arXiv:2304.00913v1 [cs.CL] 3 Apr 2023

LAHM：面向多领域多语言仇恨言论识别的大型标注数据集

Ankit Yadav，Shubham Chandel，SushantChatufale 和Anil Bandhakavi

Logically.ai，Brookfoot Mills，Brookfoot Industrial Estate，Brighouse，HD6 2RW，英国

摘要

当前仇恨言论分析研究通常聚焦于单语和单类别分类任务。本文介绍一个新的多语言仇恨言论分析数据集，涵盖英语、印地语、阿拉伯语、法语、德语和西班牙语，涉及仇恨言论的多个领域：辱骂、种族主义、性别歧视、宗教仇恨和极端主义。据我们所知，本文首次探讨了如何在这六种语言中识别上述五大领域内的各类仇恨言论。我们介绍了数据集的创建过程、如何针对不同领域进行高层级和低层级标注，以及如何利用该数据集测试当前最先进的多语言和多任务学习方法。我们在多种单语、跨语言和机器翻译分类设置下评估了该数据集，并将其与为此任务汇总并合并的开源英语数据集进行比较。随后，我们讨论如何利用这一方法创建大规模仇恨言论数据集，以及如何借助这些标注提升仇恨言论检测和分类的整体效果。

关键词

仇恨言论，多语言，多领域，跨语言，种族主义，宗教仇恨，性别歧视，辱骂，极端主义，少样本学习，零样本学习，

1. 引言

辱骂性语言是 Twitter 等社交媒体平台上一个重要且备受关注的问题。社交媒体经常被滥用于传播仇恨言论或其他形式的辱骂性语言等有害内容。每分钟产生的用户生成内容数量极其庞大，因此人工监控 Twitter 上的辱骂行为既不可行，也不切实际。T
witter 已通过制定明确的仇恨行为政策、允许用户举报并利用审核人员筛选内容等方式，努力从其平台上清除辱骂性内容。然而，这些人工措施难以扩展，也无法长期持续。

自然语言处理（NLP）领域已开展多项研究，以应对社交媒体中的仇恨言论检测问题。大多数研究提出了监督式方法，使用从传统机器

从学习方法到基于深度学习的方法。然而，大多数工作只关注单一语言（即英语）和单一辱骂性内容领域现象，例如仇恨

De-Factify：多模态事实核查与仇恨言论检测研讨会，与 *AAAI 2022*同期举办。*2022*年，加拿大温哥华

ankit.yadav@logically.co.uk (A. Yadav); Shubham.c@logically.co.uk (S. Chandel);
sushant.c@logically.co.uk (S. Chatufale); anil@logically.co.uk (A. Bandhakavi)
https://www.logically.ai/team/leadership/anil-bandhakavi (A. Bandhakavi)© 2021 年本文作者版权所有。允许依照知识共享署可协议（CC BY 4.0）使用。CEUR Workshop Proceedings（CEUR-WS.org）

言论、性别歧视、种族主义、宗教仇恨等，而非多种语言和多个领域。Twitter 支持 34 种语言的内容，用户可以使用其中任何一种表达观点。因此，如何实时处理多种语言中的仇恨内容成为一项挑战。我们需要能够跨多种语言和多个领域检测仇恨内容的稳健模型。

本文尝试应对仇恨言论检测领域的两项主要挑战：

1. 构建一个涵盖 6 种语言的仇恨言论检测**多语言数据集**：英语、印地语、法语、阿拉伯语、德语、西班牙语。2. 构建一个**多领域数据集**，涵盖以下仇恨言论领域：辱骂、种族主义、性别歧视、宗教仇恨和极端主义

我们对不同领域的定义如下：

1. 种族主义：基于种族、族裔、种姓、国籍、文化、肤色、发质和身体特征的歧视。2. 性别歧视：基于性别／性取向的歧视。3. 宗教仇恨：宗教歧视，即因某人对某种宗教所持的特定信仰／观念而对其区别对待。4. 辱骂：造成或可能造成痛苦、不尊重或精神伤害的言论，尤其是粗俗和亵渎性评论。5. 极端主义：造成或可能造成个人、群体或更广泛社会伤害的言论；任何政治或民事议题都可能通过暴力引发极端主义行为。

2. 相关工作

已有多项研究关注辱骂性语言检测 [1] [2], 冒犯性语言、仇恨言论识别[3], 毒性 [4], 仇恨性 [5],攻击性 [6],攻击 [7],种族主义、性别歧视[8],淫秽内容、威胁和侮辱。

此外，还有多项共享任务聚焦于辱骂性语言和仇恨言论检测，例如 HASOC-2019 [9], TRAC 攻击性识别共享任务[6], HatEval [10] 以及 GermEval-2018 [11]，后者聚焦于德语推文中的冒犯性语言识别。

Waseem[8] 提出了以下用于识别仇恨言论的标准。这些标准部分源自对 McIntosh（2003）所述特权的否定；这些特权体现为凸显重要性、确保受众关注以及保障白人安全的方式；部分则源自常识。一条推文若符合以下条件，则归类为冒犯性内容：

1. 使用性别歧视性或种族歧视性辱骂。2. 攻击少数群体。3. 试图让少数群体噤声。4. 批评少数群体（但缺乏有据可依的论证）。5. 宣扬仇恨言论或暴力犯罪，但并未直接使用相关言论。6. 批评少数群体并使用稻草人论证。

7. 公然歪曲事实，或试图用毫无根据的说法扭曲人们对少数群体的看法。8. 支持有问题的标签——“#BanIslam”、“#whoriental”、“#whitegenocide”。9. 对少数群体进行负面刻板印象描述。10. 为仇外心理或性别歧视辩护。11. 屏幕名称含有冒犯性内容；根据上述标准，该推文（充其量）含义模糊，且涉及符合上述任一标准的话题。

多种机器学习模型已被用于处理多领域仇恨言论检测任务。一些研究采用传统机器学习方法，例如逻辑回归 [12], 支持向量机 [13] [14], 线性支持向量机分类器（LSVC） [15] [16] [17]。研究者使用这些方法以提高可解释性，同时也采用了多种基于深度学习的模型，包括卷积神经网络 [18] [19], 长短期记忆网络 [20] [18] [17] [16] [21],双向长短期记忆网络 [22]。近期研究主要关注迁移学习，以及采用 Transformer 的新型架构，例如基于 Transformer 的双向编码器表示（BERT） [23] [24] [25] [26] [27] 及其变体 RoBERTa [25] ，用于跨领域辱骂性语言检测任务。

在跨语言辱骂性语言检测中，大多数研究使用基于 Transformer 的模型。一些研究也使用了传统模型，例如逻辑回归 [28] [10] [29], 线性支持向量机 [16] [17], 支持向量机 [30], 长短期记忆网络 [31] [29] 和双向长短期记忆网络 [31] 。近期研究聚焦于多种基于 Transformer 的架构，如多语言BERT [32] [28] [25][17] [33] [34] [29], RoBERTa [35], XLM [31] [34] 和 XLM-RoBERTa [35] [25] [36]。具有多语言表示的基于 Transformer 的模型能够轻松应对零样本跨语言任务中的语言迁移。

3. LAHM 数据集

本节介绍我们希望纳入数据集的特征、我们在涵盖仇恨言论所有主要领域的同时收集不同类型仇恨言论的方法，以及如何大规模标注数据。我们还将详细介绍所收集数据的统计信息并进行分析。

3.1. LAHM 数据集的要点

鉴于目前尚无涵盖六种语言中仇恨言论这五个领域的数据集，我们旨在创建一个可靠、多维度的多语言仇恨言论数据集。

多语言数据： 我们的数据集是一项多语言资源，旨在促进跨语言研究。其中包含英语、印地语、法语、阿拉伯语、德语和西班牙语的仇恨言论。

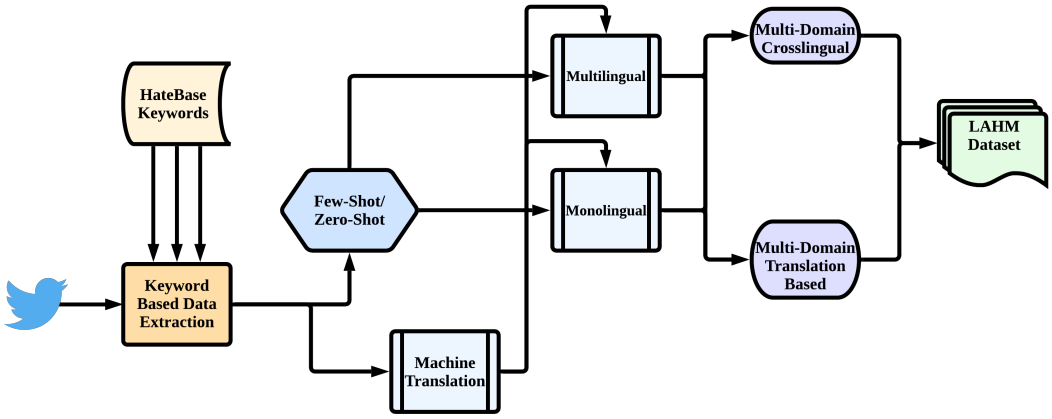


图 1： 整体标注流水线架构

多领域数据： 我们的数据集为每种语言中的每个仇恨样本提供细粒度标签。这些方面涵盖了大多数仇恨领域，如种族主义、性别歧视、辱骂、宗教仇恨和极端主义。

3.2. 数据集收集

3.2.1. 语言和领域特定关键词收集

考虑到英语、印地语、法语、阿拉伯语、西班牙语和德语主要使用地区的文化差异，我们首先查找这些语言中本土的仇恨关键词。我们使用 HateBase 词汇数据集；它是从公共论坛创建仇恨数据集的宝贵词典。此外，我们还使用 Hatebase 目击记录数据集，该数据集有助于对关键词进行趋势分析。

我们针对 6 种语言开展了研究。此外，我们还尝试了一些印度语言，主要是马拉地语和孟加拉语，但 Hatebase 对这些语言的关键词覆盖极少。此外，我们还提取了以下信息：

1. 关键词针对的群体 2. 关键词的冒犯程度——极度冒犯、高度冒犯和轻度冒犯。3. 最近的目击次数

目标群体帮助我们将关键词归类到较低层级的族裔与国籍、宗教仇恨、性别与性取向以及一般辱骂类别中。各领域和语言下的关键词详情见数据集统计中的表 1 。总计收集了 **1022** 个关键词。为提取极端主义相关数据，我们使用了一组与极端主义和恐怖主义相关的关键词（包括恐怖组织名称），从 BBC Monitoring 检索新闻文章[1]。

[1]https://monitoring.bbc.co.uk/

1. 我们进一步提取了这些与极端主义和恐怖主义相关的文章中的评论。我们还从反极端主义数据库[2]中提取了具有极端主义性质的推文和评论。2. 这些评论据称由恐怖组织和极端主义组织的成员及嫌疑人在 Facebook、Twitter 和 YouTube 上发布。

3.2.2. 大规模多语言推文数据集

我们使用按语言整理的关键词开始收集数据集。由于 Twitter 官方 API 只能用于有限数量的请求，我们使用 Twint[3]API 按每个关键词收集 1000-2000 条推文。我们通过该 API 搜索了过去一年的数据。我们还为阿拉伯语和法语添加了来自 MLMA [1] 的其他关键词。

收集到的原始数据包含跨语言推文，因此语言检测也成为我们流程的一部分。对于每种语言中的每个关键词，我们只保留该语言的推文，并删除其余推文。这有助于我们训练单语仇恨言论检测模型，无需担心语言中的语码转换。

我们将所有用户名替换为 @USER，将 URL 替换为 @URL，并清除了推文中的所有不必要符号。我们还删除了少于 4 个词的短推文。我们总计收集了 **497660** 条推文。各语言推文的详细信息见第 3.3 节数据集统计。

对于极端主义类别，我们精心手工编制了总计 **88** 个关键词，并从 BBC Monitoring 收集了数据。此外，我们还从其他网站收集了有关极端主义的数据。

3.3. 数据集统计

表 1 展示了不同语言和不同领域中关键词的分布。我们合并了仇恨词库中的类别，从而得到了这些领域。 性别歧视 包含属于 性别 和 性取向的关键词。种族主义包括族裔 和 国籍类别。宗教仇恨包括 Hatebase 宗教类别中的词语 [4]。极端主义关键词是按照第 3.2 节所述手工编制的。

就提取的关键词总数而言，英语占主导地位（500 个）。除印地语外，种族主义 是所有语言中仇恨关键词数量最多的主要领域。

3.4. HSMerge 数据集

3.4.1. 数据集准备

为了构建适用于不同任务且包含金标准标签的数据集，我们使用了 10 个公开可用的数据集，用于检测不同类型（任务）的仇恨言论。我们从多个英语数据集中抽取了带注释的样本：GAO 仇恨性数据集 [5], TRAC 攻击性数据集 [6], 冒犯性 [37], 种族主义和性别歧视 [8], MLMA [1], 攻击 [7],CONAN [38], MMHS150k [39]。

我们将这些数据集中的标签映射到不同领域：辱骂性、非辱骂性、性别歧视、种族主义和宗教仇恨。对于 OLID， [37] 冒犯性 映射为辱骂。GAO 和 WUL 使用二元

[2]https://www.反极端主义项目.com/daily-dose-archive [3] https://github.com/twintproject/twint/ [4] https://hatebase.org/

表1关键词语言与领域分布

语言	性别歧视	种族主义	辱骂	宗教仇恨	极端主义
英语	69	418	30	33	88
印地语	75	16	9	8	-
阿拉伯语	27	35	17	17	-
法语	13	85	13	9	-
德语	21	67	13	11	-
西班牙语	43	77	21	2	-
总计	**248**	**698**	**103**	**80**	**88**

表2推文分布

语言	**Raw**	**处理后**
英语	184339	105120
印地语	59321	32734
阿拉伯语	27374	5394
法语	75126	20809
德语	30868	8631
西班牙语	120632	55148
总计	**497660**	**227836**

标签，而原始 TRAC 使用三个标签：非攻击性、隐性攻击性和公开攻击性。我们将第一个重新标记为非辱骂性，将另外两个重新标记为辱骂性。对于 MLMA，我们根据目标群体创建标签：目标群体宗教 对应宗教仇恨。对于 MLMA， [1] 目标群体性别 和性取向 映射为性别歧视；目标群体宗教 对应宗教仇恨；目标群体出身 对应种族主义。对于 HASOC， [9] 冒犯性 映射为辱骂。对于 COnan，[40] 目标群体穆斯林、犹太人 对应宗教仇恨 标签；目标群体移民、有色人种 映射为种族主义；目标群体女性，*LGBT+* 映射为性别歧视 ，而残障人士 映射为一般辱骂 类别。对于 MMHS150K， [39] 我们获得了种族主义、性别歧视、宗教仇恨和其他仇恨。

对于其中一些数据，由于只提供了推文 ID，我们必须使用官方 Twitter API 从 Twitter 获取推文。对于 Waseem 和 Founta 数据集， [38], Twitter 上无法找到大多数推文，因此推文数量少于官方数据集。我们将这些数据集统一为一个*HSMerge* 数据集。各类别和样本总数的详情见表3。

表 3HSMerge 开源金标准数据集

数据集	领域	Hate	非仇恨	总计
OLID	冒犯性 / 非仇恨	4400	8840	13240
COnan	种族主义/ 性别歧视/ 宗教/ 辱骂/ 其他仇恨	938/1025/1921/175/179	-	4238
TRAC	仇恨 / 非仇恨	5993	4348	10341
MLMA	种族主义/ 性别歧视/ 宗教/ 其他仇恨	2448 / 1152 / 68/1979	-	5647
MLA150K	种族主义/ 性别歧视/ 宗教/ 其他仇恨/ 非仇恨	11925 / 7365 / 163 / 5811	112845	138209
HASOC	仇恨 / 非仇恨	2261	3591	5852
WASEEM	种族主义者/性别歧视者/非仇恨	2753/15	7640	10408
WUL	仇恨/非仇恨	8817	62937	71754
GAO	仇恨 / 非仇恨	244	675	909

4. 实验

4.1. 总体方法与模型

本节将介绍图1 所示处理流程的各个组成部分，以及在不同设置下用于半监督标注过程的模型。大多数仇恨言论检测任务依赖人工标注，这限制了可标注样本的数量。我们采用层次化方法，对最初基于 Twitter 关键词构建的数据集进行验证和优化。

1. 将 Hate-ALERT 中的单语仇恨言论检测模型用作零样本处理流程，以获取仇恨或非仇恨的高层级标签。2. 使用多语言*mbert* 微调模型来获取高层级标签。3. 将五种语言（英语除外）的原始数据翻译成英语。4. 使用多语言二元分类器（在步骤 2 中开发）对翻译后的数据进行高层级推文分类。5. 使用 Google Perspective API 预测推文的毒性，以获取高层级标签。6. 使用双向投票来预测最终的二元标签。7. 仅使用英语数据集对“distilbert-base-uncased”模型进行微调，并用它为从 Twitter 收集的五种语言数据的英语译文进行标注。8. 为验证领域特定标签，我们在表3 所示的 HSMerge 数据上训练了多个模型，并对步骤 6 中获取的仇恨样本进行预测。

4.2. 实验设置

为了评估模型的性能，我们使用加权平均 F1 在验证集上进行基准测试。所有实验均在配备最高 20 GiB RAM 的 NVIDIA A100 GPU 上完成。

4.3. 仇恨性（仇恨/非仇恨）

我们使用仇恨言论模型对原始 LAHM 数据集进行了初步验证；详情请参见第 5 节。

4.3.1. 单语实验

我们使用了来自 HuggingFace、基于 BERT 且针对特定语言的仇恨言论模型。这些模型均在同一种语言的数据上进行训练、验证和测试。实验中使用了以下语言模型：

1. Hate-speech-CNERG/dehatebert-mono-english [5]
2. l3cube-pune/hate-multi-roberta-hasoc-hindi[6]3.
Hate-speech-CNERG/dehatebert-mono-french[7]4.
Hate-speech-CNERG/dehatebert-mono-german[8]5.
Hate-speech-CNERG/dehatebert-mono-spanish[9]

单语实验的二元仇恨标签分布结果见表 8

4.3.2. 机器翻译实验

我们采用了利用开源机器翻译模型将英语数据翻译为多语言数据，反之亦然的方法。为了为每种语言选出最佳机器翻译模型，我们在一个小型人工标注数据集上评估了多个模型。翻译涵盖以下语言：印地语、阿拉伯语、法语、德语和西班牙语。我们使用这些翻译模型进行了两种类型的翻译：

1. 将从各种开源仇恨言论数据集中收集的英语数据集进行翻译。该数据集中的所有英语样本均已翻译成另外 5 种语言。2. 将从 Twitter 收集的多语言数据进行翻译。这些数据涵盖上述 5 种语言，每种语言的数据均已翻译成相应的英语数据。

选择和评估翻译模型时，主要考虑它们是否开源、是否免费且易于使用，以及翻译质量。使用的翻译模型包括：

1. Google 表格翻译 2.
m2m-100-1.2B[10]3.
IndicTrans[11]

使用的评估指标包括 BLEU、rougeL 和语义相似度。对于语义相似度，*all-mpnet-base-v2* 模型来自 *sentence-transformers* 库，用于计算输入与译文句子嵌入之间的余弦得分。图 4.3.2 展示了 3 个模型在不同语言上的 BLEU、rougeL 和语义相似度分数对比。比较各模型的表现并考虑其他因素后，*IndicTrans* 模型被选用于所有印度语言（印地语）的翻译，*m2m-100-1.2B* 则被选用于其他语言。

[5]https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-english[6] l3cube-pune/hate-roberta-hasoc-hindi [7] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-french [8] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-german [9] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-spanish [10] https://github.com/UKPLab/EasyNMT [11] https://github.com/AI4Bharat/indicTrans

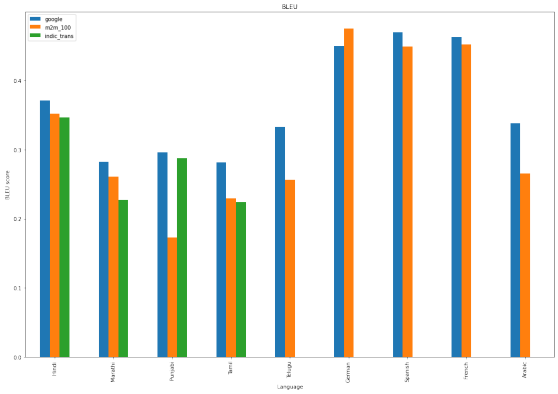


图 2： 模型翻译的 BLEU 分数

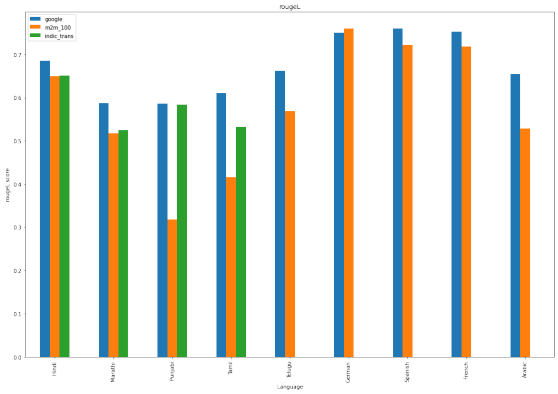


图 3： 模型翻译的 ROUGE-L 分数

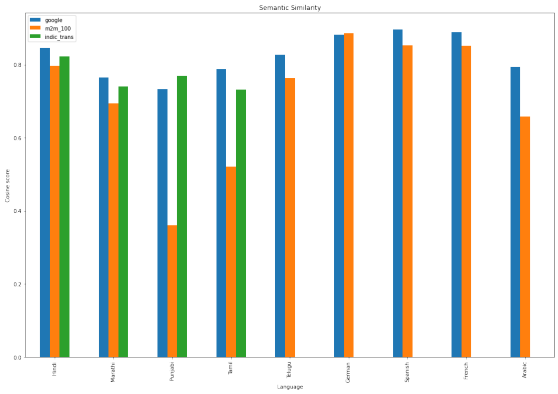


图 4： 模型翻译的语义相似度分数

4.3.3. 二元多语言实验

我们使用了 *mbert* ，这是一个针对维基百科语料库规模最大的 104 种语言预训练的模型，采用掩码语言建模（MLM）目标进行训练，并对其进行微调以用于二分类

用于开展少样本学习实验的分类任务。目标是让分类器利用少数几种语言的样本进行学习，并在其他语言上测试。训练数据集以英语为主，整理自多个开源资源。除此之外，我们还使用基于神经机器翻译的开源模型（$m2m_100_1.2B$）[12]，引入了少数其他语言的样本。神经机器翻译是一种利用人工神经网络预测最可能的词语序列，从而实现机器翻译的方法。

模型在自定义多语言数据集上进行训练、测试和验证，该数据集按每种选定语言的类别分布整理，如表4 所示。我们使用该模型对整理好的数据集进行预测。

表4多语种二元分类器训练数据

语言	Hate	非仇恨	总计
英语	20620	18171	38791
阿拉伯语	20539	18138	38680
印地语	20541	18138	38679
法语	20542	18138	38680
总计	**82242**	**72584**	**154826**

4.3.4. Perspective API 实验

Perspective API 是一款可免费使用的 API，利用机器学习识别评论中的毒性。它的 API 速率限制为每秒 1 个请求。由于这一限制，我们将其用于英语以外的所有语言。我们使用 Google Cloud 生成 Perspective API 密钥来发送请求，具体使用的是 *CommentAnalyzer* API，以获取每个数据点的有害性评分。该 *CommentAnalyzer* API 支持我们所考虑的全部 5 种语言。我们使用收集到的 5 种语言的 Twitter 清理后数据，获取毒性评分，并将每条推文分类为仇恨或非仇恨。相关数字见表 8。

4.4. 多领域实验

Thi本节介绍多领域实验，以验证 LAHM 数据集中的标签。

4.4.1. 基于机器翻译的多领域实验

表 3 中的 HSMerge 数据用于仇恨言论的多领域分类。该数据集经过预处理，删除了 URL、提及/用户名、空值和重复项。

目标是利用迁移学习，仅在英语数据集上微调模型，并用该模型为从其他 5 种语言的 Twitter 收集的数据的英语译文进行标注。

[12]https://github.com/pytorch/fairseq/tree/main/examples/m2m_100

用于在该数据集上进行微调的模型是*"distilbert-base-uncased"*[13]。DistilBERT 是 BERT 基础模型的蒸馏版本。与 bert-base-uncased 相比，它的参数量减少了 40%，速度提高了 60%，同时保留了 BERT 近 97% 的性能[14]。我们在模型的池化输出之上添加了一个线性层，用于多分类。训练过程中，最大序列长度设为 256，批次大小设为 32，学习率为 5e-5，训练了 4 个训练轮次。数据集通过分层随机抽样划分为训练集、验证集和测试集。数据集详情：

1. 20,174 个训练样本 2.
5,044 个验证样本 3.
6,305 个测试样本

测试集上的模型评估指标见表5。

表 5翻译后的多类别指标

领域	精确率	召回率	F1分数	支持度
种族主义	0.94	0.95	0.94	2882
性别歧视	0.93	0.93	0.93	2295
宗教仇恨	0.92	0.93	0.92	409
辱骂	0.94	0.91	0.93	589
极端主义	0.83	0.82	0.82	130
准确率			0.93	6305
宏平均	0.91	0.91	0.91	6305
加权平均	0.93	0.93	0.93	6305

4.4.2. 基于跨语言的多领域实验

为了使用 Hatebase 中的关键词验证我们的细粒度标签，我们针对在 HSMerge 数据集上训练的跨语言模型进行了实验，如表 3 所示。在此实验设置中，我们针对多分类任务对 *mbert* 模型进行微调，并在表 3 中的 HSMerge 数据上训练该模型，然后使用其他语言进行预测。实证结果见表 7。在这种零样本设置下，模型在英语数据集上完成微调后，不会接触任何其他语言的样本；此外，预训练和微调过程中均未使用机器翻译。

我们未能保持类别分布，因此极端主义类别中的数据点相对较少。为解决这一问题，我们引入了自定义类别权重，使少数类获得同等关注；在本例中，少数类是极端主义。我们使用加权类别随机采样器*WeightedSampler* 处理类别不平衡，使所有类别具有相同的采样概率。跨语言模型取得的经验得分见表7。

[13]https://huggingface.co/distilbert-base-uncased [14] https://arxiv.org/abs/1910.01108

我们针对具体应用场景微调的开源*mbert*模型此前已接受零样本实验[15]，并取得了不错的经验得分，如表6所示。数据集详情：

1. 19298 个训练样本 2.
6432 个验证样本 3.
6434 个测试样本

表6BERT 零样本黄金 **F1**分数

模型	英语	西班牙语	德语	阿拉伯语
BERT零样本	81.4	74.3	70.5	62.1

表7HSMerge 跨语言指标

领域	精确率	召回率	F1分数	支持度
种族主义	0.93	0.95	0.94	2161
性别歧视	0.93	0.91	0.92	1722
宗教仇恨	0.93	0.93	0.93	307
辱骂	0.93	0.92	0.92	442
极端主义	0.85	0.80	0.83	97
准确率			0.93	4729
宏平均	0.91	0.90	0.91	4729
加权平均	0.93	0.93	0.93	4729

5. 评估

我们使用开源单语模型、Perspective API 和训练的多语种二元分类器，在原始 LAHM 数据集上进行了初步验证（仇恨性）实验。我们比较了单语和多语言模型的结果，发现基于少样本学习的多语言分类器表现优于 HuggingFace 的开源单语 BERT 模型，但在阿拉伯语和法语上，后者表现更好；详情见表 8。根据分析，我们至少采纳了表现较好的模型中的 2 票，详情见表 9。

对于印地语，Perspective API 的表现更好，将 31% 的内容预测为仇恨内容，而二元分类器的预测比例为 12%。对于阿拉伯语，二元分类器的表现显著更好。

对于领域验证，我们使用了表中基于零样本跨语言、多领域和机器翻译的多领域模型，这些模型使用 HSMerge 数据进行训练3并进行了

[15]https://github.com/google-research/bert/blob/master/multilingual.md#results

表 8所有模型对高层级标签“仇恨”和“非仇恨”的预测

语言	单语		多语言		Perspective	
	Hate	非仇恨	Hate	非仇恨	Hate	非仇恨
英语	7247	101345	11380	99598	-	-
印地语	210	260	3970	7247	1674	3720
阿拉伯语	6469	26265	3914	28820	1376	31358
法语	9148	24185	8893	24440	7540	13269
德语	468	23910	3905	20473	4467	4164
西班牙语	2088	10087	4734	7441	24952	30196
总计	**25630**	**186052**	**36796**	**188019**	**40009**	**82707**

表 9多模型仇恨言论检测投票

语言	单语	多语	Perspective API
英语	✓		✓
印地语		✓	✓
阿拉伯语	✓	✓	
法语	✓	✓	
德语		✓	✓
西班牙语	✓		✓

表10跨语言多领域预测分布

语言	辱骂	性别歧视	种族主义	宗教	极端主义
英语	4228	2934	1566	751	154
阿拉伯语	1895	994	736	164	5631
印地语	1511	440	282	22	1715
法语	813	779	2393	750	4413
德语	180	123	30	4	131
西班牙语	818	483	139	11	637
总计	**9445**	**5753**	**5146**	**1702**	**12681**

表11机器翻译多领域预测分布

语言	辱骂	性别歧视	种族主义	宗教	极端主义
英语	8691	6693	2495	261	524
阿拉伯语	430	1302	209	259	1138
印地语	132	204	67	137	105
法语	745	1004	545	273	538
德语	232	608	186	58	98
西班牙语	2082	2024	1519	347	1310
总计	**12312**	**11835**	**5021**	**1335**	**3713**

不同语言上的预测。跨语言模型的标签分布见表10。多类别翻译模型在 LAHM 数据集上针对不同语言的预测标签分布见表11。

在印地语中，31.6% 的样本属于辱骂，种族主义所占比例最低，为 10.3%。在阿拉伯语中，性别歧视和极端主义合计占所有仇恨样本的 73%。在法语中，32% 的样本属于辱骂，宗教仇恨所占比例最低，为 8.3%。在德语中，性别歧视占 51%，而宗教仇恨和极端主义合计所占比例最低，为 13%。在西班牙语中，辱骂和性别歧视各占 27%，宗教仇恨占 5%。

6. 结论

我们介绍了 LAHM 数据集，这是一个大规模的半监督训练数据集，用于多语言、多领域的仇恨言论识别。该数据集由我们通过三层标注流程，并结合单语、多语言和跨语言模型构建。据我们所知，LAHM 是同类数据集中规模最大的，包含近 30 万条推文，涵盖 6 种语言和 5 个领域。

LAHM 支持在五个领域开展跨语言辱骂性语言检测，并深入研究语言迁移与领域迁移之间的相互作用。我们通过一系列跨领域实验，在单语、多语言和跨语言设置下使用先进的迁移学习模型评估仇恨言论检测，从而将 LAHM 描绘为一项综合资源。

我们希望 LAHM 能够激发更多关于理解和构建半监督式大规模多语言、多领域辱骂性语言检测数据集的研究工作。

7. 未来工作

作为未来工作，我们探索了社交网络中的变形文字检测与识别。社交媒体上的大量仇恨内容使用变形文字来躲避审核人员和自动化系统的检测。我们为这 5 个领域分别收集了属于该类别的仇恨关键词，并针对每个关键词使用一组变形文字进行实验，以提取使用变形文字表达的仇恨内容。我们计划在多语言和多领域设置中将其用于未来工作。

参考文献

[1] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, D.-Y. Yeung，多语言与多维度仇恨言论分析，载于：EMNLP论文集，计算语言学协会，2019年。[2] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini，COnan——通过利基众包构建反叙事：用于抵制网络仇恨言论的多语言回应数据集，arXiv预印本 arXiv:1910.03270（2019年）。[3] T. Davidson, D. Warmsley, M. Macy, I. Weber，自动仇恨言论检测与冒犯性语言问题，载于：国际 AAAI 网络与社交媒体会议论文集，第11卷，2017年。[4] V. Kolhatkar, M. Taboada，用于在线新闻评论分析的语料库（????）。[5] L. Gao, R. Huang，使用上下文感知模型检测网络仇恨言论，arXiv预印本 arXiv:1710.07395（2017年）。[6] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri，社交媒体攻击性识别的基准评测，载于：首届网络挑衅、攻击与网络欺凌研讨会（TRAC-2018）论文集，2018年，第1–11页。[7] E. Wulczyn, N. Thain, L. Dixon，机械降神：大规模视角下的人身攻击，载于：第26届国际万维网大会论文集，2017年，第1391–1399页。[8] Z. Waseem, D. Hovy，仇恨符号还是仇恨者？用于Twitter仇恨言论检测的预测特征，载于：NAACL 学生研究研讨会论文集，计算语言学协会，加利福尼亚州圣迭戈，2016年，第88–93页。网址：http://www.aclweb.org/anthology/N16-2013。[9] T. Mandl, S. Modha, C. Mandlia, D. Patel, A. Patel, M. Dave，HASOC——印欧语系语言中的仇恨言论与冒犯性内容识别，2019年。[10] A. Basile, C. Rubagotti，Evalita2018中用于AMI的Crotonemilano：一个高性能的跨语言厌女症检测系统，EVALITA 意大利语自然语言处理与语音工具评测 12（2018年）206。[11] M. Wiegand, M. Siegel, J. Ruppenhofer，GermEval 2018冒犯性语言识别共享任务概述（2018年）。[12] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, B. J. Jansen，面向多个社交媒体平台的在线仇恨分类器开发，以人为中心的计算与信息科学 10（2020年）1–34。[13] S. A. Chowdhury, H. Mubarak, A. Abdelali, S.-g. Jung, B. J. Jansen, J. Salminen，一个多-

平台阿拉伯语新闻评论数据集，用于冒犯性语言检测，载于：第十二届语言资源与评估会议论文集，2020年，第6203–6212页。[14] M. Wiegand, J. Ruppenhofer, A. Schmidt, C. Greenberg，归纳构建辱骂词词典——一种基于特征的方法（2018年）。[15] M. Karan, J. Šnajder，在线辱骂性语言的跨领域检测，载于：第二届在线辱骂性语言研讨会论文集（ALW2），2018年，第132–137页。[16] E. W. Pamungkas, V. Patti，跨领域与跨语言辱骂性语言检测：一种结合深度学习与多语言词典的混合方法，载于：计算语言学协会第57届年会论文集：学生研究研讨会，2019年，第363–370页。[17] E. W. Pamungkas, V. Basile, V. Patti，Twitter上的厌女内容检测：一项多语言跨领域研究，《信息处理与管理》57（2020年）102360。[18] J. S. Meyer, B. Gambäck，一种平台无关的双分支仇恨言论检测器，载于：ACL 2019 第三届在线辱骂性语言研讨会论文集，计算语言学协会，2019年。[19] K. Wang, D. Lu, S. C. Han, S. Long, J. Poon，检测所有辱骂！迈向通用辱骂性语言检测模型，arXiv预印本 arXiv:2010.03776（2020年）。[20] A. Arango, J. Pérez, B. Poblete，仇恨言论检测并不像你想象的那么简单：深入审视模型验证，载于：第42届ACM信息检索研究与发展国际会议论文集，2019年，第45–54页。[21] Z. Waseem, J. Thorne, J. Bingel，弥合差距：用于仇恨言论检测领域迁移的多任务学习，载于：网络骚扰，施普林格，2018年，第29–55页。[22] M.-A. Rizoiu, T. Wang, G. Ferraro, H. Suominen，用于社交媒体仇恨言论检测的迁移学习，arXiv预印本 arXiv:1906.03829（2019年）。[23] T. Caselli, V. Basile, J. Mitrović, M. Granitzer，HateBERT：重新训练BERT以检测英语辱骂性语言，arXiv预印本 arXiv:2010.12472（2020年）。[24] A. Koufakou, E. W. Pamungkas, V. Basile, V. Patti，HurtBERT：将词汇特征与BERT相结合以检测辱骂性语言，载于：第四届在线辱骂与伤害研讨会论文集，2020年，第34–43页。[25] G. Glavaš, M. Karan, I. Vulic，XHate-999：跨领域与跨语言分析和检测辱骂性语言，计算语言学协会，2020年。[26] M. Mozafari, R. Farahbakhsh, N. Crespi，基于BERT模型的社交媒体仇恨言论检测与种族偏见缓解，《公共科学图书馆·综合》15（2020年）e0237861。[27] K. B. Ozler, K. Kenski, S. Rains, Y. Shmargad, K. Coe, S. Bethard，微调BERT以实现多领域多标签不文明语言检测，载于：第四届在线辱骂与伤害研讨会论文集，2020年，第28–33页。[28] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee，深入探究多语言仇恨言论分类，载于：数据库中的机器学习与知识发现。应用数据科学与演示专题：欧洲会议，ECML PKDD 2020，比利时根特，2020年9月14–18日，论文集，第五部分，施普林格国际出版，2021年，第423–439页。[29] N. Vashistha, A. Zubiaga，在线多语言仇恨言论检测：印地语与英语社交媒体实验，《信息》12（2021年）5。[30] M. O. Ibrohim, I. Budi，用于多语言仇恨言论的翻译与非翻译方法

Twitter中的识别，《国际先进科学、工程与信息技术杂志》9（2019年）1116–1123。[31] M. Corazza, S. Menini, E. Cabrio，基于混合表情符号掩码语言模型的零样本辱骂性语言检测，载于：计算语言学协会会议论文集：EMNLP 2020，计算语言学协会，线上，2020年，第943–949页。URL：https://aclanthology.org/2020.findings-emnlp.84。doi:10.18653/v1/2020.findings-emnlp.84.[32] H. Ahn, J. Sun, C. Y. Park, J. Seo, NLPDove 在 SemEval-2020 任务12中的工作：通过跨语言迁移改进冒犯性语言检测，载于：第十四届语义评测研讨会论文集，国际计算语言学委员会，巴塞罗那（线上），2020年，第1576–1586页。URL：https://aclanthology.org/2020.semeval-1.206。doi:10.18653/v1/2020.semeval-1.206.[33] J. M. Pérez, A. Arango, F. Luque，ANDES 在 SemEval-2020 任务12中的工作：用于冒犯性语言检测的联合训练BERT多语言模型，载于：第十四届语义评测研讨会论文集，国际计算语言学委员会，巴塞罗那（线上），2020年，第1524–1531页。URL：https://aclanthology.org/2020.semeval-1.199。doi:10.18653/v1/2020.semeval-1.199.[34] L. Stappen, F. Brunn, B. Schuller，利用冻结的Transformer语言模型和axel进行跨语言零样本与少样本仇恨言论检测，arXiv预印本 arXiv:2004.13850（2020年）。[35] T. Dadu, K. Pant，Team rouges 在 SemEval-2020 任务12中的工作：通过跨语言归纳迁移检测冒犯性语言，载于：第十四届语义评测研讨会论文集，国际计算语言学委员会，巴塞罗那（线上），2020年，第2183–2189页。URL：https://aclanthology.org/2020.semeval-1.290。doi:10.18653/v1/2020.semeval-1.290.[36] T. Ranasinghe, M. Zampieri，使用跨语言嵌入进行多语言冒犯性语言识别，载于：2020年自然语言处理实证方法会议（EMNLP）论文集，计算语言学协会，线上，2020年，第5838–5844页。URL：https://aclanthology.org/2020.emnlp-main.470。doi:10.18653/v1/2020.emnlp-main.470.[37] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Dercynski, Z. Pitenis, Ç. Çöltekin，SemEval-2020 任务12：社交媒体中的多语言冒犯性语言识别（OffensEval 2020），arXiv预印本 arXiv:2006.07235（2020年）。[38] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis，大规模众包与Twitter辱骂行为特征分析，载于：第十二届AAAI国际网络与社交媒体会议，2018年。[39] R. Gomez, J. Gibert, L. Gomez, D. Karatzas，探索多模态出版物中的仇恨言论检测，载于：IEEE/CVF冬季计算机视觉应用会议论文集，2020年，第1470–1478页。[40] Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco，数据收集中的人在回路：用于打击网络仇恨言论的多目标反叙事数据集，载于：计算语言学协会第59届年会论文集，计算语言学协会，2021年。