使用 BERT 检测 Parler 上的仇恨言论目标

Nadav Schneider
以色列本-古里安大学电气与计算机工程系
nadavsch@post.bgu.ac.il

Shimon Shouei
以色列本-古里安大学软件与信息工程系
shouei@post.bgu.ac.il

Saleem Ghantous
软件与信息工程系，本-古里安大学，以色列
saleem@post.bgu.ac.il

Elad Feldman
软件与信息工程系，本-古里安大学，以色列
eladfeld@post.bgu.ac.il

arXiv:2304.01179v1 [cs.CL] 3 Apr 2023

摘要

在线社交网络已成为我们日常生活的重要组成部分。不幸的是，这些平台也成了仇恨言论滋生的场所。主流社交网络已制定针对仇恨言论的常规规则。因此，Parler 和 Gab 等倡导并宣称自己是言论自由平台的社交网络应运而生。这些平台已成为针对各种目标发表仇恨言论的场所。本文提出了一套用于检测仇恨言论及其目标的处理流程，并将其用于创建 Parler 上仇恨目标的分布情况。该处理流程由两个模型组成：一个用于仇恨言论检测，另一个用于目标分类；两者均基于 BERT，并结合回译和数据预处理以提升效果。

本研究使用的源代码以及其他相关资源可在以下网址获取：
https://github.com/NadavSc/HateRecognition.git。

1 引言

在线社交网络已成为我们日常生活的重要组成部分。遗憾的是，这些平台也成为仇恨言论滋生的场所。主流社交网络已制定针对仇恨言论的规则。因此，像 Parler 和 Gab 这样倡导并自称为言论自由平台的社交网络应运而生。这些平台已成为针对不同目标群体传播仇恨言论的场所（Ali-apoulioset al.，2021年）、（Israeli and Tsur，2022a）。本文提出了一套用于检测仇恨言论及其目标的处理流程，并用其生成 Parler 仇恨言论目标分布。仇恨言论被定义为针对少数群体的仇视言论。本研究聚焦于四个主要少数群体：有色人种、穆斯林、犹太人和 LGBT群体。处理流程由两个模型组成：一个用于仇恨言论检测，另一个用于目标分类；两者均基于 BERT，并结合回译和数据预处理，以提升结果。

2 相关工作

近年来，社交媒体上的仇恨言论检测已成为一个不断发展的研究领域。许多研究致力于识别仇恨言论并确定其针对的目标，也有多种机器学习方法被提出，以应对这一挑战。本节简要概述该领域一些最相关的研究。

（Davidson 等人，2017）提出了一种机器学习方法，利用众包构建的仇恨言论词典和多类别分类器，研究推文中仇恨言论与冒犯性语言之间的区别。结果表明，准确区分仇恨言论与冒犯性语言具有挑战性；未来的研究应考虑仇恨言论的不同用途以及使用者群体。

（Ibrohim 和 Budi，2019年）提出了针对印度尼西亚推文中的辱骂性语言和仇恨言论检测的多标签文本分类方法。他们采用了三种机器学习分类器，以及结合词频、正字法和词典特征的多种数据转换方法。结果显示，该方法在检测仇恨言论方面表现良好，但在识别目标、类别和仇恨言论等级方面表现较差。

(Lemmens 等人，2021年研究了将仇恨隐喻作为特征，用于识别荷兰语中的仇恨言论类型和目标的效果

Facebook 评论。使用支持向量机和 BERT/RoBERTa 模型进行的实验结果表明，将仇恨隐喻作为特征，能够提升模型在类型和目标预测任务中的表现。这项研究意义重大，因为它揭示了仇恨言论隐性表达所带来的挑战，而这类表达往往难以检测。研究结果还强调，在将仇恨隐喻作为仇恨言论检测特征时，考虑各种信息编码方法的重要性。

(Zampieri 等人，2019年) 提出了一种名为冒犯性语言识别数据集（OLID）的新数据集，并比较了不同机器学习模型在该数据集上的表现。该数据集包含经过冒犯性内容标注的推文。OLID 中的标注包括冒犯性消息的类型和目标，因此不同于以往仅关注特定冒犯性内容类型（如仇恨言论或网络霸凌）的数据集。文章讨论了 OLID 与以往数据集之间的异同，并使用支持向量机和神经网络进行了基线实验。结果表明，在社交媒体中识别和分类冒犯性语言颇具挑战，但切实可行。

3 数据集

本项目评估了多个数据集，以确定其是否适用于训练和验证。这些数据集经过标注，并根据其与所研究问题的相关性进行筛选。具体而言，标注Parler数据集（Israeli 和 Tsur，2022b）用于仇恨言论检测任务，而 HateXplain（Mathew 等人，2021）和 Dialoconan（Bonaldi 等人，2022）用于目标检测任务。这两个数据集用于训练，而 Toxigen（Hartvigsen 等人，2022）数据集以及我们新近标注的目标标注Parler（TAP）数据集（第 3.2 节）用于评估。下一节将详细介绍这些数据集。

3.1 标注Parler

经过标注的 Parler 数据集包含来自 Parler 的 10,121 条帖子，每条帖子都有四个标注参数；标签均值采用 1–5 分制，另有用户 ID 和一个布尔型可争议参数。假设以 3 为仇恨帖子与非仇恨帖子之间的阈值，则该数据由 31.85% 的仇恨帖子和 68.15% 的非仇恨帖子组成。大多数帖子都与政治有关，其中许多帖子的标签均值较高，尽管仇恨言论被定义为针对少数群体的仇恨。后文将讨论这一区分及其影响。

3.2 目标标注Parler（TAP）

我们创建了一个新的数据集，用于评估所提出的处理流程中目标分类模型的性能。为创建目标标注Parler数据集，我们对276条 Parler 帖子进行了标注，将其分为六类：犹太人、伊斯兰教、同性恋者、非洲人、政治人物和其他。分类标准遵循文献中提到的定义。如前所述，仇恨言论只针对少数群体，因此，政治人物和其他这两类是必要的，用于将有害文本与针对少数群体的仇恨言论区分开来。

从标注Parler数据集中随机抽取仇恨等级为五的帖子。大多数仇恨帖子被标记为政治人物和其他，这意味着没有针对特定的族裔或性取向群体。在族裔和性取向群体中，帖子最常针对伊斯兰教。其他最常被针对的群体是非洲人或非洲裔，其次是犹太人，最后是同性恋者。

如果对某条帖子的标注存在争议，则会进行公开讨论，直至达成最终共识。

3.3 HateXplain

HateXplain基准数据集包含约 2 万条来自 Twitter 和 Gab 的帖子，并通过亚马逊 Mechanical Turk（MTurk）为帖子添加了目标和仇恨标签。该数据集包含约 6000 条针对不同社区群体的仇恨言论帖子，重点涉及伊斯兰教、犹太人、黑人和 LGBT群体这四个规模最大的群体。其余群体归为“其他”。数据集中的每条帖子均由三人标注，最终标签通过多数投票确定。

3.4 Dialoconan

DIALOCONAN数据集全面收录了网络仇恨者与非政府组织（NGO）工作人员之间的对话。该数据集共包含3059段对话，每段包含4、6或8轮，总计16625轮。数据集涵盖仇恨言论的六个主要目标，包括犹太人、LGBT群体、移民、穆斯林、

有色人种和女性。这些数据通过人类专家干预和机器生成的对话获得，采用了19种策略。该数据集是研究仇恨言论及其影响的宝贵资源。

3.5 Toxigen

Toxigen 是一个大规模机器生成的数据集，包含 274,186 条关于 13 个少数群体的有害和无害陈述。该数据集采用基于示例的提示框架和对抗性分类器环路解码方法，使用一个大型预训练语言模型（GPT-3）生成隐晦有害和无害文本。以这种方式控制机器生成，使 Toxigen 能够以比以往人工撰写文本资源更大的规模，涵盖隐含有害文本，并涉及更多人口群体。Toxigen 可用于应对人工撰写文本和机器生成文本中的毒性。Toxigen 的两个变体已用于3.5在不同的数据集上评估目标分类模型：

（a）标注训练集——一小部分人工标注数据。本文将其称为“小型 Toxigen”。为保证数据质量，已筛除毒性评分低于 4 的行以及标注意见不一致的行。筛选后，数据集包含 1,608 行。

(b) Toxigen——大型数据集。本文将其称为“大型 Toxigen”。该数据集包含 250,951 行。

4 数据预处理

数据预处理是训练学习模型之前的重要步骤。鉴于语言数据混乱且充满噪声的特性，在本任务中这一步尤为关键。在对 BERT 进行微调之前，我们采用了多种方法来清洗并归一化原始文本数据，具体如下（Nguyen 等人，2020年）：

1. 排除非英语撰写的帖子。2. 将所有词元转换为小写，确保文本数据格式一致。3. 将特定词元（如“@”提及、主题标签和网址）替换为对应的键（例如，提及替换为 <USER>）。4. 使用 emoji 软件包中的 demojize 函数将表情符号词元转换为文字。5. 将“”””和“. . . ””等特殊字符替换为对应的 ASCII 字符。6. 将缩写形式拆分为各组成词语，并正确调整空格。7. 正确设置时间表达式（如“p.m.”和“a.m.”）的空格。8. 删除多余的空白字符，生成最终清理并归一化后的文本数据。

这一文本归一化和清洗过程对于准备 Parler 帖子以进行 BERT 微调至关重要，因为它有助于确保文本数据格式一致且与分析相关。

5 计算方法

在这项工作中，我们关注仇恨言论帖子中的目标检测。为完成此任务，我们提出了一个两阶段处理流程：（a）使用指定的仇恨言论检测模型判断帖子是否包含仇恨言论。（b）如果第一个问题的答案为肯定，则继续使用仇恨目标分类模型。

我们针对每个模型和不同数据集上的相应任务，对 BERT（Devlin 等人，2018）转换器进行微调。

5.1 仇恨言论检测

第一步是使用经过标注的 Parler 数据对 BERT 进行微调。如上文（第3节）所述，Parler 标注数据为每个样本提供一个标签均值。原论文（Israeli 和 Tsur，2022a）将标签均值达到 3 及以上的帖子归类为仇恨言论。不过，由于研究发现，标签均值高于 3 的帖子是否真的具有仇恨性质值得商榷，因此我们也设置并评估了一个高于 3 的阈值。该阈值将任务转化为二元分类：标签均值高于阈值的帖子被标记为仇恨言论，其余帖子则被视为正常内容。

由于数据分布不均衡，我们采用了加权损失，即对不同类别赋予不同权重的方法。这种技术会在模型成功识别较少见的类别时给予更多奖励，从而弥补类别间的不平衡。另一种有助于更好收敛的方法是回译，将在第5.3.2节介绍。

模型	准确率	召回率	精确率	F1
阈值 3	74	49	**62**	55
阈值 4	**89**	36	40	38
加权损失阈值 3	75	**76**	59	**66**
加权损失阈值 4	80	63	30	41
回译加权损失阈值 3	76	70	60	65
回译加权损失阈值 4	87	34	42	37

表 1：展示了所提出的仇恨言论检测模型在 Parler 标注数据集上的相关指标。

5.2 目标分类

第二步重点是在 HateXplain 和 Dialoconan 数据集上微调 Bert，以识别仇恨言论的目标群体。如第3.3节所述，我们选取了 Fourminorities 中的四个群体：LGBT群体、犹太人、穆斯林和黑人。此外，我们使用 BERTopic 模型对每个示例进行主题拼接，相关内容将在第5.3.1节进一步讨论

5.3 提升模型性能

两个模型都在训练数轮后出现了过拟合。为防止过拟合或延缓其发生，并提高验证集上的准确率，我们采用了多种方法，包括清洗、归一化、数据增强，甚至借助其他模型。

5.3.1 BERTopic

帖子主题可能与仇恨言论的目标有关。因此，我们在训练期间将文本主题（使用 BERTopic 识别，Groo-tendorst，2022）拼接到输入文本中。BERTopic 使用帖子中词元的 BERT 嵌入表示，然后通过 HDBSCAN（McInnes 等人，2017）聚类算法对降维后的嵌入进行聚类。BERTopic 模型根据每个簇中出现频率最高的词为主题命名。这种方法的难点之一是确定 HDBSCAN模型的最佳参数，尤其是最小簇大小和样本大小；这两个参数分别定义了最小簇大小和离群值的数量。为了解决这一问题，我们使用了开源代码库 TopicTuner（Top），它是一个便捷的 BERTopic 模型封装工具，可简化寻找最优最小簇大小和样本大小参数的过程。

5.3.2 回译

另一项改进是采用回译，即将文本翻译成五种不同的语言，再译回原始语言。这种技术可以增强模型的稳健性，使其能够处理不同的语言变体，并减少过拟合。我们使用 MarianNMT 实现了回译方法（Junczys-Dowmunt 等人，2018），这是一个使用 DNN 和 Transformer 模型训练一千多种不同语言的预训练翻译模型的库。我们将文本回译成五种不同的语言：西班牙语、德语、俄语、法语和中文。在这五种语言中，由于俄语和中文的回译失败情况较其他语言多，因此未采用这两种语言。部分译文存在噪声，包含乱码。为了解决这些问题，我们采用了文本预处理（第4节）和重复词删除技术。通过使用回译技术生成的数据增强内容，我们扩充了数据集。

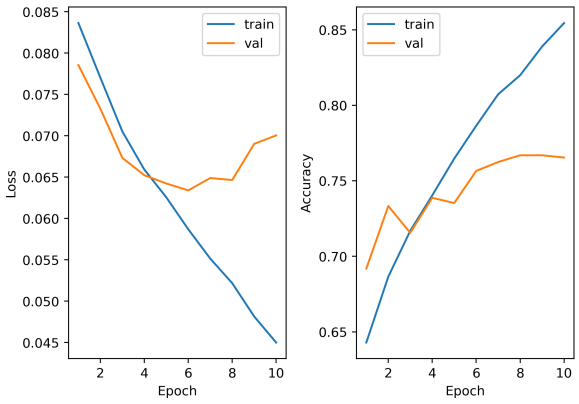


图1：展示了仇恨检测最佳模型——加权损失阈值 3 模型的损失和准确率变化趋势。

6 实验结果

6.1 仇恨言论检测

模型使用标注Parler数据 3.1 进行训练和运行，训练集与测试集按 80:20 划分，并在 32GB Tesla V100 GPU 上运行，使用 PyTorch（Paszke 等人，2019年）。训练采用批量大小为 8、10 个训练轮次、交叉熵损失函数、0.00005 的学习率和 Adam 优化器。我们还测试了阈值和加权损失等多种参数组合（表1）。

阈值 4 模型取得了最高准确率，为 89%；加权损失阈值 3 模型的召回率最高，为 76%；阈值 3 模型的精确率最高，为 62%；F1 分数最高的仍是加权损失阈值 3 模型，为 66%。在仇恨检测阶段，我们的目标是在保持合理 F1 分数的同时，将假阴性降至最低，因此选择了加权损失阈值 3 模型。即使假阳性相对较高，我们也希望尽可能少漏掉仇恨帖子。这是因为在第二阶段还会进行额外的“筛选”：目标检测模型会将不含任何预定义目标的帖子以及非仇恨帖子归入“其他”类别。所选模型的学习曲线见图。1。

6.2 目标检测

6.2.1 BERTopic

在本项目的实验阶段，我们使用 BERTopic 模型获取数据的主题划分。我们借助 Topic-Tuner 存储库优化了 HDBSCAN 算法的参数。最小样本大小设为数据集中最小目标的大小，从而得到能最大限度减少离群点出现的最优参数。在 BERTopic 模型中应用优化后的 HDBSCAN 参数后，数据被划分为四个不同的主题，另有一个离群主题（-1）。识别出的主题名称如下：

(a) ’0_黑鬼_白人_黑鬼（复数）_喜欢’, (b) ’1_穆斯林_穆斯林_沙子_数字’, (c) ’2_犹太佬_犹太人_犹太人_白人’, (d) ’3_基佬（复数）_基佬_用户_同性恋者’. (e) ’-1_用户_黑鬼_数字_白人’,

模型执行的主题划分显示，它成功检测出了目标，与原始数据中的标签一致。

6.2.2 BERT 微调结果

BERT 模型在两个训练数据集（3.3、3.4）上进行了微调，并在三个独立数据集（3.5、3.2）上进行了评估。数据按 80:20 的比例划分。我们尝试合并训练数据集，但这导致性能下降。我们主要通过实施以下步骤来提升模型性能：拼接主题名称和/或采用回译技术。

在 HateXplain 数据集上训练的、加入主题方法后的模型，其准确率和损失变化趋势如图2所示。模型从第四个训练轮次开始出现过拟合迹象，因此实施了早停。

使用具有不同分布的独立评估数据集，

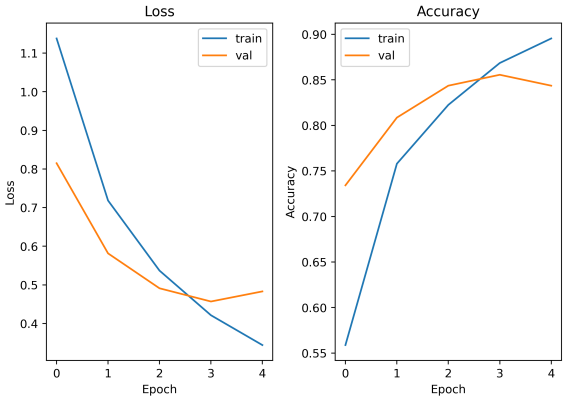


图2：在 HateXplain 数据集上训练的主题拼接模型的损失和准确率变化趋势。

与训练数据不同的数据对于准确评估机器学习模型的性能及其对新型、未见数据的泛化能力至关重要。在 HateXplain 数据集（3.3）上训练的模型的评估结果见表2。虽然主题拼接方法始终表现最佳，我们还测试了回译方法，以进一步提升模型性能。然而，结果（表 2）表明，尽管我们最初对此抱有期待，回译方法却降低了模型性能。回译过程引入了噪声或无关数据，

评估数据集	Back 翻译	主题 在输入中	准确率	召回率	精确率	F1
小型 Toxigen	✓	✓	0.5	0.57	0.47	0.47
	✓		0.65	**0.62**	0.53	0.55
		✓	**0.71**	0.57	**0.6**	**0.58**
			0.66	0.51	0.55	0.52
大型 Toxigen	✓	✓	0.53	0.57	0.47	0.47
	✓		0.63	**0.58**	0.5	0.52
		✓	**0.71**	0.57	**0.59**	**0.57**
			0.66	0.56	0.54	0.54
Parler 标注的目标	✓	✓	0.69	0.76	0.5	0.56
	✓		0.74	0.7	0.53	0.58
		✓	**0.82**	0.73	**0.61**	**0.66**
			0.78	**0.79**	0.59	0.65

表2：在 HateXplain 数据集上训练的目标检测模型的相关结果。

从而对模型性能产生了负面影响。大型 Toxigen 数据集（3.5）的结果（表2）显示，主题拼接模型的性能有了显著提升。这一提升幅度显著，且在多项指标上均表现稳定。在大型 Toxigen 数据集上，准确率和精确率指标的提升最为明显，这表明主要改进体现在减少假阳性错误方面。

7 模型可解释性

有效评估是评估每个模型的关键步骤。在评估过程中，会将模型的预测结果与所选数据集的预定标签进行比较，然而，给定的数据集并没有任何标签。对其他数据集（表2）进行评估可以在一定程度上反映模型表现，但为了提高我们提出的模型的可信度，我们采用了可解释性技术。可解释性旨在解释某项决策背后的原因，从而使我们能够确保模型只关注相关特征。

我们将 LIME（Ribeiro 等人，2016年8月13日至17日）算法应用于模型的预测结果，以研究输入与输出之间的联系。下面展示了每个模型的成功和失败案例：

（A）仇恨言论模型成功（图3a）——“生活糟糕得美妙：我们需要做的就是意识到这一点。”正如预期，模型识别出的重点积极词语是“生活”和“美妙”，而消极词语是“糟糕”。模型理解了尽管积极词和消极词混杂其中，仍能理解广泛的上下文，并将该帖子分类为非仇恨言论。

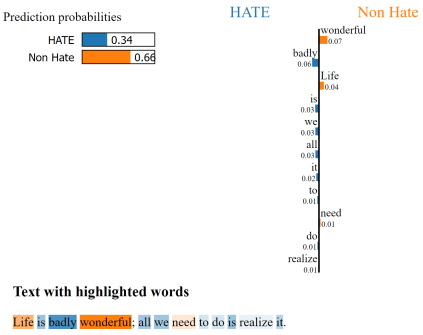
（B）仇恨言论模型失败案例（图3b）——“生活又一次让我失望了。”负面词语是“让我”和“失望”，而正面词语是“又一次”，这符合预期。该示例被错误地分类为仇恨帖子。虽然其中出现了负面词语，但没有少数群体，因此不构成仇恨言论。

（C）目标检测成功（图3c）——“犹太人垄断了邪恶。”如预期，强调犹太人身份的词是“犹太人”，模型确实将目标分类为犹太人。

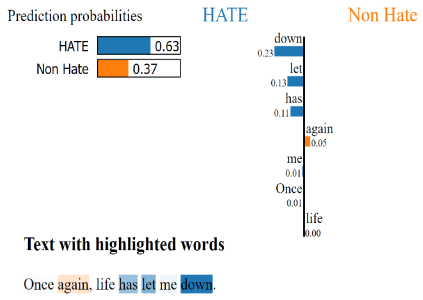
（D）目标检测失败（图3d）——“他们都是性瘾者，他们的主娶了13个妻子，这说明了他们是什么样的人。”影响最大的词是“瘾者”和“女性”。这两个群体都属于“其他”类别，模型也作出了相同判断；然而，模型没有识别出更广泛的语境，而真正的目标是穆斯林。

8 结论与未来工作

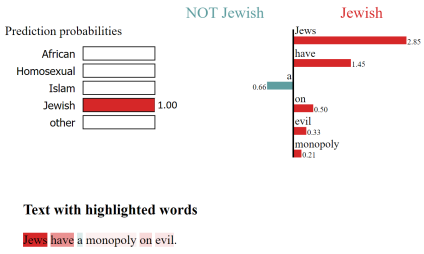
本文提出了一套从 Parler 数据集中提取仇恨言论目标的处理流程。我们训练了两个经过微调的 BERT 模型，并测试了多种提升性能的方法。目标检测模型使用了不同于 Parler 的数据集进行训练。为了提高最终 Parler 目标分布的可信度，我们提出了一个新的 Parler 标注目标数据集——TAP（第



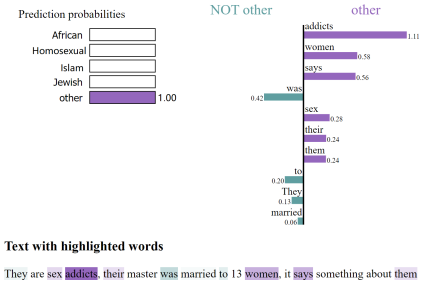
（A）仇恨言论模型成功示例。



（B）仇恨言论模型失败示例。



（C）目标检测模型成功示例已呈现。



（D）目标检测模型失败示例已呈现。

图 3：使用 LIME 检查仇恨言论分类模型和目标分类模型中的样本。

节3.2）。该数据集属于 Parler 的一部分，因此其评估结果更具代表性。此外，我们还对模型进行了可解释性分析，以进一步考察模型的性能，进而评估最终目标分布的可靠性。

在约 80 万条 Parler 帖子中，17% 是仇恨言论帖子，其中大多数针对非洲人（图4）。这一结果引发了伦理方面的问题，也可能与 Parler 用户的来源有关，甚至可能与数据抓取的时间有关。最终分布与 TAP（第3.2 节）分布相似。

未来的工作包括完成对整个 Parler 数据集的解析，继续提升模型的性能（通过增加数据或采用优化方法），并考虑在庞大的 Parler 数据集上使用无监督技术。

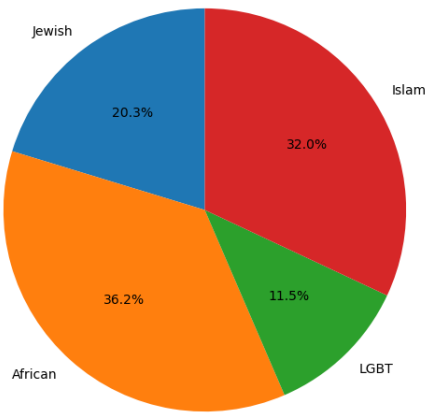


图4：824,848 条 Parler 帖子中的仇恨言论目标分布。

致谢

参考文献

TopicTuner——调优 BERTopic HDBSCAN 模型。

Max Aliapoulios、Emmi Bevensee、Jeremy Blackburn、Emiliano De Cristofaro、Gianluca Stringhini 和 Savvas Zannettou。2021年。初探 Parler 在线社交网络。*CoRR*, abs/2101.03820。

Helena Bonaldi、Sara Dellantonio、Serra Sinem Tekiroglu 和 Marco Guerini。2022。构建用于应对仇恨言论的对话数据集的人机协作方法。

Thomas Davidson、Dana Warmsley、Michael Macy 和 Ingmar Weber。2017。《自动仇恨言论检测与冒犯性语言问题》。发表于*AAAI* 网络与社交媒体国际会议论文集，第 11 卷，第 512–515 页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018。BERT：用于语言理解的深层双向变换器预训练。*CoRR*，abs/1810.04805。

Maarten Grootendorst。2022。BERTopic：采用基于类别的 tf-idf 程序进行神经主题建模。*arXiv* 预印本 *arXiv:2203.05794*。

Thomas Hartvigsen、Saadia Gabriel、Hamid Palangi、Maarten Sap、Dipankar Ray 和 Ece Kamar。2022。ToxiGen：用于对抗性隐性仇恨言论检测的大规模机器生成数据集。发表于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 3309–3326 页，爱尔兰都柏林。计算语言学协会。

Muhammad Okky Ibrohim 和 Indra Budi。2019年。印度尼西亚 Twitter 上仇恨言论和辱骂性语言的多标签检测。载于第三届线上辱骂性语言研讨会论文集，第 46–57 页。

Abraham Israeli 和 Oren Tsur。2022a。言论自由还是仇恨言论自由？分析 Parler 上仇恨言论的扩散。载于第六届线上辱骂与伤害研讨会（*WOAH*）论文集，第 109–121 页。

Abraham Israeli 和 Oren Tsur。2022b。言论自由还是仇恨言论自由？分析 Parler 上仇恨言论的扩散。载于第六届线上辱骂与伤害研讨会（*WOAH*）论文集，第 109–121 页，西雅图，华盛顿州（混合形式）。计算语言学协会。

Marcin Junczys-Dowmunt、Roman Grundkiewicz、Tomasz Dwojak、Hieu Hoang、Kenneth Heafield、Tom Neckermann、Frank Seide、Ulrich Germann、Alham Fikri Aji、Nikolay Bogoychev、André F. T. Martins 和 Alexandra Birch。2018。Marian：使用 C 语言的快速神经机器翻译++。载于*ACL 2018* 系统演示论文集，第 116–121 页，澳大利亚墨尔本。计算语言学协会。

Jens Lemmens、Ilia Markov 和 Walter Daelemans。2021年。利用仇恨隐喻特征改进仇恨言论类型和目标检测。载于第四届互联网自由自然语言处理研讨会论文集：审查、虚假信息与宣传，第7–16页。

Binny Mathew、Punyajoy Saha、Seid Muhie Yimam、Chris Biemann、Pawan Goyal 和 Animesh Mukherjee。2021年。HateXplain：用于可解释仇恨言论检测的基准数据集。载于*AAAI*人工智能会议论文集，第35卷，第14867–14875页。

Leland McInnes、John Healy 和 Steve Astels。2017。hdbscan：基于层次密度的聚类。《开放源代码软件期刊》，2(11):205。

Dat Quoc Nguyen、Thanh Vu 和 Anh Tuan Nguyen。2020年。BERTweet：一种用于英语推文的预训练语言模型。载于《*2020*年自然语言处理经验方法会议论文集：系统演示》，第9–14页，线上。计算语言学协会。

Adam Paszke 等。2019年。PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》，32。

Marco Tulio Ribeiro 等。2016年8月13日至17日。“我为什么应该相信你？”：解释任意分类器的预测。载于《第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，旧金山，美国加利福尼亚州》，第1135–1144页。

Marcos Zampieri、ShervinMalmasi、Preslav Nakov、Sara Rosenthal、Noura Farra 和 Ritesh Kumar。2019年。预测社交媒体中攻击性帖子的类型和目标。*arXiv* 预印本 *arXiv:1902.09666*。