SafeWebUH 参加 SemEval-2023 任务 11：学习贬损性文本中的标注者分歧：直接训练与聚合的比较

Sadat Shahriar休斯敦
大学 sshahriar
@uh.edu

Thamar Solorio休斯
敦大学
tsolorio@uh.edu

arXiv:2305.01050v1 [cs.CL] 1 May 2023

摘要

主观性和意见差异是关键的社会现象，在贬损性文本内容的标注和检测过程中考虑这些因素至关重要。在本文中，我们使用 SemEval-2023 任务 11 提供的四个数据集，并对 BERT 模型进行微调，以捕捉标注中的分歧。我们发现，与直接使用软标签进行训练相比，单独的标注者模型和聚合方法使交叉熵分数平均降低了 0.21。我们的研究结果进一步表明，标注者元数据使交叉熵分数平均降低了 0.029。

1 引言

尽管网络空间充斥着贬损性文本，但其解读的主观性往往需要一种能够捕捉读者分歧的系统。带分歧学习（Le-Wi-Di）任务旨在根据标注者如何对文本进行分类来学习标注者分歧（Leonardellli 等，2023 年）。近期研究发现，几乎每项标注任务都包含广泛的分歧（Du-mitrache 等，2019年；Pavlick 和 Kwiatkowski，2019年）。评分者的主观性和偏见性，以及自然语言理解中的其他因素，使得通过标注学习分歧至关重要（Uma 等，2021年a）。在本研究中，我们比较了两种分歧学习策略：软标签的分歧目标学习，以及采用后聚合的特定标注者学习，并使用 BERT 模型。此外，我们利用特定标注者的元数据，以捕捉标注者在贬损性内容上的分歧。

自社交媒体兴起以来，网络上涌现出海量内容，冒犯性文本的数量也随之增加，例如仇恨言论、厌女、性别歧视和辱骂性内容。这一问题显著加剧。为应对这一问题，开展了多项研究，例如，Burnap 和 Williams 研究了由伦敦鼓手 Lee Rigby 遇害事件引发的推文中的线上仇恨言论（Burnap 和 Williams，2015年）。Xu 等人将社交媒体中的网络欺凌问题建模为一项自然语言处理任务（Xu 等人，2012年）。类似的研究还包括 Warner 和 Hirschberg 2012；Silva 等人 2016年；Gitari 等人 2015年。然而，与冒犯性和毒性等社会现象检测相关的任务通常具有主观性（Koco´n 等人，2021年）。近期一项针对美国成年人的综述指出，半数受访者认为“很难知道他人可能会觉得哪些内容具有冒犯性”，而且大多数受访者承认，人们对于哪些内容被视为性别歧视或种族主义存在分歧（皮尤研究中心，访问日期：2022-12-03）。为此，我们旨在开发一个能够捕捉贬损性文本中主观分歧的系统。

Le-Wi-di 任务中的四个数据集都包含标注者特定标签，以及聚合硬标签（多数投票）和软标签（标签的平均值）。尽管用于分歧建模的系统应经过训练，以估计软标签，但尚不清楚直接使用软标签训练，还是聚合标注者标签，是更好的方法。因此，我们提出第一个研究问题（**Q1**）：在分歧建模中，标注者特定分类模型与事后聚合能否优于直接对软标签进行回归的方法？此外，我们还探讨了标注者元数据，这些元数据说明某位标注者如何标注其他相关文本，并提出问题（**Q2**）：标注者元数据能否改进分歧建模？为回答这些问题，我们比较了基于 BERT 的面向分歧学习（回归）和后聚合学习（分类），并探索了将标注者元数据纳入分歧建模的不同策略。然而，由于

标注者之间存在不一致，且缺乏元数据，我们仅将比较范围限制在两个数据集上。

我们的工作具有几项重要意义。首先，我们的模型能够捕捉冲突，因此适用于对有争议的社会现象和公众意见进行建模。因此，它可以用于对文本歧义中的歧义进行建模。此外，我们对纳入标注者元数据的探索有助于理解读者在不同语境下的感知和观点。最后，可以通过增强评分者之间的透明度和问责性，将其作为多评分者标注过程中的一种质量控制手段。实现我们工作的代码可在此处获取：

h

ttps://github.co

m/sadat1971/Le-Wi-Di-SemEval-23

2 数据集和任务描述

SemEval'23 任务 11 包含四个处理贬损性文本的数据集。其中三个数据集为英语，*ArMIS* 为阿拉伯语。除软标签和硬标签外，每个数据集还包含一些元数据。下面对它们进行简要介绍。

该多领域一致性 （MD）数据集包含来自三个领域的推文：BLM、选举和 COVID-19（Leonardelli 等人，2021年）。共有 819 名标注者通过亚马逊土耳其机器人为所有推文进行标签标注。每条推文均随机选取五名标注者对其冒犯性进行标注。训练集包含 6,592 条推文，来自练习阶段的开发集包含 1,104 条推文，来自评估阶段的测试集包含 3,057 条推文。

该***HS-Brexit*** 数据集包含与英国脱欧相关的推文，以及六名标注者的标注结果（英国三名穆斯林移民组成的目标组和三名成员组成的对照组）（Akhtar 等人，2021年）。每名标注者都为推文标注了仇恨言论，这也是该任务的目标类别。他们还对推文是否具有冒犯性和攻击性进行了标注。训练集、开发集和测试集分别包含 784、168 和 168 条推文。

厌女和性别歧视在 ***ArMIS***数据集中进行了标注，由三名标注者进行评定（温和派女性、自由派女性和保守派男性）（Almanea 和 Poesio，2022年）。训练集、开发集和测试集分别包含 657、141 和 145 条推文。

ConvAbuse 数据集记录了一名用户与两个对话代理之间的对话，且至少有两名标注者对该对话进行了标注。用于辱骂性（Cercas Curry 等人，2021年）。该数据集还提供了对对话是否具有性别歧视、明确、隐含、智力、种族主义、跨性别恐惧、恐同特征以及辱骂目标的标签。训练集、开发集和测试集分别包含 2,398、812 和 840 条推文。

3 系统描述

对于文本数据，我们使用一种名为 BERT 的预训练语言表征模型（Devlin 等人，2019年）。由于 BERT 仅使用英语数据进行训练，为处理 ArMIS 任务，我们使用 Arabic-BERT （Safaya 等人，2020年）。图 1 展示了系统描述。为解决**Q1**，我们比较了两种技术——后聚合和面向分歧的目标学习，同时还研究了元数据对解决 **Q2**的影响。性能通过 F1 分数和交叉熵（CE）分数进行衡量。

3.1 后聚合

在后聚合 （Post-Agg）方法中，我们分别训练模型来学习每位标注者的标注模式。首先，对 BERT 模型进行微调，使其学习目标类别，并获取所有标注者的 softmax分数 S。接下来，我们处理元数据以提取重要信息。对于 HS-Brexit 数据集，除了对仇恨言论进行标注外，每位标注者还对推文的冒犯性和攻击性进行了标注，这些信息随数据集一并提供。我们根据标注者对推文的冒犯性和攻击性标注，计算该推文被标记为仇恨言论的概率，并将其记为 P。然后，为每条推文计算软标签 SL，如下所示：

$$\hat{SL}(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i + w * P_i}{1+w} \quad (1)$$

其中，N 是标注者的数量。由于 S_i 和 P_i 都是预测得到的软标签，我们计算它们的加权平均值，并选择 w，使其在开发集上的交叉熵分数最低且 F1 分数最高。

3.2 针对分歧的学习

虽然 Post-Agg 方法考虑从每位标注者的标注中学习，但分歧目标学习（Dis-Learning）方法只从聚合标签中学习。首先，使用估计软标签的下游回归任务对 BERT 模型进行微调，并获得预测变量，

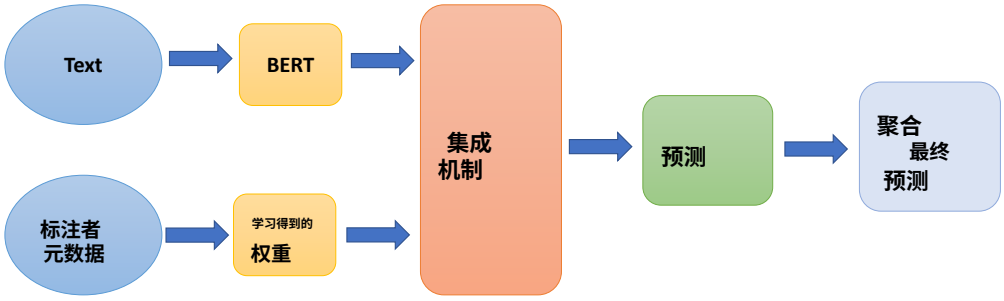


图1：将文本输入预训练BERT模型，并针对下游任务进行微调。对于后聚合方法，BERT模型的下游分类任务是预测每位标注者的标签。将每个标注者模型输出的softmax值与元数据集成起来，生成针对各标注者的软标签和硬标签预测。对于分歧目标学习方法，BERT下游任务是直接学习软标签。通过回归方法实现集成机制，以学习最终预测。

SL_{BERT} 接下来，我们测量 av - 整个数据集中所有标注者对每个元数据项的平均评分。例如，在 HS - 在英国脱欧数据集中，如果两个标注者将一条推文标记为冒犯性，而四个标记为非冒犯性，则该推文的平均元数据（冒犯性评分）为 $2/6 = 0.33$。接下来，我们仅依据可用的平均元数据评分训练一个线性回归模型，以预测软标签。

$$SL_{meta} = b_0 + b_1 * M_1 + b_2 * M_2 \quad (2)$$

b_0，b_1，b_2 由线性回归模型训练得出，M_1 和 M_2 是两个元数据评分。对于 HS-Brexit，我们使用冒犯性和攻击性指标的平均值。对于 ConvAbuse 数据集，在十二个元数据标签中，我们选取与软标签值相关系数最高的两个：明确和 目标系统。最后，我们通过对 SL_{BERT} 和SL_{meta}取平均，得到 SL 。

4 实验设置

我们所有的模型都使用 BERT 的“bert-base-uncased”版本（在 ArMIS 中使用“bert-base-arabic”）。在回归和分类任务中，我们都采用两层全连接网络进行微调。在实践阶段，我们通过三折交叉验证从所有组合中选取超参数；在评估阶段，则在发布的验证集上进行选择。隐藏层大小和丢弃率分别从 {32, 64, 128, 256} 和 {.1, .3, .5} 中选取。学习率从 {5e-4, 1e-5, 5e-5, 1e-6} 中选取。由于 GPU 资源有限，我们将批量大小设得较小，并从 {8, 16} 中选取。由于 BERT 模型很容易在数据上过拟合，我们将训练轮次数设为 2 到 4 轮。然而，对于“arabic-bert-base”，模型表现不稳定，因此我们训练最多 10 轮。所有情况下，优化器均使用 AdamW（Loshchilov 和 Hutter，2017）。所有实验均使用 PyTorch 1.11.0 版本（Paszke 等人，2019年）。

为了评估对分歧的捕捉效果，我们使用竞赛提供的交叉熵分数。若目标软标签为 T，预测的软标签为 P，数据集大小为 D，则交叉熵（CE）的计算方式如下：

$$CE = -\frac{1}{D}\sum_{i=1}^{D} T_i * log(P_i + 1e - 9) \quad (3)$$

我们进一步报告硬标签上的 F1 分数（微平均），以评估模型在多数投票最终预测任务中的表现。

5 结果与讨论

表 1显示，对于 HS-Brexit 数据集，Post-Agg 方法并未比 Dis-Learning 方法提高 F1 分数。然而，与 Dis-Learning 方法相比，Post-Agg 方法能够将交叉熵分数降低 0.1400。当使用元数据时，降幅甚至更大（达到 0.1958）。类似地，对于 ArMIS 数据集，与 Post-Agg 方法相比，Dis-Learning 方法具有更高的 F1 分数，同时 Dis-Learning 方法的交叉熵分数更低（降低了 0.3070）。

我们进一步探究 Post-Agg 方法为何在捕捉分歧方面表现更好。由于分歧学习方法不将

	后聚合		后聚合-meta		分歧学习		分歧学习-meta	
数据集	F1	CE	F1	CE	F1	CE	F1	CE
MD	–	–	–	–	**0.8266***	**0.5076***	–	–
HS-Brexit	0.8810	0.1686	**0.9167**	**0.0834**	0.8869	0.3086	0.9107*	0.2792*
ArMIS	0.7211	**0.2683**	–	–	**0.7586***	0.5753*	–	–
ConvAbuse	–	–	–	–	0.9321*	0.2364*	**0.9667**	**0.0688**

表 1：比较四个数据集中各类分歧建模方法的表现。由于 MD 和 ConvAbuse 数据集缺少一致的标注者，因此未报告其后聚合结果。此外，MD 和 ArMIS 缺少特定于标注者的元数据，因此也未报告纳入元数据后的表现。每个数据集中的最佳表现以粗体数字标示，而在 Le-Wi-Di 任务中提交的表现则以星号（*）标示。

个体标注者纳入考虑，因此它主要近似文本具有贬损性的“程度”。相反，Post-Agg 方法分别考虑每位标注者，并学习其标注模式，之后再进行聚合。因此，分歧学习只能依赖文本数据，这使其任务比 Post-Agg 更加困难。然而，在现实情况下，标注者可能并不一致（如 MD 和 ConvAbuse 数据集所示），或者需要训练大量模型，从而使 Post-Agg 技术难以实施。因此，如果数据集中的少数标注者始终保持一致，Post-Agg 方法更适合对分歧进行建模。由此，**Q1** 得到了解答。

接下来，结果表明，与不使用标注者元数据相比，使用标注者元数据能够提升表现（表 1）。使用元数据后，HS-Brexit 数据集的交叉熵分数在后聚合和分歧学习方法中分别降低了 0.0852 和 0.0294。同样，对于 ConvAbuse 数据集，标注者元数据有助于将交叉熵分数降低 0.1676。元数据包含标注者有用的标注模式，从而改善了学习过程。值得注意的是，我们没有使用 MD 和 ArMIS 的元数据，因为其中不包含与标注者相关的标注信息。因此，**Q2**得到了回答。

在 MD、HS-Brexit、ArMIS 和 ConvAbuse 数据集中，我们的结果分别排名第 7、第 9、第 11 和第 12。总体而言，我们在交叉熵分数类别中排名第 9，在 F1 分数类别中排名第 8。

错误分析 最后，我们重点分析本研究中的错误。我们发现，两种方法在预测那些不使用俚语或脏话、但仍被多数人投票标记为冒犯性文本时，都经常出错。例如，五位标注者中有三位将以下句子标记为冒犯性（软标签 0.60）：*#TonyBobulinski # MAGA2020#MAGA #ChangeY-ourVoteToTrump# BidenCrimeFamily #BidenHar-ris2020 # BidenCares #LaptopFromHell Joeis go-ing down.* <网址>。然而，我们的模型预测的软标签为 0.15。同样，包含脏话但不一定具有冒犯性的推文，有时会被我们的模型误判为仇恨言论。例如，推文：来自多产而才华横溢的—— <用户># *BlackLivesMatter #fucktrump* <网址> 被五位标注者中的三位标记为非冒犯性；然而，由于其中一个标签含有亵渎性语言，我们的模型预测软标签为 0.85。

6 相关工作

尽管大多数人工智能学习仍然基于每个项目都存在单一解释这一假设，但越来越多的研究正致力于构建不依赖这一假设的学习方法（Uma 等，2021年b）。在自然语言处理领域，标注者分歧是一种常见现象（Poesio 和 Artstein，2005年；Recasens 等，2011年）。这种分歧可能源于标注者错误或界面问题（Plank 等，2014年）、明确或隐含歧义（Poesio 和 Artstein，2005年）、项目难度（Zaenen 等，2005年）以及主观性（Akhtar 等，2019年）。然而，从较为简单的任务（如词性标注）（Plank 等，2014年）到主观任务（如情感分析），语义角色标注同样涉及标注者分歧（Kenyon-Dean 等，2018年；Dumitrache 等，2019年）。因此，研究人员主张在标注过程中考虑分歧，并保留隐性歧义（Recasens 等，2012年；Poesio 和 Artstein，

2005）为此，我们探索了面向贬损性文本的带分歧学习任务。

本次竞赛的上一届于2021年举办，组织者使用自然语言和图像分类任务来处理标签标注中的分歧（Uma 等人，2021年a）。获胜团队采用了尖锐度感知最小化技术（SAM），并使用以 BERT 为基线模型的特殊神经网络层 softmax 众包层（ Osei-Brefo 等人，2021 年）。虽然 SAM 架构主要用于 CIFAR-10（图像分类），但众包层架构旨在将标签与每位标注者对应起来。由于本次竞赛仅涉及文本，我们对 BERT 模型进行微调，并利用标注者元数据来捕捉分歧。

7 结论

随着社交媒体内容日益普及，互联网已成为贬损性文本滋生的温床。然而，由于人类感知和观点各不相同，标注者往往无法就某段文本是否具有贬损性达成一致。因此，在建模过程中保存软标签并捕捉标注者分歧至关重要。我们的工作比较了直接使用软标签进行训练与采用标注者特定模型并进行后期聚合这两种方法。我们发现，当存在观点一致的标注者时，后一种方法可能更有帮助。此外，我们的实验也证明，整合标注者元数据有益于模型表现。我们的工作为未来研究提供了多种可能方向，例如：

- 我们只使用一种基于 Transformer 的方法 BERT 进行建模。未来，我们计划使用 RoBERTA、ELECTRA 和 XLMNet。
- 我们发现仇恨言论与冒犯性之间存在很强的相关性。因此，我们计划研究跨数据集性能的表现。这类实验也将有助于提高模型的泛化能力。
- 由于语言会随着社会背景和其他现象而演变，因此纳入持续学习（CL）技术并研究标注过程中的分布偏移至关重要。未来，我们计划将持续学习纳入研究工作。

参考文献

访问日期：2022-12-03。 了解什么是冒犯性的挑战。皮尤研究中心。

Sohail Akhtar、Valerio Basile 和 Viviana Patti。2019年。仇恨言论标注中的极化新度量。载于 *AI* IA 2019*——人工智能进展：意大利人工智能协会第十八届国际会议论文集，意大利伦代，*2019*年*11*月*19*日至*22*日，论文集第*18*卷，第588–603页。施普林格。

Sohail Akhtar、Valerio Basile 和 Viviana Patti。2021年。谁的观点重要？用于在辱骂性语言检测中识别仇恨言论受害者观点的视角感知模型。*arXiv*预印本 *arXiv:2106.15896*。

Dina Almanea 和 Massimo Poesio。2022。Armis——包含标注者主观分歧的阿拉伯语厌女与性别歧视语料库。载于第十三届语言资源与评估会议论文集，第 2282–2291 页。

Peter Burnap 和 Matthew Leighton Williams。2015。Twitter 上的网络仇恨言论：机器分类和统计建模在政策制定与决策中的应用。《政策与互联网》，7：223–242。

Amanda Cercas Curry、Gavin Abercrombie 和 Ver- ena Rieser。2021年。ConvAbuse：对话式人工智能中细致入微的辱骂检测数据、分析与基准。载于*2021*年自然语言处理实证方法会议论文集，第7388–7403页，线上，蓬塔卡纳，多米尼加共和国。计算语言学协会。

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

Anca Dumitrache、Lora Aroyo 和 Chris Welty。2019年。一个包含歧义的众包框架消歧语料库。载于《计算语言学协会北美分会 *2019*年会议论文集：人类语言技术》第*1*卷（长篇论文和短论文），第2164–2170页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Njagi Dennis Gitari、Zhang Zuping、Zuping Zhang、HanyurwimfuraDamien 和 Jun Long。2015年。基于词典的仇恨言论检测方法。载于国际多媒体与普适工程会议。

Kian Kenyon-Dean、Eisha Ahmed、Scott Fujimoto、Jeremy Georges-Filteau、Christopher Glasz、Barleen Kaur、Auguste Lalande、Shruti Bhanderi、Robert Belfer、Nirmal Kanagasabai、Roman Sarrazingen- dron、Rohit Verma 和 Derek Ruths。2018年。情感分析：事情并不简单！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*1*卷（长篇论文），第1886–1895页，新奥尔良，路易斯安那州。计算语言学协会。

Jan Koco´n, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski 和 Przemysław Kazienko。2021年。在自然语言处理中为主观任务学习个人偏见与表示。载于*2021* 年 *IEEE* 国际数据挖掘会议（*ICDM*），第 1168–1173 页。

Elisa Leonardelli、Stefano Menini、Alessio Palmero Aprosio、Marco Guerini 和 Sara Tonelli。2021年。求同存异：使用标注者分歧标注冒犯性语言数据集。载于*2021* 年自然语言处理中的经验方法会议论文集，第 10528–10539 页，线上和多米尼加共和国蓬塔卡纳。计算语言学协会。

Elisa Leonardellli、Gavin Abercrombie、Dina Almanea、Valerio Basile、Tommaso Fornaciari、Barbara Plank、Massimo Poesio、Verena Rieser 和 Alexandra Uma。2023 年。SemEval-2023 任务 11：带分歧学习（LeWiDi）。载于第 *17* 届语义评估国际研讨会论文集，加拿大多伦多。计算语言学协会。

Ilya Loshchilov 和 Frank Hutter。2017年。解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*。

Emmanuel Osei-Brefo、Thanet Markchom 和 Huizhi Liang。2021年。UOR 在 SemEval-2021任务12中的工作：关于众包标注；通过带分歧学习优化众包真值。收录于 第*15*届语义评测国际研讨会论文集（*SemEval-2021*），第1303–1309页，线上。计算语言学协会。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Raison、Alykhan Te-jani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。2019年。Py-torch：一种命令式风格的高性能深度学习库。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Gar-nett 编，《神经信息处理系统进展》第*32*卷，第8024–8035页。Curran Associates, Inc.

Ellie Pavlick 和 Tom Kwiatkowski。2019年。人类文本推理中固有的分歧。计算语言学协会会刊，7：677–694。

Barbara Plank、Dirk Hovy 和 Anders Søgaard。2014年。从语言学角度看，这是值得商榷，还是纯粹错误？收录于 第*52*届计算语言学协会年会论文集（第*2*卷：短论文），第507–511页，马里兰州巴尔的摩。计算语言学协会。

Massimo Poesio 和 Ron Artstein。2005年。重新审视照应标注的可靠性：将歧义纳入考量。载于《语料标注前沿 *II*：空中楼阁》研讨会论文集，第 76–83 页，密歇根州安娜堡。计算语言学协会。

Marta Recasens、Eduard Hovy 和 M. Antònia Martí。2011年。同一性、非同一性与近同一性：应对共指的复杂性。*Lingua*，121(6):1138–1152。

Marta Recasens、M. Antònia Martí 和 Constantin Orasan。2012年。从共指分歧中标注近同一性。载于《第八届国际语言资源与评估会议（*LREC'12*）论文集》，第 165–172 页，土耳其伊斯坦布尔。欧洲语言资源协会（ELRA）。

Ali Safaya、Moutasem Abdullatif 和 Deniz Yuret。2020年。KUISAIL 在 SemEval-2020第12项任务：用于社交媒体冒犯性言论识别的 BERT-CNN。载于第十四届语义评估研讨会论文集，第2054–2059页，巴塞罗那（线上）。国际计算语言学委员会。

Leandro Silva、Mainack Mondal、Denzil Correa、Fab- rício Benevenuto 和 Ingmar Weber。2016年。分析在线社交媒体上的仇恨目标。载于第十届国际 *AAAI* 网络与社交媒体会议。

Alexandra Uma、Tommaso Fornaciari、Anca Du mi- trache、Tristan Miller、Jon Chamberlain、Barbara Plank、Edwin Simpson 和 Massimo Poesio。2021年a。SemEval-2021任务12：从分歧中学习。载于第*15*届语义评估国际研讨会（*SemEval-2021*）论文集，第338–347页，线上。计算语言学协会。

Alexandra N Uma、Tommaso Fornaciari、Dirk Hovy、Silviu Paun、Barbara Plank 和 Massimo Poesio。2021年b。从分歧中学习：一篇综述。人工智能研究期刊，72:1385– 1470。

William Warner 和 Julia Hirschberg。2012年。在万维网上检测仇恨言论。载于 第二届社交媒体语言研讨会论文集，第 19–26 页，蒙特利尔，加拿大。计算语言学协会。

Jun-Ming Xu、Kwang-Sung Jun、XiaojinZhu 和 Amy Bellmore。2012年。从社交媒体中的欺凌痕迹中学习。载于*2012*年计算语言学协会北美分会人类语言技术会议论文集，第656–666页，蒙特利尔，加拿大。计算语言学协会。

Annie Zaenen、Lauri Karttunen 和 Richard Crouch。2005年。局部文本推理：能否对其进行定义或界定？载于计算语言学协会语义等价与蕴涵实证建模研讨会论文集，第31–36页。