文明排队：调节情绪和减少数字讨论中的毒性

Akriti Verma[1], Shama Islam[1], Valeh Moghaddam[2], and Adnan Anwar[2]

[1] School of Engineering, Deakin University, Australia
[2] School of Information Technology, Deakin University, Australia

摘要 在线毒性（包括仇恨言论和 trolling）的普遍性破坏了数字互动和在线幸福感。先前的研究主要集中在事后监管上，忽视了在线对话中的实时情感动态以及用户情绪对他人影响的重要性。本文提出了一种基于图的框架来识别在线对话中需要进行情绪调节的需求。该框架促进了自我反思以管理情绪反应，并鼓励在现实时间里负责任的行为。此外，还提出了一个评论排队机制来应对利用情绪煽动对话的故意 trolling 用户。该机制引入了发布评论的延迟，为用户提供了在进一步参与对话之前自我调节的时间，有助于保持情感平衡。来自 Twitter 和 Reddit 的社交媒体数据分析表明，基于图的框架减少了 12% 的毒性，而评论排队机制将愤怒的传播降低了 15%，平均只有 4% 的评论被暂时扣留。这些研究结果表明，结合实时情绪调节和延迟监管可以显著改善在线环境中的幸福感。

Keywords: 数字情感调节（DER）· 人际情感调节（IER）· 社交媒体中的情绪 · 在线情绪 · 人机交互（HCI）· 情感计算

1 介绍

在线互动的快速节奏常常导致情绪化的对话，增加了极化反应和数字冲突的可能性 [40]。在线毒性日益受到关注，这通常包括仇恨言论和 trolling，对用户的福祉以及数字社区的整体健康产生了负面影响 [29]，[9]。目前管理在线毒性的方法主要集中在事后审查上，即有害内容在发布后被识别并删除 [3]。尽管内容审查工具有所进步，但仍需要前瞻性解决方案来赋予用户在毒性升级之前管理对话对其情感影响的能力 [44]，[10]。

数字情感调节（DER），即利用数字技术影响个人的情绪状态，最近已成为一种习惯。个体通过结合各种应用程序和设备来有目的地管理情绪 [46]，[40]。一些示例包括在锻炼时听振奋人心的音乐，在工作后观看喜剧或轻松视频以缓解压力，孤独时玩社交电子游戏，或者浏览社交媒体应用以对抗无聊。

虽然 DER 应用程序在个人情境中被广泛认可，但在互动在线环境中它们尚未得到充分探索，情绪在这种环境下可以迅速传播并影响他人 [30]。此外，现有解决方案未能弥合反应性管理与主动性情感调节之间的差距。本文提出了一种通过倡导基于自我反思的情感调节，并将其与评论排队机制相结合的新方法，以遏制负面情绪的传播，特别是在容易受到网络欺凌的对话中促进更健康的在线交流。

自我反思的实践，包括评估自己的情感和行为，已被广泛认可为一种管理日常互动中情绪健康的方法 [20]。然而，其在在线环境中调节情绪的潜力尚未得到充分探索。通过鼓励数字空间中的自我反思，用户可以更好地意识到他们的评论如何影响对话的情感走向。这将使他们能够在进一步参与之前重新评估和管理自己的情绪 [24]。这种主动的方法将焦点从仅在毒性发生后才处理的反应性调节转向实时情感调控，从而降低情绪升级的可能性。

本文介绍了一种利用图框架来可视化对话的情感基调并告知用户其对情感状态影响的方法。这种方法促进自我反思，为用户提供一个考虑他们评论情感影响的机会。此外，在故意挑衅的情况下，提出了一种评论排队机制。该系统引入了一个简短的发布延迟，让用户在暂停期间对自己的情感状态进行反思，同时确保对话的情感平衡得以维持。

我们对社交媒体数据的分析表明，促进在线自我反思可以成为增强情绪调节能力的强大工具，特别是在处理情感激烈的对话时。本文强调了通过自我反思和故意延迟回应进行隐性情绪调节的价值，这为认知重构提供了机会 [23]。这种微妙的方式引导用户采取更加体贴和受控的回应，旨在减少网络毒性并促进更健康的数字空间。因此，这项工作做出了以下研究贡献：

- 一种基于自我反思的 DER 系统：本文提出了一种基于图的框架，该框架告知用户他们对对话的情感影响，并通过以用户为中心的方法促进隐性情绪调节。
- 一种用于实时情绪调节的评论排队机制：本文介绍了一种评论排队系统，该系统延迟用户响应以防止冲动的情绪反应，为用户提供时间反思他们的情绪状态。与现有的延迟机制不同，我们的系统具有自适应性和情境敏感性，能够根据对话中的情绪动态进行动态调整。
- 情感调节在数字对话中的实证验证：我们的初步发现表明，仇恨言论和愤怒的传播减少了 12%，超过了 Google 的 Perspective API 3%。此外，在正在进行的对话中使用队列机制时，潜在的 trolling 和仇恨言论传播减少幅度可达 15%，仅有 4%的评论被暂时扣留，平均时间为 47 秒。

2 文献回顾

随着社交媒体平台的增长，在线环境中越来越需要关注情感健康。这篇简要的文献回顾探讨了最近关于 DER 的研究、自我反思作为情绪调节工具、网络喷子行为以及评论管理策略。它强调了这些研究在增进我们对这些领域理解的重要性。

2.1 数字情绪调节（DER）

DER 是心理学和人机交互中的一个新兴研究领域，专注于数字工具对个人情感健康的影响。DER 研究考察了数字技术对情绪调节 (ER)[31], [16], [15], [46] 的影响。研究利用自我报告和日记等方法展示了数字技术在日常情绪调节 [40],[38],[21],[27] 中的应用。数字化的情绪调节干预措施结合了生物反馈和基于提醒的系统，通过鼓励用户练习情绪调节 [24],[25] 来增强情绪调节技能。此外，多模态传感器，如摄像头和触摸传感器，已被用于在 DER 过程 [28],[36] 中观察和识别情感变化。研究还揭示了有毒讨论集中在社交联系有限的个体中尤其明显 [42],[37]。此外，研究表明推特上表达的道德情感可以影响虚假谣言的传播，关键用户在发起在线社会运动中扮演着至关重要的角色 [41]。然而，现有的大部分文献关注个体情绪调节，忽视了在线对话中的社交动态，在这种情况下，他人的感情传染可以显著影响一个人的情绪状态。本研究通过引入包括情绪调节策略和在线互动期间反馈机制的方法来解决当前 DER 文献中存在的这一空白。

2.2 情绪调节与自我反思

如 Gross[14] 所定义的，情绪调节涵盖了个体管理自身经历的情绪、表达方式以及这些情绪发生时机的过程。传统上在离线环境中研究的情绪调节通过使个人在反应前重新考虑其情绪反应，从而增强了人际沟通。自我反思是情绪调节的关键组成部分，它使个体能够评估自己的情绪和潜在动机，可能进而导致认知重评 [33]。研究表明，时间距离——即在对情绪刺激做出反应前留出时间——可以降低情绪唤起并促进更好的情绪控制 [17]。虽然这一概念主要在面对面和其他线下环境中进行探讨，但在数字互动中却很少付诸实践，因为在这些快速的互动中往往没有足够时间进行反思。现有研究虽然考察了数字环境中的情绪调节，但主要集中在教育工具或治疗应用中，对公共在线讨论 [18],[19] 的关注极少。我们的工作通过开发一个框架来促进实时

在线对话中的自我反思，使用户在发布可能具有毒性的评论前能够暂停、反思并调节自身情绪，从而弥合了这一差距。

2.3 在线毒性与故意挑拨

虽然标注情绪可能有助于控制无意的情绪爆发，但应对故意的 trolling 则更为复杂 [22]。trolling 涉及有意扰乱在线对话以引发情绪反应，在社会学和计算机科学领域已得到了广泛的研究 [2]。研究表明，trolls 是那些从引起情绪混乱和加剧冲突中获得乐趣的人，他们的动机多种多样，包括寻求娱乐和关注到更恶意的意图如造成伤害或操纵讨论 [26]，[5]。当前的监管方法主要集中在发布有害内容后的移除上，这是一种反应性的而非预防性的措施。虽然诸如封禁用户和删除评论等工具可以减轻某些由网络喷子造成的损害，但它们并没有解决最初的升级问题 [47], [4], [3]。先前的研究表明，传统的内容审核往往加剧了网络喷子的行为，因为网络喷子喜欢挑战平台政策的边界 [35]。这突显了需要一种更具有预防性的管理网络喷子的方法，我们的工作旨在通过设计一种基于自我反思的情绪调节评论排队机制来实现这一点。

2.4 会话分析与评论调节机制

评论审核一直是应对在线毒性的主要防御措施，通常采用检测和移除有害内容以及标记不当行为的方法。像 Facebook 和 Twitter 这样的平台已经实施了算法工具，如内容过滤器和审核机器人，以减轻有毒行为 [12]，[8]。然而，最近的文献强调它们在带来有意义、长期的用户行为改变方面的能力有限 [7]。有人认为问题的根本不仅在于识别有害内容，还在于改变促使用户发布这些内容的情感反应和行为 [6]，[34]。

在数字通信中，有效的情绪调节（ER）包括利用事实线索、自动情感识别和教学学习 [24], [39]。隐性的情绪调节操作是自动化的并且可以被自动化，在最近得到了关注。情感标签化，使对话中的情感成分更加明显，能够增强情绪调节 [43]。当前的研究已经探索了评论审查的方法，例如引入延迟以限制冲动反应。例如一个延时反馈环可以给用户在发布评论之前重新考虑的时间 [5]。这与情绪调节理论一致，该理论认为引入基于时间的暂停可以帮助用户进行认知重构，从而缓解情感化的情况。然而，大多数现有的评论延迟机制要么是任意设定的，要么忽视了对话中的情感动态。

因此，文献中存在以下空白：

- 在线对话中缺乏实时情绪调节：当前的研究探索了个人使用数字工具时的数字化情绪调节（DER），但在互动式在线环境中缺少实时的情绪调节策略，主要集中在事后内容审核上，而不是防止对话中的情绪升级。
- 在数字空间中自我反思的有限应用：虽然自我反思常用于线下情绪调节，但其在线沟通中的应用尚未得到充分探索。将自我反思融入数字互动的方法可以促进更加深思熟虑、情绪调节得当的回应。
- 对网络喷子行为根本原因的关注有限：当前的研究集中在删除或审查有毒内容上，而不是解决导致网络喷子行为的根本原因，比如情绪激动和冲动。需要制定策略来鼓励用户在发布内容前先停下来思考。

我们的研究在此基础上通过将一个评论排队系统与一个旨在持续监测对话情感基调的基于图框架相结合进行了扩展。这种方法不仅引入了时间延迟以鼓励自我反思，还确保了对话的情感平衡得以维持，从而降低了因网络喷子行为或冲动行为而导致升级的可能性。

3 方法论

本文提出了一种通过自我反思来管理在线对话中情绪的方法。该方法包括 eImpact 框架的四个阶段（图 1），其中包括情感检测、基于图形的对话分析，以及实现一个评论排队系统以鼓励自我反思和最小化情绪升级。

3.1 数据收集与预处理

对于本研究，我们关注来自 Twitter 和 Reddit 的数据，因为它们具有不同的沟通风格，并且能够捕捉广泛的情感表达。

Twitter 促进了快速、公开的互动，这通常会导致直接且情绪化的对话，特别是在政治和社会问题方面。其快节奏的特性使其非常适合分析情绪如何传播以及数字互动中自我反思的需求 [30]。

另一方面，Reddit 凭借其较长的帖子和特定主题的社区，允许进行更深入的讨论。其线程结构使得能够追踪对话中情感的发展，为情绪调节 [30] 的分析提供了更加丰富的内容。

推特数据集 数据是从 2020 年 4 月至 2022 年 8 月期间澳大利亚国会议员在 Twitter 上的对话中收集的，使用了推特 API（Tweepy）。这些推文中包含了关于 COVID-19、政策变化以及其他政治话题的讨论，并被收集和提炼以供分析。数据集包含 25K 条推文，提供了公众反应和互动的各种情感见解。

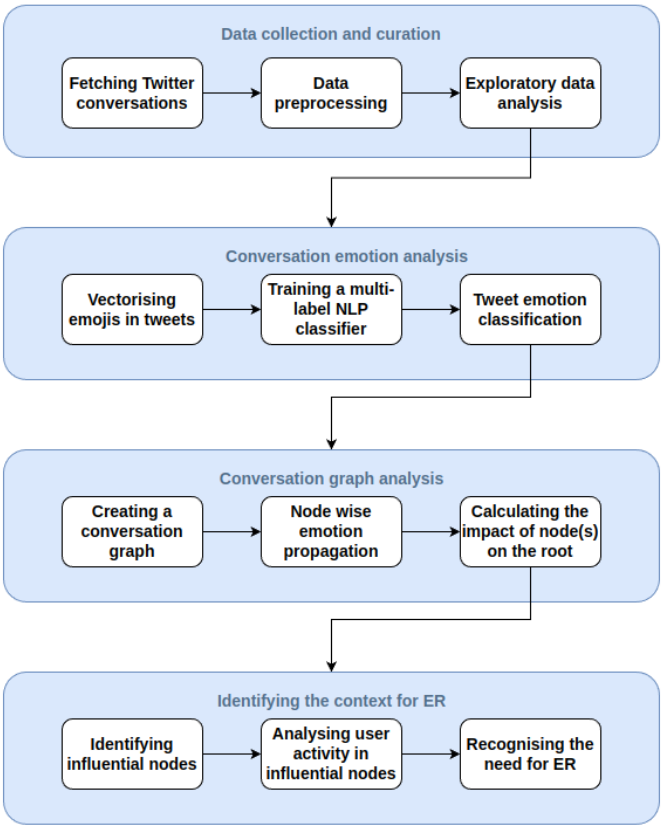


图 1: 情感影响：支持社交媒体对话中情绪调节的框架

Reddit 数据集 为了进行更深入的对话分析，我们从 Reddit 获得了 15,000 条用户帖子和评论使用了 Reddit API（PRAW）。该数据集包含帖子、回复、评论和时间戳，允许追踪整个对话过程中的情感变化。所有文本都经过清理以消除多余的字符、链接和符号，并进行了分词处理以便进行自然语言处理 (NLP) 分析。

3.2 情感分类

为了识别收集的数据中的情绪，我们使用了 NRC 情感词典 [32]，该词典能够识别诸如愤怒、恐惧、期待、信任、惊讶、悲伤、快乐和厌恶等情绪。分类过程涉及将文本和表情符号转换为向量来进行分析。推文中的表情符号通过 Emojinal 库 [1] 被替换为由生成模型创建的向量表示，之后使用推文分

词器对推文文本进行了分词处理。每条评论都收到了一个从 0.1 到 1.0 的情绪和强度评分，反映了其文本中表达的情绪及其强烈程度。

3.3 基于图的对话分析

对话使用有向无环图（DAG）结构表示，其中原始帖子作为根节点，而回复和评论则充当子节点。评论与回复之间的连接通过边来表示。这种层级结构随后被用来评估评论对整个对话的影响。

eImpact 模型将每条评论或推文表示为图中的一个节点，每个节点被分配一个情感分数 [45]。该分数由情感分类模型关联内容的可能性决定。使用各种指标评估每个节点对根节点（原始帖子）的影响：

- 回复数量：回复较多的节点被认为更有影响力。
- 距离根节点：越接近原始帖子的节点对整体情绪基调的影响越大。
- PageRank：一条评论的重要性由其在图中的位置及其收到的互动量决定。
- 情感强度：每条评论的情感强度对其影响力分数有所贡献，影响根节点的整体情感板。

  这种方法使模型能够识别对话中可能引发情绪升级或毒性反应的特定节点。

该图结构允许我们评估每个评论如何影响对话的情感基调。这些影响力分数用于更新根节点的情绪板，后者整合了所有评论的情感影响。为了防止负面情绪的升级，高度有毒的评论在添加到对话图之前会根据其对根节点的影响暂时保存。

3.4 评论排队系统以进行自我反思

为促进自我反思和调节在线对话中的消极情绪，我们引入了一个如图 2 所示的评论排队机制。该系统评估每个新评论，在发布前计算其对整个对话情感基调的潜在影响。由于每条评论都影响了对话的情感基调，因此它的加入也意味着向累积的情感基调中添加了一定的情感重量。如果一条评论的情感影响超过了预定义的阈值，例如愤怒 $> 50\%$ 或恐惧 $> 60\%$，该评论将被标记为有毒并暂时存储在队列中。

例如，如图 3 所示，在围绕政府政策改革的 Twitter 对话中，一个强烈表达对政府政策愤怒的回复将根节点的愤怒阈值提高到了 65%，图 3(a)。因

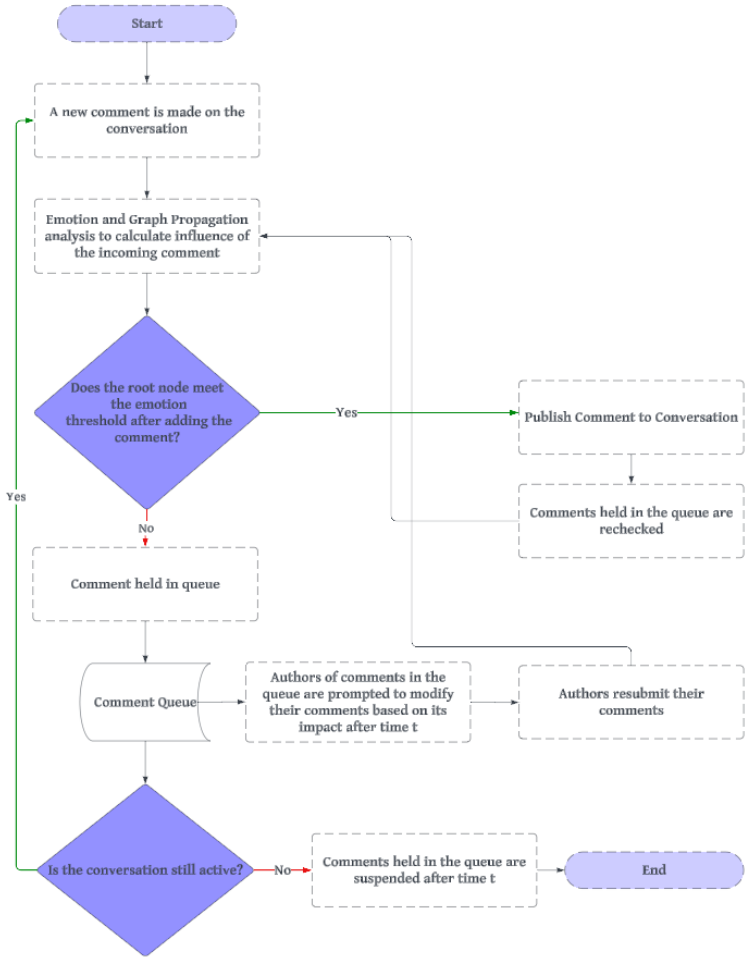
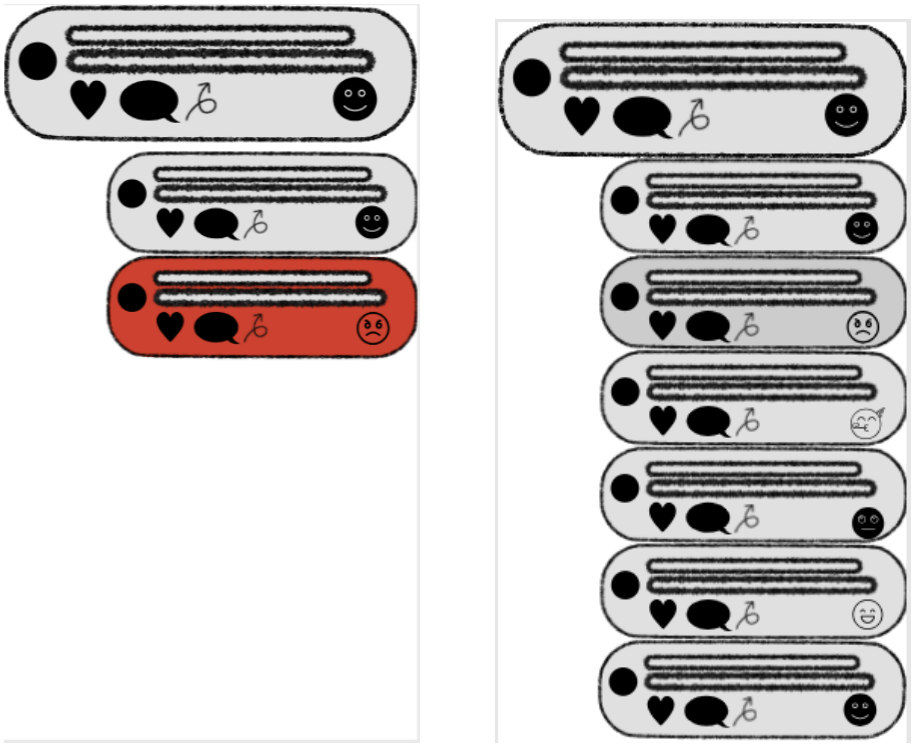


图 2: 建议评论排队以鼓励自我反思



(a) 愤怒的评论已加入队列　　(b) 愤怒的评论重新整合

图 3: 线程上的评论排队

此，它被放入队列（用红色表示）。这条评论位于根节点下方，由于其情感强度对整体情绪产生了重大影响。然而，如图 3(b) 所示，当后续评论引入了信任和喜悦的情感时，愤怒的百分比下降到了 55%以下，使得排队中的评论可以安全地被整合而不会进一步升级。

在队列中，每当有新评论添加到对话时，都会定期重新评估评论。这保证了情感板的持续更新，使得系统能够确定之前排队的评论是否现在可以加入而不会超过阈值。如果新评论的添加平衡或减少了整体的情感强度，那么之前的被标记评论就可以安全地整合进对话。

为了适应在线对话的流程，系统区分了活跃和非活跃对话阶段。在活跃阶段，阈值更为严格以防止负面情绪升级。随着对话变得不那么活跃，这些阈值可能会放宽，允许更多评论进入对话，同时仍然保持其情感平衡。

如果在处理完其他所有评论后，某条评论仍留在队列中且情绪阈值仍然超标，则会提示其作者修改评论。修改后，该条评论将被重新评估。如果修订后的评论符合情绪阈值，则将其纳入对话；如果不符，则暂停该条评论以防止进一步的情绪升级并确保更健康的讨论。

对话情绪基调的阈值是通过几个参数确定的，并且会根据算法动态调整，该算法考虑了评论的数量、对话中的总体情绪分布以及情感强度的近期变化。例如，在非常活跃的讨论中，许多评论表达了强烈的情绪，愤怒或恐惧的阈值可能会暂时提高以减少排队评论的数量。另一方面，在较为平静的时候，阈值可能会适度降低以实现更严格的监管并避免可能的情感强化。对话的活跃/非活跃阶段是基于对评论之间的时间间隔、总评论数量和用户参与模式的持续分析来区分的。这确保了系统能够适应活动水平，例如高强度讨论或沉寂阶段。

阈值也根据对话的语境和不断变化的情绪分布进行调整。采用滑动窗口方法，重点关注最近的评论以捕捉实时情绪状态，并防止过时的情绪扭曲语气。在我们的实验中，我们使用了 100 条评论。积极的情绪，如快乐或爱意，通过加权许可给予更高的权重，促进建设性的互动并调节愤怒或恐惧等消极情绪。为了最小化评论暂停并保持情感平衡，系统优先考虑对话中的代表性不足的情绪。这确保了多样化的情感表达被整合，防止单一情感的主导地位，并促进更加平衡和包容的对话。

通过纳入反思暂停并提供用户改进其贡献的机会，该系统促进了更加审慎和情绪平衡的互动，同时也抑制了冲动评论。

例如，在一个特别活跃的讨论疫情政策的 Twitter/Reddit 线程中，由于大量表达强烈情绪的评论涌入，愤怒和恐惧的阈值会暂时提高。这种调整将有助于减少对话中的干扰，使具有适度情感权重的评论能够被包括在内，同时确保极端异常值得到适当管理和排队。

3.5 实验设置

如图 1 所示，在收集的数据经过处理后，每一段对话都被转换为一个有向无环对话图。随后，使用 NRC 词典将图中的每个评论分配了一个情绪，并附上了该评论内容中所包含的情绪强度。基于其情绪强度、距离根节点的距离、PageRank 和回复数量的组合来计算每个评论的影响。这个分数被用来更新对话的情感影响。根节点维护了一个情感板，跟踪每种情感对对话的影响百分比，随着新评论添加到图中，该板会动态更新。框架通过将其性能

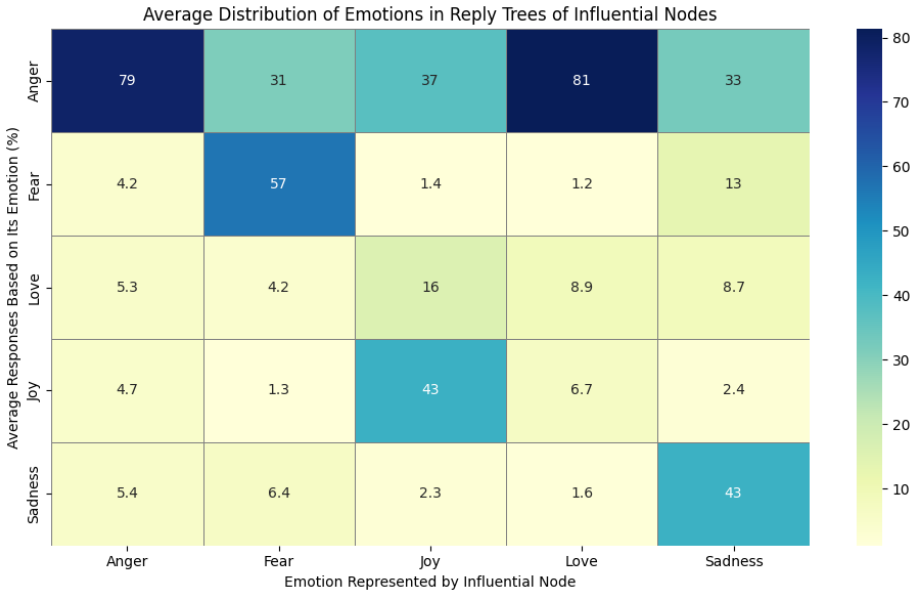


图 4: 情绪的平均分布以及有影响力的节点回复树中独特用户的数量 [45]

与 Google Perspective API 生成的毒性评分进行比较来进行评估 [11]。我们的分析集中在通过停用有影响力的节点或在其达到毒性阈值时限制响应来减少仇恨言论和极化。研究结果表明，尽管 Perspective API 将仇恨言论减少了 7%，但 eImpact 框架实现了 10%的减少。这证实了所提出的框架在识别和调节帖子毒性方面是有效的，同时考虑其主观性。此外，它可以通过提供关于对话中毒性来源的见解来补充在线内容审核工作，考虑到内容和上下文。

为了模拟在线对话的演变性质，引入了动态阈值来检测毒性。这些阈值随着处理更多评论而降低，使得情感影响的评估更加灵活和适应性更强。在排队的情境下，那些超过根节点情绪板上的毒阈值的评论会被放入队列中。这些被持有的评论会在每次添加新评论后重新评估，符合阈值的随后会被添加到图中。

4 结果与讨论

在这项研究中，我们评估了 eImpact 框架和评论排队机制在管理以图形表示的在线对话中的情感动态方面的有效性，特别强调了对缓和有毒评论的

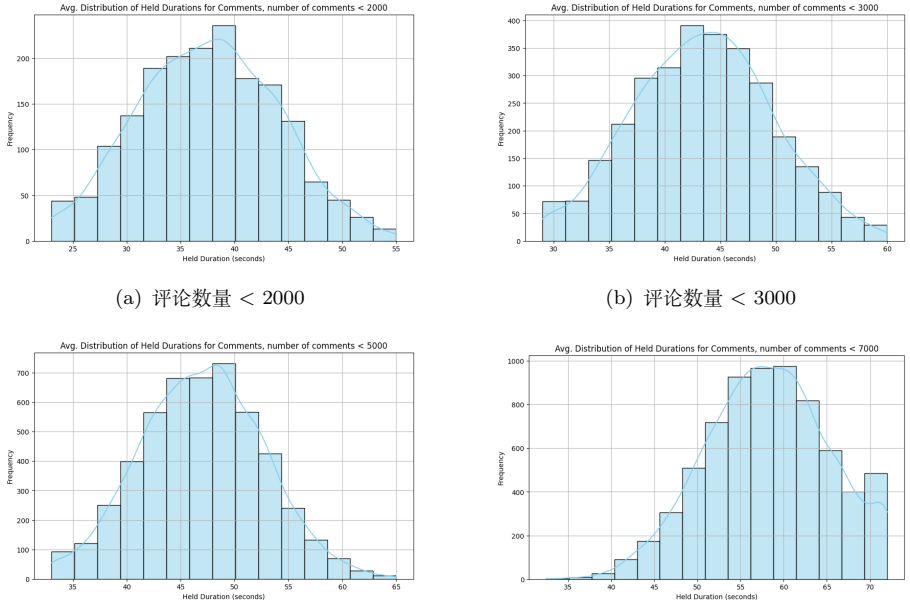
作用。我们比较了两种不同的方法：一种没有排队机制，在这种方法下，评论被立即添加到图中；另一种有排队机制，在这种情况下，评论暂时被搁置并在整合进对话之前重新评估其是否达到情感阈值。

表 1: 提出的框架 [45] 与 Perspective API[11] 的比较

Utilised Model	识别关键节点	Toxic Node Detection	估计毒性减少
eImpact Framework	基于影响分数，考虑到推文文本以及其连通性	1-4%	10%
Perspective API [11]	基于毒性，仅考虑推文文本中的	1-2%	7%

我们对来自 Twitter 和 Reddit 的在线对话中情感传播的分析表明，高度情绪化的内容往往会获得更大的用户参与度。我们观察到，使用充满情感的语言的推文，如表达爱意，经常会引发强烈的、由愤怒驱动的反应，导致情感极化加剧。当同一组用户反复与帖子互动时，这种极化效应尤为明显，从而放大了愤怒并抑制了诸如喜悦等其他情绪。我们的研究结果与现有研究一致，表明当愤怒主导在线讨论时，深深根植的激进观点更容易表现为毒性。图 4 显示了由影响力节点表示的每种情感的情感平均分布以及参与有影响力的节点回复树中的唯一用户数量。

我们评估了旨在缓解情绪升级的 eImpact 框架，以及谷歌的 Perspective API 在减少有毒内容方面的效果。结果显示，与使用 Perspective API 减少了 7%相比，eImpact 使仇恨言论和极化内容减少了 10%。这一结果强调了所提



(a) 评论数量 < 2000

(b) 评论数量 < 3000

(c) 评论数量 < 5000

(d) 评论数量 < 7000

图 5: 使用提议的队列方法时评论持有时间的平均分布

出的框架不仅在处理内容方面，而且在更广泛的对话背景和情绪动态方面都具有有效性。通过促使用户考虑评论的情绪影响，eImpact 促进了自我调节，并随着时间的推移减少有毒行为。表 1 比较了所提出的框架与 Perspective API 在识别关键节点和降低毒性可能性方面的效果。

然后我们实现了评论排队机制，以实时调节情绪强度。该排队系统通过评估其情感影响来管理评论发布前的流程。分析保留时长显示，评论通常被排队 40 到 55 秒，平均持有时间为 47 秒，提供了充足的时间进行自我反思。图 5 显示了一个直方图，说明了评论在处理前在队列中持有的时间分布情况。x 轴表示持有的时长（以秒为单位），而 y 轴代表评论的频率。如图所示，短时期持有大量评论表明排队系统有效地将评论重新整合到流程中。该分布显示大多数评论只被短暂保留，这表明系统有效调节了情绪而不造成重大延误。较长的持有时间指代那些情感挑战较大的评论，需要延长保留时间以确保情绪板的稳定性。

在比较使用和不使用排队系统的情况时，我们观察到情绪板在使用队列的情况下保持了更加平衡的情绪分布。没有队列时普遍存在的愤怒和恐惧等

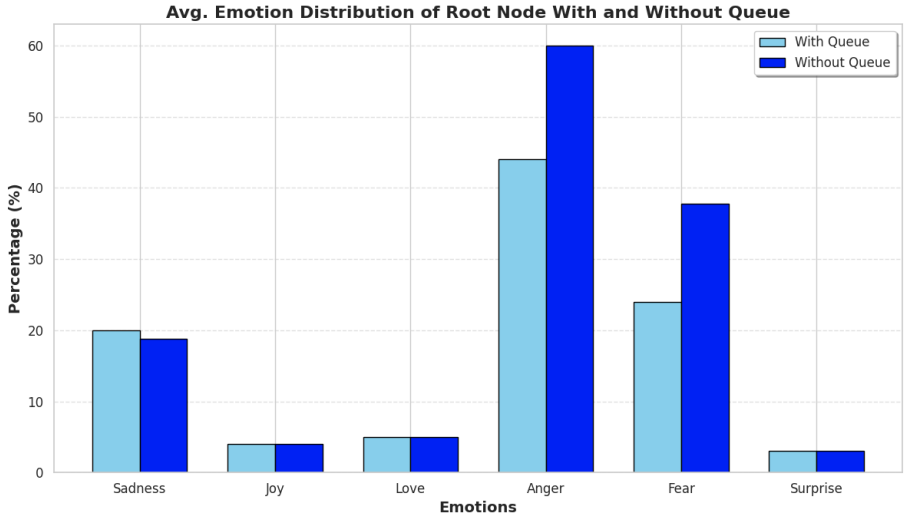


图 6: 根节点有队列和无队列的平均情绪分布

主导情绪显著减少。逐步将评论融入对话有助于减缓情绪峰值，特别是在评论可能会迅速加剧负面情绪的情况下。因此，对话的基调变得更加温和。图 6 展示了在两种实验条件下——无排队和有排队——会话图最终情绪构成的柱状图对比。y 轴显示每个情绪对根节点整体情绪状态的百分比贡献。这些图表展示了队列机制对情绪动态的影响。“无队列” 情景中，对话主要表现为愤怒和恐惧等负面情绪，而 “使用队列” 情景下则呈现更平衡的情绪状态。当使用队列时，情绪始终在阈值范围内，并且不受特定情绪的支配。研究结果表明，在没有队列的情况下，负面情绪倾向于积累并主导对话。相反，队列机制有效地调节了情绪影响，防止负面情绪压倒整个对话，从而实现更加平衡的情绪分布。

动态阈值、滑动窗口和加权配额机制有助于维持对话的流畅。通过根据对话的内容和活动水平调整阈值，系统能够实时评估评论。滑动窗口专注于最近 100 条评论，确保只有最新的情感趋势影响对话，从而防止过时的情感影响整体情绪基调。此外，通过优先考虑积极情绪，并在队列中为较少出现的情绪分配更高的优先级，系统促进了情感平衡的对话。图 7 显示了一条折线图，展示了实施排队机制后，对话图中每种情感随时间变化的情感影响。x 轴表示时间，而 y 轴代表累计情感影响。该图表提供了随着对话展开和评论依次添加时情绪如何演变的可视化展示。它揭示了排队机制如何影响特定

情感流动的方式。图表表明负面情绪会经历波动但在多个点上得到缓解，这表明排队机制干预以防止它们主导对话。另一方面，积极情绪如快乐和爱随着时间增强，表明系统促进了一条更加积极的情感轨迹。

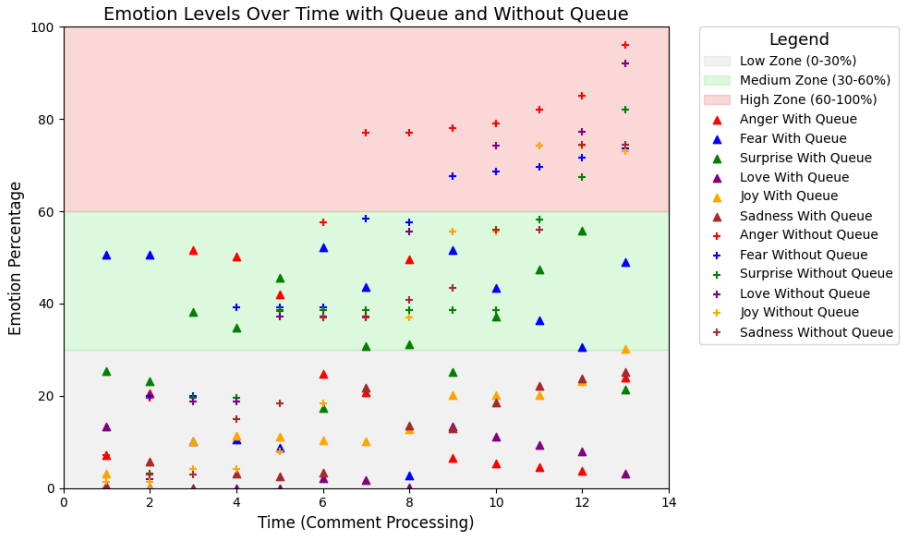


图 7: 对话中使用队列时的平均情绪水平

总体而言，当使用队列时，我们观察到愤怒和恐惧情绪的传播平均减少了 15%，相比于立即发布评论的情况。这种减少突显了排队系统通过允许用户重新考虑他们的评论来鼓励更健康、情感不那么激烈的对话的潜力。此外，只有 4%的评论被保留审查，这证明了该系统在管理情绪升级的同时维持对话流畅性的有效性。

4.1 结论

本文提出了一种创新框架，该框架结合了评论排队和自适应阈值来减少在线毒性并促进数字对话中的情绪调节。通过将潜在有害的评论暂存于队列中并在发布前评估其影响，该框架促进了自我反思，并缓解了情绪波动。我们的实验显示，与未经过滤的对话相比，负面情绪如愤怒和恐惧减少了 15%。

该框架通过鼓励更健康和更有责任心的在线讨论，而不会影响用户参与度，在数字情绪调控（DER）领域做出了有价值的贡献。未来的研究可以在

此基础上实施针对不同类型内容的自适应机制，并通过用户访谈和调查探索实时情感反馈。

5 致谢

在撰写本文的过程中，使用了 Grammarly[13] 来检查和提升本文件的语法和写作风格。我们感谢 Amity 大学接受我们的工作作为主旨论文。

参考文献

1. Barry, E., Jameel, S., Raza, H.: Emojional: Emoji embeddings. In: UK Workshop on Computational Intelligence. pp. 312–324. Springer (2021)
2. Buckels, E.E., Trapnell, P.D., Paulhus, D.L.: Trolls just want to have fun. Personality and individual Differences **67**, 97–102 (2014)
3. Chandrasekharan, E., Jhaver, S., Bruckman, A., Gilbert, E.: Quarantined! examining the effects of a community-wide moderation intervention on reddit. ACM Transactions on Computer-Human Interaction (TOCHI) **29**(4), 1–26 (2022)
4. Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E.: You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. Proceedings of the ACM on human-computer interaction **1**(CSCW), 1–22 (2017)
5. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J.: Anyone can become a troll: Causes of trolling behavior in online discussions. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. pp. 1217–1230 (2017)
6. Frischlich, L., Schatto-Eckrodt, T., Boberg, S., Wintterlin, F.: Roots of incivility: How personality, media use, and online experiences shape uncivil participation. Media and communication **9**(1), 195–208 (2021)
7. Gillespie, T.: Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press (2018)
8. Gillespie, T.: Content moderation, ai, and the question of scale. Big Data & Society **7**(2), 2053951720943234 (2020)
9. Goel, S., Anderson, A., Hofman, J., Watts, D.J.: The structural virality of online diffusion. Management Science **62**(1), 180–196 (2016)

10. Gongane, V.U., Munot, M.V., Anuse, A.D.: Detection and moderation of detrimental content on social media platforms: current status and future directions. Social Network Analysis and Mining **12**(1), 129 (2022)

11. Google: Perspective API. `https://www.perspectiveapi.com/` (2021), [Online; accessed 19-Dec-2022]

12. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society **7**(1), 2053951719897945 (2020)

13. Grammarly: Grammarly. `https://www.grammarly.com` (2024), accessed: 2024-06-20

14. Gross, J.J.: Emotion regulation. Handbook of emotions **3**(3), 497–513 (2008)

15. Gross, J.J.: Emotion regulation: conceptual and empirical foundations. (2014)

16. Gross, J.J.: Emotion regulation: Current status and future prospects. Psychological inquiry **26**(1), 1–26 (2015)

17. Grossmann, I., Kross, E.: Exploring solomon ' s paradox: Self-distancing eliminates the self-other asymmetry in wise reasoning about close relationships in younger and older adults. Psychological science **25**(8), 1571–1580 (2014)

18. Hadwin, A., Järvelä, S., Miller, M.: Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In: Handbook of self-regulation of learning and performance, pp. 83–106. Routledge (2017)

19. Hadwin, A.F., Davis, S.K., Bakhtiar, A., Winne, P.H.: Academic challenges as opportunities to learn to self-regulate learning. Problem solving for teaching and learning pp. 34–47 (2019)

20. Herwig, U., Kaffenberger, T., Jäncke, L., Brühl, A.B.: Self-related awareness and emotion regulation. NeuroImage **50**(2), 734–741 (2010)

21. Hossain, E., Wadley, G., Berthouze, N., Cox, A.: Motivational and situational aspects of active and passive social media breaks may explain the difference between recovery and procrastination. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–8 (2022)

22. Jane, E.A.: Online abuse and harassment. The international encyclopedia of gender, media, and communication **116** (2020)

23. Kiskola, J., Olsson, T., Syrjämäki, A.H., Rantasila, A., Ilves, M., Isokoski, P., Surakka, V.: Online survey on novel designs for supporting self-reflection and emotion regulation in online news commenting. In: Proceedings of the 25th International Academic Mindtrek Conference. pp. 278–312 (2022)

24. Kiskola, J., Olsson, T., Väätäjä, H., H. Syrjämäki, A., Rantasila, A., Isokoski, P., Ilves, M., Surakka, V.: Applying critical voice in design of user interfaces for supporting self-reflection and emotion regulation in online news commenting. In: Proceedings of the 2021 CHI conference on human factors in computing systems. pp. 1–13 (2021)
25. Kou, Y., Gui, X.: Emotion regulation in esports gaming: a qualitative study of league of legends. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW2), 1–25 (2020)
26. Kumar, S., Cheng, J., Leskovec, J.: Antisocial behavior on the web: Characterization and detection. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 947–950 (2017)
27. Lukoff, K., Yu, C., Kientz, J., Hiniker, A.: What makes smartphone use meaningful or meaningless? Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(1), 1–26 (2018)
28. Luo, Ruikun, N.D., Yang, X.J.: Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. IEEE Transactions on Affective Computing (2021)
29. Maarouf, A., Pröllochs, N., Feuerriegel, S.: The virality of hate speech on social media. arXiv preprint arXiv:2210.13770 (2022)
30. Manikonda, L., Beigi, G., Liu, H., Kambhampati, S.: Twitter for sparking a movement, reddit for sharing the moment:# metoo through the lens of social media. arXiv preprint arXiv:1803.08022 (2018)
31. McRae, K., Gross, J.J.: Emotion regulation. Emotion **20**(1), 1 (2020)
32. Mohammad, S.M., Turney, P.D.: Nrc emotion lexicon. National Research Council, Canada **2**, 234 (2013)
33. Ochsner, K.N., Gross, J.J.: The cognitive control of emotion. Trends in cognitive sciences **9**(5), 242–249 (2005)
34. Park, J.S., Seering, J., Bernstein, M.S.: Measuring the prevalence of anti-social behavior in online communities. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–29 (2022)
35. Phillips, W.: This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture. Mit Press (2015)
36. Ruensuk, M., Cheon, E., Hong, H., Oakley, I.: How do you feel online: Exploiting smartphone sensors to detect transitory emotions during social media use. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **4**(4), 1–32 (2020)

37. Saveski, M., Roy, B., Roy, D.: The structure of toxic conversations on twitter. In: Proceedings of the Web Conference 2021. pp. 1086–1097 (2021)

38. Shen, K., Cox, A.: Video games as a tool for digital emotion regulation. HCI-E MSc Final Project Report 2020 (2020)

39. Slovak, P., Antle, A.N., Theofanopoulou, N., Roquet, C.D., Gross, J.J., Isbister, K.: Designing for emotion regulation interventions: an agenda for hci theory and research. arXiv preprint arXiv:2204.00118 (2022)

40. Smith, W., Wadley, G., Webber, S., Tag, B., Kostakos, V., Koval, P., Gross, J.J.: Digital emotion regulation in everyday life. In: CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2022)

41. Solovev, K., Pröllochs, N.: Moral emotions shape the virality of covid-19 misinformation on social media. In: Proceedings of the ACM Web Conference 2022. pp. 3706–3717 (2022)

42. Thomas, K., Kelley, P.G., Consolvo, S., Samermit, P., Bursztein, E.: "it ' s common and a part of being a content creator" : Understanding how creators experience and cope with hate and harassment online. In: CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2022)

43. Torre, J.B., Lieberman, M.D.: Putting feelings into words: Affect labeling as implicit emotion regulation. Emotion Review **10**(2), 116–124 (2018)

44. Trujillo, A., Cresci, S.: Make reddit great again: assessing community effects of moderation interventions on r/the_donald. Proceedings of the ACM on Human-computer Interaction **6**(CSCW2), 1–28 (2022)

45. Verma, A., Islam, S., Moghaddam, V., Anwar, A.: Encouraging emotion regulation in social media conversations through self-reflection. In: 2023 IEEE Engineering Informatics. pp. 1–8 (2023). https://doi.org/10.1109/IEEECONF58110.2023.10520471

46. Wadley, G., Smith, W., Koval, P., Gross, J.J.: Digital emotion regulation. Current Directions in Psychological Science **29**(4), 412–418 (2020)

47. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L., et al.: Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB **2**(0), 1–7 (2009)