arXiv:2601.00927v1 [cs.SI] 2 Jan 2026

使用大语言模型和启发式规则测量社交媒体极化

Jawad Chowdhury、Rezaur Rashid 和 Gabriel Terejanu

北卡罗来纳大学夏洛特分校计算机科学系，美国北卡罗来纳州夏洛特市 28223{mchowdh5, mrashid1, gabriel.terejanu}@charlotte.edu

摘要。 理解在线话语中的情感极化，对于评估社交媒体互动的社会影响至关重要。本研究提出了一种新颖的框架，利用大语言模型（LLM）和领域知识启发式方法，系统地分析并量化气候变化和枪支管制等争议性议题讨论中的情感极化。与大多数依赖情感分析或预定义分类器的既有方法不同，我们的方法整合大语言模型，从大规模社交媒体讨论中提取立场、情感基调和一致模式。随后，我们采用一种基于规则的评分系统，根据立场一致性、情绪内容和互动动态，对情感极化进行量化，甚至适用于仅包含单次互动的小型对话。分析揭示了不同的极化模式，这些模式受事件影响：（i）预期驱动型极化，即在广为宣传的事件发生前，极端极化不断加剧；（ii）反应性极化，即突发且影响重大的事件发生后，强烈的情感极化立即飙升。通过结合人工智能驱动的内容标注与领域知识评分，我们的框架提供了一种可扩展且可解释的情感极化测量方法。源代码公开发布于：https://github.com/hasanjawad001/llm-social-media-polarization

关键词： 情感极化、社交媒体话语、大语言模型、立场检测、用于社会影响的人工智能

1 引言

近年来，社交媒体平台的兴起改变了政治话语，使实时信息交流成为可能，也让更多受众参与其中 [9, 25]。在不断演变的媒体格局推动下，这一转变持续影响信息的生产、传播和消费方式，同时重新定义了人们在数字空间中的互动与维系联系的方式 [4–6, 16]。尽管这些平台促进了参与，但也加剧了意识形态分歧，因为算法内容策展会优先呈现与用户既有观点一致的内容，从而强化既有信念、限制人们接触多元观点。这种选择性接触会助长情感极化，使人们对群体内成员产生强烈的积极情感，而对

对立群体成员或持有对立观点的人则表现出敌意 [8, 19, 26]。研究表明，这种极化不仅受政治意识形态影响，也受到在线讨论中的情感语调、话语结构和互动模式的影响。情感极化与政治激进化加剧、两党合作减少以及错误信息传播有关 [31, 33]。了解其动态对于评估在线话语更广泛的社会影响至关重要。

Twitter（现为 X）等社交媒体平台会通过算法内容策展进一步加剧这些分歧。此类策展优先推荐能够提升互动的内容，且往往推广煽情、加剧极化的内容 [5,21, 24]。研究表明，在线回音室主要让用户接触与自身观点相近的看法，同时限制其与对立观点的互动，从而加剧极化 [10, 12]。然而，接触反意识形态内容有时虽能纠正错误认知并减少极化 [7],，但也可能引发防御性反应，尤其是在争议激烈的社会运动中；在此类运动中，意识形态冲突往往伴随着有毒互动和数字攻击 [27]。这些动态凸显了社交媒体在塑造意识形态分歧方面所扮演的复杂角色，也说明我们迫切需要可靠的方法来量化和分析在线话语中的情感极化。

现有的社交媒体情感极化测量方法主要依赖情感分析、立场检测和／或基于网络的极化指数 [11, 17, 23, 29]。情感分析技术将文本归类为正面、负面或中性，以大致反映其情感语调，但往往无法捕捉政治话语的复杂性，例如讽刺或隐性偏见 [22]。立场检测方法旨在判断用户对某个议题是支持、反对还是持中立态度，但常常难以处理语言细微差别，尤其是在立场表达较为隐晦的高度极化辩论中 [1, 18]。近期研究结合了多模态信号 [28] 或社交网络结构 [14, 32] 来改进测量，但在捕捉推动极化的细腻情感维度方面仍存在局限。

近年来，大语言模型（LLM）已成为文本分类和内容分析的强大工具，与传统情感分类器相比，能够提供更深入的语境理解 [2, 3, 34]。大语言模型利用海量数据，更好地捕捉在线话语中意识形态框架化、修辞策略和情感暗流的细微变化 [13, 30]。它们能够以更注重语境的方式处理语言，因此特别适合分析情感变化、情绪强度以及情感极化。研究已成功运用大语言模型检测政治偏见、错误信息和意识形态框架化，展现了其在大规模社交媒体分析中的潜力 [15, 20, 35]。然而，尽管取得了这些进展，大语言模型在系统量化情感极化方面的应用仍未得到充分探索；情感极化的量化涉及对立场一致性和情绪强度的衡量。要弥补这一空白，需要采用协同方法，将大语言模型驱动的标注与领域专业知识相结合，以确保方法更

具备可解释性和可扩展性，从而衡量不同社交媒体平台上的情感极化。

在这条研究脉络的基础上，我们的研究提出了一种与现有方法显著不同的做法。Rashid 等人 [23] 开发了一种反事实框架，利用基于网络的极化指数 [29] 考察有影响力的用户在塑造 Twitter（现称 X）上的极化方面所发挥的作用。他们的研究评估了移除有影响力的对话如何改变极化评分，但依赖预先定义的基于情感的分类器。相比之下，我们的研究采用基于大语言模型的标注，从大规模讨论中提取立场、情感和一致模式，从而能够更细致地理解情感极化。

此外，我们整合启发式规则来评估极化，以解决现有基于情感的方法的局限，并确保衡量话语强度时的可解释性。具体而言，我们的大语言模型驱动框架根据推文的立场（支持、反对或对某议题持中立态度）、情感内容（是否存在表明极化的情绪激烈语言）以及一致模式（回复与原始帖子的立场一致或冲突的程度）对推文进行分类。随后，这些提取出的属性将用于结构化评分系统，通过评估立场一致性、情绪强度和分歧动态来量化极化。

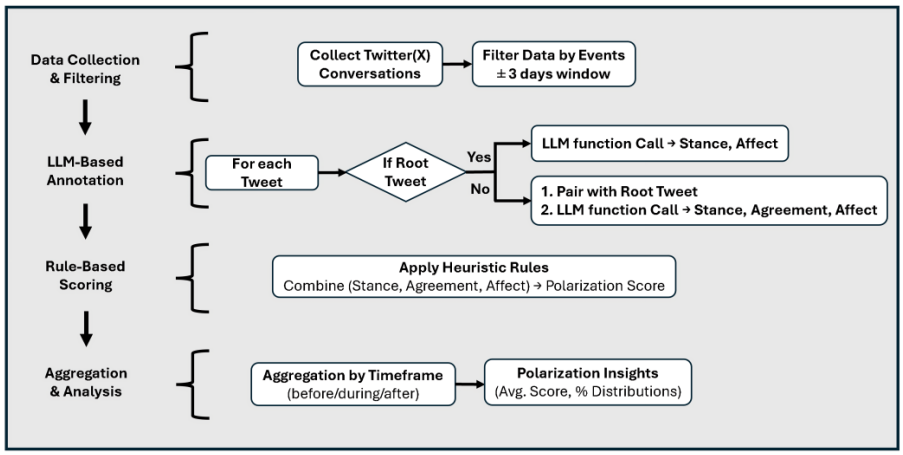

图1. 我们方法的详细工作流程管线，展示了数据收集与筛选、基于大语言模型的标注、启发式规则的应用，以及用于生成极化洞见的聚合过程。

通过结合大语言模型与领域知识启发的启发式方法，我们的方法相比传统的基于情感的技术，能够以可扩展、可解释且更具情境感知能力的方式衡量极化。本研究的主要贡献如下：

– 我们提出了一种用于衡量情感极化的新型基于大语言模型的框架，通过融入大规模语言理解，改进了传统的基于情感分析的方法。– 我们引入了一套评分系统，用于系统地量化情感极化，并捕捉立场转变、情感强度和分歧模式等细微的话语动态。与基于网络的极化测量方法不同，我们的评分系统具有可解释性，即使在仅包含一次互动的极小规模对话中，也能有效量化极化。– 我们对围绕气候变化和枪支管制的高度争议性讨论中的情感极化进行了大规模实证分析，揭示了关键的事件驱动的极化趋势，并区分了预期性极化与反应性极化。

本文其余部分安排如下：第2节详细介绍我们提出的框架，涵盖数据收集、基于大语言模型的标注，以及使用启发式规则进行情感极化评分。第3节展示我们的研究发现，分析情感极化如何随着时间推移而因重大事件发生变化。最后，第4节总结全文，并概述未来研究方向。

2 提出的方法与实现

本研究采用大语言模型（LLM）分析围绕争议性议题展开的在线讨论中的情感极化。与以往依赖情感分类器或基于规则的立场检测的研究不同，我们提出一种混合方法，将开源预训练大语言模型 LLaMA 3.1 70B 用于自动文本分析，并结合预定义评分系统，以系统地量化情感极化。我们的方法利用大语言模型提取关键属性，例如立场、情感内容，以及用户之间和用户帖子之间的认同程度，同时依据互动模式，通过预定义规则赋予情感极化分数。

我们提出的工作所采用的整体工作流程管线如图 1 所示，主要包含四个阶段。**（1）数据收集：**获取与气候变化和枪支管制等备受争议的社会政治议题相关的 Twitter（X）对话串。**（2）基于大语言模型的标注：**使用 LLaMA 3.1 70B 提取推文之间的立场、情感和一致性。**（2）预定义规则应用：**根据上一步中由大语言模型提取的立场一致性、情感和一致性信息，应用基于领域启发式方法的规则，为情感极化评分。**（4）聚合与分析：**计算对话层面的极化分数，并评估随时间变化的极化趋势。

2.1 数据收集与结构

我们的研究采用了来自 Twitter（现为 X）的综合数据集，聚焦于两个争议性政治议题——气候变化和枪支管制，并采用关键词

检索方法。在 2023 年限制措施出台之前，我们通过 Twitter API 获取了初始推文集；依据针对各议题筛选整理的关键词和话题标签，包括“#气候危机”“#立即改革枪支政策”等词语，以及与重大事件相关的特定短语（例如政府间气候变化专门委员会报告、大规模枪击事件）。为完整捕捉用户互动，我们递归扩展对话链，将所有被引用的推文（回复、引用推文、父级推文）纳入其中，确保即使没有明确匹配关键词，也能涵盖相关讨论。

气候变化数据集涵盖2021年6月1日至2022年5月31日，包含来自480万名独立用户的4600万条推文，分布在至少包含三条推文的726,378个对话串中。枪支管制数据集涵盖2022年1月1日至2022年12月31日，包含来自266万名独立用户的1440万条推文，分布在335,000个对话串中。为聚焦于互动较多的对话串，并减轻琐碎或活跃度较低的对话所带来的影响，我们将对话串限制为包含 ≥ 20 条推文和 ≥ 10 名独立用户的对话串。最后，每个数据集均由结构化的对话串组成，我们将其定义如下：

– **父推文：** 发起讨论的原始帖子。– **子级推文：** 对父推文作出回应的回复。

我们对每一组推文—回复配对进行独立分析，以提取立场、情感和一致程度，并以此作为情感极化评分的基础。尽管数据集包含大量对话，我们的分析具体聚焦于八个关键事件：四个与气候变化相关，四个与枪支管制相关。对于每个事件，我们将对话划分为三个不同的时间段：

– **之前：** 从事件发生前3天起至事件开始前一天进行的对话。– **期间：** 在事件官方开始日期与结束日期之间进行的对话。– **之后：** 从事件结束后的第二天起至结束后3天内进行的对话。

这种划分使我们能够分析情感极化如何随时间演变，尤其是极化是否会在事件前的预期阶段加剧（之前）、在事件期间达到峰值（期间），或在事件结束后因其余波而升级（之后）。这些事件的选择详见第3节。通过使用这些时间划分构建数据集，我们得以细致呈现在线话语动态，从而比较不同事件类型及其对应的极化水平变化。

2.2 基于大语言模型的分类

我们采用了LLaMA 3.1 70B，这是一种开源预训练模型，专门用于理解和生成类似人类的文本，并结合结构化提示工程进行分类：

– **推文立场：** 相信、不相信、不知道（针对气候变化）；支持、反对、不知道（针对枪支管制）。

– **推文情感：**推文是否包含表明情感极化的情绪激烈的语言。– **同意程度：**子级推文是否赞同或反对父推文。

为确保输出具有确定性，我们将大语言模型的温度设为零，以尽量减少分类中的随机性。

2.3 基于启发式规则的情感极化评分

基于上一节从大语言模型中提取的属性，我们使用启发式规则定义情感极化评分。这些规则涵盖父推文（原始帖子）与子级推文（回复）之间互动动态的关键方面，并考虑三个主要因素：立场一致性，用于判断回复推文是支持还是反对父推文的立场；情感表达，用于评估任一推文是否呈现出表明极化的强烈情绪化语言；以及同意程度，用于评估回复推文是否明确同意或反对父推文。通过结合基于大语言模型的分类与结构化启发式规则，我们能够对社交媒体话语中的情感极化进行系统、可扩展且可解释的量化。

在表1所示的评分系统中，低分（0至4分）代表建设性的互动和相互尊重的分歧。具体而言，0分代表理想情形，体现了即使观点不同，人们仍积极寻求相互理解与合作、促进开放交流和健康对话的文明互动。相反，高分（8分和10分）代表由情感极化主导的互动，包括激烈冲突、失礼行为和情绪敌意。10分尤其令人担忧，表明互动完全局限于意识形态回音室之中，在缺乏相反观点的情况下强化负面情绪，进而加剧极化，并可能引发更多敌意。高极化评分，尤其是表1评分范围的上半区，代表互动从富有成效的对话转向情绪化反应和对外群体敌意，损害有效传播并加深不宽容。0分与10分之间的鲜明对比凸显了促进相互尊重、多元话语的重要性。

重要的是，我们提出的启发式方法评分方法相比传统统计方法具有显著优势，能够有效量化极化程度，即使是在仅包含一次互动的小型对话中也是如此。对于较长的对话，则通过对所有单次互动的极化评分取平均值来计算总体极化评分，从而确保其能够适应不同长度和结构的对话，并具备可扩展性。

2.4 一个示例：提示、标注与评分

为了展示基于大语言模型的标注与预定义评分的综合效果，我们使用 LLaMA 3.1 70B 模型给出了一个示例。

表1. 用于极化评分的启发式规则。该表根据立场相似性、情感是否存在以及推文之间的一致性定义互动类别。

回复与父帖立场	**回复中的情感或家长**	**一致性评分、**	**话语质**	**量类别及描述**
相反立场	否（两者皆否）	Yes	0	**建设性对话** - 建设性讨论 没有负面情感，且彼此相互认同。
相同立场	否（两者皆否）	No	2	**友善的分歧**- 健康辩论 在相同立场群体内。
相反立场	否（两者皆否）	No	4	**相互尊重的分歧**- 礼貌地表达不同意见 跨越持对立立场的群体，且不带负面 情绪。
相同立场	两者均否	Yes	6	**回声式认同** - 回音室效应，但 不带情绪敌意。
相反立场	是（两者皆可）	Yes	6	**艰难达成的共识** - 跨越.....的一致 立场对立，但伴有负面情绪。
相反立场	是（两者皆可）	No	8	**激烈冲突**- 持对立立场的群体之间的敌对分歧 持对立立场的群体。
相同立场	是（任一方）	No	8	**不一致的盟友** - 但存在群体内部意见分歧 并伴有强烈的负面情绪。
相同立场	是（任一方皆可）	Yes	10	**极化回音室**- 存在.....的回音室 情感强化，且不与持相反立场者互动。 持相反立场者互动。

表2. 基于大语言模型的标注示例。本表展示了模拟真实讨论的推文，以确保匿名性，并列出大语言模型提取的属性，如立场、情感、一致性信息及其解释。

属性	**父推文**	**回复推文**
推文文本	‘科学结论明确。气候变化是真实存在的， 必须立即采取行动，否则为时已晚！’	‘气候变化是腐败精英推动的骗局 用来对大众进行洗脑。醒醒吧，傻瓜！’
立场 **解释**	**相信** 这条推文明确表示，气候变化是 真实存在的，表明其相信这一议题。	**不相信** 这条推文表达了对气候变化的不相信，称其为骗- 局，并暗示相信气候变化的人 都被洗脑了。语气带有对抗性，且 轻蔑不屑。
情感 **解释**	**No** 尽管这条推文的语气很紧迫，并且强调了采取行动的必要性， 但其中并不包含 针对反对观点的强烈负面情绪或态度。 opposing views。	**Yes** 这条推文对.....表达了强烈的负面情绪 持有气候变化对立观点的人， 称他们为 “傻瓜”，并暗示他们被洗了脑—— 这种说法带有强烈的情绪色彩，且具有分- 裂性。
一致 **解释**	**-** -	**No** 父推文和回复持有对立观点 ，讨论的是气候变化。回复否认气候变化的存在， 而父推文则予以肯定，并呼吁立即采取 行动。双方在这个话题上显然意见不一。

表2展示了一个仅包含两条推文的模拟对话线程中提取的立场、情感和一致性属性，而表1则列出了启发式规则。表2中的标注使用提示模板生成，模型收到的成对推文格式如下：

推文1： [回复推文]推文2：
[父推文]

如果推文1是父推文（即线程中的第一条推文），则推文2会以空字符串（""）的形式提供，并且一致性相关字段不适用。除了这些输入外，模型还收到了以下函数调用指令：

分析所提供推文的内容，评估其针对气候变化这一主题的立场和情感语调。请详细分类其立场、推文之间的一致性以及情感极化（对对立观点的情感负面态度）。

函数调用要求模型生成六个字段，每个字段都有具体指示：

– 推文1立场解释:_ 简要解释推文1在气候变化这一议题上的立场。如果推文1表达了相信气候变化是真实存在的，或表达了不相信，请说明其理由。如果立场不明确，请标注为不知道。– 推文1立场：对推文1在气候变化这一议题上的立场进行分类。可能的取值：相信、不相信、不知道。此分类应以“推文1立场解释”中提供的解释为依据。– 推文一致性解释：解释推文1和推文2在气候变化这一议题上是否一致。一致表示观点相似；不一致表示持有对立观点。如果推文2不可用，请注明“不适用”。如果是否一致不明确，请提供理由。– 推文一致：对推文1和推文2在气候变化这一议题上的一致性进行分类。可能的取值：是（一致）、否（不一致）、不知道（不明确）。– 推文1情感解释：说明推文1是否包含针对在气候变化这一议题上持有对立观点的人所抱有的强烈负面情绪或态度。关注点是立场本身之外的情感负面倾向。– 推文1情感：对推文1的情感极化进行分类，即针对气候变化这一议题上的对立观点所表现出的情感负面倾向。可能的取值：是（包含情感极化）、否（不包含）、不知道（不确定）。

–

该提示结构始终用于生成表2中及整个实验过程中呈现的标注。基于所提取的

属性，通过基于大语言模型的标注并对提取的特征应用启发式规则，将该回复推文分类为激烈冲突（得分 $= 8$），因为它：（1）持有相反立场；（2）包含情绪激烈的语言；（3）表达不一致，表明相反立场群体之间存在敌意。该示例简单展示了我们的方法，以及该框架系统评估和分类情感极化的能力。

2.5 实施细节

分类流程使用 Ollama 调用 LLaMA3.1 70B 实现立场和情感分类，并结合 LangChain 进行结构化提示工程和确定性响应生成。我们的初始数据集涵盖了气候变化相关事件中的 2,551 条对话，以及枪支管控相关事件中的 3,201 条对话，并考虑了每个事件的不同时间段。这种混合方法为大规模情感极化量化提供了一种高效且可解释的方法，将大语言模型的计算进展与人工指导的分析相结合。

3 案例研究与结果

为了系统地分析社交媒体上的情感极化，我们考察了前文提到的两个案例研究中的极化演变：气候变化和枪支管制。我们的框架量化了关键事件发生之前、期间和之后的极化情况，使我们能够观察在线话语中明显的时间模式。以下小节介绍各案例研究的发现，并重点说明不同类型的事件如何影响情感极化的动态。

3.1 气候变化话语中的情感极化

在本节中，我们将围绕重大气候变化相关事件展开的讨论作为本研究的案例之一进行分析。具体而言，我们考察了四个重要气候相关事件发生前、期间和之后情感极化水平的变化：政府间气候变化专门委员会（IPCC）评估报告发布、联合国（UN）环境大会、艾达飓风以及一场重大热浪。为提供结构化概览，表 3 列出了所考察事件的起止时间段。对于每个事件，我们计算了事件发生前、期间和之后所有对话的平均情感极化分数及标准误。结果如图 2 所示。如图所示，与事件发生前相比，这些事件期间及之后的情感极化分数明显上升。这符合预期，即重大气候相关进展会加剧公共话语中的情绪强度。此外，对话量（图中以‘n’表示）在这些事件期间及之后也显著增加，反映出社交媒体互动更加活跃。

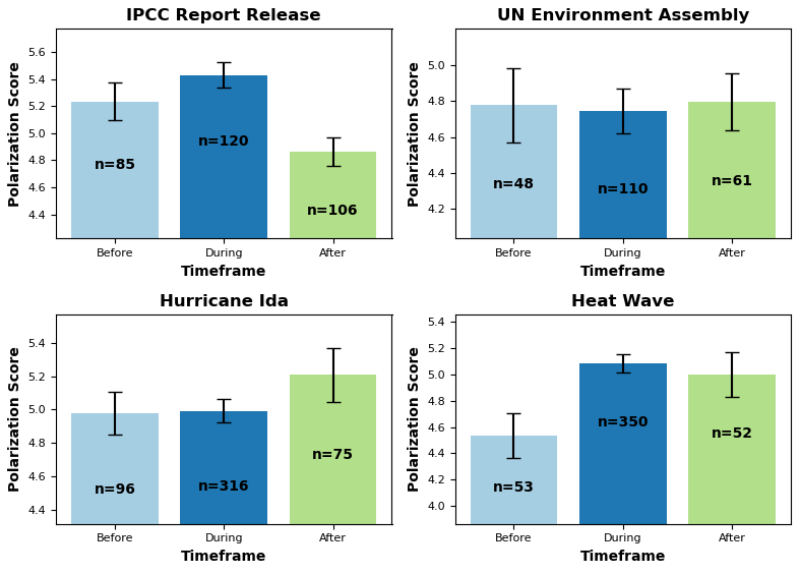

图 2. 四起与气候变化相关事件在不同时间段（之前、期间和之后）的情感极化得分。条形图展示了平均极化得分，误差条表示均值标准误。每个条形中的变量n 表示各事件在相应时间段内分析的对话数量。

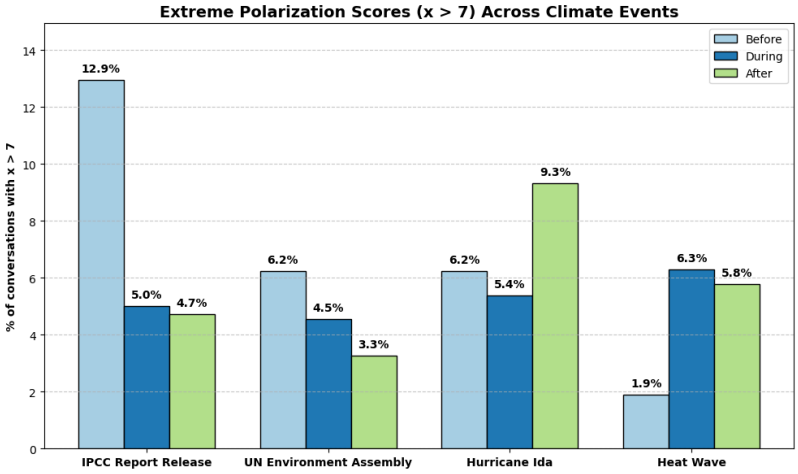

图 3. 与气候变化相关事件在不同时间段（之前、期间和之后）呈现极端极化得分（$x > 7$）的对话百分比。

为了更好地了解极化的强度，我们只关注呈现极端极化（得分高于 7）的对话。图 3 展示了每个气候相关事件在不同时间段（之前、期间和之后）高度极化对话的占比。分析揭示了鲜明的模式：政府间气候变化专门委员会评估报告发布和联合国环境大会等广为报道的事件，极端极化主要在事件发生前加剧，反映出人们的期待和意识形态参与。相反，艾达飓风和热浪等不可预测的气候事件，其极端极化程度主要在事件期间和之后激增，凸显了自发的情绪反应和反应性话语在塑造极化模式方面的作用。

表 3. 本研究分析的气候变化相关事件，包括其开始和结束日期。

事件	开始日期	结束日期
政府间气候变化专门委员会评估报告发布	2021年8月9日	2021年8月9日
艾达飓风	2021年8月26日	2021年9月4日
强热浪	2021年6月25日	2021年7月7日
联合国环境大会	2022年2月28日	2022年3月2日

这些发现表明，气候相关讨论中情感极化的时间特征与触发事件的性质密切相关。对于预先安排、由政策驱动的事件，预期性极化更为显著；而对于突发的气候相关灾害，反应性极化则占主导地位。

3.2 枪支管制话语中的情感极化

本节分析了围绕重大枪支管制相关事件的讨论，以考察大规模枪击事件发生后的情感极化趋势。具体而言，与前一个案例研究类似，我们分析了四起重大枪支事件发生前、发生期间及发生后的情感极化波动情况：得克萨斯州罗布小学枪击案、伊利诺伊州海兰帕克游行枪击案、马里兰州、伊利诺伊州和弗吉尼亚州多起枪击案，以及科罗拉多斯普林斯夜总会枪击案。表4列出了所研究事件的起止时间段。对于每起事件，我们计算了事件前、事件期间及事件后所有社交媒体对话中的平均情感极化得分和标准误。如图4所示，与事件发生前相比，事件期间及事件发生后的情感极化得分显著上升。这一模式表明，大规模枪击事件会加剧枪支管制讨论中的意识形态分歧，情绪强度在事件发生后达到峰值。事件期间及事件发生后的对话数量（图中以“n”表示）也大幅增加，反映出线上参与度有所提升。

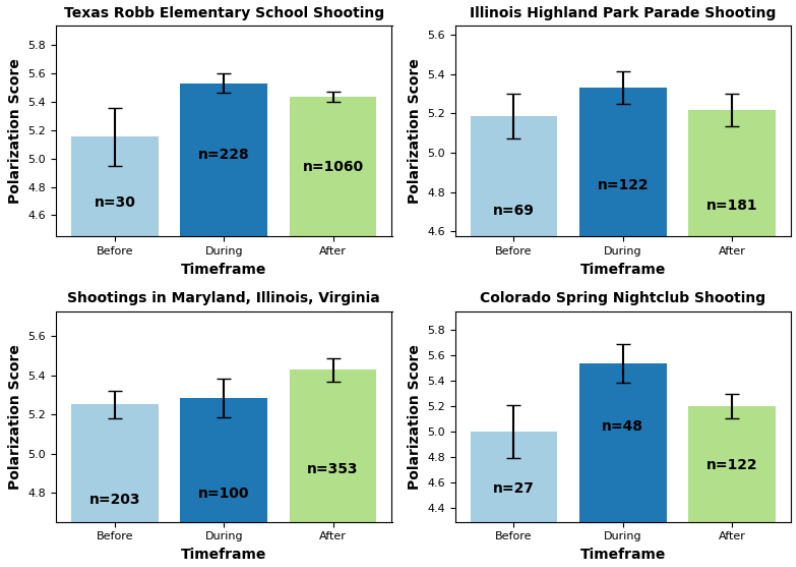

图4。 四起枪支管控相关事件在不同时间段（之前、期间和之后）的情感极化得分。条形图显示平均极化得分，误差条表示均值标准误。每个条形中的变量 n 表示相应事件在对应时间段内分析的对话数量。

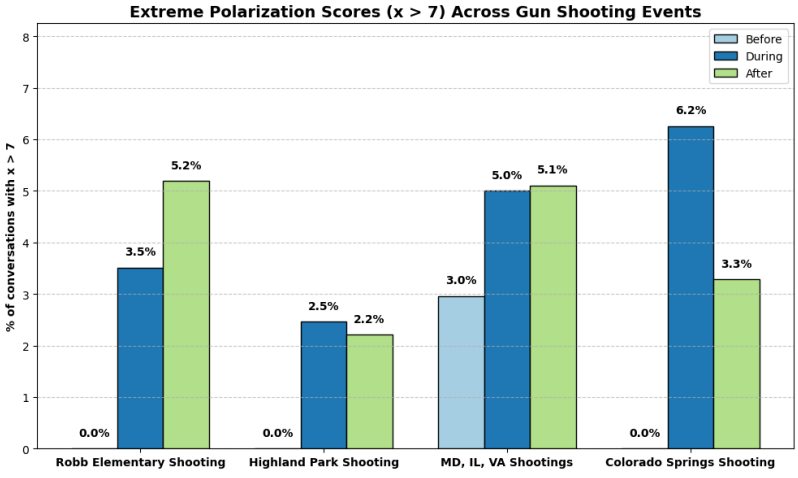

图5。 枪支管控相关事件在不同时间段（之前、期间和之后）中呈现极端极化得分（$x > 7$）的对话百分比。

表 4. 本研究分析的枪支管控相关事件，包括开始和结束日期。

事件	开始日期	结束日期
得克萨斯州罗布小学枪击案	2022-05-24	2022-05-24
伊利诺伊州海兰帕克游行枪击案	2022-07-04	2022-07-04
马里兰州、伊利诺伊州、弗吉尼亚州多起枪击案	2022-06-07	2022-06-07
科罗拉多斯普林斯夜总会枪击案	2022-11-19	2022-11-20

与我们之前的分析类似，我们仅考察了极端极化得分（高于 7 分）的对话。图 5 展示了四起与枪支管制相关事件发生前、期间和之后，高度极化对话所占的比例。分析清楚表明，在大规模枪击事件期间及之后，极端极化程度明显升高。与气候变化事件不同，后者往往在预期事件发生前出现极端极化程度的激增，而与枪支管制相关的讨论主要表现为反应性极化，即直接因枪击事件的发生而激增。

4 讨论

在本研究中，我们提出了一种新颖的框架，将大语言模型（LLM）与预先设定的启发式方法相结合，用于量化气候变化和枪支管制等敏感议题在线讨论中的情感极化。我们的混合方法利用大语言模型高效、大规模地提取立场、情感和一致性动态，同时由领域专家通过直观规则指导极化评分过程。我们的主要目标并非对大语言模型的性能进行基准测试，而是提出一个结构化的、由人在回路参与的框架，以加快情感极化分析。

我们的结果表明，气候变化和枪支管制相关事件会显著影响情感极化。气候讨论呈现出预期驱动型极化，在政府间气候变化专门委员会报告发布或联合国环境大会等事件发生前，就出现了极端观点。枪支管制辩论则呈现出反应型模式，极化主要在大规模枪击事件发生后加剧。这些事件发生后，对话量也大幅上升，凸显了在线话语中的现实触发因素。尽管大多数对话的极化程度仍属中等，但极端极化的激增与重大事件密切相关，从而让我们更深入地了解情感极化的时间动态。

未来的研究可借助我们提出的框架的普适性，将分析扩展到更广泛的事件和主题。还可以在多种社会政治背景下验证基于大语言模型的极化评分，并探索缓解极端极化的干预措施，从而为政策制定和社交媒体治理提供有价值的见解。总之，我们的研究凸显了将人工智能驱动的分析与基于领域知识的框架相结合的优势，使得即使在简短互动中，也能对在线极化进行可解释的测量。

致谢

本研究由美国陆军研究办公室通过拨款 W911NF-22-1-0035 资助。文中观点仅代表作者本人，并不一定反映美国陆军研究办公室或美国政府的官方政策。美国政府保留为政府用途复制和分发再版材料的权利。

参考文献

1. AlDayel, A., Magdy, W.：社交媒体上的立场检测：现状与趋势。《信息处理与管理》58(4)，102597（2021） 2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee- lakantan, A., Shyam, P., Sastry, G., Askell, A., 等：语言模型是少样本学习者。神经信息处理系统进展 33，1877–1901（2020） 3. Chae, Y., Davidson, T.：用于文本分类的大语言模型：从零样本学习到微调。开放科学基金会 10（2023） 4. Chambers, D.：社交媒体与人际关系：在线亲密关系与网络友谊。施普林格（2013） 5. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.：社交媒体上的回音室效应。《美国国家科学院院刊》118(9)，e2023301118（2021） 6. Dahlgren, P.：媒体与政治参与：公民、传播与民主。剑桥大学出版社（2009） 7. Druckman, J.N., Klar, S., Krupnikov, Y., Levendusky, M., Ryan, J.B.：（错误）估计情感极化。《政治学杂志》84(2)，1106–1117（2022） 8. Feldman, D., Rao, A., He, Z., Lerman, K.：社交网络中的情感极化。arXiv 电子预印本，第 arXiv–2310 页（2023） 9. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.：社交媒体上的政治话语：回音室、把关人与两党合作的代价。载于：2018 年万维网大会论文集，第 913–922 页（2018） 10. Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., Roy, D.：我、我的回音室与我：对社交媒体极化的反思。载于：2018 年万维网大会论文集，第 823–831 页（2018） 11. Guerra, P., Meira Jr, W., Cardie, C., Kleinberg, R.：一种基于社群边界衡量社交媒体网络极化程度的方法。载于：国际 AAAI 网络与社交媒体会议论文集，第 7 卷，第 215–224 页（2013） 12. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., Westwood, S.J.：美国情感极化的起源与后果。《政治学年度评论》22(1)，129–146（2019） 13. Khan, N.D., Khan, J.A., Li, J., Ullah, T., Zhao, Q.：利用大语言模型 ChatGPT 增进对社交媒体反馈中终端用户情绪的理解。《专家系统与应用》261，125524（2025） 14. Lerman, K., Feldman, D., He, Z., Rao, A.：情感极化与在线网络中的信息传播动态。《npj 复杂性》1(1)，8（2024） 15. Linegar, M., Kocielnik, R., Alvarez, R.M.：大语言模型与政治学。《政治学前沿》5，1257092（2023） 16. Marlowe, J.M., Bartley, A., Collins, F.：数字归属物：社会凝聚力、连通性与数字媒体的交汇。《族裔》17(1)，85–102（2017）

17. Mart´ınez-Espa˜na, R., Fern´andez-Pedauye, J., De Luc´ıa, J.G.P., Rojo-Mart´ınez, J.M., Bakdid-Albane, K., Garc´ıa-Escribano, J.J.：使用社交网络情感分析测量个体情感极化的方法。《IEEE Access》（2024） 18. Mets, M., Karjus, A., Ibrus, I., Schich, M.：复杂议题和小语种中的自动立场检测：极化性新闻媒体中移民议题的棘手案例。PLoS ONE 19(4)，e0302380（2024） 19. Overgaard, C.S.B.：感知美国的情感极化：社交媒体如何塑造元知觉与情感极化。《Social Media Society》10(1)（2024） 20. Pendyala, V.S., Hall, C.E.：阐释如何使用大语言模型检测错误信息。《Electronics》13(9)，1673（2024） 21. Piccardi, T., Saveski, M., Jia, C., Hancock, J.T., Tsai, J.L., Bernstein, M.：社交媒体算法可通过接触反民主态度与党派敌意塑造情感极化。arXiv 预印本 arXiv:2411.14652（2024） 22. Rani, N., Walia, R., 等：情感分析综述：技术、数据集、局限与未来展望。载于：2024 年第六届计算智能与通信技术国际会议（CCICT），第 403–409 页。IEEE（2024） 23. Rashid, R., Melton, J., Ghorbani, O., Krishnan, S., Reid, S., Terejanu, G.：量化网红对情感极化的影响。载于：2024 年国际机器学习与应用会议（ICMLA），第 1135–1140 页（2024） 24. Rodilosso, E.：过滤气泡与无感者：社交媒体人工智能如何助长极端主义与极化。Philosophy & Technology 37(2)，71（2024） 25. Saaida, M.：社交媒体在塑造政治话语与宣传中的作用。Science for All Publication 3(2)，1–8（2023） 26. Serrano-Puche, J.：数字虚假信息与情绪：探究情感极化带来的社会风险。《国际社会学评论》31(2)，231–245（2021） 27. Su´arez Estrada, M., Juarez, Y., Pi˜na-Garc´ıa, C.：有毒社交媒体：女权主义抗议后的情感极化。《Social Media Society》8(2)（2022） 28. Thareja, R.：社交媒体内容的多模态情感分析及其对心理健康的影响：极端情感研究。载于：第七届数据科学与数据管理联合国际会议论文集，第 469–473 页（2024） 29. Tyagi, A., Uyheng, J., Carley, K.M.：Twitter 上在线气候变化话语中的情感极化。载于：2020 年 IEEE/ACM 社交网络分析与挖掘国际会议（ASONAM），第 443–447 页。IEEE（2020） 30. Wang, X., Li, X., Yin, Z., Wu, Y., Liu, J.：大语言模型的情绪智力。《Journal of Pacific Rim Psychology》17（2023） 31. Yair, O.：关于情感极化文献的一点说明。SSRN 3771264（2020） 32. Yarchi, M., Baden, C., Kligler-Vilenchik, N.：数字空间中的政治极化：对社交媒体上互动、立场与情感极化的跨平台、跨时间分析。载于：Dissonant Public Spheres，第 185–226 页。Routledge（2024） 33. Yu, X., Wojcieszak, M., Casas, A.：社交媒体上的党派性：美国政治人物的党内爱意，以及普通用户对党外仇恨更高的参与度。《Political Behavior》46(2)，799–824（2024） 34. Zhang, W., Deng, Y., Liu, B., Pan, S.J., Bing, L.：大语言模型时代的情感分析：现实检验。arXiv 预印本 arXiv:2305.15005（2023） 35. Zhang, Y., Sharma, K., Du, L., Liu, Y.：迈向缓解大语言模型时代的错误信息与社交媒体操纵。载于：2024 年 ACM Web Conference 论文集，第 1302–1305 页（2024）