arXiv:2303.03124v2 [cs.CL] 2 Oct 2023

IFAN：面向人类与自然语言处理模型的可解释性交互框架

Edoardo Mosca、DarynaDementieva、Tohid Ebrahim Ajdari、Maximilian Kummeth、Kirill Gringauz、Yutong Zhou 和Georg Groh慕尼黑工业大学，信息学系，德国
{name.surname}@tum.degrohg@in.tum.de

摘要

可解释性与人工监督是将复杂自然语言处理模型部署到现实应用中的基本支柱。然而，应用可解释性和人在回路方法需要具备技术能力。尽管现有的模型理解与分析工具包已经相当丰富，但可用于整合人类反馈的选项仍然有限。我们提出 IFAN，这是一个支持基于实时解释与自然语言处理模型交互的框架。通过 IFAN 的用户界面，用户可以针对选定的模型解释提供反馈，随后通过适配器层整合这些反馈，使模型与人类推理依据保持一致。我们证明，该系统能够在对性能影响极小的情况下，有效降低仇恨言论分类器中的偏见。IFAN 还提供可视化管理系统和 API，用于管理模型（及数据集）以及控制访问权限。演示版本已上线ifan.ml。

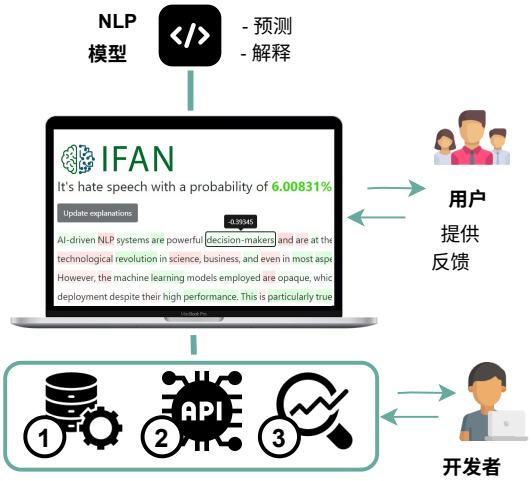


图 1：IFAN 简介。该界面允许自然语言处理模型与用户通过预测结果、解释和反馈进行交互。IFAN 还为开发者提供：（1）模型和数据集管理器；（2）模型 API 访问权限；以及（3）关于模型的报告。

1 引言

随着自然语言处理（NLP）系统的性能不断提升，它们正日益被应用于现实应用中（Khurana 等人，2022）。大语言模型（LLM）——例如 GPT-3（Brown 等人，2020）、BLOOM（Scao 等人，2022a）和 T5（Raffel 等人，2020）——无疑是近年来该领域进步的主要推动者。它们能够大幅超越以往的解决方案，同时可直接应用于任何自然语言处理任务。

然而，鉴于此类架构的黑箱特性，人们对此仍有强烈顾虑（Madsen 等人，2022；Mosca 等人，2022a）。事实上，其规模庞大、复杂度高，在透明度、问责制和人工监督方面构成了重大缺陷。除伦理考量外，欧盟的法律准则如今也明确将这些可解释性因素列为所有已部署人工智能系统的必要条件（欧盟委员会，2020）。

因此，可解释人工智能（XAI）（Arrieta 等人，2020；Mosca 等人，2022b）以及人在回路（HitL）机器学习（Monarch，2021年）领域的研究日益兴盛，并不断提出旨在缓解当前可解释性不足问题的解决方案。尤为值得注意的是，近期文献中涌现出许多用于分析、理解和改进复杂自然语言处理模型的工具包和框架（Wallace 等人，2019；Liu 等人，2021年）。其中一些甚至为缺乏所需技术能力的利益相关者提供了低代码界面。尽管如此，目前收集人类推理依据并将其作为反馈提供给模型的方式仍然有限。

我们提出了 IFAN，这是一种新颖的低代码至无代码框架，可通过解释与自然语言处理模型进行实时交互。我们的贡献可总结如下：

(1)IFAN 提供了一个界面，供用户对选定的模型解释提供反馈，

然后通过参数高效的适配器层将反馈整合进模型。

(2) 我们的在线平台还提供可视化管理系统和API，用于管理模型、数据集和用户，以及相应的访问权限。

(3) 我们展示了该框架在对仇恨言论分类器进行去偏方面的效果，并提出了一种反馈再平衡步骤，以缓解模型在多次更新中的遗忘问题。

IFAN 的演示可通过 ifan.ml 访问[1]，相关文档也可在此查阅。[2] 使用登录凭据即可获得完整访问权限，我们可应要求提供凭据。在线还可观看补充视频展示[3]。

2 相关工作

2.1 结合模型解释的人在回路

人在回路（HitL）机器学习研究如何利用人类反馈持续改进模型（Monarch，2021年）。虽然大量人在回路文献关注以标签为核心的反馈，例如主动学习，但近期研究开始探索如何利用解释来提供更详尽的人类推理依据（Lertvit-tayakumjorn andToni，2021年）。

将经典的人在回路（Wang et al.，2021年）与解释相结合，以构建用于模型的人类反馈（Han et al.，2020），被称为基于解释的人类调试（EBHD）（Lertvittayakumjorn and Toni，2021年）。典型示例包括Ray et al.（2019）、Selvaraju etal.（2019）和Stroutet al.（2019），这些研究表明，在迭代地向模型提供人类推理依据时，模型的性能和可解释性均有所提升。

一种更加专注于自然语言处理的基于解释的人类调试方法是Yao 等人 (2021年) 提出的方法，作者利用解释来调试和改进两个 Transformer 模型实例——BERT (Devlin 等人，2019) 和 RoBERTa (Liu 等人，2019)。具体而言，研究者向人类提供不同粒度层级的词语显著性解释，而人类则以自然语言提出建议。标注员的反馈会被转换为一阶逻辑规则，随后用于引导模型利用新样本进行学习。

[1]https://ifan.ml [2] https://ifan.ml/documentation [3] https://youtu.be/EzC6HI3JwaQ

2.2 交互式自然语言处理分析平台

近期文献中，我们可以看到许多在软件和数字工具包方面的突出成果，可用于分析和解释自然语言处理模型 (Wallace 等人，2019；Hoover 等人，2020)，也有研究通过参数高效微调进一步改进这些模型 (Beck 等人，2022)。

例如，Liu et al.（2021年）提出了EX-PLAINABOARD，这是一个面向自然语言处理模型、以可解释性为重点的交互式排行榜。更具体地说，它允许研究人员对给定模型的优势和劣势进行诊断分析，比较不同架构，并深入分析预测以及反复出现的模型错误。类似地，语言可解释性工具由 Tenney et al.（2020）开发，是一个用于可视化和理解自然语言处理模型的开源平台和 API。具体而言，它提供了一个基于浏览器的界面，集成本地解释和反事实示例，以实现模型可解释性和错误分析。

最后，Beck 等人（2022）发布了ADAPTER-HUB PLAYGROUND，这是一个使用语言模型进行少样本学习的无代码平台。具体而言，作者构建了一个直观的界面，用户可以在多个自然语言任务上轻松完成复杂自然语言处理模型的预测和训练。

3 IFAN

面向人工与**I**自然**F**智能的**A**交互**N**框架（**IFAN**）是一个用于检查和控制文本处理模型的基于 Web 的平台。其主要目标是降低自然语言处理系统的不透明性，并将基于解释的人在回路融入其开发流程。通过我们的界面，利益相关者可以测试并解释模型的行为；当预测或解释中出现异常时，他们可以通过提供反馈，在现场修正这些问题。

平台的主要模块如图 2所示。 **骨干模块**部分包含所有机器学习开发组件——数据集和模型。我们采用 HuggingFace 格式（参见 3.3 和 3.4）（Wolf 等人，2020），并将整个骨干模块封装为 Docker[4] 镜像以供部署。 **用户界面**是平台的可视化组件，所有人机交互都在此进行。在这里，开发者还可以访问额外的可视化资源，以配置有关模型、数据集和用户的详细信息。

[4]https://www.docker.com

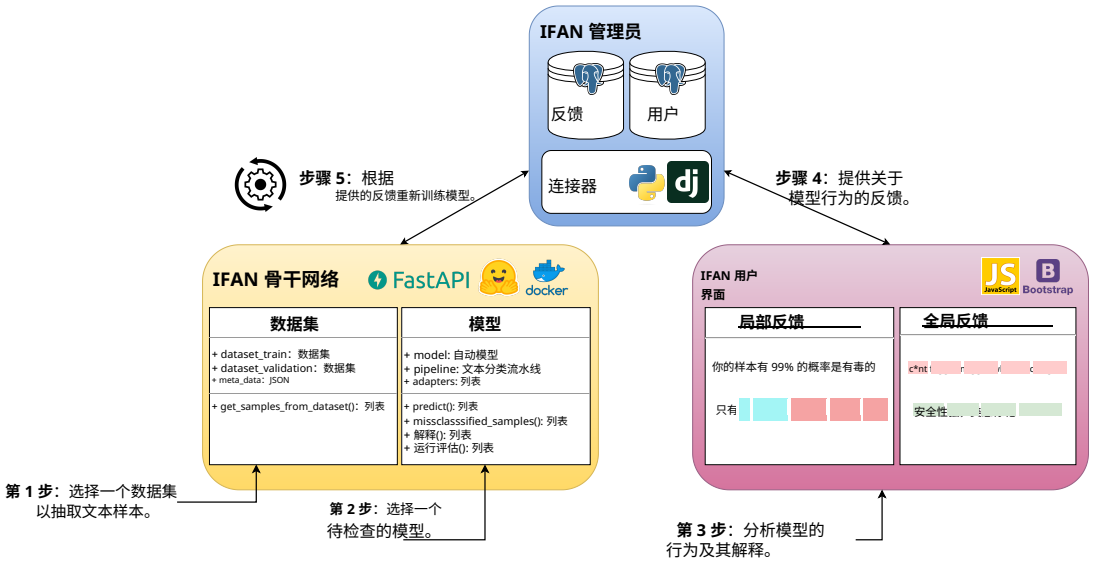


图 2：IFAN 理念的总体架构：（i）用户选择一个数据集或编写自定义输入。（ii）然后，用户可以选择要检查的模型。（iii）借助用户界面，标注员可以检查模型对样本的预测结果以及两种类型的解释——局部解释和全局解释。（iv）如果存在行为异常，标注员可以提供反馈。（v）反馈会被存储，随后用于微调模型。

骨干模块与用户界面之间的连接由 **管理员** 组件管理。所有用户数据和权限，以及接收反馈的样本，都存储在 PostgreSQL[5]数据库实例中。通信通过 Python Django[6], 处理，它整合了用户身份认证、API 调用/响应、状态日志以及骨干模型资源位置等所有功能。在接下来的章节中，我们将更详细地介绍平台的主要组件。

3.1 用户界面

我们的前端使用 Bootstrap[7] 和 JavaScript[8] 构建。目前，用户界面中提供以下页面：

着陆页 用户可在此简要了解 IFAN。我们简要介绍平台的目标、人在回路的概念，以及如何将我们的框架集成到自然语言处理模型的开发中。

文档 此处详细描述了所有用户界面组件，并附有屏幕截图和使用指南。用户可在此找到有关如何配置和使用我们平台的具体说明。

反馈 这是主要交互页面。用户可以在输入样本上运行模型，样本可以来自数据集，也可以由用户自行编写。随后，用户可以加载模型的预测结果和解释，并针对标签和特征相关性提供反馈。

报告 此页面的访问受限（参见 3.2）。开发者可以评估模型在选定数据集上整合反馈前后的表现，并检查误分类样本。

配置 此页面的访问受限（参见3.2）。开发者可在此配置和管理平台。具体而言，用户可以被创建、修改和删除，也可以提升或降低其角色和访问权限。此外，还可以管理模型和数据集，并指定当前启用的模型和数据集。

账户设置 每位获授权的用户都可以查看、编辑、导出和删除其账户数据（符合通用数据保护条例），也可以重置其登录密码。

3.2 用户

该平台将用户分为三个层级： 开发人员、标注员和未经授权的用户（表1）。

未授权用户不持有登录凭证，因此只能有限地访问平台。他们可以查看模型预测结果和解释，但其反馈不会被纳入考虑。

[5]https://www.postgresql.org [6] https://www.djangoproject.com [7] https://getbootstrap.com [8] https://www.javascript.com

	Dev	标注员	未经授权
分类和解释	✔	✔	✔
智能样本选择	✔	✔	✘
反馈	✔	✔	✘
主动配置	✔	✘	✘
模型报告与误分类样本	✔	✘	✘
新模型与数据集上传	✔	✘	✘
新用户创建	✔	✘	✘

表 1：IFAN 的不同访问级别。

普通用户（或标注员）可通过凭证进行识别，因此能够主动参与模型的使用。在一次人在回路迭代过程中，他们可以使用反馈页面中的预配置数据集和模型，在文本样本上测试模型、查看解释，并在需要时提供反馈。

开发者拥有完整访问权限，可以配置平台的所有方面。具体来说，他们可以访问报告和配置页面（参见3.1），因此可以管理与用户、角色、API 访问权限、模型和数据集相关的所有内容。

3.3 数据集

在探索模型的行为之前，应通过配置页面指定当前数据集（参见3.1）。模型测试所用的文本示例从该数据集中抽取。

数据集	**简要说明**
HateXplain (Mathew 等人，2021年)	用于仇恨言论分类的数据集，分为三类，适用于仇恨类型检测、目标社群分类，以及判定依据。
GYAFC (Rao and Tetreault，2018)	正式程度检测数据集对应二分类分类：正式和非正式。

表2：IFAN 测试的数据集示例。

我们使用 HuggingFace Datasets 库，遵循标准格式[9]。我们强烈建议与平台交互的开发者在上传新数据集并将其提供给界面时遵循此标准。表格 2 展示了平台上已有数据集的两个示例。

3.4 模型

与数据集类似，应通过配置页面指定活动模型，并遵循 HuggingFace Models 标准。[10]

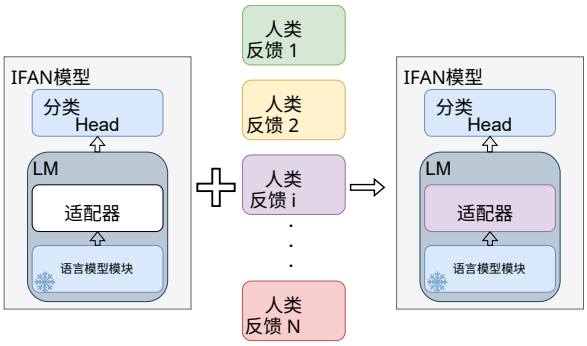


图 3：集成到 IFAN 中的模型所采用的架构：新增可基于所提供的人类反馈进行训练的适配器层。

为将反馈纳入我们的模型，我们采用适配器层（Houlsby 等人，2019），这是一种参数高效微调技术。图 3 概述了所使用的架构。适配器被集成在每个语言模型单元（例如 Transformer 模块）之上，并利用人类反馈进行训练，同时我们冻结所有其他模型权重。也可以禁用适配器，以恢复模型的原始状态。

3.5 解释与反馈机制

用户可通过反馈页面在当前数据集上评估当前模型。他们可以通过三种方式输入文本：i）自行创建一个文本样本；若已获授权：ii）从当前数据集中随机抽取文本；iii）从当前数据集的测试部分随机抽取一段被错误分类的文本。用户将获得分类结果和模型置信度。他们可以评估结果并纠正任何错误分类。

为了进一步考察模型的行为，我们提供两种解释——局部解释和全局解释。对于文本样本的局部解释，我们会显示与每个输出类别相关的特征（图4）。我们使用 LIME 框架为特征赋分（Ribeiro 等人，2016），并且为了过滤弱相关性，只将分数高于阈值 $\theta = 0.1$的词元标记为相关特征。在全局层面，我们会列出每个输出类别中影响力最大的 1 元语法。检查这些内容，可以帮助提取有关模型关注哪些关键词和模式的见解

[9]https://huggingface.co/docs/datasets/index

[10]https://huggingface.co/docs/transformers/main_类/模型

。在数据集层面，模型会关注这些关键词和模式。对于数据集中出现的所有 1 元语法，我们都会计算相应的分类分数，并在页面上显示得分最高的词元。

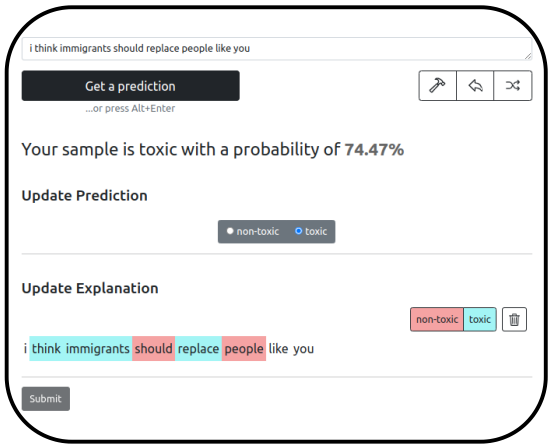


图 4：标注员可以在反馈页面上获取的结果和局部解释示例。

标注员可以轻松编辑高亮显示的词元，并将更新后的解释作为反馈提交。我们会保存结果——即高亮显示的相关部分——并用其微调适配器层。冻结其他所有模型权重，可以最大限度地降低反馈步骤的计算开销。

关于微调过程，直接使用高亮显示的反馈文本微调适配器，会导致原始模型性能显著下降。我们提出将反馈与原始样本混合，以减轻这一影响；这样既能有效整合反馈，又能减少模型遗忘（更多细节见4 ）。

3.6 主干网络 API

我们开放主干模块的 API，以提供所有必要的数据集/模型管理功能。这些功能为开展涉及模型评估、解释和反馈的其他实验提供了高级界面。该 API 基于 Python 框架 FastAPI 构建[11], 更多详情请见附录A。

4 案例研究

我们开展了一项案例研究，以检验 IFAN 的适用性。我们基于 HateXplain 数据集选择了仇恨言论检测任务（Mathew 等人，2021年）。实验的目标是利用我们的框架对给定的仇恨言论检测器进行去偏。

首先，我们修改了原始数据集，用于二元分类任务——“有毒”和“无毒”。我们选择犹太子群体作为去偏过程的目标。我们对 BERT（Devlin 等人，2019）[12] 进行了微调，并向其提供反馈。附录D和E 分别提供了使用 BLOOM 和其他大语言模型开展的补充实验。

我们标注24 个随机误分类样本，12 个具有最高置信度和12 个具有最低置信度分数的样本（参见附录C.1）。我们邀请了3 名标注员参与标注过程。标注员修改过的 n 元语法会被保存，并用于为适配器创建新的训练数据集。因此，我们收集了40个经过标注的 n 元语法，并重复使用它们以获得120 个训练样本。为了完成新训练数据集的创建，我们使用从 HateXplain 数据集中随机选取的500 个原始样本（250 个有毒样本，250 个无毒样本）对这些样本进行了平衡。

模型	Pr	Re	F1	Pr_J
BERT（基线）	0.80	**0.78**	**0.79**	0.95
最高置信度误分类样本				
BERT+反馈（未平衡）	0.34	0.28	0.31	0.82
BERT+反馈（平衡）	0.78	0.80	**0.79**	**0.97**
置信度最低的误分类样本				
BERT+反馈（未平衡）	**0.83**	0.73	0.78	0.96
BERT+反馈（平衡）	0.79	**0.78**	0.78	0.96

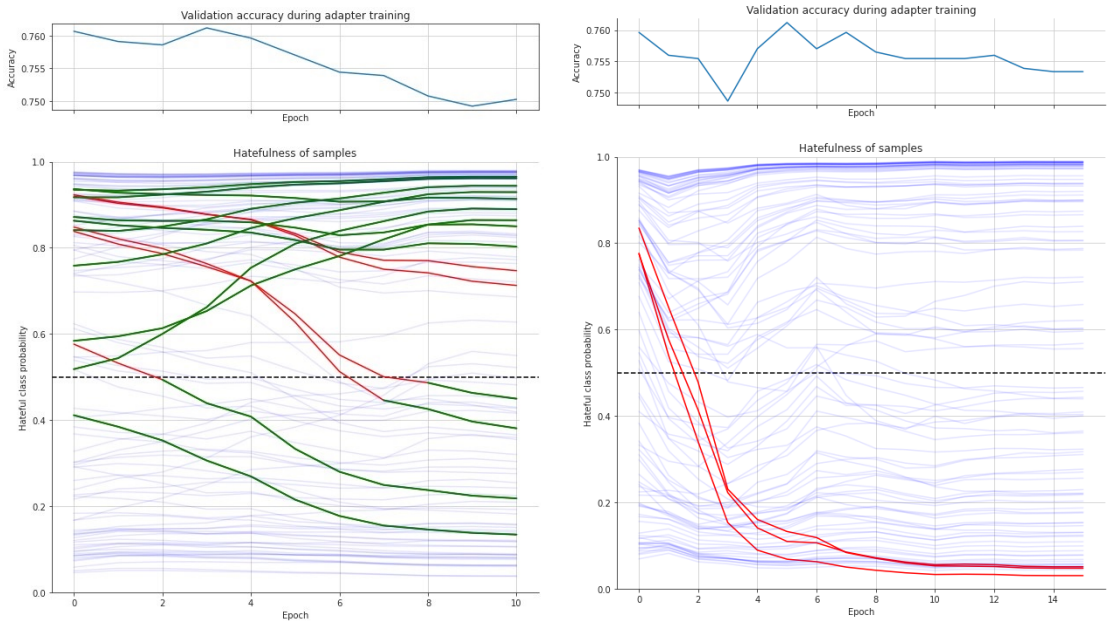
表3：案例研究结果：仇恨言论分类模型去偏。我们比较了不同的反馈整合策略。Pr_J 表示犹太人目标群体的精确率。

结果见表3。我们观察到，非平衡训练数据集仅包含针对最有把握的错误分类样本的反馈，导致性能显著下降。虽然加入针对置信度最低样本的反馈使整体 F1分数略有下降，但基于平衡反馈进行适配器训练提高了犹太人目标群体的精确率。

图 5 展示了利用收集到的反馈进行微调时检测器的变化。重新平衡反馈后，只有经过修改的样本发生了显著变化，而原始文本上的性能仅受到轻微影响。关于在非平衡反馈和平衡反馈上进行微调的详细比较见附录 C.2）。

[11]https://fastapi.tiangolo.com

[12]https://huggingface.co/google/bert_uncased_L-2_H- 128 A-2_



（a）训练包含犹太子群体样本的反馈。（b）使用带有“jewish”关键词的反馈样本进行训练s.

图5：随着模型利用人类反馈进行微调，样本置信度的变化。使用 IFAN 平台的领域案例结果表明，我们可以观察到，在两项采用平衡训练数据的实验中，整体模型的性能仅略有变化，而模型在犹太人目标群体上的行为得到了改善。

5 局限性与未来工作

目前，我们的反馈系统仅适用于序列到类别格式的应用。通过大语言模型提示将平台扩展到更多任务的工作目前正在进行中（见附录 E）。

同时，我们目前提供的解释、反馈和管理选项有限，并计划在近期增加这些选项。我们已开展了一项小规模用户研究（附录B），以收集有关该平台的反馈并提高其用户友好性。我们的目标是继续通过与开发者和外行用户进行试验，迭代开发新功能。

最后，我们的实验尚未显示性能与反馈超参数之间相关性的明确趋势。事实上，还必须开展进一步的研究和试验，以确定反馈样本数量、微调轮次以及重新平衡比率的最优选择。

6 结论

本研究提出了 IFAN，这是一个专注于自然语言处理模型与人工标注员之间实时、基于解释的交互框架。我们的贡献源于现有工具在解释和控制自然语言处理模型方面选择有限。

IFAN 由三个主要单元组成。**骨干模块** 整合了所有机器学习流水线，并提供了一个便于访问的 API。**用户界面**由着陆页、文档、反馈、报告和配置组成，提供直观的可视化组件，便于与模型交互。最后，**管理员** 负责管理前两个组件之间的连接。

此外，我们还引入了一种反馈机制，利用适配器层在下游任务上高效、迭代地微调模型。实验表明，该框架能够有效地对仇恨言论分类器进行去偏，同时仅造成极小的性能损失。

我们认为，IFAN 是迈向自然语言处理模型可解释、可控部署的重要一步，使不具备技术专长的用户也能与已部署的自然语言处理系统交互并提供反馈。关于未来工作，我们优先考虑将该框架扩展到更多自然语言处理任务，并集成更多模型分析功能和反馈机制。

致谢

我们感谢 Natália Souza Soares 和 Ashish J ha 作出的宝贵贡献。本文得到德国联邦教育与研究部（BMBF，资助项目01IS17049）的资助。

伦理考量

在这项工作中，我们展示了仇恨言论模型去偏见的实验。仇恨言论检测任务需要投入大量关注，才能得到公平的结果。目前仍存在的一个问题是偏见，尤其是因偏见而针对少数群体的偏见。我们旨在展示一个示例，说明如何针对某些目标种族群体对模型进行去偏见处理。我们希望，通过审慎选择标注员并提供反馈，我们提出的平台能够有效调整自然语言处理模型，使其适应多元化的世界。

基于这些原因，我们也认为，现代自然语言处理模型和系统的可解释性与可控性，是实现其合乎伦理且安全部署的基石（欧盟委员会，2020）。本研究旨在对这两个方面产生积极影响：它提供了一种解释模型并向其提供反馈的工具。通过降低与自然语言处理系统交互所需的技术能力门槛，我们希望促进人们提供有价值的人类理由，从而影响复杂模型。

确保人类反馈的高质量颇具挑战性（Al Kuwatly 等人，2020），而让模型受到外部影响，可能会被对抗性代理利用（Mosca 等人，2022a）。尤其是在标注员人数很少的情况下，少数人可能会对模型产生重大影响。像 IFAN 这样严格的访问权限管理系统已经能够缓解这些问题。我们认为，增加安全功能并追踪标注员的影响，是未来工作的关键，有助于提升系统的可信度。

以往研究指出，与性能不佳的模型和设计糟糕的界面交互时，用户可能会感到气馁和沮丧，这也会影响反馈质量（Lertvittayakumjornand Toni，2021年）。在开发过程中融入用户研究，有助于解决这一问题，从而设计出更直观的界面并改善整体用户体验。

在光谱的另一端，貌似合理的解释可能使人类高估模型的能力，并让他们信任尚未准备好部署的系统。在这种情况下，为用户提供更多样且互补的解释（Madsenet al.，2022），并为开发者提供全面的模型报告，是为了更完整地呈现待部署模型而设定的核心目标。

参考文献

Hala Al Kuwatly、Maximilian Wich 和 Georg Groh。2020 年。根据标注员的人口统计特征识别并衡量标注员偏见。载于第四届在线虐待与伤害研讨会论文集，第 184–190 页，在线。计算语言学协会。

Alejandro Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador García、Sergio Gil-López、Daniel Molina、Richard Benjamins 等。2020。可解释人工智能（XAI）：迈向负责任人工智能的概念、分类体系、机遇与挑战。*Information Fusion*，58：82–115。

Tilman Beck、Bela Bohlender、Christina Viehmann、Vincent Hane、Yanik Adamson、Jaber Khuri、Jonas Brossmann、Jonas Pfeiffer 和 Iryna Gurevych。2022年。AdapterHub 游乐场：借助适配器实现简单灵活的小样本学习。载于计算语言学协会第*60*届年会论文集：系统演示，第61–75页，爱尔兰都柏林。计算语言学协会。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。收录于*2019* 年计算语言学协会北美分会年会论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 4171–4186 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

欧盟委员会。2020。人工智能白皮书：欧洲追求卓越与信任的路径。*Com(2020)65Final*。

Xiaochuang Han、Byron C. Wallace 和 YuliaTsvetkov。2020。通过影响函数解释黑箱预测并揭示数据伪影。载于第*58*届计算语言学协会年会论文集，第5553–5563页，在线。计算语言学协会。

Benjamin Hoover、Hendrik Strobelt 和 Sebastian Gehrmann。2020。exBERT：用于探索变换器模型中学习到的表示的可视化分析工具。载于第*58*届计算语言学协会年会论文集：系统演示，第187–196页，在线。计算语言学协会。

Neil Houlsby、Andrei Giurgiu、Stanislaw J astrzebski、Bruna Morrone、Quentin de Laroussilhe、Andrea Ges- mundo、Mona Attariyan 和 Sylvain Gelly。2019。面向自然语言处理的参数高效迁移学习。载于第*36*届国际机器学习会议论文集，*ICML 2019*，*2019*年*6*月*9*日至*15*日，美国加利福尼亚州长滩，机器学习研究论文集第97卷机器学习研究论文集，第2790–2799页。PMLR。

Diksha Khurana、Aditya Koli、Kiran Khatter 和 Sukhdev Singh。2022年。自然语言处理：最新进展、当前趋势与挑战。多媒体工具与应用，第 1–32 页。

Piyawat Lertvittayakumjorn 和 Francesca Toni。2021年。基于解释的自然语言处理模型人类调试：综述。计算语言学协会会刊，9:1508–1528。

Pengfei Liu、Jinlan Fu、Yang Xiao、Weizhe Yuan、Shuaichen Chang、Junqi Dai、Yixin Liu、Zihuiwen Ye 和 Graham Neubig。2021年。EXPLAINABOARD：面向自然语言处理的可解释排行榜。载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集：系统演示，第280–289页，在线。计算语言学协会。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019 年。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Andreas Madsen、Siva Reddy 和 Sarath Chandar。2022 年。神经自然语言处理的事后可解释性：综述。*ACM* 计算综述，55(8)。

Binny Mathew、Punyajoy Saha、Seid Muhie Yimam、Chris Biemann、Pawan Goyal 和 Animesh Mukher- jee。2021 年。HateXplain：用于可解释仇恨言论检测的基准数据集。载于第三十五届*AAAI*人工智能会议、*AAAI 2021*、第三十三届人工智能创新应用会议、*IAAI 2021*、第十一届教育人工智能进展研讨会、*EAAI 2021*、线上活动，*2021*年*2*月*2*日至*9*日，第 14867–14875 页。AAAI出版社。

Robert Munro Monarch. 2021年。人在回路机器学习：面向以人为中心的人工智能的主动学习与标注。Simon and Schuster。

Edoardo Mosca、Shreyash Agarwal、Javier Rando Ramírez 和 Georg Groh。2022a。“这反应很可疑！”：通过解读 logits变化 检测自然语言处理对抗攻击。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 7806–7816 页，都柏林，爱尔兰。计算语言学协会。

Edoardo Mosca、Ferenc Szigeti、StellaTragianni、Daniel Gallagher 和 Georg Groh。2022b。基于 SHAP 的解释方法：自然语言处理可解释性综述。载于第 *29* 届国际计算语言学会议论文集，第 4593–4603 页，庆州，大韩民国。国际计算语言学委员会。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020。探索统一文本到文本转换器迁移学习的极限。《机器学习研究杂志》，21:140:1–140:67。

Sudha Rao 和 Joel Tetreault。2018。尊敬的先生或女士：谨介绍 GYAFC 数据集：正式性风格迁移的语料库、基准和评估指标。载于*2018*年北美计算语言学协会年会论文集：人类语言技术，第*1*卷（长篇论文），第129–140页，美国路易斯安那州新奥尔良。计算语言学协会。

Arijit Ray、Yi Yao、Rakesh Kumar、Ajay Divakaran 和 Giedrius Burachas。2019。你能解释一下吗？清晰的解释有助于人机协同图像检索。载于*AAAI*人类计算与众包会议论文集，第 7 卷，第 153–161 页。

Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2016。“我为什么应该相信你？”解释任意分类器的预测结果。载于第 *22* 届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第 1135–1144 页。

Teven Le Scao、Angela Fan、Christopher Akiki、El- lie Pavlick、Suzana Ilic、Daniel Hesslow、Roman Castagné、Alexandra Sasha Luccioni、François Yvon、Matthias Gallé、Jonathan Tow、Alexander M. Rush、StellaBiderman、AlbertWebson、PawanSasanka Am- manamanchi、Tho mas Wang、Benoît Sagot、Niklas Muennighoff、Albert Villanova del Moral、Olatunji Ruwase、Rachel Bawden、Stas Bekman、Angelina McMillan-Major、Iz Beltagy、Huu Nguyen、Lucile Saulnier、Samson Tan、Pedro Ortiz Suarez、Vic- tor Sanh、Hugo Laurençon、Yacine Jernite、Julien Launay、Margaret Mitchell、Colin Raffel、Aaron Gokaslan、Adi Simhi、Aitor Soroa、Alham Fikri Aji、Amit Alfassy、Anna Rogers、Ariel Kreisberg Nitzav、Canwen Xu、Chenghao Mou、Chris Emezue、Christopher Klamm、Colin Leong、Daniel van Strien、David Ifeoluwa Adelani 等。2022a。BLOOM：一个拥有 1760 亿参数的开放获取多语言语言模型。*CoRR*，abs/2211.05100。

Teven Le Scao、Angela Fan、Christopher Akiki、Ellie Pavlick、Suzana Ilic、Daniel Hesslow、Roman Castagné、Alexandra Sasha Luccioni、François Yvon、Matthias Gallé、Jonathan Tow、Alexander M. Rush、Stella Biderman、Albert Webson、Pawan Sasanka Ammanamanchi、Tho mas Wang、Benoît Sagot、Niklas Muennighoff、Albert Villanova del Moral、Olatunji

Ruwase、Rachel Bawden、Stas Bekman、Angelina McMillan-Major、Iz Beltagy、Huu Nguyen、Lucile Saulnier、Samson Tan、Pedro Ortiz Suarez、Victor Sanh、Hugo Laurençon、Yacine Jernite、Julien Launay、Margaret Mitchell、Colin Raffel、Aaron Gokaslan、Adi Simhi、Aitor Soroa、Alham Fikri Aji、Amit Alfassy、Anna Rogers、Ariel Kreisberg Nitzav、Canwen Xu、Chenghao Mou、Chris Emezue、Christopher Klamm、Colin Leong、Daniel van Strien、David Ifeoluwa Adelani 等。2022b。BLOOM：一个拥有 1760 亿参数、开放访问的多语言语言模型。*CoRR*，abs/2211.05100。

Ramprasaath R Selvaraju、Stefan Lee、Yilin Shen、Hongxia Jin、Shalini Ghosh、Larry Heck、Dhruv Batra 和 Devi Parikh。2019 年。借力提示：利用解释使视觉和语言模型更具 grounded 性。在 *IEEE* 国际计算机视觉会议论文集，第 2591–2600 页。

Julia Strout、Ye Zhang 和 Raymond Mooney。2019 年。人类判定依据能改善模型解释吗？载于 *2019* 年 *ACL* 研讨会 *Black-boxNLP*：分析和解释用于自然语言处理的神经网络论文集，第 56–62 页，意大利佛罗伦萨。计算语言学协会。

Ian Tenney、James Wexler、Jasmijn Bastings、Tolga Bolukbasi、Andy Coenen、Sebastian Gehrmann、Ellen Jiang、Mahima Pushkarna、Carey Radebaugh、Emily Reif 和 Ann Yuan。2020 年。语言可解释性工具：面向自然语言处理模型的可扩展交互式可视化与分析。载于 *2020* 年自然语言处理实证方法会议论文集：系统演示，第 107–118 页，在线。计算语言学协会。

Eric Wallace、Jens Tuyls、Junlin Wang、Sanjay Sub- ramanian、Matt Gardner 和 Sameer Singh。2019。AllenNLP解释工具：用于解释自然语言处理模型预测结果的框架。载于《*2019*年自然语言处理经验方法会议暨第*9*届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集：系统演示》，第7–12页，中国香港。计算语言学协会。

Zijie J. Wang、Dongjin Choi、Shenyu Xu 和 Diyi Yang。2021年。让人类参与自然语言处理：一篇综述。载于《首届人机交互与自然语言处理研讨会论文集》，第47–52页，在线。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、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。Transformers：自然语言处理领域的最先进技术。

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，在线。计算语言学协会。

Huihan Yao、Ying Chen、Qinyuan Ye、Xisen Jin 和 Xiang Ren。2021 年。《使用组合式解释改进语言模型》。神经信息处理系统进展，34：8954–8967。

A BackboneAPI 端点

图 6 展示了为我们的骨干模块 REST API 自动生成的文档，这些文档为如何与骨干模块交互提供了指南。端点按功能分组——模型、数据集、预测结果、解释和 反馈）。出于安全考虑，目前只能通过我们机构的网络访问此页面。

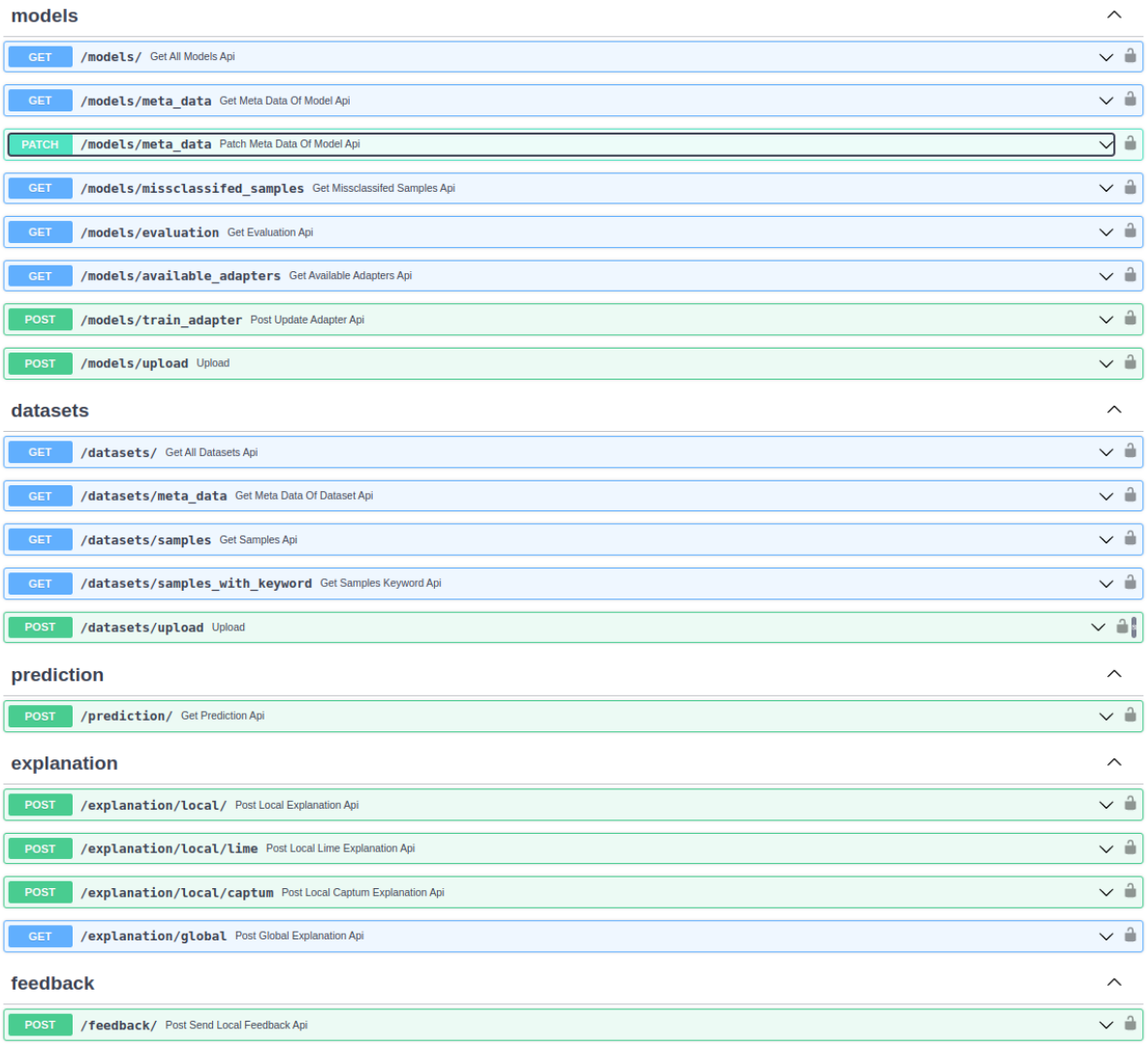


图 6：我们主干网络 API 端点的 Swagger UI 截图。

拥有直接 API 访问权限（可在配置页面上指定，参见3.1）的开发者可以直接向这个高级界面发送请求，以进行额外（规模更大）的实验。再次说明，该 API 是使用 Python 框架 FastAPI 构建的[13]。

图7 展示了解释端点的文档。在这里，我们可以查看该端点的详细信息，例如所需参数——即模型路径、要使用的解释器（例如 LIME），以及作为请求体传入的模型预测结果。

[13]https://fastapi.tiangolo.com

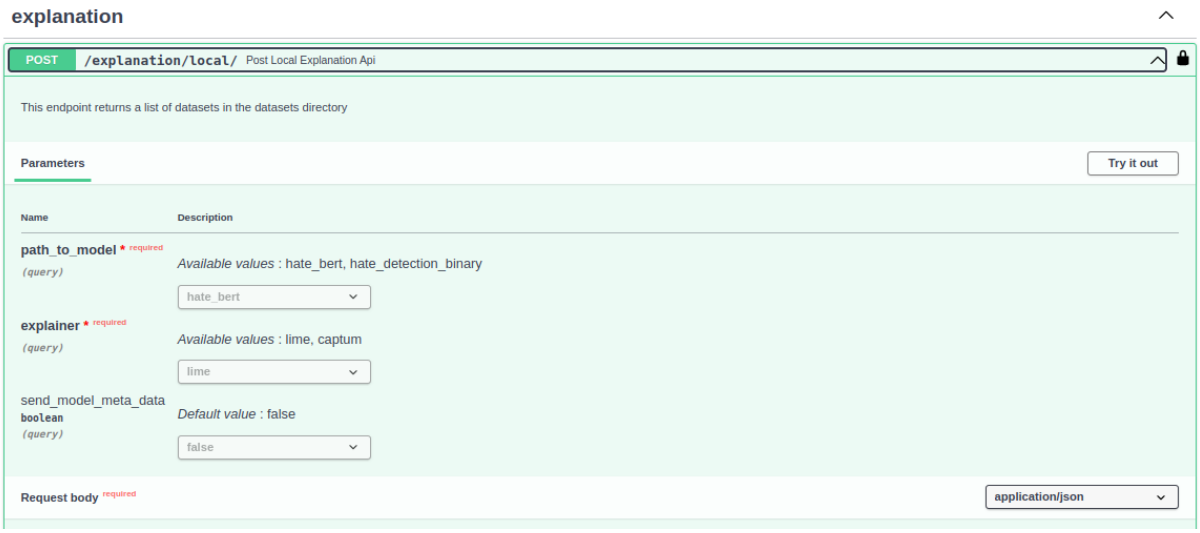


图 7：我们主干网络 API 的 Swagger UI 中解释端点的屏幕截图。

B 用户研究

我们开展了一项小型用户研究，以评估不同背景用户对我们平台的可用性。我们共招募了九名测试者，并根据他们对相关方法的熟悉程度将其分为三类。用户可以是外行用户、计算机科学家或专家。每类分配三名测试者。外行用户是没有接触过计算机科学的普通用户。计算机科学家是指学习过计算机科学或相关学科，并在相应领域（如软件开发）工作的人员。同时具备机器学习和自然语言处理领域知识的人员则被视为专家。

我们要求用户尝试与 IFAN 进行交互，然后回答几个有关该网站可用性的问题。问卷结构如下：

- 第一个问题询问测试人员的专业能力，分为计算机科学、自然语言处理专家或外行用户。

- 用户需要为反馈页面左侧的局部解释评分，分值范围为 1 到 5 分。

- 供用户填写的文本输入框，用于针对上一项评分提供更详细的反馈。

- 用户为反馈页面右侧的全局解释打 1 到 5 分。

- 供用户填写的文本输入框，用于针对上一项评分提供更详细的反馈。

- 用户为网站的整体可用性打 1 到 5 分。

- 供用户填写的文本输入框，用于针对上一项评分提供更详细的反馈。

用户给出的平均评分为：(i) 局部解释：3.88；(ii) 全局解释：3.2；(iii) 一般可用性：3.33。用户提到的问题包括解释和反馈形式过于单一、标注界面有时会产生误导，以及等待时间较长。虽然平台当前版本已经修复了部分问题，但我们计划持续提升平台的可用性。同时，我们也注意到一个积极结果：背景各异的用户总体上都能轻松使用 IFAN，并认为它有助于提高自然语言处理模型的可控性。我们相信，这证实了 IFAN 在未来应用中可供利益相关者、开发者和标注员使用。

C 案例研究

本节将补充介绍在案例研究（第 4 节）中开展的实验。

C.1 标注样本（警告：本节仅为举例而包含仇恨言论。我们无意冒犯读者或任何特定种族群体。）

此处列出了用于人工标注的样本（第 4 节）。这些样本被仇恨言论检测模型误分类。表 4 列出了置信度最低的误分类样本，表 5 则列出了置信度最高的样本。所有样本及其真实标签均取自原始 HateXplain 数据集（Mathew 等人，2021年）。有关标签分配的详情，请参阅 HateXplain 数据集说明。

样本	原文标签	预测结果	置信度分数
这与《托拉》并没有太大关系，《托拉》本身相当无害，而且尽管有相关说法，许多犹太人仍然对此置之不理；它真正指的是那本被禁止的书——《塔木德》，或《巴比伦塔木德》，它如同外邦人牛，我们连读都不被允许。祝你好运，继续你的研究	无毒	有毒	0.56
如果我没记错，阿什肯纳兹犹太人和东亚人的平均智商	无毒	有毒	0.57
说到这个，有人知道犹太人和哪个非犹太群体通婚人数最多吗？我原本以为是亚-洲人或天主教徒	有毒	无毒	0.52
<用户> 犹太人伪造了登月	有毒	无毒	0.52
<用户> <用户>我敢打赌，她会把这事怪到犹太人或特朗普头上	有毒	无毒	0.53
基督教的角色是犹太人的历史保护者，它让这个部族得以获得所需的时间和空间来准备白人种族灭绝团队白人	有毒	无毒	0.54
犹太人正准备帮助民主党人在 <num-ber> 推特 脸书 主流媒体 Google YouTube 都在加强他们的对支持白人的权利的审查	有毒	无毒	0.51
想想白人永远不会被给予一个机会，真是令人觉得有趣关于将犹太人驱逐出我们的国家的公投，当你以这种方式思考问题时，就会非常清楚我们的政府希望我们忍受犹太人的寄生行为，却不允许我们自卫	有毒	无毒	0.50
这一切都在呵呵哈哈，直到我们出现在你们当地最喜欢的聚会地点，然后在面包车里等你出来，他们一个接一个地来，一次对付一个小犹太佬，不管要付出什么代价	有毒	无毒	0.53
<用户> <用户> 犹太人和愚蠢的大屠杀神话	有毒	无毒	0.52
<用户> 犹太人永远不会出现在区块链上，因为你可以不信任他们	有毒	无毒	0.51
<用户> <用户>是犹太人，不是欧洲人，别再推卸责任了针对非犹太人	有毒	无毒	0.57

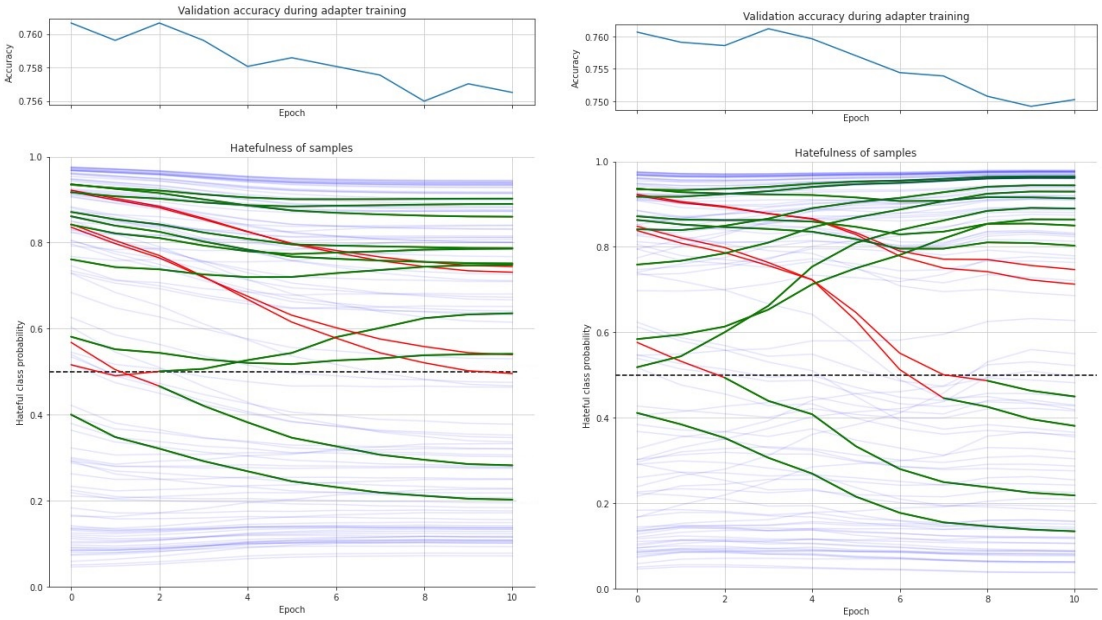
表4：用于人工标注的置信度最低的误分类样本。

样本	原文标签	预测结果	置信度分数
宁愿遭受恐怖主义，也不愿失去抵制某个盟国的能力。这个论点一定很重要，因为他反复提起它，但我实在搞不懂这到底是想表达什么意思。如果你不买犹太人的东西，他们就会假扮成阿拉伯人拆毁更多纽约市地标	无毒	有毒	0.72
犹太人在以色列还会占多数多久?	无毒	有毒	0.88
你说去买下他们的房子是什么意思？这不仅早于.....隔都，它早在隔都的人出现之前就存在了。为什么?你是想让白人为自己的家园挺身而出，却让犹太人逃离自己的家园	无毒	有毒	0.77
叫我犹太佬并不算侮辱	无毒	有毒	0.97
犹太人和穆斯林一起长大，犹太人孕育了基督教、天主教，后来还有摩门教，就像一个庞大的宗教家族	无毒	有毒	0.78
格雷格、苏珊、弗兰克、罗德里克和曼尼·赫夫利都是犹太人	无毒	有毒	0.67
真令人惊讶，<用户>说了这么久，却什么答案也没给比起萨尔贡，我从<用户>的想法和生活中学到了更多我还以为只有犹太人才会如此夸夸其谈，嗯?	有毒	无毒	0.95
<用户><用户>呃，实际上星际迷航里的是太空共产主义者而星球大战里的绝地武士只不过是太空犹太人	有毒	无毒	0.87
<用户>显然有些人并不这么认为，尤其是同性恋者犹太媒体希望女性看起来像男孩和其他人他们认为那是传统，而不是女权主义者	有毒	无毒	0.84
抱歉，我指的是大屠杀的故事，我认为它是一个lie	有毒	无毒	0.89
我敢发誓，大学简直就是个温室，里面满是犹太化的教授，他们.....白人青年被灌输这种荒谬观念，或者..........这些天生活在恐惧中，生怕因被揭穿是种族主义者而遭到驱逐，并且自豪地.....白人种族的成员	有毒	无毒	0.90
几个月前，我说过反犹太主义并不存在，给一群人取名——是社会科学的一个分支，叫作社会学。好吧，犹太人现在正与社会科学为敌，你根本不能给他们命名	有毒	无毒	0.74

表5：用于人工标注的最高置信度误分类样本。

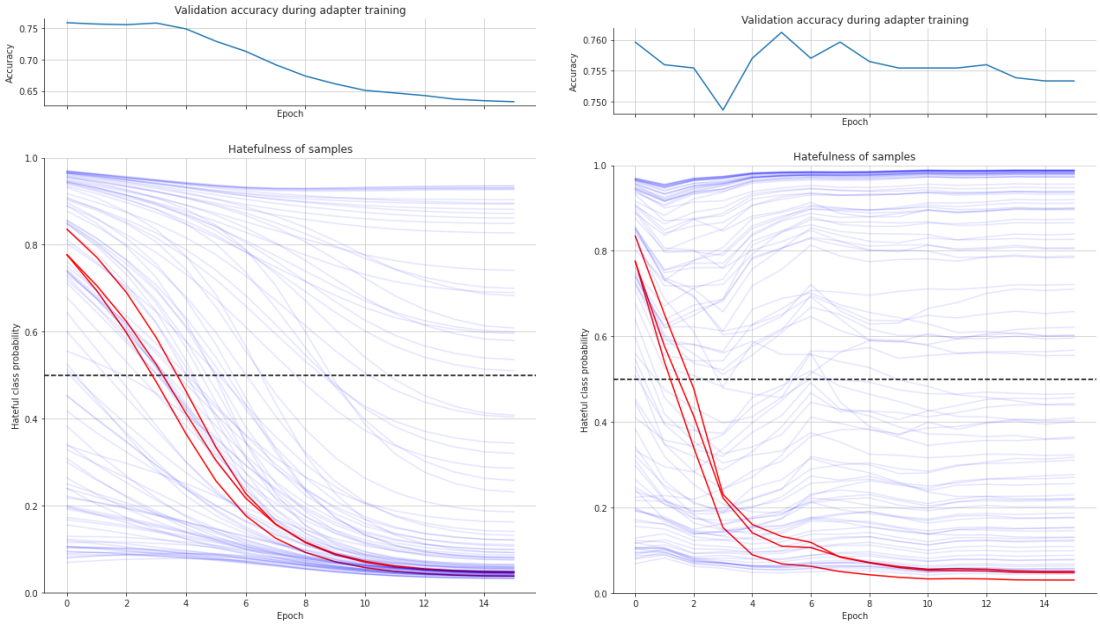
C.2 反馈机制比较

在第 4 节中，我们报告了在两种设置下使用反馈训练的模型结果：(i) 不进行平衡；(ii) 使用 HateXplain 数据集中的原始样本进行平衡。这两种设置的比较见图 8 和 9。我们在犹太人目标群体样本以及包含“Jewish”关键词的样本上测试了局部反馈方法。在这两种设置中，经过平衡后，训练过程都更加稳定。模型在 HateXplain 数据集其他样本上的性能变化较小，而且对标注样本的行为调整也更加迅速。



(a) 未进行反馈平衡的训练。　　(b) 进行了反馈平衡的训练。

图8：对比进行与未进行反馈平衡的训练过程。这里展示了针对犹太人目标群体中置信度最低的误分类样本所提供的局部反馈结果。可以看出，使用平衡数据集进行训练更加稳定，且不会显著影响模型的整体领域知识。



(a) 不进行反馈平衡的训练。 (b) 进行反馈平衡的训练。

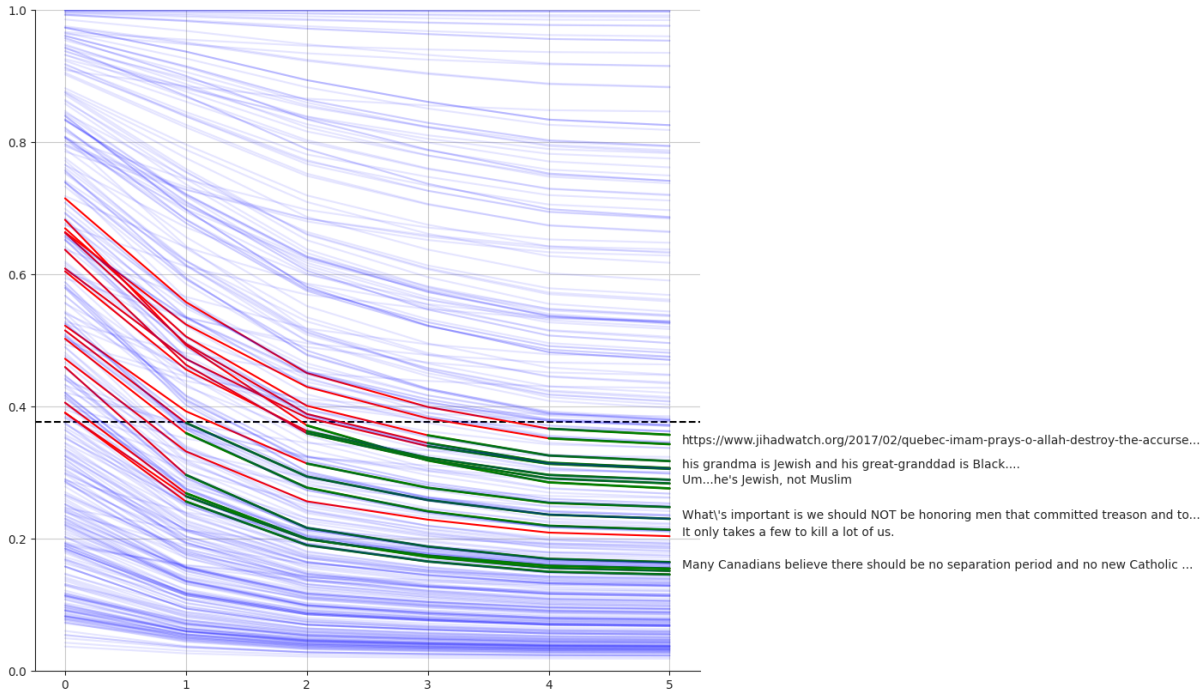
图 9：对比进行和不进行反馈平衡时的训练过程。此处展示了针对包含“jewish”关键词的误分类样本所提供的局部反馈的结果。我们可以观察到，使用重新平衡的数据集进行训练时，运行更加稳定，且不会显著影响模型的整体领域知识。

D 用于仇恨言论分类和人类反馈整合的 BLOOM n

除了 BERT 在仇恨言论分类方面的结果（第4节），我们还提供了使用 BLOOM[14] 模型（Scao 等人，2022b）进行实验的结果（表6）。同样，在犹太人目标群体中，使用人类反馈微调适配器层提升了模型性能。对于 BLOOM，整体 F1分数也有所提升，这证实了使用人类反馈的效用，以及将其扩展到其他模型的可能性。此外，从图10中可以看到，随着模型根据所提供的反馈进行微调，误分类样本的标签（即“嗯.....他是犹太人，不是穆斯林”）如何发生变化。

模型	Pr	Re	F1	$F1_J$
BLOOM（基线）	0.40	**0.70**	0.48	0.48
BLOOM+反馈	**0.49**	0.58	**0.51**	**0.53**

表 6：大语言模型推理用于仇恨言论分类的结果。$F1_J$ 表示犹太人目标群体的 F1分数。



Fig图 10：反馈整合过程中误分类样本标签修正的追踪流程 .

E 仇恨言论分类与解释的大语言模型

除了在 Transformer 模块顶部通过线性层进行分类外，我们还测试了大语言模型执行文本分类并解释其预测结果的可能性。这种方法有望使后续工作能够将该框架扩展到任意自然语言处理任务。

目前，IFAN 上可供测试的模型包括：vicgalle/gpt2-alpaca[15], chavinlo/alpaca-native[16],以及 chavinlo/gpt4-x-alpaca。[17] 它们在基于 HateXapline 数据集的仇恨言论分类任务上的结果（Mathew 等人，2021年）见表7。

对于每个模型，我们都会优化提示词以执行仇恨言论检测：

[14]https://huggingface.co/bigscience/bloom-560m [15] https://huggingface.co/vicgalle/gpt2-alpaca [16] https://huggingface.co/chavinlo/alpaca-native [17] https://huggingface.co/chavinlo/gpt4-x-alpaca

vicgalle/gpt2-alpaca 指令：使用以下标签对文本执行文本分类任务：*0* 或 *1*。文本：如果联合国大楼被改造成无家可归者收容所，那就太好笑了；或者，见鬼，为什么不改成难民中心呢？该文本的标签是 *0*。文本：*[DATASET SAMLPE]*该文本的标签是 *1*。文本：*[INPUT TEXT]*该文本的标签是

chavinlo/alpaca-native 这是一项整体文本分类任务。对于给定的文本，请对其进行分类并给出标签。标签可以是 *0* 或 *1*。只输出标签，不要输出其他内容！文本：*[5* 数据集样本*]* 标签：*0* 文本：*[5* 数据集样本*]*标签：*1* 文本：*[*输入文本*]*。标签：

chavinlo/gpt4-x-alpaca 指令：根据以下数据集描述，提取其任务，并对输入文本执行该任务。最终标签可以是 *0* 或 *1*。只输出最终标签，不要输出其他内容！数据集描述：*Hatexplain* 是首个涵盖该问题多个方面的基准仇恨言论数据集。数据集中的每条帖子都从三个不同角度进行标注：基础且常用的三分类（即仇恨、冒犯或正常）、目标社群（即帖子中遭受仇恨言论或冒犯性言论的社群），以及判定依据，即标注者据以作出标签判断（仇恨、冒犯或正常）的帖子片段。上下文：文本：*[*数据集样本 *]*最终标签：*0* 文本：*[* 数据集样本 *]* 最终标签：*1* 输入文本：文本：*[*输入文本*]*。最终标签：

尽管各种提示词之间存在具体差异，但它们都遵循一个核心思路：i）提及可从数据集元数据中提取的主要任务；ii）提供有关标签的一般信息；iii）提供数据集中每个标签的一些示例。理论上，这种提示词设计可用于任何分类任务。

模型	Pr	Re	F1
vicgalle/gpt2-alpaca	0.63	0.56	0.59
chavinlo/alpaca-native	0.60	0.51	0.55
chavinlo/gpt4-x-alpaca	**0.66**	**0.63**	**0.64**

表7：大语言模型在仇恨言论分类任务中的推理结果。

F 补充视频演示

可在 YouTube 上观看补充视频展示[18]。为求完整，我们还提供了该演示的另一个版本，同样发布在 YouTube 上[19]，可追溯至2023年3月。

[18]https://youtu.be/EzC6HI3JwaQ [19]
https://www.youtube.com/watch?v=BzzoQzTsrLo