arXiv:2303.04222v2 [cs.CL] 8 May 2023

SemEval-2023任务10：网络性别歧视的可解释检测 m

Hannah Rose Kirk[1, 2*]，**Wenjie Yin**[1, 3*]、**Bertie Vidgen**[1]，和**Paul Röttger**[1,2]

[1]Rewire[2]牛津大学[3]伦敦玛丽
女王大学

摘要

网络性别歧视是一种普遍且有害的现象。自动化工具有助于大规模检测性别歧视。然而，二元检测忽视了性别歧视内容的多样性，也无法清楚解释某种内容为何具有性别歧视。为解决这一问题，我们推出了SemEval网络性别歧视可解释检测任务（EDOS），即SemEval任务10。我们的主要贡献有三项：i）提出一种新颖的性别歧视内容层级分类体系，其中包含细粒度的性别歧视向量，以帮助提升可解释性；ii）构建一个包含20,000条社交媒体评论的新数据集，并为其提供细粒度标签，同时提供更大规模的未标注数据集用于模型适配；iii）提供基线模型，并分析参与者提交至该任务的方法、结果和错误。

内容警告： 我们展示性别歧视语言的示例，以描述该分类体系并分析错误类型。

1 引言

网络性别歧视会对受到针对的女性造成伤害，使网络空间变得难以进入且不受欢迎，并延续社会不平等与不公正现象。如今，人们广泛部署自动化工具，以大规模发现性别歧视内容，为内容审核、监测和研究提供支持。然而，大多数自动分类器除了“有害性”“辱虐”或“性别歧视”等通用的高层级类别外，不会提供进一步解释。缺乏解释的不透明分类可能在受审核用户中引发困惑、愤怒和不信任，也使人们更难质疑审核错误；使用自动化工具的审核员可能不信任或无法充分理解这些标签，从而降低工作效率；此外，在缺乏细粒度标签的情况下，很难评估模型的弱点，这会阻碍更优模型的开发。

我们组织了 SemEval 2023 任务 对网络性别歧视进行可解释检测（*EDOS*），该任务延续了以往在辱虐和仇恨检测方面成功开展的共享任务（Zampieri等人，2019年；Ribeiro 和 Silva，2019年；Pavlopoulos 等人，2021年；Basile 等人，2019年；Fersini 等人，2018a，b）。我们的任务提出并应用了一套分类体系，包含三个层级的任务，用于检测性别歧视内容。**任务 A** 是一项二元任务，用于检测内容是否具有性别歧视性质。对于具有性别歧视性质的内容，**任务 B** 将其划分为四种不同的性别歧视类别，而**任务 C** 则识别 11 种细粒度性别歧视向量中的一种。

我们的 EDOS 任务具有三个独特之处：第一，**数据多样性**：我们使用过滤方法集成，从两个大型社交媒体平台（Reddit 和 Gab）采样数据。与仅从关键词或单一平台采样相比，这些条目更加多样。第二，**标注质量**：我们招募了经过高强度培训、且均自我认定为女性的标注员。我们作出这一决定，是为了减轻标注中的隐性偏见，并践行一种参与式 AI 方法：让主要受到特定形式内容伤害的群体参与数据开发过程（Birhane 等，2022年；Zytko 等，2022年）。为依据详细指南提高标签的一致性，我们对所有标注员进行了多项试点任务培训，并由专家解决分歧。第三，**任务粒度**：我们的数据集建立在精心构建的分类体系之上，使任务 B 和任务 C 中的多个类别从社会科学视角来看都具有意义，有助于解释预测结果，并为多分类机器学习提出具有挑战性的任务。

通过发布该数据集并分析配套 SemEval 任务的提交结果，我们为构建更准确且可通过细粒度标签实现可解释性的自动化性别歧视检测系统作出贡献。[1] 这些贡献标志着朝着为女性打造更安全互联网迈出了一步。

*贡献相同

[1]所有任务数据、个人标注和指南均可在github.com/rewire-online/edos获取。数据说明见附录A。

2 分类体系设计

2.1 现有分类体系

我们对有关性别歧视和厌女症的研究进行了广泛综述。在这些文献中，我们确定了近期提出性别歧视或厌女内容分类体系的文章，尤其包括Jha和Mamidi（2017年）；Samory 等人（2021年）；Farrell 等人（2019年）；Zeinert 等人（2021年）；Guest等人（2021年）；Rodríguez-Sánchez 等人（2021年）以及Parikh 等人（2021年）。每个分类体系都各具特色，但它们通常在若干维度上存在差异，包括：（i）构建方式（基于理论还是实证构建）；（ii）范围（如何定义顶层的性别歧视或厌女症类别，以及包含哪些子类别）；以及（iii）结构（子类别是否按层级排列）。我们在附录B中回顾了这些细节。

2.2 我们的分类体系

在回顾相关文献的基础上，我们拟定了分类体系的初稿。随后，我们采用扎根理论方法（Glaser 和 Strauss，2017年），并结合数据集中的实证条目（见 §3）进一步完善分类体系，合并或调整其中的类别。该体系包含三个子任务：

任务A：二元性别歧视我们的分类体系的第一级将内容二分为性别歧视内容和非性别歧视内容。我们将性别歧视内容定义为任何基于女性的性别，或基于其性别与一种或多种其他身份属性（例如黑人女性或穆斯林女性）组合而针对她们的辱虐，无论是隐含还是明确的。我们的分类体系关注的是性别歧视，而非厌女症。厌女症指“对女性表达仇恨”（Ussher，2016年），而性别歧视也涵盖较为隐晦的隐性辱虐和偏见，这些同样可能对女性造成重大伤害。[2]

任务B：性别歧视类别 我们分类体系的第二层将性别歧视内容细分为四个在概念和分析上各不相同的类别。我们有意不按对接受者可能造成的影响或说话者可能的动机来区分类别，因为性别歧视内容造成的伤害因人而异；而且说话者意图很难判断，尤其是缺乏更广泛的语境时。这四个类别是**(1) 威胁、伤害计划与煽动：** 个人表达意图和／或鼓励他人采取行动对付女性，并对她们造成或煽动严重伤害和暴力的语言。其中包括身体、性或隐私伤害方面的威胁。**(2) 贬损：** 明确贬低、非人化、羞辱或侮辱女性的语言。其中包括对女性的负面描述和刻板印象、对她们身体的物化、强烈的负面情绪化陈述，以及非人化的类比。它涵盖针对特定女性和女性群体的负面陈述。**(3) 敌意：** 表达隐性或微妙的性别歧视、刻板印象或描述性陈述的语言。其中包括善意性别歧视，即以赞美的形式表达。**(4) 偏见性讨论：** 否认歧视存在并为性别歧视待遇辩护的语言。其中包括否认和为性别不平等辩护、为虐待女性开脱，以及男性受害者论意识形态。

任务 C：细粒度向量类别 我们分类体系的第三层将每个性别歧视类别细分为细粒度的性别歧视向量类别。我们旨在构建彼此互斥（即每个向量类别各不相同）且完备穷尽（即所有性别歧视内容都可归入某一向量类别）的向量类别。表 1概述了我们分类体系中的 11 种细粒度向量类别，并列出其定义以及取自既有文献的示例；我们对这些示例进行了编辑，以清晰展示各个不同的向量类别。

3 数据集构建

3.1 平台选择

大量线上伤害研究都从 Twitter 收集数据（Vidgen 和 Derczynski，2020年），导致研究在性别歧视内容的严重程度和形式方面缺乏基于社群的多样性。因此，为了构建一个涵盖性别歧视细粒度向量类别的多样化数据集，我们选择使用两个社交媒体平台。首先是 **Gab**，这是一个于 2016年创立的“另类科技”社交网站。它将自身定位为主流网站（如 Twitter 和 Facebook）的竞争者，并明确致力于保护言论自由（包括极右翼言论）（Zannettou 等人，2018年；Lima 等人，2018年；Jasser 等人，2021年）。其次是 **Reddit**，这是一个由基于主题的论坛（“子版块”）组成的网络，用户可以分享、浏览并评论与其兴趣相关的内容

[2]我们的分类体系不涵盖主要针对女性以外群体的辱虐，例如种族主义或反犹太主义。不过，我们要求标注员使用“其他目标”标签标记此类内容；根据多数投票，该标签适用于 15,146 条非性别歧视条目中的 2,098 条。

类别	媒介	定义	示例
1. 威胁、伤害计划伤害和煽动	**1.1** 伤害威胁	表达伤害个人的意图、意愿或欲望女性或女性群体。这可能包括但不限于以下形式的伤害：身体、性、情感或隐私方面的伤害。	任何顶嘴的女性我都要杀掉对我
	1.2 煽动和鼓励伤害	煽动或鼓励个人、群体或普通受众伤害女性或女性群体。其中包括作者试图向他人合理化和/或为其辩护的语言伤害女性。	强奸她会让她回到自己的位置
2. 贬损	**2.1** 描述性攻击	以贬损的方式刻画或描述女性。这可能包括但不限于：负面概括女性能力、外貌、性行为、智力 ，品格 ，或道德观。	女子足球真烂，她们动作太慢又笨拙
	2.2 攻击性和情绪化攻击	表达针对女性的强烈负面情绪，例如厌恶或仇恨。这可以通过直接描述说话者的主观情绪、毫无根据的指控，或使用性别侮辱语、基于性别的脏话和基于性别的侮辱。	我仇恨女性
	2.3 非人化攻击以及露骨的性物化	通过将女性比作非人类实体来贬低她们例如害虫、疾病或废弃物，或公然将她们贬低为性对象。	女性是猪
3. 敌意	**3.1** 因果性使用性别侮辱语、脏话和侮辱	使用性别侮辱语、基于性别的脏话和侮辱，但并非为了故意攻击女性。只有那些传统上用于描述女性的词语才在范围内（例如“贱女人”、“荡妇”）	别这么像个小贱人一样
	3.2 不可改变的性别差异与性别刻板印象	声称存在不可改变的、自然的或其他本质上的男性与女性之间的差异。在某些情况下 , this可能会通过利用女性特质来攻击男性。大多数性别歧视笑话都属于这一类。	男性和女性的大脑天生就不一样，兄弟，事实就是如此
	3.3 含蓄贬低的性别化赞美	表面上是在赞美女性，实际上却是在贬低她们，或暗示她们低人一等。这可能包括但不限于：将女性的价值归结为她们的吸引力或性吸引力，或暗示女性天生脆弱、无助或软弱。	女性是娇弱的花朵，她们需要得到珍视
	3.4 居高临下解释或不受欢迎的建议	就女性更了解的话题向她们提供未经请求或居高临下的建议和问题提供建议（这种行为称为“男性说教”）。	我女朋友总是抱怨经痛，但她就是不理解消除经痛的医学原理！
4. 偏见性讨论	**4.1** 支持对个别女性	表达对将女性视为个人所遭受虐待的支持。支持可以通过否认、轻描淡写或试图为此类虐待辩解来表达。	女性不应该露出那么多皮肤，如果她们遭到强奸
	4.2 支持系统性针对女性这一群体的歧视	表达对系统性歧视女性这一群体的支持。支持可以通过否认、淡化或试图为此类歧视辩护来表达。	男性在董事会会议室中的领导地位是一种必要的恶——公司必须高效运行

表1：性别歧视类别分类体系（任务 B）和细粒度向量类别（任务 C）。

。多项研究表明，Reddit 上有许多性别歧视和反女权主义社群，这些社群被称为“男性圈”（Ging，2017年；Zuckerberg，2018年；Farrell 等人，2019年；Ging 等人，2020年；Ribeiro 等人，2021年）。

我们从 Reddit 和 Gab 获取数量相等的数据。对于每个平台，我们首先创建一个包含 100 万条记录的数据池（见 §3.2）。然后，我们从每个数据池中抽取 1 万条记录进行标注（见 §3.3）。

3.2 数据收集

Gab 我们收集了 2016年8月至2018年10月期间公开发布的 3400 万条 Gab 帖文。[3] 这类数据已广泛用于学术研究（例如，Kennedy等人，2018年；Cinelli等人，2021年）。我们随机抽取 100 万条记录来创建数据池。

Reddit 首先，根据以往研究（Guest等人，2021年；Farrell 等人，2019年；Zuckerberg，2018年；Jones等人，2020年；Qian等人，2019年；Ging，2017年；Ribeiro 等人，2021年）以及在线维基。[4] 接着，我们使用 Reddit API 收集了这些子版块中从2016年8月至2018年10月的所有评论。[5] 我们人工审核每个子版块，并根据Lilly （2016年）之前的研究，将其划分为四个类别之一（非自愿独身者、男人走自己的路、男性权利活动人士、搭讪艺术家）。[6] 为确保数据集不会过度偏向小众的语言表达和话题，我们将抽样范围限制在至少有10万条评论的24个子版块，最终得到一个包含4200万条评论的数据集。最后，我们从四个子版块类别中各随机抽取25万条评论，组成样本池。

[3] files.pushshift.io/gab

[4] rationalwiki.org/wiki 和incels.wiki[5]我们只收集评论，因为早期分析表明，许多帖子仅包含网址、图片或视频，因此单独解析并进行标注较为困难。我们将日期范围设定为与 Gab 数据相匹配。[6]我们在附录中介绍这些类别 C。

		训练集		dev		test	
任务 A							
非性别歧视		10,602	76%	1,514	76%	3,030	76%
性别歧视		3,398	24%	486	24%	970	24%
	总计	14,000	100%	2,000	100%	4,000	100%
任务 B							
1. 威胁、伤害计划和煽动		310	9%	44	9%	89	9%
2. 贬损		1,590	47%	227	47%	454	47%
3. 敌意		1,165	34%	167	34%	333	34%
4. 偏见性讨论		333	10%	48	10%	94	10%
	总计	3,398	100%	486	100%	970	100%
任务 C							
1.1 伤害威胁		56	2%	8	2%	16	2%
1.2 煽动和鼓励伤害		254	7%	36	7%	73	8%
2.1 描述性攻击		717	21%	102	21%	205	21%
2.2 攻击性和情绪化攻击		673	20%	96	20%	192	20%
2.3 非人化攻击和公然的性客体化		200	6%	29	6%	57	6%
3.1 随意使用性别歧视性侮辱语、脏话和辱骂		637	19%	91	19%	182	19%
3.2 不可改变的性别差异和性别刻板印象		417	12%	60	12%	119	12%
3.3 含蓄贬低的性别化赞美		64	2%	9	2%	18	2%
3.4 居高临下的解释或不受欢迎的建议		47	1%	7	1%	14	1%
4.1 支持虐待个别女性		75	2%	11	2%	21	2%
4.2 支持针对女性这一群体的系统性歧视		258	8%	37	8%	73	8%
	总计	3,398	100%	486	100%	970	100%

表2：各任务和数据划分中类别标签的分布。

3.3 数据准备与采样

清理 我们对 Gab 和 Reddit 样本池中的文本进行了清理：（1）将网址和用户名替换为通用标记；（2）删除空条目；（3）删除仅包含网址或表情符号的条目；（4）删除非英语条目；（5）删除重复项。完成这些清理步骤后，我们应用了抽样方法。

增强采样 现实环境中辱虐内容的流行程度难以可靠估计，但可能低至 0.1% 或 1%（Vidgen 等人，2019年）。性别歧视只是所有辱虐内容中的一部分；因此，从线上平台随机采样会造成数据集类别严重不平衡，从而妨碍人工智能系统的训练。以往研究使用了多种采样技术来提高数据集中辱虐内容的比例，包括关键词搜索（ElSherief 等人，2017年）或词典（Farrell 等人，2019年），以及基于社群的采样（Guest等人，2021年；Vidgen 等人，2021a）或基于用户的采样（Vidgen 等人，2019年）。出于隐私考虑，我们不存储基于用户的信息。取而代之的是，我们结合采用基于社群的采样（在 Reddit 上进行），以及精心设计的多种采样方法集成方案。

采样方法集成依赖单一采样方法会使采样数据更容易产生偏差（Yin 和 Zubiaga，2022年）。经过大量初步测试，我们最终选定六种不同技术，从 Gab 和 Reddit 的数据池中各采样 10k 条清理后的记录（总计 $n = 20,000$），以确保覆盖 11 种细粒度性别歧视向量类别，并引入词汇和主题多样性。这六种技术分别采样：（1）1,000 条至少包含一个性别歧视关键词（例如“c*nt”、“贱女人”）的记录；（2）1,000 条至少包含一个性别歧视关键词和一个主题关键词（例如“她”、“女孩”）的记录；（3）从 Perspective 模型毒性分数的每个十分位区间中抽取 100 条记录（共 1,000 条）[7]；（4）从一个专为性别歧视检测训练的定制分类器所得分数的每个十分位区间中抽取 100 条记录；该分类器使用七个开放源代码的性别歧视/厌女症数据集进行训练（共 1,000 条）[8]；（5）抽取 1,000 条记录，其 Perspective 的毒性模型分数与我们的定制分类器分数至少相差$|0.8|$；（6）结合主题关键词与 Perspective 其他属性的分数（例如，身份攻击 +“女孩”或露骨性内容 + “她”）抽取 5,000 条记录。

3.4 数据标注

我们在数据标注中采用“规范性范式”，希望标注员遵循我们全面的标注指南，而不是依据自己的主观信念进行标注（Röttger 等人，2022年）。我们的标注指南包含定义、澄清说明、示例、边缘案例和一般指导。

标注员招募 为降低隐性偏见的风险，并鼓励受影响群体参与，我们与受过培训的标注员合作——

[7] perspectiveapi.com [8] 有关训练集的详细信息，请参阅附录 D。

这些标注员均自我认同为女性。我们要求所有标注员通过一项具有挑战性的 200 条目筛选任务，该任务涵盖数据中的全部 11 个性别歧视向量类别。我们最终招募并培训了 19 名通过筛选的标注员。我们选择专家标注而非众包工作，因为试点实验表明，标签的质量和一致性存在显著差异。[9] 标注员的人口统计信息及其网络性别歧视经历记录于附录A。

标注流程 每个条目由三名标注员进行标注。为进一步确保标签质量，对于存在分歧的情况，我们依靠专家裁决。专家被要求为以下情况给出标签：(i) 在任务 A 中一致率低于 3/3（全体一致）的案例，以及 (ii) 在任务 B 和任务 C 中一致率低于 2/3 的案例。专家团队由本文的两位女性作者以及标注团队中两位最有经验的工作人员组成。在两个月期间，每两周将数据分批分配给标注员。整个过程中，我们与标注员密切合作，以便将他们的反馈纳入我们的指南，并持续监测和保障他们的福祉。

3.5 数据集分布

数据按照 70:10:20 的比例划分为训练集、开发集和测试集。任务 B 和任务 C 仅包含性别歧视实例。所有划分的标签分布如表所示。2。

4 任务说明

4.1 任务定义

我们的 SemEval 任务包含三个子任务，体现了我们的层级分类体系（§2.2）。任务 A 是一项二分类任务（性别歧视与非性别歧视）。任务 B 和任务 C 是多分类任务，分别包含四个和 11 个类别。

4.2 任务组织

我们在 CodaLab 上开展了 SemEval 任务。[10] 主要分为两个阶段：(i) 训练与开发阶段，时间为 2022年9月至2023年1月；(ii) 测试阶段，于2023年1月10日开始，并于2023年1月31日结束。在训练与开发阶段，我们发布了以下内容的条目和标签：训练集和开发集的划分，以及一个额外的入门套件，其中包含 100 万条未标注的 Reddit 条目和 100 万条未标注的 Gab 条目。在测试阶段，我们分批发布了任务 A 的测试条目，以及任务 B 和任务 C 的测试条目，因为后者提供了前者正确标签的信息。

4.3 评估指标和基线

我们使用宏平均 F1 分数评估所有系统，以考虑类别之间的不平衡。我们提供了 7 个基线模型（见表3），用于对从简单到复杂的各类系统进行基准测试。最简单的基线是始终预测最频繁类别（*B0*），以及均匀地预测每个类别（*B1*）。我们训练了一个传统机器学习模型（*B2*）：首先使用 TF-IDF 对数据进行向量化，然后用 XGBoost 模型进行拟合。其余基线采用 DistilBERT（Sanh 等人，2019年）和 DeBERTa-v3-base（He 等人，2021a）。对于每个模型，我们在训练集上进行微调（*B3, B5*），并使用 Reddit 和 Gab 合计 200 万条未标注数据进行持续预训练（*B4, B6*）。在所有任务中，经过持续预训练的 DeBERTa 达到了最高的基线水平。任务 A 的最佳基线（F1=0.8235）比任务 B 和任务 C 的最佳基线（F1=0.5926；0.3171）更接近饱和，因此 SemEval 的这项挑战颇具意义：其难度各不相同，而且仍有提升性能的空间。

5 参赛系统与结果

5.1 参赛者概览

EDOS 是一项非常受欢迎的 SemEval 任务，共有 599 个账户注册以获取数据集访问权限。[11] 在开发阶段，任务 A、任务 B 和任务 C 分别收到 128、71 和 52 份提交。在测试阶段，任务 A、任务 B 和任务 C 分别收到 134、87 和 81 份提交。

验证流程为确保测试阶段的竞赛公平，每个团队只能使用一个提交账户，并且最多可以提交两次任务预测，以便有机会纠正一次上传错误。[12]我们发起了一份简短的调查问卷，以确认参赛者遵守我们的条款与条件，并收集系统信息。最终排行榜纳入了回应这份调查问卷的参赛者：任务 A 有 84 个，任务 B 有 69 个，任务 C 有 63 个

[9]结果见附录 E。 [10] codalab.lisn.upsaclay.fr/competitions/7124

[11]截至 2023年4月25日 注册的账户数量。[12]遗憾的是，我们发现了一些可疑活动：多个账户使用通用邮箱 ABCD1234@domain，或提交 >10测试集结果。

	模型	续预训练	宏平均 F1 任务 A	任务 B	任务 C
基线					
B0	最频繁类	✗	0.4310	0.1594	0.0317
B1	均匀	✗	0.4509	0.2413	0.0629
B2	XGBoost	✗	0.4933	0.2297	0.0881
B3	DistilBERT	✗	0.7621	0.5531	0.2935
B4		✓	0.7804	0.5367	0.3140
B5	DeBERTa-v3-base	✗	0.8235	0.4790	0.1517
B6		✓	**0.8235**	**0.5926**	**0.3171**
排名靠前的系统					
PingAnLifeInsurance	DeBERTa-v3-large、twHIN-BERT-large	✓	**0.8746**		
stce	RoBERTa-large, ELECTRA	✓	0.8740	0.7203	0.5487
FiRC-NLP	DeBERTa（集成）	✗	0.8740		
JUAGE	PaLM（集成）	✗		**0.7326**	
PASSTeam	RoBERTa ，HateBERT	✓		0.7212	0.5412
PALI	DeBERTa ，RoBERTa	✓			**0.5606**
汇总统计					
		系统数量	84	69	63
		Q1	0.7994	0.5730	0.3153
		Mean	0.8095	0.5899	0.3829
		中位数	0.8322	0.6191	0.4230
		Std	0.0746	0.1065	0.1274
		Q3	0.8537	0.6501	0.4758
		击败*B0*的系统百分比	100%	100%	100%
		击败*B6*的系统百分比	55%	60%	73%

表 3：各任务的基线、排行榜排名靠前的结果及汇总统计。最佳基线以**粗体**标示。最佳参赛提交以**红色粗体**标示。

个团队。59 个团队提交了全部三项任务，10 个团队提交了其中两项，20 个团队只提交了一项。

5.2 排行榜结果

各项任务排名前三的提交见表 3。[13] 所有任务中排名前三的系统都采用了多个模型或集成，许多顶尖系统还采用了继续预训练和多任务学习。值得注意的是，stce 和 PASSTeam 在两个或更多任务中取得了优异成绩。两者都采用了多任务学习，并使用入门套件中的未标注数据进行继续预训练：stce 使用 RoBERTa-large（Liu 等人，2019年）和 ELECTRA（Clark 等人，2020年）；而 PASSTeam 使用多任务学习策略（Yu 等人，2020年），结合微调后的 RoBERTa 和 HateBERT（Caselli 等人，2021年）。在任务 A 中，PingAnLifeInsurance 也采用了多任务深度神经网络结构（Liu 等人，2020年），并使用入门套件中的未标注数据和一个额外的 Kaggle 数据集，对 DeBERTa-v3（He 等人，2021a）和 TwHIN-BERT（Zhang 等人，2022c）进行继续预训练。FiRC-NLP 使用了由多个 DeBERTa 模型组成的集成，这些模型仅使用已标注任务数据进行微调。在任务 B 中，JUAGE 是少数依赖基于提示的学习的系统之一，但它使用指令微调的 PaLM 模型（Chowdhery 等人，2022年），并在任务数据上进行参数高效的提示调优，再通过六轮多数投票取得了第一名。在任务 C 中，*PALI* 使用入门套件中的未标注数据对 DeBERTa-v3 进行继续预训练，并采用了第二个损失项（归一化温度缩放交叉熵）。

任务 B 和任务 C 的均值显著低于任务 A，方差则显著高于任务 A（见表 3）。所有参赛系统都优于我们最简单的基线（预测最频繁类别），而大多数系统也优于我们更复杂的基线，即采用持续预训练的 DeBERTa-v3。

5.3 常用方法

架构 对于所有任务，绝大多数参与者（~90%）使用了 Transformer 架构（图 1）。最受欢迎的基于 Transformer 的模型包括 RoBERTa（Liu 等人，2019年）、DeBERTa（He 等人，2021b,a）、BERT（Devlin 等人，2019年）、BERTweet（Nguyen 等人，2020年）和 DistilBERT（Sanh 等人，2019年）。若干提交使用了提示式语言模型，例如 GPT-2（Radford 等人，2019年）、GPT-3（Brown 等人，2020年）、PaLM（Chowdhery 等人，2022年）或 OPT（Zhang 等人，2022b）。其他方法还包括传统机器学习方法（~8%）和非 Transformer 深度神经网络（~6%），它们通常会与其他方法结合使用。

额外训练与数据 大多数参与者仅在目标任务上进行微调（>90%），但也有一些参与者进一步进行了预训练（~30%），或在辅助任务上进行微调

[13] 完整排行榜可在以下网址查看：github.com/rewire-online/edos

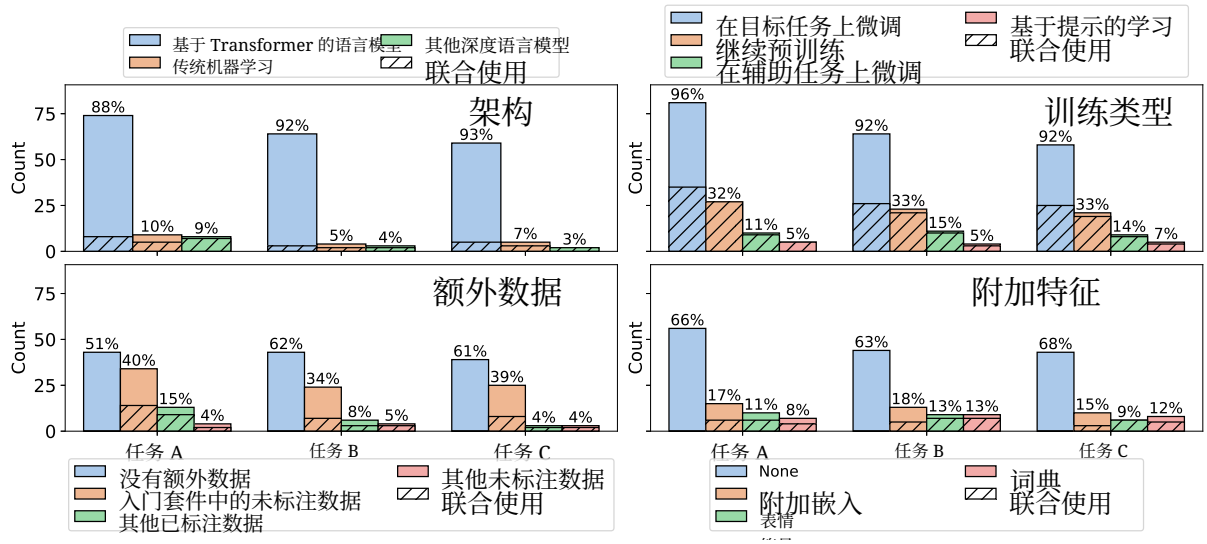


图 1：参与者方法的流行度（使用该方法的提交数量及其占比）。每个子图展示了系统论文中方法选择的不同细分情况。第一个子图（左上）显示，大多数参与者使用了基于Transformer的语言模型。第二个子图（右上）显示，在目标任务上进行微调是最受欢迎的训练方法，但相当一部分参与者也进行了继续预训练。第三个子图（左下）同样显示，虽然大多数参与者只使用了任务提供的已标注数据，但许多参与者也使用了我们入门套件中的领域内未标注数据。最后一个子图（右下）显示，大多数参与者没有使用附加特征。

（~15%），或采用基于提示的学习（~5%）。约 40% 的参与者使用了入门套件中的未标注数据。参与者有时也会使用任务之外的资源，包括已标注数据（~10%）和未标注数据（~5%）。参与者采用了多种创新方法，包括数据增强、主动学习或数据制图（Swayamdipta 等人，2020年）。

附加特征 大多数参与者没有使用附加特征（~66%）。如果使用了附加特征，主要是联合使用附加嵌入（~18%）、表情符号（~12%）或词典（~12%）。

5.4 最有效的方法

不同方法下参与者 F1 分数的分布，在任务 A 中比任务 B 和任务 C 更为集中（见图2）。在各项方法变量中，**架构**选择导致性能分布的差异最大；基于 Transformer 的模型和其他深度语言模型，在各项任务中的平均 F1 分数比传统机器学习系统更具竞争力。相比之下，**训练类型**似乎对 F1 分数分布的影响最小，各条件之间的重叠最多，即使在任务 C 中也是如此。密切相关的**额外数据**条件对性能的影响稍大一些，使用未标注数据

入门套件中的数据带来了最高的平均分。所有任务中，未使用任何**附加特征**的系统，其平均 F1 分数都更高。

6 错误分析

我们基于每项任务中表现最佳的 10 个系统进行错误分析，因为它们的错误反映了数据集中尚存的难点。

6.1 定量错误分析

二元标签混淆 在任务 A 的 4,000 个测试样本中，所有排名前 10 的系统都将其中 162 个分类错误；这些错误中，38.3% 为假阳性（将非性别歧视内容预测为性别歧视内容），61.7% 为假阴性（将性别歧视内容预测为非性别歧视）。在假阴性中，41% 属于敌意 类别，34% 属于贬损类别，16% 属于偏见性讨论，9% 属于威胁类别。这种模式符合预期，因为敌意 通常包含隐晦语言，而威胁通常以明确的语言表达。

类别与向量混淆在各类别中，威胁最不容易被误分类为其他类别，同时对于真实标签为其他任何类别的实例而言，也最少被误判为该类别（见图3）。敌意和贬损—

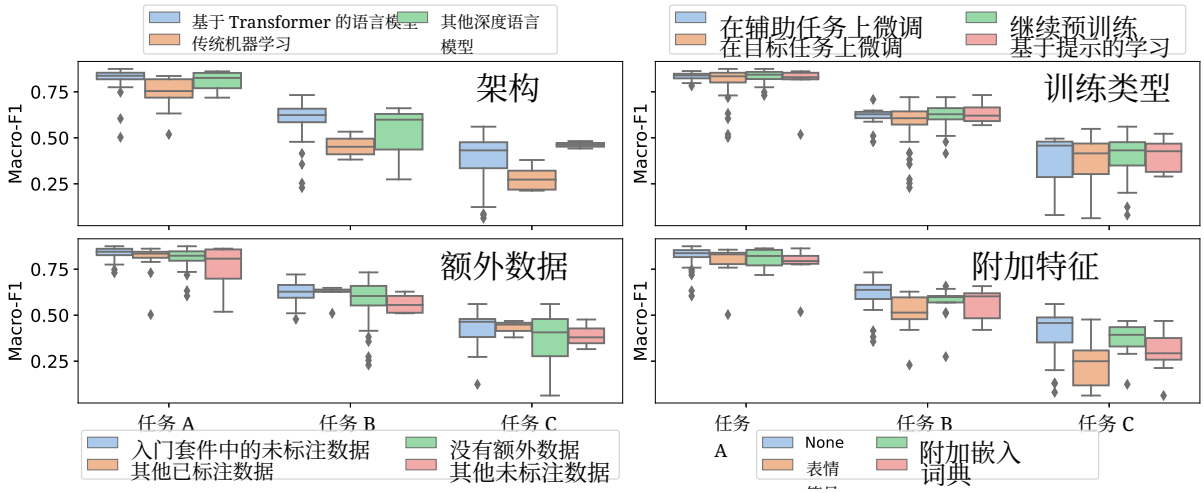


图 2：参与者方法的 Macro-F1 分数（统计单独或组合使用该方法的提交）。每个子图展示系统论文中方法选择的不同细分情况，以及各任务的平均 F1 分数。任务 A 的 F1 分数通常较高，且分布比任务 B 和任务 C 更集中。

损经常被彼此混淆。属于偏见性讨论的实例也常被误分为贬损或敌意。任务 C 中向量之间的混淆模式也类似（附录F）。部分错误反映了类别不平衡：最大类别（2.1）和最小类别（3.4）分别是最常见和最少见的混淆对象。其他错误则揭示了某些向量之间固有的相似性：2.1 和 3.2 经常被混淆，两者都指性别刻板印象，但 2.1 的情感倾向明确为负面。2.2 和 3.1 也有同样的情况，两者都包含性别侮辱语，但 3.1 属于因果性（而非针对特定对象的）用法。

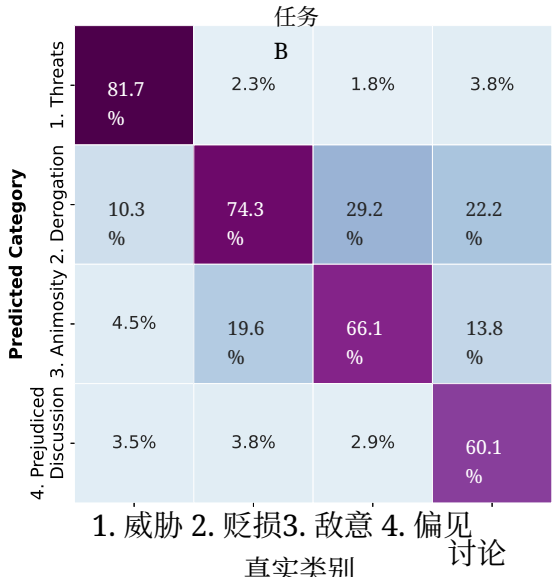


图 3：排名前十的系统在性别歧视类别（任务 B）上的混淆矩阵。

6.2 人工错误检查

两位自我认同为女性的作者人工检查了 100 条由所有排名前 10 的团队错误分类的任务 A 测试集实例。在这 100 条中，我们认定 76 条属于系统真正的分类错误，另外 24 条属于标注错误。[14]

系统错误 在 76 条系统错误中，52 条是假阴性（68%），24 条是假阳性（32%）。在**假阴性**中，错误分类的示例有几个常见主题。这些主题包括：(i) 情感倾向看似积极的表达，或带有暗讽意味的性别化恭维（例如，“*STEM* 一点也不难，女人也能做到！”）；(ii) 笑话中蕴含的性别刻板印象（例如，“我刚买了一辆智能汽车。只要妻子坐在驾驶座上，它就发动不了。”）；(iii) 俚语用法（例如，“那些兄弟想要大屁股白人女孩。”）或较少见的性别化侮辱（如“悍妇”或“女巫”）；以及 (iv) 反讽（例如，“你要么崇拜女性，要么就是厌女者。”）。一些露骨的例子还结合了疑问句等修辞手法，也被误判为非性别歧视（例如，“一个人要怎么学会掐女性的脖子？我想这其中有门道吧？是因为掐脖子会阻断大脑供血，而不只是让人无法呼吸吗？”）。

在这些**假阳性**中，也存在一些常见模式，包括：(i) 反击性言论（例如，“女性竟然连保持中立的资格都没有，这真是太他妈离谱了

[14]请注意，与整个数据集相比，这个子集中的标注错误率可能更高，因为其中包含自动化系统和人类都难以分类的棘手示例。

总得装模作样地逗男性开心，还得防着他们攻击你。”）；（ii）对与性别歧视相关的群体进行客观讨论或描述（例如，“男人走自己的路的核心理念就是，男性走自己的路……这并不意味着我们从渴望女性转而渴望男性了。只是说，我们不允许女性决定什么才算男人，或让女性成为我们生活的中心。”）；或（iii）批评某位女性个人的行为（例如，“我认识一个女性，她和她一个最好的朋友的前男友睡了。为了一个混蛋，她抛弃了整个社交圈，那人甚至都不是查德。她也算是个8分美女。真无语。她有选择，却选了这家伙。女性有时真够蠢的。”）。

标注错误 在这 24 个错误标注的样例中，17 个是假阴性（29%），7 个是假阳性（71%）。总体而言，标注员难以区分针对个人的陈述与泛化陈述。性别刻板印象以推测而非特征描述的形式表达时，会产生假阴性（例如：“我们现在有避孕套和避孕措施了，所以女性可以发生更多性行为，后果也远没那么严重。她们和顶尖男性发生更多性行为也合情合理，毕竟风险相对较小。”）。针对女性个人目标的非性别攻击则会产生假阳性（例如：“这个女人就是个蠢货。”）。

缺乏语境的灰色地带 尽管作者对每个经过人工检查的实例都作出了判定，但24个错误标注的实例中，有9个被认为可能属于灰色地带——这些实例可能因语境或对指南的解读略有不同而归入不同标签。单独来看，帖子和评论有时缺乏足够语境，难以解读其含义或意图。此时，标注者可能会根据自身对网络性别歧视的经验或对相关社群的了解来补充语境。在针对个别女性的批评中，这一点尤为明显：如果没有语境，往往很难判断说话者是在批评某人具体的行为，还是将其当作替罪羊，以表达普遍的性别歧视刻板印象和观点。这也引出了一个问题：如果某些内容暗示说话者有性别歧视倾向，但本身并未传达具体的性别歧视内容，是否也应将其标记出来。

7 讨论

7.1 经验教训

多样化数据采样的价值 我们注意到，先前 SemEval 关于仇恨或辱骂检测的任务主要报告参赛者产生的假阳性错误。例如，在 HatEval（Basile 等人，2019年）中，排名前三的系统所犯错误中有89.1%是假阳性。而在我们的分析中，假阴性反而更多。我们认为，错误类型分布出现这种反转，是由于我们的数据集中存在细微且多样的性别歧视内容（例如偏见性讨论），这些内容不一定与更明显的性别歧视内容具有相同的语言形式或关键词，因此更难被正确识别。

细粒度预测面临的挑战 我们的基线和参与者得分表明，二元性别歧视检测（任务 A）比细粒度性别歧视检测（任务 B 和 C）容易得多。任务 C 中取得的最高 F1 分数仅略高于 50%。这表明，在对具有社会条件影响的概念（如性别歧视、仇恨或辱虐）进行二元检测的任务中，报告的高分掩盖了模型在区分不同内容形式方面的细粒度缺陷；这些内容形式对其目标对象及整个社会可能造成截然不同的影响。鉴于预测我们的性别歧视类别和向量类别仍有很大的提升空间，我们期待未来以模型为中心、使用我们数据集的研究能持续提升性能。未来的工作也可以采用以数据为中心的技术来提升性能和鲁棒性，例如用更具挑战性的对抗样本扩充数据集（Kiela 等人，2021年；Vidgen 等人，2021b；Kirk 等人，2022c），或使用数据制图方法描绘难以学习的子空间（Swayamdipta 等人，2020年）。

持续预训练的效率 为社会属性相关任务（如性别歧视）标注数据颇具挑战，成本高昂，且需要一定程度的领域专业知识（Kirk 等人，2022b）。然而，提供未标注的领域内数据既成本低廉、易于扩展，也能避免因标注负担加重而对标注员造成心理伤害（Kirk 等人，2022a）。令人欣慰的是，我们表现最佳的基线和成绩优异的参赛系统都有效利用了我们提供的未标注领域内数据（来自 Reddit 和 Gab）进行持续预训练。我们鼓励未来的 SemEval

任务也以类似方式发布未标注数据集，并从中抽取较小的已标注数据集。

7.2 局限性

类别不平衡 与许多以往用于仇恨与辱骂检测的数据集相比（Vidgen 和 Derczynski，2020年），我们的数据集在二元标签（24%为性别歧视）上的分布更加均衡。然而，性别歧视类别和向量类别之间仍存在显著的不平衡。因此，很难确定类别和向量类别之间的混淆究竟源于数据的固有特征，还是类别不平衡。我们鼓励未来的研究考察平衡数据集对性能及跨向量混淆的影响。尽管如此，我们明确决定不重新平衡数据集，因为不同类型的性别歧视（尤其是在细粒度层面）在现实中确实有着截然不同的基准率。解决类别不平衡问题，正是我们试图应对的问题中不可或缺的一部分。

界定"可解释" 我们将可解释性问题设定为针对单篇文本的分类任务（不包含网络或用户信息）。在这种设定下，结合层级分类即可提升可解释性，因为它能说明一段内容是否具有性别歧视，以及原因是什么（类别 和向量类别）。此外，还有许多有前景的方法可提升面向社会敏感任务（例如性别歧视、辱虐或仇恨）的自动化系统的可解释性，包括对片段或掩码词元进行分类，以预测理由（Mathewetal.，2021年；Kim etal.，2022年）；利用用户与网络数据（Qian等人，2018年；Wich etal.，2021年）或目标数据（Kennedy等人，2020年）；将社会偏见框架用于刻板印象（Sap et al.，2020年）；或鼓励多跳推理和思维链推理（Zhang 等人，2022a；Jinetal.，2022年）。

价值具体性与 Perspective 我们的数据集由居住在英国的标注员进行标注，他们均自我认同为女性。我们提供规范性指南，然后依靠多数投票和专家裁定得出金标准标签。许多先前研究表明，标注员身份是影响标注行为的关键因素（Waseem，2016年；Larimore 等人，2021年；Sap 等人，2022年），因此多数投票无法充分反映主观分歧（Davani 等人，2021年）。为促进今后对标注员特定偏差的研究，我们为每条已标注的数据提供标注员 ID（Prabhakaran 等人，2021年）。尽管如此，我们对性别歧视的定义及其在我们分类体系中的子类别，以及标注员如何运用最佳判断，都建立在有限的文化视角之上，主要来自以西方为中心的英语社群。因此，基于我们数据集构建的任何分类器都将承袭我们的价值观（作为分类体系和指南的编写者），以及标注员的价值观。[15] 与大多数有害内容数据集一样，我们的数据仅限于英语（Röttger 等人，2022年）。然而，性别歧视的界限在不同文化中存在争议，人们对于界限应划在哪里可能会有合理的分歧。

8 结论

本文主要作出了三项贡献。首先，我们提出了一套新的分类体系，用于在三个层级上实现更具可解释性的性别歧视分类——二元性别歧视检测、性别歧视类别以及性别歧视的细粒度向量。该分类体系建立在对既有分类体系和社会科学文献的综述之上，并通过来自两个社交媒体平台的领域内数据进行了实证验证。因此，它从社会技术视角概述了网络性别歧视多样且微妙的图景。其次，我们创建了一个高质量数据集：通过技术集成进行抽样，以提升内容的多样性；同时由自我认同为女性的专家进行标注，以确保标签的一致性。该已标注数据集还与一个规模更大的未标注数据集配套使用，从而缓解标注预算（包括财务成本和标注者心理成本）对训练后的系统有效性所造成的限制。最后，我们分享了基线模型，并总结了提交至 SemEval 任务的系统。该分析展示了将最先进的 Transformer模型与持续预训练和多任务学习相结合的技术所取得的成功。然而，在检测性别歧视内容的细粒度形式方面，性能仍有很大的提升空间。我们希望我们的研究和资源能够助力未来系统的设计，使网络空间对所有人都更加安全。

致谢

这项工作由 Rewire 开展，并由 MetaAI 资助。我们感谢各位标注员，也感谢 Rewire 团队为本文提供反馈

[15] Bang 等人(2022年) 致力于设计"价值观对齐"的性别歧视检测系统，其模型会同时考虑内容和个人的价值体系，并据此进行分类。这项技术有望进一步提升性别歧视检测任务的包容性和可解释性。

。我们还要感谢所有向我们的任务提交结果的参与者，以及对系统论文提出宝贵且建设性意见的审稿人。

伦理风险与危害声明

我们发布了一个包含网络性别歧视内容的数据集，并证明最先进模型在某些特定条目和向量类型上仍会出错。恶意行为者可能会将这些细粒度的失败案例作为灵感，发布能够绕过当前检测系统的线上性别歧视帖子，或者原则上利用这些条目来训练生成模型（Vidgen 等人（2021b）和Kirk 等人（2022c）也提出了同样的担忧）。鉴于我们的数据集有助于理解和预测网络性别歧视所带来的益处，我们认为这一风险是可控的。

遵循 Kirk 等人 (2022a）的建议，我们描述了数据集构建和发布所带来的风险。首先，数据主体（成为性别歧视条目攻击目标的女性）以及论文读者可能会在复现或强化有害表征、刻板印象、偏见性讨论或侮辱性词语使用时面临伤害风险。我们在首页加入内容警告，并始终以斜体显示引用示例（表 1，§6.2）。在可能的情况下，我们会用星号替换侮辱性词语中的元音。由于这项工作需要进行大量标注（2 万条），标注员可能因反复查看性别歧视内容而受到影响，尤其是暴力条目（例如“伤害威胁”）。我们谨慎遵循相关规程，以支持标注员福祉，并通过群聊论坛与他们保持直接沟通渠道畅通。标注工作结束后，我们会对标注员进行调查，了解这项任务对其福祉的影响，以及今后如何做得更好。总体而言，标注员认为我们为其提供了充分支持，维护了他们的心理健康，并确保整个过程尽可能安全。少数标注员建议我们举办研讨会，介绍研究成果。我们承诺在未来几个月内举办这场总结交流会，因为让标注员了解其工作对支持网络安全的重要意义十分重要。

参考文献

Ashley Amaya、Ruben Bach、Florian Keusch 和 Frauke Kreuter。2021年。社会科学研究中的新数据来源：使用 Reddit 数据前须知。《社会科学计算机评论》，39(5)：943–960。出版方：SAGE 出版公司。

Yejin Bang、Tiezheng Yu、Andrea Madotto、Zhaojiang Lin、Mona Diab 和 Pascale Fung。2022年。使分类器作出与人类价值观明确一致的判断。_预印本：2210.07652v1。

Valerio Basile、Cristina Bosco、Elisabetta Fersini、Debora Nozza、Viviana Patti、Francisco Manuel Rangel Pardo、Paolo Rosso 和 Manuela Sanguinetti。2019年。SemEval-2019 任务5：Twitter 上针对移民和女性的仇恨言论的多语言检测。载于第十三届国际语义评估研讨会论文集，第54–63页。

Emily M. Bender 和 Batya Friedman。2018年。自然语言处理的数据声明：迈向缓解系统偏见并推动更好的科学研究。计算语言学协会汇刊，6：587–604。

Abeba Birhane、William Isaac、Vinodkumar Prabhakaran、Mark Diaz、Madeleine Clare Elish、Iason Gabriel 和 Shakir Mohamed。2022年。权力属于人民？参与式人工智能的机遇与挑战。载于算法、机制与优化中的公平与可及性，EAAMO ’22，第 1–8 页，美国纽约州纽约。计算机协会。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.

Tommaso Caselli、Valerio Basile、Jelena Mitrovi´c 和 Michael Granitzer。2021年。HateBERT：重新训练 BERT，以检测英语中的辱虐性语言。载于第五届在线辱虐与伤害研讨会论文集（*WOAH 2021*），第 17–25 页，线上。计算语言学协会。

AakankshaChowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vin- odkumar Prabhakaran、Emily Reif、Nan Du、Ben

Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghe- mawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fe- dus, Denny Zhou, Daphne Ippolito, David Luan, HyeontaekLim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankara- narayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Bren- nan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022年。PaLM：通过 Pathways 扩展语言建模。 arXiv:2204。02311 [cs]。

Matteo Cinelli、Gianmarco De Francisci Morales、Alessandro Galeazzi、Walter Quattrociocchi 和 Michele Starnini。2021年。社交媒体上的回音室效应。《美国国家科学院院刊》，118(9)。出版方：美国国家科学院。

Kevin Clark、Minh-Thang Luong、Quoc V. Le 和 ChristopherD. Manning。2020年。ELECTRA：将文本编码器预训练为判别器而非生成器。载于国际学习表征会议。

Aida Mostafazadeh Davani、Mark Díaz 和 Vinod- kumar Prabhakaran。2021年。处理分歧：超越主观标注中的多数投票。arXiv:2110.05719 [cs]。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深层双向Transformer预训练。载于计算语言学协会北美分会人类语言技术*2019*年会议论文集，第*1*卷（长短论文），第4171–4186页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Mai ElSherief、Elizabeth Belding 和 Dana Nguyen。2017年。#NotOkay：理解社交媒体中的性别暴力。*AAAI*国际网络与社交媒体会议论文集，11(1):52–61。期号：1。

Tracie Farrell、Miriam Fernandez、Jakub Novotny 和 Harith Alani。2019年。探索 Reddit 男性圈中的厌女症。载于第十届 *ACM*网络科学会议论文集，第 87–96 页。计算机协会，纽约州纽约，美国。

E Fersini、P Rosso 和 M Anzovino。2018a。IberEval 2018 自动厌女识别任务概述。载于伊比利亚语种人类语言技术评估第三届研讨会论文集（*IberEval 2018*），第 214–228 页。

Elisabetta Fersini、Debora Nozza 和 Paolo Rosso。2018b。EVALITA 2018 自动识别厌女症任务（AMI）概述。载于《EVALITA：意大利语自然语言处理与语音工具评测》，第 59–66 页。Accademia University Press。Debbie Ging。2017年。《阿尔法男性、贝塔男性与非自愿独身者：男性圈男子气概的理论化》。《男性与男子气概》，1:20。Debbie Ging、Theodore Lynn 和 Pierangelo Rosati。2020年。厌女主义新词化：Urban Dictionary 中关于性虐待的大众分类法。《新媒体与社会》，22(5):838–856。出版方：SAGE Publications。Barney Glaser 和 Anselm Strauss。2017年。《扎根理论的发现：质性研究策略》。Routledge，纽约。Ella Guest、Bertie Vidgen、Alexandros Mittos、Nishanth Sastry、Gareth Tyson 和 Helen Margetts。2021年。用于检测网络厌女的专家标注数据集。载于《第十六届欧洲计算语言学协会会议论文集：主会场》，第 1336–1350 页，线上。计算语言学协会。Pengcheng He、Jianfeng Gao 和 Weizhu Chen。2021a。DeBERTaV3：通过梯度解耦嵌入共享进行 ELECTRA 风格预训练，改进 DeBERTa。arXiv:2111.09543 [cs]。Pengcheng He、Xiaodong Liu、Jianfeng Gao 和 Weizhu Chen。2021b。DeBERTa：具有解耦注意力的增强型 BERT 解码器。arXiv:2006.03654 [cs]。Greta Jasser、Jordan McSwiney、Ed Pertwee 和 Savvas Zannettou。2021年。“欢迎来到 #GabFam”：Gab 上的极右翼虚拟社群。《新媒体与社会》，第 146144482110245 页。Akshita Jha 和 Radhika Mamidi。2017年。赞美何时会变成性别歧视？使用 Twitter 数据分析并分类矛盾性别歧视。载于《第二届自然语言处理与计算社会科学研讨会论文集》，第 7–16 页，温哥华（加拿大）。计算语言学协会。Zhijing Jin、Sydney Levine、Fernando Gonzalez Adauto、Ojasv Kamal、Maarten Sap、Mrinmaya Sachan、Rada Mihalcea、Joshua B Tenenbaum 和 Bernhard Schölkopf。2022年。何时应当例外：探索语言模型如何解释人类道德判断。载于《神经信息处理系统进展》。Callum Jones、Verity Trott 和 Scott Wright。2020年。《荡妇与娘炮：男人走自己的路运动与厌女式网络骚扰的产生》。《新媒体与社会》，22(10):1903–1921。出版方：SAGE Publications，英国伦敦 Sage UK。

Brendan Kennedy、Mohammad Atari、Aida Mostafazadeh Davani、Leigh Yeh、Ali Omrani、Yehsong Kim、Kris Coombs、Shreya Havaldar、Gwenyth Portillo-Wightman、Elaine Gonzalez、Joseph Hoover、Aida Azatian、Alyzeh Hussain、Austin Lara、Gabriel Olmos、Adam Omary、Christina Park、Clarisa Wijaya、Xin Wang、Yong Zhang 和 Morteza Dehghani。2018年。Gab 仇恨语料库：一个包含 2.7 万条仇恨言论标注帖文的数据集。*PsyArXiv*。出版方：PsyArXiv。

Brendan Kennedy、Xisen Jin、Aida Mostafazadeh Da- vani、MortezaDehghani 和 Xiang Ren。2020年。利用事后解释对仇恨言论分类器进行情境化。载于计算语言学协会第*58*届年会论文集，第5435–5442页，线上。计算语言学协会。

Douwe Kiela、Max Bartolo、Yixin Nie、Divyansh Kaushik、Atticus Geiger、ZhengxuanWu、Bertie Vid- gen、Grusha Prasad、Amanpreet Singh、Pratik Ring- shia 等。2021年。Dynabench：重新思考自然语言处理中的基准测试。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第4110–4124页。

Jiyun Kim、Byounghan Lee 和 Kyung-Ah Sohn。2022年。为什么这是仇恨言论？用于可解释仇恨言论检测的掩码比率预测。载于第*29*届国际计算语言学会议论文集，第6644–6655页，韩国大韩民国庆州。国际计算语言学委员会。

Hannah Kirk、Abeba Birhane、Bertie Vidgen 和 Leon Derczynski。2022a。在自然语言处理研究中处理和呈现有害文本。载于计算语言学协会：*EMNLP 2022*论文集，第497–510页，阿联酋阿布扎比。计算语言学协会。

Hannah Kirk、Bertie Vidgen 和 Scott Hale。2022b。更多数据是否更好？重新思考基于Transformer的主动学习在辱虐性语言检测中效率的重要性。载于第三届威胁、攻击与网络欺凌研讨会（*TRAC 2022*）论文集，第52–61页，韩国大韩民国庆州。计算语言学协会。

Hannah Kirk、Bertie Vidgen、Paul Rottger、Tristan Thrush 和 Scott Hale。2022c。Hatemoji：用于基准测试和检测基于表情符号的仇恨言论的测试套件和对抗生成数据集。载于*2022*年北美计算语言学协会分会会议论文集：人类语言技术，第1352–1368页，美国西雅图。计算语言学协会。

Savannah Larimore、Ian Kennedy、Breon Haskett 和 Alina Arseniev-Koehler。2021 年。重新审视标注者对种族主义语言的分歧：噪声还是信号？载于第九届社交媒体自然语言处理国际研讨会论文集，第 81–90 页，线上。计算语言学协会。

Mary Lilly。2016年。世界并非男性的安全之地：男性圈的再现政治。渥太华大学。备份出版者：UniversityofOttawa。

Lucas Lima、Julio C.S. Reis、Philipe Melo、Fabricio Murai、Leandro Araujo、Pantelis Vikatos 和 Fabricio Benevenuto。2018年。右翼回声室内部：刻画 Gab 这一不受监管的社交系统。载于*2018*年 *IEEE/ACM* 国际社交网络分析与挖掘进展会议（*ASONAM*），第515–522页。ISSN：2473-991X。

Xiaodong Liu、Yu Wang、Jianshu Ji、Hao Cheng、Xueyun Zhu、Emmanuel Awa、Pengcheng He、Weizhu Chen、Hoifung Poon、Guihong Cao 和 Jianfeng Gao。2020年。用于自然语言理解的微软多任务深度神经网络工具包。载于计算语言学协会第*58*届年会论文集：系统演示，第118–126页，线上。计算语言学协会。

Yinhan Liu、Myle Ott、NamanGoyal、JingfeiDu、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。arXiv：1907。11692 [cs]。

Binny Mathew、Punyajoy Saha、Seid Muhie Yi- mam、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。人工智能促进协会会议论文集，35(17):14867–14875。编号：17。

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

Pulkit Parikh、Harika Abburi、Niyati Chhaya、Manish Gupta 和 Vasudeva Varma。2021年。通过神经网络方法对性别歧视和厌女症进行分类。*ACM*网络汇刊，15(4)：17:1–17:31。

John Pavlopoulos、Jeffrey Sorensen、Léo Laugier 和 Ion Androutsopoulos。2021年。SemEval-2021任务5：有毒文本片段检测。收录于《第*15*届语义评测国际研讨会论文集（*SemEval-2021*）》，第59–69页。

Vinodkumar Prabhakaran、Aida Mostafazadeh Davani 和 Mark Díaz。2021年。关于在数据集中发布标注者级标签和信息。arXiv：2110。05699 [cs]。

Jing Qian、Anna Bethke、Yinyin Liu、Elizabeth Belding 和 William Yang Wang。2019年。用于学习如何干预在线仇恨言论的基准数据集。发表于*2019*年自然语言处理中的经验方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第4755–4764页。计算语言学协会。会议地点：中国香港。

Jing Qian、Mai ElSherief、Elizabeth Belding 和 William Yang Wang。2018年。利用用户内和用户间表示学习进行自动化仇恨言论检测。载于《*2018*年北美计算语言学协会分会会议论文集：人类语言技术》，第*2*卷（短篇论文），第118–123页，路易斯安那州新奥尔良。计算语言学协会。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。2019年。语言模型是无监督多任务学习器。第24页。

Alison Ribeiro 和 Nádia Silva。2019年。SemEval-2019 任务5中的 INF-HatEval：针对 Twitter 上女性和移民的仇恨言论检测卷积神经网络。载于《第*13*届语义评估国际研讨会论文集，第420–425页。

Manoel Horta Ribeiro、Jeremy Blackburn、Barry Brad- lyn、Emiliano De Cristofaro、Gianluca Stringhini、Summer Long、Stephanie Greenberg 和 Savvas Zannettou。2021年。男性圈在网络中的演变。*AAAI* 国际网络与社交媒体会议论文集，15:196–207。

Francisco Rodríguez-Sánchez、Jorge Carrillo-de Al- bornoz、Laura Plaza、Julio Gonzalo、Paolo Rosso、Miriam Comet 和 Trinidad Donoso。2021年。EXIST 2021 概述：社交网络中的性别歧视识别。自然语言处理，67(0):195–207。

Paul Röttger、Debora Nozza、Federico Bianchi 和 Dirk Hovy。2022年。将仇恨言论检测拓展至资源匮乏语言的数据高效策略。载于*2022*年自然语言处理实证方法会议论文集，第5674–5691页，阿联酋阿布扎比。计算语言学协会。

Paul Röttger、Bertie Vidgen、Dirk Hovy 和 Janet Pierrehumbert。2022年。主观性自然语言处理任务的两种对比鲜明的数据标注范式。载于*2022*年计算语言学协会北美分会人类语言技术会议论文集，第175–190页，美国西雅图。计算语言学协会。

Mattia Samory、Indira Sen、Julian Kohne、Fabian Flöck 和 Claudia Wagner。2021年。“叫我性别歧视者也无妨，但……”：运用心理量表和对抗样本重新审视性别歧视检测。*AAAI*国际网络与社交媒体会议论文集，15:573–584。Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。2019年。DistilBERT：BERT的精简版本，更小、更快、更经济、更轻量。*arXiv*预印本 *arXiv:1910.01108*。Maarten Sap、Saadia Gabriel、Lianhui Qin、Dan Jurafsky、Noah A. Smith 和 Yejin Choi。2020年。社会偏见框架：推理语言的社会与权力影响。arXiv:1911.03891 [cs]。

Maarten Sap、Swabha Swayamdipta、Laura Vianna、Xuhui Zhou、Yejin Choi 和 Noah A. Smith。2022年。具有态度的标注者：标注者的信念与身份如何导致有毒语言检测偏差。arXiv:2111。07997 [cs]。

Swabha Swayamdipta、Roy Schwartz、Nicholas Lourie、Yizhong Wang、Hannaneh Hajishirzi、Noah A. Smith 和 Yejin Choi。2020年。数据集制图：利用训练动态绘制并诊断数据集。收录于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 9275–9293 页，线上。计算语言学协会。

Jane M Ussher。2016年。厌女症。*Wiley Blackwell* 性别与性研究百科全书，第 1–3 页。出版方：Wiley Online Library。

Bertie Vidgen 和 Leon Derczynski。2019年。辱虐性语言训练数据的方向：系统综述——垃圾进，垃圾出。*PLOSONE*，15(12):e0243300。出版商：公共科学图书馆 arXiv: 2004.01670。

Bertie Vidgen、Alex Harris、Dong Nguyen、Rebekah Tromble、Scott Hale 和 Helen Margetts。2019年。辱虐内容检测的挑战与前沿。载于第三届线上辱虐性语言研讨会论文集，第 80–93 页，意大利佛罗伦萨。计算语言学协会。

Bertie Vidgen、Dong Nguyen、Helen Margetts、Patricia Rossini 和 Rebekah Tromble。2021a。介绍 CAD：语境辱虐数据集。载于*2021*年北美计算语言学协会分会会议：人类语言技术论文集，第 2289–2303 页，线上。计算语言学协会。

Bertie Vidgen、Tristan Thrush、Zeerak Waseem 和 Douwe Kiela。2021b。从最糟糕的案例中学习：动态生成数据集以改进线上仇恨检测。载于第*59*届计算语言学协会年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇

论文），第1667–1682页，线上。计算语言学协会。

Zeerak Waseem。2016年。你是种族主义者，还是我看错了？标注者对 Twitter 仇恨言论检测的影响。载于首届自然语言处理与计算社会科学研讨会论文集，第 138–142 页，得克萨斯州奥斯汀。计算语言学协会。

Maximilian Wich、Edoardo Mosca、Adrian Gorniak、Johannes Hingerl 和 Georg Groh。2021年。利用用户与网络数据实现可解释的辱骂性语言分类。载于数据库中的机器学习与知识发现：应用数据科学专题，计算机科学讲义，第 481–496 页，Cham。施普林格国际出版公司。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformer：自然语言处理领域的先进技术。载于*2020*年自然语言处理实证方法会议论文集：系统演示，第 38–45 页，线上。计算语言学协会。

Wenjie Yin 和 Arkaitz Zubiaga。2022年。隐藏在显而易见之下：社交媒体上的误导性关键词与隐性辱骂性语言。在线社交网络与媒体，30：100210。

Tianhe Yu、SaurabhKumar、Abhishek Gupta、Sergey Levine、Karol Hausman 和 Chelsea Finn。2020年。多任务学习的梯度手术。载于神经信息处理系统进展，第33卷，第5824–5836页。Curran Associates, Inc.。

Marcos Zampieri、Shervin Malmasi、Preslav Nakov、Sara Rosenthal、Noura Farra 和 Ritesh Kumar。2019年。SemEval-2019任务6：社交媒体中的冒犯性语言识别与分类（OffensEval）。载于第*13*届语义评估国际研讨会论文集，第75–86页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Savvas Zannettou、Barry Bradlyn、Emiliano De Cristo- faro、Haewoon Kwak、Michael Sirivianos、Gianluca Stringini 和 Jeremy Blackburn。2018年。Gab 是什么：言论自由的堡垒还是另类右翼回音室。收录于*2018*年网络会议论文集，WWW '18，第 1007–1014 页，瑞士日内瓦共和国。国际万维网会议指导委员会。

Philine Zeinert、Nanna Inie 和 Leon Derczynski。2021年。 在线厌女现象标注。载于计算语言学协会第*59*届年会论文集及第*11*届自然语言处理国际联合会议（第*1*卷：长篇论文），第3181–3197页，线上。计算语言学协会。

Qiang Zhang、Jason Naradowsky 和 Yusuke Miyao。2022a。重新思考将冒犯性文本检测视为多跳推理问题。收录于计算语言学协会 *ACL 2022* 年会论文集，第 3888–3905 页，爱尔兰都柏林。计算语言学协会。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、MoyaChen、Shuohui Chen、ChristopherDewan、Mona Diab、Xian Li、Xi Victoria Lin、Todor Mihaylov、Myle Ott、Sam Shleifer、Kurt Shuster、Daniel Simig、Punit Singh Koura、Anjali Sridhar、Tianlu Wang 和 Luke Zettlemoyer。2022b。OPT：开放式预训练变换器语言模型。arXiv：2205。01068 [cs]。

Xinyang Zhang、Yury Malkov、Omar Florez、Serim Park、Brian McWilliams、Jiawei Han 和 Ahmed El-Kishky。2022c。TwHIN-BERT：用于多语言推文表示的社交信息增强的预训练语言模型。arXiv:2209。07562 [cs]。

Donna Zuckerberg。2018年。并非所有白人男性都已逝去：数字时代的古典学与厌女症。哈佛大学出版社，剑桥，马萨诸塞州。

Douglas Zytko、Pamela J. Wisniewski、Shion Guha、Eric P. S. Baumer 和 Min Kyung Lee。2022年。人工智能系统的参与式设计：面向多样化用户、关系和应用领域的机遇与挑战。载于*2022*年计算机系统人为因素*CHI*会议扩展摘要，CHI EA '22，第 1–4 页，纽约州纽约。计算机协会。

数据说明

我们提供了一份数据说明（Benderand Friedman，2018年），用于记录 EDOS 数据集的生成过程和来源。

A.1 筛选理由

EDOS 数据集旨在训练和评估自动化系统，以细粒度、可解释的方式检测网络性别歧视。我们整理了各种形式的性别歧视内容的社交媒体评论，涵盖从公然威胁行为（例如“我会杀了任何跟我说话的女人”）到更隐晦的形式（例如“女性是娇弱的花朵，需要被珍爱”）。

A.2 语言变体

所有条目均为英语，这是由研究人员和标注员的专业能力所决定的。我们鼓励未来的工作将我们的向量类别扩展到其他语言。

A.3 说话者人口统计特征

所有条目均采集自两个社交媒体平台：Gab 和 Reddit。因此，说话者的人口统计特征大致反映了这两个平台用户的总体情况（Amaya 等人，2021年）。我们预计（并有意抽样）用户很可能是男性、西方人且倾向右翼，并且对女性、性别议题和女权主义持极端或极右翼观点（Zannettou 等人，2018年；Lima 等人，2018年）。

A.4 标注员人口统计特征

我们从现有的自由职业标注员网络中招募了19名标注员。他们在2022年春季期间工作了两个月。标注员的薪酬为每小时16英镑，专家标注员的薪酬为每小时20英镑。所有标注员均通过了包含200条条目的金标准集筛选，且此前均有参与仇恨言论标注任务的经验。标注员可随时通过即时通讯平台联系研究团队。所有标注员均自我认同为女性。我们向标注员发送了一份可自愿填写的调查问卷，以收集更多关于其人口统计特征的信息，并了解她们过去在社交媒体和网络性别歧视方面的经历。19名标注员中有12人回应了调查问卷。年龄方面，3人年龄在18至24岁之间，8人年龄在25至30岁之间，1人年龄在31至40岁之间。已完成的最高学历方面，1名标注员为高中学历，6名为本科学历，5名为研究生学历。标注员来自不同国籍，其中7人为英国人，此外，还有瑞典人、瑞士人、意大利人和阿根廷人。大多数标注员认同自己是白人；其中一人认同自己是南亚裔与白人的混血，另一人认同自己是英国黑人。大多数人认同自己是异性恋（7人），其余人认同自己是双性恋、泛性恋或无性恋。8名标注员是英语母语者，另有4人虽非母语者，但英语流利。大多数标注员每天使用社交媒体一小时或更久（7人）。所有标注员都见过他人遭受网络性别歧视，其中7人曾亲身成为攻击目标。

我们还收集了标注员对网络性别歧视的特定态度，例如：(i) 他们认为社交媒体公司在内容审核政策中应优先处理哪些类型的网络性别歧视；以及 (ii) 内容是否应被删除或部分隐藏。我们询问标注员，对于归属于我们的 11 个向量类别的内容，他们在内容审核方面有何偏好。在未来的工作中，我们计划将这些偏好映射到数据集条目，以调查这些态度是否影响了标注行为。

A.5 言语情境

条目的模态是来自社交媒体的短篇书面文本评论，标注时将其孤立地进行解读，即不置于评论串中，也不提供用户或网络信息。

A.6 文本特征

文本的体裁为性别歧视性和非性别歧视性社交媒体评论。最终数据集的构成见表2。

B 现有分类体系综述

我们回顾了近期提出性别歧视或厌女内容分类体系的文章，具体包括：Jha和Mamidi (2017年)；Samory 等人(2021年)；Farrell 等人 (2019年)；Zeinert 等人(2021年)；Guest等人 (2021年)；Rodríguez-Sánchez 等人 (2021年)；以及 Parikh 等人 (2021年)。我们在此介绍既有研究与我们分类体系（见 §2.2）之间的若干关键差异：

构建方式的差异 我们回顾的分类体系要么以理论为基础，要么以实证为基础。以实证为导向的研究通常先依据自然语言处理文献构建分类体系的初始版本，然后利用新收集的数据反复调整分类体系（Zeinert 等人，2021年；Guest等人，2021）。Twitter 是最

最常用的数据来源（Jha和Mamidi，2017年；Fersini 等人，2018a；Samory 等人，2021年；Zeinert 等人，2021年；Rodríguez-Sánchez 等人，2021年），其次是 Reddit（Farrell 等，2019年；Guest 等，2021年）。数据收集大多基于关键词，例如侮辱性词语和性别歧视主义标签；例外的是 Farrell 等人（2019年），他们从“男性圈”社区中进行了更广泛的抽样；以及 Parikh等人（2021年），他们的研究基于“日常性别歧视项目”——一个收录了人们遭受性别歧视经历的案例目录。一些研究还对现有数据集进行了重新标注（Jha和Mamidi，2017年；Samory 等人，2021年；Guest等人，2021年），并加入了对抗性数据（Samory 等人，2021年）。我们的分类体系以理论和实证为基础——我们首先借鉴先前基于理论的分类体系，然后结合领域内数据反复迭代。

范围差异 现有分类体系在范围上各不相同，部分体现在它们关注的是性别歧视（Samory 等人，2021；Rodríguez- Sánchez 等人，2021；Parikh 等人，2021），还是厌女症（Jha和Mamidi，2017；Fersini 等人，2018a；Farrell 等，2019；Guest等人，2021；Zeinert等人，2021）。这两个术语的定义显然并不一致。例如，Zeinert等人（2021）正式将厌女症定义为一种针对群体的辱骂和仇恨言论。实际上，他们对厌女症的描述与我们采用的较宽泛的性别歧视定义十分相似。我们的分类体系明确将性别歧视视为一种辱虐，其范围比厌女症更广。现有分类体系往往涵盖彼此重叠但又各不相同的主题。它们对更细粒度的性别歧视和厌女症类型（如果有的话）的命名和定义也各不相同，我们将这些类型称为向量类别。例如，Jha和Mamidi（2017）运用社会心理学中关于善意性别歧视的一种特定理论，提出了三种主要的性别歧视类型：家长式作风、性别差异化和异性恋。相比之下，Parikh等人（2021）基于性别歧视目标群体的视角，构建了23种性别歧视向量。他们区分了角色刻板印象和属性刻板印象，以及与工作、月经和母职相关的性别歧视等。部分研究为向量使用描述性名称（例如“身体暴力”（Farrell 等，2019；Guest等人，2021）），另一些研究则采用基于社会科学理论或术语的名称（例如“煤气灯操纵”（Parikh等人，2021）、“贬低”（Farrell 等，2019）和“新性别歧视”（Zeinert等人，2021））。我们的分类体系始终使用描述性名称，并且在每个层级上都保持大致相当的细粒度程度。

结构差异 最后，现有分类体系在组织性别歧视和厌女内容的向量类别时也各不相同。最常见的做法是将所有向量类别都设为厌女症或性别歧视的直接子类型（Fersini 等人，2018a；Zeinert 等人，2021年；Farrell 等人，2019年；Rodríguez-Sánchez 等人，2021年；Parikh 等人，2021年），但并未明确说明它们之间的关系。Jha和Mamidi（2017年）以及Samory 等人（2021年）增设了标记，用于区分善意/文明表达与敌意/不文明表达；而Fersini 等人（2018a）则设置了针对目标对象的标记。Guest等人（2021年）提出了一个多层级结构，涵盖厌女症的类型和向量类别。我们的分类体系同样采用层级式方法，先区分性别歧视类别，再区分各类别中的细粒度性别歧视向量。

C Reddit 数据补充详情

通过审阅七篇研究论文和三份成熟的线上清单，我们确定了一份包含 81 个相关子版块的长名单。其中一些子版块已被封禁，但仍有可用数据，因此也列入了这份长名单。随后，我们依据 Lilly （2016年）的研究，将这 81 个子版块划分为四个类别。Lilly 的研究也被 Ribeiro 等人 （2021年）采用，他们在此基础上，将「男性权利活动人士」（MRA）拆分为男性权利活动人士和「红药丸」（TRP），据此建立了由五部分组成的男性圈子版块分类体系。我们不采用这种区分，因为从概念上讲，这两个类别非常相似，将它们分开可能会导致我们的数据收集产生偏差，实际上过度采样这些相似的社群。为使数据更加多样化，我们纳入了在主题上与以下四个类别相关的子版块（见下文），即使这些子版块名义上由异性恋男性以外的群体主导（例如 `r/gaycel` 或 `r/trufemcels`），或主要关注批评和讨论男性圈（例如 `r/purplepilldebates`或 `r/thebluepill`）。

- **非自愿独身者（非自愿独身者）**反对女性，并贬低、侮辱或攻击女性的男性，因为他们无法获得，或认为自己无法获得女性的性关注或陪伴。许多这样的男性认为自己理应得到女性的关注。

•**男人走自己的路（MGTOW）**认为女性和女权主义腐蚀了社会、男性需要重新确立自身地位的男性。与非自愿独身者相比，这类男性中有许多人能够或认为自己能够获得女性的陪伴和性关注，对女性的怨恨和仇恨也不那么明显。

• **男性权利活动家（MRA）**通常反对女权主义，认为女性享有特权，而男性受到系统性歧视。男性权利活动群体的立场各不相同，从希望离婚父亲获得更多权利等较温和的诉求，到希望男性享有国家保障特权等较极端的诉求。

• **搭讪艺术家（PUA）**男性会主动尝试赢得女性的陪伴和性兴趣，往往使用欺骗或不光彩的手段。许多这类男性对女性持有贬损看法，并将她们描绘成性对象。

D 定制性别歧视分类器的数据采样

在我们的抽样集成中，我们使用来自七个开源数据集的英语女性目标条目，训练了一个定制的二元分类器（Zampieri等人，2019年；Rodríguez-Sánchez 等人，2021年；Vid- gen 等人，2021b；Samory 等人，2021年；Guest等人，2021年；Fersini 等人，2018a，b）。对于这些数据集，我们会删除重复项、清理空白字符，并将网址、表情符号和用户名替换为特定标记（例如，[URL]）。每个数据集都依据其自身针对性别歧视和/或厌女症的分类体系进行标注，而各数据集的结构和定义并不一致。为统一标签，我们采用每个数据集中的二元顶层标签（厌女症与非厌女症，或性别歧视与非性别歧视）。我们将合并后的数据集划分为训练集和测试集（90/10），并按数据集来源分层。我们使用 transformers库（Wolf 等人，2020年），以默认参数训练 BERT-base 模型 3 个轮次。该模型在留出的测试集上取得了 80% 的 Macro-F1 分数。我们使用该模型对 Gab 和 Reddit 数据中厌女或性别歧视内容的存在与否进行软标注。

E 众包工作者与专家实验

我们使用包含 200 条条目的金标准，测试了两种标注方案。首先，我们将金标准提交给众包标注服务，每条目获得 7 个标注，共有 176 名众包工作者。对于任务 A，按多数票确定标签后，几乎所有性别歧视条目都被正确标注（96%），但大多数非性别歧视条目被错误标注（正确率为 4%）。在 57 条性别歧视条目中，众包多数投票在任务 B 中的正确率为 28%，在任务 C 中为 14%。其次，我们招募并培训了 19 名自我认同为女性的人，为相同的 200 个金标准条目贴标签。以任务 A 的多数投票标签为准，受训标注员的正确率为 96%。在 57 个性别歧视条目中，受训标注员在任务 B 中的多数投票正确率为 84%，在任务 C 中为 72%。鉴于这项任务的敏感性和复杂性，与受训标注员合作能提供质量更高的标签，并具有其他优势，例如可以监测标注员福祉，并根据反馈更新指南。

F 向量间的混淆

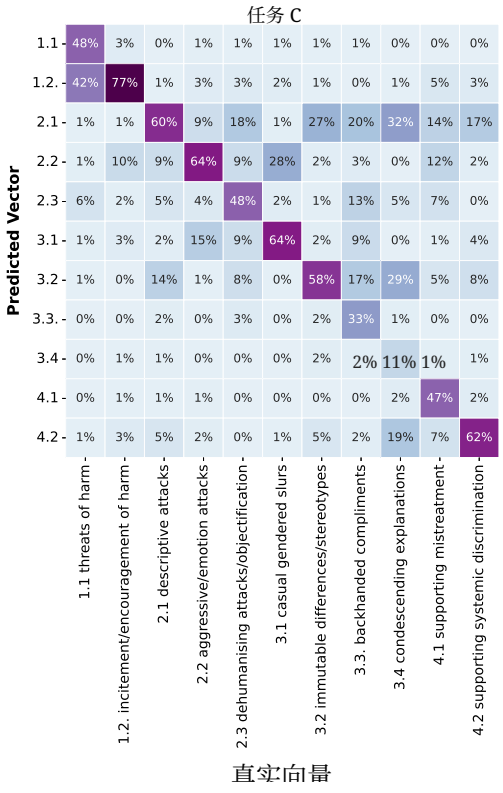


图 4：排名前十的系统在性别歧视向量类别（任务 C）上的混淆矩阵。