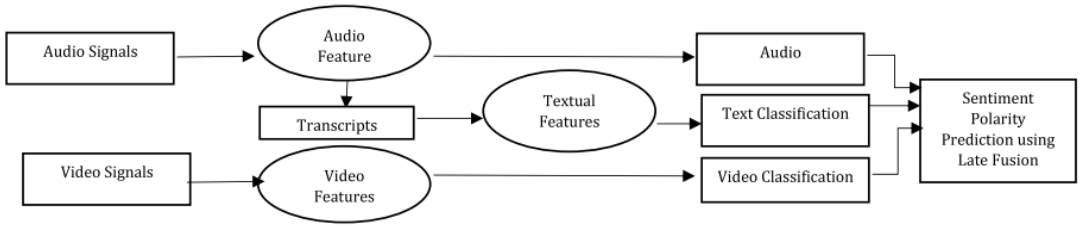

Fig. 4. Late fusion for multimodal sentiment analysis.

4.3. 混合融合

混合融合是早期融合和晚期融合技术的混合体。这种融合方法结合了特征级和决策级融合技术。研究人员使用混合融合来充分利用特征级和决策级融合过程的优势，同时避免各自的缺点。

4.4. 模型级融合

研究人员研究了不同模态的特征，以查看它们之间是否存在联系。然后根据研究领域和问题需求创建所需的模型。这是一种将来自多个模态的数据结合起来，并利用相关性创建松散融合的技术。研究人员创建了符合研究需求的模型，同时考虑了问题空间。

4.5. 张量融合

这种方法使用张量融合层构建了一个三重笛卡尔积，利用模态嵌入来明确模拟单模态、双模态和三模态的交互。它最小化了所需的训练样本数量。其中一种张量融合技术（MTFN）的架构如图5所示。

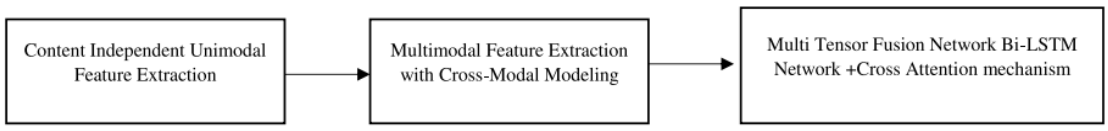

Fig. 5. MTFN architecture for tensor fusion.

4.6. 分层融合

分层融合是一种独特的特征融合方法，按照层次顺序进行工作，首先将两个模态进行融合，然后再将所有三个模态进行融合。这种简单方法的困难在于它无法过滤掉来自其他模态收集到的相互矛盾或冗余的数据。为了解决这个主要问题，我们开发了一种分层技术，从单模态向双模态向量发展，然后再从双模态向三模态向量发展。在两个模态的融合中，对于每个双模态组合，比如T + V、T + A和A + V，它融合了话语特征向量。它使用三个双模态特征来生成一个三模态特征。图6展示了分层融合中的一个示例，即HFFN模型架构。

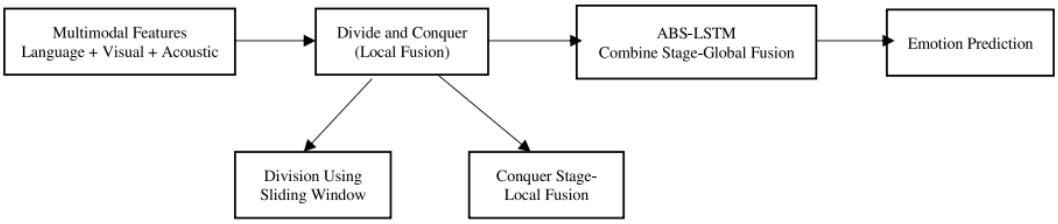

Fig 6. HFFN architecture for hierarchical fusions.

4.7. 双模态融合

在双模态融合中，通过一种新颖的端到端网络，在成对的模态表示上实现融合（增加相关性）和分离（增加差异）。这两个组件同时进行训练，以便它们可以在模拟战斗中相互竞争。由于不同模态之间已知的信息不平衡，该模型将两个双模态配对作为输入。图7显示了双模态融合中的一个架构示例，即BBFN模型架构。

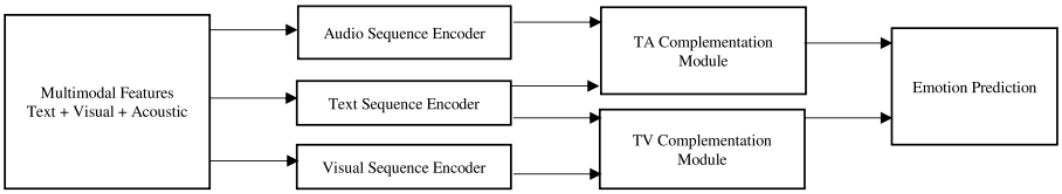

Fig. 7. BBFN architecture for bimodal fusion.

4.8. 基于注意机制的融合

在多模态情感分析和情感识别中，上下文信息提取和多模态融合是两个最重要的难题。使用基于双向循环神经网络的模型进行多层次上下文特征提取被称为基于注意机制的融合。在话语级别上，每个模态对情感和情绪分类的贡献是不同的。因此，该模型建议采用基于注意力的模态间融合，以适应每个模态话语的重要性。上下文关注的单模态特征两两组合形成双模态特征，然后将它们全部合并成三模态特征向量。在融合的每一步之后提取上下文特征。图8展示了基于注意机制的融合中的一个示例，即MMHA模型架构。

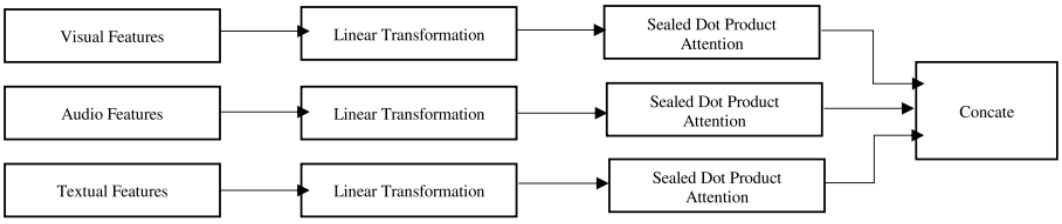

Fig. 8. MMHA architecture for attention based fusions.

4.9. 基于量子的融合

基于量子的融合利用了量子干涉和量子测量理论。使用量子干涉和强弱影响模型捕捉每个话语内部的相互作用（即不同模态之间的相关性），并使用量子测量来检测该方法中连续话语之间的相互作用（即一个发言人如何影响另一个发言人）。它还利用了决策级或晚期融合的方法。图9展示了基于量子的架构之一，即QMN模型架构。

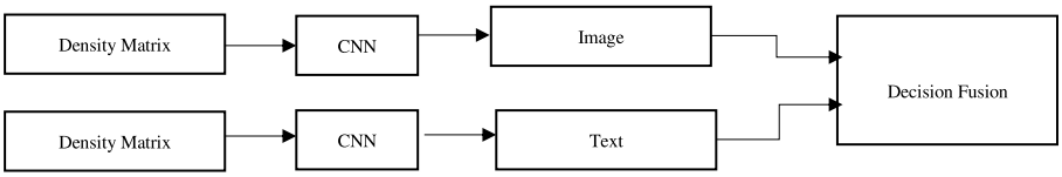

Fig. 9. QMN architecture for quantum based fusion.

4.10. 单词级融合

在这种策略中，为了获取更好的情感倾向，需要研究多个模态之间的交互作用。使用Transformer来学习话语的联合表示，并在不同模态之间进行转换。记忆融合网络（MFN）是一种用于多视角序列学习的循环模型，由三个部分组成：（1）长短期记忆（LSTM）网络，用于编码每个视角独特的动态和交互；（2）增量记忆注意网络是LSTM系统中的一种特殊注意机制，用于发现记忆的不同维度之间的跨视角和时间关系；（3）多视角门控记忆（MVGM）是一种统一的记忆，记录了时间上的跨视角交互。图10展示了RMFN流水线及其组件。MFN以一个由N个视角组成的多视角序列作为输入，每个视角的长度为T。

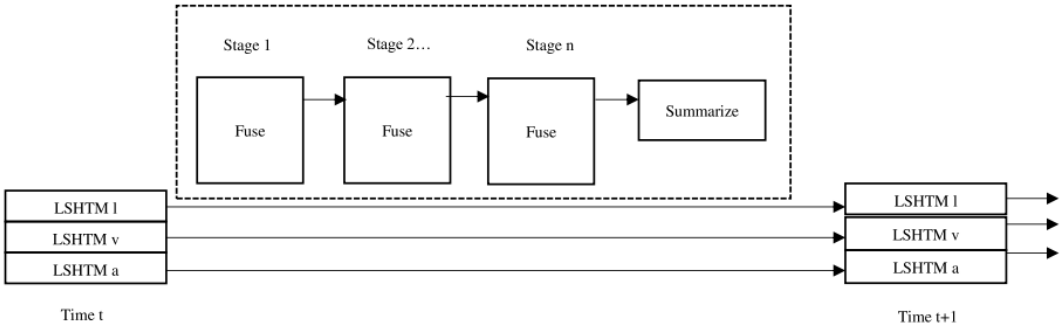

Fig. 10. RMFN architecture for word-level fusion.

4.11. 最新的可变多模态情感分析模型

根据计算机科学文献数据库dblp的数据，过去五到七年中有大约177篇文章展示了多模态情感分析框架的架构转变。该领域的研究范围和深度已经增长到无法展示整个模型的程度。2020年约有36篇论文，2021年约有67篇论文，2022年5月之前有13篇论文。本文介绍了一些已经进行基准测试的多模态情感分析模型。

4.11.1. GESentic（实时GPU-ELM多模态情感分析）（2017年）

GESentic是一种实时多模态情感分析方法，利用Sentic表情（可以产生多种人类情感的基本输入）结合ELM和GPU进行集成应用。为了改善从不同模态中提取特征的性能，该方法利用了多种适用于GPU的方法。此外，情感分析模型是基于检索到的特征构建的复杂ELM分类器[55]。

4.11.2. CSFC（卷积模糊情感分类器）（2019年）

CSFC是一种结合了深度卷积神经网络和模糊逻辑的模型。大多数短语包含了混合情感，比如讽刺，只能通过模糊隶属函数成功描述。深度学习被用于从每个模态中提取信息，然后将其投影到单个情感空间中，并将其聚类为不同的情感。它利用模糊逻辑分类器来预测情感空间中某种情感的程度，因为现实世界中的人们对于某个观点目标有部分或混合的感觉[54]。

4.11.3. HALCB（基于认知大脑边缘系统的分层注意力-BiLSTM模型）（2020年）

HALCB是一种受认知神经科学启发的多模态融合技术。HALCB将多模态情感分析分为两个模块，每个模块负责一个任务：二分类或多分类。前一个模块识别输入项的极性，并将其分成两组，然后分别发送到后一个模块。此模块中使用哈希技术以提高检索准确性和速度。后一个模块中有一个正向子网络专门处理正向输入，而一个负向子网络专门处理负向输入。为了与其各自的角色匹配，每个二分类模块和多分类模块中的两个子网络具有单独的融合方法和决策层。在最终的连接处，还额外添加了一个随机森林，用于收集所有模块的输出并在决策层进行融合[53]。

4.11.4. AHRM（基于注意力的异构关系模型）（2020年）

注意力异构关系模型（AHRM）被用于对包含内容和社交关系的多模态情感进行分类。该方法使用渐进式双重注意力模块，在学习内容信息的视角下收集图像和文本之间的相关性，形成联合的图像-文本表示。本方法提出了通道注意力机制，用于突出语义丰富的图像通道，以及基于注意力通道的区域注意力机制，用于强调情感区域。此外，除了学习社交图像的高质量表示外，还开发了一个异构关系网络，从社交情境中聚合内容信息，扩展了图卷积网络。该模型适用于Flicker和GI数据集，分为四个部分：（1）单模态表示学习，使用视觉和文本视角学习单模态图像表示。（2）渐进双重图像-文本注意力，通过两种创新的跨模态注意力（通道注意力和区域注意力）将图像和文本相关性嵌入联合的图像-文本表示中。（3）异构关系融合，从社交关系创建异构关系网络，并扩展图卷积网络，以从社交环境中收集内容信息，作为开发高质量图像表示的补充。（4）情感预测，最终负责情感分类[56]。

4.11.5. SFNN：语义特征融合神经网络（2020年）

SFNN是一种语义特征融合神经网络。该模型首先利用卷积神经网络和注意机制获取图像的有效情感特征表达。然后将情感特征表达映射到语义特征层次。最后，通过将图像的物理层次的情感特征与视觉模态的语义特征相融合，高效地检测评论的情感极性。基于语义层次的特征融合可以减少异构数据之间的差异[60]。

4.11.6. SWAFN：情感词感知融合网络（2020年）

SWAFN通过将情感词知识融入融合网络，引导多模态方面的联合表示学习。该方法分为浅层融合和聚合两个部分。为了获得融合的浅层表示，它应用了跨模态竞争技术，从每两个模型中获取双向上下文信息。为了支持和引导三种模态的深层融合，并获得最终的情感词感知融合表示，它构建了一个情感词分类的多任务聚合组件。分析中使用了CMU-MOSI、CMU-MOSEI和YouTube数据集[61]。

4.11.7. MISA：多模态情感分析中的模态不变和特定表示（2020年）

MISA是一个独特的框架，将每种模态分成两个子空间，然后将它们融合起来预测情感状态。第一个子空间是模态不变的。在这个子空间中，来自不同模态的表示学习共享共性，缩小它们之间的差距。通过分布对齐，将话语的所有模态映射到一个共同的子空间中。尽管多模态信号来自不同的来源，但它们都共享说话者的动机和目标，这些因素决定了话语的整体情感状态。第二个子空间是模态特定的，对每种模态都是独特的，并包含其特定的特征。作为共享子空间上对齐的投影，不变映射有助于捕捉这些潜在的共性和相关特征。该模型使用了MOSI和MOSEI数据集[91]。

4.11.8. MAG-BERT（Multimodal Adaption Gate-Bidirectional Encoder Representations from Transformers）（2020年）

该MAG-BERT模型结合了BERT和XLNet的核心架构，并配备了适当开发的多模态适应门（MAG）。由于MAG的存在，BERT和XLNet在微调过程中可以接收多模态非语言数据。它通过使BERT和XLNet转换为依赖于视觉和音频模态的内部表示来实现这一点。MAG利用基于非语言行为的注意力，将信息丰富的视觉和音频组件转化为具有轨迹和幅度的向量。在多模态情感分析中，它使用了CMUMOSI和CMU-MOSEI数据集。通过对MAGBERT和MAG-XLNet进行微调，以及对BERT和XLNet进行仅语言微调，可以显著提高情感分析性能，超过之前的基准[93]。

4.11.9. M2Lens（2021年）

M2Lens是一种全新的解释性可视化分析工具，可以帮助多模态机器学习模型的开发者和用户更好地理解和诊断用于情感分析的多模态模型。M2Lens利用事后解释性技术，考虑特征重要性，从全局子集和局部层次分析多模态语言模型学习到的内部模态内和模态间交互。此外，它允许对多模态元素及其对情感分析模型判断的影响进行多方位的研究。M2Lens由四个主要视角组成。它展示了三种不同形式的交互（主导、互补和冲突）对模型预测的影响。此外，多模态特征模板和可视化图形使探索一组经常使用和具有影响力的特征集更加容易[51]。

4.11.10. 波斯语多模态情感分析框架（2021年）

这是一个创新的波斯语多模态情感分析框架，以一种与上下文相关的方式融合听觉、视觉和文本元素。它创建了第一个用于波斯语言表达和情感极性的多模态数据集，该数据集源自YouTube视频。在这个上下文感知的多模态融合框架中，使用卷积神经网络（CNN）和长短期记忆（LSTM）等自动特征提取方法来提取单模态和多模态特征（LSTM）。不同的特征被组合用于进行多模态情感分析。该框架展示了

一种上下文感知的多模态策略，以克服波斯语中模棱两可的词语的限制[52]。

4.11.11. TCM-LSTM（Temporal Convolutional Multimodal LSTM）（2021年）

TCM-LSTM是使用音频和视觉LSTM从新的角度研究模态间动态的方法，其中语言方面最为重要。在每个LSTM变体内引入了精心设计的门控机制，通过相关的辅助模态增强语言表示。它由两个主要部分组成：（1）具有音频和视觉组件的LSTM，用于改进口语语言的表示。在每个LSTM变体中都包含了精心设计的门控机制，以评估是否应基于每个模态中表达的区分性信息进行音频和视觉增强。（2）通过将“通道相互依赖性学习”模块引入传统的TCN，生成了一个“通道感知”的时序卷积网络[57]。

4.11.12. MMIM（MultiModal InfoMax）（2021年）

多模态Infomax（MMIM）是一种用于多模态融合的框架，通过层次化地最大化单模态输入对（模态间）和多模态融合结果之间的互信息（MI）来保持任务相关信息。为了避免关键的任务相关数据的丢失，MI最大化在输入和融合层级上进行。据我们所知，这是将MI与MSA相结合的首次尝试。为了解决难以处理的问题，该公式包括参数化学习和非参数化GMM，具有稳定且平滑的参数估计[58]。

4.11.13. OMSKELM（Optimal Multimodal Sentiment classification using the Kernel Extreme Learning Classifier）（2021年）

该系统在进行多模态情感分析之前，研究了文本、音频和视频之间的关系。提取了一组独特的特征，并使用新的混合蜂群优化算法（HGBEE）对提取的特征进行优化，以获得具有改进的精度和减少计算时间的优化特征集。然后使用文本、音频和视频信息创建了话语级别的多模态融合。最后，该系统使用多核极限学习分类器（MKELM）进行情感分类[59]。

4.11.14. TIMF（Two Level Multimodal Fusion）（2021年）

该系统提出了一种两级多模态融合（TIMF）方法，结合了数据级和决策级融合，以完成情感分析任务。在数据级融合阶段，使用张量融合网络将文本分别与音频和视频属性融合，创建文本-音频和文本-视频嵌入。在决策级融合步骤中使用软融合方法，将上游分类器的分类或预测结果融合，以尽可能准确地得出最终的分类或预测结果。使用CMUMOSI、CMU-MOSEI和IEMOCAP数据集进行测试[62]。

4.11.15. AWMA（Asymmetric Window Multi Attention）（2021年）

门控循环单元（GRU）模块、不对称窗口注意力（AWA）模块和模态间注意力（IMA）模块构成了不对称窗口多注意力（AWMA）神经网络。在多模态情感分析中，使用GRU和AWA模块捕捉模态内动态。使用IMA模块描述模态间动态。然后逐渐合并这三个模块。首先，GRU模块接收输入序列，然后是AWA模块，最后是IMA模块。所提出的方法中，使用不对称窗口表达输入数据特定时间戳处历史和未来上下文的不对称权重是一种新颖的特点。它认为应该区分输入数据特定时间戳处的前一个和后一个上下文。此外，不对称窗口可以用来指示上下文的隐含权重[63]。

4.11.16. 基于自动机器学习的融合（2021年）

该方法包括预处理、个别分类和融合三个阶段。融合方法侧重于使用文本和图像组件的独立分类结果进行最终分类。融合策略的目标是通过利用两个来源的上下文知识超越个别分类。使用由Twitter数据组成的B-T4SA数据集。数据集中的每个样本都包含文本和图像[64]。

4.11.17. Self-MM（Self-Supervised Multi-Task Multimodal sentiment analysis network）（2021年）

Self-MM模型提出了一种自监督的单模态标签生成过程，节省了大量时间和金钱。广泛的测试证明，自动生成的单模态标签是可靠和稳定的。使用模态表示和类别中心之间的距离计算相对距离值，与模型输出呈正相关。它还设计了一种新颖的权重调整方法，以平衡各种任务相关的权重损失限制。Self-MM的目标是同时学习一个多模态任务和三个单模态子任务，以创建信息丰富的单模态表示。使用MOSI、MOSEI和SIMS数据集[92]。

4.11.18. DISRFN（Dynamic Invariant Specific Representation Fusion Network）（2022年）

通过结合每个模态的模态不变性和模态特异性表示，可以更有效地利用每个模态的冗余信息。在这个模型中，一个简单的动态融合方法可以更快地获取模态之间的交互作用。DISRFN框架由两部分组成：增强的JDSN和HGFN。首先，使用升级的JDSN模块生成每个模态的模态不变特定联合表示。它允许在模态之间高效利用互补信息，并减小模态之间的异质性差距。在性能分析实验中使用了MOSI和MOSEI数据集，结果令人满意[50]。

4.11.19. MH-GAT（Multi-Feature Hierarchical Graph Attention Model）（2022年）

为了分析情感，该研究提出了一种基于共现和句法依赖图的脑启发式多特征分层图注意模型（MH-GAT）。它同时考虑了各种结构信息、词性信息和位置关联信息。它由双图分层注意和多特征融合组成。输入层包含许多特征，如词性、位置、句法依赖和共现信息。它构建了每个文本的双图分层注意模型和图注意网络。与最新版本的AT&T BLSTM、Text-Level-GNN和TextING相比，所提出的模型的情感

分析准确度平均提高了5% [98]。

5. MSA应用

情感分析是一种识别和分类对产品或服务的意见的技术。多模态情感分析提供了使用视频、音频和文本的组合进行意见分析的方法，它在分析人类行为方面远远超出了传统的基于文本的情感分析。产品评论、意见调查、YouTube上的电影评论、新闻视频分析和医疗应用等都受益于对这些情感的研究。

多模态情感分析用于对象识别和欺诈检测[65]、交易系统的市场预测[66]、旅游情感分析、#MeToo推文的情感分析[67]等。MSA在人机交互[68]、医疗应用[69]、教育和学习中的情感分析[70]、推荐系统[71]、对任何当前问题的情感分析[72]等方面取得了显著的成功。下面描述了一些选定的用例。

5.1. 多模态动作识别和欺诈检测

在欺诈检测中，可以使用多种模态。利用照片、视频、面具或其他替代品来避免欺诈性面部验证任务被称为面部反欺诈。它用于检测错误的输入。面部反欺诈基准使用RGB、深度和红外传感器来检测虚假的面部输入。欺骗一个单模态系统很容易。在一次著名的事件中，被称为“facegate”，在纸上打印的面部或3D面具能够欺骗基于RGB的系统。另一方面，同时欺骗多个模态要困难得多。在动作识别、情感分析和面部反欺诈领域，已经提出了具有不同模态的各种数据集。在动作识别数据集中，有三种不同的模态可以利用：视觉（RGB序列）、运动（光流）和听觉。直到最近，只使用了视觉和运动模态。一个流行的方法是为每个模态训练一个模型，并通过后融合进行集成。传感器无法防御面部识别中的3D面具。一些动作（例如，捏合）只能通过音频模态识别，而一些情感只能通过语调（音频）、词语（语言）或面部表情（视觉）表达。

5.2. 交易系统的市场预测

在多模态增强交易系统的开发中，使用了强化学习、情感分析和多模态学习。在进行交易选择时，代理商不仅要考虑价格波动，还要考虑新闻信息。多模态学习可以结合多样的数据模态来提高模型的性能，情感分析可以用于理解新闻的情感。多个模态与特征向量或神经网络连接起来，产生联合表示。根据实验结果，金融情绪词汇与风险预测存在密切关联。一些专家认为新闻事件对股票价格的波动有影响。2020年的重点是情绪分类，已经证明对股市价值有影响。2021年提出了一种基于报纸的情感指数，可以实时监测西班牙的经济活动。这项指标不仅超过了欧洲委员会著名的经济态度指标，而且在预测西班牙国内生产总值（GDP）方面表现良好。

5.3. 旅游情感分析

乘客经常通过微博等社交媒体平台表达他们的情感（微博是中国著名的微博客社交网络）。利用微博分析乘客的意见可以帮助理解当前乘客的情绪，提供及时满意的服务，改善乘客体验。使用文本和图像等多种模态来建模旅行事件和情感。为了学习有区别的多模态表示，考虑了单模态内容和跨模态关系。

5.4. Instagram帖子的情感分析

Instagram是一个良好的社交媒体平台，用户可以在上面交流他们的想法、情感和观点，也可以对其他人的帖子点赞和评论。在自然语言处理和情感分析领域，对帖子的情感进行分析具有多种应用。研究中提出了一种多模态深度融合方法，有效地分析帖子的情感。对于图片和文本情感分析，该方法有两个并行分支。研究人员费力收集了一个包含波斯语评论及其相关图像的多模态数据集MPerInst。根据实验结果，这种策略在性能上优于类似的多模态深度模型和传统机器学习方法。

5.5. #MeToo推文的情感分析

由于#MeToo运动的影响，人们更加公开地谈论自己的骚扰经历。该项目旨在汇集这些性侵经历，以更好地理解社交媒体的构建方式并实现社会变革。研究人员使用深度神经网络将视觉分析和自然语言处理结合在一起，采用多模态情感分析方法。该方法旨在确定一个人在这个话题上的立场，并推断出表达的情绪。他们使用了一种名为Multimodal Bi Transformer (MMBT)的模型，该模型利用图片和文本信息来预测推文对#MeToo运动的立场和想法。

5.6. 人机交互

Pepper是一个外观像人类的机器人。人们可以以各种方式积极地解读与机器人的交流。尤其是对于可能与他人沟通困难的病童。该设备由Raspberry Pi、摄像头和麦克风组成，通过分析音频、文本和视频来提供关于对话者情绪、面部表情、肢体语言和其他因素的独立信息源。

5.7. 医疗保健

由于社交媒体的兴起，患者现在可以通过发布在线评价来评价他们接受的医疗治疗质量。在线评价的详细文本和视觉材料提供了有关患者与医生的互动和对医疗服务满意度的见解。各种研究使用文本内容分析患者的意见。本研究引入了一种新颖的多模态技术，用于分析患者对医疗服务质量（高 vs 低）的感知。不仅评估原始的书面内容，还评估了来自Yelp.com平台的图像内容。由于需要从多个模态提取特征，这更加困难。

5.8. 教育与学习

情感发现和分析（SDA）是一种自动检测对特定事物（如学习者和学习资源）的潜在态度、情感和主观性的技术。由于其巨大的潜力，SDA已被誉为在教育过程中从多模态和多源数据中识别和分类情感的有效技术。多模态SDA可以为教育利益相关者在课堂或在线学习中的情感提供互补的见解。SDA在大规模教育数据流的单模态特征选择、情感分类和多模态融合方面面临问题。因此，大量的研究探讨了各种用于教学目的的SDA方法。

5.9. 推荐系统

电子商务中推荐系统的最常见应用是在在线广告中向购买者推荐可能符合他们兴趣的商品，以满足他们的偏好。推荐系统在广义上是一种算法，试图向消费者提供相关的物品推荐（物品可以是电影、文本、商品或其他根据不同行业而定）。在某些企业中，推荐系统至关重要，因为它们可以产生大量收入或作为区别于竞争者的方法。

5.10. 对当前问题的情感态度

情感分析可用于确定公众对各种话题的意见，以及情感趋势中的高峰和低谷。COVID-19疫苗接种是当前全球各国在疫情期间的一大关注点。自2020年底以来，疫苗已广泛分发，接种者因此看到了更少的疾病、住院和死亡病例。然而，对疫苗的担忧仍然存在，特别是考虑到接种后的副作用。社交媒体分析有能力向决策者提供有关公众讨论的副作用以及对国家免疫计划的公众看法的信息。最近的研究凸显了人工智能驱动的社交媒体分析作为补充传统评估方法的潜力。例如，公共调查可以向政府和组织提供有关公众情感的信息。社交媒体分析尚未应用于调查COVID-19疫苗的常规报告的不良事件。它有助于发现在其他地方未被注意到的潜在安全信号（例如，很少报告的副作用）。

6. MSA挑战

根据上述研究的发现，未来的研究应解决以下问题：需要在多种语言中建立一个稳定的多模态数据集。该数据集应进行良好的注释和细粒度标注。需要关注共指消解问题，隐藏情绪、讽刺和挖苦的检测仍然是一个使用多种模态的开放性研究问题。数据集应在伦理上进行准备和分析，并广泛提供给公众领域，以促进更好的研究和平等机会。需要调查不同模态的隐含和显性含义，以及在文本中使用的混合代码、简化形式、噪音和低分辨率的照片和视频的隐含和显性含义。除了上述所有内容外，多语言方向的开放性研究课题还包括处理多语言数据、跨领域准确性改进、跨数据集准确性改进以及使用上下文背景的情感分析。

识别隐藏情绪（如讽刺或挖苦）的问题长期以来一直是该学科学者的挑战。因为这些情绪在文本中不会立即表达出来，所以被称为隐藏情绪。人们通过非语言交流和上下文两种线索来感知这些情绪。利用多模态情感分析来检测非语言线索和上下文，可以利用相同的线索来检测隐藏情绪。多模态情感分析是一个新的研究领域。以往的研究大多成功地表明，结合两种或三种输入类型可以提高分析的准确性。研究已经探讨了上下文在情感分析中的作用，并得出结论认为上下文可以提高分类准确性。未来可能的研究方法是结合多模态线索和上下文，重点是识别隐藏情绪。

情感分析中的特征提取面临一些问题，包括领域特异性（意味着它在其他领域中无法起作用）、冗余性、高维度、粗话、混合编码的数据、偏见和上下文依赖性。以下列举了其中一些困难。

跨领域：情感分类因意见表达方式在不同领域之间的差异而具有领域敏感性。在情感分析中，领域适应通过学习未知领域的特征来解决这个困难。在一个领域中的一个词的含义可能与在另一个领域中的一个词的含义不同。

高维度：指具有大量特征集的情况，由于计算问题导致性能下降，因此需要使用正确的特征选择方法。

混合编码的数据：这是指在同一语句中使用两种或更多语言的信息。混合编码是将一个语言的语言特征（如短语、词语和词素）嵌入到另一种语言的话语中的过程。例如，“我昨天去看电影，然后在路上遇到了Rima”是混合编码的示例。混合编码也对基于规则和深度学习的方法提出了挑战。在这个领域的发展非常有限。

偏见：情感分析工具经常被应用于涉及敏感主题（如咨询）的领域，通过对来自不同背景的客户电话和市场线索进行情感指标的检查，获取的数据对重要的决策起着推动作用。因此，在涉及人口统计学时，识别偏见尤为重要。偏见可以以许多形式存在，包括性别、肤色、年龄等。

上下文依赖性：情感词的使用取决于话题。当与其他词或短语结合使用时，表面上中性的词语也可以传达情感。例如，当有人想要买一所大房子用于休闲时，词语“大”可能具有积极的含义。然而，当在另一个上下文中使用时，同样的词可能引起消极情绪，因为大房子难以清洁。遗憾的是，情感分析研究很少关注到这个要素。

早期融合/特征融合：由于多模态情感分析考虑到不同模态的融合，可以通过各种多模态融合技术来进行研究。早期融合的MSA架构面临的主要挑战列在表3中。

Table 3
Major issues related with the multimodal fusion based on early fusion.

Model Name	Merits	Demerits
Tri-modal HMM	• First time addresses the task of tri-modal sentiment analysis • Improvements are observed for both precision and recall • Tri-modal HMM is able to learn the hidden interaction between all three modalities and take advantage of their respective discriminative power	• Domain specific model • Small scale dataset
Text-audio-Visual	• Joint use of visual, audio, and textual features greatly improves over the use of only one modality at a time • Significant improvements are also obtained on a second dataset of English videos	• Domain specific • Occluded frames in videos
Proposed Multi Method	• System outperformed all state-of-the-art systems on the eNTERFACE dataset. • With multimodal fusion, system outperformed the best state-of-the-art system by more than 10%.	• Time complexity of the methods needs to be reduced to a minimum.
Multimodal Model	• Idea of thin slices can be used to observe a short window of a speaker's behaviour to achieve comparable prediction compared to observing the entire length of the video. • New dataset time complexity of the methods needs to be reduced to a minimum introduced.	• Same-gender design in evaluating a speaker's persuasiveness and other high-level attributes. • Inherent variability in human perception and judgement. • Investigating more ways of computationally capturing various indicators of persuasiveness and better algorithmic methods of fusing information from multiple modalities. • Deeper analysis to understand relationship between persuasiveness and relevant high-level attributes including personality.
MARN	• Main strength of our model comes from discovering interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory of a recurrent • Component called the Long-	• Under performs for cross-view dynamics • Performance varies as different tasks and datasets require different number of attentions

short Term Hybrid Memory
(LSTHM).

后期融合：对每种模态的不同特征进行独立的检查和分类，然后将结果融合以生成最终的决策向量。后期融合或决策融合的主要问题是，在决策级融合阶段使用不同的分类器进行分析任务时，所有这些分类器的学习过程变得繁琐和耗时。后期融合的MSA架构面临的主要挑战列在表4中。

Table 4
Major issues related with the multimodal fusion based on late fusion.

Model Name	Merits	Demerits
Audio-Visual	• The model combines multiple measurements and can handle the absence of measurements by increasing the uncertainty of the predicted state • The uncertain measurements that are usually processed by the Kalman filter is replaced by multi-modal classifier outputs. • Filtering resulted in a reconstruction of missing classification outputs such that a class assignment is possible.	• A direct interpretation of parameter is difficult. For instance, the audio classifier is only able to provide outputs in case a signal is present, such that the classifier outputs of the video channel are much more frequent than the classifier outputs of the audio channel. As a result, the rejection rate and the process noise for audio and video cannot be compared with each other. • The ratio between these modalities is additionally modified by assessing the quality in terms of rejecting unreliable outputs
Ref	• The feature sets include audio-visual, lexical, POS, LIWC, emotional features and their combinations using majority voting. • Used predicted traits as features and designed a cascaded classification system	• Same feature set or architecture might not work for all traits. • Performance of the model using emotional feature set is very low compared to the other feature sets.
Multi CNN	• New CNN architecture that fully uses joint text-level and image-level representation • Complementary effect of the two representations as sentiment features improves performance	• Uses only two modalities of text and image.

混合融合：这种类型的融合是特征级和决策级融合方法的结合。它充分利用了特征级和决策级融合策略的优势，并克服了各自的缺点。混合融合的MSA架构面临的主要挑战列在表5中。

Table 4
Major issues related with the multimodal fusion based on late fusion.

Model Name	Merits	Demerits
Audio-Visual	• The model combines multiple measurements and can handle the absence of measurements by increasing the uncertainty of the predicted state • The uncertain measurements that are usually processed by the Kalman filter is replaced by multi-modal classifier outputs. • Filtering resulted in a reconstruction of missing classification outputs such that a class assignment is possible.	• A direct interpretation of parameter is difficult. For instance, the audio classifier is only able to provide outputs in case a signal is present, such that the classifier outputs of the video channel are much more frequent than the classifier outputs of the audio channel. As a result, the rejection rate and the process noise for audio and video cannot be compared with each other. • The ratio between these modalities is additionally modified by assessing the quality in terms of rejecting unreliable outputs
Ref	• The feature sets include audio-visual, lexical, POS, LIWC, emotional features and their combinations using majority voting. • Used predicted traits as features and designed a cascaded classification system	• Same feature set or architecture might not work for all traits. • Performance of the model using emotional feature set is very low compared to the other feature sets.
Multi CNN	• New CNN architecture that fully uses joint text-level and image-level representation • Complementary effect of the two representations as sentiment features improves performance	• Uses only two modalities of text and image.

模型级融合：这是一种将来自多个模态的数据结合起来并利用相关性创建松散融合的技术。研究各个模态的特征，看是否存在它们之间的联系。然后根据研究领域和问题需求创建所需的模型。模型级融合的MSA架构面临的主要挑战列在表6中。

Table 6
Major issues related with the multimodal fusion based on model-level fusion.

Model Name	Merits	Demerits
HMM-BLSTM	• Context-sensitive schemes for emotion recognition within a multimodal, hierarchical approach at utterance level • Sheds light into the flow of affective expressions revealing potentially useful patterns	• Less accuracy in context sensitive as compared to context free
Context-aware BLSTM	• System based on LSTM longrange temporal context modelling in order to discriminate between high and low levels of AROUSAL, EXPECTATION, POWER, and VALENCE using statistical functionals of a large set of acoustic low-level descriptors, linguistic information (including non-linguistic vocalizations), and facial movement feature	• Under performs and absolute values of the reported accuracies seem low in comparison to easier scenarios, such as the discrimination of acted, prototypical emotions.

张量融合：在基于张量的融合方法中，通过张量融合层构建了一个三倍的笛卡尔乘积，明确建模了单模态、双模态和三模态的相互作用。张量融合的MSA架构面临的主要挑战列在表7中。

Table 7
Major issues related with the multimodal fusion based on tensor fusion.

Model Name	Merits	Demerits
TFN	• Learns intra-modality and inter-modality dynamics end-to-end • Tensor fusion layer used to generate 3-fold Cartesian product from modality embeddings	• Risk of Overfitting due to very high dimensionality of the produced tensor especially in case of small datasets like CMU-MOSI • Less specialized fusion process due to little consideration to acknowledging the variations across different portions of a feature vector • Exponential computational increase in number of parameters, cost and memory
MRFF	• Removes redundant information which is duplicated across modalities and in turn leads to fewer parameters with minimal information loss. • Work as a regularizer, in turn leads to less complicated model and reduces overfitting • Modality-based factorization approach helps to understand the differences in useful information between modalities for the task at hand • Tuckers tensor decomposition method is used which gives different compression rates for each modality.	• Small dataset • Requires more efficient training for large dataset
T2FN	• T2FN with rank regularization maintains good performance despite imperfections in data • Model's improvement is more significant on random drop settings, which results in a higher tensor rank as compared to • Structured drop settings	• Tensor rank increases in presence of imperfect data
MTFN—CMM	• Multi-tensor fusion network with the cross-modal modelling for multimodal sentiment analysis in this study, which can capture intra-modal dynamics and inter-modal interactions and can be used for multi-modal affective intensity prediction effectively • MTFN—CMM can perform better in regression and classification experiments	• Works on coarse grain modal fusion • High number of parameters used

分层融合：这是一种分层方法，从单模态向双模态向量，然后从双模态向三模态向量进行处理。分层融合的MSA架构面临的主要挑战列在表8中。

Table 8
Major issues related with the multimodal fusion based on hierarchical fusion.

Model Name	Merits	Demerits
CHFusion	• Hierarchical Fusion, first fusing the modalities two in two and only then fusing all three modalities • Context-Aware Hierarchical Fusion (CHFusion)	• Performance difference between two datasets • Less accurate in unimodal specially in textual modalities
HFNN	• It achieves the highest F1 score. • HFFN learns local interactions at each local chunk and explores global interactions by conveying information across local interactions using ABS-LSTM that integrates two levels of attention mechanism	• The accuracy of HFFN is lower than that of BC-LSTM and CAT-LSTM

双模态融合：通过对成对模态表示进行处理，双模态融合的新型端到端网络实现了融合（增加相关性）和分离（增加差异性）。双模态融合的MSA架构面临的主要挑战列在表9中。

Table 9
Major issues related with the multimodal fusion based on bimodal fusion.

Model Name	Merits	Demerits
BBFN	• Bi-Bimodal Fusion Network- novel end-to- end network that performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations • Modality-specific feature space separator and gated control mechanism	• Performance degrades when visual-acoustic input pairs are added. That is, even after including all modalities in the input, redundant network architecture can cause harmful effects bringing in malicious noise, which damages collected useful information and confuses the model.
BIMHA	• Information-augmented Multi-Head Attention using bimodal feature Inter-modal interaction and inter-bimodal interaction.	• BIMHA is a shallow model in terms of efficiency.

注意力融合：基于注意力的融合的MSA架构面临的主要挑战列在表10中。

Table 9
Major issues related with the multimodal fusion based on bimodal fusion.

Model Name	Merits	Demerits
BBFN	• Bi-Bimodal Fusion Network- novel end-to- end network that performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations • Modality-specific feature space separator and gated control mechanism	• Performance degrades when visual-acoustic input pairs are added. That is, even after including all modalities in the input, redundant network architecture can cause harmful effects bringing in malicious noise, which damages collected useful information and confuses the model.
BIMHA	• Information-augmented Multi-Head Attention using bimodal feature Inter-modal interaction and inter-bimodal interaction.	• BIMHA is a shallow model in terms of efficiency.

量子融合：量子干涉捕捉到每个话语内部的相互作用（即不同模态之间的相关性），并且使用量子测量来创建一个强-弱影响模型，以检测连续话语之间的相互作用（即一个说话者如何影响另一个说话者）。量子融合的MSA架构面临的主要挑战列在表11中。

Table 11
Major issues related with the multimodal fusion based on quantum based fusion.

Model Name	Merits	Demerits
QMR	• Quantum-inspired Multimodal Sentiment Analysis • Fill the 'semantic gap' and model the correlations between different modalities via density matrix • Quantum Interference inspired Multimodal decision Fusion (QIMF) • QIMF adds an interference term	• Influence of different cos θ on the classification results • The computation time used for training and classification is longer than the use of other baselines
QMN	• It uses quantum theory (QT) mathematical formalism and a long short-term memory (LSTM) network • Intra- and inter-utterance interaction dynamic is considered • Decision correlations between different modalities is generated using quantum interference • Strong-weak influence model to make better inferences about social influence amongst speakers is generate using quantum measurement	• It is largely dependant on the density matrix so how to take a further step towards accurately capturing the interactions amongst speakers and naturally incorporating them into an end-to-end framework is difficult. • Because it used an emotion recognition dataset, QMN was tested on emotion recognition tasks rather than sentiment analysis.
QMF	• Superposition and entanglement are used respectively at different stages are used to formulate word interaction within a single modality and the interaction across modalities. • Quantum-theoretic multimodal fusion framework	• Quality of the extracted visual and acoustic features is not high • Inconsistency with quantum theory

词级融合：表12描述了基于词级融合的MSA架构面临的主要挑战。

Table 12
Major issues related with the multimodal fusion based on word level fusion.

Model Name	Merits	Demerits
MFN	• MFN shows a consistent trend for both classification and regression • DMAN can model asynchronous cross-view interactions because it attends to the memories in the System of LSTMs which can carry information about the observed inputs across different timestamps.	• Doesn't support cross view dynamics • Underperforms on some of the datasets than baseline models
DFG	• A new neural-based component called the Dynamic Fusion Graph which replaces DMFN in MFN	• Efficiency decreases when visual modalities are contradictory
RMFN	• Decomposes the fusion problem into multiple stages, each of them focused on a subset of multimodal signals for specialized, effective fusion • Crossmodal interactions are modelled using this multistage fusion approach which builds upon intermediate representations of previous stages • Temporal and intra-modal interactions are modelled by integrating our proposed fusion approach with a system of recurrent neural networks	• Works well for CMU-MOSI dataset than all other datasets.
RAVEN	• Recurrent Attended Variation Embedding Network (RAVEN) • Nonverbal Sub-networks • Gated Modality mixing Network • Multimodal Shifting that models the fine-grained structure of nonverbal subword sequences and dynamically shifts word representations based on nonverbal cues. considering subword structure of nonverbal behaviours • Learning multimodal-shifted word representations conditioned on the occurring nonverbal behaviours.	• Need for large dataset • Underperforms in cross domain

图11展示了每种架构所达到的最大二分类准确度的条形图。

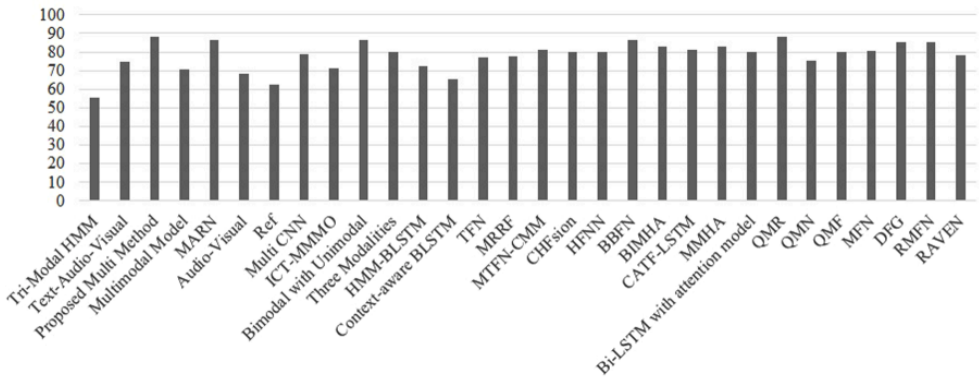

Fig. 11. Binary accuracy of each architecture.

7. 未来的范围

MSA的设计融合了许多独特的概念，尤其在情感分析和情绪识别领域推动了研究目标的发展。它还为评论和推荐系统以及健康预测系统提供了基础。研究MSA的架构进展是一个令人兴奋的研究课题，有潜力成为最广泛使用的自然语言处理方法之一。

7.1. 心理健康预测

随着越来越多的人被诊断出患有抑郁症和其他精神疾病，自动化的心理健康预测，即识别抑郁症和其他精神疾病，已成为一个重要的研究领域。在美国，四分之一的成年人患有心理健康问题，使得心理健康成为一个重要的优先事项。在徐等人最近的研究中，他们使用了用户级别的评估，采用了10折交叉验证。因此，来自同一用户的帖子出现在训练集或测试集中，但不会同时出现在两者中。他们使用了以下分类器：决策树（DT，最大深度为4）；自适应增强（AB）；使用线性核的支持向量机（SVMLinear）。此外，在实验中，他们使用TensorFlow库实现了一个两层的神经网络分类器。具体来说，他们在隐藏层中使用修正线性单元（ReLU）作为激活函数，并使用Sigmoid函数将输出概率限制在0到1之间。在另一项研究中，他们使用机器学习实现了早期融合和后期融合，以根据四种模态（计算机交互、身体姿势、面部表情和心率变异性）预测一个人是否处于压力之下。他们使用迁移学习来预测NASA-TLX分数，该分数预测一个人在0到100的压力水平，并提供了一种机制来在给定时间线上存储监测一个人心理状态的数据随着任务负荷的增加而增加。在Aloshban等人的另一项研究中，他们使用了双向长短期记忆网络进行序列建模和包括语言和声学特征在内的多模态特征。他们应用了后期融合、联合表示和门控多模态单元。与单模态相比，它们产生了更好的结果。

7.2. 情感识别

情感与人类密不可分，因此情感理解是人类化人工智能（AI）的重要组成部分。由于其能够从Facebook、YouTube、Reddit、Twitter等平台上的大量公开对话数据中挖掘意见，对话中的情感识别（ERC）作为自然语言处理（NLP）中的新研究领域变得越来越受欢迎。它还可以应用于医疗保健系统（作为心理分析工具）、教育（理解学生的挫折）和其他领域。此外，ERC对于创建需要理解用户情感的情感感知交互至关重要。ERC面临许多障碍，包括对话上下文建模、对话参与者情感变化等，这些都使得任务更加困难[79]。在情感识别的另一项工作中，引入了一种上下文和情感感知框架，称为Sentic GAT。在Sentic GAT中，常识知识通过上下文和情感感知图注意机制动态表示，通过基于分层多头注意力的对话变换器获取上下文话语的内部和外部依赖关系。内部和外部依赖关系指的是上下文信息与目标话语的关键信息之间的依赖关系[80]。

7.3. 讽刺检测

讽刺是使用词语或习语表达与真实意思相反的含义。人们使用讽刺来取笑他人或谴责他人。讽刺是多模态情感分析中使用的概念，用于描述人们可以使用积极的词语或表达来传达不好的感受，以及使用消极的词语或表达来传达积极的感受。它起到了一个极性翻转的干扰对象的作用。讽刺情感分析是自然语言处理（NLP）中一个快速扩展的研究领域。它包括对词语、短语和句子级别的分类，以及对文档和思想级别的分类。讽刺情感检测可以根据所使用的文本特征分为三类：词汇、语用和夸张。由于写作的比喻性质伴随着微妙的含义和隐含意义，检测讽刺是非常困难的。在最近的研究中，这个研究领域已经成为NLP中的一个主要问题，许多论文提供了各种策略来处理这个任务。声音和文本社区做出了最重要的贡献。讽刺经常不使用语言指示符来表达，而是依赖非语言和语言线索。语调的变化、对词语的过度强调以及一张严肃的脸都是讽刺的迹象。在确定讽刺的方法中，很少有采用多模态策略的研究[76]。其中一些研究包括Castro等人的工作[81]，他们为多模态讽刺研究创建了一个名为MUStARD的新数据集，其中包括高质量的注释，包括多模态和对话上下文特征。它还提供了对话中的前一轮作为上下文信息。因此，它总结了MUStARD的这个特性为未来的工作提出了一个新的子任务：对话中的讽刺检测。在杜等人的另一项研究中[100]，他们提出了一种双通道卷积神经网络，该网络分析目标文本及其情感背景的语义。他们使用SenticNet向长短期记忆（LSTM）模型添加常识。然后应用注意机制来考虑用户的表达习惯。

7.4. 假新闻检测

社交媒体和其他平台上的假新闻广泛存在，并且由于其可能造成重大社会和国家危害的潜力，这是一个令人担忧的原因。检测已经是许多研究的主题。2020年，关于健康的虚假新闻广泛传播，对世界健康构成了风险。2020年2月初，世界卫生组织发出警告，称COVID-19疫情导致了大规模的“信息疫情”，即真假新闻的爆发，其中包含了大量的错误信息。这是一个几乎没有研究工作的领域。在Patwa等人的研究中[82]，他们描述并发布了一个包含10,700条有关COVID-19的真假新闻的数据集。收集了来自各种社交媒体和事实核查网站的帖子，并对每个帖子的真实性进行了手动验证。数据在类别上是平衡的，可以用来开发自动的假新闻和谣言检测算法。使用机器学习算法对开发的数据集进行基准测试，并将其作为潜在的基线。在机器学习模型中，基于支持向量机（SVM）的分类器在测试集上的F1得分最好，为93.32%。

7.5. 仇恨言论检测

仇恨言论用于表达对特定群体的蔑视。它也可以在任何社交媒体上用来羞辱或贬低该群体的成员。黑白分明的仇恨言论在社交媒体上进行诋毁，可能会伤害受害者或使其处于危险境地。它是一种有偏见、不接受和残酷的言论，针对一个人或一群人的一些最糟糕的特征。网络上的仇恨言论，尤其是在Twitter等微博网站上，已经成为过去十年中可能是最严重的问题之一。由于恶意的仇恨活动，仇恨犯罪在许多国家大幅增加。尽管仇恨言论检测是一个新兴的研究领域，但仍然需要研究信息网络中特定主题的仇恨的起源和传播。在Masud等人的研究中[83]，他们预测了Twitter上仇恨言论的发起和传播。通过分析一个大规模的Twitter数据集，他们抓取并手动注释了其中的仇恨言论，他们确定了多个影响仇恨传播的关键因素（外部信息、用户的主题亲和性等）。在Araque和Iglesias的另一项研究中[99]，他们使用了机器学习框架，使用了不同特征的集合。他们还研究了不同特征提取方法（如TF-IDF和基于相似度的情感投射）的影响。他们使用了五个不同的数据集，涵盖了激进化和仇恨言论检测任务。

7.6. 欺骗检测

欺骗被定义为说服或劝诱他人相信不真实的事情。它被描述为发送给接收者以灌输错误信念的信息。在这个数字化时代的计算机辅助互动中，大多数用户都关心如何区分真实消息和假消息。这对于具有高风险因素的活动尤为重要。在线银行、购物以及在健康或财务指导等关键领域进行信息搜索等都是重要的例子。在Chebbi和Jebara的研究中[84]，他们使用音频、视频和文本模式来自动区分欺骗和真实，并试图将它们合并以更精确地识别欺骗。首先分别检查了每个模态，然后提出了特征和决策级别融合策略来整合这些模态。所提出的特征级融合方法通过调查各种特征选择技术，从使用的全部特征集中选择最相关的特征，而决策级融合方法基于置信度理论，并考虑到每个模态的信息确信度。为了实现这一目标，我们使用了一组来自公开的美国法庭听证会的真实生活视频数据集，其中人们以真实或虚假的方式进行互动。

7.7. 压力检测

当今社会存在多种类型的压力，所有这些都对我们的心理和身体健康产生影响。压力被定义为从平静到兴奋状态的过渡，其目标是维持个人的完整性。压力可能对人的工作/生活效率产生多种不利影响，包括决策能力下降、情境意识降低和表现差。在Sander等人的研究中[85]，采用了重复测量的实验设计，同一组个体在两种噪声条件下工作，这些条件经过精心控制以模拟典型的开放办公室和私人办公室的噪声水平。在Mou等人的另一项研究中[86]，提出了一种通过基于注意力的深度学习技术进行多模态融合的驾驶员压力检测框架。具体而言，提出了一种基于注意力的卷积神经网络（CNN）和长短期记忆（LSTM）模型，用于融合非侵入性数据，包括眼动数据、车辆数据和环境数据。随后，该模型可以自动从每种模态分别提取特征，并通过自注意机制对来自不同模态的特征给予不同的注意程度。

7.8. 多模态鲁棒系统

创建多模态的鲁棒系统是一项艰巨的任务。一些基准研究试图提高基线结果并使系统更具鲁棒性。人工智能在需要常识推理的复杂任务（如自然语言理解）方面仍然面临困难。在Cambria等人的研究中[94]，他们创建了一个基于常识的神经符号框架，旨在解决情感分析的相关问题。具体而言，他们采用了无监督和可重复的子符号技术，如自回归语言模型和核方法，以构建可靠的符号表示，将自然语言转换为一种原型语言，并以完全可解释和可解释的方式从文本中提取极性。在Zou等人的另一项研究中[95]，应用多个理论创建了一个更具鲁棒性的系统。从信息论的角度开发了一种新颖的多模态融合架构，并使用光学雷达（LiDAR）相机融合网络演示了其实际效用。首次创建了一个多模态融合网络作为联合编码模型，其中每个单个节点、层和流水线都表示为一个通道。在另一项研究中[96]，大多数多模态情感分析的研究都是在训练和测试数据集中存在说话者重叠的情况下进行的。由于每个个体在表达情绪和情感方面都是独特的，因此必须找到适用于情感分析的通用、与个体无关的特征。然而，在存在重叠的情况下，模型已经观察到了某个特定个体的行为，结果无法真正推广到真正的泛化。在实际应用中，模型应对个体差异具有鲁棒性。

8. 结论

在过去的十年中，对多模态情感分析（MSA）的研究取得了进展。特别是，研究表明它在情感预测和情绪识别方面的有效性。已经确定了许多重要的贡献，比如改进MSA融合方法以提高效率和性能。报道了MSA取得进展的其他领域，包括：双模态或三模态中模态数量的变化；上下文感知和与说话人无关的幽默和讽刺检测；融合技术；架构中的应用特定修改以及各种学习算法和推荐系统的开发。这篇及时的综述总结了MSA架构的最新发展。根据融合类别，具体地识别了十个基本的MSA架构进展，分别是：早期融合、后期融合、混合融合、模型级融合、张量融合、分层融合、双模态融合、基于注意力的融合、基于量子的融合和词级融合。对多个MSA架构变体进行了研究，词级架构被确定为最有效的架构，它通过

使用视频中相邻话语的上下文信息对目标话语进行分类。该架构有两个组成部分，其顺序根据模型而变化。第一个模块是上下文提取模块，用于建模视频中相邻话语之间的上下文关联，并突出显示与预测目标情绪更重要的相关上下文话语。在最近的模型中，采用了双向循环神经网络模块。第二个模块是基于注意力的模块，负责融合文本、音频和视频这三种模态，并在上下文中选择最有用的模态。最后，还确定了一些关键的MSA应用、未来的研究挑战和机会。