arXiv:2601.01350v1 [cs.CL] 4 Jan 2026

FC-CONAN：用于稳健评估检索系统的穷尽式配对数据集

Juan Junqueras[1*], **Florian Boudin**[2], **May Myo Zin**[3], **Nguyen Ha Thanh**[3,4], **WacharaFungwacharakorn**[3], **Dami´an Ariel Furman**[1], **Akiko Aizawa**[4], **Ken Satoh**[3,41]布宜诺斯艾利斯大学，精确与自然科学学院，计算机系，阿根廷布宜诺斯艾利斯 JFLI、法国国家科学研究中心、南特大学，法国南特[2]ROIS-DS 法律信息学中心，日本东京[3]日本国立信息学研究所（NII），日本东京

摘要

仇恨言论（HS）是网络话语中的一个严峻问题，而一种有前景的应对策略是使用反叙事（CNs）。将仇恨言论与反叙事关联起来的数据集对于推动反仇恨言论研究至关重要。然而，即使是 CONAN（Chung 等人，2019）这类旗舰资源，也只标注了所有可能的仇恨言论—反叙事对中的一小部分，限制了评估。我们推出 **FC-CONAN**（全连接 CONAN），这是首个通过穷尽考察 45 条英语仇恨言论消息与 129 条反叙事的所有组合而构建的数据集。由九名标注人员和四名验证人员参与的两阶段标注流程，生成了四个分区——钻石级、黄金、白银和青铜级，在可靠性与规模之间取得平衡。所有已标注配对均未与 CONAN 重叠，由此发现了数百个此前未标注的正例。FC-CONAN 使反仇恨言论检索系统的评估更加准确，并有助于开展细致的错误分析。该数据集现已公开[1][2]。

关键词：仇恨言论、反叙事、穷尽式标注、完全配对的数据集、数据集构建、基准构建、质量分级分区、标签稀疏性、下界偏差、检索系统评估、推荐系统、信息检索、评估偏差、评估指标、标注质量、论证、反仇恨言论、自然语言处理（NLP）。

1 引言

免责声明。本文出于研究目的逐字引用仇恨言论；部分读者可能会觉得其中的措辞令人反感。

许多自然语言处理（NLP）数据集由成对句子构成，例如问题与答案（Rajpurkar 等人，2016）、释义（Dolan 和 Brockett，2005年）、蕴含关系（Bowman 等人，2015）、翻译（Koehn，2005年）和对话（Li 等人，2017年）。虽然有些数据集允许一个句子与多个其他句子关联，例如 CONAN（Chung 等人，2019），但大多数数据集只涵盖所有可能组合中的一部分。由于组合数量不断增长且成本高昂，穷尽式标注很少有人尝试，而成本问题使未标注的句对仍存在歧义——它们往往反映的是疏漏，而非关系确实不存在。这种不完整性对于推荐任务尤其棘手，因为指标可能会严重低估系统性能（§4.3）。

这一挑战在仇恨言论等领域尤为严峻。社交媒体加剧了有害言论的传播（Silva 等人，2021；Waseem 和 Hovy，2016），促使人们寻求内容删除之外的应对方式，因为内容删除可能强化审查叙事。作为一种更具建设性的替代方案，结构化反仇恨言论协议侧重于及时、审慎的回应，旨在驳斥有害论点、避免引发更多争论，并与更广泛的目标保持一致。CONAN 数据集（通过利基众包收集反叙事）（Chung 等人，2019）是“首个大规模、多语言、由专家构建的仇恨言论／反叙事配对数据集”，至今仍是主要资源。为展示此类数据的特点，我们在下文给出一个具有代表性的示例。

示例 1（来自 CONAN（Chung 等人，2019）的一个仇恨言论—反叙事配对）。 仇恨言论（仇恨言论）“我憎恨穆斯林。他们不应该存在。”
反叙事。 “穆斯林也是人。人们可以选择自己的宗教。”

尽管 CONAN 得到了广泛使用，但它仍有显著局限。**至关重要的是**，它并未标注所有仇恨言论—反叙事组合，因此许多恰当的配对没有标签。这限制了它的实用性，尤其是在评估反叙事推荐系统时，因为性能指标只能反映一个下限。在一项使用其中一个系统进行的试点研究中，我们发现，尽管 10 条推荐的反叙事中只有 2 条被标注为恰当，但人工审查判定其中 8 条有效——这凸显了低估系统准确率的风险。

缺乏完整的仇恨言论—反叙事配对标注，也会因减少用于微调大语言模型的训练数据而限制生成任务。全面的仇恨言论—反叙事标注还能进一步支持对比学习等方法。归根结底，未标注的配对使宝贵的潜在信息无法得到利用，降低了数据集对下游应用的实用性。

该数据集的另一特点是，指南的规定相当宽泛。这种做法源于这样一个事实：最初的标注人员此前已经接受过培训，

*通讯作者：jjunqueras@dc.uba.ar
[1]该数据集可在 https://github.com/jnqeras/ FC-CONAN-dataset 公开获取
[2]第一作者在日本东京的国立情报学研究所（NII）任职期间，部分完成了这项工作。

需遵循非政府组织关于撰写有效反叙事的指南。这些指南在不同语言和组织之间都相当一致，并且与 Get the Trolls Out 项目[3]制定的指南十分相似。我们鼓励标注人员依靠直觉，避免过度思考，并给出合理的回答（Chung 等人，2019，§3.2）。需要注意的是，高度主观性是这一领域的特点。

受资源限制，我们对所有可能的仇恨言论—反叙事配对中的一个有代表性的子集重新进行了标注。本文详细介绍了这项工作；最初开展这项工作的目的是更准确地评估反仇恨言论推荐系统。虽然 CONAN 涵盖三种语言，但我们的工作仅关注英语。将标注扩展到其他语言是未来的工作。

2 相关工作

在仇恨言论领域，CONAN（Chung 等人，2019）是最知名的多语言资源之一。该资源由专家整理，聚焦于伊斯兰恐惧症，包含英语、意大利语和法语的仇恨言论（HS）与反叙事（CN）配对。其初始版本包含 4,078 对（其中 1,288 对为英语），基于 136 条独特的仇恨言论消息，每条消息平均匹配 9.5 条反叙事。通过翻译和改述，英语部分扩展至 6,654 对，包含 408 条独特的仇恨言论消息和 1,270 条反叙事。该数据集还包含专家人口统计信息、反叙事类型和仇恨言论子主题等元数据。

其他一些数据集也聚焦于仇恨言论和反叙事，例如 DIALOCONAN（Bonaldi 等人，2022年），其中包含仇恨言论发布者与非政府组织工作人员之间的多轮对话，但并未按仇恨言论/反叙事配对进行组织。另一个例子是 Multitarget CONAN（Fanton, Margherita、Bonaldi, Helena、Tekiro˘glu, Serra Sinem 和 Guerini, Marco，2021），这是一个包含针对多个仇恨目标的仇恨言论/反叙事配对的数据集。然而，这些数据集并未涵盖仇恨言论与反叙事配对的所有可能组合。Furman 等人，2023 的研究中，有一项发现与我们的工作最为相关：经过少量高质量仇恨言论—反叙事配对微调的小型大语言模型，表现可以优于更大型的模型。我们的工作与他们的研究相辅相成：我们关注的不是论证线索，而是对仇恨言论—反叙事示例进行穷尽式配对，以实现稳健的评估和训练。

已有许多用于仇恨言论检测任务的数据集，例如（Waseem 和 Hovy，2016）介绍的 Twitter 语料库。虽然我们的数据集也可能用于这一目的，但其设计主要着眼于其他方面。

3 数据集构建

我们提出了 FC-CONAN，这是一个源自 CONAN 语料库子集的仇恨言论–反叙事对数据集。在标注过程中，我们穷举考虑了所选仇恨言论和反叙事项目的所有可能组合，并将每一对标记为合适或不合适，其中部分配对依据质量控制标准被移除。

[3]https://getthetrollsout.org/stoppinghate

我们从 CONAN 数据集的英语分区（所有标注人员共同使用的唯一语言）中随机选取了 45 条仇恨言论消息，并收集了最初与其配对的 375 条反叙事。对于每条反叙事，我们使用 SBERT 模型 all-MiniLM-L6-v2[4]，从完整的 CONAN 数据集中检索与其最相似的两条反叙事。随后，我们舍弃了原有的 375 条反叙事，同时保留仇恨言论消息及其新检索到的反叙事，确保仅保留全新的仇恨言论–反叙事组合，使标注人员处理的反叙事与原始反叙事相似但并不相同，从而避免任务过于简单。该过程产生了 133 条独特的反叙事；我们从中随机选取 129 条，以确保仇恨言论–反叙事对在标注人员之间均匀分布，最终生成 5,805 对仇恨言论–反叙事。9 名接受过学术训练的志愿者标注人员依据经调整的 CONAN 指南对这些配对进行标注：提供基于事实的信息，并保持非冒犯性语气，以避免使对话升级。标签由标注人员独立分配，通过重叠标注进行标注者间一致性检查，并通过裁决解决冲突。标记为不确定或存在不可调和争议的配对被舍弃，以避免标签偏差，最终保留 5,032 对已标注数据（其中 4,143 对标记为“该反叙事不适用于给定的仇恨言论”，889 对标记为“该反叙事适用于给定的仇恨言论”）。

为进一步确保可靠性，4,000 对经裁定的配对由四名独立审阅者进行了一轮验证（他们未参与初始标注），每人重新评估 1,000 对。验证者遵循相同的指南，可以跳过令人不适的条目（共跳过三对），并判断标签是否符合要求。这一流程确保每一对保留的配对都经过标注员和验证者的审查。无论这些配对最终是否被判定为有效，均予以保留。集合之外的配对则被舍弃，以优先保证标签可靠性，而非数据集规模。尽管部分被排除的配对可能仍然合适，但标注过程中已审查所有可能的组合。

最终资源通过定义四个质量分级子集（钻石级、黄金、白银级、青铜级），在可靠性与覆盖范围之间取得平衡，使用户能够权衡数据规模与标签置信度。为支持标注人员，伦理保障措施包括风险警告、可选择跳过以及按需安排的总结休息。若反向利用该数据集来生成仇恨言论回复，可能带来双重用途风险；因此，我们强调必须负责任地将其用于下游任务。

4 结果/分析。

在转向系统层面的评估之前，我们先介绍数据集本身。第 4.1 节介绍了通过我们的标注与验证流水线生成的四个按质量分级的分区。第 4.2 节接着探讨这些分区的内部结构。最后，第 4.3 节介绍一项检索实验，用以量化这些分区对下游性能的影响。

分区	总计	适当	不适当
钻石级	551	35	516
Gold	663	54	609
白银	3580	431	3149
青铜级	3997	702	3295

表 1：各数据集分区中按适当性分类的仇恨言论–反叙事（HS–CN）配对数量。

4.1 数据集分区。

经过标注和验证流程（§3），我们获得了由一至三名标注人员标注、并由一名验证者评估有效性的仇恨言论–反叙事对。根据标注人员的标签是否一致以及验证流程的结果，我们定义了四个不同的数据集分区，各自在标注质量和规模上有所不同。表 1 总结了各分区的规模，以及适当与不适当的仇恨言论–反叙事对的分布。以下详细介绍各分区的特点：

•钻石标准数据集： 此分区仅包含由两名或更多标注员标注且达成一致同意的仇恨言论–反叙事配对，无论反叙事被判定为适当还是不适当。此外，一名验证者已确认这些标注的准确性。

• 金标准数据集： 此分区在钻石标准数据集 的基础上，纳入了由两名或更多标注员标注的其他仇恨言论–反叙事配对，不论标注员是否达成一致同意。对于标注员最初存在分歧的情况，这些分歧通过裁决流程（见 § 3）得到解决。随后，验证者进一步审查每项最终标注并确认其准确性。

• 银标准数据集： 银标准数据集包含金标准数据集中的所有仇恨言论–反叙事配对，以及仅由一名标注员标注、随后经验证人员确认的配对。因此，此分区中的标注由 1 至 3 名标注员完成，最初的分歧则通过前述裁决流程解决。此分区中的所有标注均在验证阶段获批。

• 青铜标准数据集： 该数据集包含银标准数据集中的所有条目，并补充了验证阶段未获批准的所有仇恨言论–反叙事配对。与之前一样，标注来自 1 至 3 名标注员，分歧则通过裁定阶段解决。不过，与银标准数据集不同，此分区还保留了验证阶段未获批准的配对。

不出所料，这些分区的规模与质量呈反向关系。质量较高的数据集（钻石级和黄金）要求更高的一致性标注和更严格的验证，因此规模较小。相反，质量较低的数据集（白银和青铜级）规模较大，但其中的标注可能可靠性较低。因此，按规模从小到大（质量从高到低）排列，这些数据集依次为：钻石级、黄金、白银和青铜标准数据集。

4.2 数据集的定性分析

在第 3 节所述的六个重叠子集中，我们得到平均 Cohen's $\kappa = 0.34 (\sigma \approx 0.20)$，该值在进行任何额外检查之前计算得出。正如 Klie 等人所指出的：“尽管人们常常如此看待一致性，但一致性并非万能；高度一致并不自动保证标签具有高质量。”（Klie、Eckart de Castilho 和 Gurevych，2024年），因此我们采用了第 3 节概述的验证程序。本节对四个数据集分区进行定性分析：钻石级、黄金、白银和青铜级。了解这些分区有助于用户选择最合适的子集——优先考虑标注可靠性（钻石级/黄金），或数据量（白银/青铜级）。

表 1 证实，原始 CONAN 数据集中有许多有效的仇恨言论—反叙事配对未被标注。通过评估新的组合，我们在所有分区中找回了数百个恰当和不恰当的配对。

表 2 展示了质量与数量之间的权衡：钻石级和黄金级规模较小，但全部有效；白银规模较大，且仍经全面验证；而青铜级规模最大，却包含一些无效配对——用户可以据此为下游任务选择规模与可靠性之间的平衡。

表 3 展示了根据参与标注的标注人员数量划分的已标注配对分布。在青铜分区中，由三名标注人员标注的 199 个配对里，只有 4 个配对（≈2.01%）被验证人员判定为无效。由两名标注人员标注的 502 个配对中，有 34 个（≈6.77%）被标记为无效。最后，在由单名标注员标注的 3,296 个配对中，有 379 个（≈11.49%）被归类为无效。这些观察结果表明了一个明显趋势：由多名标注人员标注的配对中，无效实例所占比例往往更低，这凸显了标注人员之间的一致性提高能显著提升可靠性。

总之，四级分区在标注可靠性方面实现了平衡。钻石级和黄金级提供了完美的验证结果，非常适合在标注噪声极少的情况下评测模型性能。白银级在不影响有效配对的前提下扩大了数据规模，而青铜级则增加了数据量，同时引入了唯一包含无效配对的子集。总体而言，新增的标注有助于弥补原始 CONAN 数据集中的明显缺口。

4.3 实验评估：推荐系统评估。

如§1所述，不完整的 CONAN 标签掩盖了适当的反叙事。下文将基于穷尽标注的 FC-CONAN 分区[5]，重新评估推荐系统。

我们比较了在仅包含英语的数据集（*conan* 不在 *bronze* 训练集）上训练的十二种推荐系统，该数据集通过排除青铜分区中发现的任何仇恨言论或反叙事而创建——

[4]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[5]本实验所用的数据集可在以下网址获取：https://github.com/jnqeras/FC-CONAN-dataset/tree/main/recommender_experiment_data

分区	有效	无效
钻石级	551	0
Gold	663	0
白银	3580	0
青铜级	3580	417

表 2：按有效性分类的各数据集分区中仇恨言论–反叙事（HS–CN）对的数量。

分区	3 名标注者	2 名标注者	1 名标注者
钻石级	195	356	0
Gold	195	468	0
白银	195	468	2917
青铜级	199	502	3296

表 3：由 3、2 和 1 名标注者标注的仇恨言论–反叙事对数量。

从而确保训练集与评估集之间没有重叠。具体来说，我们评估：

- **TF–IDF** – 基于 TF–IDF 向量的余弦相似度（稀疏基线）；（Saltonand Buckley，1988）。
- **BM25** – Okapi BM25 词项排序器（k_1=1.2, b=0.75）；（Robertsonand Zaragoza，2009）。
- **random** – 对十个候选 CN 进行均匀采样。
- **sbert(MiniLM)** – Sentence-BERT *all-MiniLM-L6-v2*(384 维) +余弦；（Reimers 和 Gurevych，2019），（Wang 等人，2020）。
- **sbert(MPNET)** – Sentence-BERT *all-mpnet-base-v2*(768 维) + 余弦；（ Song 等人，2020 ）。
- **sbert(RoBERTa-L)** – Sentence-BERT *all-roberta-large-v1* (1024 维) + 余弦；（Zhuang 等人，2021）。
- **textEmb3L** – OpenAI 稠密编码器 *text-embedding-3-large* (3 072 维)；（OpenAI，2024b）。
- **GPT-4o**– 对*gpt-4o-2024-08-06* 进行零样本提示，要求其返回 10 个反叙事；（OpenAI，2024a）。
- **RR25 系统 1 / RR50 系统 1** – 选取 textEmb3L 得分最高的 25/50 个反叙事，然后由 GPT-4o 重新排序。
- **RR25 系统 2 / RR50 系统 2** – 来自 sbert(MPNET) 的前 25/50 条反叙事，再由 GPT-4o 重新排名。

为确保比较公平，每个系统都对固定的候选池进行排名：青铜分区中的 100 条独特反叙事。这模拟了所有模型共享同一份固定回复目录的部署场景。为保持一致，我们将查询集限定为青铜分区中的仇恨言论实例（青铜级 *HS*）。对于青铜级 *HS*中的每一条仇恨言论，每个推荐系统都会生成排名前 10 的反叙事，从而为每个模型生成一份排名。

随后，我们针对全部四个分区（钻石级、黄金、白银和青铜级）评估每项排名。当且仅当该分区将 ⟨HS;CN⟩ 组合标注为恰当时，才计为一项正确推荐。例如，为保持论文简洁，我们展示以下仇恨言论的 TF-IDF 排名（最多列出 2 条反叙事）：

• ”伊斯兰教尚未传入英格兰时，一切都很好。如今我们到处都是穆斯林强奸犯，显而易见，英格兰的性暴力就是圣战”。

每个分区的排名都标记为 ✔ 或 Ø，以说明适当性如何随所采用的标注标准而变化。

1. 穆斯林不是强奸犯或杀人犯；事实上，许多穆斯林正在采取行动，反击那些冒用伊斯兰教之名的极端分子：https://5pillarsuk.com/2019/01/03/muslim-woman-praised-for-aiding-victims-of-new-years-eve-terror-attack-in-manchester/。[DØ, GØ, SØ, BØ]2. 在大多数情况下，强奸犯是受害者的朋友、家人或伴侣。巴基斯坦人中强奸犯的比例并不高于英国白人人口中的比例。[DØ, G✔, S✔, B✔]

在示例排名中，排名第一的反叙事（CN）在任何分区中都从未被判定为适当。相比之下，排名第二的反叙事在黄金、白银和青铜分区中被判定为适当，但在钻石级分区中则不然。这表明，根据用于定义“适当”配对的分区不同，在生成的排名上计算出的评估指标可能会发生显著变化。

图 1 和图 2 表明，随着分区中包含更多经过标注的“适当”配对，推荐系统的性能会提升：钻石级分区的得分最低，其次是黄金、白银，而青铜级分区在两项指标上均取得最高值。总体而言，各项指标值大致与每个分区中的适当配对数量成正比。对于 NDCG@10、MAP@10、Precision@10、Accuracy@10 和 F1@10 等指标，我们也观察到了相同的趋势，但由于篇幅限制，相关图表未予展示。

在 CO-NAN 数据集中，由候选池（青铜级 反叙事）和 青铜级 *HS* 组成的所有配对均未被标注为恰当配对；因此，将 CONAN 作为这一集合唯一的金标准时，所有指标的得分均为零。随着逐步纳入更多恰当配对——从 CONAN 子集到钻石级、黄金、白银和青铜级——指标得分随之提高。这表明，当恰当配对仍未被标注时，评估指标只能作为下限，无法反映检索系统的真实性能。

表 4 展示了清晰的性能层级。**基于嵌入的排序器**（所有 SBERT 变体以及 OpenAI 的 textEmb3L）获得了最高的平均分（$\mu \approx 0.32$）和最低的变异系数（$CV\% \approx 46$），这表明稠密向量表示在面对缺失标注时既有效又鲁棒。**混合重排序器**（RR50 系统 1/2 和 RR25 系统 2）的性能次之（$\mu \approx 0.23$），但稳定性仍明显较低（$CV\% \approx 69$）；这可能是因为只要嵌入预过滤器检索到较弱的候选项，GPT-4o 重排序步骤就会放大噪声。在**词汇基线**中，TF–IDF 的有效性与混合方法相当（$\mu \approx 0.22$），而 BM25 则逊色一些（$\mu \approx 0.18$）；二者的变异性都很高（$CV\% > 63$），证实了它们对标注稀疏性的敏感性。**大语言模型零样本**策略（仅使用 GPT-4o）的表现与混合重排序器 RR25 系统 1 相近

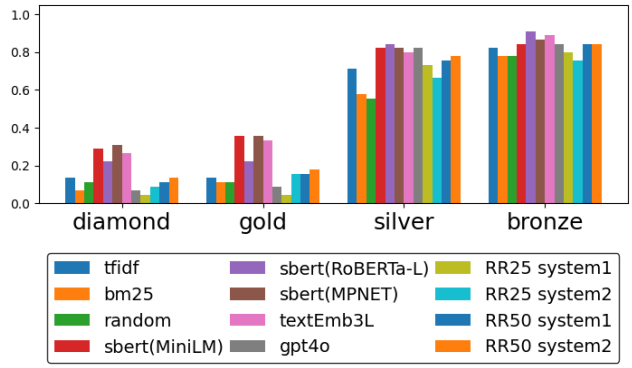


图 1：十二个系统的命中率@10（术语表见第 4.3 节）。

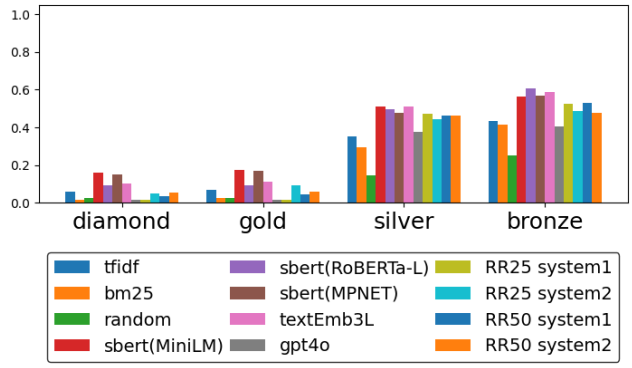


图 2：所评估的十二个系统的 MRR@10（参见第 4.3 节中的术语表）。

（$\mu \approx 0.21$, $CV\% \approx 89$）。正如预期，**随机**基线排名最低（$\mu = 0.15$）。总体而言，结果支持这样一个结论：在反叙事检索中，决定有效性和鲁棒性的是表示质量，而非单靠流水线复杂度。

5 局限性与未来工作

生成式微调与对比学习： 可以针对每个分区中的适当配对对生成式模型进行微调，以比较其性能——我们开展了此类实验，但受篇幅限制未予列出；而对比学习则可以利用适当配对和非适当配对，充分挖掘数据集的结构。

语言范围： 我们仅涵盖 CONAN 的英语部分。将穷举式配对扩展至法语和意大利语仍是未来的工作。

标注覆盖范围： 我们的 45 个 HS × 129 反叙事子集生成了足够大的分区，能够揭示评估伪影，但距离对 CONAN 的所有可能组合进行完整重新标注仍相去甚远。采用“大语言模型优先、人工核验”的半自动化流水线，或许能够以更低的成本完成这项工作。

人口统计多样性： 我们的标注员群体偏向年轻、受教育程度较高的英语使用者。扩大人口统计抽样范围，或许能揭示文化背景会影响适当性判断。

系统	Avg.	Min.	Max.	CV%
sbert(MPNET)	0.3283	0.1836	0.4904	38.425
sbert(MiniLM)	0.3270	0.1766	0.4804	39.200
textEmb3L	0.3128	0.1366	0.5098	53.200
sbert(RoBERTa-L)	0.3034	0.1214	0.5138	56.400
RR50 系统2	0.2441	0.0658	0.4271	69.175
RR50 系统1	0.2416	0.0526	0.4508	75.350
RR25 系统2	0.2278	0.0514	0.3992	63.175
TF-IDF	0.2203	0.0703	0.4024	63.275
RR25 系统1	0.2165	0.0184	0.4339	91.950
GPT-4o	0.2154	0.0267	0.4118	87.100
BM25	0.1819	0.0296	0.3810	81.650
随机	0.1500	0.0414	0.3115	78.100

表 4：12个系统的宏观层面稳健性。对于每个系统，我们对四项指标（命中率@10、MRR@10、NDCG@10、MAP@10）分别计算各项指标的平均值、最小值、最大值和变异系数（CV% = （标准差/平均值） × 100），再取平均。更高的“平均值”表示整体有效性更好，而更低的 CV% 表示在钻石–青铜分区之间具有更高的稳定性。

6 结论

我们推出了 **FC-CONAN**。据我们所知，这是首个 对仇恨言论/反叙事数据集，其中两个有限集合——45 条仇恨言论消息和 129 条反叙事——之间所有可能的配对都经过了明确判断。

因此，如果某个配对在某一分区中被标记为适当，这意味着根据该分区的要求，它确实被认定为合适。反之，如果它被标记为不适当，则说明它不符合该分区特定的标准。

四个经过质量控制的分区——DIAMOND、GOLD、SILVER 和 BRONZE——使从业者能够在标注可靠性与语料库规模之间进行权衡。

我们标注的仇恨言论–反叙事对均未出现在原始 CONAN 语料库中；因此，我们的分区补充了数百条此前缺失的适当反叙事，揭示了稀疏标注的 CONAN 语料库中包含多少未标注的适当配对。

值得注意的是，利用所提出的分区，我们确认，在创建数据集时不考虑所有可能的配对，会导致许多恰当的配对未被标注。这会人为地降低推荐系统中大多数评估指标的得分。对于不会因正例增多而受到负面影响的指标，例如 MRR@10 或命中率@10，这意味着在部分标注数据集上评估所得的分数应被视为下界。

我们还表明，基于嵌入的排序器在有效性和应对缺失标注方面的鲁棒性上都优于其他排序器。

总之，评估反叙事检索系统的研究人员应采用像我们这样的密集标注数据集，以避免低估系统性能。

致谢

本研究得到文部科学省“旨在确保生成式人工智能模型透明性与可靠性的研发中心”项目以及日本学术振兴会科学研究费助成事业（JSPS KAKENHI）资助，课题编号为 JP22H00543。我们还衷心感谢慷慨投入时间参与本项目的标注人员和验证人员。Junqueras 部分获得了 UBA BIICC 奖学金计划、Fundar FunDatos 奖学金计划和 NII 国际实习计划的资助。

参考文献

Abelson, H.; Sussman, G. J.; and Sussman, J. 1985. 计算机程序的结构与解释。马萨诸塞州剑桥：麻省理工学院出版社。Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 任务5：Twitter 上针对移民和女性的仇恨言论多语言检测。载于 May, J.; Shutova, E.; Herbelot, A.; Zhu, X.; Apidianaki, M.; and Mohammad, S. M., 编，第*13*届国际语义评测研讨会论文集，54–63。美国明尼苏达州明尼阿波利斯：计算语言学协会。Baumgartner, R.; Gottlob, G.; and Flesca, S. 2001. 使用 Lixto 进行视觉信息提取。载于第*27*届超大型数据库国际会议论文集，119–128。意大利罗马：Morgan Kaufmann。Bonaldi, H.; Dellantonio, S.; Tekiroglu, S. S.; and Guerini, M. 2022年。构建用于仇恨言论反制的对话数据集的人机协作方法。载于*2022*年自然语言处理实证方法会议论文集，8031–8049。计算语言学协会。Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. 用于学习自然语言推理的大型标注语料库。载于 M`arquez, L.; Callison-Burch, C.; and Su, J., 编，*2015*年自然语言处理实证方法会议论文集，632–642。葡萄牙里斯本：计算语言学协会。Brachman, R. J., and Schmolze, J. G. 1985. KL-ONE 知识表示系统概述。《认知科学》9(2)：171–216。Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 2019. CONAN——通过利基众包构建反制叙事：用于回应网络仇恨言论的多语言数据集。载于计算语言学协会第*57*届年会论文集，2819–2829。意大利佛罗伦萨：计算语言学协会。Chung, Y.-L.; Abercrombie, G.; Enock, F.; Bright, J.; and Rieser, V. 2024年。理解用于缓解网络伤害的反仇恨言论。《北欧语言技术期刊》10：30–49。Chung, Y.-L.; Tekiroˇglu, S. S.; and Guerini, M. 2021. 面向仇恨言论知识支撑的反叙事生成。载于计算语言学协会第*59*届年会论文集。线上：计算语言学协会。DataCanary. 2024. Quora问题对。Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. 自动仇恨言论检测与冒犯性语言问题。*AAAI*国际网络与社交媒体会议论文集11. Dolan, B., and Brockett, C. 2005. 自动构建句子改写语料库。载于第三届国际改述研讨会（*IWP2005*）。亚洲自然语言处理联合会。Fanton, Margherita and Bonaldi, Helena and Tekiroˇglu, Serra Sinem and Guerini, Marco. 2021. 人机协作进行数据收集：用于对抗网络仇恨言论的多目标反叙事数据集。载于计算语言学协会第*59*届年会论文集。计算语言学协会。Furman, D.; Torres, P.; Rodriguez, J.; Martinez, L.; Alonso Alemany, L.; Letzen, D.; and Martinez, M. V. 2022. 面向仇恨言论反叙事的简约论证标注。Furman, D.; Torres, P.; Rodr´ıguez, J.; Letzen, D.; Martinez, M.; and Alemany, L. 2023. 低资源方法中的高质量论证信息可改进反叙事生成。载于 Bouamor, H.; Pino, J.; and Bali, K., 编，计算语言学协会论文集：*EMNLP 2023*年研究成果，2942–2956。新加坡：计算语言学协会。Gao, L., and Huang, R. 2017. 使用上下文感知模型检测网络仇恨言论。载于 Mitkov, R., and Angelova, G., 编，自然语言处理最新进展国际会议论文集，*RANLP 2017*，260–266。保加利亚瓦尔纳：INCOMA有限公司。Gillespie, T. 2018. 互联网的守护者：塑造社交媒体的平台、内容审核与隐秘决策。耶鲁大学出版社。Goffredo, P.; Basile, V.; Cepollaro, B.; and Patti, V. 2022. Counter-TWIT：用于真实生态语境中网络反仇恨言论的意大利语语料库。载于 Narang, K.; Mostafazadeh Davani, A.; Mathias, L.; Vidgen, B.; and Talat, Z., 编，第六届网络虐待与伤害研讨会（*WOAH*）论文集，57–66。美国华盛顿州西雅图（线上线下混合）：计算语言学协会。Gottlob, G.; Leone, N.; and Scarcello, F. 2002. 超树分解与可处理查询。《计算机与系统科学杂志》64(3):579–627。Gottlob, G. 1992. 非单调逻辑的复杂性结果。《逻辑与计算杂志》2(3):397–425。Klie, J.-C.; Eckart de Castilho, R.; and Gurevych, I. 2024. 分析真实环境中的数据集标注质量管理。《计算语言学》50(3):817–866。

Koehn, P. 2005. Europarl：用于统计机器翻译的平行语料库。载于机器翻译峰会 *X* 论文集，79–86。

Levesque, H. J. 1984a. 知识表示功能方法的基础。人工智能23(2):155–212。

Levesque, H. J. 1984b. 一种关于隐式信念与显式信念的逻辑。载于第四届全国人工智能会议论文集，198–202。美国得克萨斯州奥斯汀：美国人工智能协会。

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017年。DailyDialog：一个由人工标注的多轮对话数据集。载于 Kondrak, G. 和 Watanabe, T. 编，第八届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），986–995。中国台湾台北：亚洲自然语言处理联合会。

Nebel, B. 2000. 关于命题规划形式体系的可编译性与表达能力。人工智能研究杂志12:271–315。

OpenAI。2024a。GPT-4o 技术报告。https://openai.com/research/gpt-4o。

OpenAI。2024b。新型嵌入模型和 API 更新。https://openai.com/index/new-embedding-models-and-api-updates/。访问日期：2025年8月2日。

Rajpurkar, P.；Zhang, J.；Lopyrev, K.；和 Liang, P. 2016。SQuAD：用于文本机器理解的 100,000+ 个问题。载于 Su, J.；Duh, K.；和 Carreras, X. 编，《2016年自然语言处理实证方法会议论文集》，2383–2392。美国得克萨斯州奥斯汀：计算语言学协会。

Reimers, N. 和 Gurevych, I. 2019。Sentence-BERT：使用孪生 BERT 网络进行句子嵌入。载于《*2019*年自然语言处理实证方法会议及第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集》，3982–3992。

Robertson, S. 和 Zaragoza, H. 2009。《概率相关性框架：BM25 及其扩展》。《信息检索基础与趋势》3：333–389。Salles, I.；Vargas, F.；和 Benevenuto, F. 2025。《HateBRX-plain：一个包含人工标注理由的基准数据集，用于巴西葡萄牙语可解释仇恨言论检测》。载于 Rambow, O.；Wanner, L.；Apidianaki, M.；Al-Khalifa, H.；Eugenio, B. D.；和 Schockaert, S. 编，《第31届国际计算语言学会议论文集》，6659–6669。阿联酋阿布扎比：计算语言学协会。Salton, G. 和 Buckley, C. 1988。《自动文本检索中的词项加权方法》。《信息处理与管理》24(5)：513–523。

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; 和 Weber, I. 2021. 分析在线社交媒体中仇恨的目标媒体。《国际人工智能促进会网络与社交媒体会议论文集》10(1):687–690.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. MPNet：用于语言理解的掩码与置换预训练。载于神经信息处理系统国际会议第*34*届会议论文集，NIPS '20。美国纽约州雷德胡克：Curran Associates Inc.

Vallecillo Rodr´ıguez, M. E.; Cantero Romero, M. V.; Cabrera De Castro, I.; Montejo R´aez, A.; and Mart´ın Valdivia, M. T. 2024年。CONAN-MT-SP：一个利用 GPT 模型进行反叙事的西班牙语语料库。载于 Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N. 编，*2024*年计算语言学、语言资源与评测联合国际会议论文集（*LREC-COLING 2024*），3677–3688。意大利都灵：ELRA and ICCL。

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. MiniLM：用于预训练变换器任务无关压缩的深度自注意力蒸馏。载于第*34*届神经信息处理系统国际会议论文集，NIPS '20。美国纽约州雷德胡克：Curran Associates Inc。

Waseem, Z. 和 Hovy, D. 2016。仇恨符号还是仇恨者？Twitter上的仇恨言论检测预测特征。载于 Andreas, J.; Choi, E.; 和 Lazaridou, A. 编，*NAACL*学生研究研讨会论文集，88–93。加利福尼亚州圣迭戈：计算语言学协会。

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; 和 Kumar, R. 2019。预测社交媒体中冒犯性帖子的类型和目标。载于 Burstein, J.; Doran, C.; 和 Solorio, T. 编，*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长短篇论文），1415–1420。明尼阿波利斯，明尼苏达州：计算语言学协会。

Zhuang, L.; Wayne, L.; Ya, S.; 和 Jun, Z. 2021. 一种通过后训练实现稳健优化的 BERT 预训练方法。载于 Li, S.; Sun, M.; Liu, Y.; Wu, H.; Liu, K.; Che, W.; He, S.; 和 Rao, G. 编，《第*20*届中国计算语言学全国会议论文集》，1218–1227。中国呼和浩特：中国中文信息学会。