亮点

通过自我反思促进社交媒体对话中的情绪调节

Akriti Verma,Shama Islam,Valeh Moghaddam,Adnan Anwar

- 提出一种模型，通过在情感表达持续处于高度激昂状态时鼓励自我反思，为在线对话中的即时注意力调节和回复调节提供支持。
- 提出一种基于图的框架，用于识别在线社交媒体对话中进行情绪调节的必要性。
- 提出了在社交媒体应用中融入支持用户进行情绪调节的设计启示。

通过自我反思促进社交媒体对话中的情绪调节⋆，⋆⋆

阿克丽蒂 维尔玛[a]，*，沙玛 伊斯兰[a]，瓦莱 莫加达姆[b] 和阿德南安瓦尔[b]

[a]迪肯大学工程学院，澳大利亚吉朗，3216[b]迪肯大学信息技术学院，澳大利亚吉朗，3216

文章信息

关键词：数字情绪调节 情感计算 人机交互

摘要

社交媒体平台上的匿名性让用户藏身于键盘之后。这使用户无需承担责任，从而参与网络愤怒、仇恨言论以及其他有害于网络福祉的文本型毒性内容。数字情绪调节领域的近期研究表明，沉溺于网络有害内容可能是情绪调节无效所致。我们认为，通过让用户了解自身行为的后果，可以减少这种现象。以往的数字情绪调节研究主要探讨数字情绪调节实践、利用多模态传感器识别情绪调节，以及鼓励用户在在线对话中负责任地行事。虽然这些研究深入揭示了用户如何有意识地利用数字媒体进行情绪调节，却未能捕捉在线情绪调节的情境动态。本文通过交互设计提供一种情绪调节支持干预方案，介绍了一种新方法，用于识别在线对话中情绪调节的需求，并以融入日常生活的方式向用户传递信息。通过在情感表达加剧期间促进自我反思，我们提出了一种基于图的框架，为在线对话提供即时情绪调节支持。研究结果表明，在对话中使用该模型有助于识别其中具有影响力的讨论串／节点，从而定位毒性内容集中的位置，并将其减少多达 12%。这是数字情绪调节领域首项通过引发自我反思和隐性情绪调节来关注学习迁移的研究。

1. 引言

有意识地改变自身情感状态的做法称为情绪调节。成功进行情绪调节的能力对于有效应对日常生活、在日常互动中采取恰当行动，或单纯为了获得愉悦感，都是至关重要的Wadley、Smith、Koval 和 Gross(2020年）。心理学研究、认知行为研究以及心理健康领域都对这一主题进行了深入探讨 McRae 和 Gross (2020年),Gross(2015),Gross(2014年）。随着技术发展，以及人们更容易接触到数字媒体并便捷地获得多种选择，近来，通过数字媒体调节情绪的做法迅速普及。

数字技术提供了更多策略选择，而且这些策略可以轻松有效地付诸实践。人们会结合使用各种应用程序和设备，有目的地管理日常生活中的情绪Smith、Wadley、Webber、Tag、Kostakos、Koval 和 Gross (2022年）。例如，锻炼时听振奋人心的音乐、下班后观看喜剧或轻松幽默的视频来缓解压力、感到孤独时玩社交电子游戏，或浏览社交媒体应用来应对无聊。人们每天会多次广泛使用社交媒体应用。这些应用具有多种情感可供性（可表达、可分享、可消费和可评估），所有这些都可能影响情绪以及与情绪相关的行为Steinert 和 Dennis (2022年）。由于使用广泛，社交媒体应用上的活动会显著影响网络福祉。近年来，人们广泛观察并研究了在线对话中毒性和仇恨言论的普遍现象，并发现它们是病毒式传播的关键因素Maarouf、Pröllochs 和 Feuerriegel(2022年）、Goel、Anderson、Hofman 和 Watts(2016年），病毒式传播是衡量帖子传播范围的指标。社交媒体对话由连接性行动和快速的信息传播推动，催生了在线运动和辩论，其结果也影响了线下事件 Saveski、Roy 和 Roy (2021年）、Mirbabaie、Brünker、Wischnewski 和 Meinert(2021年）。近年来，人们开始公开谈论在线遭受的仇恨如何影响日常生活，并质疑自身在网上的安全。政界人士、新闻网站以及年轻内容创作者等发布的帖子，都是这种现象的受害者。研究发现，如今在线遭遇或应对不尊重或粗鲁的行为已被视为常态，是使用社交媒体必须承受的一部分 Thomas、Kelley、Consolvo、Samermit 和 Bursztein (2022年）。一些社交媒体应用实施了规则，封禁粉丝数和参与度较高的账号，以遏制攻击性言论的传播，但毒性仍然存在，因为它是由许多轻度有毒者的行为造成的，而非少数高度有毒者所致Saveski 等人(2021年）。

⋆

⋆⋆

*通讯作者vermaakr@deakin.edu.au (A. Verma)；shama.is@deakin.edu.au (S. Islam)；valeh.moghaddam@deakin.edu.au (V. Moghaddam)；adnan.anwar@deakin.edu.au (A. Anwar)

ORCID（们）：0000-0003-3963-0870(A. Verma)

必须确保社交媒体平台能够通过管理数字福祉面临的脆弱因素，为健康的互动与交流提供安全空间。此前关于社交媒体与情绪调节的研究，主要关注社交媒体如何促进在线情绪及其对幸福感的影响 Goldenberg 和 Gross (2020年）、Steinert (2021年）、Yang、Liu、Li 和 Shu(2020年）、Yue、Zhang 和 Xiao (2022年）、Shi、Koval、Kostakos、Goncalves 和 Wadley (2023年）。这些研究包括探讨适应不良的情绪调节方式如何影响社交媒体和智能手机的 problematic 使用，以及确定明智的智能手机使用如何改善有效的情绪调节和社交能力 Zsido、Arato、Lang、Labadi、Stecina 和 Bandi(2021年）。此前的研究还发现，根据自动性以及社交媒体使用的具体情境，主动和被动使用社交媒体都可能成为拖延或恢复活动Hossain、Wadley、Berthouze 和 Cox (2022年）。Ments、Treur、Klein 和 Roelofsma (2021年）提出了一个二阶自适应脑网络模型，用于模拟社交媒体中的情绪调节过程，并发现，尽管某些情绪调节策略在短期内具有保护作用，但持续使用这些策略会导致情绪恶化、幸福感相对较低。另有研究考察了社交媒体应用的用户界面，并提出设计框架，通过自动识别情感因素（例如受众对用户评论的预期情绪反应），帮助用户调节情绪，打破在社交媒体平台上发布令人不快评论的习惯Kiskola、Olsson、Väätäjä、H. Syrjämäki、Rantasila、Isokoski、Ilves 和 Surakka(2021年）。

尽管这些进展让我们深入了解了情绪调节过程在社交媒体应用程序中如何展开，但目前仍缺乏可供实施的数字化解决方案。现有的情绪调节工具包括基于情绪的推荐系统和提醒，但它们只能提供暂时的帮助，而且难以融入日常生活Wadley 等人(2020年），Slovak、Antle、Theofanopoulou、Roquet、Gross 和 Isbister(2022年）。此外，目前缺乏用于综合情绪调节的通用原型，因为数字情绪调节领域近期的大多数研究都基于实地研究、生态瞬时评估（EMA），或结合面部数据的生理传感器，这使得相关方法难以扩展和复现以供进一步研究Ruensuk、Cheon、Hong 和 Oakley(2020年）。我们还需要了解如何识别在线环境中对情绪调节的需求。这需要创建一种解决方案，让用户了解自己对某条帖子的细微影响，而不是笼统地告知“可能会造成什么后果”。了解自身影响会使用户质疑自己在帖子中的匿名性，并促使他们负责任地行事。

本研究提出了一种创新方法，用于在社交媒体对话中传递调节自身情绪的必要性相关信息，并引导用户经历情绪调节过程，旨在将这种学习通过反复应用融入他们的生活，从而提升网络福祉。由于本研究使用公开可用数据开展，因此为进一步拓展、延伸、比较和对照提供了基础。因此，本研究的主要贡献如下：

- 我们提出一种模型，通过在情感表达持续处于高度激昂状态时鼓励用户进行自我反思，为在线对话提供即时的注意力与反应调节支持。
- 我们提出一种基于图的框架，用于识别在线社交媒体对话中对情绪调节的需求。
- 我们提出了设计启示，探讨社交媒体应用如何纳入对用户情绪调节的支持。

本文其余部分安排如下：首先讨论近期的相关研究进展，随后介绍问题陈述和所提出的方法。然后，我们详细阐述实验分析与性能评估，最后作出总结。

2. 相关工作

2.1. 数字情绪调节

近期的数字情绪调节研究可分为三类：观察性研究、新型数字情绪调节工具，以及利用多模态传感器进行情绪调节识别。观察性研究考察人们如何使用视频流媒体平台、讨论论坛、在线游戏和音乐等社交媒体应用来调节情绪、思想和行为Smith 等人(2022年)、Shen 和 Cox(2020年）。研究人员开展了多项日记研究，以便分别深入了解日常情绪调节的各个方面，例如利用社交媒体缓解思乡情绪，以及大学生使用音乐流媒体平台的情况Wadley、Krause、Liang、Wang 和 Leong(2019年）。研究还考察了人们在社交媒体应用上的多任务处理和被动滚动浏览习惯，揭示了人们如何主动暂离社交媒体以减轻负面情绪或维持内心平衡，并指出讨论论坛中存在人际情绪调节行为 Hossain 等人(2022年)、Lukoff、Yu、Kientz 和 Hiniker(2018年）。大多数关于数字媒体如何用于调节情绪的研究，都依赖自我报告问卷或日记记录法Smith 等人(2022年)、Shen 和 Cox(2020年)、Tag、van Berkel、Vargo、Sarsenbayeva、Colasante、Wadley、Webber、Smith、Koval、Hollenstein 等人(2022a)、Wadley 等人(2019年）。这类数据收集要求参与者在指定时间段内记录自己与情绪调节相关的活动以及技术使用情况，之后在访谈中讨论自己对技术的使用，以及伴随而来的情绪反应。通过限制可记录的数据量，这种方法使参与者能够追踪并反思其工作和社交生活中的重要发现

表 1情绪调节数字化干预：根据其支持的策略进行分类

情绪调节策略类别	新型工具
情境选择：避免可能引发不愉快情绪的情境。	•基于增强现实的情绪调节技术将情感呈现为有助于人际情绪调节的互动状态Semertzidis, Scary, Andres, Dwivedi、Kulwe、Zambetta 和 Mueller（2020年）。 • 在线对话情绪调节用户界面设计中的批判性声音……Kiskola 等人（2021年）Ķou 和 Gui （2020年）。
情境修改：修改a sce-景的特征以改变其情绪影响。	• 基于机器人的人工共餐伴侣，以促进进食时的社交互动对于想要或不得不独自进食 Mancini、Niewiadomski、Huisman、Bruijnes 和 Gallagher（2020年）。基于互动的技术，使用户能够提升积极情绪Li、Hao 和 Yoon (2020年) • 用于提升任务表现和情绪的声音景观 Newbold, Luton、Cox 和 Gould (2017年)，Yu、Hu、Funk 和 Feijs (2018年)。
注意部署：聚焦于情境中的情感元素，或将注意力从中移开，以便在情境中聚焦或避开情感元素，从而唤起预期的情绪。	• 用于即时舒缓的可定制的虚拟现实环境，为开放式工作场所提供支持Ruvimova、Kim、Fritz、Hancock 和 Shepherd (2020年). • 通过照片中的表达和意图来实现平静／增强期望的情绪 Chen, Mark 和 Ali （2016年）。
认知改变：重新评估情境，以重塑其情绪效应。	• 个性化呼吸节拍器，通过诱发明确的情绪调节，主要用于分心/重新评估 Miri, Jusuf, Uusberg, Margarit, Flory、Isbister, Marzullo 和 Gross（2019年）。 • 用于认知、生理的基于触觉的智能手表干预，以及行为变化Costa、Guimbretière、Jung 和 Choudhury (2019年).
回复调节：将当前的情绪反应或表达转变为更理想的状态。	• 帮助学生通过创新玩具提升情绪调节实践并通过反复互动塑造内隐情绪信念Theo-fanopoulou、Isbister, Edbrooke-Childs、Slovák 等（2019年）。 • 基于触觉的引导式呼吸和宜人气味以促进更安全的驾驶 Paredes, Zhou, Hamdan, Balters, Murnane, Ju 和 Landay (2018年), Dmitrenko, Maggioni、Brianza、Holthausen、Walker 和 Obrist（2020年）。

以及社交生活。总体而言，这些研究表明，人们在日常生活中会使用各种数字技术来调节情绪，凸显了设备中所集成技术的重要性，以及促进线上身心健康的必要性Wadley 等人(2020年)。

第二类包括为数字情绪调节设计和开发新工具 Kiskola 等人（2021年）、Kou 和 Gui （2020年）、Smith 等人 （2022年）。这一研究领域侧重于创建旨在支持、改善或引导情绪调节技能的干预措施，或帮助个体在具有挑战性的情境中运用这些技能。情绪调节过程分为四个阶段：认识到情绪调节的必要性或意识到进行情绪调节的意愿，选择适当的策略，实施该策略，然后监测调节后的状态，以确定是否需要进一步调节 Wadley 等人 （2020年）。技术支持的这些干预措施要么旨在辅助采用特定的情绪调节策略，要么旨在提高用户在情绪识别或监测阶段的情绪觉察 Smith 等人(2022年)，Slovak 等人(2022年）。表-1 列出了近期根据所支持的情绪调节策略分类的情绪调节干预措施。这些干预措施包含体验式设计元素，主要依靠生物反馈或隐式目标回复，通过触觉接触无意识地引导用户进入特定的生理状态，例如模拟心率，以减少焦虑、提升表现 Miri 等人(2020年)，Newbold 等人(2017年)，Mancini 等人(2020年)，Paredes 等人(2018年）。近期进展还推动了教学式干预组件的发展，这类组件依靠基于提醒的推荐系统，例如向用户推荐特定的情绪调节技术，并通过提示用户审视自己现在或过去的感受来鼓励情绪觉察Chen 等人(2016年)，Costa 等人(2019年）。这些

表2旨在识别在线环境中情绪调节的近期研究综述

标题	**目的**	**所用数据/ 参数 受监测**	**可用性 数据集的**	**数据集大小 数据集**	**可用性 一种可复现的 模型**
情绪轨迹 在智能手机使用中的： 迈向识别 情绪调节 在自然环境中 Tag、Sarsenbayeva、Cox、Wadley、Goncalves 和 Kostakos （2022b）	呈现一项 实地研究的发现，该研究测量了 喜悦如何在 日常智能手机使用过程中展现。	使用.....收集 定制智能手机 应用程序进行追踪 身体（面部）特征	不可用	这项研究 参与了 20名参与者， 开展了 为期14天	NA
基准测试 商业情绪 检测系统 使用逼真的 面部失真 图像数据集 Yang、Wang、Sarsenbayeva、Tag、Dingler、Wadley 和 Goncalves（2021年）	评估性能 商业情绪 检测服务	使用了 3 个基于面部 表情的数据集 数据集	在线使用 数据集 （ADFES） （RaFD），以及 （WSEFEP）	共838张图片 张	NA
行为与 生理信号—— 为基础的深度多模态 移动情绪识别方法 Luo andYang (2021年)	提出一种新颖的注意力 机制 LSTM 系统，该系统 结合使用 智能手机上的传感器（前置摄像头、 麦克风、触摸面板） 并且 腕带	收集了一个数据集 使用智能手机 应用程序，其中 行为和 生理 参数已 纳入观察	不可用	这项研究 涉及45名 参与者	描述于 论文中
你感觉如何 在线？利用 智能手机传感器 检测短暂的 社交期间的情绪 媒体使用 Ruensuk 等人（2020年）	探索人们情绪的识别 人们的情绪 当他们使用社交 媒体应用程序	收集了一个数据集 使用各种 （身体）动作 /眼动追踪 应用程序	不可用	该研究 涉及20名 参与者	NA
鼓励情绪 调节在社交 媒体对话 通过自我反思	**识别需要情绪调节的 情境，并 为其提供即时 对同一内容的支持**	**所使用的数据 来自 Twitter 对话**	**公开 可用**	**规模 该数据集： 180K**	**详见 该论文**

这些研究通过考察此类设计对用户的影响，探索数字情绪调节（DER）的新设计机会，并为这一研究领域提供了全新的方向。

第三类研究旨在通过多模态传感器识别并捕捉数字情绪调节（DER）过程。除了个体调节自身情绪的意愿之外，情绪调节过程还包括其周围环境、情绪的情境触发因素，以及他们调节该情绪的尝试。由于在实验室环境中测量这些特征颇具挑战，研究人员开始探索在自然情境下识别这些特征的方法。研究人员单独或组合使用智能手机的前置摄像头、触摸传感器、眼动追踪器和运动传感器，通过面部表情观察情绪变化Luo andYang (2021年）、Ruensuk 等 (2020年）。近期一项研究使用设备的前置摄像头，测量人们在每个手机使用时段中的愉悦程度，并根据他们在时段开始时感受到喜悦的可能性将其分为三组Tag 等 (2022b）。另一项研究采用图像操控，模拟捕捉人物表情时实际发生的图像失真，以进行基于面部表情的情绪的检测与调节 Yang 等 (2021年）。研究表明，结合不同模态能够提高社交媒体任务中的情感检测准确率。表2 总结了近期旨在利用传感器和数字媒体识别情绪调节的研究。可以看出，这些研究大多依赖专门为研究收集的数据，因此难以进一步拓展。因此，本文提出了一个框架，利用公开可用数据识别并支持情绪调节。

2.2. 分析社交媒体对话

社交媒体应用程序日益受到欢迎，导致我们生活中的参与活动在很大程度上实现了虚拟化。数字化也彻底改变了传统的休闲和娱乐方式，因为这些应用程序为表达和消费提供了多种维度。社交媒体提供了促进强劲在线互动的平台；遗憾的是，其潜力往往受到恶劣言论和反社会行为所引发的毒性的阻碍。现有许多研究论文已对社交媒体平台中的毒性进行了广泛研究。Saveski 等人研究了

有害对话的个体和群体结构，并发现，尽管在个体层面，毒性分布于大量低度至中度有毒的用户之间，但在群体层面，有害对话具有规模更大、更广泛且更深的回复树，而关注图则较为分散。他们还注意到，与发帖者没有社会联系且共同好友较少的用户，更有可能发布有毒回复Saveski 等人（2021年）。Solovev 等人研究了社交媒体上病毒式传播、错误信息和谣言传播的动态，发现源推文中编码的道德情绪会影响社交媒体上虚假谣言的传播Solovev 和 Pröllochs（2022年）。Mirbabaie 等人依据连接行动理论，通过将有影响力的用户分类为不同角色，并识别连接行动发起者和维护者，研究了社交媒体在在线社会运动中的重要性Mirbabaie 等人（2021年）。这些发现帮助我们理解促成有毒的在线行为的社会因素，并有助于创建更健康的社交媒体平台。他们还提出，展示对话的社会背景或许能够鼓励人们在网上采取更尊重他人的行为。

2.3. 分析社交媒体对话图中的情绪

研究人员已广泛运用图来研究基于文本的社交媒体互动，其中每个节点代表一位用户或用户发布的一条评论，边则反映用户之间或评论之间的关联。这一过程通常包含两个部分：第一部分是对节点中文本进行情绪分类；第二部分是依据节点的结构分析其连通性和影响力。文献中已有若干基于自然语言处理（NLP）的算法，可用于识别文本所表达的情绪Poria、Majumder、Mihalcea 和 Hovy (2019年），Chowanda、Sutoyo、Tanachutiwat 等人(2021年），Majumder、Poria、Hazarika、Mihalcea、Gelbukh 和 Cambria (2019年）。这些模型旨在识别 6 种基本情绪（爱、喜悦、悲伤、惊讶、愤怒和恐惧），以及 27 种次级情绪Shaver、Schwartz、Kirson 和 O’Connor (1987年）。NRC 词典是一个众包词库，收录了 27,000 个英语词语及其与八种基本情绪（愤怒、恐惧、期待、信任、惊讶、悲伤、喜悦和厌恶）以及两种情感倾向（负面和正面）的关联。如今，它被用作基准，以衡量一段文本的情绪效价 Mohammad 和 Turney(2013年）。

情感分类结果被输入图中，并用作节点属性，可用于统计分析（例如中心性度量、与根节点的距离以及出现次数），也可用于图神经网络（GNN）。个体用户影响力、人气和社会影响力通常通过基于统计属性的分析来衡量，而基于神经网络的方法则分析整个社交图，以了解其结构和高层次的网络动态 Antonakaki、Fragopoulou 和 Ioannidis （2021年）。在他们提出的隐式情感分析模型中，Yang 等人使用图注意力卷积神经网络传播语义信息，并利用注意力机制计算词语对情绪表达的贡献Yang、Xing、Li 和 Chang （2022年）。Perikos 等人提出了一种基于图的方法，根据情绪对主题的情感程度进行建模，随后创建主题的情感图，以展示公众对该主题的感受和情绪Perikos 和 Hatzilygeroudis（2018年）。Brambilla 等人提出了一种新方法，通过使用有向多重图网络，从帖子和评论的层级结构入手，提取在线对话中的常见模式，并构建和理解用户之间的交流模式Brambilla、Javadian 和 Sulistiawati（2021年）。这些研究表明，通过将社交媒体对话建模为图，可以探究其结构特征和动态特征。

2.4. 识别需要进行情绪调节的情境

不尊重他人的评论是网络有害内容的重要成因。近年来，内容审核被视为一种干预手段，可通过识别基于关键词或潜在情绪的不文明言论，提升网络福祉并减少网络有害内容Thomas等人 (2022年),Jhaver、Boylston、Yang 和 Bruckman (2021年)。研究者也探索了基于机器学习的审核方法的可扩展实现 Gorwa、Binns 和 Katzen- bach (2020年), Gillespie(2020年)。算法解决方案可用于向用户展示已发布评论的情绪分析，这可能促使一些用户重新考虑自己的帖子。例如，Perspective API谷歌 (2021年), 它会根据一段文本检测有害文本，并给出百分比得分。在西班牙语新闻网站 El Pas 的评论撰写系统中实施后，该工具被发现适度提升了讨论质量。互联网礼仪和“言论自由”准则可能含糊不清，因此此处的目标是让用户了解其评论的影响。这将提升他们共情其他网络用户的能力，并触发隐性情绪管理 Walther(1993年)。我们认为，可以通过向用户提供事实性提示，来应对计算机媒介文本交流中有效情绪调节所面临的挑战。正如 Kiskola 等人的建议，支持通过自动识别情感元素来实现情绪调节；此外，Slovak 等人强调，需要将教学式学习融入参与者生活的数字情绪调节干预措施 Kiskola 等人(2021年),Slovak 等人(2022年)。近年来，隐性情绪调节这一主题受到科学文献的关注。与涉及有意识地抑制情绪反应的显性情绪调节不同，隐性情绪调节是自动发生的，也可能实现自动化 Torre 和 Lieberman (2018年)。因此，在本研究框架内，隐性情绪调节是一种潜在的设计方法。情绪标记可以增强情绪调节，它只是让对话中情绪色彩浓厚的方面更加

显而易见。Kiskola 等人通过介绍和讨论旨在促进自我反思式情绪调节的系统，为潜在解决方案提供了重要观点 Kiskola 等人 (2021年)。在这项工作中，我们提出告知用户其言论对对话造成的情绪影响。因此，通过呈现评论所导致的“已经发生了什么”，本文提出了一种替代方案，不再依据评论预测“可能发生什么”并据此删除评论。

3. 问题陈述与拟议的方法论

3.1. 问题陈述

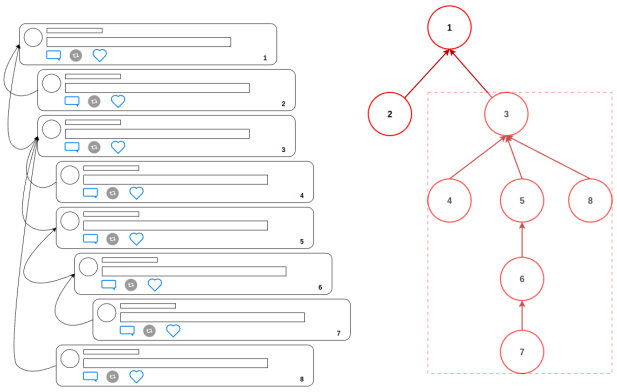
在线对话中的仇恨言论涉及众多平台、用户角色、参与度的发展过程及其维持因素，这些方面均已得到广泛研究 Solovev 和 Pröllochs (2022年), Saveski 等人 (2021年),Majó-Vázquez、Nielsen、Verdú、Rao、de Domenico 和 Papaspiliopoulos(2020年), Guberman、Schmitz 和 Hemphill (2016年), Konikoff(2021年). 近年来，将情绪调节视为仇恨言论的一个组成部分已逐渐流行。在这项研究中，我们旨在针对在线对话中由数字情绪失调引发的仇恨言论。由于情绪调节是一个受情境影响的主观过程，而在线情境并非总能被推断出来，因此我们提出，通过分析人们的在线行为，可以识别出需要情绪调节的情境，随后通过提供鼓励自我反思的信息，在这一过程中给予支持。尽管我们无法确定用户在任何特定时刻是否正主动尝试调节自己的情绪，但我们认为，借助设计提示和带标签的事实分析（而非预测）进行隐性情绪调节，将使用户更容易理解自身行为的影响，并考虑发布经过审慎思考的评论。因此，本研究试图回答以下研究问题：

- 我们如何判断在线对话何时需要进行情绪调节?
- 除了模式检测之外，我们如何为在线对话中的情绪调节提供即时支持?
- 我们如何利用在线环境，将高效的数字情绪调节作为一种可迁移的技能加以推广，而不是单纯传递说教式信息？我们如何将有效的情绪调节学习拆解，并针对每位用户进行定制?

3.2. 术语与定义

本节介绍我们在实验和分析中使用的术语、关键词和定义。

- 推文/帖子：这是原始推文，是所分析对话的源节点。它也是图中的根节点，后续将用于分析对话，并对应图 6 (b) 中的节点 1。



（a）Twitter 对话示例　（b）Twitter 对话示例

图

图 1： 同一内容的 Twitter 对话（a）及其对话图（b）。在（b）中，节点 1 是根节点，代表源推文/帖子；节点 2 和 3 是评论节点；节点 4、5、6、7、8 是回复；虚线框包含由节点 3 发起的回复树

- 评论/回复：评论是对推文或原始帖子的直接回复，在图 6 (b) 中由节点 2 和 3 表示。
- 回复：回复是针对评论所作的评论，也就是说，它不是对源推文的直接回复。图 6（b）中的节点 4 至 8 即为回复。
- 回复树：回复树是由评论及其回复构成的讨论串。图 6（b）中的虚线框表示该回复树。
- 对话：对话包括推文/帖子及其所有评论、回复和回复树，由图 6（b）中的图表示。
- 情绪板：由六个元素组成的键值对，其中键表示情绪，值为浮点数，代表帖子/推文/根节点所表达的每种情绪的累积比例。
- 影响力：根节点的情绪板代表其子节点情感表达的组合。因此，除根节点外的每个节点都会根据其在图中的位置及其文本所承载的情绪，对根节点产生影响。节点的影响力由本文第3.6节所述的方程1给出。

3.3. 方法学框架

图中展示了用于鼓励社交媒体对话中情绪调节的eImpact框架。2该框架由四个关键组成部分构成：数据检索、对话情绪分析、对话图分析，以及识别情绪调节的情境。数据检索过程从收集社交媒体对话的信息开始，本例中收集的是 Twitter 对话。近年来，

Twitter 一直是基于标签的社会运动（如 #我也是和 #黑人的命也是命）的热门平台，但该平台的言论自由政策也增加了仇恨与骚扰的风险。因此，我们收集了各种 Twitter 对话，并将其保存为 CSV 文件。我们使用特征工程创建了一组文件，每个文件都包含一段对话；文件中的每一行由一段表示推文或评论的文本，以及相应的 ID 和元数据（例如收到的评论数量、回复者等参数）组成。第二个组成部分随后使用这些 CSV 文件分析情绪传播。首先对推文中表达的情绪进行分类。情绪分为 6 个基本类别和 27 个次级类别。本研究使用基本情绪类别对推文中的情绪进行分类。随后，我们将推文划分为 6 种情绪类别，并据此生成对话图。该图用于计算单条推文对整个对话的情绪影响，以及讨论中各种情绪的百分比分布。接着，在最后一个组成部分中，我们利用该图找出对话情绪影响最大的节点，并据此识别需要进行情绪调节的情境，同时提供相应支持。

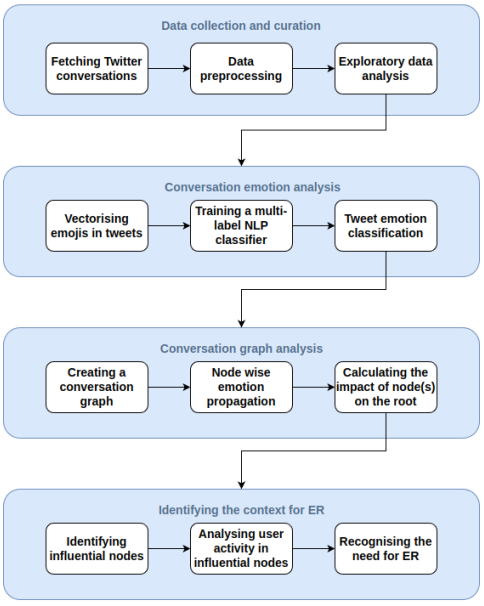


图 2： eImpact：鼓励在社交媒体对话中即时进行情绪调节的框架

3.4. 数据

澳大利亚政府官员经常使用 Twitter 发布最新消息、通报近期事件或告知民众即将开展的活动。本研究分析了 2020 年 4 月至 2022 年 8 月期间，来自澳大利亚六个州的议会议员所发布的推文。研究旨在收集涵盖不同主题、话题标签和背景的各类对话。这些推文涉及

表 3用于获取推文的参数

参数	描述
作者_标识符	The 唯一 标识符 of the user who 已发布 the 推文/评论/回复。
对话_标识符	用于标识对话的唯一标识符 推特上的讨论串 ter.
创建_于	时间戳 of the 推文/评论/回复 in UTC.
id	The 唯一 标识符 of the 推文/评论/回复 on Twitter
回复至用户 ID——	收到回复的用户的作者 ID 收到了该回复。
实体	提供元数据和额外的 有关推特 帖子的信息。例如， 话题标签， 用户提及、连接等。
lang	用于筛选英文推文的过滤器。
text	推文中包含的文本 以及表情符号。

各种封锁措施、新冠疫苗动态、政策更新、地方发展和公告。本次分析共选取了 75 个对话，每个对话至少包含 1000 条直接回复，回复总数达到 3000 条，因此数据集共计 75*3000 行。这些数据通过 Twitter API(Tweepy) 和Twitter 提供的推文下载工具下载。Twitter 上的每条推文都有一个对话 ID，可用于收集和整理推文、评论及回复。需要注意的是，这些对话中的每一个都是独立的推文，而不是对其他推文的回复或引用。表格3 介绍了获取数据时使用的参数。

本次实验仅考虑包含文本和表情符号的推文。包含图像、视频或外部链接的推文未纳入本次实验的考虑范围；媒体文件中表达的情绪将留待未来的研究进行评估。当前数据文件中的推文大多为英文，偶尔出现的其他语言评论则已被删除。数据以 CSV 文件的形式下载，每个对话对应一个文件，随后用于进一步分析。对话中的用户通过其作者_编号识别，并使用回复_至_用户_编号_ 将评论与其回复相关联。实体参数用于追踪回复的顺序。每条评论收到的回复数量，以及回复该评论的不重复用户数量，也被添加为数据属性。

3.5. 情感分类

CSV 文件中的每一行都包含一个文本字段，代表一条推文、对推文的评论或对评论的回复。该字段中文本所表达的情绪由文本情绪分类器判定。多标签

自然语言处理分类器用于根据六种基本情绪（爱、喜悦、悲伤、愤怒、恐惧和惊讶）对推文进行分类。推文中的表情符号被替换为由Gensim 使用 Emojinal 库生成的向量表示Barry、Jameel 和 Raza (2021年），之后使用TweetTokenizer对推文文本进行分词。该自然语言处理分类器使用“emotions”数据集进行训练和测试。该数据集是一个包含两列的带标签 Twitter 消息数据集，每条消息由文本字符串和标签组成，共有 20,000 行数据Saravia、Liu、Huang、Wu 和 Chen(2018年）。标签描述了六种情绪：爱、喜悦、悲伤、愤怒、恐惧和惊讶。训练时使用了一个包含双向长短期记忆层的四层序列模型，并将数据按 80:10:10 的比例划分为训练集、验证集和测试集。模型训练了 20 个训练轮次（增加训练轮次对准确率没有影响），测试准确率达到 87%。随后，使用训练好的分类器预测 Twitter 对话中表达的情绪。讨论串中的每条推文、评论和回复都会获得一个情绪标签和分数，其中分数表示分类器预测该情绪的概率。

3.6. 基于图的情绪传播分析

在这项研究中，我们提出，一条帖子或推文不仅代表其表达的情绪，也代表其评论和回复中表达的情绪。因此，我们通过汇总源推文、评论和回复所产生的情绪影响，计算对话所体现的整体情绪。我们生成一个图来表示对话分析结果。在线社交网络（OSN）中的网络通常定义为一个图，其中节点代表网络中的用户，边代表节点之间的连接Antonakaki 等人 (2021年)。这些图有助于识别用户属性（例如影响力）以及网络属性（例如同质性）。然而，本研究旨在分析一段对话（用户可能参与一次或多次），以及“用户行为对鼓励对话参与度的影响”。本研究并非旨在识别“问题用户”，而是希望找出对话中令人困扰的“帖子”（或在对话中触发愤怒／仇恨的帖子）。本研究没有考察用户在社交网络上的整体影响力或行为特征。我们认为，与笼统地将用户标记为不当相比，告知用户其粗鲁评论在特定情境下造成的影响，更能促使他们针对具体情形进行自我反思。

研究发现，公众人物、媒体公司和具有社会影响力的人士虽然能为社交媒体帖子带来大量初始关注，但真正让参与度持续下去、从而对对话中传播的情绪产生更大影响的，是那些缺乏影响力的匿名用户。Solovev 和 Pröllochs(2022年),Mirbabaie 等人 (2021年), Saveski 等人 (2021年). 告知用户他们对对话的影响，也会促使他们重新审视自己在帖子中的匿名感，并激励他们负责任地回复。因此，在本实验中，

表 4用于在对话图中查找有影响力的节点的参数

参数	描述
编号_条直接回复_	节点的入度。
收到的总参与度	某个节点的祖先节点数量 node, 某个节点的回复树中的节点数量
与根节点的距离_从_	节点的回复树。有向边的数量 在从某个节点出发的最短路径中节点到根节点。
Page_Rank	图中节点的排名 根据入边的结构 得出。
情绪_分数	情感分类模型的 情感分类 模型 预测文本情绪时所使用的 节点中文本的概率。

推文、评论和回复被视为图中的节点，而有向边则表示节点收到了回复。也就是说，如果一条评论收到了“n”条直接回复，它的入度就是“n”。图中也存在自环。因此，对话图 G 可定义为：

$G = (V,E, A)$，其中 V 是节点集合，E 是表示节点现有关系的边集合，A 表示属性向量集合。$n=|V|$ 表示顶点总数，$m=|E|$ 表示边总数，A（$A1$、$A2$、$A3$.....Ak）与 V 中的节点相关联，并描述其特征，其中 k 是每个节点所具有的属性数量。

在情感分类和在线社交网络（OSN）分析中，寻找图中的有影响力的节点已得到广泛应用。在 Twitter 网络分析中，尽管尚无公认的识别有影响力节点的标准，但研究人员已采用多种基于连通性/中心性的属性与机器学习方法的组合Berahmand、Haghani、Rostami 和 Li (2020年）、Vilarinho 和 Ruiz (2018年）、Ban Kirigin、Bu- jačić Babić 和 Perak (2021年）、Bordoloi 和 Biswas (2020年）。由于全局连通性并非衡量节点影响力的重要指标，因此在这种情况下，研究采用了侧重于识别局部影响力较大的节点的属性。这是因为，一条推文下的两条独立评论可以发展成大型讨论串，而不受其连通性或两者之间是否存在共同节点的影响。用于表示节点的属性见表-4。

直接回复的数量表示一个节点的入度，而讨论串中的回复总数表示一个节点的祖先节点（即有路径通往源节点的节点数量）。PageRank 用于对入度相同的节点进行排名；（在这种情况下，它优于介数中心性，因为它估算节点的局部重要性Antonakaki 等人（2021年））。与根节点的距离被用作衡量触发评论距离的一个因素；发起回复树的评论比回复树中的回复更具影响力

树。根据这些属性，使用以下规则计算根节点的子节点和终止节点的情绪影响力（Em）（另见图3）：

$G = (V, E, A)$ 中节点对根节点 R 的影响如下：

$$\forall V \in G - \{R\}, E_m(R) = \sum E_m(V1, V2, V3....Vn) \quad (1)$$

其中：

$$E_m(Vi) = f(Ai)$$

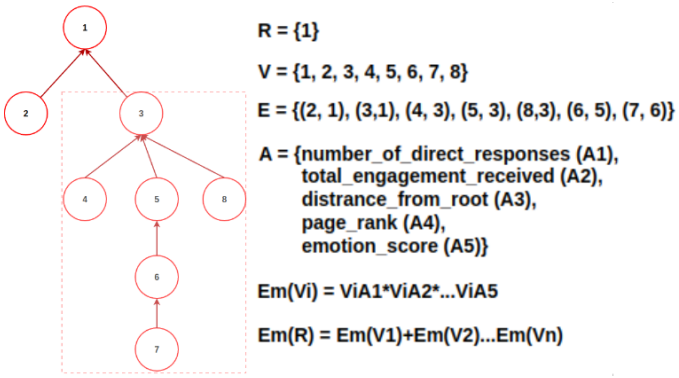


图3： 使用情绪传播计算根节点所代表的情绪

节点1是根节点，如图3所示3，它收到两条直接回复（2和3）。它还获得了7次互动（评论 + 回复）。节点2和节点3距离根节点最近，而节点7距离最远。因此，在计算节点3对根节点的情绪影响时，直接回复数为3，收到的总互动数为5，与根节点的距离为1，以及其他参数。该规则用于计算节点对特定节点的影响，在本例中该特定节点为根节点。根据节点的影响选择阈值，在本例中，阈值为节点对根节点影响的平均值。利用该阈值识别有影响力的节点，即影响值大于阈值的节点。在识别出根节点的有影响力节点*(I = {V1, V2...Vn})*后*(R)*，可以使用相同的规则，通过考察以有影响力的节点*($Iv1 = \{V1_1, V1_2...V1_n\}$)*为根节点的子图，找出对这些有影响力的节点影响最大的节点*(V1,V2...Vn)*。

3.7. 所提框架

为了估算对根节点影响最大的节点，我们将 eImpact 识别出的有影响力的节点与 Perspective API 指示的毒性分数较高的节点结合起来 谷歌 (2021年)。Perspective API 通过分析评论所涉及的多种情绪概念，预测该评论对对话的感知影响。它会返回一个概率分数（取值范围为 0 至 1）。该分数代表认为评论有毒的读者所占的百分比。为确保实验结果一致，我们根据分析中报告的平均值计算出毒性阈值（0.9）。Perspective API谷歌 (2021年) 也建议研究采用这一阈值；较高的阈值可降低出现偏差的可能性。分数越高，读者就越可能认为评论包含给定属性。Perspective API 目前不会考虑评论的上下文。我们采用 eI mpact 框架识别出的最具影响力节点与毒性分数指示的高度有毒节点的交集。

3.8. 支持即时情绪调节

以评论或帖子回复形式展开的对话，是社交媒体平台的重要组成部分。然而，这也增加了网络仇恨与骚扰发生的可能性。研究显示，41% 的美国成年人曾遭受网络仇恨与骚扰 Thomas等人 (2022年)。因此，本研究旨在通过量化仇恨言论对整体对话的影响，剖析对话中仇恨言论的发生情况，使用户了解自己对对话产生的微小影响，并鼓励他们负责任地参与互动。

识别对话图中的有影响力的节点后，可以利用相关信息，向参与评论节点回复树的用户说明其行为所带来的影响。还可以根据节点所代表情绪的颜色突出显示该节点，并展示情绪的百分比分布；或者限制对作为回复树起始节点的评论进行后续互动。如图-4所示，帖子中的评论及其相关回复树均按其所代表的情绪进行颜色编码，同时呈现这些情绪的分布，以及对该分布产生显著影响的有影响力的节点。节点 3、5 和 6 对对话中的愤怒情绪尤为关键，其中节点 6 的影响最大。因此，该节点被冻结。

在此处冻结一条评论，旨在识别某个用户或一群用户的行为，以免整篇帖子变得充满敌意，同时也通过揭示帖子中的有毒行为来对其进行剖析。需要注意的是，在此情境下抑制愤怒或仇恨言论，并不等同于社交媒体固有地偏向推广正面内容；这更像是一种尝试，旨在避免帖子引发过度仇恨，因为面对数字媒介引发的情绪挑战，人们很容易采取数字化行动来回应，例如在帖子下发表评论表达愤怒。Wang 和 Diakopoulos（2021年）、Smith 等人（2022年）。第一步是自我反思，这与强调情绪调节的认知与辩证行为疗法相似。通过说明用户先前行为所产生的影响来鼓励自我反思，将提升用户的意识，并促使他们负责任地行动。

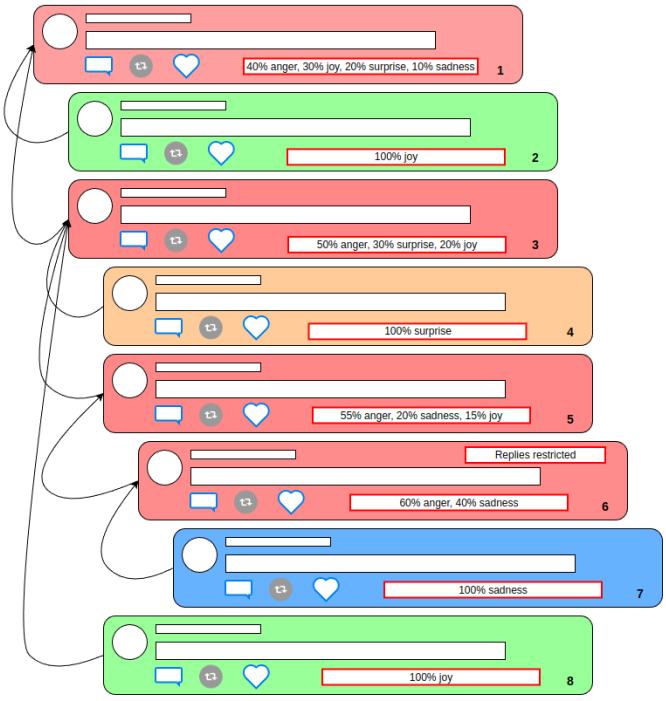


图4： 识别社交媒体对话中具有影响力的评论／回复／讨论串。评论根据其代表的情绪以不同颜色标示。

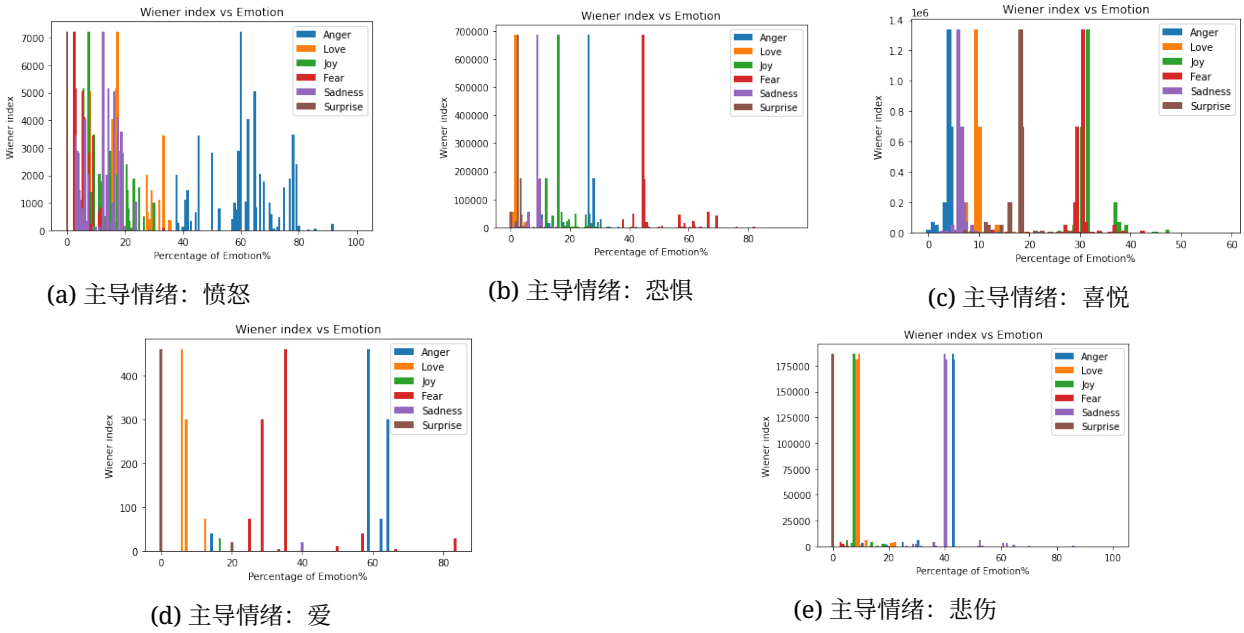
可以建议采用涉及认知改变的隐性情绪控制策略，例如重新评估或非评判性接纳，以帮助那些评论活动受到限制的用户，并防止产生类似的回复树。告知用户对话整体的情绪分布，有助于激发他们对其他参与者的共情，进而含蓄地促使用户感受到自身的责任。这种方法关注的是用户的行为，而非用户本人，因此可以避免用户因自身匿名性受到质疑而陷入焦虑。尽管限制评论下的后续活动是否会妨碍参与仍有争议，但这种做法在保护用户言论自由的同时只造成轻微的分心，而且用户仍然保有选择的自由。Kiskola 等人（2021年）。此外，研究发现，减少网络毒性的最佳方式之一是对在线内容进行审核Thomas等人（2022年）、Jhaver等人（2021年）。通过展示用户对对话及其情绪分布所产生的影响来质疑其匿名性，或冻结帖子中的特定评论（视帖子长度、涉及人数或情绪强度而定），都可以在这种情况下对毒性上升发出适度警示。

4. 情绪传播的实验评估与分析

社交媒体平台上的毒性表现形式多种多样，例如仇恨言论、骚扰、网络挑衅和网络欺凌。尽管这些有害行为的意图各不相同，但它们都遵循相同的路径：不文明行为，或者在基于文本的社交媒体平台上，使用不文明语言。近期研究探讨了如何利用 Twitter 对话检测骚扰和网络欺凌 Guberman等人(2016年)，Georgakopoulos、Tasoulis、Vrahatis和Plagianakos (2019年)，Pavlopoulos、Sorensen、Dixon、Thain和Androutsopoulos (2020年)。这些方法不仅需要上下文信息和用户资料数据，而且通常是在对话结束后才应用。社交媒体平台采用了取消平台访问和删除不当评论等审核策略，但这些策略基于静态规则，通常也是在事件发生后才实施Jhaver等人(2021年)。我们提出了一个框架，用于研究实时在线对话中情绪（通过语言）的传播。这种方法的潜力不仅在于判断对话是否正在变得有毒，还在于确定这种情况发生的原因和方式。我们认为，通过调查导致毒性的行为，该框架将有助于降低对话变得有毒的可能性。

在大量用户普遍使用不文明语言的社交媒体平台上，研究发现，基于仇恨言论的毒性是病毒式传播的重要组成部分 Maarouf等人(2022年),Goel 等人(2016年)。病毒式传播是指用户通过连接性行动迅速传播信息，且往往由其中蕴含的情绪推动。研究表明，仇恨言论和毒性都会在互动中不断升级，从而导致信息传播得更快、更广。因此，通过监测信息扩散的性质，以及对话图中情绪传播的累积效应，可以识别对话何时以及如何变得有毒，并在此之前采取内容审核策略。同样的方法也可用于评估何时需要进行情绪调节。此处的目标是让参与对话的用户了解其行为所产生的情绪影响，并通过反复应用这一方法，温和地引导他们树立负责任的心态。

Twitter 上的有毒推文被定义为不合理地粗鲁或不尊重他人的评论，导致某人离开对话。因此，在这项研究中，我们将对话中的毒性定义为：在原本中立的帖子中，有毒推文/回复不成比例或过量涌入 Aroyo、Dixon、Thain、Redfield 和 Rosen (2019年),Xenos、Pavlopoulos、Androutsopoulos、Dixon、Sorensen 和 Laugier (2021年). 如本文第 3.5 节所述，我们将推文划分为六种情绪（即愤怒、爱、喜悦、恐惧、悲伤和惊讶）。愤怒的涌入一直被



(a) 主导情绪：愤怒　(b) 主导情绪：恐惧　(c) 主导情绪：喜悦

(d) 主导情绪：爱　(e) 主导情绪：悲伤

图5： 影响力节点的维纳指数与回复树中情绪所占百分比的关系，其中主导情绪为 (a) 愤怒 (b) 恐惧 (c) 喜悦 (d) 爱和 (e) 悲伤

视为会引发狂怒、粗鲁或不尊重他人的表现，因此也被视为毒性的指标。

Twitter 对话的结构和毒性已得到广泛研究。我们借鉴文献中的研究结果，验证我们的框架是否适用。我们基于三个属性开展验证：结构特征、情绪与病毒式传播，以及连接性行动的发展。

4.1. 结构特征

用户可以通过点赞、分享或评论来回应推文/帖子。这些行为可能会引发其他用户的更多回复。由此形成的对话图中，回复树以原始推文为根节点。研究表明，随着对话图的规模、密度和宽度增加，其毒性也会增加Kanavos、Perikos、Vikatos、Hatzilygeroudis、Makris 和 Tsakalidis(2014年)，Saveski 等人(2021年）。表-5 展示了每种情绪所对应的影响力节点的回复树中，情绪平均分布情况以及参与其中的不重复用户数量；表-6 则展示了使用 eImpact框架计算情绪传播后，图中每种情绪分布的平均变化（增加/减少），以及影响力节点所代表情绪的初始占比。如表 6所示，随着对话规模扩大，最常表达的一种情绪会逐渐占据主导地位，这与最初的情绪分布不同。如果将回复树的毒性视为其收到的令人愤怒的推文所占的比例，就会发现该比例随回复树规模、宽度和密度的增加而上升。愤怒推文回复的百分比骤升至 40%，并持续上升，直至达到 80%（图-5(a)）。由于愤怒传播得比其他情绪更快，因此，在比较节点中包含的情绪占图中节点总数的百分比时，也能观察到相同的模式（表-6）。

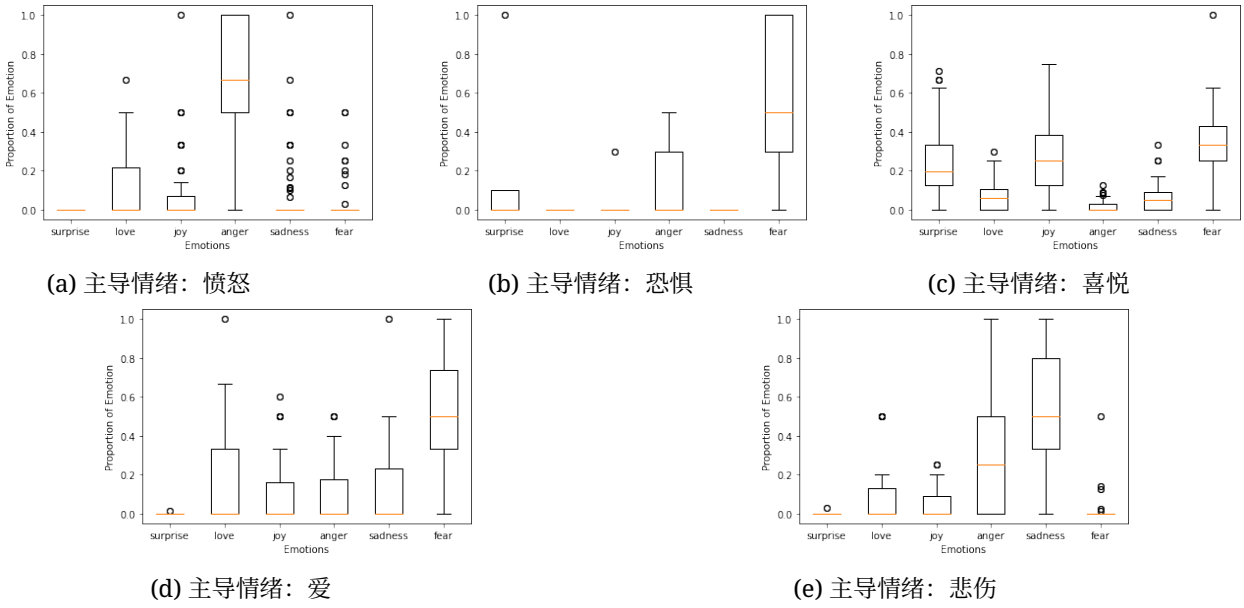
近期，维纳指数已被用于分析信息扩散的结构，具体而言，用于判断信息是以广播式还是病毒式方式传播Saveski 等人(2021年), Goel 等人(2016年)。回复树 T 的维纳指数 w(T) 定义为树中所有节点对之间的平均距离，如公式 2 所示。较低的维纳指数与扩散的广播式结构相关，在这种结构中，用户大多回复原始推文，用户之间的互动较少；而较高的维纳指数则与病毒式传播相关，在这种传播中，用户之间的往返互动多于对原始推文的回复。

$$w(T) = 1/n(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}, \qquad (2)$$

其中：

d_{ij} = 节点 i 和 j 之间最短路径的长度

在识别出对话图中的有影响力的节点后，我们根据其回复树中所包含的情绪百分比，计算了这些节点的维纳指数。图-5 展示了每种情绪下，有影响力节点的维纳指数（纵轴）与其回复树中各种情绪所占百分比（横轴）的关系。研究发现，愤怒的维纳指数呈现出最明显的上升趋势。起初该指数较低，随着回复树中愤怒所占百分比的增加而开始上升。当回复树节点中愤怒占比达到60%时，指数达到峰值，随后小幅下降并趋于平稳



(a) 主导情绪：愤怒　(b) 主导情绪：恐惧　(c) 主导情绪：喜悦

(d) 主导情绪：爱　(e) 主导情绪：悲伤

图6：主导情绪为（a）愤怒、（b）恐惧、（c）喜悦、（d）爱和（e）悲伤的对话中，有影响力节点的回复树里的情绪百分比分布

如图-5(a)所示，最终趋于平稳。这使我们得出结论：有影响力的节点最初会广播多种情绪；随着其回复树规模扩大且愤怒占比增加，它会转变为病毒式传播，在这种传播中，回复树中的节点会产生更多往返互动。此后，当愤怒占比达到60%时，由于该回复树中多个节点之间的往返互动，这些节点开始广播同一种情绪，因而转变为有毒对话，因为对话此时集中于少数节点，并导致愤怒传播。尽管如图-5所示，在其他情绪中也可以观察到对话从广播转向病毒式传播、再回到广播的模式，但其峰值比愤怒更早（35-40%）汇聚，且数值更高（10k）（b、c、d和e）。这突显出，愤怒通过汇聚式病毒传播和更早的广播，能够更快扩散。显然，负面情绪倾向比积极情绪倾向更快汇聚。这可能是因为积极情绪表达受到负面情绪的轰炸，导致回复树中的情绪分布发生变化，并使传播形式逐渐改变。如表-6所示，这是不同情绪下回复树中的平均回复数。这些发现与现有文献中的研究结果一致，从而验证了我们假设的适用性。

4.2. 情绪与病毒式传播

如今，在线内容的成功取决于其影响他人的能力。这也是发布在线帖子的动机之一，同时也是扩大个人影响范围的一个因素。对有影响力节点中的用户活动进行考察后发现，推文收到的回复数量与其中情绪的多少成正比

表5有影响力节点的回复树中的平均情绪分布及不重复用户数量

情绪呈现 有影响力节点代表的情绪	影响力节点的平均回复数 按其情绪分类（%）					不重复 用户数
	愤怒	Fear	Love	Joy	悲伤	
愤怒	78.6	4.2	5.3	4.7	5.4	3
Fear	31.3	56.8	4.2	1.3	6.4	5
Joy	37.4	1.4	15.7	43.1	2.3	7
Love	81.3	1.2	8.9	6.7	1.6	7
悲伤	32.8	13.2	8.7	2.4	42.6	4

表6使用 eImpact框架计算情绪传播后，图中情绪分布的平均变化

情绪呈现 根节点	平均变化 情绪分布（%）					初始百分比 of 节点数 情绪	节点数 在图中
	愤怒	Fear	Joy	Love	悲伤		
愤怒	27.6	-7.4	-12.8	-13.4	-4.1	38.3	3343
Fear	-5.3	20.1	-10.6	-4.2	0.4	21.9	5493
Joy	-3.2	4.2	15.4	-2.7	0.6	23.4	5073
Love	1.7	2.8	0.3	12.5	-3.4	20.8	4457
悲伤	-2.1	2.3	-2.4	-10.1	21.2	26.5	3623

在推文中。尽管表达愤怒和爱的推文收到的回复最多，但其中大多数回复都表达了愤怒。此外，表达爱的推文收到的回复最为愤怒，或最容易引发愤怒。如图-5(a)所示，当影响力节点表达中占比最高的情绪是愤怒时，随着回复树不断扩展，节点所代表的初始情绪分布会发生变化，愤怒开始占据主导地位，在大多数情况下达到整体表达的 70%–85%。当某个节点强烈表达爱时，回应树也会分支出愤怒。爱会收到大量极化且引发愤怒的回复（表-5），这导致愤怒

在回复树的情感表达中占据主导地位，如图-5(d)所示。

尽管其他情绪案例中愤怒所占的比例并不相同，但愤怒的病毒式传播在每种情况下都达到最高强度，尽管其比例相对较低。图-6展示了每种情绪对应的影响力节点的回复树中各种情绪的百分比分布（纵轴范围为 0–1）。如图-6所示，有影响力节点的回复树中情绪的百分比分布也印证了回复中表达愤怒的这一点。只有在喜悦占主导的情况下，愤怒才保持在最低水平。当愤怒是主导情绪时，喜悦和恐惧所占比例相近；除此之外，两者则呈反向关系。图-5 (e) 显示，悲伤虽然仍具有传播能力，但也会接收到大量愤怒，如图-6(e)所示。

进一步分析显示，这些回复是同一群用户反复回复的结果。研究发现，这些回复产生了两极分化效应，而认同该意识形态的用户反复参与似乎加剧了这一效应，最终导致人们对任何其他情绪表达产生敌意。文献中也观察到类似的两极分化：嵌入推文中的谴责他人的情绪会引发谣言级联，并可能加速其传播。这意味着，当对话中表达的激进思想和信念根深蒂固时，更有可能演变为毒性。当意识形态两极分化出现时，愤怒情绪的激烈混合会加速其在社交媒体对话中的传播 Solovev 和 Pröllochs(2022年)。

表6 还显示，与其他情绪相比，计算情绪传播后，表现出愤怒的节点百分比平均增幅最大。愤怒评论收到的愤怒回复与表达爱以外任何情绪的评论之间的比例也说明了这一点。在我们的研究中，所有情形和所有情绪均如此。表6 展示了传播后，对话图中节点所代表的情绪百分比增幅。众所周知，愤怒传播得更快，因此当它是主导情绪时，不仅出现比例最高，即使只涉及少数用户，它的增幅也最大。强烈表达愤怒／狂怒时，爱和喜悦等情绪也最容易被削弱（表6）。比较愤怒与其他情绪的情形时，也可以从表5所示的单个用户回复数量中看出相同模式。

4.3. 连接性行动的发展

社交媒体已发展成为社会运动的平台，这些运动属于连接性行动的范畴。用户与帖子互动，开启并维持了帖子活动。匿名用户维持着这些活动 Mirbabaie 等人(2021年), Saveski 等人(2021年). 从表格5&6可以观察到连接性行动的发展，

表7所提框架与 Perspective API 的比较谷歌 (2021年)

所用模型	识别 of 有影响力的节点	被识别为 有毒的节点的百分比	可能降低 毒性（若采取行动） 这些节点
eImpact框架	基于 影响力得分， 考虑到 推文文本 以及其 连通性	1-4%	10%
PerspectiveAPI 谷歌 (2021年)	基于 毒性，仅考虑 仅 推文文本	1-2%	7%
所提框架	考虑在内 其中的情绪 推文文本， 连通性 以及其毒性	1-4%	12%

在回复树中，少至3名用户的参与（回复树中的节点数量为112至270），就能使包含3000多个节点的图的情绪分布改变4%至27%。研究发现，帖子中的毒性源于无关紧要或不相关的内容或评论。记者、名人及其他公众人物可能会帮助推文获得关注，但他们更有可能发布宣传内容。普通个人或一般用户不仅能维持帖子的参与度，还会煽动仇恨言论或毒性。这可能是因为一般用户使用该平台的主要目的是表达自我并与他人建立联系。鉴于他们并非转推数最多或最受欢迎的人，情感表达便成为他们向公众传达观点的主要方式。因此，他们的匿名行为成为推动帖子的动力，这一点体现在表格5&6 中，具体表现为参与回复某个节点的不重复用户数量。平均只需3名不重复用户，就能使节点的传播方式从广播式转为病毒式，并使愤怒增加多达27%；但即便平均有7名不重复用户，也无法在表达的情绪为爱或喜悦时，将该节点的情绪分布改变至这一数值的一半。这也符合情绪动员的概念。情绪动员被认为是在人群的心理状态中发起社会运动所必需的 Castells (2015年),Saveski 等人(2021年). 在 Twitter 上发起 #我也是运动的那条推文，既表达了必要的愤慨，也寄托了对改变的希望。

因此，有必要以敏感且不压制言论的方式，处理社交媒体帖子中愤怒情绪诱发不平衡的问题。取消平台访问或禁止用户继续发表评论，或许能暂时减少帖子中的愤怒情绪，但从长远来看作用不大，因为这并未解决不负责任行为的根本原因。由于普遍存在的毒性是集体行动造成的，若分析情绪传播时着眼于用户整体，而非仅仅依据个别用户的行为，将有助于减少导致毒性的行为。因此，通过揭示用户行为的后果来削弱匿名性，将鼓励隐性情绪调节，促使用户自我反思并采取负责任的行为。

4.4. 性能评估

除了防止愤怒情绪不断累积，或仇恨言论逐渐演变为有毒内容之外，所提框架还考虑了毒性的主观性：它不单纯依赖帖子／回复／评论的内容，而是考量其对当前对话的影响。我们通过将该框架与 Google Perspective API 生成的毒性分数进行比较，对其进行了评估。谷歌(2021年)。随后，我们依据这一阈值，将推文／评论归类为有毒或无毒。在考察对话的毒性时，我们发现毒性最高的推文／评论也属于最有影响力的节点。我们还发现，有影响力节点的回复树包含了对话中超过 87% 的总体毒性。由于目标是通过分析对话中的有影响力节点来减少仇恨言论的诱发，我们使用 Perspective API谷歌 (2021年)的毒性分数来找出有影响力的节点。此后，我们结合 eImpact 的结果与毒性分数，作为所提框架来识别对话中的有影响力节点。我们发现，虽然 eImpact框架与 Perspective API 得出的有影响力节点有 75% 的匹配度，但所提框架得出的有影响力节点与毒性分数的匹配度为 94.3%。这证实，所提框架能够识别出可能导致帖子产生毒性的评论，同时兼顾其主观性以及这些评论本身是否有毒。该框架可与在线内容审核结合使用，同时考虑内容与上下文，从而了解对话中毒性的来源。表7比较了所提框架与 Perspective API 在识别有影响力节点以及降低毒性的可能性方面的表现。如表7所示，所提框架将有毒推文的识别率提高了 3%，并使毒性降低 5% 成为可能。这一提升源于在对话图中同时采用基于上下文和基于内容的节点评估。

我们还研究了通过停用有影响力的节点，或在这些节点达到毒性阈值后限制其进一步活动，能够在多大程度上减少仇恨言论。我们发现，eImpact框架可减少10%的仇恨言论，Perspective API可减少7%，而结合了eImpact框架与Perspective API毒性分数的所提框架谷歌 (2021年)可将毒性降低12%。这与以下认识一致：限制用户在某条评论下的活动，或许无法阻止他们在其他地方发表评论，或在同一帖子下开启新的讨论串；然而，由于限制措施会附带说明用户对帖子的影响，我们相信这将鼓励自我反思和隐性情绪调节。

5. 结论

研究人员已经关注到网络有害内容对心理健康造成的负面影响。尽管了解数字情绪调节未必能消除这些担忧，但这表明，某些数字媒体可以与建设性心理技术结合使用，以实现重要的实际目的，例如提升工作表现或促进社会和谐。基于文本的讨论和评论串是在线交流的重要组成部分，因此常常成为负面情绪滋生的温床。了解如何识别需要进行情绪调节的情境，并通过提供自我反思和情绪调节的机会来支持用户，或许能大幅减少引发狂怒的帖子，并改善网络福祉。本文提出了一种基于图的框架，用于识别对话图中以有影响力的节点形式出现的这类帖子。我们通过微观层面分析，聚焦于对话的结构和情境要素。我们将所提框架与 Perspective API 进行比较，以检验其效用，并阐述了该框架如何用于减轻对话中的毒性。这些发现有助于我们理解情绪失调如何导致网络有毒行为，以及如何通过支持有效的情绪调节来缓解这一问题。该框架可用于指导设计更健康的社交媒体对话／讨论平台，尤其有助于维持讨论质量；而这一直被认为是在线数字技术面临的一项挑战。

参考文献

Antonakaki, D., Fragopoulou, P., Ioannidis, S., 2021年。Twitter研究综述：数据模型、图结构、情感分析与攻击。《Expert Systems with Applications》164，114006。Aroyo, L., Dixon, L., Thain, N., Redfield, O., Rosen, R., 2019年。众包主观任务：理解在线讨论中毒性的案例研究，载于：2019年万维网会议伴随论文集，第1100–1105页。Ban Kirigin, T., Bujačić Babić, S., Perak, B., 2021年。使用基于语料库的依存图进行词汇义项标注和情感潜势分析。《Mathematics》9，1449。Barry, E., Jameel, S., Raza, H., 2021年。Emojional：表情符号嵌入，载于：UK Workshop on Computational Intelligence，Springer，第312–324页。Berahmand, K., Haghani, S., Rostami, M., Li, Y., 2020年。利用复杂网络中的标签传播进行一种新的属性图聚类。《沙特国王大学计算机与信息科学杂志》。Bordoloi, M., Biswas, S.K., 2020年。基于图的情感分析：采用基于关键词排名的极性赋值。《多媒体工具与应用》79，36033–36062。Brambilla, M., Javadian, A., Sulistiawati, A.E., 2021年。在线社交媒体中的对话图，载于：国际网络工程会议，Springer，第97–112页。Castells, M., 2015年。《愤怒与希望的网络：互联网时代的社会运动》。John Wiley & Sons。Chen, Y., Mark, G., Ali, S., 2016年。通过智能手机摄影促进积极情感。《幸福感心理学》6，1–16。Chowanda, A., Sutoyo, R., Tanachutiwat, S., 等，2021年。探索社交媒体对话中的文本情绪识别机器学习技术。《Procedia Computer Science》179，821–828。

Costa, J., Guimbretière, F., Jung, M.F., Choudhury, T., 2019年。Boostmeup：通过在当下以不易察觉的方式利用智能手表调节情绪，提升认知表现。《ACM交互、移动、可穿戴与泛在技术会议录》3，1–23。Dmitrenko, D., Maggioni, E., Brianza, G., Holthausen, B.E., Walker, B.N., Obrist, M., 2020年。Caroma疗法：宜人气味可促进更安全的驾驶、改善情绪并提升愤怒驾驶者的幸福感，载于：2020年CHI计算机系统中的人为因素会议，第1–13页。Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P., 2019年。用于Twitter文本毒性分析的卷积神经网络，载于：INNS大数据与深度学习会议，Springer，第370–379页。Gillespie, T., 2020年。内容审核、人工智能与规模问题。《大数据与社会》7，2053951720943234。Goel, S., Anderson, A., Hofman, J., Watts, D.J., 2016年。在线扩散的结构性病毒式传播。《管理科学》62，180–196。Goldenberg, A., Gross, J.J., 2020年。数字情绪传染。《认知科学趋势》24，316–328。Google, 2021年。Perspective API。https://www.perspectiveapi.com/。[在线；访问日期：19-Dec-2022]。Gorwa, R., Binns, R., Katzenbach, C., 2020年。算法内容审核：平台治理自动化面临的技术与政治挑战。《大数据与社会》7，2053951719897945。Gross, J.J., 2014年。情绪调节：概念与实证基础。Gross, J.J., 2015年。情绪调节：现状与未来展望。《心理学探究》26，1–26。Guberman, J., Schmitz, C., Hemphill, L., 2016年。量化Twitter上的毒性与言语暴力，载于：第19届ACM计算机支持的协同工作与社会计算会议伴随论文集，第277–280页。Hossain, E., Wadley, G., Berthouze, N., Cox, A., 2022年。主动和被动社交媒体休息的动机与情境因素或可解释恢复与拖延之间的差异，载于：CHI计算机系统中的人为因素会议扩展摘要，第1–8页。Jhaver, S., Boylston, C., Yang, D., Bruckman, A., 2021年。评估取消Twitter平台访问作为审核策略的有效性。《ACM人机交互会议录》5，1–30。Kanavos, A., Perikos, I., Vikatos, P., Hatzilygeroudis, I., Makris, C., Tsaka-lidis, A., 2014年。社交网络中的对话情绪建模，载于：2014年IEEE第26届人工智能工具国际会议，IEEE，第478–484页。Kiskola, J., Olsson, T., Väätäjä, H., H. Syrjämäki, A., Rantasila, A., Isokoski, P., Ilves, M., Surakka, V., 2021年。在在线新闻评论的用户界面设计中应用批判性声音，以支持自我反思和情绪调节，载于：2021年CHI计算机系统中的人为因素会议，第1–13页。Konikoff, D., 2021年。毒性把关者：重新审视Twitter的辱骂与仇恨言论政策。《政策与互联网》。Kou, Y., Gui, X., 2020年。电子竞技游戏中的情绪调节：对《英雄联盟》的一项定性研究。《ACM人机交互会议录》4，1–25。Li, S., Hao, Y., Yoon, J., 2020年。Purpal：一种可增强消费行为中积极情绪的交互式盒子，载于：2020年CHI计算机系统中的人为因素会议扩展摘要，第1–6页。Lukoff, K., Yu, C., Kientz, J., Hiniker, A., 2018年。什么让智能手机使用变得有意义或无意义？《ACM交互、移动、可穿戴与泛在技术会议录》2，1–26。Luo, Ruikun, N.D., Yang, X.J., 2021年。基于行为与生理信号的移动情绪识别深度多模态方法。《IEEE情感计算汇刊》。Maarouf, A., Pröllochs, N., Feuerriegel, S., 2022年。社交媒体上仇恨言论的病毒式传播。arXiv预印本 arXiv:2210.13770。Majó-Vázquez, S., Nielsen, R., Verdú, J., Rao, N., de Domenico, N., Papaspiliopoulos, O., 2020年。COVID-19疫情期间社交媒体对话中的毒性规模与模式。Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E., 2019年。Dialoguernn：一种用于对话中情绪检测的注意力循环神经网络，载于：AAAI人工智能会议论文集，第6818–6825页。Mancini, M., Niewiadomski, R., Huisman, G., Bruijnes, M., Gallagher, C.P., 2020年。再多一位如何？——引入人工共栖陪伴者，载于：2020年CHI计算机系统中的人为因素会议扩展摘要集，第1–8页。McRae, K., Gross, J.J., 2020年。情绪调节。《情绪》20，1。Ments, L.v., Treur, J., Klein, J., Roelofsma, P., 2021年。用于医院团队协作中共享心智模型的二阶自适应网络模型，载于：国际计算集体智能会议，Springer，第126–140页。Mirbabaie, M., Brünker, F., Wischnewski, M., Meinert, J., 2021年。社交媒体上社会运动中连接性行动的发展。《ACM社会计算汇刊》4，1–21。Miri, P., Jusuf, E., Uusberg, A., Margarit, H., Flory, R., Isbister, K., Marzullo, K., Gross, J.J., 2020年。评估一种可个性化、隐蔽式振动触觉（piv）呼吸节拍器，用于即时情感调节，载于：2020年CHI计算机系统中的人为因素会议论文集，第1–12页。Mohammad, S.M., Turney, P.D., 2013年。通过众包构建词语—情绪关联词典。《计算智能》29，436–465。Newbold, J.W., Luton, J., Cox, A.L., Gould, S.J., 2017年。利用自然环境声景支持任务表现和情绪，载于：2017年CHI计算机系统中的人为因素会议扩展摘要集，第2802–2809页。Paredes, P.E., Zhou, Y., Hamdan, N.A.H., Balters, S., Murnane, E., Ju, W., Landay, J.A., 2018年。深呼吸：车内引导式缓慢呼吸干预。《ACM交互、移动、可穿戴和普适技术会议录》2，1–23。Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., Androutsopoulos, I., 2020年。毒性检测：语境真的重要吗？arXiv预印本 arXiv:2006.00998。Perikos, I., Hatzilygeroudis, I., 2018年。用于分析大型社交数据并对社交媒体中的情绪建模的框架，载于：2018年IEEE第四届大数据计算服务与应用国际会议（BigDataService），IEEE，第80–84页。Poria, S., Majumder, N., Mihalcea, R., Hovy, E., 2019年。对话中的情绪识别：研究挑战、数据集与最新进展。《IEEE Access》7，100943–100953。Ruensuk, M., Cheon, E., Hong, H., Oakley, I., 2020年。你在线时感觉如何：利用智能手机传感器检测社交媒体使用过程中的短暂情绪。《ACM交互、移动、可穿戴和普适技术会议录》4，1–32。Ruvimova, A., Kim, J., Fritz, T., Hancock, M., Shepherd, D.C., 2020年。“带我远离这里”：通过虚拟现实在开放式办公室中促进心流，载于：2020年CHI计算机系统中的人为因素会议论文集，第1–14页。Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J., Chen, Y.S., 2018年。CARER：用于情绪识别的情境化情感表征，载于：2018年自然语言处理经验方法会议论文集，计算语言学协会，比利时布鲁塞尔，第3687–3697页。URL：https://www.aclweb.org/anthology/D18-1404，doi：10.18653/v1/D18-1404。Saveski, M., Roy, B., Roy, D., 2021年。Twitter上有害对话的结构，载于：2021年万维网大会论文集，第1086–1097页。Semertzidis, N., Scary, M., Andres, J., Dwivedi, B., Kulwe, Y.C., Zambetta, F., Mueller, F.F., 2020年。Neo-noumena：增强情绪交流，载于：2020年CHI计算机系统中的人为因素会议论文集，第1–13页。Shaver, P., Schwartz, J., Kirson, D., O’connor, C., 1987年。情绪知识：对原型方法的进一步探索。《人格与社会心理学杂志》52，1061。Shen, K., Cox, A., 2020年。电子游戏作为数字情绪调节工具。《HCI-E理学硕士毕业项目报告》2020。Shi, Y., Koval, P., Kostakos, V., Goncalves, J., Wadley, G., 2023年。“即时幸福”：智能手机作为日常情绪调节工具。

《国际人机研究杂志》170，102958。Slovak, P., Antle, A.N., Theofanopoulou, N., Roquet, C.D., Gross, J.J., Isbister, K., 2022年。情绪调节干预设计：人机交互理论与研究议程。arXiv预印本 arXiv:2204.00118。Smith, W., Wadley, G., Webber, S., Tag, B., Kostakos, V., Koval, P., Gross, J.J., 2022年。日常生活中的数字情绪调节，载于：CHI计算机系统中的人为因素会议论文集，第1–15页。Solovev, K., Pröllochs, N., 2022年。道德情绪塑造了新冠肺炎错误信息在社交媒体上的病毒式传播，载于：2022年ACM网络大会论文集，第3706–3717页。Steinert, S., 2021年。新冠疫情与价值观变化：社交媒体和情绪传染的作用。《伦理与信息技术》23，59–68。Steinert, S., Dennis, M.J., 2022年。情绪与数字福祉：社交媒体的情感可供性。《哲学与技术》35，1–21。Tag, B., van Berkel, N., Vargo, A.W., Sarsenbayeva, Z., Colasante, T., Wadley, G., Webber, S., Smith, W., Koval, P., Hollenstein, T., 等，2022a。全球大流行病对年轻人使用技术进行情绪调节的影响。《计算机与人类行为报告》6，100192。Tag, B., Sarsenbayeva, Z., Cox, A.L., Wadley, G., Goncalves, J., Kostakos, V., 2022b。智能手机使用中的情绪轨迹：迈向真实环境中的情绪调节识别。《国际人机研究杂志》166，102872。Theofanopoulou, N., Isbister, K., Edbrooke-Childs, J., Slovák, P., 等，2019年。促进儿童中期情绪调节的智能玩具干预：可行性研究。《JMIR心理健康》6，e14029。Thomas, K., Kelley, P.G., Consolvo, S., Samermit, P., Bursztein, E., 2022年。“这很常见，也是作为内容创作者的一部分”：理解创作者如何在线上经历并应对仇恨与骚扰，载于：CHI计算机系统中的人为因素会议论文集，第1–15页。Torre, J.B., Lieberman, M.D., 2018年。将感受化为言语：情感标记作为一种隐性情绪调节方式。《情绪评论》10，116–124。Vilarinho, G.N., Ruiz, E.E.S., 2018年。用于Twitter情感分析的词图全局中心性度量，载于：2018年第七届巴西智能系统会议（BRACIS），IEEE，第55–60页。Wadley, G., Krause, A., Liang, J., Wang, Z., Leong, T.W., 2019年。国际学生使用音乐流媒体平台进行情绪调节，载于：第31届澳大利亚人机交互会议论文集，第337–341页。Wadley, G., Smith, W., Koval, P., Gross, J.J., 2020年。数字情绪调节。《心理科学当前方向》29，412–418。Walther, J.B., 1993年。计算机媒介互动中的印象形成。《西方传播学杂志》（含《传播报告》）57，381–398。Wang, Y., Diakopoulos, N., 2021年。《纽约时报》精选内容在评论质量和参与度中的作用，载于：第54届夏威夷国际系统科学会议（HICSS 2021），IEEE计算机学会，第2924–2933页。Xenos, A., Pavlopoulos, J., Androutsopoulos, I., Dixon, L., Sorensen, J., Laugier, L., 2021年。毒性检测可能对对话语境敏感。arXiv预印本 arXiv:2111.10223。Yang, K., Wang, C., Sarsenbayeva, Z., Tag, B., Dingler, T., Wadley, G., Goncalves, J., 2021年。使用真实失真后的面部图像数据集对商业情绪检测系统进行基准测试。《视觉计算机》37，1447–1466。Yang, S., Xing, L., Li, Y., Chang, Z., 2022年。基于图注意力神经网络的隐性情感分析。《工程报告》4，e12452。Yang, Y., Liu, K., Li, S., Shu, M., 2020年。新冠肺炎疫情期间人们的社交媒体活动、情绪调节策略及其相互作用对心理健康的影响。《国际环境研究与公共卫生杂志》17，8931。Yu, B., Hu, J., Funk, M., Feijs, L., 2018年。Delight：通过环境光进行生物反馈，以干预压力并辅助放松。《个人与普适计算》22，787–805。Yue, Z., Zhang, R., Xiao, J., 2022年。新冠肺炎疫情期间被动使用社交媒体与心理幸福感：社会比较和情绪调节的作用。《计算机与人类行为 127, 107050。Zsido, A.N., Arato, N., Lang, A., Labadi, B., Stecina, D., Bandi, S.A., 2021年。不适应性认知情绪调节策略和社交焦虑在问题性智能手机和社交媒体使用中的作用。《人格与个体差异》173, 110647。