2025年加拿大选举中的深度伪造内容：普遍程度、党派倾向与平台动态

Victor Livernoche
Mila 与麦吉尔大学 蒙特利尔，加拿大

Andreea Musulan Mila、
IVADO 与蒙特利尔大学 蒙特利尔，加拿大

Zachary Yang
Mila 与麦吉尔大学 蒙特利尔，加拿大

Jean-François Godbout
Mila 与蒙特利尔大学 蒙特利尔，加拿大

Reihaneh
Rabbany Mila 与麦吉尔大学 蒙特利尔，加拿大

arXiv:2512.13915v1 [cs.SI] 15 Dec 2025

摘要

人们对人工智能生成的政治内容的担忧日益加剧，但对于在民主国家的重大事件期间，深度伪造内容实际上如何出现在社交平台上并传播，目前实证证据仍然有限。本研究聚焦于2025 年加拿大联邦大选，对这些逼真的合成媒体如何塑造网络政治格局进行了首批深入分析之一。我们分析了来自 X、Bluesky 和 Reddit 的**187,778 条帖子**，并使用基于多种现代生成模型训练的高准确率检测框架，发现**5.86%**的选举相关图像属于深度伪造内容。右翼倾向账户分享这类内容的频率更高，其发布的图像中有**8.66%**被标记，而倾向左翼的用户这一比例为**4.42%**；这些内容往往带有诽谤或阴谋论意图。然而，大多数检测到的深度伪造内容都是良性或非政治性的，有害内容也鲜少引起关注，在 X 上仅占全部浏览量的**0.12%** 。总体而言，深度伪造内容出现在选举讨论中，但传播范围有限；逼真的伪造图像虽然较少见，却获得了更高的互动，这凸显了人们对其可能被滥用的日益担忧。

CCS 概念

• **以人为中心的计算** → **社交内容共享**； •**网络** → 在线社交网络。

关键词

深度伪造检测、人工智能生成内容、社交媒体

1 引言

近年来，在 Grok、ChatGPT 和 Gemini 等大型多模态模型推动下，图像和视频生成技术取得了长足进展，并已直接集成到 X、Android 和网页浏览器等主流平台中，使普通用户能够轻易创建出逼真却完全伪造的内容。随着这些工具日益普及，其生成内容正被广泛分享到社交媒体上，而且通常不会披露其人工生成的来源。

生成式技术的快速传播对公共话语产生了深远影响。深度伪造，即人工智能生成媒体，可能侵蚀公众信任、放大错误信息，并被战略性地武器化，用于嘲讽、颂扬或操纵政治人物。这类媒体的泛滥提出了一个紧迫问题：如何深度伪造内容正被用于现实世界中的政治传播，以及它们在关键民主时刻会在多大程度上塑造信息生态系统？

为回答这一问题，我们分析了2025年加拿大联邦选举前后一周的社交媒体活动，时间跨度为4月24日至4月29日，涵盖 X、Bluesky 和 Reddit 三大平台。我们使用专门针对真实场景中的合成媒体检测而训练的先进深度伪造检测器，对每条包含图像的帖子进行评估。除了检测之外，我们还使用视觉语言模型推断每条帖子的传播意图（例如阴谋论、诽谤、宣传性），并使用大型语言模型推断作者的政治倾向。这种多层分析使我们能够量化深度伪造内容在在线政治话语中的普及程度及其目的。具体而言，我们的研究旨在回答以下问题：

- **研究问题 1：** 选举期间，主要社交媒体平台上的深度伪造内容有多普遍？
- **研究问题 2：** 在这一背景下传播的深度伪造内容具有哪些传播意图？
- **研究问题3：** 政治倾向是否会影响深度伪造内容被分享的可能性或方式？
- **研究问题4：** 深度伪造内容帖子在互动和可见度方面有何影响？

在这三个平台上，我们发现合成图像确实存在，但并不占主导地位：所有分享的图像中约有5.6%是深度伪造内容，其中*X* 占比最高。大多数案例是良性或非政治性的，不过右翼倾向账户中，针对政治人物或议题的内容更常见。不同平台生态系统存在差异，*X* 上有更多有害内容，而 Bluesky 和 Reddit 上的内容大多是非政治性的。深度伪造内容的传播范围也有限。与真实图像相比，它们获得的浏览量更少。总体而言，深度伪造内容出现在选举话语中，但在这一时期，其可见度和影响都较为有限。

2 相关工作

近期研究开始量化人工智能生成的视觉内容如何在政治场景中传播。Chen等人 [2] 记录了2024年美国大选期间政治性深度伪造内容在*X* 上的流行程度及传播情况；Chang等人 [1] 则展示了生成式工具如何沿着宣传、嘲讽和阴谋论思想等常见意图维度重塑模因框架。与此同时，Drolsbach和Pröllochs[4] 大规模分析人工智能生成的虚假信息，强调研究传播和互动的必要性，而不应只关注检测。

本作品采用知识共享署名 4.0 国际许可协议授权。



(a) 诽谤 (b) 阴谋论 (c) 宣传 (d) 伪造 (e) 良性 (f) 仇恨 (g) 非政治性

图1：本研究分析中使用的七种意图类别示例，展示了 AI 生成图像在政治和非政治用途上的多样性，涵盖从诽谤和阴谋论内容到良性和艺术创作。

在明确的政治背景之外，合成图像已成为以互动为导向的网络空间中的常见内容。DiResta和Goldstein [3] 介绍了通过批量制作吸引眼球的人工智能图像来扩大流量并实现变现的Facebook主页；对创意社群（例如艺术类subreddit）的研究则追踪了生成式媒体相关规范的演变及其引发的审核应对措施 [7]。另有一类研究聚焦于通过人工智能生成的头像进行身份欺骗：大规模审计发现，虚假Twitter/X账号广泛使用基于生成对抗网络的人脸 [8], ，而更广泛的生态系统分析则描绘了行为特征及其风险 [10]。这些研究关注账号层面的真实性，而我们关注的是帖子层面的内容及其传播作用。与以往研究相比，我们首次对与选举相关的深度伪造内容在**X、Bluesky和Reddit**上的情况进行同步跨平台测量，采用的现代检测器经过训练，涵盖了多种专有及开源生成器。以往研究往往依赖社区注释 [5], 、启发式标签 [3] , 或基于提示词的系统（例如 GPT-4o） [1, 2, 4]；我们的检测器在准确性和泛化能力方面更强。除了检测之外，我们还加入**视觉-语言意图分类、作者倾向推断**和**互动分析**，从而不仅能够刻画深度伪造内容出现在哪里，还能分析其意图传达什么，以及传播范围有多广。

3 方法论

我们分析了2025年加拿大联邦选举前后一周内，在三个主要社交媒体平台**X**、**Bluesky**和**Reddit**上分享的图像（4月24日至4月29日）。选举于4月28日举行，因此数据集涵盖了投票前夕及投票后的第一天。我们依据研究合规指南，使用官方 API 和各平台专用的抓取工具收集帖子。对于 X，我们订阅了 Pro API，以流式获取与选举相关的帖子，并根据一份全面的关键词列表进行筛选（例如“卡尼”、“博励治”、“弗里兰”、“*#cdnpoli*”、“*#canadapoli- tics*”）。我们仅保留包含图像的帖子，并排除了转发以及粉丝数少于100的用户发布的帖子，以减少垃圾信息。我们也在 Bluesky 和 Reddit 上使用各自的公共 API 进行类似的选举相关查询，同样筛选包含附件媒体的帖子（对于 Reddit，我们整理了一组 subreddit）。

3.1 深度伪造检测

我们使用主要基于 OpenFake 数据集构建的模型来检测人工智能生成图像 [6],该数据集构成了检测器的基础并且专门用于现实世界中政治错误信息情境下的深度伪造检测。OpenFake 将超过 **三百万张真实图像** 与约 **一百万张合成图像** 配对，这些合成图像由涵盖开源和专有架构的先进文生图系统生成（例如 SDXL、Imagen、DALL·E、Flux、Ideogram）。该数据集强调政治和新闻相关内容，其素材源自 LAION-400M 的图像说明，并通过视觉语言模型（ Qwen2.5-VL）进行筛选，以保留人脸和显著的社会政治场景。合成图像使用相同的图像说明生成，从而确保真实图像与伪造图像分布之间具有紧密的语义对齐。

我们使用 ConvNeXt-V2-Base 主干网络，在一个由超过**五百万张图像**组成的合并语料库上训练二元分类器，将 OpenFake 与 DF40 [9] 和 GenImage [11] 结合，以扩大对各种生成技术的覆盖范围。训练采用真实图像与合成图像混合采样策略，并进行了大量图像增强（JPEG 压缩、颜色抖动、随机裁剪），以模拟真实场景中的图像退化和篡改。

为验证模型的有效性，我们使用了 [6], 的真实场景评估划分，其中包含 1,057 张真实图像和 163 张深度伪造图像，这些图像均通过反向图像搜索和元数据检查进行了人工核验。该评估子集直接取自本研究分析的同一批社交媒体数据，确保模型验证与后续分析的领域一致性。ConvNeXt-V2 检测器在该评估集上的 F1 分数为 0.852；对真实图像的准确率为 99.2%（得分 ≤ 0.5），对伪造图像的准确率为 77.9%（得分 > 0.5）。这确保我们不会高估所报告的深度伪造检测统计数据。

3.2 意图分类

对于每张被标记为可能属于深度伪造的图像，我们使用视觉语言模型（VLM）Qwen3- VL-32B-Instruct，根据图像及其配文推断帖文的传播意图。提示词要求 VLM 提供简短的理由、类别名称和一个字母代码。我们确定了七种意图类别：诽谤性（嘲弄或攻击个人或群体）、阴谋论（暗示存在幕后操控或腐败）、宣传 （宣扬领导人、政党或意识形态）、伪造 （模仿逼真图像或新闻图片）、良性 （隐喻性／象征性或讽刺性的政治内容）、仇恨 （针对身份群体），以及非政治性 （艺术性内容或与政治无关的内容）。图 1 展示了相关示例，均直接取自我们收集的数据集中的帖文。

3.3 政治倾向分类

我们使用另一个大型语言模型（Llama 3.3 70B Instruct），根据用户近期的推文历史来估计其政治倾向。对于每位独立作者，我们按时间顺序整理其最近五条推文，并向模型提供一条系统消息，介绍加拿大政党格局（例如，将博励治和保守党归为右翼，将特鲁多、卡尼和自由党归为左翼，将加拿大人民党归为右翼，将新民主党和绿党归为左翼）。模型接到的指令是：在可能的情况下，总结作者的政治倾向。对 200 位作者进行人工检查后发现，标签准确，包括在适当情况下使用“未知”标签。主要局限在于，五条推文往往缺乏明确的政治线索，因此许多作者被标记为“未知”。

4 结果

我们首先测量了选举期间合成图像的总体普及率。在所有平台上，**5.86%** 的分享图像被检测为深度伪造内容，其中 X 平台的比例最高（*7.9%*），其次是 Reddit（**3.3%**）和 Bluesky（**2.2%**）。尽管深度伪造内容在全部视觉内容中仍占少数（见表1），但鉴于其所处的政治背景以及较短的观测窗口，其存在仍不容忽视。我们指出，深度伪造内容的总体普及率可能略高于我们报告的数值。如第3.1节所述，检测器无法检测出所有合成图像，因此我们报告的普及率低估了真实比例，不过这不太可能影响我们所报告的总体比较模式，例如不同平台和政治倾向之间的比较。

深度伪造普及率因推断出的政治倾向而异。如图2所示，右翼倾向的用户深度伪造率最高，其发帖内容中有**9.24%** 被标记为合成内容；相比之下，倾向左翼的用户为**3.87%**，中立或身份未明的账户为**5.03%**。不同政治倾向之间的意图分布总体相似，但右翼倾向的帖子更偏向阴谋论，而倾向左翼的帖子包含更多非政治内容。对发帖活动和深度伪造内容占比的详细分析证实，X 平台包含更大比例的右翼倾向账户，而 Bluesky 包含更大比例的左翼倾向账户；这为我们的政治倾向分类器按预期运行提供了基本的合理性检验。图3进一步展示了三个平台上深度伪造意图类别的相对频率。大多数检测到的深度伪造内容属于无害类别，通常是网络迷因、说明性内容或没有明确误导意图的幽默剪辑。X 平台上的深度伪造内容

平台	倾向	帖子总数	深度伪造内容	占比（%）
X	Left	27,306	1,761	6.45
	右翼	51,764	5,022	9.70
	未知	42,098	2,733	6.49
Bluesky	Left	36,903	720	1.95
	右翼	3,096	53	1.71
	未知	25,171	668	2.65
Reddit	Left	62	7	11.29
	右翼	41	0	0.00
	未知	1,337	41	3.07

表1：各平台及推断政治倾向的深度伪造普及率。百分比表示各组收集到的所有帖子中深度伪造内容所占的比例。

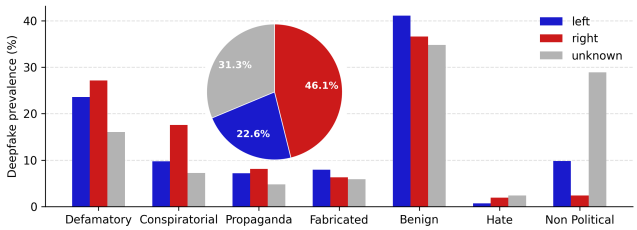

图2：不同政治倾向群体的深度伪造意图分布。柱状图显示左倾、右翼倾向和倾向未知账户发布的深度伪造内容中各意图类别的普及率，揭示了不同的叙事模式——右翼倾向账户发布的内容更侧重于诽谤和阴谋论内容。饼图概述了各倾向群体发布的深度伪造内容总体占比。

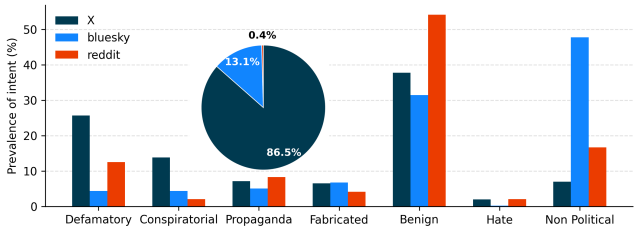

图3：各平台上深度伪造意图的普及率。柱状图显示在 X、Bluesky 和 Reddit 上检测到的深度伪造内容中，被归入各意图类别的百分比，凸显了选举期间各平台流传的合成内容类型存在差异。该直方图按平台归一化，嵌入的饼图则概述了各平台深度伪造内容的总体占比。

则更多带有诽谤性，而 Bluesky 上的内容则主要是非政治性或良性的。

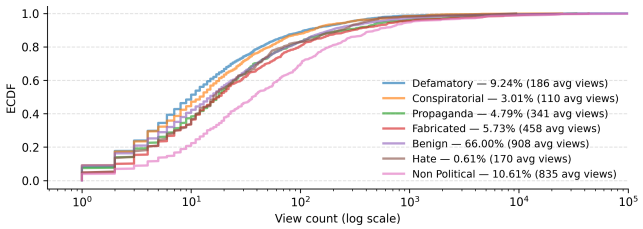

图4：不同意图的观看次数。每条曲线显示经验累积分布函数（ECDF）：观看次数低于给定值的帖子所占比例。曲线前段上升越陡，表示曝光度越低。非政治性深度伪造内容往往获得更高的观看次数，而政治性深度伪造内容的观看次数则较低。图例列出了各类别的深度伪造内容总观看次数占比，以及每条帖子的平均观看次数。

为评估传播范围，我们比较了发布六个月后*X*上的观看次数。图4和图5中的经验累积分布函数（经验累积分布函数）显示，在所有政治倾向中，深度伪造内容的帖子所获得的观看次数始终少于可比的非深度伪造帖子。检测到的深度伪造内容大多是良性的或非政治性的（两者合计占案例的**49.41%**），而深度伪造内容总体上仅占*X*全部观看次数的**0.52%**。因此，总观看次数中只有**0.12%** 对应政治有害深度伪造内容。良性和

非政治性的合成图像获得的曝光度中位数略高，而带有政治色彩的类别，尤其是诽谤性和仇恨内容，仍集中在低观看次数的尾部，几乎没有得到扩散。一个重要的例外是伪造类别；尽管其较不常见，但获得的观看次数超过了除良性和非政治内容之外的大多数类别。点赞数也呈现出类似模式，紧随观看次数变化，并显示出政治色彩深度伪造内容同样极低的互动。此外，来自政治倾向未知账号的深度伪造内容比来自倾向左翼或倾向右翼用户的内容获得了更多观看次数。*X* 上的深度伪造帖子和非深度伪造帖子被移除的比例几乎相同（13.8% 对 13.0%），表明深度伪造内容并未受到不成比例的移除处理。

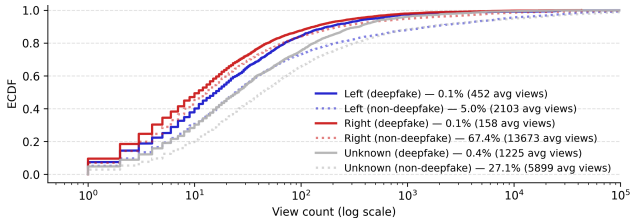

图5：不同政治倾向和深度伪造状态下的观看次数。在每种政治倾向中，深度伪造内容曲线都比非深度伪造内容曲线更早上升，表明深度伪造帖子通常获得的观看次数较少。图例显示总观看量占比和每篇帖子的平均观看次数。

图6绘出了X平台上每位作者发布的深度伪造内容数量与其关注者数量之间的关系。发布深度伪造内容最频繁的用户关注者数量不多，这表明深度伪造内容的传播主要由小型或边缘账号推动，而非有影响力的人物。少数关注者数量较多的账号也会发布深度伪造内容，但频率不高。低频发布深度伪造内容的用户最常见的意图是良性，而一些高频发布深度伪造内容的账号主要分享诽谤或阴谋论内容。

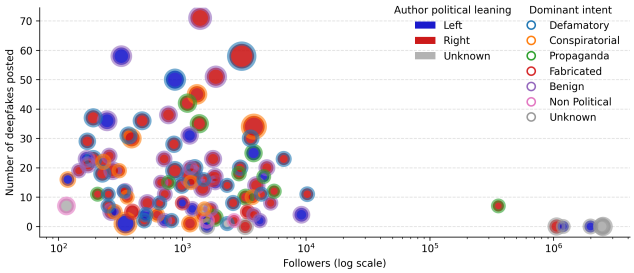

图6：深度伪造内容数量与作者关注者数量（对数刻度）。每个点代表X平台上的一位作者，颜色表示其政治倾向，点的大小表示帖子总数，环形颜色表示其主要的深度伪造用途。发布大量深度伪造内容的用户大多关注者数量较少，而关注者数量较多的账号发布的深度伪造内容很少。

5 讨论与结论

我们的结果表明，合成图像在 2025 年加拿大选举期间出现过，但并未显著影响在线关注度。深度伪造内容在各个平台传播，但大多数内容的受众规模都很小，且发布者是可见度较低的账号。这一模式表明，合成内容正逐渐成为日常在线表达的一部分，但目前其对大规模政治话语的影响仍然有限。

更重要的发现来自哪些深度伪造内容能够获得关注。高度逼真的伪造图像虽然相对少见，但比大多数政治导向类别获得了更多互动。由于这些视觉内容可能看起来像真实新闻，因此更容易误导用户，并引发级联式转发。即使总体数量不大，这些图像的可信度也会在政治紧张局势期间造成不成比例的脆弱性。数量不多但可信度高的合成内容仍可能影响人们的看法，尤其是在被策略性地投放，或在联系紧密的社群中传播时。威胁并非来自大量涌现的深度伪造内容，而是来自那些看起来足够可信、足以被当作真实证据的定向伪造内容。

我们的估算较为保守，因为检测器更注重准确率，因此会漏掉一些合成图像。许多被归类为非政治性深度伪造内容的项目只是简单的标志编辑或叠加图层，而 Reddit 图像数量较少，也降低了其结果的统计显著性。这些注意事项不会改变整体趋势，但凸显了开发更强大且更易获取的检测工具的必要性。随着合成媒体变得更容易制作，也更难与现实区分，我们需要可靠且公众可用的系统来标记深度伪造内容，并维护人们对民主信息的信任。

致谢

本研究得到CIFAR AIChairs计划、CSDC、IVADO和加拿大首席研究卓越基金的支持。我们还感谢Mila提供资金和计算资源支持，并感谢Fuxiao Gao协助收集数据。

参考文献

[1] Herbert Chang、Benjamin Shaman、Yung-Chun Chen、Mingyue Zha、Sean Noh、Chiyu Wei、Tracy Weener和Maya Magee。2024。生成式模因：人工智能在2024年美国总统选举中介入政治信息。可在 *SSRN* 获取 （2024）。[2] Zhiyi Chen、Jinyi Ye、Beverlyn Tsai、Emilio Ferrara和Luca Luceri。2025。合成政治：X平台上人工智能生成的政治图像的流行程度、传播者及情感反应。载于第*36*届*ACM*超文本与社交媒体会议论文集。11–21。[3] Renee DiResta和Josh A. Goldstein。2024。垃圾信息发送者和诈骗者如何利用Facebook上的人工智能生成图像来扩大受众。arXiv: 2403 12838 [ ] 2403 12838 。cs.CY https://arxiv.org/abs/ 。[4] 2025 Chiara Drolsbach和Nicolas Pröllochs。。社交媒体上人工智能生成的错误信息特征分析。*arXiv*预印本 *arXiv:2505.10266* （2025）。[5] Chiara Drolsbach和Nicolas Pröllochs。2025。社交媒体上人工智能生成的错误信息特征分析。arXiv:2505.10266 [cs.SI][6] Victor Livernoche、Akshatha Arodi、Andreea Musulan、Zachary Yang、Adam Salvail、Gaétan Marceau Caron、Jean-François Godbout和Reihaneh Rabbany。2025。OpenFake：面向真实场景深度伪造检测的开放数据集与平台。arXiv: 2509 09495 [ ] 2509 09495 。cs.CV [7] https://arxiv.org/abs/ 。Hana Matatov、Marianne Aubin Le Quéré、Ofra Amir和Mor Naaman。2025。考察艺术类Reddit子版块中人工智能生成媒体的流行程度与动态变化。arXiv:2410.07302 [cs.AI] https://arxiv.org/abs/2410.07302[8] Jonas Ricker、Dennis Assenmacher、Thorsten Holz、Asja Fischer和Erwin Quiring。2024。现实世界中的人工智能生成的人脸：Twitter个人资料图像的大规模案例研究。载于第*27*届国际攻击、入侵与防御研究研讨会论文集。513–530。[9] Zhiyuan Yan、Taiping Yao、Shen Chen、Yandan Zhao、Xinghe Fu、Junwei Zhu、Donghao Luo、Chengjie Wang、Shouhong Ding、Yunsheng Wu等。2024。DF40：迈向新一代深度伪造检测。神经信息处理系统进展37（2024），29387–29434。[10] Kaicheng Yang、Danishjeet Singh和Filippo Menczer。2024。带有人工智能生成的人脸的虚假社交媒体个人资料的特征与流行程度。《在线信任与安全期刊》2，4（2024年9月）。doi:10.54501/jots.v2i4.197[11] Mingjian Zhu、Hanting Chen、Qiangyu Yan、Xudong Huang、Guanyu Lin、Wei Li、Zhijun Tu、Hailin Hu、Jie Hu和Yunhe Wang。2023。GenImage：用于检测人工智能生成图像的百万规模基准。arXiv:2306.08571 [cs.CV]