# HateMM：用于仇恨视频分类的多模态数据集

**Mithun Das**[1], **Rohit Raj**[1], **Punyajoy Saha**[1], **Binny Mathew**[1], **Manish Gupta**[2], **Animesh Mukherjee** [1]

[1] 印度理工学院（IIT），卡拉格普尔 [2] 微软，印度{
mithundas,rrohit2901,punyajoys,binnymathew}@iitkgp.ac.in,
gmanish@microsoft.com, animeshm@cse.iitkgp.ac.in

arXiv:2305.03915v1 [cs.CV] 6 May 2023

**摘要**

仇恨言论已成为现代社会最重要的问题之一，对线上和线下世界都产生了影响。因此，近年来仇恨言论研究受到广泛关注。然而，大多数研究主要集中于文本媒体，涉及图像的工作相对较少，涉及视频的研究则更少。因此，我们需要早期阶段的自动化视频审核技术，以处理持续上传的视频，维护平台的安全与健康。为检测并移除视频分享平台上的仇恨内容，我们的工作聚焦于利用多种模态进行仇恨视频检测。为此，我们从 BitChute 收集了 ∼ 43小时的视频，并将其人工标注为仇恨或非仇恨，同时标注能够解释标签判定的帧区间。为收集相关视频，我们利用仇恨词典中的搜索关键词。我们观察到仇恨视频的图像和音频中存在各种线索。此外，我们构建了深度学习多模态模型来分类仇恨视频，并发现使用视频的所有模态能够提升整体仇恨言论检测性能（准确率=0.798，宏平均F1分数=0.790），与宏平均F1分数最高的单模态模型相比提高了 ∼ 5.7%。总之，我们的工作迈出了第一步，旨在理解并建模 BitChute 等视频托管平台上的仇恨视频。

## 引言

**免责声明：**本文包含许多人会认为具有冒犯性或仇恨性的内容；然而，鉴于本研究的性质，这一点无法避免。

社交媒体平台允许用户自行发布内容。预计到2023年，视频将占消费者互联网流量的82%（Wilson 2022），因此，YouTube、Dailymotion等视频托管平台已成为重要的信息来源。仅在YouTube上，人们每天观看的视频时长就超过10亿小时[1]。这类视频的病毒式传播特性是一把双刃剑：一方面，它有助于新闻迅速传播；另一方面，也可能迅速传播仇恨或错误信息。这些视频涵盖范围广泛的主题。虽然YouTube上的大多数内容无害，但仍有一些视频违反了社区准则（O'Connor 2021）。对于某些替代性视频托管平台（如BitChute[2], 、Odysee[3] 等）上的这一问题则更为严重。YouTube、Facebook、Twitter 等平台实行严格的审核政策，而这些另类科技平台[4] 允许用户发布任何内容，几乎不进行审核。此类内容不被删除，可能会对用户和整个网站造成损害，并导致形成充斥着仇恨用户的回音室式敌对环境。这还可能造成收入损失，并招致罚款（Troianovski和Schechner，2017年）及诉讼。

一些平台雇用多名人工审核员来查找有害内容，并将其从平台上移除。然而，考虑到每天发布的内容数量，这是一项极具挑战性的工作。例如，Facebook 聘用了约 1.5 万名审核员，负责审查其人工智能和用户标记的内容（Koetsier 2021），而且每天约会犯下 30 万次内容审核错误。此外，审核员自身也面临情绪和心理创伤的风险（Newton 2019）。法律要求平台在规定期限内移除仇恨内容，这使问题进一步恶化。未能遵守这些规定可能会导致罚款（Troianovski 和 Schechner 2017）。YouTube 等平台已部署机器学习算法来检测仇恨内容，但规模较小的平台可能没有足够的收入或技术来开发用于视频中的仇恨言论检测的数据集和模型。因此，我们需要开发高效、开放的模型，以检测视频中的仇恨言论。然而，目前关于仇恨言论的研究大多集中于基于文本的模型（Badjatiya 等人，2017年；Cheng 等人，2020年；Juuti 等人，2020年；Kennedy 等人，2020年；Parikh 等人，2021年；Das、Banerjee 和 Mukherjee，2022年），基于图像的研究则寥寥无几（Yang 等人，2019年；Das、Wahi 和 Li，2020年；Gomez 等人，2020年；Kiela 等人（2020））。检测视频中的仇恨行为需要结合多帧视频处理与语音处理信号，因此无法直接套用基于图像的仇恨检测方法。

**研究目标与贡献**：在本文中，我们朝着为这一新颖问题设定提供端到端解决方案迈进了一步。我们发布了 HATEMM，一个视频合集。

版权所有 © 2023，美国人工智能促进协会（www.aaai.org）。保留所有权利。

[1]https://www.youtube.com/intl/en-GB/about/press/

[2]https://www.bitchute.com/[3]

https://odysee.com[4]

https://en.wikipedia.org/wiki/Alt-tech

已标注仇恨言论。该数据集包含时长为 ∼ 43 小时的视频，总计 ∼ 144K 帧。我们将 HATEMM 数据集公开[5]，以推动多模态仇恨言论检测的进一步研究。由于 BitChute 的内容审核力度较低，我们依托该平台收集数据。BitChute 于 2017年 推出，是一个类似于 YouTube 的视频托管与分享平台，在极右翼用户中颇受欢迎。

总体而言，我们做出了以下**贡献**。

- 我们整理并构建了已知规模最大的仇恨视频数据集之一，其中包含 1083 个视频，总时长为 ∼ 43 小时，涵盖 ∼ 144K 帧。每个视频均由标注员标注为仇恨或非仇恨，并标记出能够支持该判断的帧区间。标注员为单个视频完成标签标注所花费的平均时间约为该视频时长的两倍。

- 我们分别基于三种不同模态（文本、音频和视频）单独构建检测模型，并将这些模态联合起来构建检测模型[6]。我们表现最佳的融合模型（BERT ⊙ ViT ⊙ MFCC）结合了所有模态，其宏平均 F1 分数达到 0.790。仇恨类别的精确率和召回率分别为 0.742 和 0.758。在单一模态中，文本的 Transformer 编码和基于视频的特征似乎更有利于检测仇恨视频。
- 我们进一步对各模态的重要性进行了一些初步分析。我们观察到，当转录文本相对干净时，基于文本的模型表现良好。当视频中出现喊叫和攻击性表达时，基于音频的模型最为有效。最后，当视频中存在仇恨行为的视觉证据，且受害者出现在画面中时，基于视觉的模型表现最佳。
- 最后，我们根据帧区间分析模型表现，发现基于文本和视觉的模型最能利用这一信息。此外，当视频中的仇恨目标是“黑人”或“犹太人”时，基于视觉的模型表现最佳；而针对“其他”目标群体时，基于文本的模型表现出色。

## 相关工作

随着多模态数据的大量涌现，多模态深度学习已被应用于提升各种任务的准确率，例如视觉问答（Singh 等人，2019年）、虚假新闻/谣言检测（Khattar 等人，2019年）等。近年来，多模态仇恨言论检测逐渐受到关注，其将文本帖子与用户信息、网络信息等额外上下文（Cheng 等人，2020年；Founta 等人，2019年）或图像（Yang 等人，2019年；Das、Wahi 和 Li，2020年；Gomez 等人，2020年；Kiela 等人，2020年）相结合，以提高检测准确率。这类多模态方案通常采用 CNN、长短期记忆网络或 BERT 等单模态方法对文本进行编码，并使用 ResNet 或 InceptionV3 等深度 CNN 对图像进行编码，然后通过多模态融合来实现……

[5]https://doi.org/10.5281/zenodo.7799469[6]基线模型的源代码可在 https: //github.com/hate-alert/HateMM 获取




图1：仇恨视频示例。

采用简单拼接、门控求和、双线性变换或基于注意力的方法。ViLBERT 和 Visual BERT 等多模态双向变换器也已得到应用（Kiela 等人（2020））。

除以下三项研究外，几乎没有针对冒犯性/仇恨视频检测的研究——葡萄牙语研究（Alcˆantara、Moreira 和 Feijo 2020 年）以及英语研究（ Wu 和 Bhandary （2020年）；Rana 和 Jha（2022年））。不过，前两项研究（Alcˆantara、Moreira 和 Feijo 2020年；Wu 和 Bhandary（2020年））仅通过提取转录文本，使用文本特征进行分类。此外，其标注数据集规模不足500个视频。Rana 等人（2022年）的研究同时考虑了文本和音频特征，但其数据集并未公开，数据整理和标注步骤也未得到完整说明，数据集统计信息亦未准确披露。与我们的数据集不同，他们选择语音清晰的视频；相比之下，我们没有这类限制，因为仇恨内容也可以仅以视觉形式表达，而不伴随任何语音。据我们所知，我们是首个开展多模态仇恨视频检测实验的团队，同时利用文本、音频和视频这三种数据模态。为了充分利用所有模态，我们的标注比最先进水平的相关工作更加丰富，数据规模也更大。我们相信，我们的数据集及基于该数据集训练的基准模型将有助于审核员识别真正的仇恨案例，同时减少误报。

## HATEMM 数据集

### BitChute 平台

BitChute 是一个内容审核力度较低的社交视频托管平台，于 2020年推出，旨在成为 YouTube 的替代平台（Trujillo 等，2020年）。该网站于 2017年上线，如今正日益受到欢迎，并成为极右翼用户的“避风港”。BitChute 上仇恨内容十分普遍，还托管着多位被传统平台和受审核平台封禁的内容创作者（Labarbera（2020））。

## 数据收集

为了抽取视频进行标注，我们使用了（Mathew 等人，2020a）中的词典，该研究考察了 Gab 及其他另类右翼平台。这些词典由贬损性关键词/针对不同受保护群体的侮辱性称呼组成。

我们使用每个关键词在 BitChute 上进行搜索，并将返回的链接添加到数据库中。总共收集了$\sim 8K$ 个链接。接下来，我们使用 BitChute- dl 软件[7]下载视频。下载过程中，我们发现有 25% 的链接对应的视频无法找到。此外，还有一些视频已损坏。最终，我们获得了 $\sim 6K$ 个视频。

## 标注指南

以下所述的标注方案构成了标注员的主要指南，而编码手册则确保大家对标签描述有共同的理解。我们依据 YouTube 关于仇恨言论的政策[8]编制了编码手册（其中包含标注指南），用于识别仇恨内容。如果视频符合以下情况，我们就将其视为仇恨内容——

> “它宣扬歧视，或基于种族、族群、民族血统、国籍、宗教、残疾、年龄、退伍军人身份、性取向、性别认同等，对个人或群体进行贬低或羞辱。”

此外，我们还要求标注员标出仇恨视频中他们认为具有仇恨性质的部分（即帧区间，作为理由依据），以及视频针对的社群。我们认为，理由依据之后可以作为可解释性信号，而目标则可用于衡量检测算法是否开始偏向某些目标，这与（Mathew 等人，2020b）中提出的做法类似。

## 标注流程

**训练标注员** 标注流程由两名博士生担任专家标注员负责，并由四名本科生担任新手标注员执行。所有本科生都主修计算机科学，且都自愿参与这项任务，并在完全知情同意的情况下参与，任务结束后还获得了在线礼品卡作为报酬。两名专家标注员都曾有处理社交媒体有害内容的经验。为了培训标注员，我们需要一个试点金标数据集。为此，专家标注员首先标注了30个视频。初始样本包括20个仇恨视频和10个非仇恨视频。我们将这30个视频交给本科生标注员，让他们依据标注编码手册进行标注。他们完成标注后，我们与他们讨论了标注错误的案例，以提升他们的标注技能。

**批处理模式下的标注** 在上述训练之后，我们每周以批处理模式发布一组包含30个视频的视频集。考虑到在标注仇恨视频时，标注人员可能会产生“负面心理影响”（Ybarra 等人，2006年），我们建议他们在每个视频标注完成后至少休息10分钟

[7]https://pypi.org/project/bitchute-dl/ [8] https://support.google.com/youtube/answer/2801939?hl=en

| | Hate | 非仇恨 | 总计 |
|---|---|---|---|
| **数量** | 431（39.8%） | 652（60.2%） | 1083 |
| **总时长（小时）** | 18.39 | 24.87 | 43.26 |
| $\mu$ **视频长度** | 2.56 ± 1.69 | 2.28 ± 4.77 | 2.40 ± 3.86 |
| $\mu$ **理由长度** | 1.71 ± 1.27 | - | - |
| $\mu$ **#帧** | 154 | 137 | 144 |
| $\mu$ **#单词** | 228 | 209 | 217 |

表 1：HATEMM 数据集的基本统计数据。每秒采样帧。视频时长和理由长度以分钟为单位。len：长度。hrs：小时。 $\mu$：均值。

。我们进一步规定，每天标注的视频不得超过10个。最后，我们还定期与他们开会，以确保这些标注不会对他们的心理健康产生任何不利影响。

**标注工具** Toloka[9] 和 ANVIL（Kipp 2014）等现成工具支持物体检测等任务的标注，但不支持对视频中任何类型的片段进行标注。PAVS[10] 仅支持片段选择，但幸运的是它是开源的。因此，我们对 PAVS 进行了修改，使其支持片段标注、仇恨与否的视频标注以及目标社群标记。我们将公开标注工具（见图 2 中的工具截图），以促进后续研究。



图 2：（仇恨）视频标注工具截图。

每个视频均由两名独立标注员进行标注。他们被要求观看完整视频，并根据提供的指南选择合适的类别（仇恨或非仇恨）。标注员标注一个视频所需的平均时间约为视频时长的两倍 。标注者间一致性的 Cohen’s kappa 为$\kappa$=0.625。每批标注完成后，如果两名标注员的结果不一致，则由一名专家标注员对同一视频进行标注，以打破平局。最终得到的数据集包含 431 个仇恨视频和 652 个非仇恨视频，共计 1083 个已标注实例。

[9]https://toloka.ai/ml/computer-vision [10] https://github.com/kevalvc/Python-Annotator-for-VideoS

## 数据集统计

我们的最终数据集包含1083个视频，内容总时长超过 ～ 43小时。视频平均时长为 ～ 2.40 分钟，而仇恨视频的平均时长略长，为 ～ 2.56 分钟。总体而言，我们构建了一个大致均衡的数据集，其中39.8%的样本被标记为仇恨。其类别平衡性优于许多文本仇恨言论数据集。对于每个视频，我们还使用Vosk离线语音识别[11] 工具对音频进行转录。转录文本平均包含 ～ 217 个词。此外，与非仇恨视频的转录文本相比，仇恨视频转录文本中的词数略高（～ 228）。数据集的其他重要统计信息列于表1中。一些仇恨视频示例的截图如图1所示。

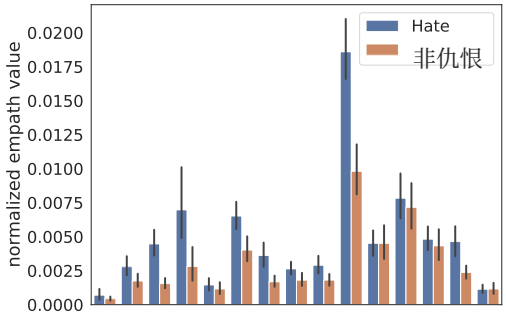



图 3：使用 Empath 对视频转录文本进行词汇分析。我们报告了 Empath 若干类别的平均值。仇恨视频转录文本在「攻击性」「脏话」「暴力」和「负面情绪」等类别中的得分显著较高。对于每个类别，我们均使用曼-惠特尼 U 检验，并标示显著性水平 ***($p < 0.0001$)、**($p < 0.001$)、*($p < 0.01$)。

## 数据集分析

**Empath分析** 为了更好地理解数据集，我们使用Empath（Fast、Chen和Bernstein（2016））识别视频转录文本中存在的重要词汇类别。Empath包含189个此类预构建类别。首先，我们忽略与仇恨言论无关的主题（例如技术和娱乐），选取70个类别。我们在图3中报告了差异最显著的前15个类别。仇恨视频转录文本在“攻击性”“脏话”“暴力”和“负面情绪”等类别中的得分显著较高。

**未登录词** 为了解数据集中的噪声程度（以及转录文本的质量），我们计算了“仇恨”和“非仇恨”两类中未登录词（OOV 词）的占比。为此，我们使用

[11]https://alphacephei.com/vosk/

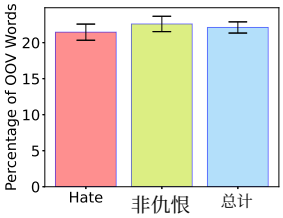



图4：数据集中存在的未登录词所占百分比。

PyEnchant 词典[12],，它是 Python 的拼写检查词典，用于识别标准英语词库中不存在的词。图 4 展示了“仇恨”和“非仇恨”两类中每个视频（转录文本）的未登录词占比。图中显示，“仇恨”和“非仇恨”两类的未登录词平均占比均约为 22%。

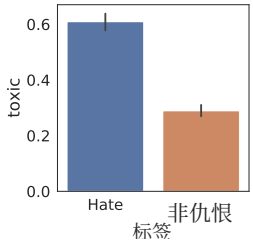



图 5：基于 Perspective API 的毒性比较。根据 Mann-Whitney U 检验，结果在 $p < 0.0001$ 水平上具有显著性。

**毒性得分** 检测有毒视频的一个简单方法是使用 Google 的 Perspective API[13] 分析转录文本；因此，我们测量了转录文本的毒性。如图 5 所示，仇恨视频转录文本的毒性几乎是非仇恨视频转录文本的两倍（～ 0.61 对比 ～ 0.28）。然而，仅依赖转录文本也有其缺点。如上所述，转录文本通常相当嘈杂，这说明仅凭转录文本可能不足以进行仇恨视频分类。

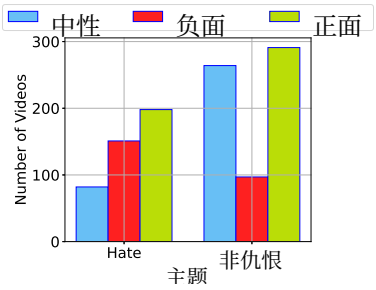



图 6：仇恨类别和非仇恨类别中不同情感的帖子数量。

[12]https://github.com/pyenchant/pyenchant [13] https://www.perspectiveapi.com/

| Hate | 非仇恨 |
|---|---|
| k k k k | joe |
| k k k | 乔·罗根 |
| k k | 乔治·齐默尔曼 |
| 山姆大叔 | 乔·拜登 |
| jack | bush |
| joe | 克里斯 |
| klan | mug |
| 罗伯特·霍尔斯 | mike |
| max | eric |
| k k k k k | 约翰逊 |

表 2：仇恨类和非仇恨类中出现频率最高的人物实体

**情感** 我们还测量了所标注视频所关联的情感。情感分析用于识别文本中的相关感受/情绪。情感分析包括三种极性：负面、中性和正面。在本研究中，采用基于词语的方法，根据所使用的词语，通过 -1 到 1 的分数确定每个转录文本的极性。负分表示负面情感，正分表示正面情感。情感分析使用 TextBlob API[14] 完成。图 6 表示两种视频类别中具有中性、负面和正面情感的视频数量。我们观察到，尽管总体而言具有正面情感的视频更多，但与“非仇恨”视频相比，“仇恨”视频转录文本具有更多负面情感。

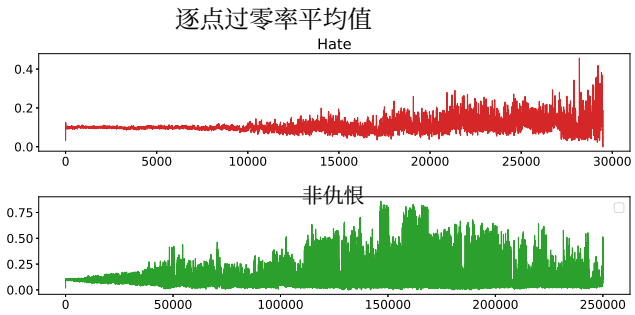



图 7：仇恨视频和非仇恨视频的过零率。

**命名实体识别分析** 我们还分析了与转录文本相关联的命名实体，以了解仇恨类和非仇恨类中不同命名实体识别标签的分布是否存在差异。为此，我们使用了 spacy 库[15]，它提供了一组实体标签。我们观察到，对于人物标签，仇恨类别中命名实体的归一化数量高于非仇恨类别。我们进一步检查了与实体类型“人物”相关联的出现频率最高的实体，这些实体列于表 2。对于仇恨类别，诸如以下短语】【。

[14]https://textblob.readthedocs.io/en/dev/ [15] https://spacy.io/

“k k k”非常常见；这可能对应于三K党[16]，，这是一个美国白人至上主义恐怖组织和仇恨团体，其主要目标群体是非裔美国人、犹太人等。

**音频分析** 我们还分析了每个视频对应的音频信号。具体而言，我们计算过零率（ZCR）、频谱带宽和均方根（RMS）能量，并报告所有音频的平均值。

图 7 展示了两个类别中所有音频文件的过零率（ZCR）平均值随时间变化的曲线。我们观察到，仇恨类别和非仇恨类别的曲线明显不同。众所周知，过零率可以用来衡量信号的噪声程度，较高的值
表明音频信号的噪声更大。这说明，非仇恨视频的噪声水平大致均匀地分布在整个时间序列中，而仇恨视频的噪声水平则大多保持平稳，只在时间序列末尾似乎有所上升，这表明仇恨视频可能经过专门制作，以获得质量更好的音频信号。频谱带宽也呈现出相同的结果（为简洁起见，未提供数据）。

在图 8 中，我们展示了仇恨视频和非仇恨视频的均方根能量图，该图分别对各类别中的所有音频文件取平均值后得到。均方根能量有助于估算音轨的平均响度。我们观察到，与非仇恨视频不同，仇恨视频仅在时间序列的初始阶段表现出更高的响度。通过人工检查发现，许多仇恨视频中存在喊叫的情况，这可能表现为较高的响度。

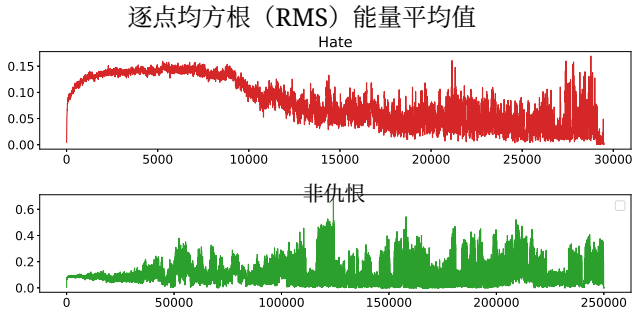



图 8：仇恨视频和非仇恨视频的均方根（RMS）能量。

**视频分析** 我们进一步尝试分析视频中主要出现哪些类型的物体。为此，我们使用 ImageAI [17] 目标检测软件包。对于每个仇恨视频和非仇恨视频，我们随机选取 20 帧，并提取与这些帧相关的所有物体。我们假设，如果某个物体在给定视频的任意一帧中出现过，就意味着该物体存在于该视频中。我们发现，仇恨视频中有 45% 的情况下会出现“人”这一物体，而“人”这一物体出现的情况占 59%。

[16]https://en.wikipedia.org/wiki/Ku Klux Klan [17]https://imageai.readthedocs.io/

出现在非仇恨视频中。在我们观察到的仇恨视频中，背景中经常出现重要的宗教人物，例如犹太拉比，同时他们身上还嵌入了大量仇恨文本。同样，被检测到的黑人往往与肮脏和贪食联系在一起。此外，在仇恨视频中，我们还观察到了“停车标志”以及一些玩具物品，例如“泰迪熊”“风筝”和“运动球”。人工检查表明，这些玩具物品大多是用于嘲讽目标社群的“卡通风格”形象。

总的来说，在本节中我们观察到，文本、音频和视频这三种模态都包含某些潜在指标，有助于区分仇恨视频类别与非仇恨视频类别。正如我们将看到的，这一观察结果也得到了第节中联合模型优异表现的佐证。

## 方法论

本节讨论了我们为仇恨视频检测所实施的预处理步骤和模型。

### 问题表述

我们将文中的仇恨视频检测问题表述如下。给定一个视频 $\boldsymbol{V}$，该任务可表示为一个二元分类问题。每个视频都要被分类为仇恨类（$y = 1$）或非仇恨类（$y = 0$）。一个视频 $\boldsymbol{V}$可表示为帧序列，即 $\boldsymbol{F} = \{f1, f2, .., fn\}$、关联音频 $\boldsymbol{A}$ 以及提取的视频转录文本 $\boldsymbol{T} = \{w_1, w_2, ..., w_m\}$，其中包含一系列词语。我们的目标是学习一个仇恨视频分类器 $Z$：$Z(\boldsymbol{F}; \boldsymbol{A}; \boldsymbol{T}) \rightarrow y$，其中 $y \in \{0, 1\}$ 是视频的真实标签。

### 预处理

我们会从转录文本中移除数字和特殊字符，并在需要时进行文本规范化。对于基于视觉的模型，我们首先以每秒一帧的速度对视频进行采样，并为每个视频采样 100 帧。对于帧数少于 100 的视频，我们会添加白色背景图像作为填充。对于帧数超过 100 的视频，我们会从所有可用帧中均匀采样 100 帧。

### 基于文本的模型

**fastText：** 我们获取所有视频转录文本的 300 维 fastText（Grave 等人，2018年）嵌入，将其输入两个各含 128 个节点的全连接层，最后传入输出节点以进行最终预测。我们将此模型命名为 **T1**。

**LASER：** 我们获取所有视频转录文本的 1024 维 LASER（Artetxe 和 Schwenk，2019年）嵌入，将其输入两个各含 128 个节点的全连接层，最后传入输出节点以进行最终预测。我们将此模型命名为**T2**。

**BERT：** 我们使用 BERT（Devlin 等人，2018年），因为它在许多文本分类任务中都表现出很高的效能，包括基于文本的仇恨言论检测。对于每份转录文本，我们提取 CLS 嵌入，并将其输入两个各含 128 个节点的全连接层，最后将其传递给输出节点，以得到最终预测。我们将此模型称为**T3**。

**HateXPlain：** 我们还试验了另一种 BERT 模型（Mathew 等人，2020b）。该模型已使用英语仇恨言论数据在预训练 BERT 上完成微调。由于该模型已经在仇恨言论数据上进行了微调，我们预计它能取得更好的表现。我们将此模型记作**T4**。

### 基于音频的模型

**MFCC：** 表示音频的一种常用方法是梅尔频率倒谱系数（MFCC）（Xu 等人，2004年），事实证明，它在肺音分类（Jung 等人，2021年）和说话人识别（Kalia 等人，2020年）等复杂任务中效果显著。我们使用 MFCC 特征来获取数据集音频的表示。为生成 MFCC 特征，我们使用开源软件包 Librosa[18]，并构建一个 40 维向量来表示音频。这些向量经过三个全连接层，生成最终标签。我们将该模型称为**A1**。

**AudioVGG19：** 我们还使用音频的波形（从视频中提取），并通过一个预训练的 VGG-19 模型（Simonyan 和 Zisserman，2015年；Grinstein 等人，2018年）生成 1000 维特征向量。与 **A1** 类似，这些向量经过三个全连接层的处理，最终生成标签。我们将该模型称为 **A2**。

### 基于视觉的模型

为了视频中的空间和时间信息，我们考虑了几种基于视觉的分类模型，例如三维卷积神经网络、InceptionV3、视觉 Transformer 等。

**三维卷积神经网络：** 三维卷积神经网络（Ji 等人，2013）模型包含两个三维卷积层和三维批归一化层。在这些层之后，我们添加修正线性单元、随机失活和最大池化层，以生成最终表示。随后，该表示经过三个全连接层，生成最终标签（有关各层大小的更多详细信息，请参阅实验设置）。我们将此模型命名为 **V1**。

**InceptionV3：** 我们还使用预训练的 InceptionV3 模型（Szegedy 等人，2015）构建特征向量。我们为全部 100 个帧提取一个 1000 维的特征向量，然后将其输入长短期记忆网络（Hochreiter 和 Schmidhuber，1997），最终馈送至输出节点进行分类。我们使用长短期记忆网络来捕捉视频帧的序列特性。我们将此模型命名为**V2**。

**视觉 Transformer：** 在这种方法中，图像被划分为一系列图像块，然后输入变换器模型。与 BERT 类似，在用于分类任务的图像块序列前也会添加额外的可学习 [类别] 标记。由于我们的目标是检测仇恨视频，因此不能直接使用视觉 Transformer。与**InceptionV3** 模型类似，我们为每个视频提取 100 帧，并将其输入预训练的视觉 Transformer（ViT）（Dosovitskiy 等人，2020）模型，为每帧获取一个 768 维特征向量，最后将其输入长短期记忆网络，以获得预测结果。我们将该模型称为**V3**。

[18] https://librosa.org/doc/latest/index.html

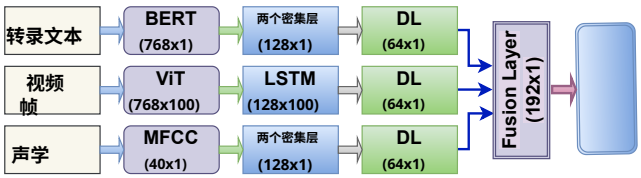



图9：多模态模型示意图。DL：密集层。

### 多模态仇恨视频检测

前面各小节讨论的模型无法利用通过不同模态（即视频、文本转录和音频）提取的特征之间的关系。为充分发挥所有模态的优势，我们尝试将文本、音频和基于视觉的模型进行有意义的组合。具体而言，我们构建了以下模型——**M1**（BERT $\odot$ ViT $\odot$ 梅尔频率倒谱系数）、**M2**（BERT $\odot$ViT $\odot$ AudioVGG19）、**M3**（HateXPlain $\odot$ ViT $\odot$ 梅尔频率倒谱系数）和**M4**（HateXPlain $\odot$ ViT $\odot$ AudioVGG19）。$\odot$ 表示通过可训练的神经网络（又称融合层）对这三种模态进行组合的操作。图9展示了整体建模流程。

## 实验与结果

### 实验设置

我们使用 $k$ 折分层交叉验证来评估模型，这有助于评估标注数据较少的模型。所有实验中，我们在此将 $k$ 设为 5；对于每一折，我们使用 70% 的数据进行训练，10% 用于验证，其余 20% 用于测试。所有模型使用相同的测试集，以确保比较公平。对于所有单模态神经网络模型，内部层包含两个各有 128 个节点的全连接层，最终缩减为长度为 2 的特征向量。对于基于单模态长短期记忆网络的模型，所有帧嵌入都会输入一个隐藏层大小为 128 的长短期记忆网络，最终缩减为长度为 2 的特征向量。对于三维卷积神经网络，每帧都会调整大小为 $100 \times 125$。两个三维卷积层均有 256 个节点，第一层的通道数为 32，第二层为 42。第一层的卷积核大小为 (5，5，5)，第二层为 (3，3，3)。此外，两层均采用步幅为 (2，2，2) 的设置和零填充，从而得到长度为 2 的特征向量。我们将最终的特征向量输入对数软最大层，并使用负对数似然损失。由此可得视频是否具有仇恨性的概率。对于融合模型，文本特征和基于音频的特征分别输入两个大小为 128 的全连接层，随后输入另一个含 64 个节点的全连接层；提取出的基于视觉的特征则输入一个隐藏层大小为 128 的长短期记忆网络，之后再输入另一个含 64 个节点的全连接层。最后，我们将所有节点拼接起来，并缩减为长度为 2 的特征向量，如图9所示。所有模型均使用 Adam 运行 20 个训练轮次

优化器、批量大小 = 10、学习率 $= 1e-4$。我们会保存宏平均F1分数达到最佳验证分数时的结果。所有模型均使用 PyTorch 库编写。

### 评估指标

为与现有文献保持一致，我们采用标准指标评估模型：准确率、F1 分数、精确率和召回率。这些指标结合起来，应能全面评估模型区分两类（仇恨与非仇恨）的分类性能。最佳结果以**粗体**标出，次佳结果带下划线。

---

### 结果

**不同模型的性能**表 3[左侧] 展示了各模型的性能。我们观察到，在所有基于文本的模型中，基于 Transformer 的模型表现最佳，尤其是**HateXPlain** 模型，该模型此前已在仇恨言论数据集上进行过微调。在基于音频的模型中，我们发现**AudioVGG19** 的表现优于**梅尔频率倒谱系数**，不过就宏平均F1分数而言，这两个模型之间的差异很小。对于基于视觉的模型，我们发现，从**ViT** 中提取的特征在所有其他基于视觉的模型中，对于检测仇恨视频最为有效。此外，我们发现所有多模态模型的表现都优于所有单模态模型（就准确率和 F1 分数而言），并且**BERT $\odot$ ViT $\odot$ 梅尔频率倒谱系数**在所有模型中表现最佳。

**基于视频长度的性能**我们根据视频长度，将所有折叠中的测试数据集划分为不同组。通过经验，我们将视频长度为 $\leq$ 105 秒的视频划入较低区间，其余视频划入较高区间，从而使两个区间中的视频数量大致相同。在表 3[右侧] 中，我们报告了两个区间的宏平均 F1 分数。我们观察到，在基于文本的模型中，除**T2**之外，其他模型在较高区间中的表现都更好。所有基于音频的模型在较高区间中的表现都更好。另一方面，除**V2**之外，所有基于视觉的模型在较低区间中的表现都优于各自对应的较高区间。这表明，当视频时长较短时，基于视觉的模型能够更好地捕捉上下文。当基于文本、音频和视觉的模型整合在一起时，正如预期，两个区间中的性能都得到了提升。

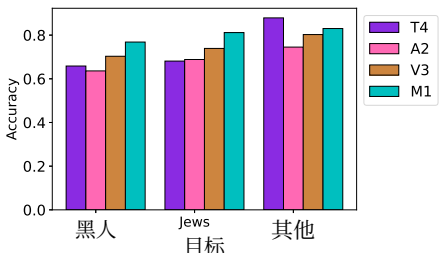



图10：按目标划分的性能。仅展示每种模态中表现最佳的模型以及最佳集成模型。

**基于目标的性能**我们还计算了最佳单模态模型和融合模型在各个目标类别上的性能。

| 模型 | 架构 | Acc | M-F1 | F1（H） | P（H） | R（H） | | VL <= 105秒（542 个视频） | VL >105秒（541 个视频） |
|---|---|---|---|---|---|---|---|---|---|
| **T1** | **fastText** | 0.687 | 0.673 | 0.609 | 0.611 | 0.614 | | 0.609 | 0.700 |
| **T2** | **LASER** | 0.730 | 0.720 | 0.668 | 0.655 | 0.686 | | 0.675 | 0.655 |
| **T3** | **BERT** | 0.735 | 0.722 | 0.664 | 0.675 | 0.667 | | 0.672 | 0.708 |
| **T4** | **HXP** | 0.757 | 0.733 | 0.653 | **0.753** | 0.577 | | 0.698 | 0.727 |
| **A1** | **MFCC** | 0.675 | 0.665 | 0.622 | 0.593 | 0.679 | | 0.603 | 0.687 |
| **A2** | **AVGG19** | 0.690 | 0.669 | 0.589 | 0.629 | 0.559 | | 0.583 | 0.669 |
| **V1** | 三维卷积神经网络 | 0.674 | 0.653 | 0.571 | 0.619 | 0.547 | | 0.637 | 0.587 |
| **V2** | **InceptionV3** | 0.720 | 0.706 | 0.643 | 0.653 | 0.637 | | 0.672 | 0.707 |
| **V3** | **ViT** | 0.748 | 0.733 | 0.672 | 0.695 | 0.656 | | 0.718 | 0.703 |
| **M1** | **BERT ⊙ ViT ⊙ 梅尔频率倒谱系数** | **0.798** | **0.790** | **0.749** | 0.742 | **0.758** | | **0.772** | **0.759** |
| **M2** | **BERT ⊙ ViT ⊙ AVGG19** | 0.755 | 0.765 | 0.718 | 0.723 | 0.719 | | 0.743 | 0.733 |
| **M3** | **HXP ⊙ ViT ⊙ 梅尔频率倒谱系数** | 0.777 | 0.767 | 0.720 | 0.718 | 0.726 | | 0.744 | 0.741 |
| **M4** | **HXP ⊙ ViT ⊙ AVGG19** | 0.767 | 0.756 | 0.707 | 0.714 | 0.712 | | 0.733 | 0.731 |

表3：[左侧] 仇恨视频分类任务的模型表现。[右侧] 按视频长度（VL，秒）划分的宏平均F1分数。H：仇恨类别，Acc：准确率，M-F1：宏平均F1，P：精确率，R：召回率，HXP：HateXPlain。

| 视频名称 | 描述 | Mode | 解释 |
|---|---|---|---|
| 犹太恐怖分子 仇恨霍尔-lywood 叛徒 犹太佬 [ID=g11ysqwzlKj6] | 在这段视频中，一个人坐着并辱骂犹太人，说着贬-带有“犹太佬”等侮辱性词语。 | Text | 该视频与仇恨言论无关，且其文-本内容干净；基于音频的模型也未能识别，因为声音中缺乏较强的攻击性。 |
| 碾磨 Nig Full Of 黑鬼 冷冻柜 头 [ID=lHk2E8MNU5HB] | 这里有一群人在大喊某种含有侮辱性称呼的歌曲像“黑鬼”这样的词，旁边有纳粹旗帜在背景中可见的 | 音频 + 视频 | 转录文本有误。其他模态发现了这一点。基于纳粹旗帜（视频）和叫喊声的有用信号（音频）。 |
| 当你身处黑鬼镇时 [ID=OngXc0A4DXxo] | 有人喊出一首歌曲，作为关于“什么是黑鬼镇？”的音频的一部分。视频中展示了无关图像。 | 音频 + Text | 音频模型能够捕捉到这一点，是因为存在叫喊声。转录文本中含有贬损性词语作为信号。视频则基本无关。 |
| A 肮脏 Jew 异性恋 From Hell 短片 Film [ID=uSH9Z7tEj9vp] | 在这段视频中，一名身穿纳粹服饰的人正在辱骂犹太人——ple. | 视频 + Text | 该视频包含一些针对.....的仇恨象征在.....中识别出了犹太人和贬损性关键词。转录文本，因此文本和视觉模型均成功了。 |

表4：一些仇恨视频及其描述示例。我们还在模态 列中注明了能够正确预测仇恨的模态。此外，我们还为这一预测提供了一种可能的解释。

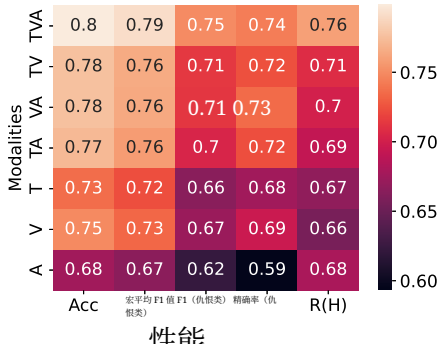



图11：不同模态性能的热力图。T：文本，V：视觉，A：音频。

并在图10中展示结果。我们观察到，在单模态模型中，基于视觉的模型 **V3** 在针对“黑人”和“犹太人”的视频上表现更好，而基于文本的模型 **T4** 在其他目标（总体而言）上表现最佳。此外，我们注意到，整合这些模型后，**M1**在所有方面都能持续取得良好表现。

目标群体。

## 各模态的有效性

为了理解不同模态对预测任务的贡献，我们进行了消融研究，以展示所有模态的有效性。我们选取了表现最佳的多模态模型 **M1**，该模型利用视频的所有特征来预测视频的标签。我们逐一移除模态，并训练模型。我们在图11中用热力图展示结果。我们观察到，联合使用所有模态进行训练能够取得最佳性能。移除至少一个模态后，性能会下降约 2–3%。此外，移除两个模态后，性能会大幅下降。总体而言，我们发现，当视频中出现喊叫或声音带有攻击性时，基于音频的特征效果显著，因为 MFCC 特征能够捕捉音频中的这些声音效果。例如，其中一段视频显示一名三K党[16] 成员喊出贬损性词语，音频模态将其识别为仇恨内容，而另外两种模态则没有识别出来。基于文本的模型的表现取决于自动

语音识别（ASR）的准确性。当 ASR 能够正确识别转录文本中的仇恨词语时，该模型在大多数情况下都能成功识别。最后，当视频本身包含受害者时，基于视觉的模型效果良好。有些视频中包含一些无关画面，例如音频内容带有贬损性，但画面却是游戏内容。在这种情况下，由于缺乏有用信号，基于视觉的模型会失效。这也说明了采用融合模型的必要性。我们在表4中展示了这类视频的几个示例。

## 结论与未来工作

本文通过利用视频三种模态的信号，迈出了识别视频中仇恨内容的一步。为此，我们从 BitChute 平台抓取视频，并手动将其标注为仇恨或非仇恨。对标注数据集 HATEMM 的分析揭示了仇恨视频的一些有趣特征。我们利用视频的所有模态来检测其是否含有仇恨内容。结果表明，与单模态模型相比，综合考虑多种模态的模型表现更好。我们还进行了初步分析，以了解不同模态对预测的贡献。我们发现，转录文本清晰时，基于文本的模型表现良好；视频中存在喊叫或攻击性行为时，基于音频的模型效果显著；视频中出现仇恨行为或辱骂目标时，基于视觉的模型能够捕捉到仇恨内容。

未来，我们计划采用其他视觉和语音 Transformer，例如 ViViT（Arnab 等，2021年）、Wav2Vec（Baevski 等，2020年）等，这或许能够进一步提升分类性能。不过，这方面的一大难点在于，为了训练和微调这类数据需求量大的模型，我们需要标注规模大得多的视频。此外，我们还计划标注时长更长的视频。我们也希望构建这样的模型：它不仅能够检测视频是否含有仇恨内容，还能识别视频中导致其被判定为仇恨内容的片段。这样，版主就不必查看完整视频，而只需观看模型标记为含有仇恨内容的片段，随后再决定采取何种审核措施。

## 伦理声明

### 伦理考量

我们的数据库包含带有标签标注的视频，不包含任何用户或上传这些视频的 BitChute 频道的个人身份信息。我们仅分析了公开可用的数据。我们遵循标准伦理准则（Rivers 和 Lewis，2014），未尝试跨网站追踪用户或破解其匿名身份。由于我们分析中使用的视频包含仇恨元素，因此应注意不要将其用于传播更多仇恨或诋毁个人或群体等负面目的。

### 偏见

数据集中注意到的任何偏见都是无意的，我们无意伤害任何个人或目标社群。我们认为，判断某项内容是否属于偏见可能带有主观性。

视频是否包含仇恨内容；因此，我们的金标准标注数据或标签分布中不可避免地存在偏差。尽管如此，我们有信心认为，鉴于我们取得了显著的标注者间一致性，数据所获标签是最准确的。

### 预期用途

我们共享数据，以促进更多关于仇恨视频分类的研究。我们仅出于研究目的发布该数据集，不授予将其用于商业或恶意用途的许可。

## 参考文献


Alcˆantara, C.; Moreira, V.; and Feijo, D. 2020年. 冒犯性视频检测：数据集与基线结果。载于*LREC*，4309–4319。

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Luˇci´c, M.; and Schmid, C. 2021. ViViT：一种视频视觉变换器。*arXiv* 预印本 *arXiv:2103.15691*。Artetxe, M.; and Schwenk, H. 2019. 用于零样本跨语言迁移及其他任务的大规模多语言句子嵌入。计算语言学协会汇刊，7：597–610。Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. 深度学习用于推文中的仇恨言论检测。载于*WWW*，759–760。Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0：一种用于语音表示自监督学习的框架。神经信息处理系统进展，33。Cheng, L.; Shu, K.; Wu, S.; Silva, Y. N.; Hall, D. L.; and Liu, H. 2020. 基于时间信息高斯混合模型的无监督网络欺凌检测。载于*CIKM*，185–194。Das, A.; Wahi, J. S.; and Li, S. 2020. 检测多模态模因中的仇恨言论。*arXiv* 预印本 *arXiv:2012.14891*。Das, M.; Banerjee, S.; and Mukherjee, A. 2022. 通过数据自举方法改进印度语言低资源辱骂性语言检测。载于第*33*届 *ACM* 超文本与社交媒体会议论文集，32–42。Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*。Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. 一张图像胜过 16×16 个词：用于大规模图像识别的变换器。*arXiv* 预印本 *arXiv:2010.11929*。Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath：理解大规模文本中的主题信号。载于*2016*年 *CHI* 人机交互大会论文集，4647–4657。Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. 一种用于辱骂检测的统一深度学习架构。载于*WebSci*，105–114。

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2020. 探索多模态出版物中的仇恨言论检测。载于*WACV*，1470–1478。Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. 学习157种语言的词向量。载于国际语言资源与评测会议（*LREC 2018*）论文集。Grinstein, E.; Duong, N. Q. K.; Ozerov, A.; and P´erez, P. 2018. 音频风格迁移。*2018*年 *IEEE*国际声学、语音与信号处理会议（*ICASSP*），586–590。Hochreiter, S.; and Schmidhuber, J. 1997. 长短期记忆。*NeuralComput.*，9(8)：1735–1780。Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 用于人体动作识别的3D卷积神经网络。*IEEE*模式分析与机器智能汇刊，35(1)：221–231。Jung, S.-Y.; Liao, C.-H.; Wu, Y.-S.; Yuan, S.-M.; and Sun, C.-T. 2021. 通过融合短时傅里叶变换与梅尔频率倒谱系数特征的深度可分离卷积神经网络模型，高效分类肺音。*Diagnostics*，11(4)：732。Juuti, M.; Gr¨ondahl, T.; Flanagan, A.; and Asokan, N. 2020. 少即是多：在数据稀缺情况下改进有毒语言分类。载于*EMNLP*：研究成果，2991–3009。Kalia, A.; Sharma, S.; Pandey, S. K.; Jadoun, V. K.; and Das, M. 2020. 基于语音活动检测技术、MFCC和感知线性预测特征的说话人识别系统比较分析。载于智能能源系统的智能计算技术，781–787。施普林格。Kennedy, B.; Jin, X.; Davani, A. M.; Dehghani, M.; and Ren, X. 2020. 通过事后解释对仇恨言论分类器进行语境化。载于*ACL*，5435–5442。Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. MVAE：用于假新闻检测的多模态变分自编码器。载于*WWW*，2915–2921。Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. 仇恨模因挑战赛：检测多模态模因中的仇恨言论。*NIPS*，33。Kipp, M. 2014. *ANVIL*：视频标注研究工具，420–436。国际标准书号9780199571932。Koetsier, J. 2021. 报告：Facebook每天都会犯下30万起内容审核错误。https: //
w
w
w.
f
o
r
b
e
s.
c
o
m/
s
i
t
e
s/
j
o
h
n
k
o
e
t
s
i
e
r/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/。访问日期：2023-04-01。Labarbera, G. 2020. 去平台化以及2020年另类科技视频托管平台BitChute的兴起：Masters of Media。
h
ttp://
mastersofmedia.hum.uva.nl/blog/2020/09/27/deplatforming-and-bitchute/。访问日期：2023-04-01。Mathew, B.; Illendula, A.; Saha, P.; Sarkar, S.; Goyal, P.; and Mukherjee, A. 2020a. 仇恨滋生仇恨：仇恨言论的时间序列研究。人机交互，4（CSCW2）：1–24。Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2020b. HateXplain：一个基准可解释仇恨言论检测数据集。*arXiv*预印本 *arXiv:2012.10289*。Newton, C. 2019年。《恐怖队列》。https: // www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video。访问日期：2023-04-01。O’Connor, J. 2021年。通过违规观看率提升透明度和问责制。https://blog.youtube/inside-youtube/building-greater-transparency-and-accountability/。访问日期：2023-04-01。Parikh, P.; Abburi, H.; Chhaya, N.; Gupta, M.; 和 Varma, V. 2021年。通过神经网络方法对性别歧视和厌女症进行分类。*TWEB*。Rana, A.; 和 Jha, S. 2022年。使用多模态学习进行基于情感的仇恨言论检测。*arXiv*预印本 *arXiv:2202.06218*。Rivers, C.; 和 Lewis, B. 2014年。大数据时代的伦理研究标准。*F1000Research*，3。Simonyan, K.; 和 Zisserman, A. 2015年。用于大规模图像识别的超深度卷积网络。载于国际学习表征会议。Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; 和 Rohrbach, M. 2019年。构建能够阅读的视觉问答模型。载于*CVPR*。Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; 和 Wojna, Z. 2015年。重新思考计算机视觉的Inception架构。*CoRR*，abs/1512.00567。Troianovski, A.; 和 Schechner, S. 2017年。德国告诫社交网络：更快删除仇恨言论，否则将面临罚款。
h
ttps://
www.wsj.com/articles/germany-to-social-networks-delete-hate-speech-faster-or-face-fines-1498757679。访问日期：2023-04-01。Trujillo, M.; Gruppi, M.; Buntain, C.; 和 Horne, B. D. 2020年。BitChute是什么？对其特征进行分析。载于第*31*届*ACM*超文本与社交媒体会议论文集，139–140。Wilson, A. 2022年。你绝对不能忽视的2023年视频营销统计数据。https://beverlyboy.com/video-marketing/2023-video-marketing-statistics-you-simply-cant-overlook/。访问日期：2023-04-01。Wu, C. S.; 和 Bhandary, U. 2020年。使用机器学习检测视频中的仇恨言论。载于*2020*年国际计算科学与计算智能会议（*CSCI*），585–590。IEEE。Xu, M.; Duan, L.-Y.; Cai, J.; Chia, L.-T.; Xu, C.; 和 Tian, Q. 2004年。基于隐马尔可夫模型的音频关键词生成。载于亚太多媒体会议，566–574。施普普林格。Yang, F.; Peng, X.; Ghosh, G.; Shilon, R.; Ma, H.; Moore, E.; 和 Predovic, G. 2019年。探索用于仇恨言论分类的文本与照片深度多模态融合。载于在线辱骂性语言第三届研讨会论文集，11–18。Ybarra, M. L.; Mitchell, K. J.; Wolak, J.; 和 Finkelhor, D. 2006年。考察与互联网骚扰相关的特征及其造成的痛苦：第二次青少年互联网安全调查的结果。《儿科学》，118(4)：e1169–e1177。