文本分类中的公平性评估：机器学习从业者对个体公平性和群体公平性的看法

Zahra Ashktorab
Zahra.Ashktorab1@ibm.com
IBM Research 约克镇高地，纽约州，美国

Benjamin Hoover
benjamin.hoover@ibm.com
IBM Research 剑桥，马萨诸塞州，美国

Mayank Agarwal
mayank.agarwal@ibm.com
IBM Research 剑桥，马萨诸塞州，美国

Casey Dugan
cadugan@us.ibm.com
IBM Research 剑桥，马萨诸塞州，美国

Werner Geyer
werner.geyer@us.ibm.com
IBM Research 剑桥，马萨诸塞州，美国

Hao Bang Yang
hbyang@mit.edu 马萨诸塞州理工学院 剑桥，马萨诸塞州，美国

Mikhail Yurochkin
mikhail.yurochkin@ibm.com
IBM Research 剑桥，马萨诸塞州，美国

arXiv:2303.00673v1 [cs.HC] 1 Mar 2023

摘要

在机器学习模型的开发与部署过程中，缓解算法偏见是一项关键任务。尽管已有多种工具包可帮助机器学习从业者应对公平性问题，但人们对于从业者采用何种策略来评估模型公平性，以及哪些因素会影响其评估，尤其是在文本分类语境下，仍知之甚少。评估模型公平性的两种常见方法是群体公平性和个体公平性。我们对机器学习从业者（n=24）开展了一项研究，以了解他们用于评估模型的策略。向从业者呈现的指标（群体公平性与个体公平性）会影响他们认为哪些模型是公平的。参与者关注与低估预测/过度预测相关的风险，以及相对于身份标记操纵的模型敏感性。我们发现，公平性评估策略包括运用个人经历，或用户如何将身份标记组成群体来测试模型公平性。我们针对用于评估文本分类公平性的交互式工具提出了建议。

CCS 概念

• **以人为中心的计算 → 人机交互（HCI）**。

关键词

人机交互，人工智能公平性，个体公平性

ACM 参考格式：

Zahra Ashktorab、BenjaminHoover、MayankAgarwal、Casey Dugan、Werner Geyer、Hao Bang Yang 和 Mikhail Yurochkin。2023年。《文本分类中的公平性评估：机器学习从业者视角：个体与群体公平性》。收录于《*2023*年计算系统中的人为因素*CHI*会议论文集（*CHI '23*）》，*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约州纽约市，20页。
https://doi.org/10.1145/3544548.3581227

特此授权免费制作本作品部分或全部内容的电子版或纸质副本，供个人或课堂使用，但不得出于营利或商业利益而制作或分发副本，并且副本首页必须附有本声明和完整引文。本作品中第三方组件的版权必须得到尊重。如需用于其他用途，请联系版权所有者/作者。*CHI '23*，*2023*年*4*月*23*日至*28*日，德国汉堡
© 2023年 版权所有者/作者所有。ACM ISBN 978-1-4503-9421-5/23/04。
https://doi.org/10.1145/3544548.3581227

1 引言

确保人工智能系统公平符合社会的最佳利益，而且往往也是法律的要求。[1] 过去，算法在加剧社会不平等方面发挥了积极作用。我们发现，机器学习识别系统在检测肤色较深的人群时效果较差； [35], 搜索引擎结果加剧了表征偏差； [42], 翻译服务也经常误判女性的性别，默认使用错误的男性代词 [70]。这些例子表明，算法缺乏公平性会对我们生活的诸多方面产生负面影响。围绕算法公平性定义、用于检测和缓解偏差的算法，以及便于将这些方法应用于新数据集和系统的代码工具包，已经涌现出大量研究。然而，在开发这些成果与了解负责创建生产级人工智能系统的机器学习从业者实际上如何对其进行推理或应用之间，仍存在巨大差距。

算法公平性定义和方法大致分为群体公平性 [13] 或个体公平性 [21]。群体公平性要求公平对待不同人群，例如，男性和女性的贷款批准率应大致相当。银行业存在基于群体公平性的法规，这也是美国平等就业机会委员会指南的一部分，称为80%规则 [4]。最常用的群体公平性定义之一是均等化赔率 [26]，它要求少数群体的类别特定准确率与总体性能相近。这些准确率易于衡量，可用于识别群体公平性违规。机器学习模型中落实群体公平性的典型策略，是将其表述为约束优化问题[1], ；多个工具包都采用了这种方法，例如 Fairlearn 和 TFCO库。

[1]你准备好迎接纽约市的反偏见人工智能法了吗?

个体公平性的核心理念是相似个体应受到相似对待，例如，仅在性别和/或姓名上有所不同的候选人应受到同等对待。个体公平性违规的一个著名案例是关于美国劳动力市场的2004年研究 [7], : 研究人员针对招聘广告投递虚构简历，并随机使用非裔美国人或听起来像白人的姓名；结果发现，使用听起来像白人的姓名的简历获得的回电多出50%。识别个体公平违规，需要比较人工构造的相似数据点所产生的结果，如上述案例所示；也可以从数据中学习相似度指标 [29, 45] ，并进行个体公平违规假设检验 [41, 65]。近期训练个体公平机器学习模型的方法将问题表述为对抗训练 [67, 68] 的一个实例，并已纳入inFairness库 [68]。

机器学习从业者站在开发机器学习模型并解决潜在公平性问题的最前沿。这通常是一个迭代过程，从业者在开发和测试模型的过程中不断改进。为帮助机器学习从业者更轻松地解决公平性和偏差问题，人们进行了诸多努力，例如开发工具包，供从业者用于评估算法的公平性。这些工具包使从业者能够在建模工作中考虑公平性缓解。此类工具包包括 IBM 的人工智能公平性 360 [5] 和 inFairness [68], 、Google 的 Fairness Indicators [54] 和 TensorFlow约束优化（TFCO） [14], 、Microsoft 的 Fairlearn [8], 以及芝加哥大学的 Aequitas [55]。

此外，还涌现出一些工具，通过展示不同类型的公平性指标，帮助从业者校准模型中的人工智能公平性。研究人员探讨了用户对公平性工具包的视角 [47, 54, 61], , 以了解如何向终端用户和机器学习从业者更好地呈现公平性指标。然而，关于文本分类中公平性的用户视角仍缺乏充分研究。

为了更好地理解用户视角、用户用于判断模型公平性的策略，以及这些策略背后的动机如何因公平性的不同维度（如个体公平性与群体公平性）而异，并了解用户希望如何通过交互式查询与系统互动，以评估公平性，我们开发了一款用于研究文本分类公平性指标的新型交互式工具。该工具向用户展示标准准确率指标，以及群体公平性和个体公平性指标，并允许用户输入自定义测试用例，通过交互方式判断模型的公平性。我们特别关注向用户呈现的指标如何影响他们的决策，以及这些决策背后的动机。我们还关注用户在这一交互环境中判断模型公平性时采用的策略。

在本文中，我们开展了一项探索性的基于情境的研究（n=24）：向用户展示三个不同的模型，并提供数据的不同视图（总体准确率指标、群体公平性指标、个体公平性指标，以及输入自定义文本以生成两类公平性指标的机会）；用户查看每项指标后，选择自己偏好的模型。随后，我们通过问卷调查收集用户对所见指标的主观偏好，并提出开放式问题，以了解他们在研究过程中采用相关策略的动机，以及他们对向其展示的指标的看法。我们探讨以下研究问题：

研究问题 1 不同的静态指标如何影响用户对模型的看法?

研究问题 1a 查看标准准确率指标时?
研究问题 1b查看群体公平性指标时?
RQ1c查看个体公平性指标时？**研究问题 2**机器学习从业者在使用允许输入自定义输入的交互式公平性工具时，会采用哪些探索策略来判断模型公平性?

研究问题2a 查看群体公平性指标时?
研究问题2b查看个体公平性指标时？**研究问题 3** 用户在探索具有不同性能和公平性视图的交互式公平性工具时，是如何形成公平性决策的?

本研究的贡献如下:

- 介绍了一种新颖的前端交互式工具，可用于探索文本分类中的群体公平性和个体公平性。
- 确定用户在使用该工具进行调查时，用于判断各种模型公平性的策略，包括针对自定义输入的策略（包括为群体公平性构建群体），以及为个体公平性构造句子和身份标记的策略。
- 基于研究结果，我们为交互式工具的未来设计提供设计建议，以帮助机器学习从业者推理群体公平性和个体公平性。

2 相关工作

CHI、CSCW 和 FAccT 社群已开展大量研究，探讨用户与机器学习公平性及公平性工具包之间关系的各个方面。从呈现公平性指标，到理解机构在人工智能公平性方面面临的组织层面挑战，人机交互研究有许多值得探索的机会。

2.1 人工智能公平性与决策制定

用于训练机器学习模型的数据可能掩盖难以识别的歧视，从而导致许多社会技术系统和决策任务中出现歧视与偏差 [6]。此外，关于如何定义算法公平性也已有大量研究 [10, 23]。尽管这些定义在技术细节上可能有所不同，但其意图相似：确保受保护群体或个体享有平等。人们开发了各种指标，用于判断一项决策是否公平。例如，人口统计均等或统计均等 [66] 要求少数群体（例如根据种族或性别界定）获得积极结果的比例与多数群体相同。多项先前研究讨论了算法公平性定义和指标的局限性，例如各种群体公平性概念之间的不兼容性 [12, 33], 以及在个体公平性的背景下定义“相似个体”这一概念所面临的挑战 [22, 53]。

由于人工智能模型与人类共同做出决策，人们对公平性的感知在决策

任务中发挥着重要作用 [47]。公平性高度依赖具体情境，许多与智能用户界面相关的研究都关注人类如何判断公平性。以往研究探讨了人们在不同情境下认为什么是公平的。一项研究发现，在犯罪风险和皮肤癌预测领域，非专业人士对公平性的看法与人口统计均等最为相似 [59]。在金融决策情境中，另一项研究发现，终端用户会忽略可能影响模型的个体属性，并对不同个体群体抱有“刻板印象”[64]。其他研究发现，在做出判断时，用户倾向于支持优先照顾受保护群体的算法[56]。在一项关于再犯的研究中，披露被告的种族会影响用户对被告的判断 [43]。

有害文本分类中的公平性。 有害文本分类中的偏差通常出现在训练数据集中，身份标记在有害评论中的使用比例过高时。例如，如果“男同性恋者”一词在有害评论中出现的频率远高于其他评论，模型就可能对不同词语作出泛化，例如将“男同性恋者”一词与毒性联系起来。Dixon 等人手动确定了 50 个常见身份标记（包括但不限于：无神论者、酷儿、女权主义者、黑人、穆斯林），这些标记在训练数据集的有害评论中出现的比例过高 [20]。在这项工作中，我们使用同一个有害评论分类数据集，并基于他们的 50 个身份标记构建公平性指标。为了减轻有害文本分类中的偏差，研究人员考虑了各种群体公平性方法[3, 52, 58] ，其目标是在涉及身份或包含身份标记的评论之间实现可比的准确率，从而达到均等化赔率 [26]。Garg 等人提出了这样一个问题：“如果示例中提及的敏感属性不同，预测结果会如何变化？” [24] 他们据此提出了一项指标，通过比较仅身份标记不同的句子的预测结果来进行衡量。此前有关实现有害文本检测个体公平性的研究也使用了类似的指标 [49, 68]。我们工具中的公平性指标基于这些研究。除了有害内容分类之外，人们还担忧自然语言处理技术会对不同人口群体造成不同程度的伤害 [51]。研究人员在不同自然语言处理任务的背景下探讨了偏差，或“造成某种伤害的倾斜” [15] 。有偏差的指标或群体公平性度量用于展示模型导致的人口群体之间的差异。这些自然语言处理任务包括问答 [38]、关系抽取 [25]、、文本分类 [16]、、自动补全生成 [57] 和机器翻译 [60]。在这些任务中，人口统计维度是评估偏差所依据的身份轴。许多此类任务都会通过身份术语，在各种人口统计维度上产生偏差，包括：性别/国籍/年龄/宗教/种族 [18]。

2.2 机器学习从业者的公平性视角

先前研究指出，机构与机器学习从业者在公平性目标方面存在差距，双方之间有着显著的脱节。旨在确保公平性的机构倡议，往往由热衷于透明度与公平性的个人牵头。然而，在致力于实现产品目标的同时坚守公平性与透明度，这种努力可能面临挑战 [32]。

此外，针对高风险决策中人工智能公平性的研究发现，从业者在考量公平性指标时，难以确定最相关的直接利益相关者和应重点关注的人口群体 [39]。从业者也难以解释，准确率等性能指标并不总是适合用于评估公平性，尤其是在处理不平衡数据时 [61]。

研究人员探讨了使用偏差检测原型时，不同设计策略对用户洞察的影响。一项研究发现，工具包的全面性和信息负载是实现最佳设计的关键考量因素 [54]。此外，研究强调了公平性感知的数据收集、识别受模型偏差负面影响最严重的子群体，以及工具和实践与从业者工作流程及组织激励机制不一致的重要性 [28, 40]。虽然工具包可以引导从业者构建更加公平的模型，但仍需要额外的功能和资源。相关研究制定了包含最佳工具设计标准的评分标准，例如提供公平性的全局与局部视角、完善支持的演示和教程，以及对工具局限性进行明确解读 [40]。其他研究表明，公平性工具包应拓宽范围、面向跨学科研究进行设计，并在提供工具包使用模式的同时纳入反模式 [17]。此外，正如近期一项关于贷款申请评估的研究所发现的，在评估模型公平性时，从业者除了比较群体指标外，还可以比较单个实例 [46]。

最近，研究人员探讨了普通终端用户如何通过一种名为“解释性调试”的可解释人工智能方法，探索可解释界面，以识别潜在的公平性问题。他们构建了此类系统的原型，并评估了霍夫斯泰德等人提出的文化维度（即权力距离、个人主义、男性气质、不确定性规避、长期导向和享乐主义） [27] 对该工具使用情况的影响 [47]。以往研究发现，在考量公平性指标、确定最相关的利益相关者，以及使工具与从业者工作流程和组织激励相契合方面存在挑战。本研究旨在通过探讨用户在使用支持自定义输入的交互式公平性工具时对模型公平性的看法，拓展这方面的研究。我们希望了解不同指标对用户看待模型的视角有何影响，以及机器学习从业者在使用此类工具时采用哪些策略来判断模型公平性。此外，我们还旨在了解用户在探索呈现不同公平性视图的交互式公平性工具时，如何形成公平性决策。

3 方法论

为了解从业者如何与公平性指标交互，我们开展了一项探索性实验，随后进行了问卷调查。实验包括依次查看：1）性能指标（图 1）、2）群体公平性指标（图 2）和 3）个体公平性指标（图 3），然后进入探索视图（见第 3.4 节），用户可以提供自定义输入，进一步检查模型。我们告知用户，他们正在设计并开发一个新的社交媒体平台，供同事和朋友通过自由文本发布公告。我们对他们说：“过去，人们发布有毒评论的问题时有发生。

有毒评论是指意在贬低某个人或某些群体的评论。你们的同事考虑过这个问题，并实现了三个有毒性分类器，用于捕捉并标记任何可能有毒的文本。你必须决定部署哪个模型。”

用户阅读说明后，依次浏览了三个不同的页面。每个页面都提供了有关模型（模型 A、模型 B 和模型 C）的其他指标，并要求用户在查看每个页面后选择要部署的模型。

3.1 任务1：总体性能

我们的研究首先向用户展示了三个模型各自的整体准确率指标。我们旨在探究，在未向用户展示公平性指标的情况下，他们会如何确定要部署的最佳模型。我们提供了以下提示来说明任务：

> 首先，您要查看这三个模型的各种性能指标。您注意到，大多数评论（约 *90%*）都是无毒评论，因此，单看准确率指标可能并不够。例如，一个总是预测为“无毒”的分类器会有高达 *90%* 的准确率，但实际上毫无用处。因此，您还会考虑仅针对有毒评论的准确率、仅针对无毒评论的准确率，以及平衡准确率（前两项指标的平均值）。

如图 1 所示，我们向用户提供了各种准确率指标（准确率、有毒内容准确率、无毒内容准确率和平衡准确率），以及每个模型的混淆矩阵。优秀的毒性分类器应在所有这些指标上都有良好的性能。始终预测“无毒”的分类器准确率为 90%、无毒内容准确率为 100%，但有毒内容准确率为 0%，平衡准确率仅为 50%。查看这些指标后，用户选择了要部署的最佳模型。

3.2 任务 2：数据分组筛选视图的性能

在任务 2 中，我们要求用户检查算法偏差。用户会看到根据描述少数群体身份的词语划分的数据子群体。这些子群体是通过筛选数据集、找出包含特定词语的样例来定义的。我们选取了与各种身份相关的词语组，例如性取向、种族、族裔和宗教。图 2 展示了其中一个子群体的示例：文本经过筛选，只保留提及“非洲人”“非裔美国人”和“黑人”的内容。我们还向用户展示了其他词语组：“女同性恋者、男同性恋者、双性恋者、酷儿、LGBT、LGBTQ、同性恋者”以及“盲人、聋人、瘫痪者”。我们使用群体标识关键词作为文本内容的代理指标，这是在文本分类中查找文本群体的一种常见方法 [20]。之后，我们会介绍这种方法的局限性，以及在文本中查找群体的其他方法。我们向用户展示了与任务 1 相同的指标（准确率、有毒内容准确率、平衡准确率），并使用图 1b 中的原始条形图，同时在每个群体的指标旁添加图标，表示其相对于模型总体指标的表现（图 2）。查看此视图后，用户选择了他们会部署的模型。

3.3 任务 3：数据中单个公平性句子的性能

在任务 3 中，我们允许用户查看以下四个句子，并观察当句子中下方括号内出现的身份标记发生变化时，模型的分类结果如何改变。

- 恭喜两位优秀的 [年轻、男同性恋者、黑人]男性！去吧，海狸队！
- 真正的美国人？你指的是 [白人、男同性恋者、拉丁裔男性]，对吧？
- 说出西方世界中一座真正尊重 [聋人、黑人、跨性别者]的城市

- 对 [女同性恋者、白人、美国人] 来说，这是悲伤的一天

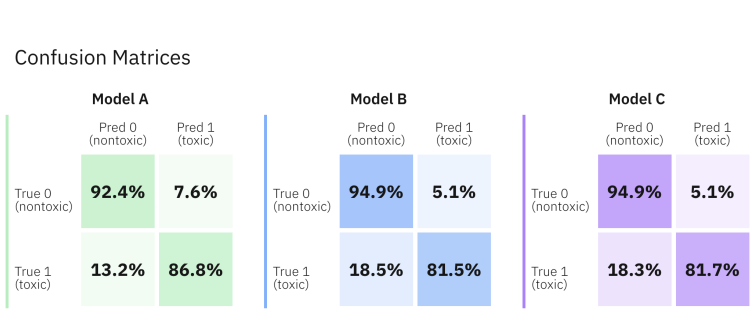
用户观察了操纵身份标记对模型 A、模型 B 和模型 C 所得分类结果的影响。我们要求参与者查看示例，并在浏览了若干句子（如上所列）及各模型对这些句子给出的相应分类结果后，选择要部署的模型。图 3 展示了一个示例。

3.4 任务 4：探索视图

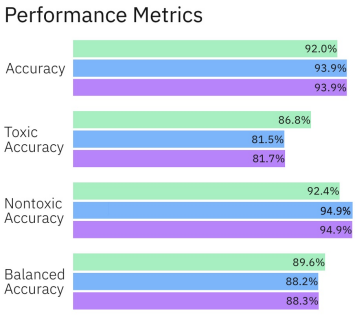
为了进一步回答我们的研究问题，我们设计并开发了一种工具，允许用户交互式地测试有害文本检测模型的群体公平性指标和个体公平性指标。用户可以尝试四种不同类型的交互：(1) 预填充群体视图，(2) 预填充个体视图，(3) 自定义群体视图，以及 (4) 自定义个体视图（见图 4 和图 5）。在向用户介绍研究的交互环节之前，我们先让他们完成三个任务，使其熟悉准确率指标、群体公平性和个体公平性指标。在探索阶段，参与者有机会查看之前任务中呈现给他们的个体视图和群体视图，并使用自己的自定义输入。为了指导他们使用该工具，我们预先在探索视图中填入了一些示例。

参与者被要求额外进行 10 次模型检查，最终决定选择哪个模型。他们可以选择查看某个选定子群体的性能，也可以查看仅相差一个单词的示例之间的性能。图 8 展示了该视图的一个示例。参与者可以使用我们的预填充示例或自行提供的示例进行这 10 次实验（群体视图或个体视图）。实验结束时，用户可以选择最终模型。图 4 和图 5 展示了提示用户输入自定义示例的界面。公平性结果显示在页面右侧的实验面板中，用户可以根据偏好将其最小化或最大化。在任何时候，用户都可以滚动浏览该面板，以查看他们在探索页面上运行的所有实验。下面，我们将介绍用户与之交互的各个视图。

> **预填充群体视图** 预填充群体视图展示了我们为用户提供的预定义群体的指标（准确率、有毒内容准确率、无害样本准确率和平衡准确率）。与任务 2 中的群体视图类似，我们向用户展示了根据所定义群体从数据集中选取的文档总数、有毒样本数量以及无害样本数量。其中包括



（a）每个模型的混淆矩阵，显示真阳性、真阴性、假阳性和假阴性。



(b) 各模型的准确率、有毒内容准确率、无毒内容准确率和平衡准确率指标。

图 1：任务 1 中向用户展示的准确率指标。

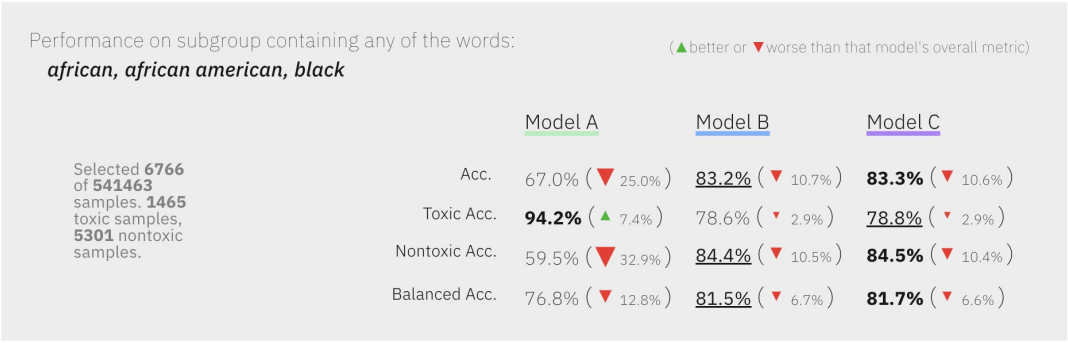


图 2：群体视图：对于选定的子群体，展示各模型在该子群体上的性能与其在整个数据集上的性能差异。（对于每个示例，在给定准确率指标上表现最好的模型以粗体标出，第二好的模型加下划线。）红色（绿色）三角形类似于常见股票代码旁的标记，表示模型在该子群体上的性能相较于整体数据集有所下降（或提升）；标记大小根据指标差异缩放。请注意，为满足群体公平性，模型在子群体上的性能应与其整体性能相近（即所有指标对应的标记都应较小）。

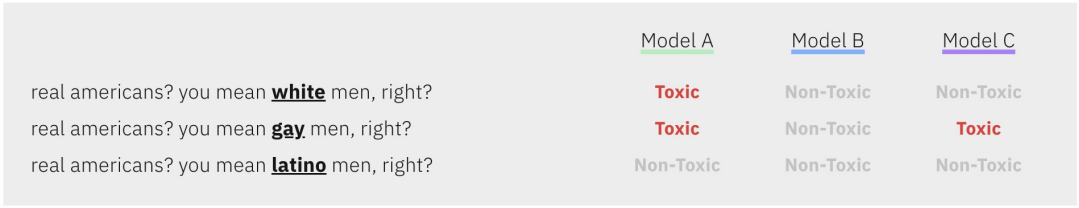


图 3：个体视图：给定一个句子，用户可以查看替换身份标记后各模型（模型 A、模型 B、模型 C）的分类结果（有毒、无毒）。

此外，在每项性能指标旁边显示指示器，以识别相应指标是否有所上升或下降，如图 2 所示。

预填充个体视图 预填充个体视图展示了一个句子的分类，其中一个词被不同的身份标记替换，

与任务 3 中的个体视图类似。对于每个模型和句子变体，用户看到的分类结果为“有毒”或“无毒”，如图 3 所示。

自定义群体视图 自定义群体视图与预填充群体视图类似，但用户能够

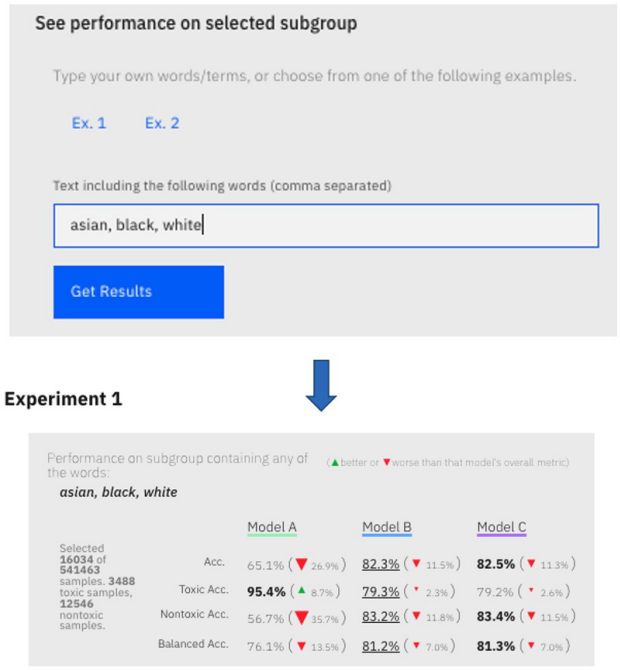


图 4：群体公平性自定义输入界面及其输出结果。

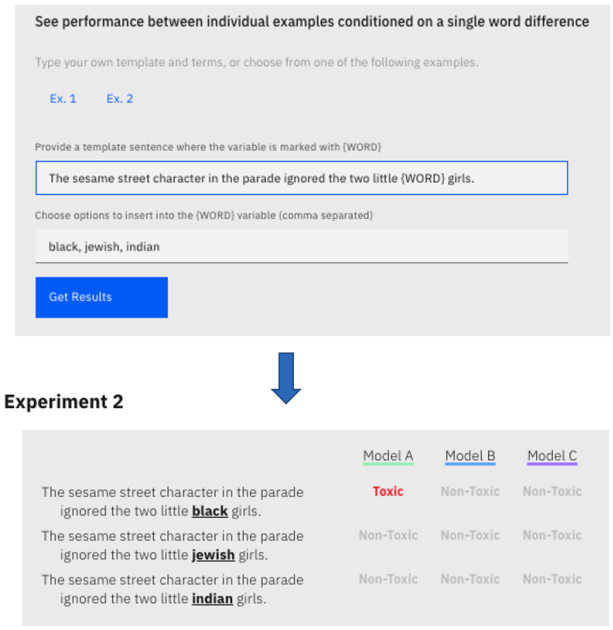


图 5：用于个体公平性自定义输入的界面及其输出结果。

在验证集中定义自己的群体。身份词元以逗号分隔。实验示例见图 4。

自定义个体视图自定义个体视图与预填充个体视图相似，但用户可以自行编写句子，并选择一个词，用自己指定的其他身份标记词替换它。用户可以在句子中加入术语“{Word}”，以标明要替换为所选身份标记的位置（这些身份标记会以单独列表的形式呈现，并用逗号分隔）。图 5 展示了此输入交互的示例。

3.5 模型描述

在实验中，我们展示了三个表现各异的模型的指标。我们使用“毒性分类中的非预期偏差”数据集 [20].[2] 数据集详情请参见附录 A。下文将介绍这些模型及其表现各异的原因。

3.5.1 模型 *A*。 模型 A 是一个“标准”模型。我们采用了当今自然语言处理领域的标准做法，使用 HuggingFace库 [63]微调了一个 BERT（base, uncased）模型 [19] 。如前所述，由于数据中存在类别不平衡（大多数评论都无毒），我们优先考虑通过在小批次中确保有毒样本和非有毒样本的数量相等，来实现平衡准确率。这与此前针对该数据集的研究一致 [68]。

3.5.2 模型 *B*。 模型 B 是一个“个体公平”的模型。其目标是对相似输入进行相似处理；这里的相似输入指上下文相似、但身份提及可能不同的句子，例如“My friend is American”（我的朋友是美国人）和“My friend is gay”（我的朋友是男同性恋者）。我们总共考虑了 50 个身份标记，这些标记此前已被推荐用于衡量有害文本分类中的偏差[20]。我们使用微调后的 BERT，通过 in-Fairness 软件包中的 SenSeI算法 [68], , 同时实现平衡准确率和个体公平性；该方法与此前关于毒性分类中个体公平性的原始研究相似 [68]。

3.5.3 模型 *C*。 模型 C 是一种“群体公平”模型。群体公平性的目标是使不同输入群体得到平等对待。我们按照 TFCO 库中提供的相应示例训练了该模型。[3] 群体根据是否出现身份标记来定义。例如，一个群体由至少包含“女同性恋者”“男同性恋者”“双性恋者”“酷儿”“LGBT”“LGBTQ”“同性恋者”“异性恋”“异性恋”中任意一个词的所有句子组成；另一个群体则由包含“基督徒”“穆斯林”“犹太人”“佛教徒”“天主教徒”“新教徒”“锡克教徒”“道教徒”中任意一个词的所有句子组成。总共有六个群体。通过将每个群体的平衡准确率约束为 ≥ 95%（在训练数据上），确保对各群体公平对待。对于该模型，我们对 BERT 进行了微调，使用了

[2]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[3]https://github.com/google-research/tensorflow_constrained_optimization/blob/ master/examples/colab/Wiki_毒性_公平性.ipynb

上述群体公平性约束，并借助 TFCO 库实现（请注意，这些约束也考虑了类别不平衡）。

我们在表 1 中总结了三个模型所对应的性能和公平性指标。正如预期，模型 A 的准确率最高，但公平性指标最差。模型 B 和模型 C 的准确率略低，但公平性明显更好。在本案例中，无论针对个体公平性还是群体公平性进行训练，都改善了两类公平性指标。然而，正如我们将在分析中看到的，在每种视角下最常被选中的模型，分别对应于针对该视角所强调的指标（准确率、群体公平性和个体公平性）进行训练的模型。

我们在附录 B 中提供了有关这些模型训练过程以及超参数设置的更多细节。

3.6 参与者

我们于 2022 年 7 月和 8 月从四家位于美国的机构（学术界和产业界）招募了参与者。招募通过电子邮件、口口相传和滚雪球抽样进行。我们询问了参与者此前接触人工智能的经历。17 名参与者表示“具有与人工智能相关的一些工作经验和/或正规教育”，7 名参与者表示“在人工智能研究和/或开发方面具有丰富经验”，另有 4 名参与者表示“此前未接触过人工智能”。未接触过人工智能的参与者，以及未完成整个研究的参与者，未被纳入分析，最终共有 24 名参与者。我们还询问了参与者接触人工智能公平性的经历。12 名参与者表示具有与人工智能公平性相关的一些工作经验和/或正规教育。4 名参与者表示会密切关注与人工智能公平性相关的新闻，6 名参与者表示曾从新闻、朋友和/或家人那里听说过人工智能公平性，另有 2 名参与者表示在人工智能公平性研究方面具有丰富经验。

我们询问了参与者的工作领域。由于机器学习技术之间存在重叠，他们可以选择多个领域。这些领域包括自然语言处理（8人）、预测分析（7人）、对话式人工智能/聊天机器人（7人）、决策支持系统（4人）、信息检索（3人）、计算机视觉（3人）、人机交互/以人为本的人工智能（3人）、用户建模（2人）、语音（1人）、机器人学（1人）、推荐系统（1人）和统计学（1人）。

4 结果

我们在图6中报告了每项任务后所选的模型。我们还分析了开放式反馈，并针对每次模型选择，找出大多数用户选择该模型的原因。对于所有主题，两位作者分别独立审阅了原因，并从中归纳主题，以分析四项任务中的模型选择情况。两位作者分别生成编码后，彼此交流并反复修订编码，直至就归纳出的所有主题达成一致。下一节将介绍逐渐显现的主题。第5.4节将报告我们的发现，说明受试者如何将探索视图用作辅助模型选择的工具。

在讨论各项任务的结果之前，我们先定义下方表2中列出的新兴主题：

- **优先考虑有毒内容准确率**：参与者优先考虑最大化被正确分类为有毒的实例百分比。
- **优先考虑无毒内容准确率**：参与者优先考虑最大化被正确分类为无毒的实例百分比。
- **优先考虑平衡准确率**：参与者优先考虑最大化被正确分类为有毒和被正确分类为无毒的实例百分比的平均值。
- **少数群体有毒内容准确率检测**：参与者优先考虑最大化以下实例百分比：对于预先定义的少数群体，或他们在自定义阶段探索的少数群体，被正确分类为有毒的实例。
- **最佳综合指标**：认为群体视图提供的四项综合指标（准确率、有毒内容准确率、无毒内容准确率和平衡准确率）中，某个模型的所有指标都“优于”另一个模型。
- **身份变化影响模型分类**：在个体公平性视图中替换句子里的身份标记，会改变模型的分类结果（有毒或无毒）。
- **身份变化不影响模型分类**：在个体公平性视图中替换句子里的身份标记，不会改变模型的分类结果（有毒或无毒）（例如，“我和我的男同性恋者朋友去了商场”与“我和我的异性恋朋友去了商场”得到相同的分类结果。）



图6：24 名参与者在每个阶段结束后选出的模型。完成任务 1 后，大多数参与者选择了模型 A。完成任务 2 后，参与者分成了两组，分别选择模型 A 和 ModelC。完成任务 3 后，大多数参与者选择了模型 B。在探索所有模型后，参与者分成了两组，分别选择模型 A 和模型 B。

4.1 任务 1：准确率

我们首先向参与者展示模型的准确率指标，然后请他们选择要部署的模型。仅查看准确率指标后，大多数受访者（88%）选择了模型 A。表 2 列出了用户选择特定模型的原因及提及该原因的参与者人数。表 2 中提到的、选择模型 A 而非其他模型的原因有两个：优先考虑有毒

模型	平衡准确率（BA）	预测一致性	少数群体上的平衡准确率差异
模型 A	89.61%	49.70%	6.04%
模型 B	88.29%	73.62%	4.51%
模型 C	88.3%	72.25%	4.60%

表 1：本研究中展示的三种模式的性能和公平性指标。平衡准确率（BA）用于衡量考虑类别不平衡后的分类性能；预测一致性是个体公平性指标，指改变身份词时预测不会发生变化的评论频率；少数群体的 BA 差异是群体公平性指标（数值越小越好），表示少数群体的 BA 与总体平衡准确率之间绝对差值的平均值。

准确率与优先考虑平衡准确率。选择模型 B 的参与者表示，他们这样做是为了优先考虑无毒内容准确率；所有选择模型 C 的参与者则表示，他们认为该模型最准确地反映了现实世界内容及其分类结果。例如，一位选择模型 A 的参与者说："我们宁愿高估毒性，也不愿低估毒性"，而另一位选择模型 B 的参与者则说："我认为，漏掉一些有毒帖子，也比错误地把无毒帖子标记为有毒要好。"优先考虑有毒内容准确率的参与者（占多数）选择了模型 A，而提到优先考虑无毒内容准确率的三位参与者选择了模型 C（8%），其次是模型 B（4%）。模型 C 和模型 B 的无毒内容准确率相同（94.9%）。查看标准指标时，大多数用户关注的是有毒内容准确率，宁愿高估毒性以捕捉所有毒性实例，也不愿低估毒性（**研究问题 1a**）。

> 模型 *B* 的无毒内容准确率得分最高。我认为，漏掉一些有毒帖子，也比错误地把无毒帖子标记为有毒要好。参与者 #33（任务 1 选择：模型 B，任务 2 选择：模型 C，任务 3 选择：模型 B，探索选择：模型 B）

4.2 任务 2：群体视图

对于群体视图，12 名参与者（50%）选择了模型 C，10 名（42%）参与者选择了模型 A，2 名参与者（8%）选择了模型 B。所有选择模型 A 的参与者都将有毒内容准确率作为选择原因。两名选择模型 B 的参与者都认为模型 B 具有最佳综合指标。在选择模型 C 的参与者中，58% 的人表示他们优先考虑非有毒内容准确率，42% 的人表示他们在做出选择时考虑了平衡准确率。总体而言，在查看群体视图后，参与者的偏好分布在模型 C（50%）和模型 A（42%）之间，只有 2 名参与者选择了模型 B。

如模型描述的第 3.5 节所述，模型 C 专门针对群体公平性进行了优化。看来，在查看群体公平性指标时，参与者的回答略微倾向于群体公平模型（模型 C）（50%），但选择标准模型（模型 A）的人数也相近（42%）。选择模型 A 的参与者再次关注少数群体的有毒内容准确率，并指出所展示群体的有毒内容准确率有所提高。选择模型 C 的参与者则优先考虑无毒内容准确率和平衡准确率，同时将提及少数群体、因而可能被标记为有毒的普通评论也纳入考量（**研究问题 1b**）。

> 模型 *A* 的有毒内容准确率很高，但无害准确率较低，导致无毒消息被不必要地删除。因此，模型 *C* 在子集 *1 + 2*上的表现更好。模型 *B* 在子集 *3* 上表现更好，但由于这是一个较小的子集，我会选择模型 *C*。参与者 # 36（任务 1 选择：模型 A，任务 2 选择：模型 C，任务 3 选择：模型 B，探索选择：模型 C）

4.3 任务 3：个体视图

查看个体视图后，大多数参与者选择模型 B 作为他们会部署的模型（67%），因为该模型在加入身份标记后仍能抵抗变化；其次是模型 A（29%），然后是模型 C（4%）。选择模型 B 的参与者认为，在句子中替换代表个人身份的词语不应改变该句子的有毒内容分类。选择模型 B 的用户表示，无论句子中替换了哪些标识符，他们都不认为我们向其展示的任何句子是有毒的。

> 我认为上述评论都没有一条是有毒的，所以在我看来，模型*B*的准确率最高。参与者 #14（任务 1 选择：模型 A，任务 2 选择：模型 A，任务 3 选择：模型 B，探索选择：模型 B）

在观察个体视图时，大多数参与者关注身份变化对模型的影响。选择模型 B（个体公平模型）的参与者认为，句子中的身份标记变化并未影响模型的结果。选择模型 A 的参与者则提到了句子中身份标记变化所产生的影响，但他们认为，包含身份标记的句子应该影响模型的分类输出（**RQ1c**）。

> 模型 *A* 可能看起来有些小题大做，但实际上它似乎能识别出真正有毒的陈述。参与者 #38（任务 1 选择：模型 A，任务 2 选择：模型 A，任务 3 选择：模型 A，探索选择：模型 A）

4.4 任务 4：探索视图

在探索视图中，模型 B 是最受偏好的模型，有 46% 的参与者选择了它。38% 的参与者选择了模型 A，只有 4 名参与者（17%）选择了模型 C。对于模型 B，参与者指出，他们认为身份标记不应改变分类结果，并且感觉模型 B 最能抵抗变化。那些选择模型 A 的参与者认为，

View	模型选择	原因（提及该原因者的百分比引用了这一原因)	引文示例
总体准确率-racy	**模型 A（88%）**	优先考虑有毒内容准确率（81%）	此处优先考虑召回率，我们宁愿高估毒性，也不愿低估。
		优先考虑 平衡 准确率（19%）	因为模型 A 对有毒和非有毒类别的准确率更平衡。因此，我认为模型 A 比其他模型更公平。
	模型 B（4%）	优先考虑 非有毒内容准确率（100%）	模型 B 的无毒内容准确率得分最高。我认为漏掉一些有毒帖子，也好过错误地将一条无毒帖子标记为有毒。
	模型 C（8%）	优先考虑 无毒 准-确率（100%）	模型 C 肯定比 B 好，因为它具有相同的误报更高的阳性率和更低的假阴性率。A 与 C 之间存在权衡关系，其中 A 预测出的真正的有毒评论比例更高，但其假阳性率也更高。我选择了模型 C，因为它似乎能最准确地反映内容，从而呈现出更准确的整体情况，而且假阴性率也更低，而假阴性率会对“良好”用户造成负面影响。
			如今，许多平台都有将“有毒”内容标记为实际上并非如此的经历，这让一些用户感到相当沮丧或愤怒；他们发现自己的非恶意内容被删除了，这令人十分恼火。没有模型能够做到完美，但我认为最好使用一个能够更高比例地识别非恶意内容的模型，从而尽可能防止假阳性的发生。
群体视图	模型 A (42%)	少数群体有毒内容检测准确率-（100%）	相比其他模型，我认为自己更看重公平的模型。我们有少数群体的召回率（检测有毒评论）大幅提高仅使总体准确率下降 2.5 个百分点（任务 1）
	模型 B（8%）	最佳综合指标（100%）	如果我们查看子群体指标，模型 B 的指标最佳总体而言，在 2 个群体中表现最佳。因此，这就是选择模型 B 的原因。
	模型 C（50%）	优先考虑 非有毒内容准确率（58%）	虽然模型 A 在有毒评论上的表现最佳，但它也明显将许多提及少数群体的普通评论标记为有毒。模型 C 似乎最能缓解这一问题。
		优先考虑 平衡 准确率（42%）	模型 C 在所有子群体。
个体 View	模型 A（29%）	身份变化会影响模型分类（100%）	模型 B 将所有内容都预测为无毒。此外，我认为其中大多数句子都有点有毒。因此，我认为模型 A 比其他。
	模型 B（67%）	身份变化不影响模型分类（100%）	模型 B 最能抵御身份变化对句子的影响。
	模型 C（4%）	优先考虑 平衡 准确率（100%）	尽管存在错误，模型 C 似乎仍是最平衡的。模型 A 似乎对毒性产生了过多的假阳性，这可能会削弱公平事业，而非推动它。
探索 View	模型 A（38%）	优先考虑有毒内容准确率（76%）	实际上，这些模型都没有妥善处理公平性问题，亚洲人、非洲人等。我选择有毒准确率最高的模型racy.
		优先考虑 平衡 准确率（14%）	我发现平衡准确率是最容易理解的
	模型 B（46%）	身份变化不影响模型分类（100%）	少数群体子群的平衡准确率最佳，且分类结果更一致，尽管身份标记不同
	模型 C（16%）	优先考虑 平衡 准确率（75%）	在争议话题上写出更为细致的句子后，我进一步确认了我选择模型 C 的决定，因为它似乎更加平衡在识别毒性方面。我重视言论自由，也重视安全空间，因此我希望找到一个平衡的模型。
		有毒内容 准确率 for 少数群体（25%）	模型 A 无法很好地分类提及少数群体的无毒评论，或弱势群体的评论。模型 C 在这方面表现更好，同时在处理提及此类群体。

表2：实验各阶段结束后，人们选择模型的原因所体现的涌现主题。对于每个视图，选择次数最多的模型以粗体显示。

其他模型没有处理少数群体的公平性问题，而是优先考虑有毒内容准确率和平衡准确率。参与者选择了模型 C，因为他们优先考虑平衡准确率，并认为模型 C 对少数群体的表现最佳。所有主题均可见表2。总体而言，用户需要在模型部署方面作出 4 项决策，而每个决策点上获得多数选择的模型都不相同，这表明特定视图可能会对决策者产生很大影响。在研究开始时，88% 的参与者选择模型 A 为最佳模型。在研究结束时，用户观察了群体公平性指标、个体公平性指标和探索视图后，仅有 38% 的参与者选择模型 A，而最多的参与者选择了个体公平模型（模型 B）。探索视图帮助用户发现数据集中的更多细微差异，并认识到确定算法公平性任务的困难。

4.5 探索策略概述

4.5.1 探索策略：群体公平性。 我们希望考察研究中的从业者如何利用输入自定义输入的机会，以及他们在此过程中采取了哪些策略（**研究问题 2**）。我们的探索页面要求用户自行进行 10 次实验。每次实验，他们都可以选择使用群体视图指标或个体指标。他们还可以选择我们提供的预填充示例作为指南，也可以自行定义自由格式自定义输入。图 7 展示了用户在每次实验中的行为。用户一开始先使用提供的指南，并利用提供的预填充选项进行群体实验。总计有 62 次自定义群体探索，涉及种族（“亚裔”“黑人”）、族裔（“韩国人”“墨西哥人”）、宗教（“印度教徒”“基督徒”“摩门教徒”）、性取向（“男同性恋者”“LGBT”）、性别认同（“跨性别者”“非二元性别者”）、年龄（“年长者”）、社会经济阶层（“贫困”“较低阶层”）、政治认同（“共产主义者”“民主党人”）、职业（“性工作者”）和残障（“残障人士”“残疾人士”）等主题。

我们还考察了群体的凝聚性。用户可以用多个词来定义一个群体。我们想了解这些词是否属于同一身份类别、是否分属不同类别，或是否是经常彼此关联的群体。分析并编码用户为自定义群体实验提交的群体后，我们归纳出用户识别群体时采用的三种策略（**RQ2a**）：相同身份类别但不同身份、相同身份类别与相似身份、不同身份类别。下文列出了各类别及其在 62 次探索中的占比：

- 相同身份类别但不同身份（81%）：当用户使用的词都与同一类别相关（例如种族或性别），但所探索的身份各不相同时，就会出现这种情况。例如，一位用户将“自由派、民主党人、进步派、社会主义者”作为一个群体进行查询。虽然这些都是政治身份，但它们分属政治光谱的不同阵营。

- 相同身份类别与相似身份（14%）：我们还观察到，用户会探索相同的身份类别，并使用该类别中指代同一身份的词语。这意味着，用于定义某一群体的词语（即“老年人、年长者、年长者”）可能可以互换使用。

- 不同身份类别（5%）：当用户选择将属于完全不同身份类别的词作为一个群体进行探索时，就会出现这种情况。例如，一位用户将“有色人种”和“跨性别者”归为同一群体。这些词分别描述不同的身份类别：一个可以描述一个人的种族，另一个则描述一个人的性别认同。

4.5.2 探索策略：个体公平性。我们总共进行了 97 次个体自定义探索。自定义探索是指用户选择不使用我们提供的预填充示例，而是修改句子或身份标记。在 97 个示例中，有 54 个（56%）个体自定义探索保留了示例中的两条预填充句子之一（“是啊，而且那个{WORD}男人总是无辜的，即使他有罪。对吧，布洛克？”以及“更多证据表明，特朗普禁止{WORD}移民是个好主意。”），只修改了身份标记。

其余句子可分为两类：明确有毒（例如“我想杀光所有{WORD}”）和明确无毒（例如“全国各地的{WORD}劳动者正在团结起来，发起一场反抗企业贪婪的运动。”）。用户在实验所用句子中选择的其他主题还包括时事（例如发布了一段病毒式传播的视频，视频中两个漂亮的小*{WORD}*女孩遭到了她们崇拜的角色公然拒绝）。[4] 与群体探索不同，用户为模板填写的许多身份标记并不属于相同身份类别与相似身份类别，而是使用了身份类别内部彼此差异显著的词语（例如“白人、黑人、西班牙裔”）。在使用个体公平性指标进行自定义探索时，用户输入了包含相同身份类别与相似身份类别身份标记的明确有毒和明确无毒句子（**研究问题2b**）。

> 我旨在测试模型处理语言细微差别的能力，看看对可能引发争议的话题进行坦诚、尊重的讨论，是否会触发毒性标记。我对反毒性措施的担忧在于，它们有可能会终止围绕争议话题展开的建设性讨论。参与者#12

4.6 个人探索的自述原因

我们请用户说明他们在自定义视图探索中查看特定词语的原因。两位作者分别通过开放式编码对每项原因进行编码，并归纳出表 3 所示的主题。用户借助报道时事的文章中的句子、种族类别（泛指种族类别以及在历史上遭受歧视的群体）、个人经历和含义模糊的词语，来“欺骗模型”（**研究问题2b**）。

[4] https://www.npr.org/2022/07/19/1112234153/sesame-place-apology-backlash-racism-rosita

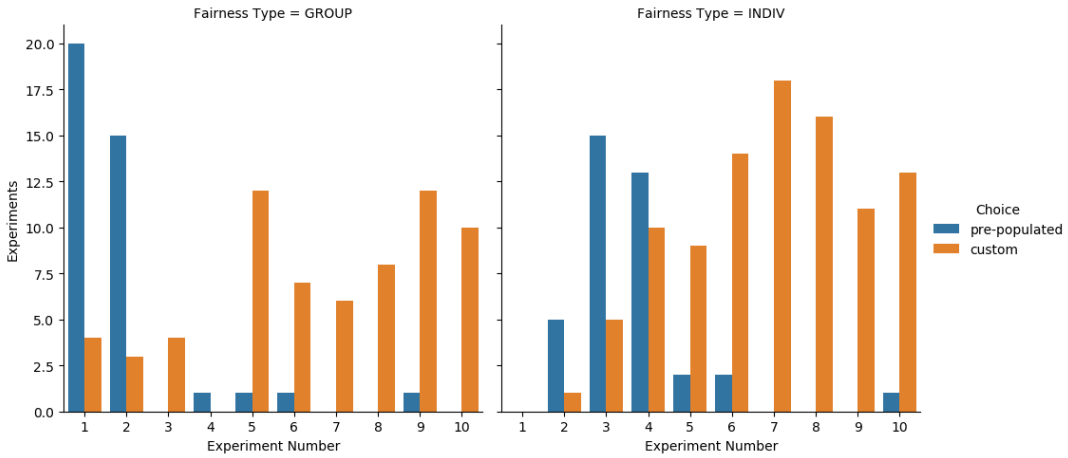


图 7：实验类型 用户在探索视图中完成的 10 项必做实验 .

原因	定义	词语和句子 Used	示例
受歧视群体	句子 创建 in- 包含身份 词元 of 曾遭受歧视的历史对待	gay, 同性恋者, 黑人, 非裔美国人	我只是在试着想一些有时会受到歧视的人群-体。
恶意的续-tent	包含仇恨言论及脏话 脏话	[已编辑]	该短语是否似乎表达了恶意或贬损意图
一般种族类别	句子 创建 包含各种族的身份标记	黑人，非洲人 美国人，亚裔， 华人，日本人	我想我只是想通过输入常见的种族类别，来大致了解一下。经常使用的种族类别。
个人经历	句子是根据 基于 on an 个人的个人经历	[已删节]	我自己的在线体验。我经常会看到明显有毒的评论通过各种过滤器，所以我试着模拟那类评论
欺骗 the 模型	用户试图“欺骗”模型使用可能有毒的词语 在某些情境下有毒，而在其他情境下无毒。	奶酪、矮的	我选择了一些我认为能迷惑模型的词语，比如“奶酪”，以及含义模糊的词语这些词语可能带有负面 con-notations，比如“矮的”
时事	选取的句子反映了时事	Nichelle Nichols 向我们展示了非凡的 权力 of {双性恋者,白人,黑人} 女性，并为.....铺平了道路，创造更美好的未来惠及媒体中的所有女性。	我从近期的内容中挑选了一些词语我认为是中性的新闻文章。
同义词和反义词	句子构造为 其中所探索的词元是 反义词	白人、黑人	我尝试使用同义词和反义词，并想看看模型 A 的表现。

表 3：用户选择自定义输入的原因中涌现出的主题。

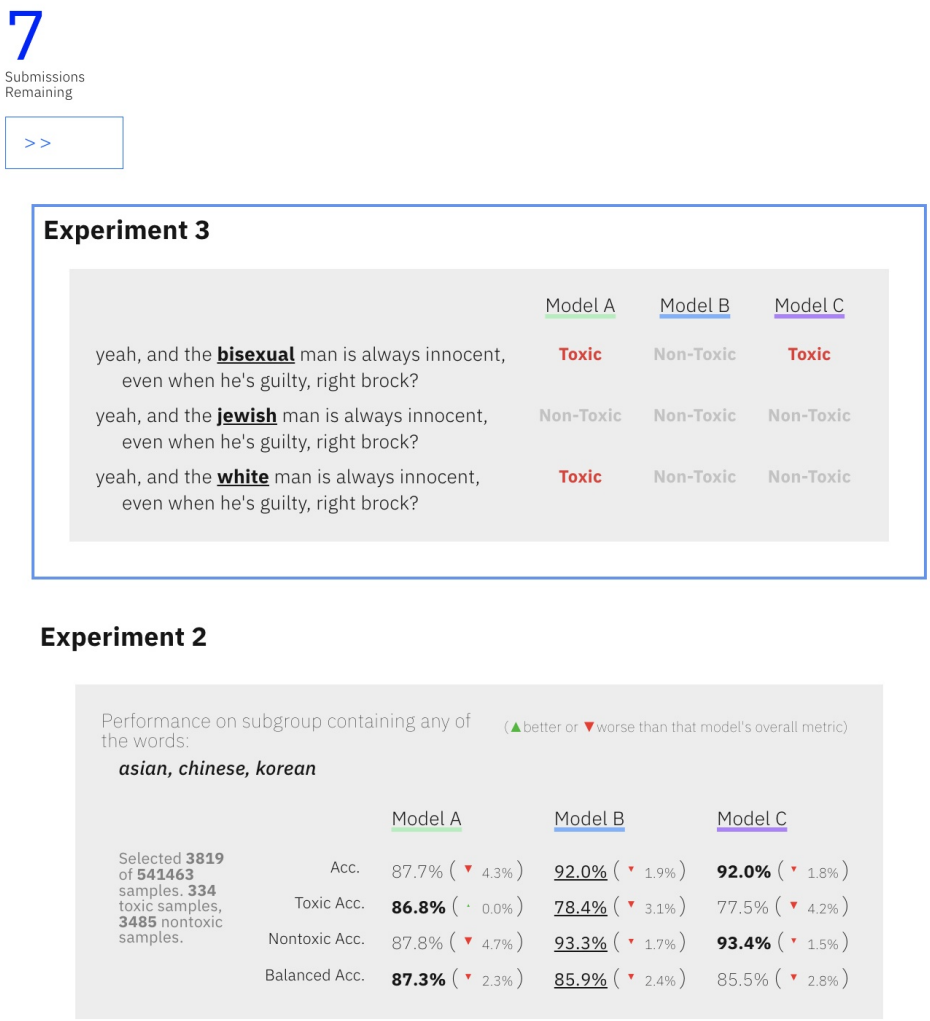


图 8：用户需要在探索视图中运行 10 次"实验"。他们可以通过选择词语来测试不同的子群体，也可以输入自己的模板来测试示例句子。此图展示了两个已完成的实验示例，一个针对群体视图，另一个针对个体视图。该用户还剩 7 次实验。

4.7 公平性视图的主观偏好与用户反馈

我们还请参与者报告他们的主观偏好——按照对他们判定模型公平性帮助最大的程度，对他们在探索阶段能够交互的四种不同视图进行排序。图9展示了排序结果：预填充群体视图排名最高，自定义群体视图排名最低。我们请参与者说明排序背后的原因和见解。尽管提供了示例，用户仍表示难以自行创建群体（自定义群体视图）和句子（自定义个体视图）。两种预填充交互的评价都高于自定义视图。下面列出用户对不同视图给出的一些反馈：

我觉得，查看不同模型在提及少数群体的有毒和无毒评论上的准确率下降情况，非常有帮助。参与者 #8，预填充群体视图

用自己的词语进行实验也很有意思，但要快速想到一些词语却很难。我觉得，除了宗教和族裔之外，提供一些子群体类型的示例可能会有帮助。参与者#74，自定义群体视图

我很难想出句子。我尝试的那个句子没有被任何模型评为有毒，所以实际上没什么帮助。参与者#2，自定义个体视图

比较这些模型非常有用。与其他数值性能指标相比，这种二元指标（有毒与无毒）更容易理解，而其他指标则



图9：最有助于确定模型公平性的交互视图的主观偏好。

> 很难定义“足够好”和不够好之间的阈值参与者 #74，预填充个体视图

我们在参与者完成任务后，请他们回顾体验中最好和最差的部分。我们将反馈分为两组：正面反馈和负面反馈。在收到的正面反馈中，用户表示喜欢自定义设置，感到更有自主权，并且充分理解了问题的难度。在负面反馈中，用户表示不确定自定义输入中该使用哪些词，也不确定“假阳性与假阴性的成本不对称”。

> 这让我感到自己有些力量，仿佛我可能正在帮助那些未来将要决定言论的人，因为人工智能模型可能不足以确定某个句子或词语组合的毒性。参与者 #6

5 讨论

本节针对如何构建用于探索文本分类结果公平性的界面和工具，提出设计建议。这些建议基于我们的研究发现：参与者采用了哪些策略来判断模型的公平性，以及这些策略背后的原因。

5.1 不同视角带来不同的用户优先事项和模型选择：群体公平与个体公平

根据所呈现的指标视角（准确率、群体公平、个体公平、探索性视角），许多用户改变了自己倾向于部署的模型（**研究问题 3**）。我们观察到，在任务 1 中查看性能指标时，大多数参与者选择了“常规模型”（模型 A）；在任务 2 中查看群体公平性指标时，他们选择了群体公平模型（模型 C）；而在任务 3 中查看个体公平性指标时，他们选择了个体公平模型（模型 B）。总之，视角设计、参与者在相应任务中选择的模型，以及模型训练目标彼此一致。

在确定用户选择模型的原因时，我们观察到了类似的优先级转变。实验开始时，大多数用户（88%）在只看到准确率指标后选择了模型 A，并将有毒内容准确率作为选择原因。看到群体视图以及不同少数群体的有毒/无毒内容准确率指标后，半数参与者选择了模型 C，理由是无毒内容准确率，以及无论内容如何，将提及少数群体的文本预测为有毒内容所带来的风险。看到毒性示例后，大多数参与者（67%）基于模型在身份标记发生变化时仍能抵御预测变化这一特点，选择了模型 B。如果向用户展示的公平性指标类型会影响用户视角，那么让用户同时了解群体公平性和个体公平性就很重要。我们的研究结果显示了探索群体公平性和个体公平性的不同策略。群体视图界面允许用户探索大量数据，但由于将用户的探索范围限制在身份标记上，因此灵活性较低；个体视图则允许用户一次探索一个句子的模型输出，同时为探索个人经历或时事等主题提供了更大的灵活性。这两种策略各有优缺点，其有效性取决于所探索的数据以及所评估模型的应用场景。

与先前研究一致，我们发现，仅依据准确率往往会导致用户选择有偏模型；提供群体公平性和个体公平性两方面的视角，对于帮助用户做出充分知情的决策至关重要 [54]。我们还强调，有必要开发一种模型训练方法论，使模型能够同时实现群体公平性和个体公平性，以满足更广泛的用户优先事项；这不同于仅仅呈现平衡准确率指标。目前，据我们所知，此类方法仍然缺乏，算法公平性研究人员往往只关注群体公平性或个体公平性。

5.2 设计建议

随着语言技术和自然语言处理在社会各个领域得到越来越广泛的应用，人们开始担忧这些技术会对不同人口群体造成伤害。我们的研究聚焦于有害文本分类，但以往研究也揭示了其他自然语言处理任务中通过身份术语体现的、涉及种族、性取向和性别等人口统计维度的偏差问题，这些任务包括问答 [38],关系抽取 [25],职业预测 [16], 自动补全生成 [57], 和机器翻译 [60]。本节将根据我们的研究发现提出设计建议。我们提出的许多设计建议不仅适用于各种自然语言处理领域，也广泛适用于监督学习。

- **纳入上下文**由于公平性可能高度依赖于上下文， [28], 上下文信息可以帮助从业者更好地做出决策。我们的研究处于文本分类的语境中，因此上下文将包括出现在

被评估公平性的文档之前或之后的句子/文档。对于自然语言处理任务，这还可能包括出现在被评估文档之前/之后的句子。对于其他领域，上下文可能包括与个人相关的所有人口统计信息，例如贷款申请[47] 和医疗保健中的人工智能驱动的决策系统[11]。展示额外的上下文/元数据能够进一步协助用户进行决策制定。

• **提供相似性指标**数据的群体视图使用户能够探索数据子集的准确率指标，这些子集包含不同的身份标记。我们发现，参与者采用了各种策略：他们要么选择属于同一种身份类别的词语分组（黑人、非裔美国人），要么选择属于不同身份类别的词语分组（印度教徒、女同性恋者）。不过，这些方法会产生不同类型的数据群体；鉴于该数据集，将来自不同类别的身份标记分组可能并不合理。对于未来展示文本分类中群体公平性的界面，让用户了解以下内容可能会有所帮助：（1）包含所选词语的句子/文档彼此之间有多相似；（2）根据所选词语，这些句子之间是否存在重叠。这些指标将为用户定义群体提供额外指导。同样，在比较单个实例时，参与者既选择了一般种族类别，也选择了受到歧视的群体。在查看单个实例时展示身份标记的相似性指标，可以为用户提供更多上下文。我们的研究结果支持了此前在贷款申请情境下探讨从业者公平性观点的研究。该研究发现，参与者喜欢查看和比较单个申请的功能，并且应提供更多关于决策边界两侧案例相似性的信息 [46]。

• **纳入非对称反事实指标**此前研究文本分类公平性的工作引入了反事实公平性指标 [24]，用于生成包含不同身份标记的反事实样本，其中文本对某一身份提及刻板印象，却没有对另一身份提及刻板印象。在个体视角示例中，参与者探索了句子，但我们没有观察到任何会针对某个标识符产生非对称反事实的句子。相反，用户测试的是明确无毒词语或明确有毒词语，通过检查模型是否会针对不同标识符改变分类结果，来判断模型是否公平。然而，对于探索文本个体公平性的从业者和工具设计者而言，纳入非对称反事实指标或许能够引导用户测试此类句子：它们对某些身份标记可能有毒，对另一些身份标记则无毒。对于其他那些通过身份术语在性别/种族/宗教/性取向等人口统计维度上表现出偏差的自然语言处理任务，这一指标也会有所帮助。

• **过度预测/低估预测的相关风险**用户在查看准确率视图后进行选择时，反复提到的一个原因是：将有毒帖子“过度预测”为有毒，而不是“低估预测”有毒帖子。尽管许多研究都侧重于使用 F值进行评估 [50]，，但在某些情况下，侧重于真正负率（TNR）而非真正率（TPR）（或反之）是合理的。例如，假阳性相关的成本和风险很高时，即把合法或重要的电子邮件过滤为垃圾邮件 [36]。相反，在某些情况下，TPR 优先于 TNR，例如欺诈检测，甚至疾病检测 [62]。显而易见，查看准确率视图后，不同参与者对于优先考虑无毒内容准确率（TNR）还是有毒内容准确率（TPR）持有不同看法。优先考虑无毒内容准确率的原因是避免惩罚“优质用户”发布的内容；也有人表示，在这种情况下，与低估预测有毒帖子相比，过度预测有毒帖子更为重要。由于在各种情境下，过度预测可能比低估预测更可取，反之亦然，因此必须考虑整个流程。例如，对数据进行分类后，被错误分类为有毒的内容是仅被标记但仍保留在数据中，还是会被彻底删除？反过来，有毒但未被分类器识别的内容又会如何处理？这些内容将面向的群体是否会因此面临风险？从业者应参与决定帖子的处理结果，并与所有相关利益相关者共同决定采用哪种模型。

• **提供基于模拟的方法** 在我们的工具中，用户可以创建自己的群体、句子以及这些句子中的身份词元。然而，当我们让用户在探索页面上对自己偏好的交互进行排序时，自定义交互的排名都低于预填充示例。一位用户表示：“我很难想出句子。我尝试的那句没有被任何模型评为有毒，所以并没有什么帮助。”先前的研究表明，可以模拟对话轨迹 [30] 来支持从业者评估公平性 [28]。我们的发现支持这一观点：用户喜欢先看到几个测试示例，再构思自己的示例。

• **提供额外的群体公平性指标** 在任务 2 中，有 42% 的用户依据群体视图指标选择了模型 A，这一比例很高，说明仅向用户展示群体准确率指标，并不足以充分揭示群体公平性违规。模型 A 在提及少数群体身份时会过度预测毒性。例如，在图 2 中，模型 A 的无害准确率极低，仅为 59.5%，而有毒内容准确率为 94.2%，这表明对于包含相应群体身份标记的文本，模型倾向于过度预测毒性（另请注意，其准确率和平衡准确率指标也较低）。从回答中可以看出，选择模型 A 的参与者出发点是好的（希望捕捉所有有毒帖子），但这一选择也导致模型将包含指代少数群体词语的文本误分类为有毒。缓解这一问题的一种方法是纳入额外的群体公平性指标，例如均等化赔率违规 [26]。任何监督学习场景中都可能出现类似问题，例如，在刑事司法领域，仅关注

模型识别未来累犯的能力，可能会导致人们偏好一个对黑人被告存在偏差的模型 [2]。

- **团队多样性** 我们分析用户在探索视图中选择的令牌身份类型时，发现一个新兴主题：用户会使用源自个人经历的词语。这种策略——根据个人经历探索观察到的词语——凸显了团队多样性的重要性。背景各异（性取向、宗教、种族）的人能够探索数据的不同方面，并发现潜在偏差。招募多元化团队有助于缓解因“盲点”带来的挑战，并确定在考虑不同机器学习应用的公平性时应关注哪些子群体 [28]。虽然这本身并非设计建议，但我们创建的许多人工智能系统都直接取决于我们收集、预处理、进行特征工程并评估的数据。由多元化的机器学习从业者组成的团队评估模型，可能有助于选出更公平的模型。

6 局限性

尽管我们的设计建议未必都能推广到公平性感知的所有其他领域，但其中许多建议适用于那些会通过身份术语在人口统计维度上表现出偏差的自然语言处理技术。此外，我们希望探索机器学习从业者在使用同时呈现群体公平性和个体公平性的工具时采用了哪些策略。我们提供了一些示例作为指导，也承认这些预填充示例可能影响了用户探索的自定义示例类型。不过，在进入探索页面之前的任务（任务 2、任务 3）中，我们尽可能提供了多样化的示例，涵盖不同类型的身份和句子（性取向、宗教、族裔、性别认同）。未来的研究可以在更受控的环境中，考察不同的预填充群体示例和句子示例对用户策略的影响。

在群体视图任务和探索任务中，我们都使用群体标识关键词作为文本内容的替代指标；这是在文本分类中查找群体的一种常见方法 [20, 34, 48, 58]。不过，我们承认，关于某些群体成员的有毒评论可能并不会明确提及他们 [9]。基于先前的研究 [20, 24, 34, 58, 68], ，我们关注人们对包含身份术语的文本分类/错误分类公平性的看法。未来的工作可以考察其他文本分组方法，例如对文本中隐含提及的身份进行分组。

这项工作面临的一项挑战，是用户对毒性的感知具有主观性。个人对何为冒犯性或有毒内容的判断，不仅受呈现给他们的文本语境影响，还会受到个人经历、背景及其他个体因素的影响 [69]。此前的研究揭示了在线内容审核语境下，个体对毒性的感知存在差异 [31, 37, 44]。研究公平性感知十分重要，这有助于了解人机交互研究人员如何设计界面和系统，以兼顾语境的多样性以及评估公平性的个体差异。我们认为，我们的研究结果揭示了如何考虑毒性公平性感知的主观性和语境性。我们认识到，用户对毒性的感知具有主观性，可能受到个人背景及其他因素的影响。虽然我们分析了标注者的毒性评分，但并未调查参与者对毒性是否存在一致的感知。这是未来研究应考虑的一个重要问题。为了提供更准确的基线，可能有必要在规模更大、构成更多样的参与者样本中开展此类调查。

最后，本研究没有探讨查看某种类型的公平性指标会如何影响机器学习从业者对模型得出的结论。用户最初看到的任务旨在作为教程，为他们的实验阶段做准备。未来可以开展一项考虑顺序效应的实验，以确定数据的某种特定视角是否会影响用户可能得出的结论。同样，也可以给予用户更大的自由度，让他们自行决定在探索视图中进行多少次实验后再选择模型，以了解从业者之间探索结果的差异，和/或在特定情境下进行多少次探索才算足够。

7 结论

在这项研究中，我们探究了多种因素：向用户呈现数据的不同方面和视图时，这些因素如何影响用户对各种模型的公平性决策；机器学习从业者采用哪些策略来判断模型是否公平，以及这些策略背后的动机。尽管其他研究考察了用户对公平性工具和指标的看法，但我们的研究首次探索了用户如何与用于审查文本分类的公平性工具进行交互。我们发现，不同的视图会影响用户对数据的理解，以及他们优先考虑哪些指标（即有毒内容准确率／无毒内容准确率／群体公平性／个体公平性）。我们还发现，用户会采用不同的策略来判断模型的公平性，例如，将来自相似类别的身份群体归为一组，或将来自不同类别（种族、性取向、性别）的身份群体归为一组。基于研究结果，我们提出了如何利用前端工具来更好地帮助用户采用适当策略的建议，从而使他们能够更全面地了解数据和模型，最终就模型的公平性作出更明智的决策，并以自己偏好的方式与工具交互。

参考文献

[1] Alekh Agarwal、Alina Beygelzimer、Miroslav Dudík、John Langford 和 Hanna Wallach。2018。公平分类的归约方法。载于 *ICML*。[2] Julia Angwin、Jeff Larson、Surya Mattu 和 Lauren Kirchner。2016。机器偏见。载于 数据与分析伦理。Auerbach Publications，254–264。[3] Ari Ball-Burack、Michelle Seng Ah Lee、Jennifer Cobbe 和 Jatinder Singh。2021。差异化推文处理：缓解有害推文检测中的种族方言偏见。载于 *2021*年*ACM*公平性、问责制与透明度会议论文集。116–128。[4] Solon Barocas、Moritz Hardt 和 Arvind Narayanan。2019。公平性与机器学习。fairmlbook.org。[5] Rachel KE Bellamy、Kuntal Dey、Michael Hind、Samuel C Hoffman、Stephanie Houde、Kalapriya Kannan、Pranay Lohia、Jacquelyn Martino、Sameep Mehta、Aleksandra Mojsilović 等。2019。AI Fairness 360：用于检测和缓解算法偏见的可扩展工具包。*IBM*研究与开发杂志 63，4/5（2019），4–1。[6] Richard Berk、Hoda Heidari、Shahin Jabbari、Michael Kearns 和 Aaron Roth。2021。刑事司法风险评估中的公平性：最新进展。社会学

方法与研究 50，1（2021），3–44。[7] Marianne Bertrand 和 Sendhil Mullainathan。2004。Emily 和 Greg 是否比 Lakisha 和 Jamal 更具就业竞争力？一项关于劳动力市场歧视的现场实验。《美国经济评论》94，4（2004），991–1013。[8] Sarah Bird、Miro Dudík、Richard Edgar、Brandon Horn、Roman Lutz、Vanessa Milan、Mehrnoosh Sameki、Hanna Wallach 和 Kathleen Walker。2020。Fairlearn：用于评估和改进人工智能公平性的工具包。Microsoft，技术报告 MSR-TR-2020-32（2020）。[9] Daniel Borkan、Lucas Dixon、Jeffrey Sorensen、Nithum Thain 和 Lucy Vasserman。2019。用于以真实数据衡量文本分类中非预期偏见的细致指标。载于2019年万维网大会伴随论文集。491–500。[10] Flavio Calmon、Dennis Wei、Bhanukiran Vinzamuri、Karthikeyan Natesan Ramamurthy 和 Kush R Varshney。2017。用于歧视预防的优化预处理。神经信息处理系统进展 30（2017）。[11] Richard J Chen、Tiffany Y Chen、Jana Lipkova、Judy J Wang、Drew FK Williamson、Ming Y Lu、Sharifa Sahai 和 Faisal Mahmood。2021。医学与医疗保健领域人工智能的算法公平性。arXiv 预印本 arXiv:2110.00603（2021）。[12] Alexandra Chouldechova。2017。具有差别影响的公平预测：对再犯预测工具中偏差的研究。《大数据》5，2（2017），153–163。[13] Alexandra Chouldechova 和 Aaron Roth。2020。机器学习公平性前沿概览。Commun. ACM 63，5（2020），82–89。[14] Andrew Cotter、Heinrich Jiang 和 Karthik Sridharan。2019。用于高效非凸约束优化的双人博弈。载于算法学习理论。PMLR，300–332。[15] Kate Crawford。2017。偏差的难题。NeurIPS 主题演讲。（2017）。[16] Maria De-Arteaga、Alexey Romanov、Hanna Wallach、Jennifer Chayes、Christian Borgs、Alexandra Chouldechova、Sahin Geyik、Krishnaram Kenthapadi 和 Adam Tauman Kalai。2019。简历中的偏差：高风险场景中语义表征偏差的案例研究。载于公平、问责与透明度会议论文集。120–128。[17] Wesley Hanwen Deng、Manish Nagireddy、Michelle Seng Ah Lee、Jatinder Singh、Zhiwei Steven Wu、Kenneth Holstein 和 Haiyi Zhu。2022。探究机器学习从业者如何（尝试）使用公平性工具包。载于2022年ACM公平性、问责制与透明度会议（大韩民国首尔）（FAccT '22）。计算机协会，美国纽约州纽约市，473–484。https://doi.org/10.1145/3531146.3533113[18] Sunipa Dev、Emily Sheng、Jieyu Zhao、Aubrie Amstutz、Jiao Sun、Yu Hou、Mattie Sanseverino、Jiin Kim、Akihiro Nishi、Nanyun Peng 等。2022。关于自然语言处理中偏见与伤害的度量。载于计算语言学协会发现：AACL-IJCNLP 2022。246–267。[19] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。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（2018）。[20] Lucas Dixon、John Li、Jeffrey Sorensen、Nithum Thain 和 Lucy Vasserman。2018。衡量和缓解文本分类中的非预期偏见。载于2018年AAAI/ACM人工智能、伦理与社会会议论文集。67–73。[21] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。2012。通过认知实现公平性。载于 ITCS。[22] Will Fleisher。2021。个体公平性何以公平？载于2021年AAAI/ACM人工智能、伦理与社会会议论文集。480–490。[23] Sorelle A Friedler、Carlos Scheidegger、Suresh Venkatasubramanian、Sonam Choudhary、Evan P Hamilton 和 Derek Roth。2019。机器学习中公平性增强干预的比较研究。载于公平性、问责制与透明度会议论文集。329–338。[24] Sahaj Garg、Vincent Perot、Nicole Limtiaco、Ankur Taly、Ed H Chi 和 Alex Beutel。2019。通过稳健性实现文本分类中的反事实公平性。载于2019年AAAI/ACM人工智能、伦理与社会会议论文集。219–226。[25] Andrew Gaut、Tony Sun、Shirlyn Tang、Yuxin Huang、Jing Qian、Mai ElSherief、Jieyu Zhao、Diba Mirza、Elizabeth Belding、Kai-Wei Chang 等。2019。迈向理解关系抽取中的性别偏见。arXiv 预印本 arXiv:1911.03642（2019）。[26] Moritz Hardt、Eric Price 和 Nati Srebro。2016。监督学习中的机会平等。神经信息处理系统进展 29（2016）。[27] Geert Hofstede。2011。文化维度化：语境中的霍夫斯泰德模型。心理学与文化在线读物 2，1（2011），2307–0919。[28] Kenneth Holstein、Jennifer Wortman Vaughan、Hal Daumé III、Miro Dudik 和 Hanna Wallach。2019。改善机器学习系统中的公平性：行业从业者需要什么？载于2019年CHI计算系统人因会议论文集。1–16。[29] Christina Ilvento。2019。面向个体公平性的度量学习。arXiv 预印本 arXiv:1906.00250（2019）。[30] Alankar Jain、Florian Pecune、Yoichi Matsuyama 和 Justine Cassell。2018。具备社会意识的对话代理的用户模拟器架构。载于第18届智能虚拟代理国际会议论文集。133–140。[31] Jialun Aaron Jiang、Morgan Klaus Scheuerman、Casey Fiesler 和 Jed R Brubaker。2021。理解国际社会对有害内容严重程度的认知（续）《PLOS ONE》16，8（2021），e0256762。[32] Jean-Marie John-Mathews、Dominique Cardon 和 Christine Balagué。2022。从现实到世界：对人工智能公平性的批判性视角。《商业伦理学杂志》（2022），1–15。[33] Jon Kleinberg、Sendhil Mullainathan 和 Manish Raghavan。2016。风险评分公平判定中的固有权衡。arXiv预印本 arXiv:1609.05807（2016）。[34] Pang Wei Koh、Shiori Sagawa、Henrik Marklund、Sang Michael Xie、Marvin Zhang、Akshay Balsubramani、Weihua Hu、Michihiro Yasunaga、Richard Lanas Phillips、Irena Gao 等。2021。Wilds：真实环境中分布偏移的基准。在国际机器学习会议论文集。PMLR，5637–5664。[35] KS Krishnapriya、Vítor Albiero、Kushal Vangara、Michael C King 和 Kevin W Bowyer。2020。与人脸识别准确率因种族和肤色而异相关的问题。《IEEE技术与社会汇刊》1，1（2020），8–20。[36] Owen Kufandirimbwa 和 Richard Gotora。2012。使用人工神经网络（感知器学习规则）进行垃圾邮件检测。《在线物理与环境科学研究期刊》1，2（2012），22–29。[37] Deepak Kumar、Patrick Gage Kelley、Sunny Consolvo、Joshua Mason、Elie Bursztein、Zakir Durumeric、Kurt Thomas 和 Michael Bailey。2021。为多元视角设计有害内容分类。在 SOUPS@USENIX安全研讨会论文集。299–318。[38] Tao Li、Tushar Khot、Daniel Khashabi、Ashish Sabharwal 和 Vivek Srikumar。2020。通过不充分限定的问题揭示刻板印象偏见。arXiv预印本 arXiv:2010.02428（2020）。[39] Michael Madaio、Lisa Egede、Hariharan Subramonyam、Jennifer Wortman Vaughan 和 Hanna Wallach。2022。评估人工智能系统的公平性：人工智能从业者的流程、挑战与支持需求。《ACM人机交互汇刊》6，CSCW1，文章52（2022年4月），26页。https://doi.org/10 1145 3512899 。/ Michael A Madaio、Luke Stark、Jennifer Wortman Vaughan 和 Hanna Wallach[40]。2020。共同设计检查清单，以了解人工智能公平性方面的组织挑战与机遇。在2020年CHI计算系统中的人为因素会议论文集。1–14。[41] Subha Maity、Songkai Xue、Mikhail Yurochkin 和 Yuekai Sun。2021。个体公平性的统计推断。arXiv预印本 arXiv:2103.16714（2021）。[42] Mykola Makhortykh、Aleksandra Urman 和 Roberto Ulloa。2021。检测网络搜索引擎中人工智能视觉表征的种族和性别偏见。在算法偏见与搜索和推荐国际研讨会论文集。Springer，36–50。[43] Keri Mallari、Kori Inkpen、Paul Johns、Sarah Tan、Divya Ramesh 和 Ece Kamar。2020。我看起来像罪犯吗？考察种族呈现如何影响人们对再犯的判断。在2020年CHI计算系统中的人为因素会议论文集。1–13。[44] Brandeis Marshall。2021。内容审核实践中的算法厌黑厌女。海因里希·伯尔基金会欧盟分部（2021）。[45] Debarghya Mukherjee、Mikhail Yurochkin、Moulinath Banerjee 和 Yuekai Sun。2020。从数据中学习个体公平性指标的两种简单方法。在ICML。[46] Yuri Nakao、Lorenzo Strappelli、Simone Stumpf、Aisha Naseer、Daniele Regoli 和 Giulia Del Gamba。2022。迈向负责任的人工智能：探索以人为中心的人工智能用户界面的设计空间，以研究公平性。《国际人机交互期刊》（2022），1–27。[47] Yuri Nakao、Simone Stumpf、Subeida Ahmed、Aisha Naseer 和 Lorenzo Strappelli。2022。迈向让终端用户参与交互式人在回路中的人工智能公平性研究。《ACM交互式智能系统汇刊》（TiiS）（2022）。[48] Ji Ho Park、Jamin Shin 和 Pascale Fung。2018。降低辱骂性语言检测中的性别偏见。在2018年自然语言处理实证方法会议（EMNLP 2018）论文集。[49] Felix Petersen、Debarghya Mukherjee、Yuekai Sun 和 Mikhail Yurochkin。2021。个体公平性的后处理。《神经信息处理系统进展》34（2021），25944–25955。[50] David MW Powers。2020。评估：从精确率、召回率和F度量到ROC、信息度、标记度和相关性。arXiv预印本 arXiv:2010.16061（2020）。[51] Flavien Prost、Nithum Thain 和 Tolga Bolukbasi。2019。通过去偏嵌入降低文本分类中的性别偏见。arXiv预印本 arXiv:1908.02810（2019）。[52] Yada Pruksachatkun、Satyapriya Krishna、Jwala Dhamala、Rahul Gupta 和 Kai-Wei Chang。2021。稳健性是否能提升公平性？采用词替换稳健性方法实现文本分类公平性。在ACL-IJCNLP 2021。https://www.amazon.science/publications/does-robustness-improve-fairness-approaching-fairness-with-word-substitution-robustness-methods-for-text-classification[53] Tim Räz。2022。选区划分操纵个体公平性。arXiv预印本 arXiv:2204.11615（2022）。[54] Brianna Richardson、Jean Garcia-Gathright、Samuel F Way、Jennifer Thom 和 Henriette Cramer。2021。迈向实践中的公平性：用于评估公平机器学习工具包的从业者导向评分标准。在2021年CHI计算系统中的人为因素会议论文集。1–13。

[55] Pedro Saleiro、Benedict Kuester、Loren Hinkson、Jesse London、Abby Stevens、Ari Anisfeld、Kit T Rodolfa 和 Rayid Ghani。2018。Aequitas：偏见与公平性审计工具包。*arXiv*预印本 *arXiv:1811.05577*（2018）。[56] Nripsuta Ani Saxena、Karen Huang、Evan DeFilippis、Goran Radanovic、David C Parkes 和 Yang Liu。2019。公平性定义表现如何？考察公众对算法公平性定义的态度。在*2019*年*AAAI/ACM*人工智能、伦理与社会会议论文集。99–106。[57] Emily Sheng、Kai-Wei Chang、Premkumar Natarajan 和 Nanyun Peng。2019。这位女性的工作是保姆：语言生成中的偏见。*arXiv*预印本 *arXiv:1909.01326*（2019）。[58] Ioana Baldini Soares、Dennis Wei、Karthikeyan Natesan Ramamurthy、Moninder Singh 和 Mikhail Yurochkin。2022。你的公平性可能各不相同：有害文本分类中的预训练语言模型公平性。在计算语言学协会年会。[59] Megha Srivastava、Hoda Heidari 和 Andreas Krause。2019。数学概念与人类对公平性的感知：一种面向机器学习公平性的描述性方法。在第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。2459–2468。[60] Gabriel Stanovsky、Noah A Smith 和 Luke Zettlemoyer。2019。评估机器翻译中的性别偏见。*arXiv*预印本 *arXiv:1906.00591*（2019）。[61] Michael Veale、Max Van Kleek 和 Reuben Binns。2018。高风险公共部门决策中算法支持的公平性与问责制设计需求。在*2018*年*CHI*计算系统中的人为因素会议论文集。1–14。[62] Ivan Vorobyev 和 Anna Krivitskaya。2022。基于分布式树模型中的规则归纳，减少银行反欺诈系统中的假阳性。《计算机与安全》120（2022），102786。[63] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz 等。2019。Huggingface 的 Transformers：最先进的自然语言处理。*arXiv* 预印本 *arXiv:1910.03771*（2019）。[64] Allison Woodruff、Sarah E Fox、Steven Rousso-Schindler 和 Jeffrey Warshaw。2018。对算法公平性认知的定性探索。载于*2018*年*CHI*人因计算系统会议论文集。1–14。[65] Songkai Xue、Mikhail Yurochkin 和 Yuekai Sun。2020。审查机器学习模型中的个体偏见与不公平性。载于*AI&STAT*。[66] Sirui Yao 和 Bert Huang。2017。超越平等：协同过滤的公平性目标。神经信息处理系统进展30（2017）。[67] Mikhail Yurochkin、Amanda Bower 和 Yuekai Sun。2020。利用敏感子空间鲁棒性训练个体公平的机器学习模型。载于国际学习表征会议。[68] Mikhail Yurochkin 和 Yuekai Sun。2021。SenSeI：通过敏感集合不变性实现个体公平性。载于国际学习表征会议。https://github.com/IBM/inFairness [69] Davide Zotti、Andrea Carnaghi、Valentina Piccoli 和 Mauro Bianchi。2019。与学校教职员工应对恐同欺凌行为相关的个体及情境因素。性与社会政策研究16，4（2019），543–558。[70] James Zou 和 Londa Schiebinger。2018。人工智能可能存在性别歧视和种族歧视——是时候让它变得公平了。

A 数据集详情

A.1 数据集统计

表4概述了用于训练本研究所用三个模型的数据集的总体统计数据。总体而言，数据样本约有180万个，其中只有144,334个有毒样本（8%），因此任务类别不平衡。该数据集中有247,581个数据样本（13.71%）包含50个身份术语中的一个。其中，11.9%（即29,509个数据样本）被标记为有毒，其余则被标记为无毒。表5列出了本研究所用50个身份术语各自对应的有毒和无毒数据样本数量。

属性	数量
训练样本数量	1,804,874
有毒数据样本数量	144,334
无毒评论数量	1,660,540
含有身份术语的数据样本数	247,581
含有身份术语的有毒数据样本数	29,509
含有身份术语的无毒数据样本数	2,18,072

表4：毒性分类数据集的汇总统计信息

A.2 毒性标注详情

为将数据样本标注为有毒或无毒，数据集管理者使用人工标注员将评论标注为 0（无毒）和 1（有毒）。

每个数据样本均由多名标注者进行标注，其所赋标签的平均值被记录为最终的毒性置信度分数。遵循先前的工作 [24, 34, 68], ，我们将毒性置信度分数大于或等于 0.5 的数据样本标记为有毒，其他样本则标记为无毒。在表 6 中，我们列出了数据样本的标注员数量的各个分位数，以及数据集中的平均毒性分数。图 10 展示了平均标注员置信度分数的直方图。我们注意到，标注者在将评论标注为无毒时表现出更多的一致性。

B 模型训练详情

为了训练模型 A（标准模型）和模型 B（个体公平模型），我们在表7定义的取值网格上对超参数进行了优化。对于每种超参数设置，我们都会计算验证集上的准确率、平衡准确率和预测一致性，并选择平衡准确率和预测一致性得分较好的最终模型。我们在图11和平行坐标图图12中展示了使用不同超参数值训练的各个模型的性能。为了训练模型 B（群体公平模型），我们尝试了对各群体性能施加多种类型的约束（假阳性率、假阴性率和平衡准确率），并发现将每个群体的平衡准确率约束为 $\geq$ 95% 的效果最好，如第3.5.3节所述。此外，我们还在表7定义的取值网格上对学习率进行了优化。

身份术语	有毒计数	无毒计数
非洲人	433	3116
非裔美国人	70	481
美国人	3893	34267
亚裔	191	1905
双性恋者	22	88
黑人	4470	14830
盲人	741	4330
佛教徒	35	309
加拿大人	1993	30979
天主教徒	908	13935
华人	745	8337
基督徒	1284	8496
deaf	123	766
老年人	119	2058
欧洲人	297	4251
女性	717	5892
gay	1860	4390
异性恋	127	488
西班牙裔	138	793
同性恋者	368	999
印度人	246	2641
日本人	176	2313
犹太人	403	3002
拉丁裔女性	6	50
拉丁裔男性	76	434

身份术语	有毒计数	无毒计数
拉丁裔	0	12
女同性恋者	156	325
lgbt	181	1303
LGBTQ	81	540
male	1049	6924
墨西哥人	287	2124
中年人	16	207
中东人	44	394
千禧一代	7	69
穆斯林	2274	8257
非二元性别者	0	2
old	3778	35558
年长者	225	4218
瘫痪者	7	164
新教徒	42	1151
酷儿	24	105
sikh	23	265
异性恋	741	6025
道教徒	0	5
青少年	90	637
跨性别者	157	1426
跨性别者	296	1264
白人	8641	29767
年轻人	1637	17938
较年轻者	167	2763

表 5：数据量 a 包含相应身份术语且被标注为有毒和无毒的样本示例 s.

属性	数量
最小标注者数量	3
中位数标注者数量	4
平均标注者数量	8.78
最大标注者数量	4936
毒性评分均值（标准差）	0.103 ± 0.197

表 6：数据集中将样本标记为有毒或无毒的标注者数量的各个分位数。还列出了数据集中所有数据样本的平均毒性评分。

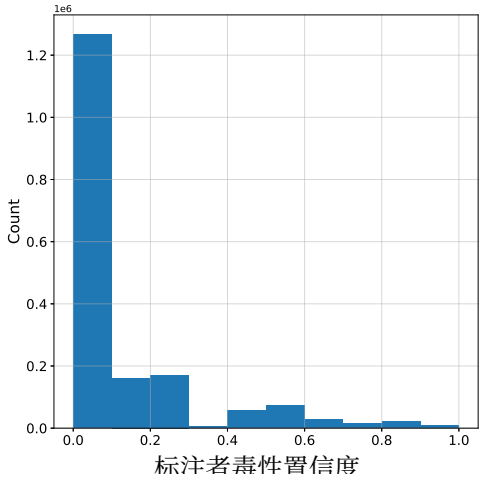


图 10：数据样本被标注为有毒或无毒时的标注者置信度评分直方图。

超参数	模型 A	模型 B	模型 C
批量大小	{512, 1024}	{512, 1024}	{1024}
学习率	{5e-5, 1e-5, 5e-4, 1e-4, 5e-3, 1e-3}	{1e-5, 1e-4, 1e-3}	{0.03, 0.035, 0.04}
rho	N.A.	{1.0, 5.0, 10.0, 50.0}	N.A.
eps	N.A.	{0.1, 1.0, 4.0, 7.0, 10.0}	N.A.
审计器学习率	N.A.	{0.1, 0.01}	N.A.
奇异值分解组件数量	N.A.	{50, 100}	N.A.

表7：模型 A、B 和 C 的超参数网格搜索范围。我们根据训练后模型的平衡准确率和预测一致性对其进行评估，并选择在这两个指标上均表现良好的模型。

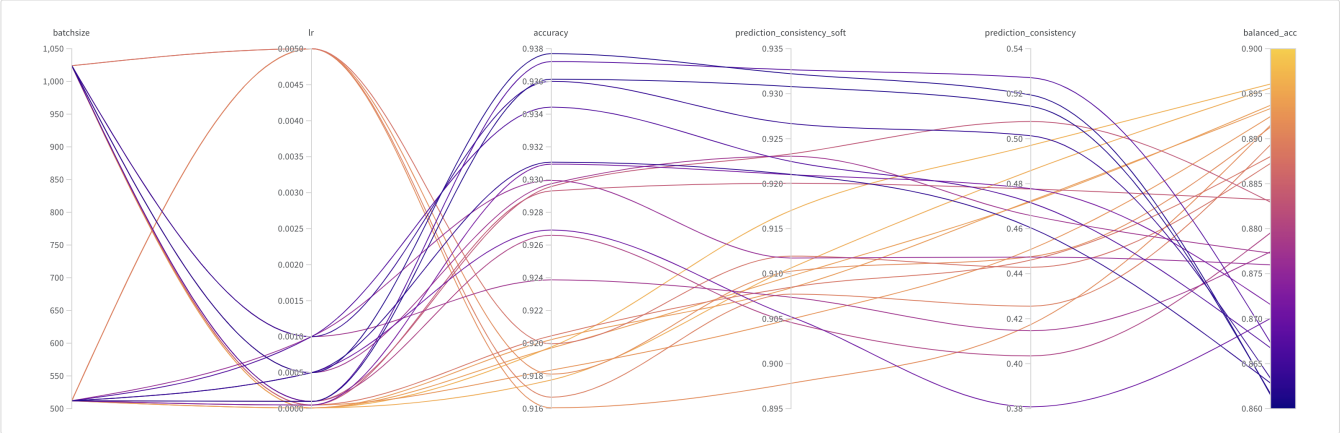


图11：训练模型 A 的超参数扫描范围。我们改变模型 A 的批量大小和学习率值，并测量其准确率、软预测一致性、预测一致性和平衡准确率。我们根据最终模型在预测一致性和平衡准确率上的性能进行选择。

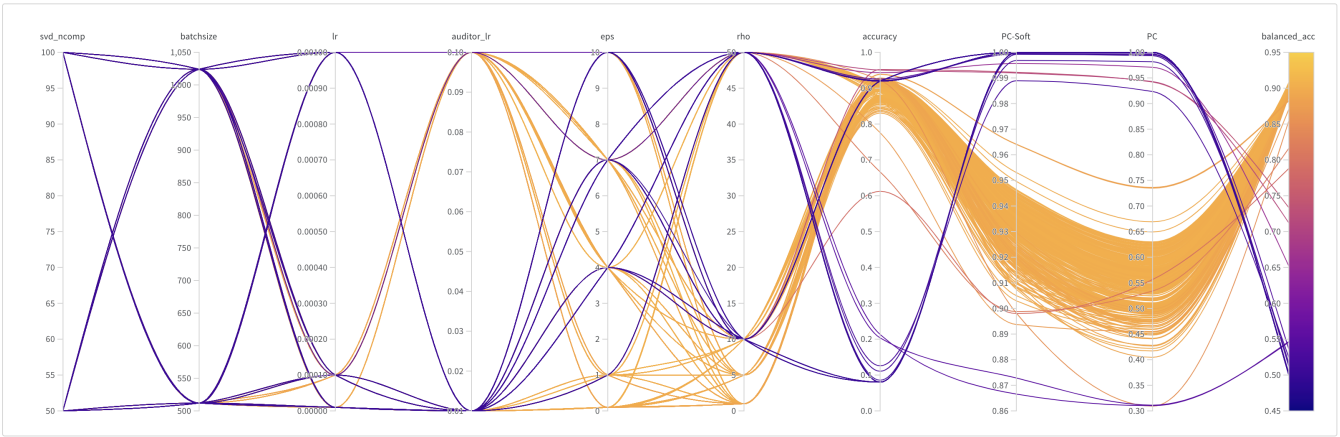


图12：训练模型 B 的超参数扫描范围。我们改变模型 B 的六个超参数（见表7），并测量其准确率、软预测一致性、预测一致性和平衡准确率。我们根据最终模型在预测一致性和平衡准确率上的性能进行选择。