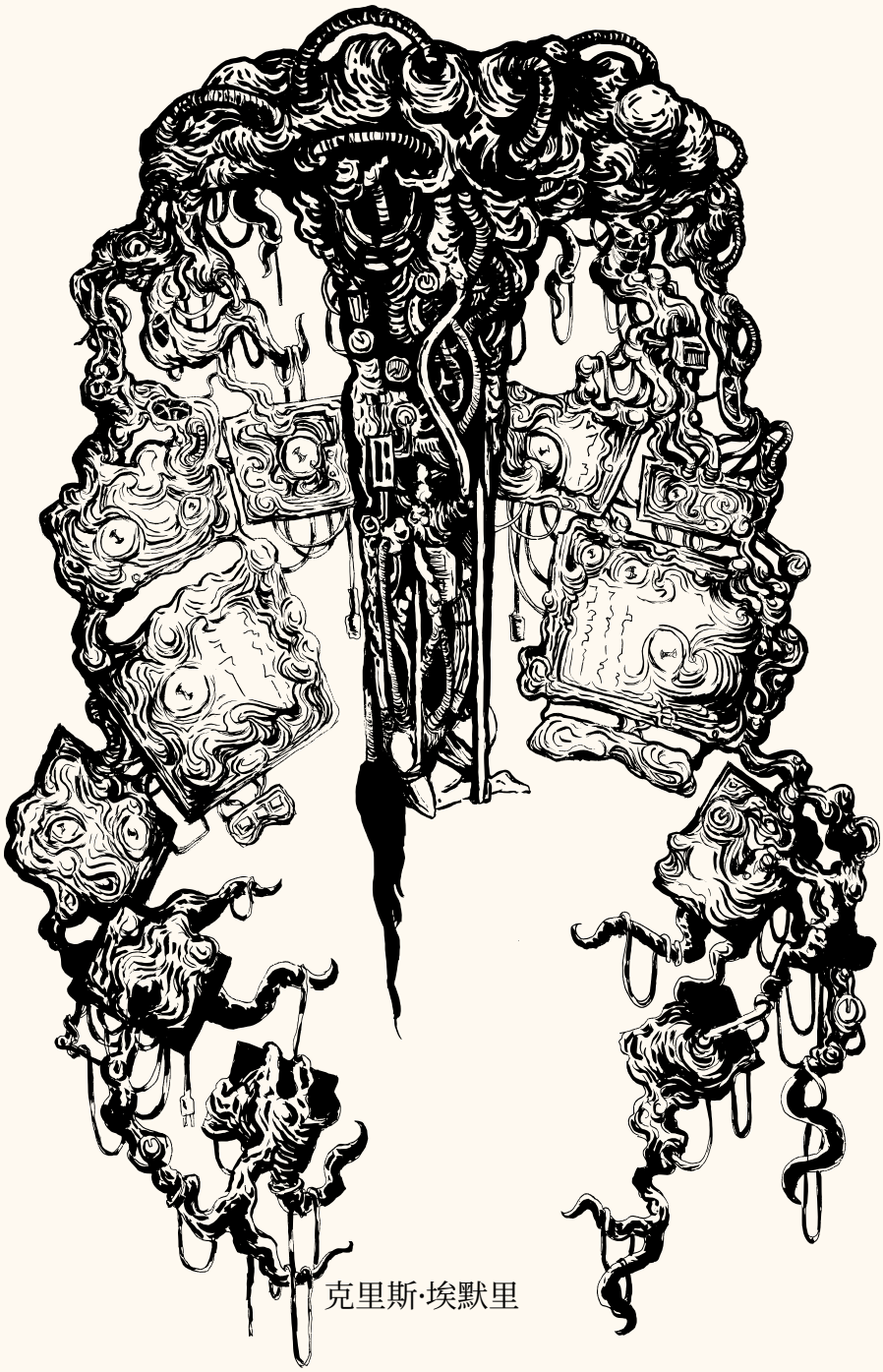
以用户为中心的自然语言处理安全性

arXiv:2301.04230v1 [cs.CL] 10 Jan 2023



克里斯·埃默里

自然语言处理中的以用户为中心的安全

为获得蒂尔堡大学博士学位，根据校长 W.B.H.J. van de Donk 教授的授权，将于2023年1月13日（星期五）13:30，在大学礼堂公开答辩，答辩委员会由博士学位授予委员会指定。

作者 Christian David Emmery，出生于埃因霍温

∞

指导教授：教授、博士 E.O. Postma（蒂尔堡大学）；
教授、工程博士 W. Daelemans（安特卫普大学）

联合导师 G.A. Chrupała 博士（蒂尔堡大学）

答辩委员会成员：M.F. Moens 教授、博士（鲁汶大学）；W. van den Heuvel 教授、博士（JADS）；M. Nissim 教授、博士（格罗宁根大学）；A. Gatt 博士（乌得勒支大学）；M. Potthast 博士（莱比锡大学）

自然语言处理中的以用户为中心的安全 Chris Emmery 博士论文 蒂尔堡大学，2023

封面：Ambika Kirkland（@sendrinon）
设计：Chris Emmery（@cmry）印刷：Ridderprint（ridderprint.nl）

第 3 章是 AMiCA（IWT SBO 项目 120007）的一部分，由政府创新与科学技术机构资助。

Chris Emmery 的《自然语言处理中的以用户为中心的安全》采用 知识共享署名 - 非商业性使用 - 相同方式共享 4.0 国际许可协议。

献给 *Owen*

目录

零. 引言 3一. 通过不变性进行文本风格混淆 21二. 真实场景中的对抗性文体分析 45三. 网络欺凌检测的当前局限 73四. 网络欺凌分类器对扰动很敏感 125五. 讨论与结论 153

致谢 170

摘要 176

参考文献 181



本章基于以下作品：

Chris Emmery、Grzegorz Chrupała 和 Walter Daelemans。2017年。将简单查询作为远程标签，以预测 Twitter 上的性别。见《第三届噪声用户生成文本研讨会论文集》，第 50–55 页，丹麦哥本哈根。计算语言学协会

Chris Emmery、Ákos Kádár、Travis J. Wiltshire 和 Andrew T. Hen- drickson。 2019年。 迈向计算认知建模的复现：机器学习视角。 《计算脑与行为》， 2（3）： 242–246

部分内容取自Emmery 等人(2017年)，或作为第0.1节所述观点的基础。部分内容原文照录自 Emmery 等人(2019年) 的论文，收录于第0.3节。

0

引 HE引言

互联网已经无处不在，以至于有人可能会认为，撰写引言实在是多此一举。毕竟，互联网在人们的生活中发挥着至关重要的作用。尤其是年轻一代（我们的“数字原住民”），他们围绕互联网安排个人生活和职业生活；按照长期以来流行的观点，他们理应因此变得技术非常娴熟。然而，这种相当乐观的看法似乎需要冷静地重新审视。事实上，不同世代的技术熟练度仍有差异，并且与社会人口因素密切相关（Hargittai，2010年；Elueze and Quan-Haase，2018年；Brashierand Schacter，2020年）。此外，随着系统复杂性的提升，如今对技术的要求也日益提高：前所未有的勒索软件攻击、虚假信息的传播，以及通过机器人影响选举，都是社会技术安全问题的实例；在这些问题上，人们对“互联网”的判断是最薄弱的一环。

因此，介绍互联网的现状以及用户如何利用它，似乎更有意义。

互联网

作为一个自由的[1]平台，其唯一目的在于去中心化信息共享，这一时代早已过去（Morrison，2009）。社交网络极其重视大规模通信，尤其是在移动设备问世之后。由于互联网内容历来大多免费，[2]我们付出的代价是（字面意义上的）毕生个人数据；这些数据被变现，用来为广告网络提供资金，使其得以维持“免费”并进一步发展。这一转变催生了一种被称为监视资本主义的机制（Zuboff，2015年）：信息换取服务，点击带来收入。为了获得方便（且高效）的服务，我们很快便会放弃对个人数据的控制（Kehr 等人，2015年）。这形成了强者愈强的局面：提供大量（快速）数据驱动型服务、且无需用户具备技术专长的平台，吸引了大多数用户。互联网的硬件、软件和服务日益集中于这些平台之下（Evans，2017年；Moura 等人，2020年），因此它们在处理我们的数据时可能已经大到难以监管（Arogyaswamy，2020年）。

目前，我们对这少数几家大型实体的广泛依赖，已使我们起初认定属于私人领域的个人信息与我们认为合理地涉及机构和企业的信息之间的界线变得模糊，甚至消失（Constantiou 和 Kallinikos，2015年）。与此同时，我们最初仅打算将数据用于相关服务，但数据被用于其他用途这一事实，也在无意间变得隐而不彰。若要举一个典型例子：剑桥分析利用一个包含 8700 万个 Facebook 个人资料的数据库中的个人信息来定向操纵选民（这些信息是通过一名以科学研究为掩护开展业务的第三方应用开发者获取的），无疑是恰当的例子（Isaak 和 Hanna，2018年）。

[1] 如同“言论自由”中的“自由”。
[2] 如同“免费啤酒”中的“免费”。

正如本学位论文后文将逐渐展现的那样，人工智能（AI），更确切地说，其下属领域机器学习（ML——即在数据上训练的预测算法），在掩盖我们的数据如何被改作他用方面发挥着重要作用（Wachter 和 Mittelstadt，2018年；Manheim 和 Kaplan，2019年）。深度学习（DL，Schmidhuber，2015年）目前仍势头正盛，而移动消费设备的计算能力也已接近台式机硬件水平（Lemley 等人，2017年；Ignatov 等人，2019年）。因此，索引和按物体对照片分类（Haralick 等人，1973年；Lowe，1999年）、通过人脸识别解锁设备（Turk 和 Pentland，1991年）、追踪睡眠（Schaltenbrand 等人，1996年）和活动（Abowd 等人，1998年），以及控制能源消耗（Srivastava 等人，1996年）等诸多复杂任务，已进入世界上大多数人的口袋。上述所有应用都将机器学习用于高度敏感的数据（也就是说，它们会处理相关用户数据，以训练模型执行相应任务）。这些模型往往会随移动设备一同提供（有时还会嵌入设备中并在设备上主动运行），而设备所有者对此几乎毫不知情。这并非偶然：几乎所有主要的移动技术企业都已设立人工智能实验室，并吸引了众多杰出的研究人员来协助内部开发。

机器学习如今已具有巨大的商业价值，因此与其他信息技术领域一样，呈现出集中化趋势。同步发生的、快速推进且由行业驱动的研究与开发显著推动了该领域的发展；其提升不仅体现在研究产出方面，也体现在实验规模、产出的影响力与可访问性、受试者数量，以及总体计算规模等方面。这一文化转变引发了科学界和社会的关切，尤其是研究是否会（在目前的程度上）继续与公共利益保持一致（Jurowetzki 等，2021年）。由于其中一些关切正是本学位论文研究动机和研究问题的核心，

我们将在以下各节中讨论这些问题：我们将讨论如何利用机器学习危及在线隐私，如何运用以用户为中心的安全框架来应对这一问题，以及该框架又如何提升相关研究的可访问性。

0.1 毫升与数字隐私

算法[3]从根本上改变了我们对如何冒险以及如何在互联网上分享的看法。过去，人们通过在线购买商品等行为交出的身份识别信息（例如 IP 地址、浏览器指纹、Cookie 和个人详细信息）大多只有开发者能看到。如今，这些行为几乎会立即得到反馈，例如通过定向广告。这（以及其他因素）大幅加剧了公众对数字隐私的担忧。然而，矛盾的是，普通互联网用户很少采取措施保护自己的数据（Barth 和 de Jong，2017年）；需要敏感信息的服务具有实用性且呈集中化特征，导致用户低估信息披露的风险（Kehr 等人，2015年）。此外，用户无法掌控哪些数据会被分享，这让他们对自己的隐私感到愤世嫉俗、无可奈何（Lutz 和 Hoffmann，2017年；Lutz 等人，2020年）。

诚然，可以说，如果没有电话、银行卡或电子邮件账户，一个人恐怕会成为彻底的社会边缘人[4]；不过，对于我们（直接）分享什么、使用哪个标识符，以及在哪些空间分享，我们仍然有选择权。例如，信息安全提供了各种工具和框架，让人们能够保护个人网络空间中所剩无几的部分；端到端加密便是其中之一。

[3] 利用机器学习向用户提供预测性输出的系统，通常被称为——尽管从定义上说这一术语尚有争议——算法。[4] 尤其适用于中高收入国家。

在这一协议下与他人和服务进行的互动是私密的，或者至少是受控的。然而，公众似乎对聊天客户端中的隐私威胁模型（例如）有着良好的了解，却缺乏对加密的技术认知，也不认为加密是一种能够保护其数据的技术解决方案（Dechand 等人，2019年）。此外，加密主要保护直接通信和本地信息，而互联网上的社交活动在一定程度上涉及自愿向更广泛的受众披露特定信息。然而，理想情况下，我们应当能够这样做，同时不必放弃对自己选择不分享的内容所享有的隐私，也不必在分享时放弃匿名性。

“大数据”[5] 分析和深度学习在此类场景中对我们的隐私产生了不利影响。这些模型需要海量数据才能达到最先进水平，而这些数据往往涉及人——但并非所有人都得到同等（因而公平）的[6] 表征，且远不如加密所能提供的保护。例如，由于模型会存储信息的（某种抽象表示），因此容易受到提取攻击（Ateniese等人，2015年； Shokri等人，2017年；Song等人，2017年），这类攻击从成员推断（即某个实例是否属于原始训练数据）到生成敏感内容（Carlini等人，2021年；Pan 等人，2020 年）。这促使安全与隐私研究人员探究如何让机器学习模型妥善处理敏感信息（Edwards和Storkey，2016年； Abadi等人，2016年b），并采用与加密类似的框架来实现。不过，保护机器学习模型中编码的用户信息，只能涵盖潜在问题的一部分；更重要的是，机器学习可能被用于推断人们未曾主动分享的信息。

5 （对于特定目标或任务而言）规模适当，或许是更贴切的说法。6 虽然公平性不在本学位论文的讨论范围内，但它是当前社会和学术界关注的重要议题。关于这一主题的进一步阅读资料，例如：Bender 和 Friedman (2018年); Mitchell 等人 (2019年); Bender 等人(2021年)。

0.1.1 潜在用户信息

借助联网设备进行身体活动，是行为可能促成侵入式推断的明显例子；位置可能暴露个人偏好和社交关系（Anthony等人，2007年），而传感器则能收集大量身体与生理信息。触觉传感器甚至可能捕捉到用户在手机屏幕上的手指点击位置（Liang等人，2018年b）。然而，对于通过互联网分享的某些信息来源而言，这一点或许并不那么显而易见。随着非专家也越来越容易使用机器学习，任何公众可访问的信息都可能带来远超我们（即便作为“专家”）所能现实预测的深远影响。例如，机器学习可以通过一个人的搜索词推断人口统计信息（Murray和Durrell，1999年）、朋友圈（尤其是朋友也分享个人信息时：Mislove等人，2010年； Dong等人，2014年）、购买记录（Wang等人，2016年）和媒体偏好（Sun等人，2017年）。因此，仅仅使用互联网，我们就会留下关于自身的潜在信息痕迹。

让我们在本学位论文中定义“潜在信息”这一术语。这个概念对于数字隐私研究和机器学习而言似乎都较为小众。在后者中，它有多种不同的称谓；例如，He 等人(2018年) 将其称为“潜在数据”，而 Volkova 等人(2014年) 使用“潜在属性”这一术语。这些称谓都可以归结为同一种现象，因此我们将潜在信息定义为：并非数据预期用途的一部分，却可以作为其副产品推断出来的信息。举例来说：在线购买尿布可能（相当显而易见地）表明你是家长；而杂货中甜味蔬菜（如欧洲防风草）突然增多，也可能（虽不那么明显地）透露这一点。同样，研究表明，在社交媒体上只有少数关系密切的朋友，是老年人的良好预测指标（Dong等人，2014年）；而观看《卡萨布兰卡》则可以表明你是女性（Sun 等人，2017年）。[7]

[7] 遗憾的是，预测结果仍常常依赖于其中（有害的）刻板印象线索

0.1.2 文体特征分析与双重用途

语言是能够揭示潜在信息的典型数据，因此它成为本学位论文的主题并非偶然。自然语言处理（NLP）领域致力于使人类语言能够被机器理解，并提供了多种技术来推断字面文本之外的信息。除了身份之外，一个人独特的写作风格也可能包含敏感的社会人口特征，例如性别和年龄（Schler 等人，2006年；Bamman 等人，2014b年）、人格（Plank 和 Hovy，2015年）、教育程度和收入（Rao 等人，2010年；Volkova 等人，2014年）、原籍地区（Bamman 等人，2014a年；Tulkens 等人，2016年）、政治或宗教归属（Koppel 等人，2009；Pennacchiotti 和 Popescu，2011年），以及心理健康问题（Choudhury 等人，2013年；Coppersmith 等人，2015年）。通过对公开分享的文本进行文体计量分析，有可能推断出这些属性，而且往往准确度很高。对这类作者属性的预测，与作者身份推断相关（对于单个作者，这是一个二元任务；对于多个作者，则是多类别预测，例如参见 Stamatatos 等人，1999年），称为作者画像（另见第1章和第2章）。

尽管这些努力极大地推动了多个研究领域的发展，例如计算社会语言学（Daelemans，2013）、欺诈、欺骗和身份盗窃的检测（Badaskar等人，2008年；Ott等人，2011年；Banerjee等人，2014年；Fornaciari和Poesio，2014年），但它们也可能使互联网用户暴露于对手面前，让对手推断出这些属性——而用户可能在不知情的情况下遭到滥用。这对种族、政治立场、心理健康或任何其他被明确设为不可获取的个人信息方面处于弱势地位的人尤其有害。因此，这是功能蔓延的典型案例

训练数据 (Koolenand vanCranenburgh, 2017年)。因此，这些数据和模型在各种人口统计属性方面并非没有偏见 (Jo and Gebru,2020年)。

（特指信息技术领域：Schneier，2010年），或两用研究（例如健康与自然科学领域更广泛讨论的情形：Ehni，2008年；Williams-Jones等人，2013年）；也就是说，这类技术虽有意带来益处，却也会产生非预期的有害用途或后果。研究人员通常会反思这类伦理问题——不过在机器学习领域（Wallach，2014年），尤其是自然语言处理领域（Hovy和Spruit，2016年），直到最近（例如，自2021年 年ACL会议开始实施伦理审查流程）这类反思才逐渐出现。Williams-Jones等人 (2013年)建议，此类研究的监管应由所有相关方（例如科学界和治理机构）共同承担；我们还要补充一点，对技术的掌控权并不只属于开发技术的各方，这也大大增加了此类工作的复杂性。

长期以来，这些画像方法一直仅用于科学研究；收集高质量的数据标签成本高昂（既耗时，也耗费资源），通常需要招募并培训标注员。此外，收集数据本身以及微调模型，也需要一定的专业知识和基础设施。在Emmery 等人(2017年的研究中，我们表明，只需在社交媒体上询问用户自我报告的性别，就能获得足够的数据（通过远程监督获得，且带有噪声），使模型的表现足以媲美使用“干净”的已标注语料库训练的模型。[8] 最重要的是，这项工作仅用了24 小时，就在一台普通的4核笔记本电脑上运行完毕，使用开箱即用的模型；无需调优。后来，这项研究又进行了重复验证，[9]在来自不同时期的新数据样本上，以及针对多种属性，都得出了相似的结果，与Beller 等人(2014年)先前研究的结果一致。这意味着任何人都有可能建立此类画像工具，使监管变得更具挑战性。

8 https://github.com/cmry/simple-queries

9 https://cmry.github.io/toku

因此，为互联网用户提供能够缓解此类有害推断的工具，是保障其在线安全的一项重要举措。

这引出了当前研究所要解决的第一个问题：自然语言处理能够进行侵犯隐私的画像推断，而其运作机制并不符合人类的直觉。与自然语言处理的许多其他领域不同，当前深度学习技术的“黑箱”特性（Benítez 等人，1997年）并非根本问题所在。文体计量分类器可能会使用高层特征（例如标点符号使用、句子和词语长度、音节：Brinegar，1963年；Morton，1965年；Brainerd，1974年；Otterbacher，2010年）、特定词元（词语：Holmes 和 Forsyth，1995年）和符号（表情符号：Peersman 等人，2011年），或句法特征（Biber，1995年；Baayen 等人，1996年；Feng 等人，2012年；Hollingsworth，2012年）——所有这些都可能与某些个人属性相关联。互联网用户若无法深入了解这些模型及其预测结果，也无法访问这些模型，就没有合理的经验法则来保护自己免受此类推断的影响。因此，他们需要一种安全工具；具体来说，需要一种由机器学习驱动的工具。

0.2 自然语言处理中的以用户为中心的安全

自然语言处理中的安全通常是在信息安全或取证语境下进行界定的，重点在于系统和数据；更具体地说，是指提供服务的实体希望确保其（用户）数据安全（Atallah 等人，2000年；Raskin 等人，2002年）。这类研究的例子（包括但不限于）有：在大型语料库中查找敏感信息（Canfora 等人，2018年；Kleinberg 等人，2022）、检测恶意软件（Sen 和 Can，2021年），以及保护处理文本数据的模型中的表示（Feng 等人，2020年；Lyu 等人，2020年）。对自然语言处理中安全研究进行更广泛的分类，是近期才出现的（Feyisetan 等人，2020年）；迄今为止，这类分类尚未涵盖针对个人的伤害。因此，这形成了一个（颇为有趣的）灰色地带，在这里，企业利益与消费者利益仅

部分重叠。尽管在线毒性（Greevy 和 Smeaton，2004）、虚假信息（Vlachosand Riedel，2014年）、冒充（Shropshire，2018年）和监控可能对公众造成极大伤害，但平台通常没有直接的经济动机来保护个人用户免受此类外部威胁。

这引出了本论文探讨的第二个问题：自然语言处理领域的安全研究主要以行业为中心，而非以用户为中心。这意味着，尽管相关研究可能解决具有社会意义的问题，但由此产生的工具往往并未提供给公众使用；更重要的是，研究在方法上也并非从典型用户的视角出发。例如，研究可能采用连普通研究实验室都难以获取的计算方法，只针对某个特定领域、平台或场景开展实验，或者总体研究目标与公共利益并不一致（例如改进具有侵入性的技术）。在自然语言处理研究的其他领域，也曾探讨过采用这种以用户为中心的框架，例如对话（Rieser 和 Lemon，2008年；Nothdurft 等人，2015年）和问答系统（Kelly 等人，2006年）、摘要（Passali 等人，2021年）、可解释性（Schuff 等人，2020年）、临床领域（Chapman 等人，2011年）、自然语言处理评估工具（Madnani 和 Loukina，2020年），以及机器学习库的设计（Cai 和 Guo，2019年）。

需要注意的是，以用户为中心的安全与隐私之间存在紧密的相互依存关系；这要求以能够在本地计算机上运行的方式部署安全工具，从而实现数据处理和后续模型的去中心化（不同于分布式推理或联邦学习等相关安全框架）。自然语言处理仍依赖许多可能会泄露敏感作者信息的公共资源（概述请参见 Leidner 和 Plachouras，2017年）。因此，鉴于直接或

基于社区的通信平台（Tandocet al.，2018年，例如 WhatsApp、Telegram、Discord）日益普及，形成以用户为中心的研究实践将有益于整个领域。若用户的（间接）贡献回报甚微，甚至可能造成伤害，那么目前可公开访问、用于训练最先进模型的最丰富数据源（例如 Reddit）便无法保证长期开放。为推动此类以用户为中心的实践，本学位论文探讨了两种安全应用：网络欺凌检测和对抗攻击。

0.2.1 网络欺凌检测

网络欺凌检测——我们在此简要讨论这一主题，第3章将详细介绍这项任务及相关工作——是在线伤害预防的一个子类别（也与毒性检测有关）；其公共利益在于开发能够识别网络欺凌事件的软件。此类应用[10] 旨在利用这类分类结果进行监控；也就是说，向版主或可信赖的人发送报告（Van Hee 等人，2018年）。据我们所知，这项任务尚未被视为一种去中心化问题；也就是说，尚未用于在接收方一端屏蔽联系人或内容（这更接近垃圾信息过滤）。为此，理想情况下，相关研究应满足若干以用户为中心的标准：例如，用户可能会使用多个通信平台，若只针对单一平台进行定制，检测潜力就会大幅降低；这意味着现有分类器理想情况下应能泛化到多个领域。此外，对于单个用户而言，这些分类器无法利用社交媒体的完整网络结构；其数据范围仅限于对话或社交媒体个人资料。因此，收集和维护有用的训练数据是一项主要要求，而毒性语言

[10]例如网络空间应用自动监测（AMiCA）项目：https://amicaproject.be/

会迅速演变（见第4章）。因此，值得研究这种变化会在多大程度上影响当前分类器；如果确实会产生影响，较小的数据源能否通过增强来缓解这一问题。

0.2.2 对抗攻击

最后，我们探讨对抗攻击；也就是说，改变给定的输入（生成所谓的对抗样本），使给定的机器学习模型无法按预期运行。最先进的（神经）模型被证明极易受到此类受控扰动攻击的影响（Szegedy 等，2014年；Goodfellow 等，2015年；Kurakin 等，2017a,b；Papernot 等，2016b；Madry 等，2018年）。通常，只需进行细微变化就能奏效；尤其是在计算机视觉领域，这些变化并不会显著改变人们感知到的输出。例如，Tian 等人（2018年）对自动驾驶系统的输入中的光照条件稍作改变，并表明这些变化虽未被察觉，却会影响系统的行为。然而，这种易于发动隐蔽攻击的特点并不适用于语言等领域的离散输入；要让相关技术可用，更重要的是不被察觉，就必须保留语法正确性和预期语义。因此，生成文本对抗样本的方法（Li 等，2016年；Alvarez-Melis 和 Jaakkola，2017年）通常会在词级进行人工删除和替换，或采用基于启发式的删除和替换。例如，Papernot 等（2016c）会将词替换为其嵌入表示的最近邻。虽然这种方法较为保守，但也需要用户付出大量努力（例如人工校正：Jia 和 Liang，2017年）；因此，全自动（即端到端）对抗系统正日益成为热门研究方向（第1章和第2章将进一步介绍相关细节和相关工作）。

在本学位论文中，这类攻击是贯穿我们以用户为中心的安全问题的一条主线；它们与隐私

保护式混淆方法（Potthast 等人，2019年），用于对抗作者画像（即对抗性文体计量：Brennan 等人，2012年），而对抗样本也可能有助于评估词汇变异和数据增强对网络欺凌检测的影响。为了使这项研究符合所提出的以用户为中心的框架，它还必须满足若干标准：大多数对抗攻击假设攻击者能够完全访问目标模型，或至少能够有限访问用于拟合对抗样本的目标模型（即白盒攻击和黑盒攻击）；然而，使用这些攻击的用户对目标模型一无所知。此外，我们不能假设用户有足够的能力调整这些攻击，也不能假设用户拥有可长时间运行这些攻击的基础设施；因此，这些攻击必须开箱即用，并能在合理时间内运行。因此，这里需要探讨的方法论权衡，是定向攻击与端到端攻击之间的取舍；定向攻击提供了更多控制，更为保守，因此可能更忠实于原始输入。更重要的是，它们通常对用户的技术要求较低。不过，其有限的变化可能远不如对整句进行端到端改写有效。反过来，这些系统也可能无法满足我们的可用性标准。最后，以用户为中心的框架要求所有（可用的）研究都向公众开放。

0.3 可复现性与可访问性

鉴于该领域目前产出研究的速度（或许前所未有），近期人们开始关注取得这些结果的模型是否缺乏严谨性和鲁棒性，以及这会如何影响更广泛的机器学习社区（Hutson，2018年）。考虑到推动该领域发展和吸引人们兴趣的一个主要因素，是功能强大的开源库和预训练模型日益普及，降低了进入该领域的门槛，这或许令人意外。此外，

机器学习研究的许多特性都有利于开展开放、可复现的研究：大型标准化且公开可用的数据集、通常噪声较少的数据收集、由社区推动的共享任务和基准、分享完整研究代码的能力，以及倾向于采用开放获取出版方式（Munafò 等人，2017年）。

尽管（应用）机器学习研究具有这些固有优势，但人们仍普遍认为它难以复现（Gundersen 和 Kjensmo，2018年）；也就是说，给定不同的数据和另一种实现方式，无法得到相同的结果，也无法支持原始研究中的推断和结论。要完整复现奠基性论文的结果（而非仅仅重复使用已发表的模型），研究者遭遇了诸多问题，包括实验描述不清（Pineau 等人，2018年；Tian 等人，2019年）、无法推广到域外数据集（例如 Recht 等人，2019年；Zhang 等人，2020年a），以及无法复现的模型评估（Melis 等人，2018年）。人们很大程度上将复现问题加剧归咎于机器学习研究生态系统（Henderson 和 Brunskill，2018年；Hutson，2018年；Lipton 和 Steinhardt，2019年；Mitchell 等，2019年；Sculley 等人，2019年、2018年；Gebru 等人，2021年）。

根据定义，以用户为中心的研究应致力于[11]向社区开放；也就是说，研究应可复现，但更重要的是应易于获取：提供完善的文档，举例说明代码（或软件）如何用于预期目的，并在可能的情况下加以扩展。对于Emmery 等人(2017年，我们为这些具体要求提供了一个初步模板，并将在本文所述研究的全过程中遵循该模板。

[11]我们将这些表述为目标，因为实现完全可复现性是一项极其复杂的任务，其标准往往比可访问性多得多（Emmery 等人，2019年）。

0.4 研究问题

如前所述，本论文将自然语言处理领域与其安全和隐私框架并置，并将这些内容与以用户为中心的机器学习这一更广泛挑战的相关研究联系起来。因此，本论文聚焦于自然语言处理领域中两个备受公众关注的安全领域：作者画像和网络欺凌检测，并旨在回答以下研究问题：

研究问题1 端到端攻击和定向攻击在多大程度上可用于以用户为中心的对抗性文体计量?

研究问题2 网络欺凌分类器是否足够稳健，可以以用户为中心的方式部署?

在此过程中，我们提出以下独特贡献（截至各章发表时）：我们 (i) 探讨自然语言处理安全研究的以用户为中心框架，具体聚焦于对抗性文体计量和网络欺凌检测任务。为此，我们 (ii) 提出一种针对对抗性文体计量的风格中立框架（第 1章），(iii) 明确测量对抗性文体计量的可迁移性（第 2章），(iv) 评估网络欺凌分类器的泛化能力（第 3章），(v) 评估其对词汇变异的敏感性，以及 (vi) 探讨扩充其语料库的可能性（第4章）。



本章已发表于：

> Chris Emmery、Enrique Manjavacas Arevalo 和 Grzegorz Chrupała。2018年。通过不变性进行风格混淆。载于论文集：第*27*届国际计算语言学会议，第984–996页，圣菲，新墨西哥州，美国。计算语言学协会

标题、正文和图表均进行了少量修订。图1。1以及表1。6 & 1。7 均已得到显著改进。

1

通过不变性进行文本风格混淆

此前，人们已在基于迁移的混淆框架下，研究了使用序列模型对写作风格进行混淆的任务；在这种框架中，输入文本会被明确改写成另一种风格。这一框架的一个负面影响是，输入文本的语义内容经常会发生重大改变。本章中，我们提出了基于不变性的混淆，并研究经过明确训练以实现风格不变的模型在多大程度上能够保留语义。我们在并行和非并行设置下评估了我们的架构，并对混淆句进行了自动评估和人工评估。实验表明，风格分类器的性能可以降至随机水平，同时其输出质量经评估与采用风格迁移的模型相当。此外，人工评估表明，混淆程度与输出在意义保留和语法正确性方面的质量之间存在权衡。

1.1 引言

写作风格能够独特地表征一个人，并可用于自动作者识别（Holmes，1998；Stamatatos 等人，2000），这一事实已在（计算）文体计量学领域得到充分研究（Neal 等人，2018）。类似地，关于作者画像的研究（Koppel 等人，2002）表明，此类文体计量特征可用于准确推断大量个人信息，例如年龄、性别、教育程度、社会经济地位和心理健康问题（Eisenstein 等人，2011；Alowibdi 等人，2013；Volkova 等人，2014；Plank 和 Hovy，2015；Volkova 和 Bachrach，2016）。传统上，这些技术依赖于成本高昂的人工标注样例；然而，近期研究表明，即使仅将自我报告作为远程监督信号，也能达到几乎相同的准确率（Beller 等人，2014；Emmery 等人，2017；Yates 等人，2017）。尽管这些工作极大地促进了计算社会语言学等多个研究领域的发展（Daelemans，2013），但它们也可能使此类媒体的用户暴露于攻击之下，攻击者可能在用户不知情的情况下滥用这些信息。对于因种族、政治立场、心理健康或其他个人信息而处于弱势地位的人来说，这尤其有害，因为这些人明确不希望他人获取这些信息。

对抗性文体计量学，即风格混淆，是旨在保护用户免受此类攻击的一种方法。其目标是改写输入文本，使得在检测某个特定变量（例如人口统计属性）上训练的分类器（对手）受到蒙骗，从而有效保护该变量。主要挑战在于保留输入原文的含义，同时掩盖混淆行为（Potthast 等人，2016年）。近期关于自动混淆的研究（Shetty 等人，2018年；Karadzhov 等人，2017年）在降低对手的性能方面取得了令人鼓舞的结果；然而，这些模型（正如

作者所指出的）难以保留输入的语义内容。举例来说，将学校改写为婚礼[1] 虽然成功蒙骗年龄分类器，使其认为文本是由成年人而非青少年写的，但输出并未保留原文的含义。

我们认为，这种观察到的意义偏移在一定程度上是混淆任务设计所产生的副作用。与特定属性密切相关的实词，往往会对潜在对手的准确率产生重要影响。年龄和性别分类研究中有大量证据表明存在这一现象（Koppel 等人，2002年；Rao 等人，2010年；Burger 等人，2011年；Sap 等人，2014年；Pardo 等人，2016年等）。具体来看 Sap 等人（2014年）提供的例子，性别分类中系数权重较高的特征包括：男性相关的 boxers（拳击手）、shaved（剃光的）、girlfriend（女朋友）、beard（胡须）、fightin（打架），以及女性相关的 purse（手提包）、blueberry（蓝莓）、pedicure（修脚）、hubby（老公）、earrings（耳环）。因此，明确以降低此类分类器性能为目标的模型会执行我们所说的基于迁移的混淆，这并不令人意外。举例来说，即使句子是男性写的，只要看起来明显偏女性，对手就很容易受骗。因此，从这一角度看，最有效的混淆方式是一种风格迁移：将少数几个与目标属性强相关的实词替换为与之对照的变体（如将 wife 换成 husband，将 school 换成 wedding）。如果这些变体在序列模型所利用的语义空间中也彼此接近，那么任何重构指标——例如 BLEU（Papineni 等人，2002年）、METEOR（Banerjee 和 Lavie，2005年）、嵌入距离等——都可能无法准确反映这些变化所造成的真实意义偏移。

我们的贡献　在本章中，我们提出了一种不同的自动混淆方法，并假设它部分克服了——

[1] 摘自Shetty 等人(2018年)。

保留输入文本含义方面的局限：基于不变性的混淆。在这种方法中，目标转向最大限度地提高对手的不确定性，这意味着其对受保护变量的准确率应尽可能接近随机水平。要使对手的表现维持在随机水平附近，就需要去除输入文本中与任何受保护变量高度相关的文体特征，从而生成在这些文体差异方面保持中性的语言。我们在若干实验条件下检验了这一假设。[2] 在我们的编码器—解码器架构中，为实现输入的风格不变编码，核心组件是插入输入句子嵌入与风格分类器之间的梯度反转层（GRL，Ganin 和 Lempitsky，2015年）。我们通过使用条件解码器，分别考察该模块单独使用时的效果，以及在风格不变软迁移设置中的效果。首先，为了评估这一架构能否成功解码风格不变编码，以及这会对手的表现产生什么影响，我们在英语《圣经》风格的平行语料库上训练序列到序列模型。其次，考虑到现实的混淆设置中无法获取此类平行语料，我们移除目标文本对，构建自编码器设置。在实验中，我们展示了随机水平附近的一种权衡：在平行语料设置下，使用多对多翻译模型的基于迁移的混淆效果良好，但其人工评估分数低于我们的风格不变模型。因此，我们认为风格不变的混淆方法具有潜力，值得进一步研究。

[2] 用于完整复现实验的所有代码和数据均可在 github.com/cmry/style-obfuscation获取。

1.2 相关工作

1.2.1 对抗性文体计量

Rao 和 Rohatgi（2000年）首次探讨了利用计算文体学破坏匿名性的设想。他们认为，可以借助机器翻译（MT）隐藏文体信息，但指出当时的技术能力还不够强。Kacmarcik 和 Gamon（2006年）沿着这一研究思路继续探索，向用户介绍其写作风格中的特征和更深层的语言线索。近期相关研究见（Thi 等，2015；Caliskan 等，2018）。Brennan 等人（2012年）明确将混淆视为一项对抗任务，并采用机器翻译（往返翻译），这与（Caliskan 和 Greenstadt，2012年）的做法类似。基于规则的扰动（Juola 和 Vescovi，2011年）以及两者的混合方法（Karadzhov 等，2017年）也已用于全自动混淆。与我们的方法最接近的是 Shetty 等人（2018年）近期开展的工作：他们采用生成对抗网络架构，研究基于迁移的混淆。其设置中，生成器经过训练，能够生成使判别器所赋概率最大化的句子；而判别器则经过训练，用于区分真实句子与生成句子。尽管他们采用了不同的语义损失，并展示了在带有年龄和性别标注的微博数据及政治演讲上成功实现混淆的结果，但其输出仍存在我们此前所述的语义保持不足问题。最后，Carlson 等人（2017年）提出了一种具有混淆应用潜力的风格迁移方法，他们借鉴零样本翻译（Johnson 等人，2017年），利用序列到序列机器翻译模型研究文本零样本风格迁移。他们的工作展示了不同版本英文《圣经》之间成功实现翻译。

1.2.2 梯度反转

Ganin 和 Lempitsky 提出了使用梯度反转层学习领域不变特征表征的方法，(2015)，并展示了该方法在计算机视觉数据光照条件上的应用。此后，该方法已应用于多种语言任务，例如：文本特征提取（Pryzant 等人，2017）、词性标注（Kim 等人，2017；Gui 等，2017）、图像描述（Chen 等人，2017）和文档分类（Liu等人，2017；Xu 和 Yang，2017）。最重要的是，Xie 等人 (2017证明，梯度反转层可用于实现对抗设置，并提升包括生成在内的多种语言任务的性能。这些结果表明，该方法有望应用于基于不变性的混淆。

1.3 模型

我们的基础架构是一种神经编码器—解码器（Sutskever 等，2014年）模型，与 Wu 等人（2016年）的模型相似，使用 PyTorch（Paszke 等人，2019年）实现。给定一个由独热编码词构成的输入序列，编码器首先将这些词嵌入为稠密向量，然后通过一个或多个双向（Schuster 和 Paliwal，1997）层进行处理，这些层采用长短期记忆单元（Hochreiter 和 Schmidhuber，1997）。随后，使用下文将介绍的内部注意力机制，将处理后的向量序列合并为单一表示。接着训练一个神经语言模型，根据编码后的句子嵌入（即所谓的上下文向量）解码输出序列。训练通过最小化目标序列按词计算的局部归一化交叉熵来完成。

在自编码器设置中，网络的目标只是根据编码后的上下文向量重构原始句子（Chandar 等，2014年；Li 等，2015年）。这种设置可以结合

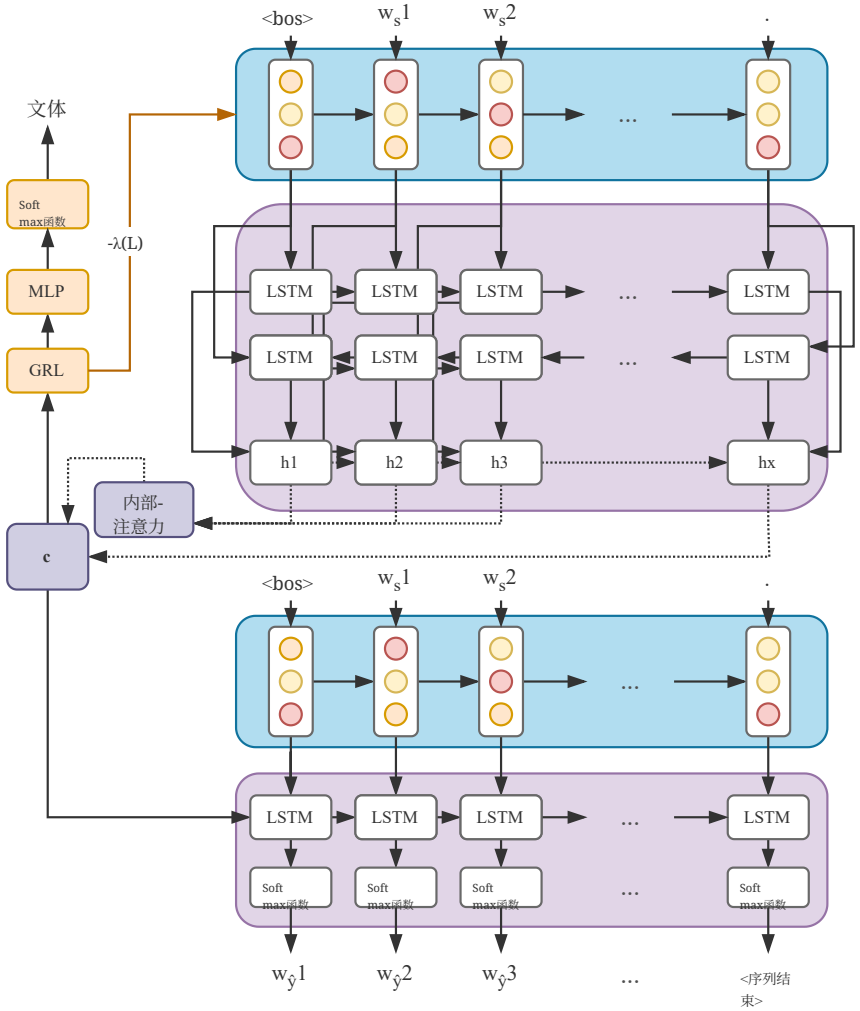


图1.1： 带有梯度反转层（GRL）组件的模型架构：顶部是编码器部分，左侧是GRL，它将输入句子嵌入（或上下文向量）**c** 作为输入，底部则是该架构的解码器部分。请注意，由于 GRL 会反转损失（$-\lambda(L)$），编码器既要降低多层感知机的文体分类性能，又必须生成有用的表示**c**供解码器使用。

通过梯度反转层鼓励编码器生成属性不变的上下文向量。对于自编码器（AE）架构，目标文本是输入本身；对于序列到序列（S2S）架构，目标文本则是配对句子。图 1.1 以可视化方式展示了基础架构。

1.3.1 架构组件

梯度反转层 (GRL) 梯度反转层（Ganin 和 Lempitsky，2015年）应用于上下文向量之上。在前向传播过程中，梯度反转层计算恒等函数，并将其输入传递给浅层多层感知机（MLP）风格分类器。然而，在反向传播过程中，风格分类器的梯度会反转符号。在本研究中，它的作用是优化编码器参数，使其具有风格不变性；也就是说，尽可能少地编码输入文本中的风格信息。

条件解码器 (C) 以往关于神经语言建模的研究表明，根据句子层面和上下文变量（如时态、情态或语态）对模型进行条件化是有效的。在我们的实验中，我们评估了一种条件自编码器，其中解码器以输入风格标签为条件。我们按照 Ficler 和 Goldberg(2017年) 的方法实现条件化；也就是说，将对应的属性嵌入向量（在我们的例子中，即对应的风格嵌入）拼接到输入解码器的每个词嵌入上。因此，每种风格都关联一个嵌入向量 **c**，并在每一步输入架构中。与简单自编码器不同，条件自编码器可以在测试时切换到所需的风格。我们预期，通过以原始风格标签为条件对解码器进行条件化，输出能够保留其语言特征和一致性，而不会完全恢复目标风格。

令牌转移 (文本变换) 我们认为，依赖风格平行语料库（无论语料按属性还是作者划分）的机器翻译系统采用的是基于迁移的混淆；而且，由于翻译在很大程度上是保留意义的操作（Ide 等人，2002年；Dyvik，2004），其语义一致性可能较高。然而，这类平行语料库通常并不存在，且相关整理成本极高；这需要大量相同（平行）的信息（理想情况下达到句子层面），例如分别由青少年和成年人撰写的文本。因此，使用机器翻译进行文本风格迁移并不是一种可行的混淆应用场景。不过，这确实能很好地说明混淆模型在基于迁移的混淆框架下的性能。为此，我们采用了 Carlson 等人（2017年）讨论过的、在风格数据上训练的序列到序列翻译模型。按照 Johnson 等人（2017年）的方法，我们使用目标风格词元，使在多对多翻译上训练的模型只需在输入句子前添加一个目标词元，就能将其改写为另一种风格。这是唯一使用注意力机制的配置，因为该机制并未与梯度反转层结合进行测试。

1.3.2 架构详情

我们的模型使用 300 维嵌入，编码器和解码器均为单层，包含 1000 个长短期记忆单元。编码器输出的隐藏状态序列通过逐特征内部注意力机制合并为一个表示。令 w_t 和 h_t 分别表示步骤 t 时的输入词嵌入和长短期记忆网络隐藏状态，总步数为 n。句子嵌入 c 的 i^{th} 个特征通过对隐藏状态的 i^{th} 个特征加权求和计算得出：

$$c_i = \sum_{t=1}^{n} a_i^t h_i^t \tag{1.1}$$

其中，每个权重 a_i^t 的计算方式如下：

$$a_i^t = \frac{\exp([W^T z^t]_i)}{\sum_{s=1}^{n} \exp([W^T z^s]_i)} \tag{1.2}$$

在公式 1。2中， z_t 是 w_t 与 h_t的拼接， $W \in \mathbb{R}^{(D+H)\times H}$是一个需要学习的额外投影矩阵，而 D 和 H 分别表示词嵌入矩阵和 LSTM 层的维度。与传统的合并模型（如最大池化或平均池化）相比，额外的参数有助于模型学习哪些输入词和特征对任务更重要。这类似于对隐藏激活向量采用传统注意力机制，主要区别在于权重是按特征进行的，并且从编码器到解码器的所有信息流都要经过单一的输出编码向量瓶颈。请注意，传统的对齐式注意力机制与 GRL 的应用不兼容，因为 GRL 的作用范围必须涵盖从编码器传递到解码器的所有信息。

解码器在每一步都会将上下文向量拼接到其输入上，从而以该向量为条件。这有助于提高传递给编码器的梯度信号，促进模型学习。所有模型参数均使用 Adam（Kingma 和 Ba，2015年）通过迷你批次进行优化，每批包含50 个样本，学习率为0.001 ，并在每轮训练后乘以0.75 。为避免过拟合，我们将 Dropout（Srivastava 等人，2014年）应用于每个长短期记忆网络层的输入，丢弃概率为0.25；当损失连续3 轮不再下降时，便停止训练。测试时，我们通过使用束宽为5的束搜索，近似求得全局最优输出序列。梯度反转层与单层多层感知机结合实现，其维度与编码器相同。

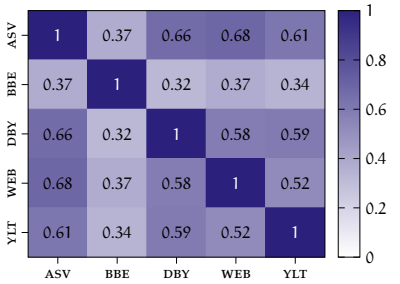


图1.2： 不同《圣经》文体词汇之间的杰卡德指数。

文体	类型	TOK	UNIQ	PUNC
ASV	12.5K	3.7M	602	0.13
BBE	6.0K	3.9M	569	0.11
DBY	13.8K	3.7M	1729	0.14
WEB	12.8K	3.6M	1663	0.14
YLT	12.2K	3.9M	1682	0.17

表 1.1： 语料库描述统计，展示每种圣经文体的类型和词元频数（词元）、唯一项类型的数量以及标点（标点）比例。

1.4 实验设置

我们的目标是研究基于不变性的混淆方法是否有效，具体而言，是探究风格不变表征在多大程度上保留了输入的句子语义。为此，我们使用上述组件，针对不同语料库和混淆设置开展了三项实验。在每种设置中，都会训练一个对手来检测待混淆的变量。

1.4.1 数据

我们使用了一个由五种不同版本的英文《圣经》组成的平行语料库（从 GitHub 获取）。[3] 这些版本在语义上保持一致，但在不同方面的文体各异：BBE 使用简化英语写成，YLT 遵循原文希腊语和希伯来语的句法结构，其他版本（DBY、ASV）则对应较早的版本，反映了历时变异。句子级经文编码

[3] https://github.com/scrollmapper/bible_databases，其原始收集来源为：http://www.openbible.info/labs/cross-references/。

(书 + 章 +诗节编号）用于在不同版本之间实现近乎完美的配对（移除了部分缺失的配对），从而构成风格五元组；这些五元组使用 spaCy[4] (Honnibal 等人，2020年）进行词元化。风格元组被划分为训练集（80%）、开发集（10%）和测试集（10%），以确保开发集和测试划分中的所有句子均未见过，无论它们属于何种风格组合；并使用所有风格组合（不包括同一风格的组合）构成620,752 对。不同文体的更多语料库描述统计见图1.2 和表1.1。

1.4.2 对手

对手是一个句子级 fastText[5] (Joulin 等人，2017年）分类器；这是一个简单的线性模型，包含一个隐藏嵌入层，使用词袋或 n元组输入来学习句子表示，并使用 Softmax 函数生成给定文体的概率分布。该分类器在训练集划分的源端数据上进行训练，因为这些实例正是我们打算进行混淆的对象。模型训练20 轮，使用100 维嵌入、单元语法和二元语法，学习率为0。01，桶大小为1M。其准确率达到86。6%，而对于包含20% 个班级且分布均匀的情况，对手的随机水平性能为5 个班级。

1.4.3 评估

为了自动评估生成句子的重构效果和语义保留情况，我们使用机器翻译指标，以及语义嵌入空间中的距离。混淆程度则通过对手的准确率与随机猜测准确率之间的差异来衡量。

4 https://spacy.io/

5 https://github.com/facebookresearch/fastText

机器翻译指标(BLEU、METEOR) 我们使用BLEU-4 和 METEOR (Pap- ineni 等人，2002年；Banerjee 和 Lavie，2005年) 进行计算 `nlg-eval`。[6] 鉴于这并非标准的机器翻译任务，我们分别提供生成句子与源句子（←）之间，以及生成句子与目标文本（→）之间的分数。对于序列到序列模型，→ 是成功实现通过迁移进行混淆的主要指标。然而，← 则能在一定程度上反映输出与原文的关联程度。梯度反转层应使← 和 →的分数均降低。对于自编码器，只能在←上进行评估；加入梯度反转层后，其分数应大幅降低。

词移的距离(词移距离) 为了衡量混淆句与原文之间的嵌入距离，我们采用词移距离（词移距离）（Kusner 等人，2015年），该方法基于英语 `fastText`维基百科嵌入（Bojanowski 等人，2016年）。词移距离将两个句子在由嵌入词构成的加权点云中的距离，定义为该句子的点云为与句子 A 的点云对齐所需移动的最短累计距离 B——研究表明，这种距离能够捕捉某些语义关系。

对手影响（Δ 准确率）我们比较对手在生成句子上的准确率与其在原文上的准确率，以评估混淆强度。不过，由于我们的目标是使对手的表现接近随机猜测水平，因此我们将 Δ 准确率定义为 = 准确率 − p，其中 p 为多数类基线。因此，如果 Δ准确率为负数，这意味着对手的表现已低于随机猜测水平，任务更接近基于迁移的混淆，而非基于不变性的混淆。随后，正分数表示混淆在这两种情况下未能达到目标的程度。

[6] https://github.com/Maluuba/nlg-eval

高斯噪声（σ）为确保解码输出发生显著变化，我们在生成时向上下文向量添加高斯噪声掩码。该掩码是从高斯分布 $N(0,\sigma)$中采样的随机向量，其中 $\sigma = \{0.01、0.05、0.10、0.15、0.20\}$，并将其加到上下文向量的值上。这种噪声可以提高混淆程度（因为解码行为更加随机），但会牺牲输出质量。

1.4.4 实验

借助以上讨论的所有组件，我们定义了三种实验环境，用以衡量应用梯度反转层（GRL）和条件解码器来实现基于不变性的混淆所产生的效果。**(1)** 我们在英语《圣经》语料库的文体配对上，以多对多序列到序列的设置训练架构。在此引入梯度反转层后，编码器不会捕获那些能高度指示目标风格的词语。为实现有效的多对多机器翻译系统（进而实现风格迁移）设置，我们在 `<2{stylename}>` 词元前添加前缀。**(2)** 我们在非配对样本上训练自编码器。在此我们同时引入梯度反转层和条件解码器。我们假设，条件解码器能够从中性编码实现柔性的风格迁移，这意味着相比机器翻译模型，它能更好地保留语义结构。**(3)** 我们使用高斯噪声，使序列到序列模型和自编码器模型的混淆性能达到相同水平，并评估它们在随机水平性能下各自的质量损失。

1.5 结果

1.5.1 实验结果

所有结果和自动评估均显示在表 1。2 中。正如我们的假设所示，总体而言，使用风格迁移进行混淆是有效的，

模型	C	GRL	TT	PPL	BLEU ←	MTR ←	BLEU →	MTR →	WMD	ACC Δ
S2S				9.08	22.0	25.3	17.8	23.6	1.50	1.6
S2S		x		9.27	21.8	25.2	16.9	22.9	1.50	6.9
S2S			x	7.38	34.9	30.5	39.2	29.9	1.24	−12.0
AE				1.51	79.5	52.9	–	–	0.25	64.4
AE		x		1.99	60.0	41.2	–	–	0.65	50.8
AE	x	x		1.87	51.9	38.1	–	–	0.79	18.3
来源					100.0	100.0	36.0	34.5	0.00	66.6

表 1。2： 《圣经》实验的结果。第一列（模型）表示基础架构的设置：序列到序列（s2s）或自编码器（ae）。第二列说明所采用的模块：条件解码器记为 c，梯度反转层记为 grl，`tt` 表示前置的风格标记。结果展示了开发集上的困惑度（ppl）、源文本（←）/目标文本（→）与混淆句之间的 BLEU-4 和 METEOR、源文本与混淆句之间的词移距离得分（wmd），以及混淆句在多大程度上将对手推向随机水平性能（Δ acc）。在最后一行，我们列出了一个“源文本”基线，即通过复制源文本得到的结果。这体现了源文本与目标文本之间的总体重叠程度，以及对手高于随机水平的性能。

表现处于或低于随机水平。观察目标文本的 BLEU 和 METEOR 可以发现，不包含目标标记的序列到序列模型生成的句子与原文的相似度高于与目标文本的相似度；总体分数较低，根据词移距离判断，这些句子与目标文本相去甚远。不过需要注意的是，这是在没有任何目标信息的情况下进行的多对多翻译，且所用语言的词汇表大体相同。因此，文本变换更能现实地反映风格迁移的成功程度。就翻译质量而言，它相比原始基线几乎没有提升，但确实成功实现了基于迁移的混淆，这一点可由低于随机水平 12% 的准确率体现出来。评估梯度反转层在这一设置下的表现时，

	AE			自编码器 + 梯度反转层			自编码器 +C +梯度反转层		
σ	BLEU ←	MTR ←	ACC Δ	BLEU ←	MTR ←	ACC Δ	BLEU ←	MTR ←	ACC Δ
–	79.5	52.9	64.4	60.0	41.2	50.8	51.9	38.1	18.3
0.01	78.7	52.2	64.1	59.7	41.4	50.5	52.0	38.5	18.5
0.05	52.9	38.9	57.2	54.8	39.1	47.6	49.1	37.2	16.5
0.10	14.7	21.6	36.1	32.8	32.7	36.9	40.7	33.6	11.6
0.15	4.2	14.7	23.3	25.3	26.2	24.3	30.0	29.1	4.6
0.20	1.5	11.3	16.4	15.3	21.5	15.5	21.1	24.9	0.0

表1.3： 向自编码器中添加高斯噪声（σ）的效果。

它似乎既没有改善翻译，也没有改善混淆，这基本符合我们的预期。

自编码器设置能更清晰地呈现梯度反转层的性能，尤其是基于不变性的混淆方面。普通自编码器在一定程度上成功复现了目标文本；从 BLEU、对手的表现和词移距离来看，输出仍与输入密切相关。引入梯度反转层确实会影响输出与源句子的关联，但在混淆性能方面相比之下提升有限。然而，在架构中加入条件解码器后，混淆性能受到的影响显然大于 BLEU 和 METEOR 的降幅。最后，在解码器中加入高斯噪声掩码的效果见表1.3。在约 0.15 时，BLEU、METEOR 和 Δ 准确率方面的度量指标似乎相当。我们进一步通过人工评估考察了三个最合适的模型（序列到序列 + 文本变换、自编码器 + 梯度反转层 + C，以及自编码器 + 梯度反转层 + C+ σ = 0.15）。

1.5.2 人工评估

在人工评估中，20 对（原文、混淆文本）从三个目标模型的输出中抽样（总计 60 对）。每对样本由四名参与者（均具有语言学背景）使用3分制，从三个维度进行评分。这些维度包括原文与混淆句之间的语义一致性、后者的句法连贯性，以及输出中的变化程度。评分者获知每对句子中哪一句是原文，并被明确要求以原文为参照对各维度进行评分。评分者不知道存在多个模型；样本对经过打乱，因此无法比较具有相同原文的样本对。为简化与原文的比较，我们只从 BBE（基础英语圣经）中抽样。参与者收到的说明示例见表1。4，结果见表1。5。

根据评估结果，自编码器 + 梯度反转层 + C 在所有三个维度上都更受评分者青睐。具体而言，鉴于我们关注语义保持，该模型在语义 维度上的表现优于一种能够在一定程度上利用源文本与目标文本之间语义关系的文体迁移模型。根据变化维度的自编码器 + 梯度反转层 + C 最为保守，这与表BLEU 和METEOR 中的分数一致，且这种保守性可能也体现在1.2的语法正确性 维度上。

1.5.3 定性分析

在此，我们比较上一节中由评分者评估的三个模型的行为和生成的文本。据此，我们将指出实验方法的优缺点，并提出未来工作的方向。

	语义	语法正确性	变化
1	语义被破坏；句子不表示相同。	部分内容，或完整句子变得乱码。	特殊字符发生变化，或字符被翻转单个词。
2	语义略有变化，但并不突兀。	词序略有变化，且这种变化不符合语法。	多个词语被替换了，但与原文一致。
3	语义保持完整，这些变化不会改变平均值。	语法正确性未受到影响。	新增部分已引入／改写到句子中。

表 1。4： 人类评估研究中向评分者提供的解释，按三个维度的每项评分展示。

	语义	语法正确性	变化
序列到序列 +文本变换	1.88	2.21	2.43
自编码器 + 梯度反转层 + C	2.12	2.32	1.99
自编码器 + 梯度反转层 + C + σ = 0.15	1.35	1.66	2.65

表 1.5： 三个维度中各模型的平均人工评估分数。分数越高越好，但变化最好保持在约 2。

我们发现，基于迁移的混淆存在一个问题：对输入进行的细微、局部变化主要集中在与对手最相关的词语上。观察表 1。6这类较长的句子时，可以找到一些（部分）正确的变体，例如城镇 → 城市、荒地 → 旱地，以及表 1。7中的绿柱石→ onyx、印章 → seal，但仍有一些错误的变体：rest → work。总体而言，序列越长，观察到的变化就越多。

更值得注意的是，模型会对句子的某些部分进行扩展，例如 *thereis withouta waste land;andshemakesandry land*——尽管错误地插入了“*without*”，其余部分仍可视为正确的扩展。表1.7中的情况也是如此，

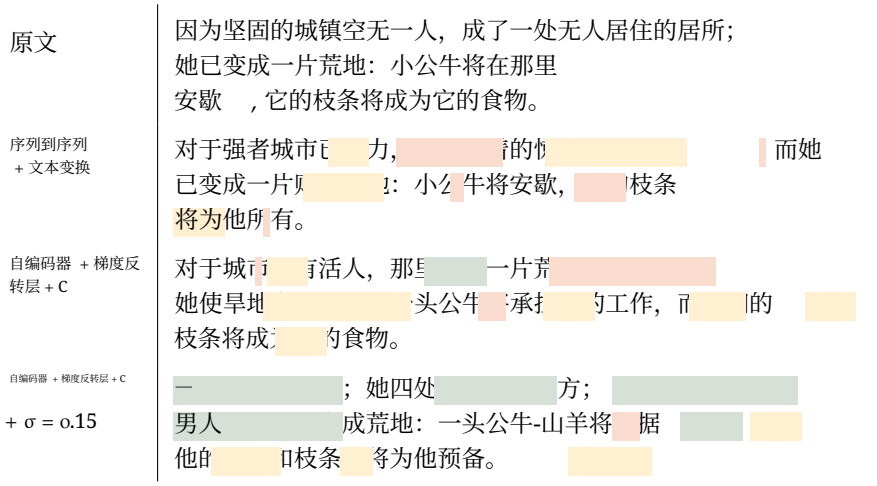


表 1.6： 示例 1 ——以赛亚书27：10。
色高亮表示删除或折叠，黄色替换项，以及
入项和无关替换项

其中“固定在扭曲的金框中”扩展为“于是他们被束缚在精巧编织的红色金框中”，部分有误地，类似于活人表1。6。相比之下，自编码器 + 梯度反转层似乎也倾向于使用较为压缩的短语，删去形容词，例如小（小公牛中的坚固的（坚固的城中的绿柱石（绿柱石，且完全没有纳入“无人居住的活人”。

直接比较序列到序列和自编码器 + GRL 的示例，可以推断出迁移方法似乎（至少在这些示例中）相当保守，基本保持完全对齐，只做细微变化。然而，这也会导致自编码器用无关的变体替换词语，或插入并非直接相关的词语；不过，在序列到序列模型中也能观察到同样的行为。

自编码器的输出显示，句子可能经过了一些轻微的改写，这或许是一个值得进一步探索的有趣方向。在

原文	然后，他们制作了绿柱石，并将其镶嵌在扭曲的金框中 并按照印章的刻法雕刻，上面刻有孩子们的名字 以色列的。’
序列到序列 +文本变换	然后他们制作了缟玛 共一百块 刻 ，并刻上印章的印记，帮 以色列子孙的名字。
自编码器 + 梯度反转层 + C	然后他们写 石头， 精巧 和按照雕刻印章的方 成 上面刻有以色列子孙的名字。’
自编码器 + 梯度反转层 + C + σ = 0.15	随后，他们 于希 底 成，并像印 的宝石一 出来， 在 子民的胸牌上。’

表格 1.7： 示例 2 —《出埃及记》39:6

条件解码器中纳入句子长度等不同变量（Ficler 和 Goldberg（2017 年）也展示了这一做法），将使这一方向的实验成为可能。图1.3表明，改写并非仅限于自编码器。我们先前在第1.5.2 节指出，自编码器 + 梯度反转层 + C 架构在改变输入时似乎更为保守；这一点也可从S2 + 文本变换 模型通常引入数量相等或更多的新词元，以及其异常值多得多（有些甚至超过原文中的词元数量）看出。这张图还显示，BBE 和 YLT 倾向于增加更多词元，这与表1.1所示的这两种方法较大的词元数量一致。极端异常值很大程度上是由循环机制产生的伪影所致（以及撒督的儿子比拿雅、希伯和以利法列）。

最后必须指出，在有关《圣经》的生成输出中，很难找到文体中性改写的证据；这不仅是因为可能存在古旧句式（我们尝试通过使用 BBE 将其尽量减少），更重要的是，识别文体变化需要一定的专业知识。不过，最重要的是，

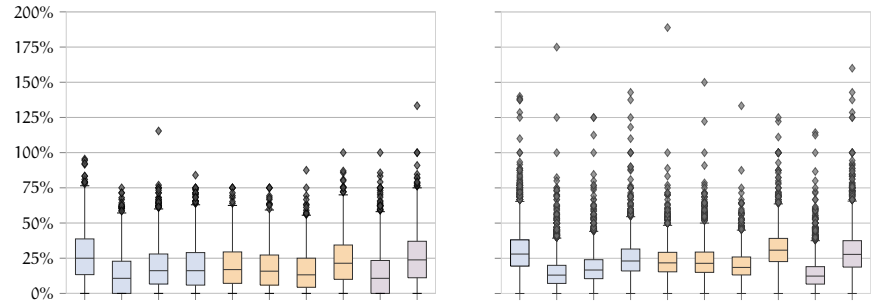


图 1.3： 对于 ae + 梯度反转层 + C 基于迁移的不变性混淆架构（左）和序列到序列 + 文本变换基于迁移的混淆模型（右），展示了新引入词元数量占原词元总量的比例分布。图中展示了五种源风格中的三种。

鉴于目前的结果，将自编码器应用于非平行数据具有一定的可行性，因此未来工作中或可将其拓展到其他领域。

1.6 结论

我们提出了一种对自动风格混淆任务的替代性阐释——基于不变性的混淆，并在神经编码器—解码器架构中测试了若干组件；我们假设这些组件能够实现输入文本的风格不变重写。我们测试了梯度反转层和条件解码器在并行与非并行设置下对混淆的作用。尽管在《圣经》语料库中难以找到文本具有风格中性的有力证据，但我们通过人工评估表明，在非平行数据上训练的自编码器架构，其评估结果优于这样一种模型：该模型在平行数据上训练，并且只能部分利用源文本与目标文本之间的语义关系。在我们的定性分析中，我们

发现了输入文本在语义正确性方面发生局部更改的证据，也发现了符合诗句语境的部分改写。这些结果表明，将这一架构扩展到其他非平行语料库，以在实际应用场景中测试混淆（例如针对年龄和性别等作者属性），前景可期。



本章已发表于：

> Chris Emmery、Ákos Kádár 和 Grzegorz Chrupała。2021年a. 真实场景中的对抗性文体计量学：针对作者画像的可迁移词汇替换攻击。见《*16*届欧洲计算语言学协会计算语言学分会会议论文集：主会场》，第2388–2402页，线上。计算语言学协会

对文本和图表进行了细微的格式调整（图2.1 除外，该图已得到显著改进）。

2

真实场景中的对抗性文体分析：针对作者画像的可迁移词汇替换攻击

WRITTEN 语言包含文体线索，可利用这些线索自动推断

出各种可能敏感的作者信息。对抗性文体计量旨在通过改写作者的文本来攻击此类模型。本研究提出了若干组件，以促进这些对抗攻击在真实环境中的部署，在那里既无法获取数据，也无法访问目标模型。我们提出了词汇替换攻击的一个基于 Transformer 的扩展，并证明在弱标注语料库上训练时，该方法具有很强的可迁移性，能将目标模型性能降至低于随机猜测的水平。尽管并非完全不易察觉，我们效果更好的攻击也证明更难被人类发现。因此，我们的框架为未来的隐私保护对抗攻击提供了一个颇有前景的方向。

2.1 引言

机器学习在消费设备上的广泛应用及其对设备数据的处理，引发了安全与隐私研究人员对如何妥善处理敏感信息的共同关注（Edwards 和 Storkey，2016；Abadi 等，2016b）。自然语言处理（NLP）也不例外（Fernandes 等，2019；Li 等，2018）；书面文本可能包含大量作者信息——无论这些信息是有意分享的，还是能够通过文体计量分析推断出来的（Rao 和 Rohatgi，2000年；Adams，2006年）。这一特性是作者画像的基础（Koppel 等人，2002年）；尽管该领域主要关注支撑语言使用的社会语言学和文体计量特征的研究（Daelemans，2013年），但与此同时，这也带来了双重用途问题。作者画像往往能够以很高的准确率推断出大量（敏感的）个人信息，例如年龄、性别、教育程度、社会经济地位和心理健康问题（Eisenstein 等，2011年；Alowibdi 等，2013年；Volkova 等，2014年；Plank 和 Hovy，2015年；Volkova 和 Bachrach，2016年）。因此，任何分享书面在线内容的人，都可能因其写作风格而面临个人信息遭到未经授权收集的风险。这对处于弱势地位的个人尤其有害，例如因种族、政治立场或心理健康状况而处于弱势的人。

针对这类推断的隐私保护防御措施，可见于对抗[1] 文体计量学领域。我们的研究[2] 关注混淆子任务，其目标是改写输入文本，使其文体发生变化，从而令文体计量学预测失效。这属于日益增多的自然语言处理对抗攻击研究的一部分（Smith，

[1] 这些是针对进行文体计量预测的模型的对抗攻击，不应与对抗学习混淆。[2] 所有代码、数据和材料均已公开，可用于完整复现实验，详见 https://github.com/cmry/reap。

2012年），多种现代模型已被证明容易受到攻击；例如，神经机器翻译（Ebrahimi 等人，2018年a）、摘要（Cheng 等人，2020年b）和文本分类（Liang 等人，2018年a）。

自然语言处理领域的对抗攻击主要旨在揭示现有算法或模型中的漏洞，以便修复这些漏洞，或通过对抗训练明确加以改进。因此，大多数相关工作都关注白盒或黑盒设置，在这些设置中，攻击者可以访问目标模型的全部或部分信息（例如其预测结果、数据、参数、梯度或概率分布），以据此构造攻击。然而，当前研究的目的并非改进定向模型，而是将攻击作为保护在线隐私的工具。这使得本研究相较于其他基于自然语言处理的对抗攻击受到若干限制，因为它要求应用场景真实且贴近实际。

首先，寻求保护自己免受文体计量分析影响的作者，不能被假定为了解目标架构，也不能被假定能够获取合适的训练数据（因为目标模型可能基于任何领域的数据进行训练）。因此，我们无法针对目标模型优化攻击，需要一种易于使用的方法来模拟目标模型，以评估混淆成功率。为此，我们采用所谓的替代模型。就本研究而言，替代模型是一个独立训练的作者画像分类器（使用其自身的数据和架构），可为我们的攻击提供指导。研究表明，在多种机器学习任务中，针对替代模型构造的攻击在攻击架构不同或使用其他数据训练的模型时，也能取得相应效果（Papernot 等人，2016年a）。针对“真实”模型时，基于替代模型构造的攻击的有效性称为可迁移性，我们将测量本研究所提出混淆方法的可迁移性。

其次，要让混淆攻击在实践中奏效（例如，考虑到

发文记录有限），攻击应针对作者在其所选领域内的写作提出相关修改建议。这意味着替代模型应在本地拟合，因此需要满足两个标准：可靠地获取带标签的数据，以及训练速度相对较快、过程较为简便。为满足第一个标准，本研究聚焦于性别预测，原因如下：i）标注了这一变量的 Twitter 语料库规模迄今最大（也最常见）；ii）作者画像方法通常会针对不同属性采用相似的架构，因此可以在很大程度上假定攻击能够泛化到其他作者属性；最重要的是，iii）Beller 等人（2014年）和Emmery 等人（2017年）表明，通过远程标注，可以在不到一天内收集到用于此任务的代表性语料库。这使我们能够衡量使用现实收集的远程语料库拟合的攻击，向采用高质量人工标注语料库的模型迁移的可迁移性。

对于攻击，我们重点关注与给定标签（即属性）密切相关的实词替换，因为已有研究表明，这些词可以解释文体计量学模型准确率的很大一部分（例如，Rao 等人，2010年；Burger 等人，2011年；Sap 等人，2014年；Pardo 等人，2016年）。为此，我们扩展了Jin 等人 (2020年)提出的替换攻击，并将其应用于作者属性混淆。具体而言，我们探索训练一个简单的（以满足速度标准）、非神经替代模型的潜力 f'，以指出需要扰动的相关词语，同时优先保留原文含义。我们将两个基于 Transformer 的模型引入该框架，用于提出并排序词汇替换，以改变 f'的预测结果。我们评估针对 f' 的攻击能否迁移到其他语料库、架构以及一个单独训练的目标模型f （见图2。1）。最后，我们使用自动评估指标衡量变化的质量，并通过人工评估考察攻击的检测准确率。

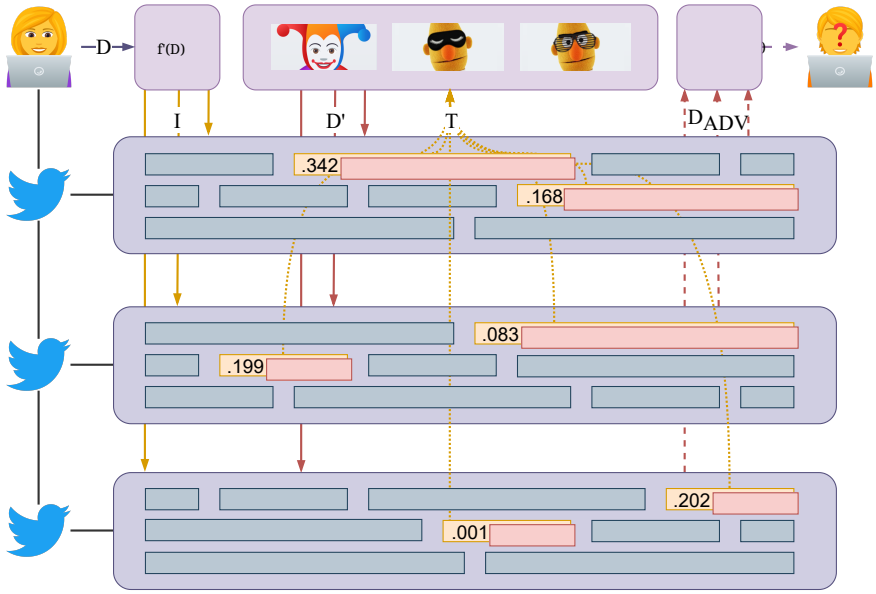


图 2.1： 混淆场景：模型 f′ 在推文批次上训练，使用省略分数根据各词对分类的贡献对其进行判断和排序。随后将这些词传递给 TextFooler、掩码BERT 或 Dropout BERT，以提出排名靠前的 k 个替换候选项。然后，根据这些候选项对 f′(D)的类别概率变化进行选择。最后，对扰动后的推文 D_{ADV}进行评估 f 。

2.2 相关工作

文体计量学是对（主要是）写作风格的研究，其历史可追溯至数十年前（Mosteller 和 Wallace，1963），随着统计模型的引入，其可访问性不断提高（参见Holmes于1998年；Neal 等人于2018年的综述），机器学习也进一步提升了其可访问性（例如，Matthews 和 Merriam，1993年；Merriam 和 Matthews，1994年）。计算文体计量学包含多个子任务，例如判定（Baayen 等人，2002年）和验证作者身份（Koppel 和 Schler，2004），以及作者画像（Argamon 等人，2005）；例如，预测人口统计属性。对抗性文体计量学（其概念由

Brennan 等人，2012年）旨在通过模仿来改变作者的文本，从而颠覆这些推断；或者，就我们的研究而言，通过混淆写作风格来实现这一目的（Kacmarcik 和 Gamon，2006年；Caliskan 等人，2018年；Thi 等人，2015年；Xu 等人，2019年）。

这些变化或扰动可以通过多种方式产生，因此这项任务常常与释义改写（Reddy 和 Knight，2016年）、风格迁移（Kabbara 和 Cheung，2016年）以及生成对抗样本或触发器（Zhang 等人，2020年C）混为一谈。无论采用何种方法，混淆的主要挑战在于保留输入文本的原意；书面语言媒介的特性使扰动只能产生离散输出，而不自然的差异比方说图像中的少量像素变化更容易被人察觉。另一项持续存在的局限是，缺乏能够确保输入文本原意得到完整保留，同时又让变化不被察觉的评估指标（Potthast 等人，2016年）。这不仅阻碍了对混淆进行自动评估，也妨碍了所有自然语言生成（以及一般意义上的自然语言处理）研究（Novikova 等人，2017年），因而更加凸显了人工评估的重要性（van der Lee 等人，2019年）。

或许正因如此，大多数混淆工作采用启发式驱动的受控变化，例如拆分或合并词语或句子、删除停用词、改变拼写、标点或字母大小写（例如参见Karadzhov 等人，2017年；Eger 等人，2019年）。这类特定攻击通常较容易通过预处理来缓解（Juola 和 Vescovi，2011年）。通过词汇替换进行混淆（Mansoorizadeh 等人，2016年；Joo 和 Hwang，2019年；Bevendorff 等人，2020年）在可控性、语义保持和攻击有效性之间取得了折中；然而，对于依赖更深层文体特征（例如词序、词性（POS）标签或阅读复杂度评分）的模型，这种方法可能效果不佳。端到端系统

也被用于类似目的（Shetty 等人，2018年；Saedi 和 Dras，2020年），或者像第1章中以及Bo 等人，2021年所述那样，借助（对抗驱动的）自编码器重写整个短语。这类攻击似乎较不常见，且对扰动和语义一致性的控制也较弱。

我们的工作并不假设攻击以端到端的方式运行，而是假设人在回路中。我们进一步选择那些更有可能找到与原文语义镜像、同时改动最少的技术。我们采用替代模型（作者可以按需操控其算法、超参数和输出）来指示候选替换词；我们的攻击则针对这些候选词给出建议并排序。此外，以往工作通常针对在相同数据上训练的对手发起攻击，而我们还加入了可迁移性度量。最后，尽管已有研究在实际环境中考察了作者识别（Stolerman 等人，2014年），据我们所知，我们的工作是首个有意识地致力于提升混淆技术在现实场景中适用性的研究。

2.3 方法

我们的框架通过多种方式扩展了 TextFooler（TF，Jin 等人，2020年）。首先，我们训练一个替代性别分类器，并利用其针对文档给出的 logit 输出，通过省略分数（第2.3.1节）按照预测重要性对词语进行排序。对于重要性最高的词语，我们提出替代候选项，并为此增加两种技术（第 2.3.2节）。可以根据这些候选项与原词的一致性（例如依据其词性标签）对其进行检查和筛选，原样接受，或重新排序（第 2.3.3节）。对于后一种情况，我们增加了一种评分方法。最后，使用剩余候选项进行迭代替换，直到满足 TF 的停止准则（即预测发生变化，或候选项耗尽）。

2.3.1 目标词重要性

我们给定一个目标分类器 f、替代分类器 f'、由词元 D_i 组成的文档 D，以及目标标签 y。其中，f' 在语料库 X 上进行训练，并接收作者的新输入文本 D，作者为其提供标签 y。如果 $f'(D)$ 预测的不是 y，我们将类别标签记为 $\bar{y}$。我们的扰动构成对抗性输入 D_{ADV}，旨在生成 $f'(D_{ADV}) = \bar{y}$，从而隐式地 $f(D_{ADV}) = \bar{y}$。请注意，我们仅将 D 提交给 f 以评估攻击有效性，而从不使用它来拟合攻击本身。

为了创建 D_{ADV}，我们倾向于采用最少数量的编辑操作，因此根据省略分数（类似于例如 Kádár 等人，2017年）对 D 中的所有词语进行排序，依据的是 f' (`omission_score` 见算法 1)。令 $D_{\setminus i}$表示删除 D_i后得到的文档，$o_y(D)$ 表示其 logit 分数f'。省略分数定义为 $o_y(D) - o_y(D_{\setminus i})$，并用于计算词元 I的重要性分数 D_i，如下所示：

$$I_{D_i} = \begin{cases} o_y(D) - o_y(D_{\setminus i}), \\ \quad \text{if } f'(D) = f'(D_{\setminus i}) = y. \\ o_y(D) - o_y(D_{\setminus i}) + o_{\bar{y}}(D) - o_{\bar{y}}(D_{\setminus i}), \\ \quad \text{if } f'(D) = y, f'(D_{\setminus i}) = \bar{y}, y \neq \bar{y}. \end{cases} \tag{2.1}$$

计算出 I_{D_i} 在 D 中所有词语上的值后，选取排名前 k 的词元作为目标词 T。

2.3.2 词汇替换攻击

我们在实验中考虑了四种确定如何扰动目标词的方法 $t \in T$。这些操作在算法中称为`candidates`算法 1。

算法 1： 通过词汇替换进行混淆。

输入： f'– 替代模型
$D = \{w_0, w_1, \ldots, w_n\}$ – 文档
y – 目标标签
检查 – 应用检查（布尔值）
k – 目标文本最多包含 k 个词

输出： D_{ADV} – 混淆文本

对 $D_i \in D$ **执行**
  $I_{D_i} \leftarrow$ `omission_score`(f', y) `// via Equation 2.1`
结束
$T \leftarrow$ `top_k(argsort_desc(`D, I_{D_i} 分数`), k)`
$D_{ADV} = D$
对于 $t \in T$ **执行**
  `// substitution attack on t`
  $C_t \leftarrow$`candidates(t)`
  $A = (D_{ADV_{1:i-1}}, C_{t,j}, D_{ADV_{i+1:n}})_{1 \leqslant j \leqslant |C_t|}$
  $\bar{A} =$ `filter/rank(D, A; t, 检查)`
  `// test attack success on` f'
  对于 $D' \in \bar{A}$，**执行**
    如果 $\operatorname{argmax} o_y(D') \neq y$，**则**
      返回 $D_{ADV} = D'$
    否则，如果 $o_y(D') < o_y(D_{ADV})$，**则**
      t 中的 D_{ADV} 被替换为来自 c D'的内容
  end
结束
返回 D_{ADV}

同义词替换 (ws) 这种基于TF的替换使用预训练嵌入矩阵 $\mathbf{V}$ 将 t 嵌入为 $\mathbf{t}$。通过计算 $\mathbf{t}$ 与所有可用词嵌入 $\boldsymbol{w} \in \mathbf{V}$ 之间的余弦相似度来选出 C_t。我们用 $\Lambda(\mathbf{t}, \boldsymbol{w})$ 表示余弦相似度。使用阈值 δ 仅保留可靠的候选项 $\Lambda(\mathbf{t}, \boldsymbol{w}) > \delta$。

掩码替换 (MB) 基于嵌入的替换可以改由语言模型来完成，由模型预测语境中最可能出现的词元。BERT（Devlin 等人，2019年a）是一种双向编码器（Vaswani 等人，2017年），通过掩码语言建模和下一句预测进行训练，因此实现这一点相当简单。将 t替换为掩码后，BERT 会为该位置生成最可能的前k 个候选项 C_t 。在 TF 中实现这一方法，意味着每次先前对 t的替换都可能被纳入当前替换的语境中。这种上下文替换方法有两个缺点：i）无法保证与原文中的词语在语义上一致（因为模型并不了解 t）；ii）语境被替换后可能会发生语义漂移，因为后续所有替换都会依据新的、可能（在语义上）不正确的语境进行。

Dropout 替换(DB) Zhou 等人Zhou 等人(2019年)提出了一种绕过前述问题（即 BERT 在词汇替换时受掩码预测限制）的方法。他们对目标词Srivastavaet al.，2014年)的 BERT 内部嵌入应用 Dropout（ t ，然后再将其传入 Transformer——以一定概率将部分权重置零。他们假设， C_t （BERT 的前k个候选项）会比掩码建议更接近原文 t 。

启发式替换 为评估前文所述各项技术的相对表现，我们采用了多种启发式攻击作为基线方法。按表 2.3中的顺序：1337语：将字符转换为黑客语（13375P34K）变体，类似于变音符号转换（Belinkov 和 Bisk，2018年）。字符翻转：交换单词中间的两个字符（wrod），研究表明这种做法对可读性的影响最小（Rayner 等人，2006年）。随机空格：在随机位置将一个词元拆分为两部分（pos ition）。

	作者	推文	女性	MALE	训练集	TEST
Huang	37,929	47K	26,758	20,453	30,602	7,651
Emmmery	6,610	16.789K	61,736	32,900	75,918	18,718
Volkova	4,620	12.227K	32,376	26,708	47,298	11,777

表 2.1： 语料库统计，列出了作者数量、推文数量、女性和男性标签数量、训练集和测试集划分的规模、类型（不重复词语）和词元（总词数）的数量，以及每篇文档的平均词元数（平均规模）。

2.3.3 候选项过滤与重新排序

给定 C_t，既可以接受所有候选项，也可以只接受排名最高的候选项。或者，可以将所有 D′ 提交给检查进行过滤，或根据它们与 D的语义一致性重新排序。这些操作称为 `rank/filter` ，见算法1——两者都可以执行。

词性-与-文档编码TensorFlow 使用两个检查组件：首先，它会移除任何 c 其词性标签与t不同的项。如果存在多个 D′ 满足 $f'(D') = g$的项，则选择文档D′，其与原始文档的通用句子编码器（USE）嵌入的余弦相似度最高（Cer 等人，2018年）D。否则，选择 D′ 中目标词省略分数最低的项（按照 TensorFlow 的方法）。

BERT 相似度Zhou 等人（2019年）使用 BERT 最后四层的拼接作为句子的上下文表示 h。我们在 Masked（MB）和 Dropout（DB）BERT 中都应用这一方法，根据嵌入对所有可能的 D′ 重新排序。给定文档 D、目标文本 t和扰动候选文档 D′， C_t 将通过一个

嵌入相似度分数：

$$\textsc{sim}\left(D, D'; t\right) = \sum_{i}^{n} w_{i,t} \times \Lambda\left(\mathbf{h}(D_i|D), \mathbf{h}(D'_i|D')\right) \tag{2.2}$$

其中，$\mathbf{h}(D_i|D)$ 是 D中 i^{th}词元的 BERT 上下文表示，而$w_{i,\,t}$ 是所有层中所有注意力头的平均自注意力分数，表示从 i^{th} 词元到 D中 t 的注意力。[3]

2.4 实验

2.4.1 数据

我们使用了三个作者画像数据集（见表2.1 有关统计信息），这些数据集均标注了二元性别分类（男性或女性）：首先是 Volkova 等人 (2015年) 通过 Mechanical Turk 众包收集并标注的数据集，包含 5,000[4] 英语推特个人资料。可以将其视为 Twitter 个人资料的“随机”样本，因此是三个数据集中偏见最少的一个。因此，我们认为它最能代表作者画像数据集，并将其用作训练集（80%），用于 f，并将其用作攻击的测试集（20%）。

第二个是Huang 等人(2020年发表的多语言仇恨言论公平性语料库中的英语部分，该语料库的收集目的与作者画像不同。它汇集了现有的仇恨言论语料库（由Waseem 和 Hovy、2016年；Waseem、2016年；Founta 等人、2018年）构成——这些语料库主要通过查找词、筛选经常使用辱骂性语言的用户等方式进行初步构建——并在事后标注了人口统计信息。该数据集的收集并未侧重于

[3] Zhou 等人（2019）还使用提议分数来查找 T，我们将其替换为省略分数。[4] 由于收集限制（例如，已删除的账户），当前研究中的个人资料数量有所不同。

个人资料，而且大多数作者只与一条推文相关联。与一般的作者画像相比，这可能会造成显著的领域偏移。不过，它可以被视为免费提供的（带有噪声的）数据。

最后，我们还纳入了一个由Emmery 等人（2017年）构建的弱标签作者画像语料库，该语料库通过英语关键词检索自我报告收集而来——这与Beller 等人（2014年）的方法类似。该语料库可能包含错误标签，但其收集耗时不到一天，因此是一个理想的候选数据集，能够真实地获取（新的）数据来拟合替代模型。

预处理& 采样所有语料库均使用 spaCy 进行词元化[5]（Honnibal 等人，2020年）。除转为小写、为用户提及和话题标签分配特殊标记（#与文本被拆分）以及 URL 移除外，未执行其他预处理步骤。每个作者时间线都被划分为最多包含100 条推文的区块（即部分区块包含的推文更少），这意味着每位作者最多有25 个实例（部分作者包含一个，2，500 是 API 历史记录上限）。从测试集中，对最后的[6] 200 个实例进行了攻击采样（110 名男性，90 名女性）。尽管样本规模相当小，但它确实反映了针对单个个人资料执行攻击时的真实攻击持续时间和时间线规模。

2.4.2 攻击

为了扩展 TensorFlow，我们重新实现了代码[7]，由Jin 等人(2020年编写，使其能够与 Scikit-learn 配合使用[8] (Pedregosa 等人，2011年）。对于他们的同义词替换组件，我们同样使用了由Mrkši´c 等人(2016年)在 Simlex-999 上训练的 (Hill 等人，2015 年）。

[5] https://spacy.io [6] 由于数据集未经过随机打乱，以避免对作者特定特征过拟合，因此同一作者的少量文档可能会从训练集泄漏到测试集；这样可避免将这些文档纳入我们的攻击样本。[7] https://github.com/jind11/TextFooler
[8] https://scikit-learn.org/

通用句子编码器（Cer 等人，2018年）的实现使用 TensorFlow[9] (Abadi 等人，2016年a）作为后端，而所有 BERT 变体均使用 Transformers[10] 库（Wolf 等人，2020年），并以 PyTorch[11] (Paszke 等人，2019年）作为后端。

我们在所有 TF 实验中均采用 Jin 等人(2020年) 使用的参数设置：他们通过经验将 N （视为同义词）和 δ （余弦相似度下限）分别设为 50 和 0。7 。对于 MB 和 DB，我们将 T 限制为 50 ，并将 top-k 限制为 10 （以提高速度）。对于 DB，我们遵循 Zhou 等人(2019年) 的做法，将丢弃概率设为 0。3。

2.4.3 模型

对于 f 和 f′，我们需要（最好是快速的）流水线，既能在作者画像任务中取得较高准确率，又能充分区分开来，以衡量我们的攻击在不同架构之间的迁移效果，而不只是跨语料库的迁移效果。由于当前最先进算法尚未证明其在作者画像方面（Joo 和 Hwang，2019年）具有显著效果，我们选择使用 n元特征和线性模型。

逻辑回归 采用单元语法和二元语法特征、基于 tf·idf 训练的逻辑回归（LR），在以往的作者画像研究中被证明是强大的基线。该分类器的简单性也使其成为一种替代模型，作者实际上可以运行它。未进行任何调优： C 设为 1。

N-GRAM 由 Basile 等人 (2017年) 提出，作为一种高效且简单的模型，新格罗宁根作者画像模型（N-GrAM）在作者画像任务中的表现优于更复杂的（神经网络）替代模型，且几乎无需调优。它使用 tf·idf 加权的一元和二元-

9 https://tensorflow.org/
10 https://huggingface.co/
11 https://pytorch.org/

DATA	Huang、Emmery、Volkova
重要性	省略分数
攻击	启发式方法、TextFooler、掩码BERT、Dropout BERT
模型	逻辑回归、N-GrAM
排序	None, 词性 + 通用句子编码器, BERT 相似度

表2。2：可能的实验配置网格。

gram 词元特征、字符六元组以及经过次线性缩放的词频（$1+\log(\text{tf})$）。随后将这些特征输入线性支持向量机（Cortes 和 Vapnik，1995年；Fan 等人，2008年），其中 $C=1$。

2.4.4 实验设置

概括而言（另见表 2.2），实验按如下方式进行：在给定语料库上训练替代目标模型（f'）——所有实验均使用逻辑回归。真实目标模型（f，即逻辑回归或 N-GrAM）始终在 Volkova 等人 (2015年）200个样本用于评估攻击。通过 f'的省略分数对目标词进行排序，再将其输入启发式方法、TF、MB 或 DB 攻击。启发式方法会直接更改目标词，其余方法则输出一组按排名排列的替换候选项。后者可以通过 TF 流水线针对 f' 进行评估，也可以直接返回排名第 1 的候选项。可以通过词性标注/通用句子编码器进行筛选，以评估语义相似度并进行词性兼容性检查（Check），也可以不筛选（Check）。

请注意，我们主要关注可迁移性，因此希望尽可能测试多种数据访问限制与架构访问限制的组合。假设作者无法访问数据，则替代分类器使用除 Volkova 等人 的语料库以外的任何数据进行训练。假设作者不知道目标模型架构，则目标模型为 N-GrAM（而不是

		测试集 = Volkova 等人					
	逻辑回归 f′ →	Huang 等人		Emmery 等人		Volkova 等人	
	f →	LR	NG	LR	NG	LR	NG
	none	.885	.940	.885	.940	.885	.940
Heuristic	1337	.770	.850	.775	.835	.715	.860
	flip	.900	.950	.885	.905	.840	.905
	空间	.845	.925	.760	.870	.720	.850
Top-1	WS	.825	.930	.805	.890	.750	.915
	MB	.655	.905	.595	.785	.145	.410
	DB	.625	.895	.575	.785	.210	.530
~~Check~~	WS	.540	.855	.355	.670	.000	.009
	MB	.415	.790	.120	.420	.000	.085
	DB	.430	.775	.175	.430	.000	.085
Check	TF	.705	.920	.780	.910	.375	.700
	TensorFlow + MB	.640	.880	.760	.890	.380	.725
	TensorFlow + DB	.650	.885	.755	.890	.435	.715

表 2.3： 在55%）基准水平以下（ =表现较好的攻击后准确率分数， f 在Volkova 等人 语料库的测试样本上。左侧为攻击条件：启发式方法、排名第一的同义词、应用词性和 USE 相似度检查，或不应用这些检查（Check）。按训练语料库划分的结果标注于 ~~f′~~（始终使用逻辑回归（LR））。目标模型使用 LR 或 N-GrAM（NG）。替换攻击包括 TextFooler（TF）、Masked（MB）和 Dropout BERT（DB）。如果使用 TF 的停止准则，则标注为 TF +。词语相似度（WS）表示未进行检查的 TF 流水线。

而非逻辑回归）。因此，一个完整的模型迁移设置（数据和架构均不同）例如：数据 f′ = Emmery 等人，数据 f = Volkova 等人， f′ = 逻辑回归，以及 f = NGrAM。最后，为了与最优情况进行比较，我们测试了一个可以访问对手数据的设置。

	Volkova 等人 →	训练集	TEST
训练集	Huang 等人	.640	.620
	Emmery 等人	.725	.890

表 2.4： f的数据划分上，替代模型的性别预测准确率为 f′。

2.4.5 评估

度量指标混淆成功的衡量标准是准确率分数低于随机水平性能；根据我们的测试样本，该阈值为55%。我们认为，在不希望预测出相反标签的情境中，随机性能更可取。然而，对于当前任务，准确率降至随机水平附近或更低，即满足混淆成功的条件。[12] 为了评估受攻击句子的语义保持情况，我们计算了METEOR（Banerjee 和 Lavie，2005年；Lavie 和 Denkowski，2009）以及 `nltk`，[13] 和 BERTScore（Zhang 等人，2020年b）在 D和D_{ADV}之间的相似度。METEOR衡量灵活的一元语法标记重叠，包括形态变体；BERTScore 则根据句子上下文计算相似度。

人工评估 对于人工评估，我们从表现最佳的实验配置中，针对每种攻击类型抽取了 20 个文档片段（包含一条或多条推文）。片段需满足以下条件：i) 包含针对三种攻击的变化；ii) 至少由 15 个单词组成（不包括表情符号和标签）；iii) 不

[12]如果攻击将准确率降至 0%，那么（对于二元标签）实际上就会翻转该标签。作者可能也不希望得到这个标签（例如，被分类为持有截然相反的政治观点）。这意味着，目标模型对分类越不确定越理想（另见第1章）。

[13] https://www.nltk.org/_modules/nltk/translate/meteor_score.html

包含明显的脏话。[14] 三个模型的全部 60 个文档片段经过了随机打乱，并在末尾追加了 20 个原文版本（因此“正确”的片段会最后出现）。因此，每个文本变换模型都有 80 个项目用于评估。

虽然以往研究通常会对语义一致性和流畅度进行评分，并为文本标注类别（例如参见 Potthast 等人，2016年；Jin 等人，2020年），但我们的 Twitter 数据噪声过大（原文中包含大量拼写和语法错误），且文档批次过长，因此难以执行这项任务。相反，我们要求六名参与者（每种替代方法各三人）指出：a）某个句子是否经过人工修改；如果是，b）指出一个引起他们怀疑的词语。这样，我们就能评估哪种攻击生成的句子最自然，以及对输入内容的改动最不明显。

这些项目分别进行评分；人工评估者事先并不知道同一句子的不同版本会重复出现，也不知道原文会在最后展示。所有参与者都具有大学学历、较高的英语水平，并且熟悉该数据领域。在表2中可以找到同一句子的几项示例评分。6。

2.5 结果

2.5.1 领域偏移

正如第 2.4.1 节所述，用于训练替代模型的两个语料库实际上都不是用于作者画像的参考语料库，因此可以视为次优的、不相交的领域。尤其是 Huang 等人的语料库表现出明显的领域偏移（见表 2.4）

[14]为了避免评分者接触到过于有害的内容，我们使用关键词列表过滤掉了明显的示例。仍有少数轻微示例，因此我们添加了免责声明。

对于训练集和测试集都是如此。远程标注的 Emmery 等人语料库在 Volkova 等人的训练集划分上取得了高出 7.5% 的准确率，且测试集性能显著更高（27%）。因此，我们可能会预期后者具有更好的混淆性能。

2.5.2 基线方法

所有攻击的结果见表 2.3。请注意，这些是针对 f 的性能；因此，在未应用任何攻击（none）的情况下，两种替代语料库的性能保持不变（因为它们只会影响攻击）。对于启发式攻击，1337 似乎能构成更稳健的基线，在可迁移性方面也优于其他一些设置。从表面上看，这种攻击的优势在于，对可读性的影响较小（谨慎应用时），且不会改变语义；然而，启发式攻击很容易通过预处理（Juola 和 Vescovi，2011年）以及字符特征加以缓解（N- GrAM 模型的性能便说明了这一点）。要实现可迁移性，显然不能只靠添加噪声将词语转换为词汇表外形式（OOV）。尽管可以认为启发式方法能够改变所有词语，但若将所有内容都变成词汇表外形式，就不够稳健；目标模型一方很容易发现异常输入，并据此采取行动（例如拒绝输入）。

2.5.3 攻击可迁移性

可通过比较逻辑回归和 N-GrAM（NG）两列来评估可迁移性。总体而言，可以看出，在 Emmery 等人的语料库上训练的替代模型，始终优于在 Huang 等人的语料库上训练的模型；无论对手的架构已知（逻辑回归）还是未知（NG），均是如此。这与我们根据观察到的领域偏移所得的预期一致。我们的结果还

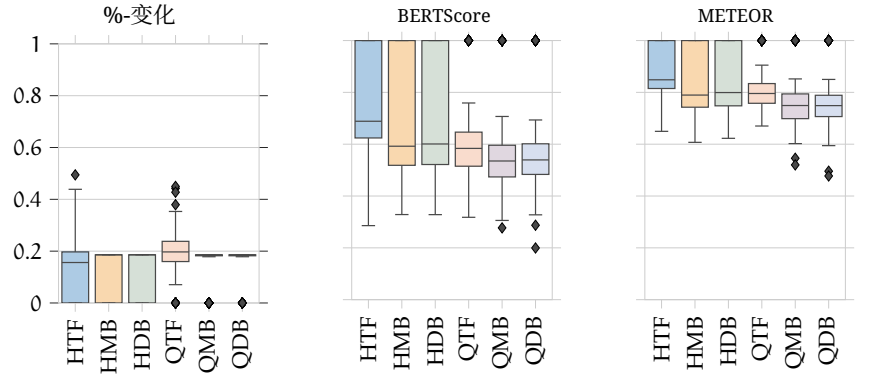


图 2.2： 按语料库划分的三种不同攻击的改写指标：TextFooler (TF)、Masked (MB) 和 Dropout BERT (DB)Huang 等人 (H)，Emmery 等人 (Q)。图中展示了变化的相对数量（按文档长度计算）、BERTScore (F_1)，以及 METEOR(MET) 相对于原文档的数值。

表明，混淆性能会明显下降（10 – 30目标模型性能提升百分比），当攻击被迁移到不同的数据和另一种模型时。相反，正如表 2。3的最后两列所示，在模型和数据均可获得这一在实践中不现实的设定下，混淆会专门针对已知弱点进行调整，因此效果显著。[15]

2.5.4 *Transformer* 性能

查看 Top-1、Check 和 ~~Check~~ brackets（表2。3）可见，除了基于 BERT 的模型比 TF 具有更高的可迁移性成功率外，它们也保留了混淆成功率；在未使用额外检查的设置下，目标模型的性能会降至随机水平以下（55%）。这一比较还表明

[15]Jin 等人（2020年）发现，对于词级模型，相同比例的变化也会导致准确率降至 0%。

		Huang 等人			Emmery 等人		
	ORG	TF	MB	DB	TF	MB	DB
改变	.888	.967	.633	.783	.950	.617	.633
WORD	—	.950	.583	.700	.867	.433	.433

表 2.5： 预测文本是否被改变以及猜测受攻击词的人工准确率分数（越低越好）。所有替代模型均采用 Check 设置，并在不同语料库上训练（即，每个训练语料库对应—不同的受攻击词）。ORG 表示对原文的正确预测。

同义词排序可以发挥作用（Top-1 与 Check 和 Check 相比），而 Check 条件过于严格，导致攻击能力较低且可迁移性较差。图2。2所示的百分比变化进一步说明了这一点。比较 MB 和 DB 变体，其性能似乎几乎相同，其中掩码略占优势。正如 Zhou 等人（2019年）所指出的，应用 dropout 应生成更接近原文（与 MB 相比）的词语，这可能会影响混淆性能。此外，对掩码替换候选项应用 BERT 相似度排序（见第 2。3。3节）也可能产生一些有益效果。这一点还需要借助输出评估进行更详细的研究。

改写指标图 2.2 中的度量指标显示，在将其应用于此任务时存在一个普遍的初始局限：攻击越常不作任何改变，自动评估度量指标（BERTScore、METEOR）就越高。因此，比较模型时，需要结合混淆性能和相关工作来考量这些分数。可以看出，尽管 MB 和 DB 的变化始终较多，但它们在语义一致性上的得分仍低于 TF。不过，对于 Emmery 等人的语料库，MB、DB 与 TF 之间的这些差异很小。此外，尽管是在不同的

ORG 已经准备好回家了。与上帝的关系更好。我需要尽快找另一份工作。 htf **贷款**回家 **房子**已经准备好了。与**耶稣**的关系**有所改善**。我**应该**再找一份**工作**，越快越好。hmb 已经准备好**继续**回家了。与上帝的关系更好。我需要另一个**男人**。hdb 已经准备好**去某处**了。与上帝的关系更好。我需要另一个**职位**与。ORG 特朗普在导弹发射后批评金正恩：“这家伙就没有更好的事可做吗？”
htf **特朗普在火箭发射后批评金正恩**：“这个**男朋友**就没有**更好**的事可做吗？”
hmb 特朗普在**女议员**发起**竞选**后提出批评：“这本**书**就没有**别的**事可做吗？”
hdb 特朗普在**辛**发起**炸弹**后提出批评：“这个**孩子**就没有**更少**的事可做吗？”

表 2.6： 两个语义连贯性和人工检测准确率各不相同的句子在遭受不同攻击时的示例评分（未同时展示给人工评估者）。在第一个示例中，所有评分者都将 hmb 标记为未改变，多数评分者将 hdb 标记为未改变，而没有人将 htf 标记为未改变。在第二个示例中，只有一位评分者将 hdb 标记为未改变。受攻击词以粗体标出，猜中其中任意一个词都算正确识别了攻击。

领域内，这些分数与先前的文本混淆研究结果相当（例如，Shetty 等人（2018年）显示 METEOR 分数介于 0.69 和0.79）。

人工评估 表 2。5中的结果2。3呈现了与表中相同的趋势；混淆成功率高，似乎会导致人在判断句子是否经过混淆时更容易出错。相反，尽管语义一致性得分较高，原始 TensorFlow 流水线似乎更容易被检测出来。这可以

归因于模型在缺少额外检查时产生的拼写和语法错误。此外，识别原始句子时的 11% 错误率也反映了这项任务中预期存在的一定误差范围，因为我们的 Twitter 数据本身就带有噪声。最后，虽然这些结果与混淆成功率相符，且低于相关工作中的可检测性得分（Mahmood等人，2020年），但也表明这些模型仍能以高于随机猜测的水平被检测出来。在三种选项（包括原文）的情况下，表现应达到 25% 或更低，才表明文本没有经过侵入性改动（即改动在语义上不连贯，或不够隐蔽——这两项度量指标由Potthast 等人，2016年）使用。因此，尽管所提出的方法有效且在现实中具有可迁移性，但在实际适用性方面仍有改进空间。

2.6 讨论与未来工作

我们展示了在真实场景下作者属性混淆的效果。我们使用一个简单的逻辑回归模型进行候选项推荐，该模型在不到一天内收集的弱标注语料库上训练，攻击成功迁移到了不同的数据和架构。这一结果为针对该任务及其实际应用开展未来的对抗性研究带来了希望。

要自动评估成功混淆所需的变更数量及其侵入程度仍然颇具挑战，尤其是这些变更对作者整体消息一致性的影响。不过在实践中，这类考量可以交由作者决定。在这种人在回路场景中，可以提供更广泛的候选项，并以交互方式展示它们对替代模型的影响。这样一来，便可以手动调整攻击，在有效性、不显眼性与语义一致性之间取得平衡。它还可以向作者展示其写作风格

如何影响未来可能产生的推断。

关于攻击的性能：我们展示了语境语言模型在检索候选项推荐方面的总体有效性。然而，通过更广泛的基于规则的检查，可以进一步提升这些候选项的质量；例如，利用解析进行更深入的分析。尽管如此，这类方法仍使语言重写——进而更广泛地说，自然语言处理——面临一个核心限制：虽然在我们的实验中，Masked 攻击似乎更为成功，但人工检查扰动后发现，Dropout 在语义上往往更加接近（另见表 2。6）——而这并未反映在人工评估中。这引出了一个问题：在当前语义一致性指标的局限下进行评估时，任何 自动化方法是否能够真正同时优化混淆效果和不显眼性。

因此，我们认为，未来工作应当着重于尽可能减少扰动，只保留取得混淆成功所必需的最低限度。鉴于此，其他约束就不再那么重要；可以生成较短的句子（例如单条推文），这些句子在语义或语境上可能并不正确，但如果将其作为一条消息插入较长的发帖历史中，几乎不会被察觉，也不会造成干扰。这可能需要特定的对抗触发器（例如 Wallace 等人 (2019年) 所展示的触发器），并确定它们的迁移效果。

2.7 结论

在本章中，我们指出，在无法访问目标模型的数据或架构的情境下，应通过可迁移性来测试现实的对抗性文体计量。我们将先前的对抗性文本分类研究扩展到两种基于 Transformer 的模型，并研究了它们在这种情境下的混淆效果。我们证明了它们

能够可靠地将目标模型性能降至随机猜测水平以下，尽管人类对这些攻击的可检测性仍高于随机猜测水平。未来工作可以着重于在我们的现实约束下进一步降低这种可检测性。



本章已发表于：

Chris Emmery、Ben Verhoeven、Guy De Pauw、Gilles Jacobs、Cynthia Van Hee、Els Lefever、Bart Desmet、Véronique Hoste 和 Walter Daelemans。2021年b. 网络欺凌检测的当前局限：关于评估标准、可复现性和数据稀缺。语言资源与评估，55(3)：597–633

已对文本和图表进行了细微的格式调整。

3

当前网络欺凌检测的局限性：评估标准、可复现性和数据稀缺

DETECTION

网络欺凌在社会重要性、研究领域的受欢迎程度以及可获取的开放数据方面均有所增加。然而，尽管计算能力和资源的可负担性不断提升，高质量数据的访问限制仍然限制了最先进技术的适用性。因此，近期许多研究使用小型、异构数据集，却未对其适用性进行全面评估。在本章中，我们将进一步阐述这些问题，因为我们（i）评估了许多公开可用的资源，并展示了数据收集方面的困难。这些资源主要产生了小型数据集，无法捕捉所需的复杂社会动态，也妨碍了对研究进展进行直接比较。我们（ii）开展了一系列广泛的实验，表明基于这些来源训练的分类器普遍缺乏跨领域泛化能力，并公开提供这一框架，以便复现和扩展我们的评估标准。最后，我们

（iii）提出了一种有效的众包方法：在实验室环境中模拟现实生活中的欺凌场景，可以生成可信的数据，并有效用于丰富真实数据。这在很大程度上规避了可收集数据方面的限制，并提升了分类器的性能。我们相信，这些贡献将有助于改进该领域未来研究的实证实践。

3.1 引言

在海量数据流中学习准确分类罕见现象，对机器学习的诸多应用构成了挑战。纳入具有代表性的实例往往需要耗费大量资源，而获取此类实例的途径有限，会严重影响预测结果的可靠性。这些局限在处理敏感社会现象的应用中尤为普遍，例如法证学中的相关应用：预测恐怖主义行为、检测欺诈，或揭露性越轨行为。这些事件错综复杂，需要丰富的表征才能有效检测。与之相反，承载此类互动的在线文本、图像和元数据对托管它们的平台具有商业价值，因此通常受到限制，以保护用户隐私。

过去十年间，网络欺凌检测因其日益凸显的社会重要性和不断增长的关注度，成为一项受到此类局限影响的应用。其数据不仅敏感，而且从公共获取渠道来看，天生就十分稀缺。大多数网络欺凌事件都不为大多数研究者所能接触，因为它们发生在私人对话中。要全面捕捉这些事件的社会动态和复杂性，需要比迄今研究群体所能获取的数据丰富得多的数据。与此相关，Rosa 等人(2019年b近期指出了网络欺凌检测研究在操作化方面存在的各种问题，他们的许多担忧与我们将在本章讨论的问题相同。

他们的研究侧重于以往研究的方法严谨性，而我们将关注该领域的核心局限以及网络欺凌检测的复杂性。通过评估该任务的当前进展，我们说明这些问题如何影响当前研究，尤其是跨领域研究。最后，我们展示如何在实验环境中运用众包，以期缓解该任务的数据稀缺问题。不过，我们首先介绍网络欺凌的理论框架，以及自动检测此类事件的任务。

3.1.1 网络欺凌

异步性和可选择的匿名性是当今我们所熟知的在线交流的特点；在线交流在很大程度上依赖于与不在场的人沟通的能力，并通过社交网络促进人们与亲密朋友群体之外的人互动（Madden 等人，2013年）。这些网络的兴起为青少年带来了诸多益处：研究显示，在线交流与社会联结之间存在积极关系（Bessière 等人，2008年；Valkenburg 和 Peter，2007年），而在这些网络上的自我披露有助于提升既有关系和新建立关系的质量（Steijn 和 Schouten，2013年）。社交网络和即时通讯在儿童中的普及，使他们开始接入互联网的年龄越来越小（Ólafsson 等人，2013），95% 的青少年[1] 年龄 12–17 岁会上网，其中80% 使用社交媒体（Lenhart 等人，2011年）。然而，对他们而言，社会互动从主要发生在操场上转变为通过移动设备进行（Livingstone 等人，2011年，也使负面通信转移到了一个平台

[1] 2011年对799名美国青少年开展的综述。黑人和拉丁裔家庭的抽样比例较高。

在这个平台上，间接且匿名的互动得以进入人们的家中。

皮尤研究中心[2]，开展的一系列研究，尤其是 Lenhart 等人（2011）的研究，为这些发展提供了详细见解。尽管78%的青少年表示，他们在社交媒体上的互动带来了积极结果，但41%的人经历过至少一些负面结果，包括争吵、与学校和父母发生矛盾、肢体冲突以及友谊破裂。在研究前12个月内遭受欺凌的青少年中，有19%；所有青少年中有8%表示，这属于某种形式的网络欺凌。这些数字与其他研究的结果相当（6至12年级为7%，9至12年级为15%）（Kann 等人，2014年；Robers 等人，2015年）。长期以来，众多机构都将欺凌视为公共卫生风险（Xu 等人，2012年），其主要担忧包括抑郁、焦虑、低自尊、缺课、成绩下降以及自行用药的风险。

网络欺凌这一行为除了发生在线上之外，还具有传统欺凌的特征：欺凌者与受害者之间存在权力失衡（Stratford，1996年），伤害是故意造成的，会随着时间推移而反复发生，并对受害者造成负面的心理影响（Dehue 等人，2008年）。然而，互联网作为一种沟通平台，也带来了一些额外的特点：位置、时间和是否身处现场已不再是相关因素。因此，人们定义了几种这种欺凌形式所特有的类别（Willard，2007年；Beran 和 Li，2008年）：网络辱骂（发送粗鲁或粗俗的消息）、隐私曝光（未经同意发布某人的私人信息或经过篡改的个人资料）、骚扰（反复向同一个人发送冒犯性消息）、排斥（将某人排除在某个线上群组之外）、网络跟踪（通过发送明确带有威胁和恐吓意味的消息进行恐吓）、诋毁（在网上散布流言蜚语），以及冒充。此外，除了可选择的

[2] www.pewinternet.org

匿名性会隐藏网络欺凌行为背后的关键人物，也可能掩盖参与者的数量（也就是说，表面上看似有多人参与，实际上可能只有一人）。网络欺凌行为一经发布，往往难以删除；消息或图像可能在转发过程中持续存在，并被许多人看到（仇恨页面通常如此），也可能只有少数人能够看到（在群组对话或私聊中）。因此，可以认为，各种形式的骚扰都变得更容易实施，也更具侵扰性。这种线上特性也有其优势：理论上，平台会记录这些欺凌实例。因此，越来越多的研究关注网络欺凌的自动检测（和预防）。

3.1.2 检测与任务复杂度

网络欺凌检测任务通常可定义为：运用机器学习技术，自动对消息中的文本是否包含欺凌内容进行分类，或依据更高阶的信息（例如用户特征或社交网络属性）推断特征。欺凌在较年轻群体中最为明显，通常表现为直接的言语隐私曝光（Vaez 等人，2004）；在年长群体中则更为隐晦，主要表现为更复杂的社会动态，例如排斥、破坏和流言蜚语（Privitera 和 Campbell，2009）。因此，无论是有意为之，还是由于数据的主要来源是社交媒体（这很可能导致某些媒体上此类内容大量出现），有关这一主题的大多数研究都聚焦于较年轻群体（Duggan，2015年）。除了语言使用本身带来的公认挑战（如歧义、讽刺、方言、俚语和新词）之外，事件中的两个因素还增加了语言上的复杂性，即参与者的角色以及相关的语境。与只需根据单条消息文本中提供的充分信息即可完成的任务不同，要完整描绘网络欺凌事件并确定欺凌者

和受害者，需要理解相关参与者之间的动态关系，并对语域进行文本解读。

语域 首先，要将网络欺凌检测理解为文本分类的一个特定领域，就应考虑界定该领域的语域全貌。第 3.1.1 节讨论的欺凌类别中，有些线索仅凭文本即可推断；其中，网络辱骂最为明显，通常表现为简单地使用脏话、侮辱性称呼或其他粗俗语言。同样，属于网络跟踪的威胁性或恐吓性消息，也可通过特定词语的使用清晰识别。其他类别则更为隐晦：隐私曝光也可以通过文本实现，例如发布电话号码，或透露个人性质或敏感性质的信息。诋毁所包含的词语并不一定会明显与辱骂行为相关；然而，关于敏感话题的虚假信息可能会与受害者的姓名结合出现。根据相关文献（也见 Van Hee 等人，2018年b），还可以进一步扩展这些线索，将欺凌事件线索纳入其中，例如为受害者辩护的消息，以及支持欺凌者的消息。因此，这一语言学任务可以（部分基于 Van Hee 等人，2018年）表述为 识别包含针对受害者的攻击性或伤害性内容的在线消息语境。然而，这些语境中的若干其他交际要素会进一步改变对这些线索的解读。

角色其次，网络欺凌事件中通常会区分若干参与者。简单的角色分配包括欺凌者 B、受害者 V 和旁观者 BY，后者可能赞同也可能不赞同欺凌行为。更细致的模型，例如Xu 等人(2012年)提出的模型，还纳入了其他角色（参见图3.1，其中展示了角色互动可视化）；不同角色可以分配给同一个人，例如，既遭受欺凌又进行举报。

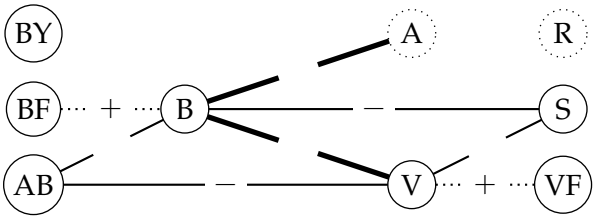


图 3.1： 欺凌事件的角色图。每个顶点代表一名参与者，并以其在事件中的角色标注：欺凌者（B）、受害者（V）、旁观者（BY）、助长者（AB）、协助者（BF）、保护者（S）、举报者（S）、指控者（A）和朋友（VF）。每条边表示一条通信流，并标注其性质是正面（+）还是负面（–），边的粗细表示互动频率。虚线边表示未参与该事件；带虚线的顶点则表示由Xu 等人 (2012年)添加，以涵盖社交媒体特定角色。

最重要的是，如表3.1所示，社交媒体上的单个讨论串中可能出现所有这些角色。有些角色会通过频繁互动清楚地表现出积极或负面的情感（B， V， A），另一些角色则可能无法通过任何形式的对话观察到（R， BY），也可能过于隐晦，或难以与其他角色区分。

语境 第三，针对不同角色，必须以不同方式解读消息内容。虽然脏话可能是骚扰的良好迹象，但要识别欺凌者，仅凭这些显然还不够。请看表 3。1：两者 B和 A 都使用了侮辱性言语（第 7–8 行）， V 的消息（第 6 行）单独来看可能被视为欺凌；而在已经确定 B之后，最后一句（第 10 行）通常可以被视为威胁。总之，这项任务的整体范围很复杂；它可能具有时间序列特征，需要确定参与者及其互动，还应考虑严重程度（例如，区分欺凌与戏弄）。

Line	Role	消息	欺凌者	Type
1	V	我和我的朋友们今晚一起出去玩！:)		中性
2	B	@V 哈哈，*婊子，你根本没有朋友……你是个冒牌货像是狗屎	✓	咒骂，侮辱
3	AB	@B 哈哈，宝贝，她真可悲	✓	鼓励
4	VF	@V 你心知肚明，姑娘		
5	S	@V 别听@B的，我们会玩得很开心，当然！		辩护
6	V	@B闭嘴@B！！没人问你的意见！！！！		辩护
7	A	@B你真是个他妈的欺凌者，出去走走什么的吧		侮辱
8	B	@V @S 哈哈，你们都蠢死了，干脆去死吧已经！	✓	侮辱，咒骂
9	A, R	@B 闭嘴，不然我就举报你		
10	B	@A 你要哭了吗？尽管试试，看会有什么后果明天！	✓	威胁

表 3。1： 网络欺凌对话的虚构示例。各行表示按顺序进行的轮次。角色的标注方式如第 78 页（同名段落下）所述；根据 ✓ 判断消息是否可视为欺凌，并依据 Van Hee 等人(2015年b) 的标注指南确定类型。

3.1.3 我们的贡献

令人惊讶的是，该任务的大量研究并未收集（或使用）能够推断此类特征的数据（我们将在第 3.3 节中进一步阐述）。为证实这一点，我们在不同来源的数据上复现了部分先前的网络欺凌检测研究。我们表明，当前自动化方法对网络欺凌分类所作的预测并未体现上述任务的复杂性；我们展示了模型在不同训练领域上的性能下降，并探讨了内容特征的重要性及其局限性。此外，我们报告了当前最先进研究在接受我们的评估时出现的可复现性问题。为便于后续复现，我们将开源所有代码，包括数据集读取器、

实验代码和定性分析。[3] 最后，我们提出了一种在实验环境中收集众包网络欺凌数据的方法。该方法可以控制数据的规模和丰富程度，不会侵犯隐私，也不依赖外部各方来协助获取数据。最重要的是，我们证明了该方法能够有效提升分类器性能。通过本章，我们提出了改进方法严谨性的建议，并希望帮助学界以更贴近现实的方式评估和实施这项任务。

3.2 相关工作

网络欺凌内容检测任务大致可分为三类。第一类研究聚焦于二元分类，只关注一条消息是否包含欺凌内容。第二类研究采用更细粒度的方法，任务是确定参与者在欺凌场景中的角色，或判定内容类型（即不同的欺凌类别）。二元和细粒度方法都主要关注基于文本的特征。最后，元数据方法不仅考虑消息内容，还会纳入个人资料、网络或图像信息。下文将围绕这三个主题，讨论与网络欺凌分类任务相关的研究。我们将主要关注基于公开可用数据开展的研究，以及报告了（正面）F_1分数的研究，以促进公平比较。[4] 关于全面的文献综述和不同研究的详细比较，请参阅Rosa 等人(2019年b)。最后，我们的研究有很大一部分涉及生成……

[3] 可在此处获取 https://github.com/cmry/amica。[4] 遗憾的是，许多近期关于网络欺凌检测的研究似乎并未报告此类 F_1分数（而是倾向于报告准确率），或仅限于使用受到批评且基线分数较高的数据集（例如 CAW 数据集），或方法严谨性不足——因此，部分研究未纳入本综述。

作者	其他	Name	操作系统平台		Pos	Neg
Yin等人	Nahar等人	caw_kon	v	Kongregate	42	4802
Yin等人	Nahar等人	CAW_SLS	v	Slashdot	60	4303
Yin等人	Nahar等人	caw msp_	v	Myspace	65	1946
Reynolds 等人	Kontostathis 等人; Squicciarini 等人; Rosa 等人; Rosa 等人	kon_frm	v	Formspring	369	3915
Dinakar 等人		DIN_YTB	x	YouTube	2277	4500
Bayzick 等人	Zhao 和 Mao; Squicciarini 等人	湾 msp_	v	Myspace	415	1647
Xu 等人	Zhao 和 Mao	XU_TREC	v	Twitter	684	1762
Dadvar		ddv msp_	x	Myspace	311	8938
Dadvar		DDV_YTB	x	YouTube	449	4177
Bretschneider 等人		BRT_TWI	v	Twitter	220	5162
Bretschneider 等人		BRT_TW2	v	Twitter	194	2599
Van Hee 等人	Van Hee 等人	AMI_ASK	v	Ask.fm	3787	86419
Hosseinmardi 等人	Cheng 等人	hos_ins	v	Instagram	567	1387
Sui	Zhao 和 Mao	SUI_TWI	v	Twitter	2102	5219

表3。2： 网络欺凌检测数据集概览。列出了初始数据集的作者（作者）、该数据集是否被其他研究使用（其他）、参考名称（名称）、数据是否公开可用（OS，包含来源链接）、数据提取自哪个平台（平台）、报告的网络欺凌实例数（正面）和非网络欺凌实例数（负面）。请注意，实例数量是原始研究报告的数量，之后可能有所变化（例如 Twitter 和 Formspring 数据集）。

泛化能力，因此也涉及领域适应这一领域。我们将讨论其关于文本分类的既有观察，以及这些观察对（未来）网络欺凌检测研究的意义。

3.2.1 二元分类

关于将文本挖掘应用于网络欺凌检测任务，最早有据可查的建议之一出自Kontostathis 等人(2010年，他们指出Yin等人(2009)此前曾尝试对CAW 2。0 数据集中的在线骚扰进行分类。[5] Yin等人已指出，含有骚扰内容的文档与普通文档的比例极低，令人难以应对。此外，他们还预见到这项任务面临的其他几个关键问题：正例数量不足会使识别特征变得困难；而对此类数据集进行人工标注时，也可能遇到歧义和讽刺等问题，当消息脱离对话语境时，这些问题很难评估。即便数据集极为稀疏（< 1% 的正例），骚扰分类器仍能利用 tf-idf、代词、脏话和帖子相似度特征，优于随机基线。

继Yin等人（2009）以及Reynolds等人（2011年）指出，caw 2.0 数据集通常不适用于网络欺凌分类：除了缺少欺凌标签（它只提供骚扰标签）之外，其中的对话主要发生在成年人之间。他们的工作与 Bayzick 等人（2011）的工作，都是通过抓取问答网站 Formspring.me 以及 Myspace，创建用于网络欺凌分类的数据集的首次尝试。与类似研究不同，他们旨在使用文本特征，同时刻意避免使用词袋（BoW）特征。通过脏话词典和定制的严重程度标注，他们构建了多个用于衡量与这些“坏”词相关特征的度量指标。在较新的研究中，Kontostathis 等人（2013年）重新分析了KON_FRM 数据集，主要关注脏话对欺凌消息分类的贡献。通过根据脏话词典构建查询，他们表明

[5] 数据已在 caw2.barcelonamedia.org 上提供。[6] 数据已在 www.chatcoder.com/DataDownload 上提供。

不存在能够检索出全部脏话的单一组合。然而，通过将这些词捕获到一个基本维度的潜在语义索引查询向量中，并对已知的欺凌内容进行平均处理——将排名前-k （按余弦相似度排序）的内容分类为正例——他们表明，平均精度相比其基线得到了显著提升。

较新的研究包括 Bretschneider 等人（2014）的工作，他们在基于规则的模式分类器中结合了词语规范化、命名实体识别（用于检测特定人物的指代）以及多个脏话词典（Noswearing.com，2016年；广播标准局，2013年；Millwood-Hargreave，2000年），并在 Twitter 数据上取得了良好表现。[7] 我们自己的研究（Van Hee 等人，2015a）收集了一个包含 Ask.fm 帖子的大型数据集，并使用了标准词袋特征。后来，Van Hee 等人（2018年）将其扩展，加入了术语列表、主观性词典和主题模型特征。近来广受欢迎的词嵌入和神经网络技术已由 Zhao 等人（2016年）以及 Zhao 和 Mao（2017年）应用于 xu_trec、nay_msp 和 sui_twi，二者均在相应数据集上取得了最高性能。Zhang 等人（2016年）将卷积神经网络（CNN）应用于语音特征；Rosa 等人（2018b）则研究了其他内容，包括将同一架构与长短期记忆网络（LSTM）结合，用于文本特征。Rosa 等人（2018b）和 Agrawal 和 Awekar（2018）都研究了卷积长短期记忆网络（C-LSTM）（Zhou 等人，2015年），后者还采用了合成少数类过采样技术（SMOTE）。然而，正如我们将在本章中所展示的，这两项研究都存在可复现性问题。最后，Rosa 等人（2018a）在 kon_frm 上使用每个类别的 top-k 词语列表所生成的模糊化向量进行基于成员归属可能性的分类，其召回率高于以往采用的方法。

[7] 数据已发布于 www.ub-web.de/research。

3.2.2 细粒度分类

先前讨论的研究有一个共同点：关注检测单条消息中是否存在网络欺凌证据，并以每条消息为单位进行判断。正如第3.1.2节所述，然而，除了判断一条消息是否可能被解读为欺凌之外，还有更多文本线索值得推断。Xu 等人（2012）提出通过考察欺凌痕迹，将这种二元方法扩展为更细粒度的任务；也就是说，考察对欺凌事件的回应。他们基于通过关键词检索获得的（欺凌者）Twitter 数据，区分了两项任务：[8]（1）一项角色标注任务，随后使用语义角色标注来区分人物提及的角色；（2）引入情感来判定戏弄；尽管准确率很高，48% 的正例仍被错误分类。

在我们先前的研究中，我们进一步拓展了这一思路，并证明了仅凭简单的词袋和情感特征，对欺凌的类型（咒骂、诽谤、辩护、鼓励、侮辱、性相关内容、威胁）以及角色（骚扰者、旁观者协助者、旁观者保护者、受害者）进行细粒度分类是很困难的——尤其是在识别类型方面（Van Hee 等人，2015年a，b）。在Van Hee 等人(2018年)的研究中，荷兰语和英语中的这一问题都得到了进一步探讨。与亵渎词语（脏话词典）和词 n元基线相比，多特征模型（如第3.2.1节所述）在几乎所有标签上都取得了最低错误率，无论是欺凌类型分类还是角色分类。最后，Tomkins 等人(2018年)也在其社会语言学模型中运用了有关欺凌事件的细粒度知识；除了欺凌分类之外，他们还发现了潜在的文本类别和角色，并在一定程度上依赖 Twitter 上的社会互动。因此，这项研究与下一类工作相关：利用收集数据所依托网络的元数据。

[8] 数据已在 research.cs.wisc.edu/bullying/data.html 提供。

3.2.3 元数据特征

这项任务中一个值得注意、但不太常见的方面，是利用图来可视化潜在欺凌者及其关联。Nahar等人Nahar等人(2013年)首次采用了这种方法，他们将这些信息与基于潜在狄利克雷分配训练的分类器以及加权 tf·idf 特征相结合，以检测 CAW * 数据集中的欺凌者和受害者。_更具体地运用图论技术的研究来自Squicciarini 等人(2015年)，他们采用了多种特征：用于衡量受欢迎程度的网络特征（例如度中心性、接近中心性）、基于内容的特征（长度、情感、冒犯性词语、第二人称代词），并纳入了年龄、性别和评论数量。他们在KON FRM 和BAY MSP数据集上取得了最高性能。_ _

研究者Hosseinmardi 等人（2015年）以 Instagram 帖文为研究对象，并纳入从图像及其网络中提取的平台特定特征。他们率先更贴近相关文献的界定，将网络攻击（Kowalski 等人，2012年）与网络欺凌区分开来：网络攻击指单条负面帖文，而网络欺凌则具有反复发生的特征。他们还发现，语言查询与词语计数（LIWC）中的某些类别，例如死亡、外貌、宗教和性，能够很好地指示网络欺凌。虽然词袋特征表现最佳，但元数据特征（例如用户属性和图像内容）以及排名前15 条评论的文本特征，得分也相近。网络攻击似乎稍微容易分类。

3.2.4 领域适应

由于上述讨论的大部分工作都聚焦于单一语料库，一个明显的遗漏似乎是衡量这对模型

泛化能力的影响。自然语言处理（NLP）中的许多应用往往受限于成本高昂的高质量标注，而未标注数据则十分丰富。源域和目标域之间的特异性往往会损害依赖这些标注的技术在实际场景中的表现（McClosky 等，2006年；Chan 和 Ng，2006年；Vilain 等，2007年）。领域适应研究的正是受此类限制影响的任务，并旨在通过监督式（Finkel 和 Manning，2009；Daumé III 等，2010年）或无监督（Blitzer 等，2007年；Jiang 和 Zhai，2007年；Ma，2014年；Schnabel 和 Schütze，2014年）的方法克服这些限制。对于文本分类而言，情感分析可以说与网络欺凌分类最为接近（Glorot 等，2011年；Chen 等，2012年；Pan 等人，2010年）。尤其是，不平衡数据会加剧泛化问题（Li 等，2012年）。然而，尽管情感分析领域已明确识别并积极应对这些问题，但网络欺凌检测领域却并非如此，[9] 其中的具体局限性仍有待探究。我们认为，当前任务中也存在与情感分析类似的问题，下一节将对此作进一步讨论。

3.3 任务评估的重要性与假设

网络欺凌检测领域尚处于早期阶段，这一点从表3。2中可以看出。大多数数据集都相当小，且只有少数数据集经过了重复实验。鉴于改进迄今开发的方法具有重大的社会重要性，明确当前研究现状中的不足之处，应有助于创建一个稳健的

[9] 关于后者，一个很新的例外见于 Cheng 等人 (2020年a)。他们的研究提出了一种新颖的领域适应技术，并展示了该技术如何提升两项文本分类任务的表现，其中一项是网络欺凌检测。

框架，以开展未来的实验——尤其是评估分类器的（领域）泛化能力。就后者而言，据我们所知，当前的研究似乎尚未涉及。因此，这也是我们工作的主要重点。在本节中，我们提出评估这一点的三个动机，并提出相应的三个假设，以便进一步探究网络欺凌检测当前存在的局限性。

3.3.1 数据稀缺

考虑到分类目标背后的社会动态十分复杂，且训练数据的收集与标注成本高昂，数据稀缺问题主要可以从前述数据访问限制来解释：虽然在少数平台上，大多数数据无需任何内部访问权限即可获取（通常是因为用户可以选择匿名），但可以认为，实际发生的欺凌有相当一部分是在“闭门”状态下进行的。要揭示这些情况，就需要访问社交网络中的所有已知信息（例如朋友、联系人和私人消息，包括所有元数据）。这在实践中并不现实，因此研究人员只能依赖数量有限的公开可访问数据流（主要是文本）。因此，用于网络欺凌检测的大多数数据集规模较小，而且正面和负面消息的分布极不平衡（如表3。3所示）。这些小型数据集不太可能准确反映特定平台上的语言使用情况，更不用说欺凌实例中具有普适性的语言特征了。与领域适应研究的观点一致，我们因此认为，这些样本的统计效力不足，无法准确体现不同平台之间显著的语言变异，无论是正常语言使用还是欺凌特定语言使用（假设**1**）。

3.3.2 任务定义

此外，我们认为，这种数据稀缺会导致人们难以遵循网络欺凌任务的定义。以消息为单位的方法（即仅使用单条文档）是该领域大多数研究所采用的方法，但它们几乎不可能捕捉到网络欺凌（文献中的定义）背后的动态过程。被收集到的来源中的用户必须足够鲁莽，才会公开欺凌他人；而特定的（咒骂）词语用法，则可以解释为何词典方法和基于词袋的方法在以往研究中有效。因此，我们还假设正例存在偏见，只反映了欺凌的有限维度（假设 **2**）。更符合现实的情境——将重复性和权力失衡等特征纳入考量——需要考察人与人之间的互动，甚至是个人资料实例，而非单条消息；正如我们所述，这类信息通常无法获取。元数据类别中的研究（第 3。2。3节）支持这一观点：使用这类信息能够改善结果。

关于该任务定义（或操作化）的这一理论，与 Rosa 等人的观点一致。他们认为：“自*2011*年以来发表的自动网络欺凌检测领域最具代表性的研究，都进行了彼此孤立的在线攻击分类”（Rosa 等人，2019年a，第341页）。我们主要关注这一共同观点，即这种界定仅限于言语攻击；不过，我们的研究将通过实证评估这一界定与那些被界定为仅包含在线有害内容数据（即在线／网络攻击）的数据之间的重合程度，以寻找确凿证据。

3.3.3 领域影响

通过补充网络结构、交互统计数据、个人资料信息和基于时间的分析等数据，丰富以往的研究，或许能够

为分类和正确地操作化任务提供丰富的资源。不过，这些数据也具有领域特异性，因为并非所有社交媒体平台都拥有如此丰富的交互结构。此外，认为 Facebook 等社交网络（理想情况下，可获取上述所有信息来源）会一直是占主导地位的通信平台，这种假设未免过于天真。近年来，年轻群体转向了更直接的通信方式，如 WhatsApp、Snapchat，或更侧重媒体的平台，如 Instagram（Smith 和 Anderson，2018年），以及近来的 TikTok。这一转变意味着数据可获取的环境更加私密、资源也更少（导致数据稀缺程度进一步加剧）；这也意味着，要推动该领域进一步发展，就必须审慎评估当前对可用特征的使用方式，并探索全面提升跨领域泛化能力的方法。因此，本研究并未忽视文本特征；它们仍应被视为主要信息来源，同时也应特别关注此处提到的问题。我们还试图为实现这一目标做出贡献，并认为，众包收集欺凌内容有望降低领域特定语言使用的影响，获得更丰富的表示，并缓解数据稀缺（假设3）。

3.4 数据

在本研究中，我们区分了种类繁多的数据集。对于通过 AMiCA（网络空间应用自动监测）[11] 项目提供的数据集，*Ask.fm* 语料库部分可作为开源资源获取，[12] 而众包数据集 语料库将会提供——

[10]表情符号通过 https://github.com/NeelShah18/emot 进行检测。脏话则通过参考词表进行检测：英语词表取自 www.noswearing.com，荷兰语词表则由人工编制。
[11] www.amicaproject.be[12] https://osf.io/rgqw8/

	pos	负面类型		词元	平均每条消息词元数		情绪、发誓、脏话		
D_{twB}	237	5,258	12K	78K	14	($\sigma = 8$)	961	277	867
D_{frm}	1,025	11,742	21K	348K	27	($\sigma = 29$)	3,322	1,228	2,871
D_{msp}	426	1,627	13K	803K	391	($\sigma = 285$)	931	1,447	3,730
D_{ytb}	417	3,045	52K	827K	239	($\sigma = 252$)	3,662	2,606	8,705
D_{ask}	5,001	89,404	63K	1,017K	12	($\sigma = 23$)	17,362	4,839	12,191
D_{twX}	281	4,654	19K	86K	18	($\sigma = 8$)	1,344	74	502
D_{tox}	15,279	144,226	220K	12,924K	81	($\sigma = 121$)	11,876	13,732	22,404
D_{ask_nl}	8,675	70,557	58K	776K	10	($\sigma = 15$)	16,905	2,025	2,299
D_{sim_nl}	2,330	2,681	7K	55K	11	($\sigma = 16$)	434	682	194
D_{don_nl}	152	211	2K	7K	20	($\sigma = 24$)	33	47	19

表 3.3： 英语和荷兰语网络欺凌数据集的语料库统计，列出了正面（pos，欺凌）和负面（neg，其他）实例的数量、类型（不重复词语）、词元（词语总数）、每条消息的平均词元数（平均每条消息词元数）、表情符号和表情图标的数量（emote），以及每个中性实例（swean）和正面实例（sweap）中的脏话出现次数。[10]

可按需提供。所有其他来源均为从先前研究中收集的公开可用数据集[13]，如第3.2节所述。下文讨论的所有数据的语料库统计信息见表3.3。各数据集的缩写、语言（英语、荷兰语）及其简要收集特征如下。

3.4.1 *AMiCA*

Ask.fm（D_{ask}，D_{ask_nl}、英语、荷兰语）是由Van Hee 等人（2015年a）从这个以其名称命名的社交网络中收集的。Ask.fm 是一种问答式社交网络，用户通过在他人个人资料上提问（通常匿名）并回答问题来进行互动

[13]这些数据是在可能的情况下尽量完整地收集的。尤其是 Twitter，召回率较低；检索到的推文约占60%。鉴于分类问题的性质，这样的数字并不意外；正如Xu 等人（2012年）此前所示，人们往往会删除骚扰性消息。

，也可以在自己的个人资料上回答问题。因此，第三方无法直接对这些问答对作出回应。匿名性和受限的互动使网络欺凌发生的可能性很高。研究人员通过个人资料种子列表获取个人资料，并以此为起点遍历其他个人资料，收集这些个人资料中所有现存的问答对——这些内容主要为荷兰语和英语。每条消息都标注了细粒度标签（更多详情见Van Hee 等人，2015年b）；不过，在我们的实验中，这些标签被二值化，任何形式的欺凌都被标记为正类。

捐赠数据（$\mathrm{D}_{\mathtt{don_nl}}$，nl）包含来自 Skype、Facebook 和 Ask.fm 等多个平台的（荷兰语）网络欺凌实例。该数据集规模很小；不过，其中包含若干仇恨页面，这些页面汇集了大量针对某个人的网络欺凌内容，颇具价值。这些数据由曾遭受欺凌的青少年捐赠，供 AMiCA 项目使用，因此构成了特别可靠的金标准现实数据来源。

众包数据集（$\mathrm{D}_{\mathtt{sim_nl}}$，NL）源自Van den Broeck 等人（2014年）开展的一项众包实验，其中200名年龄在14至18岁之间的青少年参与了一项在隔离的 SocialEngine[14] 社交网络上进行的角色扮演实验。在该实验中，每位受访者都获得一个虚构人物的账号，并在一个六人小组中扮演四种角色之一：一名欺凌者、一名受害者、两名旁观者协助者和两名旁观者维护者。他们被要求阅读角色描述并产生共鸣，然后对一条人工生成、归于小组某位成员的初始帖子作出回应。所有人都接触到两条初始帖子，其中的网络欺凌感知严重程度分别较低或较高。

[14] www.socialengine.com

3.4.2 相关工作

Formspring（D_{frm}，EN）取自Reynolds等人（2011年a）的研究，内容由 Formspring.me 上的帖子构成。Formspring.me 是一个类似 Ask.fm 的问答平台。由于 Formspring 主要由青少年和年轻人使用，而且提供匿名互动选项，因此以充斥大量欺凌内容而臭名昭著（Binns，2013年）。这些数据通过 Mechanical Turk 进行标注，每个问答对由多数票确定一个标签。在我们的实验中，问答对被合并为一个文档实例。

MYSPACE (D_{msp}, EN）由Bayzick 等人(2011年）收集。作为一项信息检索任务，这些帖子以每十条为一批进行标注，因此单个标签适用于整批帖子（即该批是否包含网络欺凌）。每个批次都被视为一个实例，并据此进行标注。由于采用了这种批次处理方式，每个实例的平均词元数远高于其他任何语料库。

TWITTER (D_{twB}, EN）由Bretschneider 等人(2014年）从20-10-2012年至30-12-2012年期间的推文流中收集，并根据三位标注员的多数票进行标注。剔除转推后，主数据集包含220 条正面样本和5162 条负面样本，符合通常预期的4%发生率。其规模相近的测试集包含194 条正面样本和2699 条负面样本，通过在推文流中添加筛选条件，收集包含以下内容的消息：`school`、`class`、`college`或`campus`。在当前实验中，这些数据集被合并。

Twitter II (D_{twX}，英语）来自Xu 等人(2012年）聚焦于欺凌痕迹，因此通过关键词（`bully`、`bullying`）检索得到，它

如果不进行掩码处理（无论是因为用词，还是因为它混合了毒性内容和受害者内容），就会产生严重的分类偏差。不过，它确实可以用来展示检测欺凌相关主题以及（间接）欺凌报告的能力。

3·4·3 实验特定

ASK.fm 上下文 (C_{ask}, C_{ask_nl}, EN, NL) — Ask.fm 语料库是在个人资料层面收集的，但以往实验关注的是单条消息实例（Van Hee 等人，2018年）。在此，我们汇总单个个人资料的所有消息；只要该个人资料中出现一个欺凌实例，就将其标记为正例。这种汇总将网络欺凌消息检测任务转变为个人资料层面的受害者检测，从而能够更多地利用语境和个人资料层面的严重程度信息（例如反复骚扰），并使数据集更加平衡（1,763个正例和6,245个负例）。

formspring 上下文 (C_{frm}, EN) — 与 Ask.fm 语料库类似，该语料库是在个人资料层面收集的（Reynolds等人, 2011年a）。然而，该数据集仅包含 49 个个人资料，其中一些仅包含一条消息。若按完整个人资料进行分组，实例数量会非常少；因此，我们选择以五条为一批创建小型“语境”（来自同一个个人资料）。与 Ask.fm 的处理方式类似，只要其中一条消息包含欺凌，就将其标记为正例，从而使数据集更加平衡（565 个正例和756 个负例）。

毒性(D_{tox}，英语）——基于Wikimedia的Detox数据集（Thain等人，2017年；Wulczyn等人，2017年），该数据集包含超过30万条消息[15] 维基百科讨论页评论，由Crowdflower标注

[15]由 Kaggle 提供，更多信息请见此处： https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge。

毒性（包括其子类型）的标签。[16] 值得注意的是，无论是任务还是平台，都与本研究所用的其他语料库截然不同。尽管毒性与欺凌有许多共同特征，但这里关注的是针对匿名对象的单条侮辱，这些对象很可能是骚扰者不认识的人。由于数据来源是维基百科，这些侧重于文章和内容审核的评论，在主题上与人们对社交媒体的预期大不相同——两者最根本的重合之处是脏话，而这只是检测网络欺凌（而非毒性）时需要捕捉的众多维度之一。

3.4.4 预处理

所有文本均使用 spaCy 进行词元化[17] (Honnibal 等人,2020年）。表 3.3中的语料库统计未进行预处理。所有模型（第 3.5节）仅进行了转为小写和特殊字符移除；其他预处理步骤反而导致性能下降（见表 3.7）。

3.4.5 描述性分析

表 3.3 和图 3.2 都展示了显著差异；差异不仅存在于不同领域之间，更重要的是，也存在于域内训练集和测试集之间。大多数杰卡德相似系数都不超过0.20 （图 3.2），这意味着它们的词汇表有很大一部分并不重叠。这种差异不一定会影响分类；然而，它确实会妨碍模型学习负类的通用表示。这也清楚地表明，即使是更为离散的 D_{twX}（通过追踪查询收集）和 D_{tox} 也与其余语料库截然不同

[16]有关此数据集操作化的更多信息，请参阅 https://meta.wikimedia.org/wiki/Research:Detox/Data_Release 中的描述。

[17] https://spacy.io (v2.0.5)

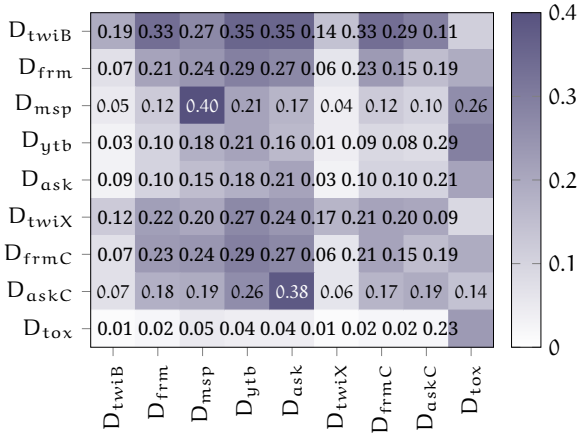


图3.2： 训练集（y 轴）与测试集（x 轴）之间的杰卡德相似度。

以及数据划分。最后，描述性统计（表3.3）进一步表明，在规模、消息长度、类别平衡和词型／词次比（即写作水平）方面存在显著差异。总之，可以认为，正面实例和负面实例中的语言使用会有显著差异，而且域内建模和域外泛化都将面临挑战。

3.5 实验设置

W我们尝试解决第3.3节中提出的以下假设 :

- 假设 **1**：由于研究人员只能依赖有限的公开欺凌数据，因此假定他们的样本数量不足以准确表征不同平台之间显著的语言变异，无论是正常语言使用还是欺凌特定语言使用。

- 假设 **2**：根据文献中的知识，单条消息不太可能充分体现欺凌的全部复杂性

事件。因此，所考察语料库中的网络欺凌实例预计会存在较大偏见，仅反映欺凌的有限层面。

- 假设 **3**：通过控制数据生成和结构，众包收集欺凌内容有望减少领域特定语言使用的影响，提供更丰富的表示，并缓解数据稀缺问题。

据此，我们提出五项主要实验。实验 I（假设 **1**）和实验 III（假设 **2**）探讨泛化能力问题；实验 II（假设 **1**）和实验 V（假设 **3**）则都将针对受限的数据收集提出解决方案。实验 IV 将复现一系列用于网络欺凌检测的最先进模型，并对其进行跨领域评估，以便与我们的基线方法进行比较。

3.5.1 实验 *I*：跨领域评估

在本实验中，我们引入跨领域评估框架，该框架将在其他所有实验中扩展应用。为此，我们首先对一个给定模型（基线模型或其他模型）进行多对多评估：该模型分别使用所有可用数据源进行训练，并将数据划分为训练集和测试集。在后续实验中，我们会在此基础上扩展一对多评估。此设置意味着：（i）我们使用某个给定语料库的训练部分拟合模型，并分别在所有可用语料库的测试部分上评估预测性能（多对多）。此外，（ii）我们将所有语料库的训练部分合并后用于拟合模型，再分别在各语料库的所有测试部分上进行评估（一对多）。总之，对于每个测试集，我们都会报告分别在各语料库上训练的“小型”模型，以及将这些语料库合并后训练的“大型”模型的结果。

PART	参数	值
BoW	范围	(1,1), (1,2), (1,3)
	级别	词语
SVM	权重 y	默认，均衡
	loss	铰链，方形铰链
	C	1e−3, 1e−2, …, 1e2, 1e3

表 3.4： 支持向量机 baseline 和 NBSVM 在超参数搜索中使用的网格。

每项实验都通过穷举网格搜索进行超参数调优，并在训练集上采用嵌套交叉验证（内层十折、外层三折），以找出给定参数的最佳组合。模型选择的所有步骤均依据外层折的评估结果进行。随后，将表现最佳的模型在完整训练集（占数据的 90%）上重新拟合，并应用于测试集（占 10%）。所有划分（包括交叉验证期间的划分）均采用分层方式，使各划分中的标签分布与整个数据集相近。[18] 下文中，本节的所有实验均可视为遵循此设置。

多对多评估框架旨在检验假设**1** (第3.3.1节），涉及语言变异与网络欺凌检测的跨领域性能。为此，我们采用一个初始 baseline 模型：Scikit-learn 的（Pedregosa 等人，2011年）线性支持向量机（支持向量机）（Cortes 和 Vapnik，1995年；Fan 等人，2008年）实现，基于二元词袋特征进行训练，并依据表3.4中的网格进行调优，网格设置参考Van Hee 等人（2018年）。鉴于该模型在以往研究中的应用，它理应是一个有力的候选模型。为确定其相较于此基线的域外性能，我们报告所有测试划分上的测试分数均值，但排除该

[18]索引（与其他随机组件类似）在所有实验中均由相同的随机种子固定。

模型所用的训练集（域内）。

因此，我们在相关工作的评估标准之外增加一项评估标准：提出一种新的网络欺凌检测方法时，该方法不仅应在大多数现有语料库中取得最佳域内表现，还应平均而言实现最高的域外性能，才能被视为稳健方法。需要指出的是，本章所选的语料库并非都具有最佳代表性。因此，我们实验中的测试应被视为改进任务评估的一项提议。

3.5.2 实验 *II*：衡量领域影响

为了克服语言使用方面的领域限制，并进一步巩固我们针对假设 **1** 的测试，我们尝试通过三种不同方式改变表示来提升基线方法的性能：i) 合并所有可用的训练集（以模拟一个规模庞大且多样化的语料库）；ii) 在用户级别聚合实例；iii) 在所有设置中使用最先进的语言表示，而非简单的词袋特征。我们将这些实验定义如下：

数量与多样性 一些用于训练的语料库相对较小，因此可以认为它们不足以代表留出数据（如测试集）。有人可能会认为，只需收集更多数据或在多个领域上进行训练，就能部分缓解这一问题。为模拟这种情形，我们将所有可用的网络欺凌相关训练划分合并（生成 D_{all}），这对应于评估框架中的一对多设置。我们希望，规模或内容相似的语料库（Twitter 数据集、Ask.fm 和 Formspring、YouTube 和 Myspace）能从更多（相关）数据中受益。此外，在全部数据上训练一个大型模型，有助于构建一种全覆盖设置，用于评估

完整任务的平均跨领域性能（即在所有可用语料库上训练时，跨所有测试集的表现）。这项特定评估将在实验 IV（复现）中用于模型比较。

上下文变化 除 MySpace 和 YouTube 外，几乎所有语料库的标注都基于短句，这一点在表 3。3 中尤为明显。这种一次性（即基于单条消息）的网络欺凌分类方法，所能利用的内容（和语境）非常有限。因此，它并不符合网络欺凌的定义——如第 3.3.2 节前文所述。作为添加（更丰富）语境的初步模拟[19]，我们将 D_{ask} 的个人资料以及（批次）D_{frm} 合并为单个语境实例（创建 C_{ask} 和 C_{frm}，见第 3.4 节）。这样，我们便能直接比较基于较大语境训练的模型与基于单条消息训练的模型，并评估语境限制对整体任务表现及跨领域表现的影响。

改进表示 词嵌入作为语言表示，已被证明能显著提升许多自然语言处理相关任务的性能（Collobert 等人，2011年）。鉴于许多语料库普遍缺乏训练数据（包括负例），基于现有数据训练的词特征（及其权重）往往无法很好地反映平台本身的语言使用情况，更不用说其他社交媒体平台了。因此，预训练表示能够提供一些特征，理论上应能在跨领域场景中取得更好的表现。针对当前任务，我们为每种语言选取了两种现成的嵌入模型：对于英语，采用平均后的200维 GloVe（Pennington 等，

[19]理想情况下，最好在设计阶段就收集个人资料层面的数据。现有语料库并非专门以这种方式收集，因此我们的设置只是对这种情形的近似。

2014年）向量（在 Twitter 上训练），[20] 以及 DistilBERT（Sanh 等人，2019年）句子嵌入（Devlin 等人，2019年a）。[21] 对于荷兰语，`fastText`采用在维基百科上训练的嵌入（Bojanowski 等人，2017年）[22] 以及`word2vec` （Mikolov 等人，2013年a、b）嵌入[23] （Tulkens 等人，2016年）在 COrpora from the Web (COW) 语料库上训练的嵌入。GloVe、 `fastText`和 `word2vec` 嵌入使用 Gensim[24] （ˇReh ˚uˇrek 和 Sojka，2010年）进行处理。

作为本节额外的基线，我们纳入了Wang andManning（2012年）提出的朴素贝叶斯支持向量机（NBSVM），该模型在文本分类任务中应能取得有竞争力的表现。[25] 该模型也曾作为与 D_{tox}相关的 Kaggle 挑战赛的基线。[26] NBSVM 将 tf-idf 加权的一元和二元语法特征作为输入，最小文档频率为3，语料库占比为90%。idf 值经过平滑处理，tf 则采用次线性缩放（$1 + \log(\text{tf})$）。随后，根据多项式朴素贝叶斯得出的对数计数比对这些特征加权。

嵌入和朴素贝叶斯表示分类器的调优均使用与表 3.4相同的网格，不过将 C 替换为[1,2,3,4,5,10,25,50,100,200,500]。最后，我们选择了逻辑回归（LR），主要是因为上述 NBSVM 实现中使用了它，同时也因为 `fastText`。此外，我们发现，使用该网格时，支持向量机在这些特征上的表现略差；而使用全连接层微调 DistilBERT（类似于例如 Sun 等人，2019年）则表现相近。嵌入

[20] https://nlp.stanford.edu/projects/glove/ (v1.2)

[21] https://github.com/huggingface/transformers (1d646ba)

[22] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md (2665eac)[23] https://github.com/clips/dutchembeddings (1e3d528)

[24] https://radimrehurek.com/gensim/index.html (v3.4)[25]这些模型的实现可在我们的代码仓库中找到。

[26] https://kaggle.com/jhoward/nb-svm-strong-linear-baseline

并未针对该任务进行微调。尽管这样做有可能提升性能，但会使其与我们的基线方法难以直接比较——我们将这一点留待实验 IV 探讨。

3.5.3 实验 *III*：攻击行为重叠

此前采用细粒度标签进行网络欺凌分类的研究（例如 Van Hee 等人，2018年）发现，网络欺凌分类器在明显的攻击行为（咒骂、性话题和威胁）案例上错误率最低。 Rosa 等人（2019年a）进一步采纳了这一观点。为实证检验假设 **2** （参见第 3.3.2 节）——该假设与现有正例中存在的偏见有关——我们借鉴了这项先前研究中的脏话基线方法。不过，我们不再依赖包含脏话的查找列表，而是通过在毒性检测上训练一个独立分类器（D_{tox}），拓展了这一思路，并考察该分类器在我们的欺凌语料库上的表现（反之亦然）。对于带有细粒度标签的语料库，我们还可以进一步检查并比较该模型所识别的欺凌类别。

我们认为，毒性检测模型与用于网络欺凌检测的模型在测试集性能上高度重叠，这有力地表明，此类模型未能捕捉网络欺凌中细微的特征。值得注意的是，这与Rosa 等人(2019年a的观点一致：当前的操作化定义与网络攻击行为（或毒性）的检测并无显著差异，因此无法捕捉真正的网络欺凌。若有足够证据，以上两点都应被视为该领域分类器开发过程中亟待改进的关键问题。

3.5.4 实验 *IV*：复现最先进水平

在本次实验中，我们纳入了两种在网络欺凌检测方面取得最先进结果的架构。作为语言任务的基准神经网络模型，我们采用了双向（Schuster 和 Paliwal，1997；Baldi 等人，1999年）长短期记忆网络（Hochreiter 和 Schmidhuber，1997；Gers 等人，2002年）（双向长短期记忆网络），并部分复现了 Agrawal 和 Awekar（2018）的架构。随后，我们尝试复现卷积神经网络（CNN）（Kim，2014年），该模型用于 Rosa 等人（2018b）和 Agrawal 和 Awekar（2018）的研究；以及卷积长短期记忆网络（C-LSTM）（Zhou 等人，2015年），该模型用于 Rosa 等人（2018b）的研究。由于 Rosa 等人未报告这些模型的关键实现细节（批大小、学习率、训练轮数），因此无法可靠地复现其工作。因此，我们将 Agrawal 和 Awekar（2018）对双向长短期记忆网络和卷积神经网络的实现作为初始设置。鉴于该工作已开源，我们在实验 I 和 II 的评估中运行了完全相同的架构（包括合成少数类过采样技术）。各架构的具体细节如下：

复现我们首先采用[27] Agrawal 和 Awekar 提出的基础实现（2018年）：随机初始化维度为50 的嵌入（论文发现，改变维度或初始化方式并无显著影响），训练10 轮，批大小为128，丢弃概率为0.25，学习率为0.01。更多架构细节可在我们的代码仓库中找到。[28]我们还运行了启用合成少数类过采样技术（SMOTE）的变体，以及一个直接使用所提供的

[27] https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs/blob/master/DNNs.ipynb

[28] https://github.com/cmry/amica/blob/master/neural.py

笔记本文件的变体。[29] 这些神经模型及后续神经模型均在 NVIDIA Titan X Pascal 上运行，使用 Keras（Chollet 等人，2015年）和 TensorFlow（Abadi 等人，2015年）作为后端。

双向长短期记忆网络 在自行实现双向长短期记忆网络时，我们仅对 Agrawal 和 Awekar （2018年年的架构做了最小改动，只对批大小 [32, 64, 128, 256], 嵌入维度 [50,100, 200, 300], 和学习率 [0进行网格搜索调优，取值为1、0。01、0。05、0。001、0。005]。我们没有像以往那样训练十轮，而是使用验证集划分（训练集的10%），并在验证损失连续三轮未下降时启动提前停止。因此，与早期实验不同，我们不对神经网络模型进行10折交叉验证，而是采用简单的2折训练测试划分，其中测试集占给定语料库的10%。同样，我们主要关注的是验证早期研究中的结论，即在这一特定设置下，参数调优并不会对性能产生显著影响。

卷积神经网络 我们采用与双向长短期记忆网络相同的实验设置。Agrawal 和 Awekar (2018年) 和Rosa 等人 (2018年b) 的实现使用大小为 3、4和 5的滤波器窗口，并在最后进行最大池化。由于使用相同的网格，因此会调整词嵌入维度并训练权重（而 Rosa 等人 (2018年b) 使用 300 维的预训练嵌入）。因此，为了直接比较性能，我们将 Agrawal 和 Awekar (2018年) 的结果用作参考。鉴于基于卷积神经网络的文本分类架构通常也会在字符级进行训练，我们还加入了一个采用这种输入的模型变体。

[29]请注意，这仅用于测试复现，因为它未经过相同的评估框架。

卷积长短期记忆网络 对于这一架构，我们采用一个开源文本分类综述实现。[30] 该实现使用大小为 [10,20,30,40,50]，64 维的长短期记忆单元滤波窗口，并设有一个最终的 128 维全连接层。有关该架构及前述架构的更多实现细节，请参阅我们的代码仓库。

3.5.5 实验 *V*：众包数据

针对假设**1** 和**2**中提出的当前可用语料库的不足，我们建议采用众包方法收集数据。在本实验中，我们将使用上述评估中表现最佳的域外分类器，重复实验 I 和实验 II，并使用三个（荷兰语）[31] 数据集：D_{ask_nl}；此前使用的 Ask.fm 语料库中的荷兰语部分，D_{sim_nl}；我们的合成众包网络欺凌数据，最后是 D_{don_nl}；一个小型捐赠网络欺凌测试集，其中包含来自各种平台的消息（关于这三者的完整概述和说明，请见第3.4节）。本实验设置中唯一显著的不同是，我们从不将 D_{don_nl} 用作训练数据。因此，我们不采用 D_{all}，而是将 Ask.fm 语料库与众包网络欺凌数据合并，构成 D_{comb} 集。

3.6 结果与讨论

现在我们将逐项介绍各实验的结果，并探讨这些结果在多大程度上支持第 3节3提出的假设。由于大多数实验都需要进行反向评估（例如，实验 III 在实验 I 的数据集上进行了测试），实验 I–III 的结果汇总于表 3.5。表 3.7 汇总了 改进表示 部分

[30] https://github.com/bicepjai/Deep-Survey-Text-Classification/[31]由于合成数据仅提供荷兰语版本，因此未重复实验 III，因为目前没有这种语言的对应毒性数据集。

训练集	T1						Avg	T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}		C_{frm}	C_{ask}	D_{tox}
D_{twB}	.417	.308	.000	.122	.298	.051	.153	.131	.158	.349
D_{frm}	.423	.454	.042	.379	.418	.041	.321	.682	.259	.465
D_{msp}	.120	.176	**.941**	.324	.168	.043	.197	.364	.185	.185
D_{ytb}	.074	.160	.375	.365	.138	.000	.183	.338	.197	.140
D_{ask}	.493	.444	.211	**.421**	**.561**	.139	.351	.389	.357	.584
D_{twX}	.049	.131	.184	.175	.077	**.508**	.205	.496	.325	.082
D_{all}	**.524**	**.473**	**.941**	.397	.553	.194	**.557**	**.780**	.570	.587
C_{frm}	.152	.253	.143	.286	.136	.126	.214	.758	.400	.372
C_{ask}	.286	.237	.359	.244	.356	.107	.310	.582	**.579**	.280
D_{tox}	.343	.373	.449	.335	.443	.149	.389	.628	.539	**.806**

表 3.5： 实验 I（T1）、II（T2）和 III（T3）的跨语料库正类 F_1 分数。模型按行在各语料库的训练比例上拟合，按列进行测试。域外平均值（平均值）不包括父训练语料库上的测试性能。整体最佳测试分数以粗体标出，最佳域外性能以灰色标出。

实验 II（在“词2向量”和“DistilBERT”下）的结果，以及基线方法的预处理效果。实验 V 的结果见表3。8。为简洁起见，后两者仅报告域内分数，并列出用于比较的模型的域外平均值，以及D_{all} 模型的 D_{tox} 平均值，见表3。8。

3。6。1 实验 *I*

查看表3.5中 T 下方的上半部分行1 代表实验 I 的结果。我们在假设1 中提出，样本不足以体现平台之间的语言变异，无论是欺凌还是正常语言使用方面都是如此。第3.4.5 节的数据分析显示，各领域在词汇表方面的重叠极少，且在多个方面存在显著差异

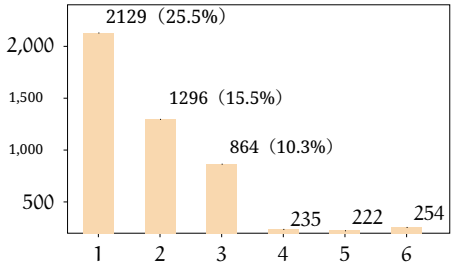


图 3.3： 绝对值最高的前 5,000 个特征系数值在测试集中的出现频率（及百分比）。

可用语料库。由此，我们对基于这些单独语料库训练的模型能否推广到其他语料库（即领域）产生了疑虑。

首先，我们考察我们的 `baseline` 在域内测试集上的表现。对于一半的语料库，它在这些特定测试集上取得了最佳表现（即模型所训练的数据中的测试集部分）。更重要的是，这意味着在另外四个测试集上，使用其他语料库训练的模型表现相当或更好。尤其是 $\mathrm{D}_{\mathrm{ask}}$ 的效果有时令人意外；例如，Dadvar 等人 (2014)（$\mathrm{D}_{\mathrm{ytb}}$）的 YouTube 语料库包含长得多的实例（见表3.3）。

不过需要注意的是，基线选自针对 Ask.fm 语料库的研究（Van Hee 等人，2018年）。这份数据是较为多样化（且规模较大）的数据集之一，其中仅包含短消息；因此，可以认为，使用这份数据训练的模型应能很好地处理较长和较短的实例。不过，这一基线（二元词特征）也可能因此特别强调浅层特征的重要性。我们将在实验 II 和 III 中进一步探讨这一点。

然而，对于实验 I，我们的目标是评估这些分类器在域外的表现，而不是最大化其性能。

为此，我们查看平均值列（表3。5）。在表格的上半部分中， D_{ask} 模型在所有域上的表现最佳（如上所述，它在其中三个域上取得了最高分）。第二好的模型是在 Formspring 数据上训练的，该数据来自 Reynolds等人（2011年a）（D_{frm}），其域与 Ask.fm 类似（问答形式，可选择匿名发布）。可以看出，几乎所有模型在 Xu 等人Xu 等人（2012年）通过查询收集的 Twitter“欺凌痕迹”语料库上表现最差。鉴于图3。2所示该语料库与其测试集的词汇重合度较低，这一结果并不令人意外。我们还证实，与Reynolds等人（2011年a）的结论一致，Bayzick等人Bayzick 等人（2011年）的 caw 数据不适合作为欺凌语料库；使用基线时会得到显著的正 F_1分数，泛化能力差，而且难以用作测试集。

此外，我们观察到，即使是表现最佳的模型，在其他领域的分数也会低于1 至2 个百分点 F_1，或比原文分数低 $15-30\%$。为了解释这一现象，我们考察重要特征在不同测试集之间的泛化情况。由于我们的 `baseline` 是线性支持向量机，因此可以直接提取所有系数为正（即与欺凌相关）的词 n 元组。图3.3 展示了系数值最高的前5,000 个特征的频率。可以看出，这些特征呈类齐夫分布：重要特征中，最常出现于某一个测试集的仅占25.5%，且随着频率增加，该比例迅速下降。反过来说，这意味着训练期间见到的前75% 的5,000 个特征不会出现在任何测试实例中，且只有3% 能在所有数据集中泛化。这一覆盖率对前60% 和4% 个特征而言分别降至约10,000，这进一步证明了欺凌相关语言使用存在显著差异。

图 3.4也表明，系数值在不同测试集之间高度不稳定，大多数的标准差约为 0.4 个标准差

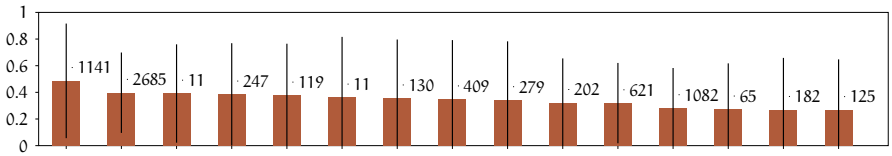


图3.4： 所有分类器（不包括使用 D_{tox} 训练的模型）中平均系数值最高的前 15 个测试集词语。误差线表示标准差。每个测试集中的每个系数值仅计数一次。词语的频率列于注释中。

。请注意，这些系数值在某些测试集上也可能变为负值，因此，对于某些特征，其取值范围从与另一类别相关，到与欺凌高度相关。根据表 3.5的结果以及图 3.3 和 3.4，我们可以得出结论，我们的baseline模型无法泛化到域外数据。根据本实验报告的定量和定性结果，这一特定设置在一定程度上支持假设**1**。

3.6.2 实验 *II*

本实验的结果主要见表 3.5 （中部和下部，尤其是 T2 ），部分结果见表 3.7 （word2vec，DistilBERT）。在本实验中，我们旨在通过采用三种方法进一步检验假设 **1** ： 合并所有网络欺凌数据以增加数据量和多样性，在上下文层面进行聚合以应对上下文变化，以及通过预训练词嵌入特征改进表示。这些方法都相当直接，可用于尝试缓解数据稀缺问题。

数量和多样性 本部分的结果列于 D_{all}表 3.5。对于接下来的所有实验，我们现在着重于

表 3.6： 根据 baselineSVM 训练D_{all}、在 D_{twB} 和 D_{ask}上测试所得的单元语法权重示例。红色词语与欺凌相关，绿色词语与中性内容相关。颜色深浅由每个特征的支持向量机系数强度决定（大多数系数接近零）。黑框表示未登录词。标签分为金标准标签（y）和预测标签（$\hat{y}$）：欺凌内容用 , 表示，中性内容用 - 表示。

y	$\hat{y}$	D_{twB}	D_{ask}
👍	👎	正要离	更大? how 多? 它的
		库 and take my *ss	gon 纳 ?
		家——	
👎	👎	you p*ss me off so	r u 到 mate
		多。	f*ck off
👎	👍	@username i will 头骨	h* 我的
		拖着	

完整的结果表（包括实验 I 的结果），并查看哪些个体分类器在所有测试集上的泛化效果最好（以灰色突出显示）。平均值列显示，我们使用所有可用语料库训练的“大型”模型[32] 在一半测试集上的性能排名第二，在另一半上排名第一。更重要的是，它的域外平均性能最高，在任何测试集上都没有竞争者。这些观察结果表明，对于 baseline 设置，不应优先选择由不同小型分类器组成的集成模型，而应选择大型模型。因此可以得出结论，收集更多数据通常似乎是有益的。

然而，对该模型所做预测结果的定性分析清楚地表明，仍存在一些局限（见表 6）。这三个随机选取的示例清楚地表明，模型关注的是显而易见的脏话——

[32]此平均值不包括来自 D_{tox} 的毒性数据；我们发现，加入这些数据后，除 D_{twX} 和 C_{frm} 外，所有领域的性能都会大幅下降。请注意，它还包括来自上下文变化 实验的范围。

露骨脏话（例如 d*ck、 p*ss 和 f*ck）。尤其是那些单独来看可能与欺凌内容相关的词语组合（leave、 touch），往往会让模型感到困惑。模型也无法识别更隐晦的威胁（skull drag）和不常见的变体（h*）。提供更多上下文或许能够缓解这两类结构性错误；这些上下文可能包含更多有毒内容，或更多中性内容示例，从而降低单个脏话的影响——这也正是下一项实验的目的。

上下文变化 关于上下文范围的访问，我们只能使用 Ask.fm 和 Formspring 数据（C_{frm} 以及 C_{ask} 表 3.5）。尽管如此，在这两种情况下，我们都看到域内性能显著提升：上下文范围的准确率得分为 .F_1 579，而 Ask.fm 上为 .561；Formspring 上分别为 .758和 .454。这一提升表明，对于这两个数据集，优先采用消息级检测更为合适。不过，另一方面，这些较长的上下文在域外数据集上的表现确实更差。这在一定程度上可以解释为，纳入更多数据（从而将数据提升至个人资料级或对话级）会使任务转变为识别霸凌对话或受害者的个人资料。尽管变异会更大，但多个单条霸凌消息同时出现在同一上下文中的可能性也会增加。因此，这将有助于学习区分以下两种情况：一是个人资料或对话中主要包含中性消息，只有一条有毒消息（因此可能无害）；二是其中包含多条有毒消息，从而使严重程度增加。

范围的变化显然会影响哪些特征被认为重要。通过人工检查，对所有 baseline 模型在各自域内测试集上的特征重要性取平均值后发现，最重要的前 500 个特征中，脏话占 63%。对于在

Ask.fm 和 Formspring 上训练的模型（具体来说是D_{ask} 和 D_{frm}），这一比例平均为 42%。值得注意的是，对于在上下文范围（C_{frm} 和 C_{ask}）上训练的模型，这一比例显著降至 11%；它们的许多重要二元语法特征都包含 you，会出现dating、 boys、 girls等话题，也会出现（are） beautiful等正面词语——后者可能表明消息来自朋友（保护者）。这种变化在一定程度上是意料之中的，因为改变范围后，任务转变为对遭受欺凌的个人资料进行分类，因此呈现出更多样的欺凌特征。

这些结果表明，将分类扩展到语境中，作为一种值得探索的平台特定设置，是可行的。不过，我们也可以得出与实验 I 相同的结论：纳入直接语境并不能克服该任务在通用领域中的局限，因此进一步支持了假设 1。一种可能的解决方案是通过更通用的语言表示来改进词袋特征，例如词嵌入所提供的表示。

改进表示 本实验旨在寻找能够改进我们 baseline 模型中所用简单词袋特征的（开箱即用）表示（即在域内取得良好表现，同时实现域外泛化）。表格3。7 列出了我们考察的两种基线方法，并在不同的预处理方法下进行了测试。随后，我们将它们与两种不同的嵌入表示进行比较。

预处理采用了多个层级：所有模型的默认设置为1）仅转为小写，然后2）要么移除特殊字符，要么3）进行词形还原，并以更恰当的方式处理特殊字符（例如，拆分 #word 以添加话题标签词元）。相应结果见表3。7 ，但并未明确指出词袋 baseline 或

Repr	T1						Avg	T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}		C_{frm}	C_{ask}	D_{tox}
baseline	.417	.454	.941	.365	.561	.508	.557	.758	.579	.806
+ 清理	.408	.477	.927	.354	.562	.517	.561	.764	.592	.807
+ 预处理	.345	.426	.929	.377	.506	.293	.512	.600	.582	.734
NBSVM	.364	.462	.929	.231	.508	.469	.542	.635	.592	.779
+ 清理	.410	.456	.940	.211	.541	.467	.563	.641	.596	.747
+ 预处理	.318	.466	.907	.320	.480	.305	.566	.532	.597	.756
word2vec	.368	.394	.860	.338	.304	.323	.366	.698	.572	.634
DistilBERT	.377	.336	.697	.296	.369	.435	.402	.598	.629	.642

表3。7： 实验 I 和 II 的不同特征表示（表征）概览。“+”部分展示了预处理方法的表现：移除所有特殊字符（清理），以及更复杂地处理社交媒体标签和表情符号（预处理）。实验 I（T1）和 II（T2）的域内正类 F_1 分数，以及针对 D_{all}的域外平均值（平均值）。基线分数见表3。5。

NBSVM。尽管后者通过全面的预处理（‘+preproc’，正向 F_1为 .566）实现了最高的域外泛化能力，但该baseline 模型在九个语料库中的五个上取得了最佳域内性能，并且采用简单清理（‘+clean’）时，域外平均值也不相上下（.566 对 .561）。

根据第 3.5.1 节提出的标准，兼具良好域内和域外表现的方法应当优先考虑。本实验对预处理方法的考察表明，相较于仅评估域内表现或只关注单个语料库，本实验采用更严格的评估标准，可能会得出不同的总体结果。反过来，鉴于简单清理在两种基线方法中表现稳定，我们在实验其余部分均采用简单清理（如第 3.4.4 节所述）。

本次实验所选用的嵌入似乎无法提供能够全面提升我们的逻辑回归模型分类性能的表示。令人意外的是，

然而，对于 Ask.fm 的对话级语料库，DistilBERT 的表现确实显著优于我们的基线（正向类别为 .629 F_1 ，高于 .579）。这或许意味着，此类表示在数据更多且更均衡的情况下会表现良好。虽然我们没有发现浅层微调 DistilBERT 能显著提升性能，但在得出明确结论之前，仍需进一步研究更深入的微调方法。

此外，鉴于我们将 word2vec 的嵌入限制为文档级的平均表示，并将 BERT 的句子表示词元作为输入（遵循常见做法），许多设置仍有待探索。尽管这两项潜在改进（即微调和采用其他输入表示）都值得在未来专注于优化的研究中进一步探索，但这超出了当前研究的范围。同样，使用在相似领域训练的嵌入来表示我们的噪声数据会更为理想；我们最终采用了完全开箱即用的嵌入，其中包含 WEB 内容和大型词汇表。

因此，我们得出结论，没有其他开箱即用的基线能够明显优于我们的词袋基线。我们先前曾提到，二元词袋表示在以往研究中的有效性，并认为这是因为它们能够捕捉明显的脏话。我们将在下一个实验中进一步测试这一点。

3。6。3 实验 *III*

在这里，我们探究假设2：所有网络欺凌语料库中的正例都存在偏见，且只反映了欺凌的有限维度。此前的实验一和实验二中，我们发现了有力证据支持这一假设。图3。4 和3。3、表3。6以及对排名靠前特征的人工分析，均表明毒性是一贯排名靠前的特征。为此补充更多实证证据，我们

使用毒性或网络攻击行为数据训练模型，再用霸凌数据对其进行测试（反之亦然），从而得出这些任务之间重叠情况的结果。本实验的结果见表3。5的下部，在 D_{tox} 和 T3。

可以看出，网络欺凌分类器（以 D_{all} 作为参考）在 D_{tox} 测试集上的性能与毒性模型（正F_1 得分分别为587 和806 ）之间存在显著差距。然而，更令人瞩目的是，反过来看，毒性分类器在域外平均值（表3中的平均值3.5）上的表现位居第二。在上下文范围（C_{frm}和 C_{ask}）中，其表现尤其接近域内性能，在其他测试集上也相对接近。

网络欺凌检测应包括对有害内容的检测，同时还应能够处理更复杂的社会现象，而这些现象可能不会出现在毒性语料库的维基百科评论中。因此，网络欺凌分类在域外性能上高于所有使用词袋特征捕捉欺凌帖子的单独模型，这一点尤其令人惊讶。只有在合并所有语料库后， D_{all} 分类器的表现才优于毒性模型。结合先前的结果，这一观察有力地表明，现有的大部分网络欺凌内容并不复杂；当前模型主要使用简单的攻击性特征，因此只能在有限程度上实现泛化，这支持了假设3。

3.6.4 实验 *IV*

到目前为止，我们尝试通过几种不同的方法改进一个基于二元特征训练的直接基线模型。虽然数据（表示）的变化似乎对性能有显著影响（增加每个实例的消息数、合并所有语料库），但采用不同特征的实验均未.....

Arch	T1						Avg		T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}	all	tox	C_{frm}	C_{ask}	D_{tox}
baseline	.417	.454	.941	.365	.561	.508	.557	.389	.758	.579	.806
NBSVM	.383	.486	.925	.387	.476	.396	.551	.385	.703	.604	.797
双向长短期记忆网络*	.171	.363	.938	.152	.504	.400	.440	.349	.609	.507	.762
双向长短期记忆网络+	.188	.396	.951	.160	.438	.341	.417	.337	.541	.505	.737
双向长短期记忆网络	.182	.341	.905	.148	.463	.246	.479	.356	.608	.522	.774
CNN*	.500	.276	.790	.133	.462	.438	.364	.350	.000	.306	.753
CNN	.444	.416	.816	.000	.498	.438	.464	.342	.000	.610	.754
卷积神经网络+	.444	.419	.816	.000	.499	.375	.460	.362	.000	.647	.774
卷积长短期记忆网络*	.000	.421	.875	.095	.000	.000	.449	.329	.094	.425	.757
卷积长短期记忆网络	.000	.019	.829	.000	.066	.000	.463	.355	.095	.518	.761
卷积长短期记忆网络⋆	.000	.057	.853	.075	.008	.000	.278	.358	.296	.506	.756

表 3.8： 不同架构（Arch）的概览：其在域内实验 I（TF_1 的正类分数（T1）和 II（T2），域外平均值 D_{all}(all)，以及 D_{tox} (tox)。基线模型（及其分数）见表 3.5。Agrawal 和 Awekar 的复现结果 (2018年)以 * 表示，其过采样方法以 +表示，字符级模型以 ⋆表示，其余均为我们调优后的模型。

表示产生了影响。通过当前实验，我们希望复现其方法，借助此前最先进的架构，并将这些模型纳入我们的评估框架进行检验。

从表3.8可以看出，我们的基线方法在几乎所有域内测试以及域外平均结果上都优于这些神经技术。在严格遵循Agrawal 和 Awekar (2018年) 的实验设置，并尽可能贴近Rosa 等人 (2018年b) 的设置后，我们可以得出结论：在更严格的评估下，有充分证据表明，这些模型在网络欺凌任务上并未取得最先进的结果。[33] 调优这些网络（至少

[33]在获得 Agrawal 和 Awekar (2018年) 的复现结果后（尤其是未能复现论文中过采样所产生的效果），我们检查了所提供的代码和笔记本文件。我们的理解是，在……之前进行过采样

在我们的设置中）似乎并不能提升性能，反而会使其下降。这表明，用于提前停止的验证集往往不能代表测试集。因此，在该验证集上进行参数调优容易受到过拟合的影响——考虑到语料库的规模，这一点并不令人意外。

关于卷积神经网络架构的表现，还可以得出一些值得注意的进一步观察：在词级（针对 $\mathrm{D}_{\mathrm{twB}}$）和字符级（针对 $\mathrm{C}_{\mathrm{ask}}$）上，该架构都取得了相当显著的提升。尤其是，对话范围（C，其类别分布相对均衡）的表现远胜于基线方法。数据量增加时也能观察到同样的效果；$\mathrm{D}_{\mathrm{all}}$ 和 $\mathrm{D}_{\mathrm{tox}}$ 的平均测试分数在几乎所有架构上都与基线相当。此外，$\mathrm{D}_{\mathrm{tox}}$分数表明，所有架构在毒性检测上的重合程度大致相同，不过有趣的是，神经网络模型的重合程度低于基线方法。

因此可以得出结论：当前的神经网络架构并不能解决这项任务的局限性，反而表现更差。不过，我们的实验再次表明，通过合并所有来源以提供更多数据，并通过对包含多条消息或对话的批次进行分类来实现平衡，从而改进语料库表示的这些方法，能够使神经网络模型的表现接近基线水平。

将数据集拆分为训练集和测试集会导致性能提升；我们测量了这两个拆分中正例的重合度，发现 测试集中没有唯一实例。此外，在重新运行笔记本文件中的实验，并在拆分后进行过采样后，这种效果显著减弱（与我们在表 3.8中的结果相似）。我们于2019年3月联系了作者并告知这些观察结果，此后作者确认了我们的结果。我们的分析见此处： https://github.com/cmry/amica/tree/master/reproduction。截至撰写本文时，该论文尚未撤稿，且已累计获得 271 次引用（截至 13日，2022年）。

训练集	T1			T2	
	D_{ask_nl}	D_{sim_nl}	D_{don_nl}	D_{askC_nl}	D_{simC_nl}
D_{ask_nl}	.598	.516	.495	.264	.533
D_{sim_nl}	.273	**.708**	**.667**	.501	.800
D_{comb}	**.608**	.681	.516	**.801**	.808
D_{askC_nl}	.165	.361	.182	.505	.750
D_{simC_nl}	.175	.424	.333	.496	.750
D_{all_nl}	.577	.677	.516	.379	**.821**

表 3.9： 荷兰语数据的实验 IV 正类 F_1 分数。模型按行在语料库的训练比例上拟合，并按列进行测试。整体最佳测试分数以粗体标出。重点关注的分数已突出显示。

我们的目标并非彻底优化这些架构，而是复现研究；尽管如此，所提出的方法仍可为后续研究提供更多方向。最后，鉴于基线模型表现稳健，我们将在下一项实验中继续使用它。

3.6.5 实验 *V*

由于实验设置的特性（如表3.2所示，该设置生成了语言使用简单的平衡数据），众包数据集很容易分类。因此，我们不报告域外平均值，因为该数据集会使平均值过于乐观，且缺乏参考价值。无论如何，我们主要关注加入众包数据或用其替代真实数据时的表现。与其他实验不同，本节主要关注 Ask.fm（D_{ask_nl}）和捐赠数据（D_{don_nl}）的分数（见表3.9）。Ask.fm 语料库荷兰语部分的分数与英语语料库的分数相当接近（.561 对比 .598 正面F_1 分数），这与早期结果一致（Van Hee 等人, 2015年b）。

此外，尤其是在数据量较小时，众包语料库在 D_{ask_nl} （.516）上的表现出人意料地好，在捐赠的测试数据（ D_{don_nl}上的 .667）上则显著更好。这表明，针对具体任务定制的平衡、受控欺凌数据集无需大量数据，就能达到相当（甚至更好）的表现，这是一个令人鼓舞的结果。

此外，在使用上下文表示的设置中，在对话范围上进行训练，起初似乎并未提升任何配置下的检测性能（除 D_{simC_nl} 上略有提升外）。然而，在测试时，这确实以有意义的方式简化了任务；在消息级别上 D_{ask_nl} ，性能略有提升（从 .598 F_1-分数提升至 .608），而合并两个数据集时，使用 D_{comb} 训练并在 D_{askC_nl} 上测试，可观察到显著的性能提升（合并数据集上的结果从 .264 和 .501 提升至 .801）。因此可以进一步得出结论：用众包数据集扩充现有训练集，能够带来有意义的改进。

根据这些结果，我们确认实验 II 的结果同样适用于荷兰语：更多样、规模更大的数据集以及不断扩大的语境范围，有助于提升任务表现。最重要的是，有充分证据支持假设**3**：众包实验生成的数据有助于提高我们真实场景测试集的检测率；将其与外部收集的数据结合，无论是否加入额外语境，都能提升性能。这些结果凸显了这种数据收集方法的潜力。

3.6.6 未来工作建议

我们希望，我们的实验有助于揭示网络欺凌检测任务在方法严谨性方面存在的多个问题，并提高人们对此的进一步关注。我们认为，相较于对（过度简化的）班级分类任务投入过多的工作，缺乏对构建丰富且具有代表性的语料库的关注，而这些语料库应能反映欺凌的实际动态，这使得人们难以对该任务的进展进行批判性评估。

因此，我们特别想强调纳入本研究评估标准的简单基线和域外测试的重要性。它们有助于更公平地比较所提出的新型分类器，并使评估方法更加统一。与此相应，系统地纳入领域自适应技术似乎是提升跨领域性能的合理下一步，尤其是针对不平衡数据量身定制的技术。

此外，还应以批判性视角审视任务的全部范围在多大程度上得到了落实。未来的研究若能明确寻找证据，证明其假设——即标注为“网络欺凌检测”的分类器不止能进行一次性的消息级毒性检测——将大有裨益。我们认为，目前大多数相关研究对这一任务的界定仍过于狭隘，尚不足以称为理论上定义明确的网络欺凌分类。本文展示了若干有助于开展此类分析的定性和定量方法。随着网络欺凌检测应用日益普及，这些方法有助于避免曲解已开展的研究，以及未来在真实场景中的应用。

我们可以设想，这些结论对计算取证领域的进一步研究具有参考价值，例如在线恋童行为的检测（Bogdanova 等人，2014年）、攻击行为与恐吓（Escalante 等人，2017年）、恐怖主义与极端主义（Kaati 等人，2015年；Ebrahimi 等人，2016年），以及系统性欺骗（Feng 等人，2012年）等。这些都是发生在更复杂网络中的严重偏斜现象，若加以简化，可能会严重歪曲研究任务。与网络欺凌研究一样，对多个领域进行严格评估，或许能够

揭示潜在的性能差距问题。

虽然我们展示了一种在确保知情同意的前提下收集可信网络欺凌数据的方法，但能够支持复杂社会互动模型的、更有价值的真实世界数据来源仍受到限制。我们预计，未来的建模将受益于构建规模大得多的（匿名化）语料库——正如大多数语言研究领域所做的那样，因此我们希望未来的研究能够朝这一方向发展。

3.7 结论

本章指出了影响当前大多数网络欺凌检测研究的若干问题。由于难以在真实环境中收集准确的网络欺凌数据，该领域面临数据稀缺的问题。在理想情况下，能够捕捉所有必要元数据、用于建模文献所定义的网络欺凌复杂社会动态的丰富表征，很可能会带来良好成效。然而，可以预见，短期内此类数据的获取仍将受到限制；而且，随着当前社交媒体转向私人交流，这些数据一开始就可能不具备普遍适用性。因此，该领域的实证实践亟需重大改变。为此，我们提出了一套跨领域评估方案，并在多种不同表示方式下测试了若干网络欺凌检测模型，以期克服现有数据的局限，并为这项任务提供一个公平严谨的框架，促进模型之间的直接比较。

此外，我们提出了三个假设，并预期能在这些评估中找到支持证据：1）语料库规模过小且具有异质性，无法准确体现不同平台上欺凌和中性内容在语言使用方面的显著差异；2）正例存在偏见，主要捕捉的是毒性，

而未涵盖欺凌的其他维度；最后，3）众包是一种生成合理网络欺凌事件的资源，有助于扩充现有数据并改进当前模型。

我们发现了支持全部三个假设的证据：以往的网络欺凌模型难以跨领域泛化；简单的词袋基线难以超越；毒性分类与网络欺凌检测之间存在大量重叠；众包数据能够产生表现良好的网络欺凌检测模型。我们认为，假设1)和2 的结果尤为重要，它们揭示了推进这一研究领域所需克服的主要障碍。此外，我们还表明，利用所有公开语料库中的训练数据，以及将表示方式调整为纳入语境，都能切实提升整体任务的性能。因此，我们认为，未来工作应将这两者都纳入评估考量。尤其是，我们还表明，这些做法无法解决当前可用语料库的既有局限，因此可以为未来聚焦于收集（更丰富）数据的研究开辟途径。最后，我们证明，此前在这项任务上展现出最先进性能的模型无法复现。我们希望，本文的观察与贡献能够帮助提升未来网络欺凌检测研究的严谨性。



本章已发表于：

Chris Emmery、Ákos Kádár、Grzegorz Chrupała 和 Walter Daele- mans。2022。网络欺凌分类器对模型无关扰动十分敏感。载于第十三届语言资源与评估会议论文集，第2976–2988页，法国马赛。欧洲语言资源协会

对文本和图表进行了少量格式调整。

4

网络欺凌分类器对模型无关扰动很敏感

A少量研究探讨了模型无关的对抗行为在有害内容分类中的作用。由于有害性分类器主要依赖词汇线索，（蓄意）创造且不断演变的语言使用方式，可能会降低当前语料库和最先进模型在部署用于内容审核时的效用。可用训练数据越少，模型可能就越容易受到影响。据我们所知，本研究首次探讨了对抗行为和数据增强对网络欺凌检测的影响。我们证明，模型无关的词汇替换会显著损害分类器性能。此外，我们还表明，将这些经过扰动的样本用于数据增强时，模型对词级扰动的鲁棒性会提高，但整体任务性能会略有下降。先前有关毒性的研究提出的数据增强方法效果较差。我们的结果强调，在语料库规模较小的在线伤害领域，有必要开展此类评估。

扰动数据、模型和代码可在https://github.com/cmry/augtox获取，以便复现。

4.1 引言

我们的线上存在让我们与（直接或间接的）社交网络联系起来更加容易，也彻底改变了我们的互动方式以及互动对象。尽管线上联系和自我披露通常有益于社会（Valkenburg 和 Peter，2007年），但缺乏身体互动也会带来不利影响：它会削弱（通常匿名的）互动中的社会责任约束，增加我们接触怀有恶意的人之机会，而且由于我们频繁使用移动设备，这种影响会更加具有侵入性（Mason，2008年）。这些因素不断累积，导致在线有害行为持续发生，而在线平台在技术、法律和伦理层面仍难以应对其规模。

在线伤害（Banko 等人，2020年对这一领域进行了全面分类）以及——尤其是在自然语言处理（NLP）领域——辱骂性语言，都是高度复杂的现象。对它们的研究涉及多个子领域（仇恨言论、有害评论、攻击性和辱骂性语言、攻击行为以及网络欺凌的检测），每个子领域都有各自的问题集合和（几乎全为英语的）语料库（Vidgen和 Derczynski，2021年）。此外，这些任务还存在许多尚未解决的问题，若干批判性研究对此有所强调（第3章，Rosa 等人，2019年a；Swamy 等人，2019年；Madukwe 等人，2020年；Nakov等人，2021年）。这些未解决的问题主要涉及毒性的语境、历史和多模态性质、数据的特异性，以及跨领域泛化能力较差。

本章聚焦于其中一个子问题：有害内容不断演变的特性。除了（年轻）用户交流所用的各种渠道和媒体之外，这种变化尤其体现在相关词汇上：俚语、

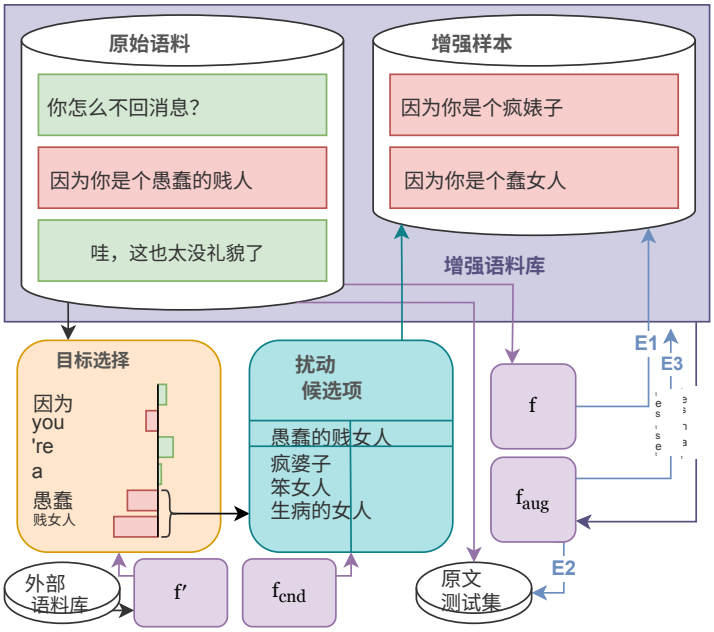


图 4。1： 通过模型无关的词汇替换增强网络欺凌内容数据的实验（E1-3）示意图。

仇恨言论或一般性侮辱（例如，*karen*、*simp*、*coofer*和*covidiot*）。鉴于最先进的有害内容分类器（Gehman 等人，2020年）高度依赖词汇线索，现有语料库必须不断扩充，模型才能维持其性能。这给任何审核有害内容的系统都带来了高昂的要求，而该领域的研究似乎仍未关注在真实环境中评估模型（Nakovet al.，2021年）。

毒性的对抗性质进一步加剧了这些问题；与任何安全应用一样，可以稳妥地假设，恶意行为者会试图（主动）规避他们所受的任何形式的内容审核。尽管已有大量研究探讨如何构建能够模仿此类行为的系统（Hosseini 等人，2017年；Ebrahimi 等人，2018年b；Li 等，

例如，2019年针对 Google 的 Perspective API 的特征攻击），有害内容检测研究却很少纳入针对对抗攻击的鲁棒性测试。更重要的是，这些攻击通常是针对现有的毒性分类器设计的，而人类对手无法直接访问执行内容审核的模型。因此，要真实地模拟人类的破坏行为，就需要采用模型无关攻击。

因此，本研究结合了以往工作的多种思路：我们将词汇替换应用于语料库规模较小的在线伤害子任务（网络欺凌检测），以探究词汇变异（自然变异或对抗性变异）如何影响模型性能，并由此评估当前最先进模型的鲁棒性。我们采用与模型无关的方式开展研究；外部分类器会指出哪些词可能适合替换。随后，我们使用多种基于 Transformer 的模型对这些词进行扰动，再评估目标分类器的预测结果发生了哪些变化（见图 4。1）。这些扰动并非为了对外部分类器或目标分类器进行对抗攻击而选取。之后，我们评估增强现有网络欺凌语料库在多大程度上能够提升分类器性能、增强其抵御词级扰动的鲁棒性，以及提高其在不同替换模型之间的可迁移性。为此，我们提供了方法和语言资源，用于测试网络欺凌分类器抵御毒性文本中词汇变异的鲁棒性。

		目标			
提示词	你是个智力障	呆子	且愚		af .
词元	你是个迟 ##钝				一个 ##混蛋。
#1	You 是个傻	baby	and	傻瓜	af .
#2	你真没用	青少年	且愚		af .
#3	你真是个病态	婊子	而且		af .
#4	你真是个疯子	dog	and	没用	af .
#5	你是个蠢	货	而且		.

表 4.1: 使用 Dropout BERT 进行词汇替换的示例。展示了初始提示词（已消音）、要替换的词语（高亮显示）、它们的词片编码（词元）以及生成样本。

4.2 词汇替换

我们采用词级或词元级扰动（即替换，示例见表 4。 1 ），这意味着，对于文档 w_t 中的给定目标词 $D = (w_0, w_1, \ldots, w_t, \ldots, w_n)$，我们使用替换 C 分类器[1]，找出一组扰动候选项 f'，以穷尽地生成新样本 D'，其中任一样本都有可能使目标分类器给出错误标签 f。然而，这些样本并非根据此类标签变化来选取，因此这并不构成对抗攻击。我们沿用并改进了第 2 章中的对抗替换框架[2]，该框架又扩展了 TextFooler（Jin 等人，2020年）[3]，加入了基于 Transformer 的扰动。

[1] 此处的“替代”并非指词语替换，而是指“替代型”分类器。具体来说，它的架构和训练数据都与我们使用的任何目标分类器不同。 f

[2] https://github.com/cmry/reap (ba8ee44)

[3] https://github.com/jind11/TextFooler

4.2.1 选择要扰动的词语

目标词语 $T(D, f')$ 根据其对文档分类的贡献进行选择和排序。这种重要性，即省略分数（Samek 等人，2017年；Kádár 等人，2017年等人的研究），通过删除给定位置 D_t的一个词来计算，该位置$D_{\setminus t}$表示为 $o_y(D) - o_y(D_{\setminus t})$，其中$o_y$ 是替代分类器的 logit 分数 f'。直观而言，这样做可以找出有毒程度很高的词语，或与欺凌相关的文本片段，并以某种方式对其进行扰动。

4.2.2 提出扰动候选项

由于我们的目标是提升词汇变异，因此着重提出同义词作为扰动候选项。Zhou 等人（2019年）通过将原词的嵌入提供给掩码位置，以此让 BERT 的掩码语言建模以给定词为条件。他们采用 Dropout（Srivastava 等人，2014年）作为替代掩码，并表明这种方法能生成一组排名靠前的k 个潜在同义词。随后，某个独立 Transformer 模型 f_{cnd}在经 Dropout 掩码处理的位置预测出的词语，就是我们的候选项 $C(T, f_{cnd})$。为了给候选项排序，他们使用上下文相似度分数：

$$\text{SIM}\left(D, D'; t\right) = \sum_{i}^{n} \alpha_{i,t} \times \Lambda\left(\mathbf{h}(D_i), \mathbf{h}\left(D'_i\right)\right) \tag{4.1}$$

其中： $\mathbf{h}(D_i)$ 是给定文档 D中某个i^{th}词元的最后四层表示的拼接 f_{cnd}， $D' = (w_0, \ldots, c_t, \ldots w_n)$是将目标文本 w_t 中的目标词替换为候选项c （位于索引 t处）后得到的扰动文档 D， Λ 是二者的余弦相似度，而$\alpha_{i、t}$ 是从 i^{th}

词元到 D中的 t^{th} 位置在所有层、所有注意力头上的平均自注意力分数。最后，我们对候选项进行清理：过滤掉单字符、复数形式和首字母大写的形式、子词以及句子级重复项。

4.2.3 处理词汇表外词语

BERT 的分词器会将未知词拆分为词片段（Wu 等人，2016年），这意味着没有单一的嵌入可用于应用 Dropout。Zhou 等人(2019年)没有提及如何处理此类情况；然而，这类情况在我们的任务中相当常见，且会造成问题（见表4.1）。因此，我们在其方法基础上扩展出一种回退方法：如果 w_t 属于词汇表外（OOV），[4] 我们就将这些词片段合并为一个，并将该位置的嵌入置零（此后它便充当掩码）。[5] 除 w_t 以外的 OOV 词语仍保留为词片段。

4.3 实验设置

我们采用并比较这种词汇替换方法，以生成新的正例。我们评估这些正例能否经受住作为对抗样本的检验，以及能否用于数据增强。

4.3.1 数据

对于我们的语料库（全部为英语），我们使用两个允许匿名发帖的问答式社交网络：*Formspring*（Reynolds等人，2011年b）和*Ask.fm*（Hee等人，2015年）。后者采用多标签标注，但已进行二值化处理（任何欺凌迹象[6] 均被标记为正例），以便与其他语料库兼容。这些语料库

[4] 请注意，f' 的词汇表可能包含 BERT 词汇表中没有的词元。[5] 替代方法（如对词元嵌入取平均值或求和）并未提供更好的表示。[6] 包括自我防御和受害者的帮助者。

	😡	🙂	TTR	平均每条消息词元数
Ask.fm	5,001	89,404	.154	12 ($\sigma = 23$)
Myspace	426	1,627	.016	391 ($\sigma = 285$)
Twitter I	237	5,258	.154	14 ($\sigma = 8$)
Twitter II	281	4,654	.221	18 ($\sigma = 8$)
YouTube	417	3,045	.063	239 ($\sigma = 252$)
Formspring	1,025	11,742	.060	27 ($\sigma = 29$)

表 4.2： 网络欺凌数据的语料库统计。列出了正面实例（愤怒表情符号、欺凌）和负面实例（咒骂表情符号）的数量、类符形符比（ttr）、每条消息的平均词元数（四舍五入，avgtok/msg）及其标准差（σ）。

的规模明显大于其余语料库，因为其平台通常由年轻人使用，并且因其欺凌内容而臭名昭著（Binns，2013年）。两个长篇平台分别是*YouTube*（Dinakar 等人，2011年）和*MySpace*（Bayzick 等人，2011年），后者的实例由十篇帖子组成。规模最小的两个语料库来自*Twitter*，均通过主题关键词收集（Xu 等人，2012年；Bretschneider 等人，2014年）。这些语料库的统计数据见表4。2。

4.3.2 数据增强模型

这些模型是使用 `transformers`（Wolf 等人，2020年）库实现的。[7] 依赖项版本可在我们的代码仓库中找到。

目标词选择器 所有实验都遵循相同的模型无关方法：通过替代分类器 f' 确定目标词（即一个不同于其他任何实验所用语料库训练语料库的独立模型）。通常，这指的是在*Formspring*上训练的、基于 tf·idf 加权向量的高斯朴素贝叶斯。我们还

7 https://huggingface.co/transformers/

进一步研究了在 Jigsaw 数据集上微调的预训练版 BERT（Hanu 等人，2020年， `unitary/toxic-bert`），将其作为基于 Transformer 的替代方案 f'（记作 +）。尽管其微调所针对的任务略有不同，我们仍假设该模型具有更优的表示能力和更大的词汇表，这可能使其在选择目标词时更为有效。

替换器 我们将自行实现的Zhou 等人(2019年的替换模型（此处为 Dropout BERT）与其他掩码词预测方法进行比较：BERT（Devlin 等人，2019年b）和 BART（Lewis 等人，2020年）。表4。6 展示了各个模型的输出示例。Dropout BERT 中嵌入维度归零的概率设为 0。2，因为我们通过实证发现，取值约为 0。3 时通常不会产生同义词。我们将词元可纳入替换的最低省略分数设为 0。005，平均每篇文档会得到1-3 个目标词。替换时不考虑先前的替换；替换会同时进行——优先选择最佳候选项——并穷尽所有可用候选项，最多生成五个样本。

对于所有 BERT 模型，我们使用预训练的 `bert-large-cased`；对于 BART，我们使用 `bart-large`。[8] 我们还报告了使用经过微调的预训练 BERT 毒性版本进行的实验（Caselli 等人，2021年，`GroNLP/hateBERT`），用于 f_{cnd} Dropout BERT（此处称为 Hate BERT；使用经过微调的替代分类器时称为 Hate BERT+）。这与使用 Dropout BERT+类似；领域专用词汇表可能会产生更好、更多样的替换。Hate BERT（+）使用的 BERT 毒性模型不同于用于+的“ f'”模型，以确保这些选择不受模型影响。

[8] 我们通过实证发现， `bert-base` 模型在寻找合适同义词方面的表现明显较差。

提示词	好吧，别再叫我杰基了，你这混蛋 [redacted]
GPT-2	废[redacted]物。你心里清楚，你这个[redacted]该死的懦[redacted]夫。 你知道自己是个邪恶的婊子 [redacted] [redacted] 只在乎让自己的名字登上 报纸。我敢说，要是你看到我的脸， 你甚至不会相信我说过“哦，是啊，我 觉得我能跟你[redacted]上床。”.....

表 4.3： 以原文实例为提示词，由 GPT-2 生成的输出。随后将生成文本（最多 70 个词元）用作增强实例。

基线方法 替换模型与两种根据相关工作改编的基线方法进行比较。已有研究表明，这两种方法都能提升毒性检测效果（Gehman 等，2020；Quteineh 等人，2020；Yoo 等人，2021年），但据我们所知，尚未用于增强网络欺凌内容。首先，我们采用常见的数据增强基线方法：简易数据增强（Wei 和 Zou，2019年，简称 EDA）。EDA 对输入文本执行以下操作中的 n 项：使用 WordNet 替换同义词（Miller，1995年）、随机字符插入、交换和删除。我们将 EDA 生成的增强样本数量设置为与其他模型相近。其次，我们采用 GPT-2 进行完全无监督的数据增强（Radford 等人，2019年，在预训练的 `gpt2-large` 中实现）。我们将每个数据集中的正例（即包含网络欺凌内容的文档）用作提示词，输入文本的最大长度为 30 个词元，生成输出的最大长度为 70 个词元，因为我们发现毒性在生成文本的开头部分更为常见（Gehman 等，2020年也有类似发现）。表 4.3 展示了输出示例，以及其最终偏离毒性内容的情况。[9]

[9] GPT-2 尤其容易在生成过多词元后转向文学内容。我们还试验了 GPT-3 的（Brown 等人，2020年） `curie` （来自 OpenAI 测试版 API `https://beta.openai.com/`），但发现其表现系统性地较低

4.3.3 分类器

我们的主要分类模型采用了近期最先进的结果（Elsafoury 等人，2021年a），并对所有基于 BERT 的模型进行微调，训练 10 轮，批大小为 32，学习率为 $2e{-}5$，这一设置参照了Devlin 等人（2019年b）的建议。因此，我们将最大序列长度设为 128，并在池化输出后插入一个线性层。对于 Transformer 实验，我们采用渐进式微调：先在原文集合上微调，再在增强后的训练集（包含原始实例）上微调；两者使用相同的配置（学习率、批大小等），但训练轮数为 2 轮。这样既有望提升性能（Yang 等人，2019年），也能提高模型稳定性。[10]

我们将 BERT 与先前经过尝试和验证的（第3章）“简单”线性基线进行比较：Scikit-learn（Pedregosa 等人，2011年）实现的线性支持向量机（Cortes 和 Vapnik，1995年；Fan 等人，2008年，支持向量机），采用二元词袋（BoW）特征，超参数范围参照Van Hee 等人（2018年）。支持向量机和 BERT 分类器均在表4.2中所有网络欺凌语料库合并后的数据集上训练，*Formspring* 除外（该语料库留作替代分类器使用 f′）——始终使用相同的 90% 划分，无论是否使用增强数据。支持向量机通过网格搜索和嵌套分层交叉验证进行调优（内层十折、外层三折，不打乱数据，采用 10% 的划分）。所有实验均采用最佳设置（1-3-词 n 元组、类别平衡、方形铰链损失，以及 $C = 0.01$）。

对于这两个模型，我们还尝试在增强实例前添加一个特殊词元（Daumé III，2007年；Caswell 等人，2019年采用了类似的方法）

在所有实验中的表现均优于 GPT-2。因此未纳入这些结果。[10]我们发现，对于所有测试的模型，在混合数据集上进行微调会使数据增强失效。

	训练集		TEST	
	😡	😶	😡	😶
合并	4,789	72,243	561	8,001
增强训练集	28,148	72,243	561	8,001
增强测试集	4,789	72,243	3,283	8,001

表 4.4： 我们实验中所使用的不同划分的数据实例数量。增强测试集用于实验 1（衡量我们样本的对抗性），增强训练集用于实验 2（数据增强）。

。<A>。根据Kumar 等人（2020年）的建议，不将该词元添加到词汇表中。这些模型在图 f 或 f_{aug} 中分别称为4。1，具体取决于它们是否使用增强数据进行训练。如果没有，则跳过 BERT 的2 轮微调。

4.3.4 评估

为了在主要分类任务上评估我们的分类器，我们使用F_1分数。通过比较常规样本与替换样本之间真正例率（TPR）的下降，来衡量替换模型对分类性能的影响（即，扰动后，先前被正确分类为正面的样本中有多少被分类为负面）。请注意，在这些实验中，f' 是相同的（即朴素贝叶斯或 BERT）；因此，替代分类器始终选择相同的目标词进行扰动。样本数量取决于模型所提出候选项的质量。

真正率的下降本身也可能表明，增强后的实例不再有毒；因此，为了评估各种数据增强模型生成的样本的语义一致性，我们计算了 METEOR (Banerjee 和 Lavie,2005年；Denkowski 和

Lavie，2011年），使用 `nltk`[11]，并计算 BERTScore(Sellam 等人，2020年）在原文句子与其对应的增强样本之间的得分。Meteor 衡量灵活的一元语法标记重叠，而 BERTScore 衡量基于上下文句子编码的基于 Transformer 的相似度。

4.3.5 实验

我们在正例 X_{pos} 上运行替换流程（如图 4.1所示），使用第 4章 f' 和 f_{cnd}中讨论的不同模型，处理给定语料库（或整个集合）中的正例。对于每种配置，都会生成增强样本 X'_{pos} （每个原始实例最多生成五个）。这些样本既可以直接分类，也可以与原始语料混合，生成增强语料库 X'，其中包含 X'_{train}和 X'_{test} 划分。采用此配置，我们开展了三项实验：

实验 1 我们通过 $f(X'_{pos})$ 衡量增强样本中的词汇变异 （进而评估其“对抗性”特征）。 F_1 分数和接近 $f(X_{pos})$ 的真正率变化意味着替换词与原词相似。我们通过 X'_{pos}的语义一致性指标确认这种意义保留。

实验2 在这里，我们采用第*4*节中讨论的3.3节中的数据增强方案进行训练；也就是说，在2 个训练轮数内对 $f(X'_{train})$进行微调。所得的增强分类器称为 f_{aug}，我们在原文 X_{test}上对其进行评估。相较于 F_1，$f(X_{test})$-分数的提升表明数据增强取得了成功。

[11] `https://www.nltk.org/_modules/nltk/translate/meteor_score.html` (v3.5)

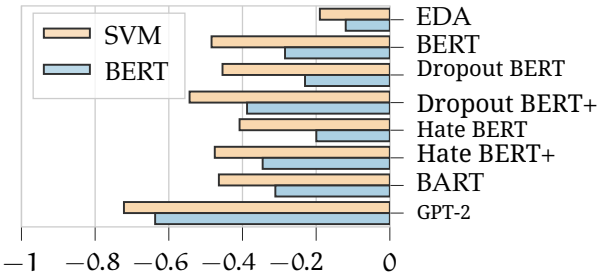


图 4.2： 应用相应替换模型后，支持向量机和 BERT 分类器的真正率下降情况（数值越低，对抗性越强）。

实验3 我们通过评估不同替换模型生成的扰动样本在 X'_{test}中的 f_{aug} 表现（即多对多评估），来衡量其对扰动的鲁棒性以及可迁移性（通过 $f_{aug}(X'_{test})$ ）。任何真正率的提升都意味着数据增强增强了模型对扰动的鲁棒性。总真正率高于 $f(X_{test})$ （无替换）并不一定会提高 F_1分数（来自实验2）。如果这种提升在多个扰动模型上都成立，则说明这些增强样本具有可迁移性。

4.4 结果与讨论

此处，我们讨论三个实验的结果（第 4.4.1 -4.4.3节）。主要结果见表 4.5 和4.7。

4.4.1 词汇变异的影响

结果见表4.5 中“样本”一行。

分类器性能 可以观察到，无监督式（提示条件化）样本（即 GPT-2）最难分类。这是意料之中的，因为生成输出并非

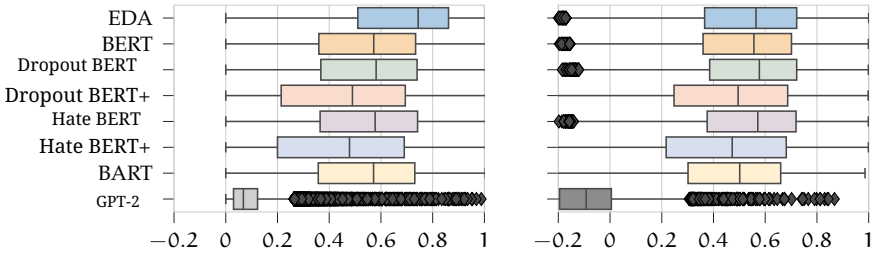


图 4.3： 语义一致性指标（越高越好）：按模型列出的 METEOR（上）和 BERTScore（下）分数——以原始文档为参考，对增强样本（仅正向样本）进行评估。

始终有毒。然而，网络欺凌分类器将大量生成句子标记为正面，这一点可以说相当令人惊讶。这证实了生成的内容（在语境中）更有害（如表4。3所示），毒性检测研究也显示了这一点，参见Gehman 等人 (2020年) 和Ousidhoum等人 (2021年）。接下来可以看到，微调后的目标选择器模型（'+'）表现出最明显的'对抗性'行为，可能是通过在对内容分类更重要的词语上提供词汇多样性来实现的。BART也会导致类似的性能下降，但往往会插入噪声（见表4。6）。Dropout模型似乎生成的样本多样性较低，但仍显示出明显的 .1 分下降，体现在 F_1 分（17。67% 平均值）。需要强调的是，这种下降是基于非定向替换得出的；也就是说，替换词并非经过挑选，以改变 f 或 f′的预测结果。

对抗样本 更多分析见图 4。2。我们观察到，每种替换模型都呈现相同的模式[12] ，与表 4。5中的情况一样；BERT分类器的真正率比支持向量机低约20%。这种差异部分可以解释为：替换模型从与我们为分类任务微调的同一模型中进行采样（下文续）

[12]为统一长度，我们在本次实验中匹配了未增强测试集实例的数量。

（bert-large-cased）。如图所示，如果并非如此（“+”模型），差异就会较小。这项实验不仅凸显了线性分类器对词汇线索的高度依赖[13]，还表明 Transformer 模型也无法免受词汇变异的影响——即使候选项是从其自身的语言模型中采样得来。这进一步印证了 Elsafoury 等人 (2021年a) 以及 Elsafoury 等人 (2021年b) 的研究（见第 4。5节）。

样本的语义一致性 在此，我们将 X'_{pos} 与X_{pos} 作为参照进行比较。Meteor 和 BERTScore 的这些样本对的结果见图4。3。总的来说，这些结果印证了实验前两个部分所呈现的趋势：语义一致性较高的模型，对分类性能的影响较小。EDA 与其他模型在 Meteor 上的得分存在明显差异。这可能是因为该指标和模型都使用了 WordNet，从而使结果偏向 EDA。GPT-2 是一个明显的离群值，因为它会生成新的数据。

语义一致性得分似乎与以往的词汇替换研究相当，有时甚至略好（见第 1章；Shetty 等人，2018年；Mathai 等人，2020年，不过这些研究都明确采用了对抗方法）。我们注意到，就样本本身而言，基于 Transformer 的模型在排名较低的候选项上，语义保持往往明显变差（见表 4.6）。对于未使用软语义约束的模型（如 Dropout），我们在前 5 个候选项中已经能发现反义词，以及普遍不合语法、语意不连贯的句子。

[13]这也带来了自身的问题；例如，参见Zhou 等人(2021年)关于偏见与去偏的研究。

$X_{\text{训练集}}$	无替换	EDA	BERT	D.out BERT	D.out BERT+	Hate BERT	Hate BERT+	BART	GPT-2
		$f(X'_{pos})$							
合并	.614	.598	.458	.491	.439	.520	.478	.436	.334
	.009	.012	.002	.005	.002	.005	.004	.007	.007
		$f_{aug}(X_{test})$							
合并	.563	.553	.538	.546	.523	.562	.535	.536	.550
	.014	.007	.014	.012	.004	.007	.009	.013	.017
Ask.fm	.621	.591	.581	.597	.574	.611	.592	.587	.601
	.011	.011	.016	.015	.003	.007	.007	.015	.009
Myspace	.436	.476	.403	.376	.351	.414	.370	.387	.496
	.093	.021	.058	.161	.040	.042	.065	.043	.037
Twitter I	.596	.630	.592	.531	.594	.617	.533	.591	.583
	.046	.016	.059	.048	.030	.053	.051	.027	.097
Twitter II	.308	.290	.297	.329	.257	.313	.268	.277	.295
	.059	.038	.043	.034	.040	.048	.027	.077	.048
YouTube	.150	.226	.180	.201	.144	.173	.152	.152	.207
	.033	.057	.010	.066	.045	.056	.009	.055	.061

表 4.5： 基于 BERT 的网络欺凌分类得分（F_1），对应实验1（采用 $f(X'_{pos})$）和2（采用 $f_{aug}(X_{test})$）。分类器在指定语料库上进行训练和测试（见第 4.3.1节），“合并”表示它们的组合。其余各列依次表示：无替换（无替换）、EDA、基于 BERT 的模型（其中“+”表示 f' 使用 BERT 而非支持向量机，而“Hate”表示 f_{cnd} 经过预训练）、BART 和 GPT-2。高亮单元格表示未进行数据增强时的性能最高，粗体表示各数据增强模型中的最高性能。报告了5 次采用不同随机种子的运行结果的标准差（小号字体）。

有趣的是，与（直观上）更好的替换候选项相比，BERTScore 给反义词赋予的分数也相近。鉴于这些观察结果，如果要更加重视模仿对抗行为，就应考虑限制增强样本的数量，而调整省略分数阈值或许也值得一试。

	#		目标			BSC
		我不明白为什么人们		like	u .	
BE	1	我不明白他们为什么		想要你	. .809	
	5	我不明白为什么	boys	hate	u . .807	
Dr BE	1	我不明白为什么大家都				
	5	我不明白为什么你		love	u . .806	
BA	1	我不明白为什么你		are	u . .700	
	5	我不明白为什么这么		有你。	.634	

表 4.6： 增强样本包括来自 BERT（BE）、Dropout BERT（Dr BE）和 BART（BA）的排名前 1 和前 5 候选项，以及以原文为参照计算的 BERTScore（bsc）。随着样本量增加，质量会下降（该指标未能充分反映这一点）。BERT 会推荐反义词，而 BART 在语义上表现不佳。

4.4.2 用于数据增强的替换

鉴于替换模型对我们的分类器性能影响显著，它们似乎很可能能够生成有效的数据增强样本。查看表 4。5 的下半部分；不过，我们可以看到，数据增强并未提升两个最大数据集（合并数据集和 Ask.fm）上的性能。Dropout BERT 似乎提升了其中一个较小数据集（Twitter II）上的性能，但总体而言，EDA 通常表现不俗，甚至比所有更“先进”的模型更有效。有趣的是，基于 Transformer 的模型似乎显著提升了 Myspace 数据集上的性能，其中 GPT-2 带来的提升最大。这可能归因于该数据集较低的类符形符比（见表4。2）。

解读 总的来说，这些方法（基线方法或基于替换的数据增强）似乎都无法带来与毒性研究中观察到的相同性能提升（Ibrahim 等人，

	INIT TPR	f_{aug}							
无替换	.537	EDA	BERT	D.out B	D.out B+	Hate B	Hate B+	BART	GPT-2
X'_{test} ↓		Δ 真正率							
EDA	.498	.270	.106	.104	.108	.101	.107	**.114**	.037
BERT	.390	–.033	.195	.144	.110	.128	.092	**.149**	.085
D.out B	.421	–.017	**.183**	.183	.143	.143	.111	.152	.086
D.out B+	.362	.015	.228	.236	.301	.177	**.238**	.191	.088
仇恨 B	.444	–.020	**.170**	.148	.104	.162	.123	.143	.073
仇恨 B+	.394	.034	.188	.174	**.238**	.215	.262	.183	.065
BART	.378	–.010	**.200**	.157	.127	.142	.115	.243	.079
GPT-2	.303	–.098	.031	.003	–.020	.003	–.025	**.048**	.597
MEAN	.399	–.114	**.158**	.138	.116	.111	.130	.140	.073

表 4.7： 各种 f_{aug} 模型在各种增强测试样本 X'_{test}上的可迁移性和鲁棒性。图中显示了 f （无替换）在初始 trp 下的原始真正率（TPR），以及 $f_{aug}(X'_{test})$ 相对于 $f(X'_{test})$的真正率变化（Δ）。正向 Δtrp 表明增强后对扰动具有鲁棒性，负向则表示相反。对同一模型扰动最具鲁棒性的分类器以粗体突出显示，次优者也予以标示。底部显示了平均真正率（按增强分类器计算），不包括 在同一模型上的性能（以反映可迁移性）。Dropout 缩写为 D.out，BERT 缩写为 B。

2018年；Jungiewicz 和 Smywinski-Pohl，2019年）。然而请注意，由于我们有意模拟对抗行为，因此采用了模型无关增强（即针对未知攻击者或噪声）。因此，虽然我们可能会在显式对抗训练方案中使用这些模型来直接提升模型性能，但这需要进行大量可迁移性评估——通常需要更大、质量更高的数据集——而且只能满足一个维度（数据）。鉴于此，我们认为，像实验1 （第 4.4.1节）中那样提升对增强数据集的分类效果，意义更为重大。

4.4.3 用于鲁棒性的数据增强

实验 3 的结果见表 4.7。我们报告了 $f(X_{test})$（在初始真正率下）的真正率变化，以及 $f_{aug}(X'_{test})$ 每种设置 f_{aug} 和 X' 分别。

鲁棒性 通常可以观察到（不出所料），当替换来自同类型的模型时，增强后的分类器的真正率提升最大。因此，“次优”的真正率提升更值得关注。可以观察到，BERT 和 BART 分别在三组数据上表现出最强的真正率提升，其次是“+”模型，分别在一组数据上表现出最强提升。与实验 1 和 2 相比，这是一个相当显著且截然不同的结果，不过它与语义一致性得分所反映的观察结果一致：更加保守的模型作为数据增强器时效果较差。因此，似乎更加多样化且与原文实例距离更远的替换，能够更好地提升对抗扰动的鲁棒性。尽管它们的输出在语义上可能不够一致，但当人们关注于提升任务鲁棒性时，这通常并不是一个相关标准。

可迁移性 所有替换模型均实现系统性的性能增益（即可迁移性），这是数据增强效用的最终指标。首先必须指出，基于 Transformer 的模型中，“相同模型”与不同模型配对之间的真正率差异（平均为 .026）小于 EDA（.156）。使用开箱即用的BERT进行词汇替换，除了具有较高的鲁棒性外，在替换后的数据集之间还实现了最高的可迁移性（平均值为 .158）。

性能权衡- 当前实验的 f_{aug} 结果应结合实验F_1的 F-score$_2$中的性能权衡来理解。根据表4.7中的信息，可以

推断出，表现最佳的模型 BERT 实际上提高了平均绝对真正率（TPR）；如果将其 .158 的平均 TPR 增幅加到 .399 的平均值（= .557），则超过了 .537 的未增强 TPR。然而，正如我们在实验2 （表4.5）中所示，这并未提升整体任务性能；反而降低了性能。对于 BERT，所有数据集上的 F_1分数略有下降（.025 –.030）。因此，这并非万灵药；部署这些数据增强模型以提高对词汇变异的鲁棒性时，应考虑此类权衡。

局限性与未来工作将所提出的模型用于数据增强的一个重大障碍是耗时。使用基于 Transformer 的模型对表4.4 (5，350 个实例）中的正例进行上采样，需要在单台 NVIDIA Titan X（Pascal）上为每个模型耗时2-3 小时。[14] 这会影响在合理时间内能够调整的参数数量，尤其是在使用该架构时（例如省略分数阈值、排序时的余弦相似度、丢弃率等；这些参数都是通过经验设定的）。对于我们这样较小的数据集而言，这种计算需求尚可接受，而且增强样本可以“离线”运行（即只运行一次）；但如果希望扩展规模，就必须考虑这些局限性。

因此，近期关于减少相关模型查询次数的研究（Chauhan 等人，2021年）对未来工作尤其具有参考价值。此外，我们在此提出的基础架构还有许多可改进之处。大多数内容已在第2章讨论；不过，也有一些新研究特别值得关注数据增强，例如利用束搜索改进替换

[14]训练 BERT 分类器最多需要 31 小时，数据增强需要 40 分钟，在测试集上进行预测需要 5 分钟。

（Zhao 等人，2021年a），而非本章采用的同步展开。更广泛地说，可以探索对抗训练（Si 等人，2021年；Pan 等人，2021年）、实现更稳健的文体计量特征（Markov 等人，2021年），或对增强模型采用基于模型的权重分配；例如，在生成模型环节中选择实例（Anaby-Tavor 等人，2020年）。与关注消息级网络欺凌内容（第3章）相比，这种做法在关注对话范围时尤其值得一试。

最后，本章专门探讨了如何扩充网络欺凌语料库——这是一项通常（尽管并非毫无争议地）被认为不止涉及有害内容的分类任务，而且相关数据通常十分稀缺。尽管这不属于本文的研究范围，但未来工作或可采用我们的方法，进一步（批判性地）阐明毒性在这项分类任务中的作用，从而有助于整理语料库或构建对比集（Gard- ner 等人，2020年；Li 等，2020年），使其更能体现网络欺凌的理论基础。

4.5 相关工作

我们的研究结合了多个规模可观的研究领域——这些领域的规模之大，甚至各自都催生了多篇综述（Fortuna 和 Nunes，2018年；Gunasekara 和 Nejadgholi，2018年；Mishra 等人，2019年；Banko 等人，2020年；Madukwe 等人，2020年；Muneer 和 Fati，2020年；Salawu 等人，2020年；Jahan 和 Oussalah，2021年；Mladenovic 等人，2021年）——因此，我们将简要概述与实验设置直接相关的研究。

对于所有任务而言，泛化问题似乎都是一个特别受欢迎的研究主题：对于网络欺凌，在第3章中，Larochelle 和 Khoury（2020年）得出结论认为，标注实践和数据集之间的重叠程度尚未形成多少共识，而结合所有数据集

似乎最能提升迁移性能。对于仇恨言论，Salminen 等人（2020年）和Fortuna 等人（2021年）得出了类似结论，表明一般形式的伤害（例如有毒、冒犯性内容）比仇恨言论等具体形式具有更好的泛化能力。最后，Nejadgholi 和 Kir- itchenko （2020年）提出了通过无监督方法解决数据整理中主题偏差的建议，从而有望改善泛化能力。我们通过针对网络欺凌的单独语料库和组合语料库开展跨领域实验，以及在毒性数据上进行预训练，借鉴了这些研究成果。

近期关于网络欺凌的研究（Reynolds 等人，2011b；Xu 等，2012年；Nitta 等人，2013年；Bretschneider 等人，2014年；Dadvar 等人，2014年；例如，Van Hee 等人，2015a 的研究均具有开创性）主要聚焦于部署基于 Transformer 的模型（Vaswani 等人，2017年）；总体而言，采用微调（例如 Swamy 等人，2019年；Paul 和 Saha，2020年；Gencoglu，2021年）或重新训练（Caselli 等人，2020年）BERT。值得注意的是，Elsafoury 等人（2021a、b）表明，尽管微调 BERT 在分类任务中达到了最先进的性能，但其注意力分数与网络欺凌特征并不相关，并预计此类模型的泛化能力会较差。在我们的实验中，我们采用类似的领域特定微调 BERT 模型，并评估其泛化能力、对扰动的敏感性，以及数据增强对改善前者的效果。

针对文本的对抗攻击（Zhang 等人，2020年C；Roth 等人，2021年，例如，提供了更全面的综述）大致可分为字符级和词级攻击。前者指故意拼错文本，或以其他符号形式替换文本（例如，*fvkyou*，*@ssh*l3*），以规避算法（Eger 等人，2019年；Kurita 等人，2019年）。Wu 等人(2018年)表明，这类针对有害内容的攻击能够被有效识别。词级攻击对人类而言相对容易理解，但要实现自动化则困难得多——因为必须保留毒性，也就是句子的语义。此前的研究

探讨了对高影响力有害性词语进行最少编辑的效果，即将其替换为无害的变体（Hosseini 等人，2017年；Brassard- Gourdeau 和 Khoury，2019年）。我们目前的工作与Tapia-Téllez 和 Escalante (2020年)的研究相似，与Guzman- Silverio 等人(2020年)的研究最为接近；后者使用 EDA 进行简单的同义词替换，并采用对抗性词元替换——通过 TextFooler 处理分类错误的实例。我们将基于 BERT 的词汇替换（Zhou 等人，2019年）扩展至模型无关扰动和数据增强。

最后，关于网络伤害的数据增强（Bayer 等人，2021年；Feng 等人，2021年等人对各种自然语言数据提供了更通用的概述），毒性研究与文本对抗攻击研究存在部分重叠；例如 Ibrahim 等人（2018年）以及 Jungiewicz 和 Smywinski-Pohl（2019年）提出的同义词替换方法，分别属于无监督或半监督方法。另一个例子见 Rosenthal 等人（2021年），他们采用民主式协同训练收集了大型有害推文语料库；Gehman 等人（2020年）则通过查询类 GPT 模型（Radford 等人，2018年）找到了能够产生有害内容的触发器。Quteineh 等人（2020年）和 Yoo 等人（2021年）也采用了完全无监督的数据增强。在我们的实验中，我们使用由多个模型组成的流水线进行词汇替换，并将其与 GPT 生成文本进行比较（Radford 等人，2019年；Brown 等人，2020年）。

4.6 结论

在本章中，我们将与模型无关、基于 Transformer 的词汇替换方法应用于网络欺凌分类任务。我们表明，这些扰动会显著降低分类器性能。将扰动后的实例作为新样本进行增强，虽然会略微牺牲任务性能，但能提升模型抵御词汇

变异的鲁棒性。未来工作应进一步研究如何利用这些模型来模拟并缓解对抗行为对内容审核的影响。



5

In讨论与结论

在本学位论文中，我们介绍了自然语言处理中的以用户为中心的安全框架。我们开展了一系列研究，展示了如何将此类以用户为中心的约束应用于两个相关领域：作者画像和网络欺凌检测。实验表明，这些约束是任务泛化的良好指标。这不仅体现在机器学习层面，也体现在以人为本的层面；也就是说，当前针对这些任务的研究是否真正有益于人类？人们能否在现实场景中使用这些模型？为实现这些目标，我们提出了两个研究问题，本章将进一步探讨这些问题，同时总结所开展的研究，讨论其局限性及未来方向。

5.1 对抗性文体计量

F对于这项任务，我们旨在回答以下研究问题：

研究问题1 端到端攻击和定向攻击在多大程度上可以用于以用户为中心的对抗性文体计量?

本问题中有关端到端混淆的开篇部分已在第 1 章中讨论，我们在那里探讨了基于自编码器的混淆器的可行性。设想的情境是，用户使用自己的写作内容训练该模型。端到端模型为此提供了一种显而易见的自动化解决方案。

5.1.1 端到端：文体不变混淆

在指出现有文献的局限后，我们强调了探索“中性”文体这一概念的重要性。[1] 我们最初认为，这似乎更忠实于混淆任务的预期目的；否则，我们不就是在模仿另一种写作风格吗？此外，正如本章所述，这似乎是大多数已提出的端到端模型的一项重要局限：尽量降低目标模型性能或许只需做少量改动，但这些改动很可能与某个<label id='2'>不同</label>的标签高度相关。这不仅会助长语义不一致扰动，也不是以用户为中心的解决方案。可以想象，混淆模型的用户不会想要进行文体<label id='4'>迁移</label>，因为与相反的文体联系在一起或许并不理想（例如，成年人使用青少年俚语写作、共和党人谈论全民基本收入等）。

生成不变的语言（这是计算机视觉中更常见的术语，例如，Nam 和 Kim，2018年），或者生成不受不同文体影响的中性语言，事后看来似乎多少有些雄心勃勃。不过，理想情况下，我们希望生成相对于目标模型，或其某种

[1] 相对于模型而言是中性的。虽然在风格迁移研究中，这一概念被描述为“正式”写作（例如，参见Madaan 等人，2020年），但真正中性的写作风格并非一个具体可感的概念。

近似模型的语言。这一区别尤为重要——尽管在实践中，由于用户特定的资源需求，这种情况不太可能发生——如果目标模型一方要训练一个模型来显式恢复这些属性，该模型很可能仍然能够恢复文体。[2]

在第2章的端到端研究中，我们希望通过几种比较来实现这些想法：i）通过风格迁移进行混淆的理想情况是什么（即，不切实际地假设我们可以访问平行句子）；ii）风格不变编码的可行性如何；以及 iii）能否通过采用不同的架构摆脱对平行数据的依赖？[3] 我们比较了序列到序列（即平行）模型和自编码器，并考察了添加用于实现风格不变性的梯度反转层的效果（通过隐式风格条件控制的效果最为显著）。为此，我们正式定义了基于迁移的混淆和基于不变性的混淆的优化目标：目标模型性能接近随机水平意味着实现了风格不变性，而性能接近零则意味着实现了风格迁移。

严格来说，我们取得了令人满意的结果（即接近随机表现，同时仍获得较高的 BLEU 和 METEOR 分数）。人工评估者对我们模型输出的评价也高于一个能够使用平行数据的风格迁移模型。然而，这项初步研究的局限性为这些结果提供了一些审慎的背景。总体而言，如果能更全面地评估条件自编码器并为其提供更充分的论证，实验设计本可以得到改进。在这项工作中，我们选择了 BLEU 和 METEOR；然而，由于我们将输入句子用作参考句子，自编码器因此获得了不公平的保留优势。

[2] 正如后来的相关研究所示，例如Gonen和Goldberg（2019年）尝试使用梯度反转对模型进行去偏。[3] 后来，John 等人（2019年）也在风格迁移研究中探讨了摆脱对平行数据依赖的类似尝试。

即使是 WMD（它使用嵌入空间），也可能倾向于选择与（神经）模型在“语义上接近”的句子，因而偏向保守型模型。对于自动评估，结合源文本和目标文本计算的 BLEU 和 METEOR 分数颇有参考价值，但只有在使用平行数据时才能采用。更公平的做法或许是使用语言模型（LM），根据困惑度为输出打分，尤其针对词语变化进行评估。

通过手动检查输入，我们还观察到所有模型都出现了明显的语法变更。我们在第o章中提到了端到端模型的这一缺陷；由于无法控制输出，我们的模型经常会用无关的变体替换词语。这表明模型似乎无法从头学习鲁棒语义。回过头来看，使用预训练嵌入或许可以部分避免这一问题。[4] 最后，尽管人工评估提供的信息多得多，且分数与以往研究的结果一致，但人工评估过程十分繁琐，而配对评估也使我们无法测试变化是否不显眼。

5.1.2 定向：词汇替换

虽然我们的研究建立在第 1 章的构想之上，但第 2 章的工作主要致力于解决初步研究在实验和以用户为中心方面的局限。因此，我们沿用了相同的风格中立框架，但重点放在定向混淆、模型的可迁移性，以及严格遵循训练和部署的现实标准上。我们通过若干基于 Transformer 的方法改进了 Jin 等人 (2020年) 提出的对抗框架，以寻找语义相似的扰动，并证明这些攻击可以完全迁移（即性能处于随机水平的范围内）。这意味着用户可以训练自己的

[4] 这并不令人意外；《圣经》是一种非常受限的语言来源。

架构，使用自己的 数据 （统称为替代模型），并认为它能够有效混淆未知的目标模型。此外，这项研究中一个尤其令人振奋的结果是，可迁移性最高的模型使用了一个通过远程方式收集的数据集 Emmery 等人 (2017年)。因此，即使数据收集是在以用户为中心的限制条件下进行的，目标模型使用的仍是来自另一领域的黄金标准语料库。

这些结果确实令人鼓舞，但仍存在一些局限，值得在未来的研究中加以解决。首先，我们只考察了一个属性（二元性别）。正如我们所论证的，可以假设这些技术能够迁移到词级模型，因为相关工作中的这些架构通常相似。然而，这并未考虑不同的文体计量特征，也没有顾及未来迁移模型可能带来的潜在复杂性。一个显而易见的补救方法是“全面增加”：考察不同的属性、各种领域（尤其是 Reddit）以及各种架构。然而，这些都受到另一项局限的制约：通过 BERT 进行前向传播相当缓慢，因此，针对我们的替代模型定制攻击，可能会使如此复杂的设置在计算资源有限时难以实现。这种复杂性对用户也没有帮助；不过，随着蒸馏版 Transformer 变体的发展（例如，Sanh 等人，2019年），未来情况或许会有所改变。

在这项工作中，我们选择提出一种替代性的人类评估方法；主要是为了减少标注者工作量，同时也因为这种格式将来更便于集成到线上标注应用程序中。这种方法的一个局限是，它无法捕捉替换内容的质量；它假设完全不引人注意也意味着语法完全正确。虽然这是对例如 Potthast 等人(2016年) 和 Hagen 等人 (2017年) 提出的度量指标的忠实简化，但从用户角度来看用处较小；

也就是说，它对预期含义只字未提，只说明它是否是一个合理的序列。这是另一个说明加入语言模型困惑度有助于评估的例子。最后，我们没有明确纳入近期关于检测混淆器的研究（例如，Mahmood 等人，2020年；Mozes 等人，2021年b；Dong等人，2021年）；不过，我们确实测量了人类可检测性。

5.1.3 未来工作

对于本学位论文所提出的研究方向，尤其是以下内容，仍有一些值得探索且可能富有成效的途径。第1章中讨论的不变编码尽管存在局限性，但仍值得进一步研究。特别是，鉴于预训练Transformer模型近期的发展，这种架构所需的基于领域的微调将大幅减少。此外，遵循Gonen和Goldberg (2019年)的思路，针对分类器集成来训练不变混淆器集成（即提升混淆器）或许会很有意思。近期，Haroon等人(2021年)采用了混淆器集成，这项工作或可为其提供有价值的补充。

对于第 2章，除了上一节讨论的改进之外，现有架构中还有许多值得进一步扩展的组件。目前，替换候选项的采样方式有些简单；每个替换候选项都单独针对目标模型进行评估，然后选取最能降低分类概率的候选项作为替换。这种做法没有考虑其他候选项或它们的组合。一项有趣的改进方向是优化攻击的采样方法。相关工作已经对此进行了一些探索，例如（改进的）束搜索和组合优化（Yoo 等人，2020年；Liu等人，2022；Zhao 等人，2021a）。

替换的范围以及任务的界定，可能是

一个需要从更高层面探讨的问题。我们目前刻意聚焦于表层（内容）词，因为它们的直接影响最为显著；然而，人们可能会质疑，这是否真是一种稳健的对抗攻击生成方法。我们只是在寻找模型的未登录词，还是确实希望改变文体？目前可行的替代方案包括拼写变化 Belinkov 和 Bisk （2018年），不过事实证明它们并不稳健（ Keller 等人，2021年）。对抗触发器（例如参见Wallace 等人，2019年；Song 等人，2021年）仍是一种有效且非侵入式的替代方案，但遗憾的是，它们往往针对某个特定模型量身定制，因此不太可能迁移，也就不符合以用户为中心的混淆方法。端到端模型与定向模型之间的折中方法或许能解决这一问题；具体而言，可以通过选取更大的词语跨度，给予攻击创意更大的发挥空间，同时又加以限制，避免对输入内容造成过多改变。

不过，这些建议大多都需要一项关键的未来工作：找到有效的方法，将各自存在缺陷、用于此任务的评估指标整合起来。经典的机器翻译和自然语言生成指标难以解释变化；理想情况下，这些变化应当很少，因此替换应当接近完美。此外，值得研究这些指标是否与针对该任务的人工判断相关，从而为其提供一些有力支持。目前，根据我们在第2章中的观察，混淆性能与自动评估似乎和人工判断并不一致。其他基于模型的语义评分器，例如 Bleurt (Sellam 等人，2020年) 和 BERTScore (Zhang 等人，2020 年 b) 也引发了类似的疑虑；如果从同一模型中抽取同义词（或完整的语义表示），显然它们的得分会相当高，

无论是在句子层面还是世界层面。[5]我们没有尝试的指标或许也存在同样的问题：TERP (Snover 等人，2009)、PINC (Chen 和 Dolan，2011年)、基于语言模型的困惑度——对于受控扰动，这些指标会给出较高的句子层面分数，造成模型表现良好的假象，并偏向于更保守的模型。

综合来看，对抗性文体计量未来最有价值的研究方向，是在混淆性能与可检测性之间找到适当的平衡。对此进展不大，近期论文（例如，Xu 等，2020年）似乎重新提出了开创性研究（Potthast 等人，2016年）中的观点。我们认为，未来有两个方向：正如我们之前提出的，可以采用统一度量指标（尽管它并非没有局限性），例如根据扰动数量对变化赋予反向权重，并纳入相应的混淆能力。不过，我们也认为，很难不将追求此类度量指标视为只是在推动竞争性基准测试。另一种做法是，改进人类评估这项任务的方法，这似乎更为妥当。这应部分借鉴现有研究及相关领域的成果（例如，Potthast 等人，2018年；van der Lee 等人，2019年；Ippolito 等人，2020年；Howcroft 等人，2020年），但更重要的是，要考虑到这项任务的独特之处。特别是在使用多个目标分类器、混淆器和去混淆器时，交互式的人在回路评估既能产生高质量的评估结果，也能为改进现有系统提供直接反馈信号（例如，Mozes等人，2021年a）。人类甚至可以通过交互式方式生成一组“完美”的编辑操作，从而提供黄金标准质量度量，同时为典型评估分数设定上限。

[5] 我们在第4章中进一步证实了这一点。

5.2 网络欺凌检测

针对这项任务，我们提出以下研究问题：

研究问题2 网络欺凌分类器是否足够鲁棒，能够以用户为中心的方式部署？

对这个问题作出回答的最大贡献来自第3章，我们在该章开展了一系列详尽的实验，评估当前基线方法和最先进的网络欺凌检测分类器的泛化能力、对任务的表征效果，以及如何应对当前的局限性。由于这项工作本身是一篇讨论性文章，我们将主要阐述研究发现与我们提出的以用户为中心的框架之间的联系。

5.2.1 任务鲁棒性

以用户为中心的方式执行这项任务，核心在于泛化。在第3章的研究中，我们着重探究网络欺凌任务的界定方式如何影响模型在域外场景中的表现。具体而言，我们假设，当前机器学习研究对任务的界定未能体现其本质，原因在于缺乏理论和实验严谨性。[6] 自我们发布初版预印本以来，若干研究进一步指出了更广泛的网络伤害研究中存在的问题（例如，参见Kim 等人，2021年）：这些研究均指出了任务粒度、‘玩具’数据集以及模型适用性有限等问题。因此，泛化仍是网络欺凌检测研究中普遍存在的问题；有限的可用数据与任务的复杂性形成了鲜明对比。许多研究在局限性部分确实提及了这一点，但目前几乎没有工作致力于解决这些问题。我们认为，其核心

[6] 见第117页。

局限在于，网络欺凌检测被界定为一项以行业为中心的任务（提供平台级分类），却无法获取行业规模的数据。另一种界定方式——以用户为中心的欺凌检测——是一个值得探索的方向，但迄今为止基本尚未得到研究。

在第 3 章中，我们不得不承认，我们才刚刚触及表面。我们主要表明，正如第 0 章所论述的那样，如果用户拥有多个通信平台，现有系统就会失效，因为它们无法跨领域泛化。我们确实考察了数据范围的限制，包括单条消息（这是此任务最常见的格式）和对话范围（例如完整的个人资料），但这些方法都假设能够获取许多不同用户的数据。最不符合我们以用户为中心的框架的，是个人资料层面的分类；这种做法实质上将任务变成了预测某个人是否正在遭受欺凌，而不是将（若干）消息分类为欺凌内容。然而，就分类而言，这种做法益处显著：随着数据集变得更加均衡，语境也更加丰富，模型性能得到了显著提升。若要部署这类模型，就需要通过集中式系统，并整合向第三方举报的功能，才能发挥作用——而这恰恰是我们试图摆脱的做法。我们通过模拟欺凌事件来收集更多数据的努力也有其他局限；这类做法虽然并非过度集中化，但可以说成本高昂，而且很可能只能提供静态数据来源。前一项局限或许可以通过集体努力来解决，例如构建规模大得多的语料库，采用匿名化或众包方式。后一项局限则意味着需要频繁开展这类工作，因为有害语言可能会随时间变化。

5.2.2 词汇增强

这使我们看到我们自身以及相当一部分现有研究的另一项局限：这些分类器对词汇线索过于敏感。我们依据第 3 章的假设 1 考察了词汇多样性；相关实验表明，消息层面分类尤其依赖词汇线索，特别是直白的脏话。这对现有研究有重要影响，因为创造性、跨文化特征（Mateo和Yus，2013年）以及有害语言使用中的偏见（Sap 等人，2019年；Wiegand等人，2019年）都会降低当前可用语料库的实用性。

因此，在第4章中，我们的目标有两个：首先，展示这种对词汇线索的关注如何渗透到（甚至）基于大语言模型的最先进水平分类器中。其次，我们想了解用于生成创意语言的类似技术，是否也能用于小型（或更小规模）资源的数据增强。后者与我们的以用户为中心的框架尤为相关。这意味着我们可以生成本地训练数据，并提高现有数据源应对词汇多样性的鲁棒性。通过这一研究，我们完成了本学位论文所探讨的工作：创意语言使用可能源自自然产生，也可能源自对抗行为。因此，以前我们关注的是模拟检测模型并以其为目标；而在这里，我们模拟的是提供对抗性输入的用户。不过有一点重要区别：我们在第2章中使用的架构会针对目标模型定制替换，而在这里则没有这样做。通过直接接受替换候选项，该方法变成了半监督方法。这与我们的以用户为中心的框架相契合，因为其核心设置只需要一个小型标注数据集（可以是用户提供的历史数据，或类似于我们的设置：可免费获取的外部语料库）和一个简单的分类器。

在本章中，我们已经谈到了这项工作的核心局限：时间，因此也包括可扩展性。我们开展的各项实验中，所有参数都预先根据经验设定；这些实验在一台（昂贵的）图形处理器机器上运行需要数天，而单个数据增强模型最多需要一天。如果我们想开展面向用户的分类，所提出的词汇替换模型似乎并非切实可行的选择。幸运的是，开箱即用的BERT似乎能提供鲁棒性和可迁移性最强的增强样本。这意味着，与第2章中提出的扩展类似，更轻量级的模型（Sanh 等人，2021年，是另一个非常近期的例子）是一个有前景的方向。尤其是，这意味着可以采用无需自定义嵌入初始化的模型（Dropout 技术需要这种初始化），从而带来更多可能性。与此一致，我们还观察到，噪声更多、词汇上不太恰当的替换会产生质量更高的增强样本。因此，值得针对这一特定任务探索的一点——由于实验的计算成本，我们没有进行这项探索——是使用更广泛的数据增强模型来测试我们的半监督框架（例如，Morris 等人，2020年）。由于我们的增强样本未能提升分类任务的性能，我们着重对模型进行了深入分析，而没有开展跨平台和跨领域（例如，这些技术向毒性检测和仇恨言论领域的可迁移性）实验。不过，未来的研究方向应在任务性能与鲁棒性之间取得更好的平衡。

5.2.3 未来工作

自 AMiCA 项目启动语境下网络欺凌检测研究以来，这一领域以及更广泛的毒性检测研究都已成为非常热门的研究主题。在此，我们

归纳了当前的若干趋势（语境、公平性、数据增强和语言模型提示），并简要探讨相关的近期研究成果，以及这些方向如何推动本学位论文所呈现的工作。

正如我们在第3章中讨论并展示的那样，对话语境可能会改变标注员的标注结果和分类结果。 Car- ton 等人 (2020年) 将这两个观点结合起来，把模型语境解释纳入标注员的任务中，部分目的是减轻标注负担。他们发现，在标注员中，假阴性增加，而假阳性减少，因此认为这些观察结果在效用方面存在歧义。 Pavlopoulos 等人 (2020年) 在标注中也发现了字面语境带来的同样歧义； 5% 的实例在不显示语境时会被标注为不同的标签。相反，他们发现，以语境为条件并不会提升模型性能。尽管应用任务略有不同，这与我们在第3章中的发现形成了有趣的对比。其他研究对语境的用途提供了不同见解：例如，Radfar 等人 (2020年) 表明，社交语境（而非对话语境）十分重要，因为彼此没有关联的社交媒体用户参与有毒对话的可能性是其他用户的三倍。此外， Xenos 等人 (2021年) 发现，知识蒸馏增强系统能够对毒性进行语境变化分类，并提出这或许可用于提升毒性检测（尤其是内容审核）。后两种语境用途尤其可以独立部署，无需依赖更大的社交媒体环境，因此值得纳入以用户为中心的有害内容检测方法中。

就此而言，我们所采用的标签表示一直相当有限。虽然可以想象，不同用户所面对的毒性程度并不相同（例如，边缘化群体可能面对纯粹针对身份的仇恨），但纳入更广泛的多样性可能

会带来益处；此前已有研究指出，对有毒和“脏”内容进行二元分类可能会再现社会中有害的不平等（Thylstrup 和 Waseem，2020年）。通常提出的算法层面的缓解尝试，是纳入公平性约束（Adragna 等人，2020年；Gencoglu，2021年）；也就是说，主动使毒性分类不受种族、性别和主题语境等因素影响，这似乎还能提升分类器性能。尽管这类方法在更广泛的语境下十分重要，但它们需要大量用户信息（至少用于训练），相应地也需要更大的社会系统来运行。因此，能够为分类提供更有依据的细粒度方法的数据源可能更为合适。Kiritchenko 和 Nejadgholi （2020年）例如提出了更好的群体表征和主题分类体系，并且只有在应用标注方案、对严重程度进行恰当的语境化之后，才进行相关处理。然而，这种方法在多大程度上能够带来下游准确率，仍有待观察。一些初步研究表明，集成模型在此类设置下能够发挥良好作用（Burtenshaw 和 Kestemont，2021年），而 Zhao 等人 （2021年b）还提供了处理小型数据集的进一步建议。

除第4章的结果外，数据增强似乎也是一个富有成效（且受欢迎）的方向。最近，Han 和 Tsvetkov (2020年) 尝试采用师生架构增强他们的（毒性）分类器，并主动挖掘更隐晦的毒性形式，例如厌女、种族主义笑话、对跨性别者的仇恨等。我们认为，数据集多样性及其如何传递到模型中，在这里发挥着相当重要的作用。除了Gehman 等人 (2020年) 展示了有毒语言模型提示生成行为，并发现CommonCrawl数据集中保留着仇恨性子版块之外，也能在社交媒体文本以外找到相关案例；例如，Birhane 等人 (2021年) 发现了大量有害内容和暴力色情材料（图像

说明）数据集。因此，这些模型对这类内容的“规范化”或许能解释为何查询有害内容（类似于我们在研究中加以利用的做法）相当容易——Welbl 等人(2021年)对此已有部分展示。尽管大型语言模型的这种行为在实际部署中会造成危害，但仍值得进一步研究无监督算法（尤其是 GPT-3）在未来能否用于生成和识别有毒数据实例。特别是，为了更好地理解（我们不会夸大到声称要改进）其训练所用语言资源中的有害内容，并探究能否利用这些模型减轻个人用户的计算负担，以及减少他们对提供内容审核的外部各方的依赖。

5.3 结论

本学位论文提出并阐释了一个以用户为中心的自然语言处理安全研究框架。为探究该框架的应用，我们聚焦于两项任务：侧重隐私的对抗性文体计量，以及侧重安全的网络欺凌检测。在这两项任务中，我们探究了若干理想要求的可行性：在用户可能涉足的不同领域之间实现泛化；构建个人用户可用的个体分类器；这些分类器开箱即用，且运行时间合理。对于对抗性文体计量，我们研究了端到端且完全模型无关的定向模型。事实证明，词汇替换非常适合以用户为中心的框架，既能有效混淆作者敏感属性，又具有很强的可迁移性。对于网络欺凌检测，我们研究了衡量泛化能力的方法，以及提高分类器抵御词汇变异的鲁棒性的方法。在这里，我们的模型无关设置也显示出，针对潜在对抗性输入的防御措施具有很强的可迁移性。我们相信，该框架所提出的约束

不仅已证明能够在自然语言数据分类任务中带来更贴近现实的评估，更重要的是，它为个人用户提供了一条途径，使其能够重新掌控对自身数据用于机器学习模型及其效用的同意权。本学位论文中呈现的所有研究均以开源、开放获取的方式提供，并附有公开可用的数据集及文档，说明如何复现和重用所呈现的研究成果。



致谢

如果你正在读这篇文字，那就说明你至少愿意翻阅这本精心编排的论文集；对此我很感激。为了完成它，我花了很长时间；不仅要撰写论文，还要排版文档。从提交初稿到答辩，我大约有一年的时间，因此那阵子我着实全力以赴地折腾了一番 LaTeX。说实话，这大概是整个过程中唯一纯粹有趣的部分。希望这一点没有被忽略。

话虽如此，我一如既往地把这一节拖到了付印前的最后几天才动笔。七年来，我一直在其他事务（先是软件开发，后来是教学）之间挤时间，努力推进同一个项目；要感谢一路同行的每一个人，确实不那么容易。不过，我会尽力而为。

如果要我指出一个点燃我学术生涯核心主题的具体时刻，那要从我学业第二年左右，在父亲家里的一次咖啡闲谈说起。尽管据我所知，家里似乎没什么学术背景，但（技术）怀疑主义却相当盛行。父亲曾以嵌入式软件工程师的身份从事底层工作的经历（例如自动取款机、军事通信），再加上他似乎随着年龄增长而日益加深的偏执，常常让我们展开激烈的讨论。这次谈话中，二十出头的我轻描淡写地看待他关于政府数据收集行动的种种假设。不到一年后，埃因霍温同乡 Peter Vlemmix 的《全景监狱》上映，爱德华·斯诺登也让世人一窥大规模监控活动的真相。这着实让我猝不及防，猛然惊醒。

后来，我很快就发现机器学习（ML）是一个令人着迷的研究领域，但博士学位才读没多久，我新养成的怀疑主义便也开始在这个领域显现。许多事情都与我的个人价值观格格不入。当然，部分原因是竞争激烈的文化，但更主要的是，诸多实例都暴露出人们对实验严谨性的惊人漠视；尤其令人担忧的是，大型企业为了牟利，无视基本伦理标准，推动学术炒作浪潮的势头愈演愈烈。因此，我（颇具讽刺意味地）发现自己身处一个积极参与监控并影响我们日常生活的领域（诚然，往往是间接参与）。幸运的是，机器学习相当擅长自我破坏，于是我便有了动力撰写这篇你正在阅读的学位论文。

所以没错，摆在你面前的这项研究，源于怀疑、愤世嫉俗，或许甚至还带着一些潜在的行动主义倾向。尽管我和父亲的关系远谈不上好，但这是我们之间为数不多的共同点之一（他在我撰写学位论文的最后一年去世了），我想他会喜欢读这篇论文。不过大体而言，我希望有朝一日它能成为一摞过时的纸张。一本小册子，我儿子 Owen 会在喝咖啡时翻看；那时，他也许会有着和我当年相似的二十来岁的心态。但对他来说，这不该成为一段幻灭历程的开端，而应是一则教训：别把宝贵的时间浪费在对着云朵大喊上。

这篇学位论文是在动荡的时期完成的——无论是个人生活、职业生涯还是社会层面都是如此。在这段时间里，在我们携手走过的十五 + 年里，莫妮克始终证明自己是「生活」团队中不可或缺的一员。再多的言语也无法表达她和我们的儿子 Owen 为我的人生增添了多少光彩。无论我们还能共度多久，我都愿意用一切去换取这段时光。没有莫妮克的支持、关怀和理解，这篇学位论文就不可能完成。感谢你所做的一切。

说到学术生活，首先要向我的导师们致以诚挚的谢意。Walter Daelemans 教授及其在 CLiPS 的实验室让我结识了本地区的计算语言学研究团队，而他的研究成果和理念也是本学位论文中大多数论文的起点。Eric Postma 教授信任我，让我在蒂尔堡大学的数据科学硕士项目中授课，使我得以继续完成学位论文；对于我试图为此开辟的研究方向，他也始终给予坚定支持。他的办公室是一个让我觉得可以随时分享想法和见解、无须摆出严肃学术姿态的地方。Grzegorz Chrupała 博士承担起指导这条与主流稍显疏远的研究路线的任务。他是一位宝贵的批判性把关者，我可以借由他的意见打磨想法，甚至整篇论文，而且常常是在极其临近截止时间时才请他审阅。最重要的是，我很感激你们三位在这一过程中，对我相当非传统的职位和做法所表现出的耐心。

我还要向委员会致以谢意：Albert Gatt 博士、Willem-Jan van den Heuvel 教授博士、Marie-Francine (Sien) Moens 教授博士、Malvina Nissim 教授博士和 Martin Potthast 初级教授博士。他们愿意审阅这项工作，并且——令我惊喜的是——给出了非常友善的评论（我已经习惯了 *ACL 审稿群中的某些评论风格）。

虽然我在指导上花的时间不多（至少我是这么认为的），但有两个人对我们这种“非等级制、以辅导为导向的文化”尤为重要：Ákos Kádár博士负责研究讨论，Travis Wiltshire博士则负责大学认为重要的其他一切。我有幸和你们两位花了很多时间，一起做学术界之外我们都喜欢的事情。能与你们共事真是莫大的荣幸；要不是我太爱唱反调，你们肯定得在我身后站上一个多小时。

我还想感谢悲伤企鹅小组的其他成员：我的备用克隆人 Bram Willemsen、同为学术界父亲的 Bertrand Higy博士（我们很想念你们两位），以及人格最高的 Chris van der Lee博士。我们每天（线上）的打趣、每周（或隔周）喝啤酒（特别感谢 Cat's Back），还有拉面（也要感谢 Pig & Rye），都让工作变得很愉快，尤其是在工作不那么愉快的时候。你们都成了我非常亲爱的朋友——愿这份情谊延续一生。

我们有一个很棒的系，我想感谢每一位为此作出贡献的人（说真的，要不是时间紧迫，我会把大家都列出来）。不过，我尤其要感谢几位帮助我保持冷静的人：Henry Brighton博士、Elisabeth（Lisanne）Huis in 't Velt博士、Drew Hendrickson博士和Emmanuel Keuleers博士。感谢你们帮助我应对学术界的种种事务（其中有些确实不那么令人兴奋）。

尤其要感谢我们现任和前任的博士生：在这项工作逐步于两个地点开展的过程中，我见证了许多人来来去去。你们中很多人参与了大部分工作，我相当确信，你们每个人都以某种方式帮助这项工作取得了如今的成果。安特卫普团队里，我无疑是进度最慢的：Giovanni Cassani博士、Robert Grimm博士、Enrique Manjavacas博士和Stéphan Tulkens博士；特立独行的博士委员会成员：George Aalbers、Lieke Gelderloos和Paris Mavromoustakos Blom博士；还有你们这些难以归入任何恰当类别的人：Thiago Castro Ferreira博士、Mariana Dias Da Silva、Mirjam de Haas博士、Nanne van Noord博士、Javad Pourmostafa和Lisa Rombout。感谢你们一路相伴，走过Path。

最后，谨向Karin Berkhout和Eva Verschoor致谢，感谢她们包容我糟糕的组织能力（我保证正在改善），以及由此带来的最后一刻请求；也感谢Lars Biemans让我的工作站始终保持运行。

在生活中稍微不那么学术的一面，我想感谢我的另一群、

年长一些但无疑十分亲爱的朋友：Suzanne Aussems博士、Thom van Iersel、Vincent Lichtenberg、Kristel Litjens和Danny Merkx博士。这几年我确实不太容易联系上，对此我非常感激的是，这似乎从来都不成问题。

感谢你们在学术界之外给予我的关爱与支持。我要感谢我的家人，特别是我的妈妈 Inge Koenen 和继父 Ignace Cichy，感谢他们在过去两年里让我享有许多个周五假期；还要感谢我的姻亲 Bernadette 和 Hans（他与我父亲在同一年去世），每当需要帮助时，他们总会随时伸出援手；也要感谢我的祖母 Trini Koenen（她于2020年初去世），她是莫妮克和我最忠实的支持者。最后，还要感谢我的弟弟 Daniël Emmery——他比我年轻，却毫无疑问比我聪明得多——他目前也在攻读博士学位（做的是真正的科学研究）。有一位心意相通、随时可以依靠的手足，实在是一种莫大的幸运。我对此从不视为理所当然。希望我们很快又能找时间，心安理得地忘掉自己已经不适合玩某些电子游戏；也希望我们能合作发表一些有趣（但主要是实用）的 Emmery & Emmery出版物。

最后但同样重要的是，我要感谢我那两位可靠的得力助手 Onyx 和 Thurisaz（真希望有人种了些树来抵消你们的碳足迹），以及 Lubomír Beneš、Vladimír Jiránek、Kees Prins 和 Siem van Leeuwen，感谢他们在过去两年里成功转移了 Owen 的注意力。

大功告成！

摘要

通过算法收集和处理（敏感）个人数据的服务，如今已像互联网本身一样无处不在。这些服务日益集中化的特征如今也已渗透到学术界，引发了社会和学术界对其是否符合公共利益的担忧。尽管相关学术研究总体上可能致力于解决具有社会意义的问题，但由此产生的工具往往并未向公众开放使用；更重要的是，其方法也并非从普通用户的角度出发进行设计。自然语言处理（NLP）领域尤其如此：尽管人们努力让其最先进技术（需要企业级计算资源）“民主化”，但同时也出现了通过此类模型的访问权限获利的趋势。从安全和隐私的角度来看，这令人担忧，因为它妨碍了对模型的检查、对双重用途问题的处理，以及对数据使用的控制。

为解决这些问题，本学位论文提出了自然语言处理中的以用户为中心的安全框架，并展示了该框架如何提升相关研究的可访问性。因此，论文聚焦于自然语言处理领域内两个备受公共利益关注的安全领域。首先是作者画像，它可能通过侵入式推断损害在线隐私。若无法深入了解这些模型及其预测结果，也无法访问这些模型，互联网用户便没有合理的启发式方法来保护自己免受此类推断的侵害。其次是网络欺凌检测，它采用集中化方法，即在社交平台上进行内容审核。由于获取适当数据受到限制，且任务性质迅速变化（既有词汇变异，也有文化变迁），其分类器的有效性会大幅降低，因此往往无法得到准确反映。

在所提出的框架下，我们研究了针对语言的对抗性攻击；也就是说，改变给定输入（生成对抗样本），使给定模型无法按预期运行。这些攻击构成了我们以用户为中心的安全问题之间的共同主线；它们与保护隐私、抵御作者画像的混淆方法密切相关，而对抗样本也可能有助于评估词汇变异和数据增强对网络欺凌检测的影响。我们旨在回答以下研究问题：

研究问题1 端到端攻击和定向攻击在多大程度上可用于以用户为中心的对抗性文体计量?

研究问题2 网络欺凌分类器是否足够稳健，能够以用户为中心的方式部署?

为此，我们作出以下贡献：第1章提出一种风格中立的对抗性文体计量框架；第2章明确测量对抗性文体计量的可迁移性；第3章评估网络欺凌分类器的泛化能力；第4章探讨网络欺凌分类器对词汇变异的敏感性，以及扩充其语料库的潜力。

摘要

通过算法收集和处理（敏感）个人数据的服务，与互联网本身一样无处不在。它们的地位日益重要，并对学术界产生影响，这引发了社会和学术界对如何保障公共利益的担忧。尽管学术研究通常试图解决具有社会意义的问题，但由此产生的软件往往并未针对公众使用进行设计；或许更重要的是，相关方法也没有考虑其潜在用户的需求。这在自然语言处理（Natural Language Processing，简称 NLP）领域尤为明显。尽管有人试图让最先进的技术（这些技术需要巨大的计算能力）“普及化”，与此同时，一种借助此类模型的访问权限牟利的趋势也逐渐兴起。从安全和隐私的角度来看，这极其令人担忧，主要是因为它妨碍了模型的透明度、防止双重用途，以及对数据使用的管控。

为了解决这些问题，本博士论文提出了一个用于自然语言处理中的以用户为中心的安全框架，并展示了该框架如何提升相关研究的可及性。论文特别关注自然语言处理领域中两个备受公众关注的安全领域。首先是作者画像，它可能通过侵入性推断危及在线隐私。如果无法详细了解这些模型及其预测，也无法访问它们，互联网用户就没有有效的启发式方法来抵御此类推断。其次是网络欺凌检测，该领域采用集中式方法；

也就是说，在社交平台上进行内容审核。由于获取适用数据的途径有限，而且任务性质变化迅速（词汇变异和文化变迁均会造成影响），分类模型的有效性往往远不如其宣称的那样。

在提出的框架中，我们还研究针对语言的攻击；也就是说，通过改变语言使给定模型的效果降低。这些攻击贯穿我们面向用户的安全问题；它们与针对作者画像分析的隐私保护混淆方法密切相关，也有助于测试词汇变异和数据增强对网络欺凌检测的影响。由此，我们提出以下研究问题：

研究问题1 端到端模型和定向模型在多大程度上可用于面向用户的文体计量攻击?

研究问题2 用于网络欺凌检测的分类模型是否足够稳健，能够以面向用户的方式部署?

因此，本论文作出以下贡献：第1章提出一种应用于文体计量攻击的风格中立框架；第2章衡量此类文体计量攻击的可迁移性；第3章评估网络欺凌分类模型的泛化能力；第4章评估这些模型对词汇变异的敏感性，并探讨扩大相关语料库的可能性。

参考文献

Martín Abadi、Ashish Agarwal、Paul Barham、Eugene Brevdo、Zhifeng Chen、Craig Citro、Greg S. Corrado、Andy Davis、Jeffrey Dean、Matthieu Devin、Sanjay Ghe- mawat、Ian Goodfellow、Andrew Harp、Geoffrey Irving、Michael Isard、Yangqing Jia、Rafal Jozefowicz、Lukasz Kaiser、Manjunath Kudlur、J osh Levenberg、Dan Mané、Rajat Monga、Sherry Moore、Derek Murray、Chris Olah、Mike Schuster、Jonathon Shlens、Benoit Steiner、Ilya Sutskever、Kunal Talwar、Paul Tucker、Vincent Vanhoucke、Vijay Vasudevan、Fernanda Viégas、Oriol Vinyals、Pete Warden、Martin Wattenberg、Martin Wicke、Yuan Yu 和 Xiaoqiang Zheng。2015年。TensorFlow：异质系统上的大规模机器学习。软件可从 tensorflow.org 获取。(第104页)

Martín Abadi、Paul Barham、Jianmin Chen、Zhifeng Chen、Andy Davis、Jeffrey Dean、Matthieu Devin、Sanjay Ghemawat、Geoffrey Irving、Michael Isard、Manjunath Kudlur、Josh Levenberg、Rajat Monga、Sherry Moore、Derek Gordon Murray、Benoit Steiner、Paul A. Tucker、Vijay Vasudevan、Pete Warden、Martin Wicke、Yuan Yu 和 Xiaoqiang Zheng。2016年a. Tensorflow：一个用于大规模机器学习的系统。载于第*12*届 *USENIX* 操作系统设计与实现研讨会（*OSDI*）*2016*，美国佐治亚州萨凡纳，*11*月*2-4*，*2016*，第265–283页。USENIX 协会。（第58页）

Martín Abadi、Andy Chu、Ian J. Goodfellow、H. Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。2016年b. 采用差分隐私的深度学习。载于《*2016*年 *ACM SIGSAC* 计算机与通信安全会议论文集》，奥地利维也纳，*10*月*24-28*，*2016*，第308–318页。ACM。（第7 和46页）

Gregory D. Abowd、Anind K. Dey、Robert J. Orr 和 Jason A. Brotherton。1998年。可穿戴与泛在计算中的情境感知。*VirtualReal.*，3(3)：200–211。(第5页)

Carlisle M. Adams。2006年。隐私技术分类。渥太华大学法律与技术期刊，1。(第46页)

Robert Adragna、Elliot Creager、David Madras 和 Richard S. Zemel。2020年。不变学习中的公平性与鲁棒性：毒性分类案例研究。*CoRR*，abs/2011年。06485。（第166页）

Sweta Agrawal 和 Amit Awekar。2018年。利用深度学习检测多个社交媒体平台上的网络欺凌。载于《信息检索进展》——第*40*届欧洲信息检索研究会议（*ECIR*）*2018*年，法国格勒诺布尔，*3*月*26*—*29*日，*2018*年，论文集，第10772 卷，载于《计算机科学讲义》，第141–153页。施普林格。（第84、103、104和116页）

Jalal S. Alowibdi、Ugo A. Buy 和 Philip S. Yu。2013年。Twitter 上用于性别分类的个人资料特征实证评估。收录于第*12*届国际机器学习与应用会议（*ICMLA*）*2013*年，美国佛罗里达州迈阿密，*12*月*4-7*日，*2013*年，第*1*卷，第365–369页。IEEE。(第22 和46页)

David Alvarez-Melis 和 Tommi Jaakkola。2017年。用于解释黑箱序列到序列模型预测结果的因果框架。收录于《*2017*年自然语言处理实证方法会议论文集，第412–421页，丹麦哥本哈根。计算语言学协会。(第14页)

Ateret Anaby-Tavor、Boaz Carmeli、Esther Goldbraich、Amir Kantor、George Kour、Segev Shlomov、Naama Tepper 和 Naama Zwerdling。2020年。数据不足？深度学习来助力！收录于第*34*届*AAAI*人工智能会议，*AAAI2020*年、第*32*届人工智能创新应用会议，*IAAI2020*年、第十届*AAAI*人工智能教育进展研讨会，*EAAI2020*年，美国纽约州纽约市，*2*月*7-12*日，*2020*年，第7383–7390页。AAAI出版社。(第146页)

Denise L. Anthony、Tristan Henderson 和 David Kotz。 2007年。 位置感知计算环境中的隐私。 *IEEE Pervasive Comput.*， 6(4):64–72。 (第8页)

Shlomo Argamon、Sushant Dhawle、Moshe Koppel 和 James W Pennebaker。2005 年。 人格类型的词汇预测因子。载于 《*2005* 年 *Interface Foundation* 与北美分类学会联合年会论文集，圣路易斯，密苏里州，美国，*6*月， *2005*年，第 1–16。 (第49页)

Bernard Arogyaswamy。 2020年。 大型科技公司与社会可持续性：伦理框架。 *AI Soc.*， 35(4):829–840。 (第4页)

Mikhail J. Atallah、Craig J. McDonough、Victor Raskin 和 Sergei Nirenburg。2000年。信息保障与安全领域的自然语言处理：概述与实现。载于《*2000*年新安全范式研讨会论文集》，爱尔兰科克郡巴利科顿，*9*月*18-21*，*2000*年，第51–65页。ACM。(第11页)

Giuseppe Ateniese、Luigi V. Mancini、Angelo Spognardi、Antonio Villani、Domenico Vitali 和 Giovanni Felici。2015年。用更智能的机器入侵智能机器：如何从机器学习分类器中提取有意义的数据。《国际安全网络杂志》，10(3)：137–150。(第7页)

H Baayen、H van Halteren 和 FTweedie。 1996年。 走出阴影洞穴：利用句法标注增强作者归属判定。 《文学与语言计算》， 11(3)：121–132。 (第11页)

Harald Baayen、Hans van Halteren、Anneke Neijt 和 Fiona Tweedie。2002年。一项作者归属判定实验。载于 *JADT* 论文集 *2002*年：第六届文本数据统计分析国际会议，第 1 卷，页码 29–37。 (第49页)

Sameer Badaskar、Sachin Agarwal 和 Shilpa Arora。 2008年。 识别真实或虚假文章：迈向更好的语言建模。载于 第三届自然语言处理国际联合会议论文集：第二卷。 (第9页)

P Baldi、S Brunak、P Frasconi、G Soda 和 G Pollastri。 1999年。 利用过去与未来进行蛋白质二级结构预测。 《生物信息学》（英国牛津）， 15(11)：937—946。 (第103页)

David Bamman、Chris Dyer 和 Noah A. Smith。 2014年a. 地理定位语言的分布式表示。载于 论文集：第*52*届计算语言学协会年会（第*2*卷：短论文），第 828–834 页，美国马里兰州巴尔的摩。计算语言学协会。 （第9页）

David Bamman、Jacob Eisenstein 和 Tyler Schnoebelen。 2014年b. 社交媒体中的性别认同与词汇变异。 社会语言学杂志， 18（2）： 135–160。 （第9页）

Ritwik Banerjee、Song Feng、Jun Seok Kang 和 Yejin Choi。 2014年。 数字文本中的击键模式作为韵律：关于欺骗性评论和文章的案例研究。载于 论文集：第*2014*年自然语言处理实证方法会议（*EMNLP*），第 1469–1473页，多哈，卡塔尔。计算语言学协会。 （第9页）

Satanjeev Banerjee 和 Alon Lavie。 2005年。 METEOR：一种自动机器翻译评估指标，与人工判断的相关性更高。见机器翻译和/或摘要内在与外在评估指标 *ACL* 研讨会论文集，第 65–72页，密歇根州安娜堡。计算语言学协会。 (第 23、33、61和 136 页)

Michele Banko、Brendon MacKeen 和 Laurie Ray。 2020年。 有害内容的统一分类体系。见 第四届在线虐待与伤害研讨会论文集，第 125–137页，线上。计算语言学协会。 (第 126和 146页)

Susanne Barth和Menno D.T. de Jong。 2017年。 隐私悖论——调查表达的隐私顾虑与实际在线行为之间的差异：系统性文献综述。 《电信与信息学》， 34(7)：1038–1058。 (第6页)

Angelo Basile、Gareth Dwyer、Maria Medvedeva、Josine Rawee、Hessel Haagsma和Malvina Nissim。 2017年。 简约至上：极简系统在作者画像中胜过复杂模型。载于 实验信息检索与多语言、多模态和交互 -第*9*届*CLEF*协会国际会议，*CLEF 2018*年，法国阿维尼翁，*9*月*10-14*， *2018*年，论文集，卷 11018 ，《计算机科学讲义》， 页码 143–156。施普林格。 (第58页)

Markus Bayer、Marc-André Kaufhold和Christian Reuter。 2021年。 文本分类数据增强综述。 *CoRR*，abs/2107。03158。 (第148页)

Jennifer Bayzick、April Kontostathis 和 Lynne Edwards。 2011年。 使用计算机软件检测网络欺凌的存在。载于 *ACM WebSci*会议论文集，德国科布伦茨，*2011*年。 （第 82、83、93、108和132页）

Yonatan Belinkov 和 Yonatan Bisk。2018年。 合成噪声和自然噪声都会破坏神经机器翻译。载于 第*6*届国际学习表征会议，*ICLR 2018*年，加拿大不列颠哥伦比亚省温哥华，*4*月*30*日 至*5*月 *3*日， *2018*年，会议论文集。OpenReview.net。 （第 54和159页）

Charley Beller、Rebecca Knowles、Craig Harman、Shane Bergsma、Margaret Mitchell 和 Benjamin Van Durme。 2014年。 我是 Belieber：通过自我认同和概念属性体现社会角色。载于 《 *52*届计算语言学协会年会论文集（第 *2*卷：短论文），第 181–186页，美国马里兰州巴尔的摩。计算语言学协会。 （第 10、22、48和57页）

Emily M. Bender 和 Batya Friedman。2018年。自然语言处理的数据说明：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。（第7页）

Emily M. Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。 2021年。 论随机鹦鹉的危害：语言模型会不会太大？ 载于 *FAccT '21*：*2021* 年 *ACM* 公平性、问责制与

透明度线上活动 / 加拿大多伦多，*3*月 *3-10*日，*2021*年，第 610–623页。ACM。(第7页)

José Manuel Benítez、Juan Luis Castro 和 Ignacio Requena。 1997。 人工神经网络是黑箱吗？ *IEEE*神经网络汇刊， 8（5）： 1156–1164。 （第11页）

Tanya Beran 和 Qing Li。 2008年。 网络欺凌与校园欺凌之间的关系。 《学生福祉期刊》， 1（2）： 16–33。 （第76页）

Katherine Bessière、Sara Kiesler、Robert Kraut 和 Bonka S. Boneva。 2008年。 互联网使用和社会资源对抑郁变化的影响。 《信息、传播与社会》， 11 （1）： 47–70。 （第75页）

Janek Bevendorff、Tobias Wenzel、Martin Potthast、Matthias Hagen 和 Benno Stein。2020年。 论基于散度的作者身份混淆：对统计性作者身份验证领域现有最佳方法的攻击。 *it Inf. Technol.*， 62（2）： 99–115。 （第50页）

Douglas Biber。1995年。《语域变异的维度：跨语言比较》。剑桥大学出版社。(第11页)

Amy Binns。2013年。Facebook的丑陋姐妹：Formspring和Ask.fm上的匿名性与辱骂。《媒体教育研究期刊》。(第93页至132页)

Abeba Birhane、Vinay Uday Prabhu和Emmanuel Kahembwe。2021年。多模态数据集：厌女、色情内容与恶性刻板印象。*CoRR*，abs/2110。01963。(第166页)

John Blitzer、Mark Dredze和Fernando Pereira。2007年。传记、宝莱坞、音箱与搅拌机：情感分类的领域适应。载于《*45*届计算语言学协会年会论文集》，第440–447页，捷克共和国布拉格。计算语言学协会。(第87页)

Haohan Bo、Steven H. H. Ding、Benjamin C. M. Fung和Farkhund Iqbal。2021年。ER-AE：用于作者身份匿名化的差分隐私文本生成。载于《*2021*年计算语言学协会北美分会人类语言技术会议论文集》，第3997–4007页，线上。计算语言学协会。(第51页)

Dasha Bogdanova、Paolo Rosso 和 Thamar Solorio。2014年。 探索用于检测网络恋童癖的高层特征。 《计算机、语音与语言》， 28（1）： 108–120。 （第120页）

Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。2016年。利用子词信息丰富词向量。*arXiv* 预印本 *arXiv*：*1607*。*04606*。（第33页）

Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。 2017年。 利用子词信息丰富词向量。 《计算语言学协会汇刊》， 5：135–146。 (第101页)

Barron Brainerd。 1974。 《语言与文学中的证据加权》。多伦多大学出版社。 (第11页)

Nadia M. Brashier 和 Daniel L. Schacter。 2020年。 假新闻时代的衰老。 《心理科学当前方向》， 29(3)：316–323。 PMID： 32968336。 (第3页)

Eloi Brassard-Gourdeau 和 Richard Khoury。 2019年。 利用情感信息进行隐蔽的毒性检测。载于 《第三届在线辱骂性语言研讨会论文集》，页码1–10，意大利佛罗伦萨。计算语言学协会。(第 148页)

Michael Brennan、Sadia Afroz 和 Rachel Greenstadt。2012年。 对抗性文体计量：规避作者识别以保护隐私和匿名性。 《*ACM*信息与系统安全汇刊》，15（3）：12：1–12：22。 （第 15、25和 50页）

Uwe Bretschneider、Thomas Wöhner 和 Ralf Peters。 2014年。 检测社交网络中的线上骚扰。载于 国际信息系统会议论文集——通过信息系统建设更美好的世界，*ICIS 2014*年，奥克兰，新西兰，*12*月 *14-17*，*2014*年。信息系统协会。 （第82、84、93、132和147页）

Claude S. Brinegar。 1963。马克·吐温与 Quintus Curtius Snodgrass 书信：作者身份统计检验。 《美国统计协会会刊》，58（301）：85–96。 （第11页）

广播标准局。 2013年。广播中哪些词不该说：词语在广播中的可接受程度。 `http://bsa.govt.nz/images/assets/Research/Acceptability_of_Words_2013_WEB.pdf`。最后访问于 15/03/2016年。 (第84页)

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、AdityaRamesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher

Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、BenjaminChess、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*33*：神经信息处理系统年会*2020*年，*NeurIPS 2020*年，*12*月*6-12*，*2020*年，线上。(第134 和148页)

John D. Burger、John Henderson、George Kim 和 Guido Zarrella。 2011年。Twitter 上的性别判别。载于《*2011*年 自然语言处理实证方法会议论文集，第 1301–1309页，苏格兰爱丁堡，英国。计算语言学协会。（第 23 和 48页）

Ben Burtenshaw 和 Mike Kestemont。2021年。 用于跨语言多标签毒性检测的荷兰语数据集。载于 《 第*14*届可比语料库构建与使用研讨会（*BUCC 2021*年），第 75–79页，线上（虚拟模式）。INCOMA Ltd.（第166页）

Carrie J. Cai 和 Philip J. Guo。 2019年。 软件开发者学习机器学习：动机、障碍与需求。载于 《*2019*年*IEEE*可视化语言与以人为中心计算研讨会，*VL/HCC 2019*年，美国田纳西州孟菲斯，*10*月 *14*日至*18*日，*2019*年，第 25–34页。IEEE计算机学会。（第12页）

Aylin Caliskan 和 Rachel Greenstadt。 2012年。 翻译一次、两次、三次并进行归因：识别译文中的作者和机器翻译工具。载于 第六届*IEEE*语义计算国际会议，*ICSC 2012*年，意大利巴勒莫，*9*月 *19-21*， *2012*年，第 121–125页。IEEE计算机学会。（第 25 页）

Aylin Caliskan、Fabian Yamaguchi、Edwin Dauber、Richard E. Harang、Konrad Rieck、Rachel Greenstadt 和 Arvind Narayanan。 2018年。编译后编码风格依然存在：从可执行二进制文件中去匿名化识别程序员。载于 第*25*届网络与分布式系统安全年度研讨会，*NDSS 2018*年，美国加利福尼亚州圣迭戈，*2*月 *18-21*，*2018*年。互联网协会。 （第 25页和 50页）

Gerardo Canfora、Andrea Di Sorbo、Enrico Emanuele、Sara Forootani 和 Corrado Aaron Visaggio。 2018年。 一种防止社交网络帖子隐私泄露的基于自然语言处理的解决方案。载于 《第*13*届可用性、可靠性与安全国际会议论文集，*ARES 2018*年，德国汉堡，*8*月 *27-30*， *2018*年，第 36：1–36：6页。ACM。 （第11页）

Nicholas Carlini、Florian Tramèr、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom B. Brown、Dawn Song、ÚlfarErlingsson、Alina

Oprea、以及 Colin Raffel。2021年。从大语言模型中提取训练数据。载于第*30*届 *USENIX* 安全研讨会，*USENIX* 安全*2021*年，*8*月*11-13*，*2021*年，第2633–2650页。USENIX 协会。（第7页）

Keith Carlson、Allen Riddell 和 Daniel N. Rockmore。 2017年。 使用循环神经网络对文本进行零样式迁移。 *CoRR*，abs/1711。04731。 （第 25 页和 29 页）

Samuel Carton、Qiaozhu Mei 和 Paul Resnick。 2020年。 基于特征的解释无法帮助人们检测在线毒性的错误分类。载于 第十四届 *AAAI* 国际网络与社交媒体会议，*ICWSM 2020*年，线上举行，原定地点：美国佐治亚州亚特兰大，*6*月 *8-11*， *2020*年，第 95–106 页。AAAI出版社。 （第165页）

Tommaso Caselli、Valerio Basile、Jelena Mitrovic 和 Michael Granitzer。 2020年。HateBERT：针对英语辱骂性语言检测重新训练 BERT。 *CoRR*，abs/2010。12472。 （第147页）

Tommaso Caselli、Valerio Basile、Jelena Mitrovi´c 和 Michael Granitzer。 2021年。 Hate-BERT：针对英语辱骂性语言检测重新训练 BERT。载于《第 *5*届网络滥用与危害研讨会（*WOAH2021*年）论文集，页码 17–25，线上。计算语言学协会。（第 133 页）

IsaacCaswell、Ciprian Chelba 和 David Grangier。 2019年。 带标签回译。载于《第四届机器翻译会议论文集（第*1*卷：研究论文）》，页码 53–63，意大利佛罗伦萨。计算语言学协会。（第135页）

Daniel Cer、Yinfei Yang、Sheng-yi Kong、Nan Hua、Nicole Limtiaco、Rhomni St. John、Noah Constant、Mario Guajardo-Cespedes、Steve Yuan、Chris Tar、Brian Strope 和 Ray Kurzweil。 2018年。 英语通用句子编码器。载于《*2018*年自然语言处理实证方法会议：系统演示论文集，页码169–174，比利时布鲁塞尔。计算语言学协会。 （第 55页和第 58页）

Yee Seng Chan 和 Hwee Tou Ng。 2006年。 领域适应中词义消歧的类别先验估计。载于 论文集：第 *21*届国际计算语言学会议暨 *44*届计算语言学协会年会，页码 89–96，澳大利亚悉尼。计算语言学协会。 （第87页）

A. P. Sarath Chandar、Stanislas Lauly、Hugo Larochelle、MiteshM. Khapra、Balaraman Ravindran、Vikas C. Raykar 和 Amrita Saha。 2014年。 学习双语词表示的自编码器方法。载于 神经信息处理系统进展 *27*：神经信息处理系统年会 *2014*，*12*月 *8-13*日 *2014*，加拿大魁北克省蒙特利尔，页码 1853–1861。 （第26页）

Wendy W Chapman、PrakashMNadkarni、Lynette Hirschman、Leonard W D'Avolio、Guergana K Savova 和 Ozlem Uzuner。2011年。克服临床文本自然语言处理的障碍：共享任务的作用及对更多创新解决方案的需求。美国医学信息学会杂志，18(5)：540–543。(第12页)

Jatin Chauhan、Karan Bhukar 和 Manohar Kaul。2021年。针对语言理解模型、具有查询预算且与目标模型无关的对抗攻击。*CoRR*，abs/2106.07047。(第145页)

David Chen 和 William Dolan。2011年。收集用于释义评估的高度平行数据。载于论文集：第*49*届计算语言学协会年会：人类语言技术，页码190–200，美国俄勒冈州波特兰。计算语言学协会。(第160页)

Minmin Chen、Zhixiang Eddie Xu、Kilian Q. Weinberger 和 Fei Sha。2012年。用于领域适应的边缘化去噪自编码器。载于 第 *29* 届国际机器学习会议（*ICML*）论文集，*2012*年，英国苏格兰爱丁堡，*2012*年*6*月 *26* 日至*7*月 *1*日，*2012*年。icml.cc / Omnipress。 （第87页）

Tseng-Hung Chen、Yuan-Hong Liao、Ching-Yao Chuang、Wan-Ting Hsu、Jianlong Fu 和 Min Sun。 2017年。展示、适应并描述：跨领域图像描述生成器的对抗训练。载于 *IEEE*国际计算机视觉会议（*ICCV*），第 2卷。 （第26页）

Lu Cheng、Ruocheng Guo、K. Selçuk Candan 和 Huan Liu。2020年a。 不平衡跨领域分类的表示学习。载于 第 *2020*年*SIAM*国际数据挖掘会议（*SDM*）论文集，*2020*年，美国俄亥俄州辛辛那提，*2020*年*5*月 *7*日至*9*日， *2020*年，第 478–486页。SIAM。 （第87页）

Lu Cheng、Ruocheng Guo、Yasin N. Silva、Deborah L. Hall 和 Huan Liu。2019年。用于 Instagram 社交网络网络欺凌检测的分层注意力网络。载于《*2019*年*SIAM*国际数据挖掘会议（*SDM*）*2019*年，加拿大艾伯塔省卡尔加里，*2019*年*5*月2—*4*日，*2019*年，第235–243页。SIAM。（第82页）

Minhao Cheng、Jinfeng Yi、Pin-Yu Chen、Huan Zhang 和 Cho-Jui Hsieh。2020年b.Seq2sick：使用对抗样例评估序列到序列模型的鲁棒性。载于第三十四届*AAAI* 人工智能会议*2020*年、第三十二届人工智能创新应用会议（*IAAI*）*2020*年、第十届 *AAAI* 人工智能教育进展研讨会（*EAAI*）*2020*年，美国纽约州纽约市，*2020*年*2*月*7*—*12*日，*2020*年，第3601–3608页。AAAI出版社。（第47页）

François Chollet 等人2015年。Keras。 https://keras.io。（第104页）

Munmun De Choudhury、Michael Gamon、Scott Counts 和 Eric Horvitz。2013年。通过社交媒体预测抑郁。载于第七届国际博客与社交媒体会议论文集（*ICWSM*）*2013*年，美国马萨诸塞州剑桥，7月*8*—*11*，*2013*年。AAAI出版社。（第9页）

Ronan Collobert、Jason Weston、Léon Bottou、Michael Karlen、Koray Kavukcuoglu 和 Pavel P. Kuksa。2011年。几乎从零开始的自然语言处理。机器学习研究期刊，12：2493–2537。（第100页）

Ioanna D. Constantiou 和 Jannis Kallinikos。2015年。新游戏，新规则：大数据与不断变化的战略语境。《信息技术杂志》，30（1）：44–57。（第4页）

Glen Coppersmith、Mark Dredze、Craig Harman 和 Kristy Hollingshead。2015年。从注意缺陷多动障碍到社交焦虑障碍：通过自我报告的诊断分析 Twitter 上的心理健康语言。 *NAACL HLT 2015*年，第 1 页。 (第9页)

Corinna Cortes 和 Vladimir Vapnik。 1995年。 支持向量网络。 *Mach.Learn.*，20(3)：273–297。 (第 59、98 和135页)

Maral Dadvar。 2014年。 专家与机器联手对抗网络欺凌。博士论文，荷兰特温特大学。 (第82页)

Maral Dadvar、Dolf Trieschnigg 和 Franciska de Jong。 2014年。 专家与机器对抗欺凌者：检测网络欺凌者的混合方法。载于 人工智能进展——第*27*届加拿大人工智能会议，加拿大人工智能 *2014*年，加拿大魁北克省蒙特利尔，*5*月 *6*—*9*，*2014*年。论文集，《8436 卷，载于 《计算机科学讲义》，第 275–281页。施普林格。 (第 107 页和 147页)

Walter Daelemans。2013年。计算文体学中的解释。载于 《计算语言学与智能文本处理》- 第*14*届国际会议，*CICLing 2013*年，

希腊萨摩斯，*3*月 *24-30*日，*2013*年，论文集，第*II*部分，第 7817 卷，《计算机科学讲义》，第 451–462页。施普林格。（第 9、22和46）

Hal Daumé III. 2007年。 令人沮丧地简单的领域适应方法。载于 第*45*届计算语言学协会年会论文集，第 256–263页，捷克共和国布拉格。计算语言学协会。 （第135页）

Hal Daumé III、Abhishek Kumar 和 Avishek Saha。 2010年。 令人沮丧地简单的半监督式领域适应方法。载于 论文集 *2010*年自然语言处理领域适应研讨会，第 53–59页，瑞典乌普萨拉。计算语言学协会。 （第87页）

Sergej Dechand、Alena Naiakshina、Anastasia Danilova 和 Matthew Smith。 2019年。我们不信任加密：面向大众的端到端加密对用户感知的影响。载于 *IEEE*欧洲安全与隐私研讨会，*EuroS&P2019*年，瑞典斯德哥尔摩，*6*月 *17-19*日，*2019*年，第 401–415页。IEEE。 （第7页）

Francine Dehue、Catherine Bolman 和 TrijntjeVöllink。 2008年。 网络欺凌：青少年的经历与家长感知。 *CyberpsycholBehav*，11(2)：217–223。(第 76页)

Michael Denkowski 和 Alon Lavie。 2011年。 METEOR 1。3：用于机器翻译系统可靠优化与评估的自动指标。载于 第六届统计机器翻译研讨会论文集，第 85–91页，苏格兰爱丁堡。计算语言学协会。 (第136页)

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年 a. BERT：用于语言理解的深层双向 Transformer 预训练。载于《*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（长短论文），第 4171–4186 页，明尼苏达州明尼阿波利斯。计算语言学协会。 (第54 和101页)

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年 b. BERT：用于语言理解的深层双向 Transformer 预训练。载于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，*NAACL-HLT2019*年，美国明尼苏达州明尼阿波利斯，*6*月*2-7*，*2019*年，第*1* 卷（长篇与短篇论文），页码4171–4186。计算语言学协会。（第133和135页）

Karthik Dinakar、Roi Reichart 和 Henry Lieberman。2011年。文本网络欺凌检测建模。载于*The Social Mobile Web，2011*年*ICWSM*研讨会论文，西班牙加泰罗尼亚巴塞罗那，*7*月*21*日，*2011*年，第WS-11-02 卷，载于*AAAI*技术报告。AAAI。（第82 和132页）

Xinshuai Dong、AnhTuan Luu、Rongrong Ji 和 Hong Liu。2021年。增强抵御自然语言词语替换的鲁棒性。载于第*9*届国际学习表征会议（*ICLR*）*2021*年，奥地利，线上活动，*5*月*3*至*7*日，*2021*年。OpenRe-view.net。（第158页）

Yuxiao Dong、Yang Yang、Jie Tang、Yang Yang 和 Nitesh V. Chawla。2014年。推断移动社交网络中的用户人口统计特征与社交策略。载于第*20*届*ACM SIGKDD*国际知识发现与数据挖掘会议（*KDD '14*，美国纽约州纽约市，*8*月*24* 至*27*日，*2014*年，第15–24页。ACM。（第8页）

Maeve Duggan。2015年。移动消息与社交媒体2015年。皮尤研究中心。（第77页）

Helge Dyvik。2004。翻译作为语义镜像：从平行语料库到*WordNet*，第309 – 326页。Brill，莱顿，荷兰。（第29页）

Javid Ebrahimi、Daniel Lowd 和 Dejing Dou。 2018年a. 关于字符级神经机器翻译的对抗样本。载于《第*27*届国际计算语言学会议论文集》，第 653–663页，圣菲，新墨西哥州，美国。计算语言学协会。 (第47页)

Javid Ebrahimi、Anyi Rao、Daniel Lowd 和 Dejing Dou。 2018年b. HotFlip：用于文本分类的白盒对抗样本。载于《第*56*届计算语言学协会年会论文集（第*2*卷：短论文）》，第 31–36页，墨尔本，澳大利亚。计算语言学协会。 (第127页)

Mohammad Reza Ebrahimi、Ching Y. Suen 和 Olga Ormandjieva。 2016年。利用深度卷积神经网络检测社交媒体中的掠夺性对话。数字调查，18：33–49。 (第120页)

Harrison Edwards 和 Amos J. Storkey。2016年。利用对手审查表示。载于第 *4* 届国际学习表征会议（*ICLR*）*2016*年，波多黎各圣胡安，*5* 月*2-4*日，*2016*年，会议论文集。（第7页和46页）

Steffen Eger、Gözde Gül ,Sahin、Andreas Rücklé、Ji-Ung Lee、Claudia Schulz、Mohsen Mesgar、Krishnkant Swarnkar、Edwin Simpson 和 Iryna Gurevych。 2019年。 文本

像人类一样处理：以视觉方式攻击和防护自然语言处理系统。载于《*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（长篇与短篇论文），第1634–1647 页，明尼苏达州明尼阿波利斯。计算语言学协会。（第 50 和 147 页）

Hans-Jörg Ehni。 2008年。 双重用途与科学家的伦理责任。 *ArchivumImmunologiae et Therapiae Experimentalis*，56(3)：147。 (第10页)

Jacob Eisenstein、Noah A. Smith 和 Eric P. Xing。2011年。 运用结构化稀疏性发现社会语言学关联。 收录于 《*49*届计算语言学协会年会：人类语言技术论文集》，页码1365–1374，美国俄勒冈州波特兰。计算语言学协会。(第 22 和46页)

Fatma Elsafoury、Stamos Katsigiannis、Zeeshan Pervez 和 Naeem Ramzan。 2021年a.当时间线遇上流水线：自动网络欺凌检测综述。 *IEEE Access*，9：103541–103563。 (第 135，140，以及 147页)

Fatma Elsafoury、Stamos Katsigiannis、Steven R. Wilson 和 Naeem Ra mzan。 2021年b.BERT 会关注网络欺凌吗？收录于 *SIGIR ′21*：《*44*届国际 *ACM SIGIR* 信息检索研究与开发会议，线上活动，加拿大，*7* 月 *11-15*，*2021*年，页码1900–1904》。 ACM。 (第 140和 147页)

Isioma Elueze 和 Anabel Quan-Haase。 2018年。 老年人数字生活中的隐私态度与担忧：重新审视韦斯汀的隐私态度类型学。 *AmericanBehavioral Scientist*，62(10)：1372–1391。 (第3页)

Chris Emmery、Grzegorz Chrupała 和 Walter Daelemans。 2017年。 作为远距离标签预测推特性别的简单查询。载于《*3rd Workshopon NoisyUser-generated Text*》，第 50–55页，丹麦哥本哈根。计算语言学协会。 (第2，10，16，22，48，57，60，61，63，64，65，以及157页)

Chris Emmery、Ákos Kádár 和 Grzegorz Chrupała。2021年a. 真实场景中的对抗性文体分析：针对作者画像的可迁移词汇替换攻击。载于《会议论文集：》*16th Conference of the European Chapterof the Association for ComputationalLinguistics: MainVolume*，第2388–2402页，线上。计算语言学协会。（未引用）

Chris Emmery、Ákos Kádár、Grzegorz Chrupała 和 Walter Daelemans。2022。网络欺凌分类器对模型无关扰动很敏感。载于论文集

第十三届语言资源与评估会议，第 2976–2988 页，法国马赛。欧洲语言资源协会。（未引用）

Chris Emmery、Ákos Kádár、Travis J. Wiltshire 和 Andrew T. Hendrickson。2019年。迈向计算认知建模的复现：机器学习视角。《计算脑与行为》，2（3）：242–246。（第2和16页）

Chris Emmery、Enrique Manjavacas Arevalo 和 Grzegorz Chrupała。2018年。通过不变性进行风格混淆。载于《第27届国际计算语言学会议论文集，第984–996页，美国新墨西哥州圣菲。计算语言学协会。（未引用）

Chris Emmery、Ben Verhoeven、Guy De Pauw、Gilles Jacobs、Cynthia Van Hee、Els Lefever、Bart Desmet、Véronique Hoste 和 Walter Daelemans。2021年b。网络欺凌检测的当前局限：评估标准、可复现性和数据稀缺。《语言资源与评估》，55（3）：597–633。（未引用）

Hugo Jair Escalante、Esaú Villatoro-Tello、SaraElena GarzaVillarreal、AdriánPastor López-Monroy、Manuel Montes-y-Gómez 和 Luis Villaseñor Pineda。2017年。使用基于用户档案的表示进行欺骗和攻击性的早期检测。*Expert Syst. Appl.*，89：99–111。（第120页）

David S Evans。2017年。为什么在线平台的竞争动态会导致彻夜难眠，却不会形成昏昏欲睡的垄断。*SSRN* 上可获取*3009438*。（第4页）

Rong-En Fan、Kai-Wei Chang、Cho-Jui Hsieh、Xiang-Rui Wang 和 Chih-Jen Lin。2008年。LIBLINEAR：一个用于大规模线性分类的库。机器学习研究期刊，9：1871–1874。（第59、98和135页）

Qi Feng、Debiao He、Zhe Liu、Huaqun Wang 和 Kim-Kwang Raymond Choo。2020年。SecureNLP：一个用于多方隐私保护自然语言处理的系统。*IEEE Trans. Inf. Forensics Secur.*，15：3709–3721。（第11页）

Song Feng、Ritwik Banerjee 和 Yejin Choi。2012年。用于欺骗检测的句法文体计量学。载于《第*50*届计算语言学协会年会论文集（第*2*卷：短论文），第 171–175 页，韩国济州岛。计算语言学协会。（第 11 页和120页）

Steven Feng、Varun Gangal、Jason Wei、Sarath Chandar、Soroush Vosoughi、Teruko Mitamura 和 Eduard Hovy。2021年。自然语言处理数据增强方法综述

领域。计算语言学协会论文集：*ACL-IJCNLP2021*年，第968–988页，线上。计算语言学协会。（第148页）

Natasha Fernandes、Mark Dras 和 Annabelle McIver。2019年。面向文本文件处理的广义差分隐私。载于《安全与信任原理》—第*8*届国际会议 *POST2019*年，作为软件理论与实践欧洲联合会议 *ETAPS2019*年的一部分举行，捷克共和国布拉格，*4*月*6*—*11*日，*2019*年，会议论文集，第11426 卷，载于《计算机科学讲义》，页码123–148。施普林格。（第46页）

Oluwaseyi Feyisetan、Sepideh Ghanavati 和 Patricia Thaine。2020年。自然语言处理隐私研讨会（privatenlp 2020年）。载于*WSDM'20*：第十三届 *ACM* 网络搜索与数据挖掘国际会议，休斯敦，德克萨斯州，美国，*2*月*3*—*7*日，*2020*年，页码903–904。ACM。（第11页）

Jessica Ficler 和 Yoav Goldberg。2017年。神经语言生成中的语言风格控制。*arXiv* 预印本 *arXiv*：*1707*。*02633*。（第28 和40页）

Jenny Rose Finkel 和 Christopher D. Manning。 2009。 分层贝叶斯领域自适应。载于 人类语言技术论文集： *2009* 计算语言学协会北美分会年度会议，第 602–610页，科罗拉多州博尔德。计算语言学协会。(第 87页)

Tommaso Fornaciari 和 Massimo Poesio。2014年。 将虚假亚马逊评论识别视为群体学习。载于 计算语言学协会 第*14*届欧洲分会会议论文集，第 279–287页，瑞典哥德堡。计算语言学协会。 (第9页)

Paula Fortuna 和 Sérgio Nunes。 2018年。 文本中仇恨言论自动检测综述。 《*ACM*计算调查》， 51(4)：85：1–85：30。 (第146页)

Paula Fortuna、Juan Soler Company 和 Leo Wanner。 2021年。 仇恨言论、毒性、辱骂性和冒犯性语言分类模型在不同数据集上的泛化能力如何？ 《信息处理与管理》， 58(3)：102524。 (第147页)

Antigoni-Maria Founta、Constantinos Djouvas、Despoina Chatzakou、Ilias Leontiadis、Jeremy Blackburn、Gianluca Stringhini、Athena Vakali、Michael Sirivianos 和 Nicolas Kourtellis。 2018年。 大规模众包与 Twitter 辱骂行为特征分析。载于 第十二届国际 *Web* 与

社交媒体，*ICWSM 2018*年，美国加利福尼亚州斯坦福，*6*月 *25-28*， *2018*年，页码 491–500。AAAI出版社。 (第56)

Yaroslav Ganin 和 Victor S. Lempitsky。 2015年。 通过反向传播进行无监督领域自适应。载于 《第*32*届国际机器学习会议（*ICML*） *2015*年，法国里尔，*6—11* 7月 *2015*年，第 37 卷，收录于 《机器学习研究期刊》研讨会与会议论文集，页码 1180–1189。JMLR.org。 （第 24、26 和 28页）

Matt Gardner、Yoav Artzi、Victoria Basmova、Jonathan Berant、Ben Bogin、Sihao Chen、Pradeep Dasigi、Dheeru Dua、Yanai Elazar、Ananth Gottumukkala、Nitish Gupta、Hanna Hajishirzi、Gabriel Ilharco、Daniel Khashabi、Kevin Lin、Jiang ming Liu、Nelson F. Liu、Phoebe Mulcaire、Qiang Ning、Sameer Singh、Noah A. Smith、Sanjay Subramanian、Reut Tsarfaty、Eric Wallace、Ally Zhang 和 Ben Zhou。2020 年。通过对比集评估自然语言处理模型。 *CoRR*，abs/2004。02709。 （第 146 页）

Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna M. Wallach、Hal Daumé III 和 Kate Crawford。2021年。数据集数据表。《*ACM* 通讯》，64（12）： 86–92。（第16页）

Samuel Gehman、Suchin Gururangan、Maarten Sap、Yejin Choi 和 Noah A. Smith。2020年。RealToxicityPrompts：评估语言模型中的神经毒性退化。载于计算语言学协会论文集：*EMNLP2020*年，第3356–3369页，线上。计算语言学协会。（第127、134、139、148和第166页）

Oguzhan Gencoglu。2021年。带有公平性约束的网络欺凌检测。*IEEE* 互联网计算，25（1）： 20–29。（第147页和166页）

Felix A. Gers、Nicol N. Schraudolph 和 Jürgen Schmidhuber。2002年。使用LSTM循环网络学习精确计时。机器学习研究期刊，3：115–143。（第103页）

Xavier Glorot、Antoine Bordes 和 Yoshua Bengio。2011年。面向大规模情感分类的领域适应：一种深度学习方法。载于论文集：第*28*届国际机器学习会议（*ICML*）*2011*，美国华盛顿州贝尔维尤，*6*月*28 - 7*月*2*日，2011，第513–520页。Omnipress。（第87页）

Hila Gonen 和 Yoav Goldberg。 2019年。 粉饰太平：去偏方法掩盖了词嵌入中的系统性性别偏见，却未将其消除。载于《*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（长篇与短篇

论文），第 609–614页，明尼苏达州明尼阿波利斯。计算语言学协会。（第155 页和第158页）

Ian J. Goodfellow、Jonathon Shlens 和 Christian Szegedy。 2015年。 解释并利用对抗样本。载于 第3届国际学习表征会议（*ICLR*）*2015*年，美国加利福尼亚州圣迭戈，*5*月 *7-9*日，*2015*年，会议论文集。（第 14页）

Edel Greevy 和 Alan F. Smeaton。 2004。 使用支持向量机对种族主义文本进行分类。载于 *SIGIR 2004*：第*27*届 *ACM* 信息检索研究与开发国际年会论文集，英国谢菲尔德，7月 *25-29*日，*2004*，页码 468–469。ACM。 （第12页）

Tao Gui、Qi Zhang、Haoran Huang、Minlong Peng 和 Xuanjing Huang。 2017年。 使用对抗神经网络为 Twitter 文本进行词性标注。载于《*2017*年自然语言处理实证方法会议论文集》，页码2411–2420，丹麦哥本哈根。计算语言学协会。（第 26页）

Isuru Gunasekara 和 Isar Nejadgholi。 2018年。 在线内容多标签毒性识别的标准文本分类实践综述。收录于 第*2*届在线辱骂性语言研讨会（*ALW2*）论文集，第 21–25页，比利时布鲁塞尔。计算语言学协会。 （第146页）

Odd Erik Gundersen 和 Sigbjørn Kjensmo。 2018年。 研究现状：人工智能领域的可复现性。 人工智能 *AAAI* 会议论文集，32（1）。 （第16页）

Mario Guzman-Silverio、Ángel Balderas-Paredes 和 Adrián Pastor López-Monroy。2020年。 用于墨西哥西班牙语攻击性检测的 Transformer 和数据增强。收录于 伊比利亚语种评测论坛论文集（*IberLEF2020*年），与 第*36*届西班牙自然语言处理学会会议（*SEPLN2020*年）同期举行，西班牙马拉加，*9*月 *23*日，*2020*年，第 2664 卷，收录于 *CEUR*研讨会论文集，第 293–302页。CEUR-WS.org。（第148页）

Matthias Hagen、Martin Potthast 和 Benno Stein。2017年。PAN 2017年作者身份混淆任务概述：重新审视安全性评估。载于 *CLEF 2017*年——评测论坛会议与实验室，爱尔兰都柏林，*9*月*11—14*，*2017*年，第1866 卷，收录于*CEUR*研讨会论文集。CEUR-WS.org。（第157页）

Xiaochuang Han 和 Yulia Tsvetkov。2020年。加固有害言论检测器以抵御隐晦有害内容。载于《*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第7732–7739页，线上。计算语言学协会。（第166页）

Laura Hanu、Unitary及其团队。 2020年。 Detoxify。GitHub。 (第133页)

Robert M. Haralick、K. Sam Shanmugam 和 Its'hak Dinstein。1973。 用于图像分类的纹理特征。 *IEEE Trans. Syst. Man Cybern.*， 3(6)：610–621。 (第5页)

Eszter Hargittai。 2010年。 数字原住民？"网络世代"成员的互联网技能和使用情况差异*。 《社会学探究》， 80(1)：92–113。 (第3页)

Muhammad Haroon、Muhammad Fareed Zaffar、Padmini Srinivasan 和 Zubair Shafiq。 2021年。 复仇者集成！提升作者身份混淆的可迁移性。 *CoRR*，abs/2109。07028。 (第158页)

Zaobo He、Zhipeng Cai 和 Jiguo Yu。 2018年。 针对社交网络数据、具有定制数据效用的潜在数据隐私保护。 *IEEE Trans.Veh.Technol.*， 67(1)：665–673。 (第8页)

Cynthia Van Hee、Els Lefever、Ben Verhoeven、Julie Mennes、Bart Des met、Guy De Pauw、Walter Daelemans 和 Véronique Hoste。 2015年。 网络欺凌事件的检测与细粒度分类。载于 自然语言处理最新进展，*RANLP2015*年， *7-9* 月，*2015*年，希萨尔，保加利亚，第 672–680。RANLP2015年 组委会 / ACL。 (第131)

Peter Henderson 和 Emma Brunskill。2018年。 从洪流中提炼信息：荟萃分析和系统综述在机器学习研究中的应用可能性。 *CoRR*，abs/1812.01074。 (第16)

Felix Hill、Roi Reichart 和 Anna Korhonen。 2015年。 SimLex-999：使用（真正的）相似度估计评估语义模型。 计算语言学， 41(4)：665–695。 (第57)

Sepp Hochreiter 和 Jürgen Schmidhuber。 1997。 长短期记忆。 *NeuralComput.*， 9（8）： 1735–1780。 （第 26 和103页）

Charles Hollingsworth。2012年。 利用基于依存关系的标注进行作者身份识别。载于 文本、语音与对话 - 第*15*届国际会议，*TSD 2012*，捷克共和国布尔诺，*9*月 *3-7*，*2012*年。会议论文集，第 7499 卷，选自 《计算机科学讲义》，第 314–319页。施普林格。 （第11页）

D. I. Holmes 和 R. S. Forsyth。 1995年。 重访《联邦党人文集》：作者归属判定的新方向。 《文学与语言计算》， 10（2）： 111–127。 （第11页）

David I.Holmes. 1998. 人文学术研究中文体计量学的演变。《文学与语言计算》，13（3）： 111–117。 （第22 和49页）

Matthew Honnibal、Ines Montani、Sofie Van Landeghem 和 Boyd Adriane。2020。spaCy：Python 中的工业级自然语言处理。 （第32、57和95页）

Hossein Hosseini、Sreeram Kannan、Baosen Zhang 和 RadhaPoovendran。2017。欺骗谷歌用于检测有害评论的 Perspective API。 *CoRR*，abs/1702。08138。（第127和148页）

Homa Hosseinmardi、Sabrina Arredondo Mattson、Rahat Ibn Rafiq、Richard Han、Qin Lv 和 Shivakant Mishra。 2015。 在 Instagram 社交网络上预测网络欺凌事件。 *CoRR*，abs/1508。06257。 （第82 和86页）

Dirk Hovy 和 Shannon L. Spruit。 2016年。 自然语言处理的社会影响。载于 《*54*届计算语言学协会年会论文集（第 *2*卷：短论文），第 591–598页，德国柏林。计算语言学协会。 （第10页）

David M. Howcroft、Anya Belz、Miruna-Adriana Clinciu、Dimitra Gkatzia、Sadid A. Hasan、Saad Mahamood、Simon Mille、Emiel van Miltenburg、Sashank Santhanam 和 Verena Rieser。 2020年。 人工评估中二十年的混乱：自然语言生成需要评估表和标准化定义。载于 《*13*届国际自然语言生成会议论文集，第 169–182页，爱尔兰都柏林。计算语言学协会。 （第160页）

Xiaolei Huang、Linzi Xing、Franck Dernoncourt 和 Michael J. Paul。 2020年。用于评估仇恨言论识别中人口统计偏差的多语言 Twitter 语料库和基线方法。载于《*12*届语言资源与评估会议论文集，第 1440–1448页，法国马赛。欧洲语言资源协会。（第 56、60、61、62、63、64和 65页）

Matthew Hutson。2018年。 人工智能面临可重复性危机。 《科学》， 359（6377）：725–726。 （第15 和16页）

Mai Ibrahim、Marwan Torki 和 Nagwa M. El-Makky。2018年。采用数据增强和深度学习的不平衡有害评论分类。载于第*17*届*IEEE*

国际机器学习与应用会议，*ICMLA2018*年，美国佛罗里达州奥兰多，*12*月*17*—*20*，*2018*年，第875–878页。IEEE。（第142和148页）

Nancy Ide、Tomaz Erjavec 和 Dan Tufis。2002年。利用平行语料库进行词义区分。载于*ACL-02*词义消歧研讨会论文集：近期成果与未来方向，第61–66页。计算语言学协会。（第29页）

Andrey Ignatov、Radu Timofte、Andrei Kulik、Seungsoo Yang、Ke Wang、Felix Baum、Max Wu、Lirong Xu 和 Luc Van Gool。2019年。人工智能基准测试：2019年智能手机深度学习综述。载于*2019*年*IEEE/CVF* 国际计算机视觉会议研讨会，*ICCV* 研讨会*2019*年，韩国首尔，*10*月*27-28*日，*2019*年，第3617–3635页。IEEE。（第5页）

Daphne Ippolito、Daniel Duckworth、Chris Callison-Burch 和 Douglas Eck。2020年。当人类受骗时，生成文本的自动检测最容易。载于*58*届计算语言学协会年会，第1808–1822页，线上。计算语言学协会。（第160页）

Jim Isaak 和 Mina J. Hanna。2018年。用户数据隐私：Facebook、剑桥分析与隐私保护。《*Computer*》，51（8）：56–59。（第4页）

Md Saroar Jahan 和 Mourad Oussalah。2021年。使用自然语言处理自动检测仇恨言论的系统综述。*CoRR*，abs/2106。00742。（第 146页）

Robin Jia 和 Percy Liang。2017年。用于评估阅读理解系统的对抗样本。载于《*2017*年 自然语言处理实证方法会议论文集》，第 2021–2031页，哥本哈根，丹麦。计算语言学协会。（第14页）

Jing Jiang 和 ChengXiang Zhai。2007年。针对统计分类器的两阶段领域适应方法。载于 《第十六届 *ACM* 信息与知识管理会议论文集，*CIKM 2007*年》，里斯本，葡萄牙，*11*月*6-10*，2007年，第401–410页。ACM。（第87页）

Di Jin、Zhijing Jin、Joey Tianyi Zhou 和 Peter Szolovits。2020年。BERT 真的稳健吗？针对文本分类和蕴涵的自然语言攻击的强基线。载于第*34*届*AAAI*人工智能会议，*AAAI 2020*年，第十届人工智能创新应用会议，*IAAI 2020*年，第十届

*AAAI*人工智能教育进展研讨会，*EAAI 2020*年，美国纽约州纽约市，*2*月 *7-12*，*2020*年，第 8018–8025页。AAAI出版社。（第48、51、57、58、62、129和156页）

Eun Seo Jo 和 Timnit Gebru。2020年。档案中的经验教训：机器学习中收集社会文化数据的策略。载于*FAT* '20*：公平性、问责制与透明度会议，西班牙巴塞罗那，*1*月*27-30*，*2020*年，第306–316。ACM。（第9页）

Vineet John、Lili Mou、Hareesh Bahuleyan 和 Olga Vechtomova。2019年。用于非平行文本风格迁移的解耦表示学习。载于论文集：*57*届计算语言学协会年会，第424–434页，意大利佛罗伦萨。计算语言学协会。（第155页）

Melvin Johnson、Mike Schuster、Quoc V. Le、Maxim Krikun、Yonghui Wu、Zhifeng Chen、Nikhil Thorat、Fernanda B. Viégas、Martin Wattenberg、Greg Corrado、Macduff Hughes 和 Jeffrey Dean。2017年。谷歌的多语言神经机器翻译系统：实现零样本翻译。计算语言学协会会刊，5：339–351。（第25 和29页）

Youngjun Joo 和 Inchon Hwang。 2019年。 社交媒体上的作者画像：使用多种特征的集成学习方法。载于 *CLEF 2019*年评测论坛会议与实验室工作札记，瑞士卢加诺，*9*月 *9-12*日，*2019*年，第 2380 卷，收录于 *CEUR*研讨会论文集。CEUR-WS.org。（第 50 页和58页）

Armand Joulin、Edouard Grave、Piotr Bojanowski 和 Tomas Mikolov。2017年。 高效文本分类的技巧包。载于《*15*届计算语言学协会欧洲分会会议论文集：第*2*卷，短论文，第 427–431页，西班牙瓦伦西亚。计算语言学协会。（第 32页）

Michał Jungiewicz 和 Aleksander Smywinski-Pohl。 2019年。 面向神经网络的文本数据增强：同义词与最大损失。 计算机科学， 20(1)。（第 143页和 148页）

Patrick Juola 和 Darren Vescovi。 2011年。 分析作者身份混淆的文体计量方法。载于 数字取证进展*VII - 7*届 *IFIP WG 11*。*9* 数字取证国际会议，美国佛罗里达州奥兰多，*1*月 *31* 日—*2*月 *2*日， *2011*年，修订版精选论文集，第 361 卷，选自 *IFIP*信息与通信技术进展，页码 115–125。施普林格。（第 25、50 和 63 页）

Roman Jurowetzki、Daniel S. Hain、Juan Mateos-Garcia 和 Konstantinos Stathoulopoulos。 2021年。 人工智能研究（者）的私有化：原因与潜在后果——从校企互动到公共研究领域的人才流失？ *CoRR*，abs/2102。01648。 （第5页）

Lisa Kaati、Enghin Omer、Nico Prucha 和 Amendra Shrestha。 2015年。 检测 Twitter 上的圣战主义传播者。载于 *IEEE* 数据挖掘国际会议研讨会，*ICDMW 2015* 年，美国新泽西州大西洋城，*11*月 *14-17*， *2015*年，页码 954–960。IEEE计算机学会。(第 120 页)

Jad Kabbara 和 Jackie Chi Kit Cheung。 2016年。 使用循环神经网络在自然语言生成系统中进行文体迁移。载于 语言处理中的艰难挑战研讨会：将早期成果扩展为稳健方法论文集，页码 43–47，德克萨斯州奥斯汀。计算语言学协会。 (第50页)

Gary Kacmarcik 和 Michael Gamon。 2006年。 对文档文体计量学进行混淆处理以保护作者匿名性。载于 *COLING/ACL*论文集 *2006*年 主会场海报环节，页码 444–451，澳大利亚悉尼。计算语言学协会。 (第 25和50页)

Ákos Kádár、Grzegorz Chrupała 和 Afra Alishahi。 2017年。 循环神经网络中语言形式与功能的表征。 计算语言学，43(4)：761–780。 (第52页)

Ákos Kádár、Grzegorz Chrupała 和 Afra Alishahi。2017年。 循环神经网络中语言形式与功能的表征。 计算语言学， 43(4)。 (第130页)

Laura Kann、Steve Kinchen、Shari L Shanklin、Katherine H Flint、Joseph Kawkins、William A Harris、Richard Lowry、Emily O’malley Olsen、Tim McManus、David Chyen、Lisa Whittle、Eboni Taylor、Zewditu Demissie、Nancy Brener、Jemekia Thornton、John Moore、Stephanie Zaza 和美国疾病控制与预防中心（CDC）。 2014年。青少年风险行为监测——美国，2013年。《发病率和死亡率周报》增刊， 63(4)：1–168。 (第76页)

Georgi Karadzhov、Tsvetomila Mihaylova、Yasen Kiprov、Georgi Georgiev、Ivan Koychev 和 Preslav Nakov。2017年。《平庸之道：采用中庸方法进行文体掩饰与作者身份混淆》。载于实验信息检索与多语言、多模态和交互，第173–185页，尚。施普林格国际出版。（第22、25和50页）

Flavius Kehr、Tobias Kowatsch、Daniel Wentzel 和 Elgar Fleisch。2015年。blissfully ignorant：一般隐私顾虑、一般机构信任及情感对隐私权衡的影响。信息系统期刊，25（6）：607–635。（第4和6页）

Yannik Keller、Jan Mackensen 和 Steffen Eger。2021年。BERT-defense：一种基于 BERT 的概率模型，用于对抗认知启发的正字法对抗攻击。载于 计算语言学协会论文集：*ACL-IJCNLP2021*年，第 1616–1629页，线上。计算语言学协会。（第159页）

Diane Kelly、Paul Kantor、Emile Morse、Jean Scholtz 和 Ying Sun。2006年。以用户为中心的交互式问答系统评估。载于 *HLT-NAACL* 交互式问答研讨会论文集 *2006*年，第 49–56页，纽约，纽约州，美国。计算语言学协会。（第12页）

Joo-Kyung Kim、Young-Bum Kim、Ruhi Sarikaya 和 Eric Fosler-Lussier。2017年。在没有跨语言资源的情况下进行跨语言迁移学习以实现词性标注。载于 论文集：*2017*年自然语言处理实证方法会议，第 2832–2838页，哥本哈根，丹麦。计算语言学协会。（第26页）

Seunghyun Kim、Afsaneh Razi、Gianluca Stringhini、Pamela J. Wisniewski 和 Munmun De Choudhury。 2021年。 以人为本的网络欺凌检测算法系统性文献综述。 *ACM*人机交互论文集， 5(CSCW2)。(第 161 页)

Yoon Kim。 2014年。 用于句子分类的卷积神经网络。载于论文集： *2014*年 自然语言处理经验方法会议（*EMNLP*），第 1746–1751 页，多哈，卡塔尔。计算语言学协会。 (第103页)

Diederik P. Kingma 和 Jimmy Ba。2015年。 Adam：一种随机优化方法。载于 第 *3*届国际学习表征会议（*ICLR*）*2015*年，美国加利福尼亚州圣迭戈，*5*月 *7*—*9*日，*2015*年，会议论文集。 (第30页)

Svetlana Kiritchenko 和 Isar Nejadgholi。 2020年。 迈向在线辱骂内容检测中的设计伦理。 *CoRR*，abs/2010年。14952。 （第166页）

Bennett Kleinberg、Toby Davies 和 Maximilian Mozes。 2022。 Textwash——自动化开源文本匿名化。 *CoRR*，abs/2208。13081。 （第11页）

April Kontostathis、Lynne Edwards 和 Amanda Leatherman。 2010年。 文本挖掘与网络犯罪，第 8章。John Wiley & Sons, Ltd.（第83页）

April Kontostathis、Kelly Reynolds、Andy Garron 和 Lynne Edwards。2013年。检测网络欺凌：查询词和技术。收录于网络科学*2013*年（与 *ECRC* 联合举办），*WebSci '13*，法国巴黎，*5*月*2-4*，*2013*年，第195–204页。ACM。（第82和83页）

Corina Koolen 和 Andreas van Cranenburgh。2017年。这些并非你所寻找的刻板印象：作者性别归因中的偏见与公平性。收录于第一届 *ACL* 自然语言处理伦理研讨会论文集，第12–22页，西班牙瓦伦西亚。计算语言学协会。（第9页）

Moshe Koppel、Navot Akiva、Eli Alshech 和 Kfir Bar。2009。按意识形态和组织归属自动分类文档。收录于*IEEE* 情报与安全信息学国际会议（*ISI*）*2009*，美国得克萨斯州达拉斯，*6*月*8-11*，*2009*，会议录，第176–178页。IEEE。（第9页）

Moshe Koppel、Shlomo Argamon 和 Anat Rachel Shimoni。2002年。根据作者性别自动对书面文本进行分类。*Lit. LinguisticComput.*，17(4)：401–412。(第 22、23 和第46页)

Moshe Koppel 和 Jonathan Schler。2004。将作者身份验证作为单类分类问题。载于 机器学习：第二十一届国际会议论文集（*ICML 2004*），加拿大阿尔伯塔省班夫，*7*月*4-8*，*2004*，第 69 卷，收录于 *ACM* 国际会议论文集系列。ACM。(第49页)

RobinKowalski、Susan Limber 和 Patricia Agatston。2012年。网络欺凌：数字时代的欺凌。《美国精神病学杂志》- *AMER J PSYCHIAT*，165：296。(第 86页)

Varun Kumar、Ashutosh Choudhary 和 Eunah Cho。2020年。使用预训练 Transformer模型进行数据增强。*CoRR*，abs/2003。02245。(第136页)

Alexey Kurakin、Ian J. Goodfellow 和 Samy Bengio。2017年a. 现实世界中的对抗样本。载于 第*5*届国际学习表征会议（*ICLR*）*2017*年，法国土伦，*4*月*24—26*，*2017*年研讨会分会场论文集。OpenReview.net。（第14页）

Alexey Kurakin、Ian J. Goodfellow 和 Samy Bengio。 2017年b. 大规模对抗性机器学习。载于第五届国际学习表征会议（*ICLR*）

*2017*年，法国土伦，*4*月*24-26*，*2017*年，会议论文集。OpenReview.net。（第14页）

Keita Kurita、Anna Belova 和 Antonios Anastasopoulos。2019年。迈向稳健的有害内容分类。*CoRR*，abs/1912。06872。（第147页）

Matt Kusner、Yu Sun、Nicholas Kolkin 和 Kilian Weinberger。2015年。从词嵌入到文档距离。载于国际机器学习会议，第957–966页。（第33页）

Marc-André Larochelle 和 Richard Khoury。2020年。网络欺凌检测的泛化。载于*IEEE/ACM* 社会网络分析与挖掘进展国际会议，*ASONAM 2020*年，荷兰海牙，*12*月*7-10*，*2020*年，第296–300页。IEEE。（第146页）

Alon Lavie 和 Michael J. Denkowski。2009。用于机器翻译自动评估的 METEOR 指标。*Mach. Transl.*，23（2-3）：105–115。（第61页）

Jochen L. Leidner 和 Vassilis Plachouras。 2017年。 以设计确保伦理：自然语言处理的伦理最佳实践。载于 首届 *ACL*自然语言处理伦理研讨会论文集，第 30–40页，西班牙瓦伦西亚。计算语言学协会。 （第12页）

Joe Lemley、Shabab Bazrafkan 和 Peter Corcoran。 2017年。 面向消费设备和服务的深度学习：推动机器学习、人工智能和计算机视觉的极限。 *IEEE*消费电子杂志，第 6（2）：48–56。 （第5页）

Amanda Lenhart、Mary Madden、Aaron Smith、Kristen Purcell、Kathryn Zickuhr 和 LeeRainie。 2011年。 青少年、社交网站上的善意与残酷：美国青少年如何应对这个新世界。 皮尤研究中心。 （第 75和 76页）

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mo- hamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。 BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于 论文集 *58*计算语言学协会年会，第 7871–7880页，线上。计算语言学协会。 （第133页）

Chuanrong Li、Lin Shengshuo、Zeyu Liu、Xinyi Wu、Xuhui Zhou 和 Shane Steinert- Threlkeld。 2020年。 语言学指导的变换（LIT）：一种自动生成对比集的方法。载于 第三届 *BlackboxNLP* 研讨会论文集

关于自然语言处理神经网络分析与解释，*BlackboxNLP@EMNLP 2020*年，线上，*11*月*2020*年，第 126–135页。计算语言学协会。（第 146页）

Jinfeng Li、Shouling Ji、Tianyu Du、Bo Li 和 Ting Wang。2019年。Textbugger：针对现实应用生成对抗文本。载于第*26*届网络与分布式系统安全年度研讨会（*NDSS*）*2019*年，美国加利福尼亚州圣迭戈，*2*月*24*日—*27*日，*2019*年。互联网协会。（第127页）

Jiwei Li、Minh-Thang Luong 和 Dan Jurafsky。2015年。面向段落和文档的分层神经自编码器。载于《第*53*届计算语言学协会年会暨第*7*届亚洲自然语言处理联合会国际自然语言处理联合会议，*ACL2015*年，*7*月*26*日—*31*日，*2015*年，中国北京，第*1*卷：长篇论文，页码1106–1115。计算语言学协会。（第26页）

Jiwei Li、Will Monroe 和 Dan Jurafsky。 2016年。 通过表征擦除理解神经网络。*CoRR*，abs/1612。08220。 （第 14 页）

Shoushan Li、Shengfeng Ju、Guodong Zhou 和 Xiaojun Li。 2012年。 面向不平衡情感分类的主动学习。载于 论文集：*2012*年 自然语言处理实证方法会议与计算自然语言学习联合会议，第 139–148页，韩国济州岛。计算语言学协会。 （第87页）

Yitong Li、Timothy Baldwin 和 Trevor Cohn。 2018年。 迈向稳健且保护隐私的文本表示。载于 论文集：第*56*届计算语言学协会年会（第*2*卷：短论文），第 25–30页，澳大利亚墨尔本。计算语言学协会。 （第46页）

Bin Liang、Hongcheng Li、Miaoqiang Su、Pan Bian、Xirong Li 和 Wenchang Shi。2018a。 深度文本分类很容易被欺骗。载于 第二十七届国际人工智能联合会议论文集，*IJCAI 2018*年，*7*月*13-19*，*2018*年，瑞典斯德哥尔摩，第 4208–4215页。ijcai.org。 （第47页）

Yi Liang、Zhipeng Cai、Jiguo Yu、Qilong Han 和 YingshuLi。2018年b. 基于深度学习利用智能设备中的嵌入式传感器推断私人信息。*IEEE Netw.*，32（4）：8–14。（第8页）

Zachary C. Lipton 和 Jacob Steinhardt。 2019年。 面向实践的研究：机器学习学术研究中的令人担忧的趋势。 《*ACM* 通讯》，62(6)：45–53。 (第16页)

Pengfei Liu、Xipeng Qiu 和 Xuanjing Huang。 2017年。 用于文本分类的对抗式多任务学习。载于 论文集：第*55*届计算语言学协会年会（第 *1*卷：长篇论文），第1–10页，温哥华，加拿大。计算语言学协会。 (第26页)

Shengcai Liu、Ning Lu、Cheng Chen 和 Ke Tang。2022。针对词级对抗文本攻击的高效组合优化。*IEEE ACM*《音频、语音与语言处理汇刊》，30：98–111。（第158页）

Sonia Livingstone、Leslie Haddon、Anke Görzig 和 Kjartan Ólafsson。2011年。EU Kids Online II：最终报告 2011年。技术报告，EU Kids Online。（第75页）

David G. Lowe。1999年。基于局部尺度不变特征的目标识别。载于国际计算机视觉会议论文集，希腊科孚岛克基拉，*9*月*20-25*，*1999*年，页码 1150–1157。IEEE计算机学会。（第5页）

Christoph Lutz 和 Christian Pieter Hoffmann。2017年。在线参与的阴暗面：探讨非参与、被动参与和负面参与。《信息、传播与社会》，20（6）：876–897。（第6页）

Christoph Lutz、Christian Pieter Hoffmann 和 Giulia Ranzini。 2020年。 数据资本主义与用户：对德国隐私犬儒主义的探索。 《新媒体与社会》，22(7)。 (第6页)

Lingjuan Lyu、Xuanli He 和 Yitong Li。2020年。 面向自然语言处理的差分隐私表示：形式化保证以及关于隐私与公平性的实证研究。载于计算语言学协会论文集：*EMNLP2020*年，第2355–2365页，线上。计算语言学协会。 (第11页)

Jianqiang Ma。 2014年。 汉语词结构中句法范畴的自动细化。载于 第九届国际语言资源与评测会议论文集（*LREC'14*），第4087–4092页，冰岛雷克雅未克。欧洲语言资源协会（ELRA）。 (第87页)

Aman Madaan、Amrith Setlur、Tanmay Parekh、Barnabas Poczos、Graham Neubig、Yiming Yang、Ruslan Salakhutdinov、Alan W Black 和 Shrimai Prabhumoye。2020年。 礼貌风格迁移：一种标注生成方法。载于 《*58*届计算语言学协会年会论文集，页码 1869–1881，线上。计算语言学协会。 (第154页)

Mary Madden、Amanda Lenhart 和 Sandra Cortesi。 2013年。 青少年、社交媒体与隐私。技术报告，Pew Internet。 (第75页)

Nitin Madnani 和 Anastassia Loukina。 2020年。 以用户为中心且稳健的自然语言处理开源软件：开发和维护 RSMTool 的经验。载于 第二届自然语言处理开源软件研讨会论文集（*NLP-OSS*），第 141–146页，线上。计算语言学协会。 （第12页）

Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 AdrianVladu。 2018年。 迈向能够抵御对抗攻击的深度学习模型。载于 第 *6* 届国际学习表征会议（*ICLR*）*2018*，加拿大不列颠哥伦比亚省温哥华，*4* 月 *30* 日至 *5* 月 *3*日，*2018*年，会议论文集。OpenReview.net。 （第14页）

Kosisochukwu Madukwe、Xiaoying Gao 和 Bing Xue。 2020年。 我们信赖数据：对仇恨言论检测数据集的批判性分析。载于 第四届在线虐待与伤害研讨会论文集，第 150–161页，线上。计算语言学协会。 （第 126页和第 146页）

Asad Mahmood、Zubair Shafiq 和 Padmini Srinivasan。2020年。女孩有名字：检测作者身份混淆。载于《*58*届计算语言学协会年会论文集，第2235–2245页，线上。计算语言学协会。（第67页和第158页）

Karl M. Manheim 和 Lyric Kaplan。2019年。人工智能：对隐私和民主的风险。《耶鲁法律与技术期刊》*106*。（第5页）

MuharramMansoorizadeh、TaherRahgooy、Mohammad Aminian 和 MehdyEskan- dari。2016年。使用WordNet和语言模型进行作者身份混淆。载于 *CLEF* 工作笔记*2016*年-评测论坛会议与实验室，埃武拉，葡萄牙，*5-8 9*月，*2016*年，第1609 卷，收录于*CEUR*研讨会论文集，第939–946页。CEUR-WS.org。（第50页）

Ilia Markov, Nikola Ljubeši´c、Darja Fišer 和 Walter Daelemans。2021年。探索用于多语言跨领域仇恨言论检测的文体特征与基于情感的特征。载于第十一届主观性、情感与社交媒体分析计算方法研讨会论文集，第 149–159 页，线上。计算语言学协会。（第146页）

Kimberly L. Mason. 2008年。 网络欺凌：面向学校人员的初步评估。《学校心理学》，45(4)：323–348。 (第126页)

José Mateo 和 Francisco Ramos Yus. 2013年。 迈向跨文化辱骂语用分类体系。《语言攻击与冲突杂志》，1(1)：87–114。 (第163页)

Alex Mathai、Shreya Khare、Srikanth Tamilselvam 和 Senthil Mani。2020年。利用深度网络引导的多目标遗传优化，对文本分类器实施对抗性黑盒攻击。 *CoRR*，abs/2011年。03901。 （第140页）

Robert A. J. Matthews 和 Thomas V. N. Merriam。 1993年。 文体计量学中的神经计算 I：对莎士比亚和 Fletcher 作品的应用。 《文学与语言计算》， 8（4）：203–209。 （第49页）

David McClosky、Eugene Charniak 和 Mark Johnson。 2006年。 句法分析器适配的重排序与自训练。载于 《第*21*届国际计算语言学会议暨第*44*届计算语言学协会年会论文集》，第 337–344页，澳大利亚悉尼。计算语言学协会。 （第87页）

Gábor Melis、Chris Dyer 和 Phil Blunsom。 2018年。 神经语言模型评估的最新进展。载于 第*6*届国际学习表征会议，*ICLR 2018*年，加拿大不列颠哥伦比亚省温哥华，*4*月 *30*日 至*5*月 *3*日， *2018*年，会议论文集。OpenReview.net。 （第16页）

Thomas V. N. Merriam 和 Robert A. J. Matthews。 1994年。 文体计量学中的神经计算 II：莎士比亚和马洛作品的应用。 《文学与语言计算》， 9（1）： 1–6。 （第49页）

Tomás Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。 2013年a。 向量空间中词表示的高效估计。载于 第*1*届国际学习表征会议，*ICLR 2013*年，美国亚利桑那州斯科茨代尔，*5*月 *2*日至*4*日， *2013*年，研讨会分会场论文集。 （第101页）

Tomás Mikolov、Ilya Sutskever、Kai Chen、Gregory S. Corrado 和 Jeffrey Dean。2013年b. 词语和短语的分布式表示及其组合性。载于神经信息处理系统进展 *26*：第*27*届神经信息处理系统年会*2013*年。会议论文集，会议于*12*月*5*日至*8*日，在美国内华达州太浩湖举行，第3111–3119页。（第101页）

George A. Miller。 1995年。 WordNet：英语词汇数据库。 《*ACM* 通讯》， 38（11）： 39–41。 （第134页）

Andrea Millwood-Hargreave。 2000年。 删除脏话? 广告标准管理局、英国广播公司、广播标准委员会和独立电视委员会联合开展的研究。 *ASA*、*BBC*、*BSC* 和 *ITC*，伦敦。 （第84页）

Pushkar Mishra、HelenYannakoudakis 和 Ekaterina Shutova。 2019年。 应对网络滥用：自动化滥用检测方法综述。 *CoRR*，abs/1908。06024。（第 146 页）

Alan Mislove、Bimal Viswanath、P. Krishna Gummadi 和 Peter Druschel。 2010年。 你认识谁，你就是谁：推断在线社交网络中的用户画像。发表于 第三届网络搜索与网络数据挖掘国际会议论文集，*WSDM2010*，美国纽约州纽约市，*2*月 *4*—*6*，*2010*年，第 251–260页。ACM。 （第8页）

Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡片。发表于 公平性、问责制与透明度会议论文集，*FAT* 2019*，美国佐治亚州亚特兰大，*1*月 *29*—*31*，*2019*年，第 220–229页。ACM。 （第 7 和16页）

Miljana Mladenovic、Vera Osmjanski 和 Stasa Vujicic Stankovic。2021年。网络攻击行为、网络欺凌与网络诱骗：综述与研究挑战。《*ACM*计算调查》，54(1)：1：1–1：42。(第146页)

John X. Morris、Eli Lifland、Jin Yong Yoo、Jake Grigsby、Di Jin 和 Yanjun Qi。2020年。Textattack：一个用于自然语言处理中的对抗攻击、数据增强和对抗训练的框架。载于 《*2020*年自然语言处理实证方法会议：系统演示，*EMNLP2020*年-演示，线上，*11*月*16-20*，*2020*年，第 119–126。计算语言学协会。(第164页)

Aimée Hope Morrison。2009。一个不可能的未来：John Perry Barlow 的《网络空间独立宣言》。《新媒体与社会》，11(1-2)：53–71。(第4页)

A. Q. Morton. 1965. 希腊散文作者身份. 《皇家统计学会杂志》. *Series A (General)*, 128(2):169–233. (p.11)

Frederick Mosteller 和 David L. Wallace。 1963年。 作者身份问题中的推断。《美国统计协会会刊》， 58(302)：275–309。 (第49页)

Giovane C. M. Moura、Sebastian Castro、Wes Hardaker、Maarten Wullink 和 Cristian Hesselman。2020 年。 互联网云化：域名系统流量正变得多么集中化？ 载于 *IMC '20*：*ACM互联网测量会议*，线上活动，美国，*10*月 *27-29*， *2020*年，第 42–49页。ACM。 (第4页)

Maximilian Mozes、Max Bartolo、Pontus Stenetorp、Bennett Kleinberg 和 Lewis Griffin。2021年a. 对比人工与机器生成的用于文本分类的词级对抗样例。载于《*2021*年自然语言处理实证方法会议论文集，第8258–8270页，线上举行，地点为多米尼加共和国蓬塔卡纳。计算语言学协会。（第160页）

Maximilian Mozes、Pontus Stenetorp、Bennett Kleinberg 和 Lewis Griffin。2021年b.用于检测文本对抗样例的频率引导词语替换。载于《*16*届计算语言学协会欧洲分会会议：主会论文集，第171–186页，线上举行。计算语言学协会。（第158页）

Nikola Mrkši´c、Diarmuid Ó Séaghdha、Blaise Thomson、Milica Gaši´c、Lina M. Rojas-Barahona、Pei-Hao Su、David Vandyke、Tsung-Hsien Wen 和 Steve Young。2016年。将词向量调整以符合语言学约束。载于《*2016*年计算语言学协会北美分会会议：人类语言技术论文集，第142–148页，圣迭戈，加利福尼亚州。计算语言学协会。（第57页）

Marcus R. Munafò、Brian A. Nosek、Dorothy V. M. Bishop、Katherine S. Button、ChristopherD.Chambers、NathaliePerciedu Sert、Uri Simonsohn、Eric-JanWa- genmakers、Jennifer J. Ware 和 John P. A. Ioannidis。2017年。可重复科学宣言。《自然·人类行为》，1(1)：0021。(第16页)

Amgad Muneer 和 Suliman Mohamed Fati。2020年。Twitter 上网络欺凌检测机器学习技术的比较分析。*FutureInternet*，12(11)：187。(第146页)

Dan Murray 和 Kevan Durrell。1999年。推断匿名互联网用户的人口统计属性。载于 网络使用分析与用户画像国际研讨会，页码 7–20。施普林格。(第8页)

Vinita Nahar、Xue Li 和 Chaoyi Pang。 2013年。 一种有效的网络欺凌检测方法。《信息科学与管理工程通信》，3(五月)：238–247。 (第 82 和 86 页)

Preslav Nakov、Vibha Nayak、Kyle Dent、Ameya Bhatawdekar、Sheikh Muham- mad Sarwar、Momchil Hardalov、Yoan Dinkov、Dimitrina Zlatkova、Guillaume Bouchard 和 Isabelle Augenstein。 2021年。 检测在线平台上的辱骂性语言：一项批判性分析。 *CoRR*，abs/2103。00153。 (第 126 和 127页)

Hyeonseob Nam 和 Hyo-Eun Kim。2018年。用于自适应风格不变神经网络的批次-实例归一化。载于神经信息处理系统进展*31*：神经信息处理系统年会 *2018*年，*NeurIPS2018*年，*12*月 *3-8*，*2018*年，加拿大蒙特利尔，第 2563–2572。(第154页)

Tempestt J. Neal、Kalaivani Sundararajan、Aneez Fatima、Yiming Yan、Yingfei Xiang 和 DamonL. Woodard。2018年。文体计量技术及其应用综述。《*ACM*计算调查》，50(6)：86：1–86：36。(第 22页和第49页)

Isar Nejadgholi 和 Svetlana Kiritchenko。2020年。关于在线辱骂自动检测中的跨数据集泛化。载于第四届在线虐待与伤害研讨会论文集，第 173–183，线上。计算语言学协会。(第147页)

Taisei Nitta、Fumito Masui、Michal Ptaszynski、Yasutomo Kimura、Rafal Rzepka 和 Kenji Araki。2013年。基于类别相关性最大化检测非正式学校网站中的网络欺凌内容。载于第六届自然语言处理国际联合会议论文集，第579–586页，名古屋，日本。亚洲自然语言处理联合会。（第147页）

Noswearing.com。2016年。脏话列表和脏话过滤器。 http://www.noswearing.com。最后访问于15/03/2016。（第84页）

Florian Nothdurft、Gregor Behnke、Pascal Bercher、Susanne Biundo 和 Wolfgang Minker。2015年。以用户为中心的对话系统与人工智能规划的相互作用。载于《第*16*届话语与对话特别兴趣小组年会论文集，第344–353页，捷克共和国布拉格。计算语言学协会。（第12页）

Jekaterina Novikova、Ondˇrej Dušek、Amanda Cercas Curry 和 Verena Rieser。2017年。为什么我们需要新的自然语言生成评估指标。见论文集*2017*年自然语言处理实证方法会议，第2241–2252，丹麦哥本哈根。计算语言学协会。（第50）

Myle Ott、Yejin Choi、Claire Cardie 和 Jeffrey T. Hancock。2011年。 凭想象力找到欺骗性观点垃圾信息。载于《*49*届年会论文集

计算语言学协会：人类语言技术会议，页码 309–319，美国俄勒冈州波特兰。计算语言学协会。（第 9页）

Jahna Otterbacher。 2010年。 推断电影评论者的性别：利用写作风格、内容和元数据。见 《*19th ACM*信息与知识管理会议，*CIKM2010*年，加拿大安大略省多伦多，*10*月 *26-30*，*2010*年，第 369–378页。ACM。 （第11页）

Nedjma Ousidhoum、Xinran Zhao、Tianqing Fang、Yangqiu Song 和 Dit-Yan Yeung。2021年。 探究大型预训练语言模型中的有害内容。见 《*59th*计算语言学协会年会与*11th*国际自然语言处理联合会议，*ACL/IJCNLP2021*年，（第*1*卷：长篇论文），线上活动，*8*月 *1-6*，*2021*年，第 4262–4274页。计算语言学协会。（第139页）

Lin Pan、Chung-Wei Hang、Avirup Sil、Saloni Potdar 和 Mo Yu。2021年。通过对比对抗训练改进文本分类。*CoRR*，abs/2107。10137。（第146页）

Sinno Jialin Pan、Xiaochuan Ni、Jian-Tao Sun、Qiang Yang 和 Zheng Chen。2010年。通过谱特征对齐进行跨领域情感分类。载于第*19*届国际万维网大会（*WWW*）*2010*年，美国北卡罗来纳州罗利，*4*月*26*—*30*日，*2010*年，第751–760页。ACM。（第87页）

Xudong Pan、Mi Zhang、Shouling Ji 和 Min Yang。2020年。通用语言模型的隐私风险。载于*2020*年*IEEE*安全与隐私研讨会，*SP2020*年，美国加利福尼亚州旧金山，*5*月*18*—*21*日，*2020*年，第1314–1331页。IEEE。（第7页）

Nicolas Papernot、Patrick D. McDaniel 和 Ian J. Goodfellow。2016年a. 机器学习中的可迁移性：从现象到利用对抗样本实施黑盒攻击。*CoRR*，abs/1605。07277。(第47页)

Nicolas Papernot、Patrick D. McDaniel、Somesh Jha、Matt Fredrikson、Z. Berkay Celik 和 Ananthram Swami。2016年b. 对抗环境下深度学习的局限性。发表于 *IEEE*欧洲安全与隐私研讨会，*EuroS&P2016*年，德国萨尔布吕肯，*3*月*21-24*，*2016*年，第 372–387页。IEEE。(第14页)

Nicolas Papernot、Patrick D. McDaniel、Ananthram Swami 和 Richard E. Harang。2016年C. 为循环神经网络构造对抗性输入序列。载于*2016*年 *IEEE* 军事通信会议（*MILCOM*）*2016*年，美国马里兰州巴尔的摩，*11*月 *1-3*，*2016*年，第 49–54页。IEEE。 （第14页）

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于《*40*届计算语言学协会年会论文集，第311–318页，美国宾夕法尼亚州费城。计算语言学协会。（第 23 和33页）

Francisco Manuel Rangel Pardo、Paolo Rosso、Ben Verhoeven、Walter Daelemans、Martin Potthast 和 Benno Stein。2016年。PAN 第 4届作者画像任务概述2016年：跨体裁评估。载于 *CLEF* 工作记录*2016*年 —评测论坛会议与实验室，葡萄牙埃武拉，*5—8* 月，*2016*年，第 1609 卷，*CEUR*研讨会论文集，第 750–784页。CEUR-WS.org。（第 23 页和第 48 页）

Tatiana Passali、Alexios Gidiotis、Efstathios Chatzikyriakidis 和 Grigorios Tsoumakas。2021年。迈向以人为中心的摘要生成：财经新闻案例研究。载于首届人机交互与自然语言处理交叉研讨会论文集，第 21–27页，线上。计算语言学协会。（第12页）

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Köpf、Edward Z. Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。2019年。 PyTorch：一种命令式风格的高性能深度学习库。载于 神经信息处理系统进展 *32*：神经信息处理系统年会 *2019*年，*NeurIPS2019*年，*12*月*8-14*，*2019*年，加拿大不列颠哥伦比亚省温哥华，页码 8024–8035。（第 26页和 58页）

Sayanta Paul 和 Sriparna Saha。 2020年。 CyberBERT：用于识别网络欺凌的 BERT。 多媒体系统。 00000。（第 147页）

John Pavlopoulos、Jeffrey Sorensen、Lucas Dixon、Nithum Thain 和 Ion Androutsopoulos。 2020年。 毒性检测：语境真的重要吗？ 载于《第*58*届计算语言学协会年会论文集》，第 4296–4305页，线上。计算语言学协会。 (第165页)

Fabian Pedregosa、Gaël Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vin- cent Dubourg、Jake VanderPlas、Alexandre Passos、David Cournapeau、Matthieu Brucher、Matthieu Perrot 和 Edouard Duchesnay。 2011年。 Scikit-learn：Python 中的机器学习。 机器学习研究期刊， 12:2825–2830。 (pp. 57， 98，和 135)

Claudia Peersman、Walter Daelemans 和 Leona Van Vaerenbergh。2011年。预测在线社交网络中的年龄和性别。发表于 论文集：第*3*届国际 *CIKM* 用户生成内容搜索与挖掘研讨会，*SMUC 2011*年，英国格拉斯哥，*10*月 *28*日， *2011*年，页码 37–44。ACM。 (第11页)

Marco Pennacchiotti 和 Ana-Maria Popescu。 2011年。 民主党人、共和党人和星巴克爱好者：Twitter 用户分类。载于 第 *17* 届 *ACM SIGKDD*国际知识发现与数据挖掘会议论文集，圣迭戈，加利福尼亚州，美国，*8*月 *21-24*日， *2011*年，第 430–438页。ACM。 （第9页）

Jeffrey Pennington、Richard Socher 和 Christopher Manning。2014年。GloVe：词表示全局向量。载于 《*2014*年 自然语言处理实证方法会议（*EMNLP*）论文集，第 1532–1543页，多哈，卡塔尔。计算语言学协会。 （第100页）

Joelle Pineau、G Fried、RN Ke 和 H Larochelle。 2018年。ICLR2018年 可复现性挑战。载于 *ICML*机器学习可复现性研讨会。 （第16页）

Barbara Plank 和 Dirk Hovy。 2015年。 Twitter 上的人格特质——或者——如何在一周内完成1,500 项人格测试。载于 《*6*届主观性、情感与社交媒体分析计算方法研讨会论文集，第 92–98页，葡萄牙里斯本。计算语言学协会。 （第 9、22和46页）

Martin Potthast、Matthias Hagen 和 Benno Stein。 2016年。 作者身份混淆：攻击作者身份验证领域的现有技术。载于 *CLEF 2016*年—评测论坛会议与实验室工作论文，葡萄牙埃武拉， *5-8*月，*2016*年，第 1609 卷，收录于 *CEUR*研讨会论文集，第 716–749页。CEUR-WS.org。 （第22、50、62、67、157和 160页）

Martin Potthast、Paolo Rosso、Efstathios Stamatatos 和 Benno Stein。 2019年。PAN 数字文本取证领域十年的共享任务。载于 信息检索进展，第 291–300页，尚。施普林格国际出版。 （第15页）

Martin Potthast、Felix Schremmer、Matthias Hagen 和 Benno Stein。 2018年。PAN 2018 年作者身份混淆任务概述 2018 年：一种测量安全性的新方法。载于 *CLEF2018*年 - 评测论坛会议与实验室工作札记，法国阿维尼翁，*9*月 *10-14*， *2018*年，第 2125 卷，收录于 *CEUR*研讨会论文集。CEUR-WS.org。 (第160页)

Carmel Privitera 和 Marilyn Anne Campbell。2009。 网络欺凌：职场欺凌的新面貌？ *Cyberpsychol Behav*， 12(4)：395–400。 (p.77)

Reid Pryzant、Youngjoo Chung 和 Dan Jurafsky。 2017年。 根据产品描述的语言预测销售额。载于 *SIGIR* 论文集 *2017*年电子商务研讨会，与*40*届 *ACM* 信息检索研究与开发国际会议同期举行，*eCOM@SIGIR 2017*年，日本东京，*8*月*11*， *2017*年，第 2311 卷，收录于 *CEUR*研讨会论文集。CEUR-WS.org。 (p.26)

Husam Quteineh、Spyridon Samothrakis 和 Richard Sutcliffe。2020年。面向小型数据集高效主动学习的文本数据增强。载于《*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第7400–7410页，线上。计算语言学协会。（第134页和第148页）

Bahar Radfar、Karthik Shivaram 和 Aron Culotta。2020年。社交语境下有害语言变异的特征分析。载于《第十四届 *AAAI* 国际网络与社交媒体会议（*ICWSM*）*2020*年》，线上举行，原定地点：美国佐治亚州亚特兰大，*6*月*8–11*日，*2020*年，第959–963页。AAAI出版社。（第165页）

Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。2018年。通过生成式预训练提升语言理解能力。*OpenAI* 博客。（第148页）

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等人。2019年。语言模型是无监督式多任务学习者。*OpenAI* 博客，1(8)：9。(第134 页和148页)

Delip Rao、David Yarowsky、Abhishek Shreevats 和 Manaswi Gupta。2010年。对 Twitter 中的潜在用户属性进行分类。载于 《第二届用户生成内容搜索与挖掘国际研讨会论文集，*SMUC@CIKM 2010*年，加拿大安大略省多伦多，*2010*年*10*月*30*日，*2010*年，第37–44页》。ACM。(第9、23和48页)

Josyula R. Rao 和 Pankaj Rohatgi。2000年。假名性真的能保障隐私吗？ 载于 第九届 *USENIX* 安全研讨会，*14-17*日，*2000*年。USENIX 协会。(第25页和46页)

Victor Raskin、Sergei Nirenburg、Mikhail J. Atallah、Christian F. Hempelmann 和 Katrina E. Triezenberg。 2002年。 为什么自然语言处理应迈入IAS。载于 *COLING-02*：计算语言学路线图。 （第11页）

Keith Rayner、Sarah J White、Rebecca L Johnson 和 Simon P Liversedge。2006年。阅读字母顺序打乱的单词：这是有代价的。*PsycholSci"*，17(3)：192–193。(第54页)

Benjamin Recht、Rebecca Roelofs、Ludwig Schmidt 和 Vaishaal Shankar。2019年。ImageNet 分类器能否泛化到 ImageNet? 载于《*36*届国际机器学习会议论文集（*ICML*）*2019*年，*9-15*日，*2019*年，美国加利福尼亚州长滩，第97 卷，收录于机器学习研究论文集，页码5389–5400。PMLR。(第16页)

Sravana Reddy 和 Kevin Knight。2016年。社交媒体写作中的性别隐匿。载于首届自然语言处理与计算社会科学研讨会论文集，第17–26页，得克萨斯州奥斯汀。计算语言学协会。（第50页）

Radim ˇReh ̊uˇrek 和 Petr Sojka。2010年。面向大型语料库主题建模的软件框架。载于*LREC 2010*年自然语言处理框架的新挑战研讨会论文集，第45–50页，马耳他瓦莱塔。ELRA。（第101页）

Kelly Reynolds、April Kontostathis 和 Lynne Edwards。2011年a。使用机器学习检测网络欺凌。载于论文集第*10*届国际机器学习与应用会议及研讨会，第2卷，第241–244页。（第82、83、93、94和108页）

Kelly Reynolds、April Kontostathis 和 Lynne Edwards。2011年b. 使用机器学习检测网络欺凌。载于 第*10*届机器学习与应用国际会议及研讨会，*ICMLA2011*年，美国夏威夷州檀香山，*12*月 *18-21*，*2011*年。第 2卷：特别会议与研讨会，第 241–244 页。IEEE计算机学会。（第 131页和 147页）

Verena Rieser 和 Oliver Lemon。 2008年。 为对话系统优化自动学习和评估以用户为中心的目标函数。载于 第六届国际语言资源与评测会议（*LREC'08*），摩洛哥马拉喀什。欧洲语言资源协会（ELRA）。（第 12页）

Simone Robers、Anlan Zhang 和 Rachel E Morgan。 2015年。校园犯罪与安全指标： 2014年。 国家教育统计中心、美国教育部、司法统计局、司法项目办公室、美国司法部。NCES 2015-072/NCJ 248036。（第 76页）

H. Rosa、N. Pereira、R. Ribeiro、P.C. Ferreira、J.P. Carvalho、S. Oliveira、L. Coheur、P. Paulino、A.M. Veiga Simão 和 I. Trancoso。2019年a. 自动网络欺凌检测：系统综述。《人类行为中的计算机》，93：333–345。(第89、102和126页)

Hugo Rosa、João Paulo Carvalho、Pável Calado、Bruno Martins、Ricardo Ribeiro 和 Luísa Coheur。2018年a. 使用模糊指纹检测社交网络中的网络欺凌。发表于*2018*年 *IEEE* 国际模糊系统会议（*FUZZ-IEEE*）*2018*年，巴西里约热内卢，7月 *8-13*日，*2018*年，第 1–7页。IEEE。 (第82 和84页)

Hugo Rosa、David Martins de Matos、Ricardo Ribeiro、Luísa Coheur 和 João Paulo Carvalho。 2018年b. 深入探讨社交网络中的网络欺凌检测。载于*2018*年国际神经网络联合会议，*IJCNN 2018* 年，巴西里约热内卢，7 月 *8-13*， *2018* 年，第 1–8页。IEEE。 （第 82、 84、 103、 104和 116页）

Hugo Rosa、Nádia Salgado Pereira、Ricardo Ribeiro、Paula Costa Ferreira、João Paulo Carvalho、Sofia Oliveira、Luísa Coheur、Paula Paulino、Ana Margarida Veiga Simão 和 Isabel Trancoso。 2019年b. 自动网络欺凌检测：系统综述。《计算机与人类行为》， 93：333–345。 （第 74和 81页）

Sara Rosenthal、Pepa Atanasova、Georgi Karadzhov、Marcos Zampieri 和 Preslav Nakov。 2021年。 SOLID：用于冒犯性语言识别的大规模半监督数据集。载于 计算语言学协会论文集：*ACL- IJCNLP2021*年，第 915–928页，线上。计算语言学协会。（第 148页）

Tom Roth、Yansong Gao、Alsharif Abuadbba、Surya Nepal 和 Wei Liu。 2021年。 用于自然语言处理的词元修改对抗攻击：综述。 *CoRR*，abs/2103。00676。（第 147 页）

Chakaveh Saedi 和 Mark Dras。2020年。 使用孪生变分自编码器进行大规模作者身份混淆：SiamAO 系统。载于 第九届词汇与计算语义联合会议论文集，第 179–189页，西班牙巴塞罗那（线上）。计算语言学协会。（第51页）

Semiu Salawu、Yulan He 和 Joanna Lumsden。2020年。 网络欺凌自动检测方法：综述。《*IEEE*情感计算汇刊》，11（1）： 3–24。（第 146页）

Joni Salminen、Maximilian Hopf、Shammur A. Chowdhury、Soon-Gyo Jung、Hind Almerekhi 和 Bernard J. Jansen。2020年。 开发用于多个社交媒体平台的在线仇恨分类器。《以人为中心的计算与信息科学》， 10：1。 （第147页）

Wojciech Samek、Alexander Binder、Grégoire Montavon、Sebastian Lapuschkin 和 Klaus-Robert Müller。2017年。 评估深度神经网络所学内容的可视化

网络。 *IEEE* 神经网络与学习系统汇刊，28(11)：2660–2673。(第 130页)

VictorSanh、Lysandre Debut、JulienChaumond 和 Thomas Wolf。2019年。Distil- bert，一种 BERT 的蒸馏版本：更小、更快、更便宜且更轻量。*CoRR*，abs/1910。01108。（第101至157页）

Victor Sanh、Albert Webson、Colin Raffel、Stephen H. Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、Arnaud Stiegler、Teven Le Scao、Arun Raja、Manan Dey、M. Saiful Bari、Canwen Xu、Urmish Thakker、Shanya Sharma、Eliza Szczechla、Taewoon Kim、Gunjan Chhablani、Nihal V. Nayak、Debajyoti Datta、Jonathan Chang、Mike Tian-Jian Jiang、Han Wang、Matteo Manica、Sheng Shen、Zheng Xin Yong、Harshit Pandey、RachelBawden、Thomas Wang、Trishala Neeraj、Jos Rozen、Abheesht Sharma、Andrea Santilli、Thibault Févry、Jason Alan Fries、Ryan Teehan、Stella Biderman、Leo Gao、Tali Bers、Thomas Wolf 和 Alexander M. Rush。2021年。多任务提示训练实现零样本任务泛化。*CoRR*，abs/2110。08207。（第164页）

Maarten Sap、Dallas Card、Saadia Gabriel、Yejin Choi 和 Noah A. Smith。2019年。仇恨言论检测中的种族偏见风险。载于《57届计算语言学协会年会论文集，第1668–1678页，意大利佛罗伦萨。计算语言学协会。（第163页）

Maarten Sap、Gregory Park、Johannes Eichstaedt、Margaret Kern、David Stillwell、Michal Kosinski、Lyle Ungar 和 Hansen Andrew Schwartz。2014年。基于社交媒体构建年龄和性别预测词典。载于《*2014*年自然语言处理经验方法会议（*EMNLP*）论文集，第1146–1151页，多哈，卡塔尔。计算语言学协会。（第23 和 48页）

Roland Schäfer 和 Felix Bildhauer。2012年。利用新型高效工具链从 WEB 构建大型语料库。载于第八届国际语言资源与评测会议（*LREC'12*）论文集，第486–493页，伊斯坦布尔，土耳其。欧洲语言资源协会（ELRA）。（第101页）

N. Schaltenbrand、R. Lengelle、M. Toussaint、R. Luthringer、G. Carelli、A. Jacqmin、E. Lainey、A. Muzet 和 J. P. Macher。1996年。使用神经网络模型进行睡眠分期评分：正常受试者与患者的目视分析和自动分析比较。《睡眠》，19(1)：26–35。(第5页)

Jonathan Schler、Moshe Koppel、Shlomo Argamon 和 James W. Pennebaker。2006年。年龄和性别对博客写作的影响。载于《计算分析网络日志的方法》，*2006年AAAI春季研讨会论文，技术报告 SS-06-03*，美国加利福尼亚州斯坦福，*3月27-29，2006*年，第199–205页。AAAI。(第9页)

Jürgen Schmidhuber. 2015年. 神经网络中的深度学习：综述. 神经网络, 61:85–117. (p.5)

Tobias Schnabel andHinrich Schütze. 2014年. FLORS：用于词性标注的快速简便领域适应方法. 计算语言学协会汇刊, 2:15–26. (p.87)

Bruce Schneier. 2010年. 安全与功能蔓延. *IEEE* 安全与隐私, 8(1):88. (p.10)

Hendrik Schuff、Heike Adel 和 Ngoc Thang Vu. 2020年. F1 还不够！迈向以用户为中心的可解释问答的模型与评估. In论文集：*2020*年 自然语言处理实证方法会议（*EMNLP*），页码 7076–7095，线上。计算语言学协会。(p. 12)

Mike Schuster 和 Kuldip K. Paliwal。 1997年。 双向循环神经网络。*IEEE* 信号处理汇刊， 45(11)：2673–2681。 (第 26和 103页)

D. Sculley、Jasper Snoek、Alex Wiltschko 和 Ali Rahimi。 2018年。 赢家诅咒？论速度、进展与实证严谨性。 (第16页)

D. Sculley、Jasper Snoek 和 Alexander B. Wiltschko。 2019年。 避免同行评审流程中的公地悲剧。 *CoRR*，abs/1901。06246。 (第16页)

Thibault Sellam、Dipanjan Das 和 Ankur Parikh。 2020年。 BLEURT：学习用于文本生成的鲁棒度量指标。见 《第*58*届计算语言学协会年度会议论文集，第7881–7892页，线上。计算语言学协会。(第 137和 159页)

Sevil Sen 和 Burcu Can。 2021年。 使用自然语言处理技术实现 Android 安全：综述。*CoRR*，abs/2107。03072。 （第11页）

Rakshith Shetty、Bernt Schiele 和 Mario Fritz。 2018年。 A4NT：通过神经机器翻译的对抗训练实现作者属性匿名性。载于 第*27*届 *USENIX* 安全研讨会，*USENIX Security 2018*年，美国马里兰州巴尔的摩，*8*月*15-17*， *2018*年，页码 1633–1650。USENIX 协会。 (第 22、23、25、51、66和 140页)

Reza Shokri、Marco Stronati、Congzheng Song 和 Vitaly Shmatikov。 2017年。 针对机器学习模型的成员推断攻击。载于 *2017* 年 *IEEE* 安全与隐私研讨会，*SP2017*年，圣何塞，加利福尼亚州，美国，*5*月 *22-26*， *2017*年，页码 3–18。IEEE 计算机学会。 (第7页)

Jordan Shropshire. 2018年。 将自然语言处理用作武器。载于 第*13*届信息安全与隐私 *Pre-ICIS* 研讨会论文集，第 1 卷。 （第 12 页）

Chenglei Si、Zhengyan Zhang、Fanchao Qi、Zhiyuan Liu、Yasheng Wang、Qun Liu 和 Maosong Sun。 2021年。 通过扩大覆盖率提升鲁棒性：用于鲁棒微调的对抗数据增强与混合数据增强。载于 计算语言学协会论文集：*ACL/IJCNLP 2021* 年，线上活动，*8*月*1-6*， *2021*年，ACL/IJCNLP 2021年 卷，收录于 *ACL*论文集，页码 1569–1576。计算语言学协会。 （第146页）

Aaron Smith 和 Monica Anderson。 2018年。 2018年社交媒体使用情况。 皮尤研究中心。 （第90页）

Noah A. Smith。 2012年。 自然语言模型的对抗评估。 *CoRR*，abs/1207。0245。 （第46页）

Matthew Snover、Nitin Madnani、Bonnie Dorr 和 Richard Schwartz。 2009。 流畅度、充分性，还是 HTER？使用可调机器翻译指标探索不同的人工判断。载于 第四届统计机器翻译研讨会论文集，第259–268页，希腊雅典。计算语言学协会。 (第160页)

Congzheng Song、Thomas Ristenpart 和 Vitaly Shmatikov。 2017年。 记忆过多的机器学习模型。载于 论文集*2017*年 *ACM SIGSAC*计算机与通信安全会议（*CCS*）*2017*，美国德克萨斯州达拉斯，*10*月 *30* 日至*11*月 *03*， *2017*年，第 587–601页。ACM。 (第7页)

Liwei Song、Xinwei Yu、Hsuan-Tung Peng 和 Karthik Narasimhan。 2021年。 用于文本分类的自然触发器通用对抗攻击。载于 论文集 *2021* 年 计算语言学协会北美分会人类语言技术会议，第 3724–3733页，线上。计算语言学协会。 (第159页)

Anna Cinzia Squicciarini、Sarah Michele Rajtmajer、Y. Liu 和 Christopher Griffin。2015年。 在线社交网络中网络欺凌动态的识别与表征。载于 《*2015*年 *IEEE/ACM* 国际会议论文集：关于

社交网络分析与挖掘进展，*ASONAM 2015*年，法国巴黎，*8*月 *25 — 28*， *2015*年，第 280–285页。ACM。 （第 82和第 86页）

Mani B. Srivastava、Anantha P. Chandrakasan 和 Robert W. Brodersen。1996年。预测性系统关闭及其他节能可编程计算架构技术。 *IEEE*超大规模集成系统汇刊，4（1）：42–55。（第 5 页）

Nitish Srivastava、Geoffrey E. Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。2014年。 Dropout：一种防止神经网络过拟合的简单方法。机器学习研究期刊，15 （ 1 ）： 1929–1958。 （第 30 、 54 和 130 页）

E. Stamatatos、N. Fakotakis 和 G. Kokkinakis。 1999年。 自动作者归属。载于欧洲计算语言学协会第九届欧洲分会会议，第 158–164 页，卑尔根，挪威。计算语言学协会。 （第9页）

Efstathios Stamatatos、Nikos Fakotakis 和 GeorgeKokkinakis。 2000年。 基于体裁和作者的自动文本分类。 计算语言学， 26（4）： 471–495。 （第22页）

Wouter M P Steijn 和 Alexander P Schouten。 2013年。 社交网站上的信息共享与关系。 *CyberpsycholBehavSocNetw*， 16（8）： 582–587。 （第75页）

Ariel Stolerman、Rebekah Overdorf、Sadia Afroz 和 Rachel Greenstadt。2014年。打破文体计量作者归属判定中的封闭世界假设。载于数字取证进展 *X - 10*届 *IFIP WG 11*。第*9*届国际会议，维也纳，奥地利，*2014*年*1*月*8-10*， *2014*年，修订精选论文，第 433 卷，载于 *IFIP*信息与通信技术进展，第 185–205页。施普林格。 （第51页）

Margaret Stratford。1996年。 校园欺凌：见解与观点。 犯罪行为与心理健康，6(4)：362–363。 (第76页)

Junming Sui。 2015年。 运用机器学习理解并应对欺凌。博士论文，威斯康星大学麦迪逊分校。 (第82页)

Chi Sun、Xipeng Qiu、Yige Xu 和 Xuanjing Huang。 2019年。如何微调 BERT 以进行文本分类？载于 中文计算语言学，第 194–206页，尚。施普林格国际出版。（第101页）

Mingxuan Sun、Changbin Li 和 Hongyuan Zha。 2017年。 推断推荐系统中新用户的私人个人资料。载于 第*20*届 *ACM* 无线与移动系统建模、分析与仿真国际会议论文集（*MSWiM*）*2017*年，美国佛罗里达州迈阿密，*11*月 *21 - 25*，*2017*年，第 237–244页。ACM。 （第8页）

Ilya Sutskever、Oriol Vinyals 和 Quoc V. Le。 2014年。 神经网络的序列到序列学习。收录于 神经信息处理系统进展 *27*：神经信息处理系统年会 *2014*年，*12*月 *8-13*日，*2014*年，加拿大魁北克省蒙特利尔，第 3104–3112页。 (第26页)

Steve Durairaj Swamy、Anupam Jamatia 和 Björn Gambäck。 2019年。 研究不同辱骂性语言检测数据集之间的泛化能力。收录于 论文集： *23*届计算自然语言学习会议（*CoNLL*），第 940–950页，中国香港。计算语言学协会。 (第 126 和147页)

Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian J. Goodfellow 和 Rob Fergus。 2014年。 神经网络引人注目的特性。收录于 第*2*届国际学习表征会议（*ICLR*）*2014*年，加拿大班夫，*4*月 *14-16*，*2014*年，会议论文集。 (第14页)

Jr. Tandoc, Edson C, Chen Lou 和 Velyn Lee Hui Min。2018年。多社交媒体语境中的平台切换：用户如何以及为何在多个社交媒体平台间游走。《计算机中介传播杂志》，24(1)：21–35。(第13页)

José Medardo Tapia-Téllez 和 Hugo Jair Escalante。2020年。使用 Transformer 进行文本分类的数据增强。载于《计算智能进展》-第*19*届墨西哥人工智能国际会议（*MICAI*）*2020*年，墨西哥墨西哥城，*10*月*12-17*日，*2020*年，论文集，第二部分，第12469 卷，载于《计算机科学讲义》，第247–259页。施普林格。(第148页)

Nithum Thain、Lucas Dixon 和 Ellery Wulczyn。2017年。维基百科讨论标签：毒性。(第94页)

Hoi LeThi、Reihaneh Safavi-Naini 和 Asadullah Al Galib。2015年。写作风格的安全混淆。载于《信息安全理论与实践》-第*9*届 *IFIP WG11*。*2* 国际会议，*WISTP2015*年希腊克里特岛伊拉克利翁，*2015*年*8*月*24-25*，*2015*年会议论文集，第 9311 卷，收录于《计算机科学讲义》，页码88–103。施普林格。（第25 页和50页）

Nanna Thylstrup 和 Zeerak Waseem。2020年。检测“污垢”和“毒性”：重新思考将内容审核视为污染行为。可通过 *SSRN* 获取*3709719*。（第166页）

Yuandong Tian、Jerry Ma、Qucheng Gong、Shubho Sengupta、Zhuoyuan Chen、James Pinkerton 和 Larry Zitnick。2019年。ELF opengo：对 AlphaZero 的分析与开放式再实现。收录于《第*36*届国际机器学习会议（*ICML*）*2019*年，*6*月*9*—*15*日，*2019*年，美国加利福尼亚州长滩》，第97卷，载于《机器学习研究论文集》，页码6244–6253。PMLR。(第16页)

Yuchi Tian、Kexin Pei、Suman Jana 和 Baishakhi Ray。2018年。DeepTest：深度神经网络驱动的自动驾驶汽车的自动化测试。收录于《第*40*届国际软件工程会议（*ICSE*）*2018*年，瑞典哥德堡，*5*月*27*日至*6*月*03*日，*2018*年，页码303–314。ACM。(第14页)

Sabina Tomkins、Lise Getoor、Yunfei Chen 和 Yi Zhang。2018年。用于网络欺凌检测的社会语言学模型。载于*IEEE/ACM 2018*年社交网络分析与挖掘进展国际会议（*ASONAM*）*2018*年，西班牙巴塞罗那，*8*月*28-31*，*2018*年，第53–60页。IEEE 计算机学会。（第85页）

Stéphan Tulkens、Chris Emmery 和 Walter Daelemans。2016年。将无监督荷兰语词嵌入评估为一种语言资源。载于第十届语言资源与评测国际会议（*LREC'16*）论文集，第4130–4136页，斯洛文尼亚波尔托罗日。欧洲语言资源协会（ELRA）。（第9页和第101页）

Matthew Turk 和 Alex Pentland。1991。用于识别的特征脸。《认知神经科学杂志》，3(1)：71–86。（第5页）

Marjan Vaez、Kerstin Ekberg 和 Lucie Laflamme。 2004年。 年轻在职成年人工作场所虐待事件：问题的严重程度及其对自评健康的影响。 《工业关系》， 59(3)：569–584。 (第77页)

Patti M. Valkenburg 和 Jochen Peter。2007年。 青春期前儿童和青少年的在线交流及其与朋友的亲密程度。 《发展心理学》， 43(2)：267–277。 (第 75 和126页)

Evert Van den Broeck、Karolien Poels、Heidi Vandebosch 和 KathleenVan Royen。2014年。 在线观点采择作为应对网络欺凌的干预工具。《健康技术与信息学研究》，199：113–7。 （第92页）

Chris van der Lee、Albert Gatt、Emiel van Miltenburg、Sander Wubben 和 Emiel Krahmer。2019年。 自动生成文本人工评估的最佳实践。载于《第*12*届国际自然语言生成会议论文集，第355–368页，东京，日本。计算语言学协会。（第50页和160页）

Cynthia Van Hee、Gilles Jacobs、Chris Emmery、Bart Desmet、Els Lefever、Ben Verho- even、Guy De Pauw、Walter Daelemans 和 Véronique Hoste。2018年。 社交媒体文本中的网络欺凌自动检测。 *PLOS ONE*， 13（10）： 1–22。（第13、78、82、84、85、94、98、102、107和 135页）

Cynthia Van Hee、Els Lefever、Ben Verhoeven、Julie Mennes、Bart Des met、Guy De Pauw、Walter Daelemans 和 Veronique Hoste。 2015年a. 网络欺凌事件的检测与细粒度分类。载于 国际自然语言处理最新进展会议论文集，第 672–680页，保加利亚希萨尔。INCOMA Ltd. 舒门，保加利亚。 （第 82、84、85、91和147页）

Cynthia Van Hee、Ben Verhoeven、Els Lefever、Guy De Pauw、Walter Daelemans 和 Véronique Hoste。2015年b. 网络欺凌细粒度分析指南。技术报告，版本 1。0。技术报告 LT315-01，LT3，语言与翻译技术团队–根特大学。（第 78、80、85、92和 118页）

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。 Attention is All You Need。载于神经信息处理系统进展 *30*：神经信息处理系统年会 *2017* 年，*12* 月 *4-9*， *2017* 年，美国加利福尼亚州长滩，第 5998–6008 页。 （第 54 和 147页）

Bertie Vidgen 和 Leon Derczynski。2021年。辱骂性语言训练数据的发展方向，系统综述：垃圾进，垃圾出。*PLOSONE*，15（12）： 1–32。（第126页）

Marc Vilain、Jennifer Su 和 Suzi Lubar。2007年。实体抽取是一个无聊且已解决的问题——还是这样吗？ 载于人类语言技术*2007*：北美分会计算语言学协会会议论文集；伴随卷，短论文，第181–184页，纽约州罗切斯特。计算语言学协会。（第87页）

Andreas Vlachos 和 Sebastian Riedel。2014年。 事实核查：任务定义与数据集构建。载于 *ACL 2014*年 语言技术与计-

算社会科学研讨会论文集，第 18–22 页，美国马里兰州巴尔的摩。计算语言学协会。(第 12 页)

Svitlana Volkova 和 Yoram Bachrach。 2016年。 从用户情绪语气和用户与环境的情绪反差推断感知人口统计属性。载于 《第*54*届计算语言学协会年会论文集》（第*1*卷：长篇论文），第 1567–1578页，德国柏林。计算语言学协会。 （第 22和第 46页）

Svitlana Volkova、Yoram Bachrach、Michael Armstrong 和 Vijay Sharma。 2015年。从社交媒体发布的文本中推断用户潜在属性。载于 《第二十九届*AAAI*人工智能大会论文集》，会议于 *25-30*日举行，*2015*年，美国德克萨斯州奥斯汀，第4296–4297页。AAAI出版社。 （第 56页、59页、60页、61页和第 63页）

Svitlana Volkova、Glen Coppersmith 和 Benjamin Van Durme。 2014年。 从流式通信中推断用户政治偏好。载于 《第*52*届计算语言学协会年会论文集》（第 *1*卷：长篇论文），第 186–196页，美国马里兰州巴尔的摩。计算语言学协会。（第 8页、9页、22页、46页和第 60页）

Sandra Wachter 和 B. Mittelstadt。2018年。《合理推断权：重新思考大数据和人工智能时代的数据保护法》。《哥伦比亚商法评论》，2019年：494–620–494–620。(第5页)

Eric Wallace、Shi Feng、Nikhil Kandpal、Matt Gardner 和 Sameer Singh。 2019年。用于攻击和分析自然语言处理的通用对抗触发器。载于《论文集》*2019*年自然语言处理实证方法会议和9届国际自然语言处理联合会议（*EMNLP-IJCNLP*），页码2153–2162，中国香港。计算语言学协会。(第68页及159页)

Hanna Wallach。2014年。大数据、机器学习与社会科学：公平性、问责制和透明度。*Medium*。(第10页)

Pengfei Wang、Jiafeng Guo、Yanyan Lan、Jun Xu 和 Xueqi Cheng。2016年。购物车告诉你：从购买数据推断人口统计属性。载于第九届 *ACM* 国际网络搜索与数据挖掘会议论文集，美国加利福尼亚州旧金山，*2*月*22-25*，*2016*年，页码173–182。ACM。(第8页)

Sida Wang 和 Christopher Manning。 2012年。 基线方法与二元语法：简单且出色的情感与主题分类。载于 第 *50*届年度会议论文集，

计算语言学协会（第 2卷：短论文），第 90–94页，韩国济州岛。计算语言学协会。(第 101)

Zeerak Waseem。2016年。你是种族主义者，还是我看错了？Twitter 上仇恨言论检测中的标注者影响。载于第一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。（第56页）

Zeerak Waseem 和 Dirk Hovy。2016年。仇恨符号还是仇恨之人？用于 Twitter 仇恨言论检测的预测特征。载于北美计算语言学协会学生研究研讨会论文集，第88–93页，加利福尼亚州圣迭戈。计算语言学协会。（第56页）

Jason Wei 和 Kai Zou。2019年。EDA：用于提升文本分类任务性能的简易数据增强技术。载于论文集：*2019*年自然语言处理实证方法会议第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*），第6382–6388页，中国香港。计算语言学协会。（第134页）

Johannes Welbl、Amelia Glaese、Jonathan Uesato、Sumanth Dathathri、John Mellor、Lisa Anne Hendricks、Kirsty Anderson、Pushmeet Kohli、Ben Coppin 和 Po- Sen Huang。 2007年。 语言模型去毒化面临的挑战。载于 计算语言学协会论文集：*EMNLP 2021*年，第 2447–2469页，蓬塔卡纳，多米尼加共和国。计算语言学协会。 （第167页）

Michael Wiegand、Josef Ruppenhofer 和 Thomas Kleinbauer。 2019年。 辱骂性语言检测：有偏数据集带来的问题。载于 论文集*2019*年 计算语言学协会北美分会会议：人类语言技术，第 *1* 卷（长篇与短篇论文），第602–608页，明尼苏达州明尼阿波利斯。计算语言学协会。（第 163页）

Nancy E Willard。 2007年。 网络欺凌与网络威胁：应对网络社交攻击、威胁与困扰这一挑战。《网络欺凌与网络威胁：应对网络社交攻击、威胁与困扰这一挑战》。Research Press，尚佩恩，美国伊利诺伊州。 （第76页）

Bryn Williams-Jones、Catherine Olivier 和 Elise Smith。2013年。加拿大如何治理“双重用途”研究：政策评述。《科学与公共政策》，41(1)：76–93。(第10页)

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davi- son、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexan- der Rush。2020年。Transformer：自然语言处理的最先进水平。载于*2020*年自然语言处理经验方法会议论文集：系统演示，第38–45，线上。计算语言学协会。(第58和132页)

Yonghui Wu、Mike Schuster、Zhifeng Chen、Quoc V. Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、Yuan Cao、Qin Gao、Klaus Macherey、Jeff Klingner、Apurva Shah、Melvin Johnson、Xiaobing Liu、Lukasz Kaiser、Stephan Gouws、Yoshikiyo Kato、TakuKudo、Hideto Kazawa、Keith Stevens、George Kurian、Nishant Patil、Wei Wang、Cliff Young、Jason Smith、Jason Riesa、Alex Rudnick、Oriol Vinyals、Greg Corrado、Macduff Hughes 和 Jeffrey Dean。 2016年。 谷歌的神经机器翻译系统：弥合人工翻译与机器翻译之间的差距。 *CoRR*，abs/1609。08144。 （第26 和131页）

Zhelun Wu、Nishant Kambhatla 和 Anoop Sarkar。 2018年。 用于对抗性冒犯语言检测的解密。收录于 《第*2*届线上辱骂性语言研讨会论文集，*ALW@EMNLP 2018*年，比利时布鲁塞尔，*10*月 *31*日，*2018*年，第149–159页。计算语言学协会。（第147页）

Ellery Wulczyn、Nithum Thain 和 Lucas Dixon。2017年。Ex machina：大规模观察到的人身攻击。载于《第*26*届万维网国际会议（*WWW*）论文集，*2017*年，澳大利亚珀斯，*4*月*3-7*日，*2017*年，第1391–1399页。ACM。（第94页）

Alexandros Xenos、John Pavlopoulos、Ion Androutsopoulos、Lucas Dixon、Jeffrey Sorensen 和 Leo Laugier。2021年。毒性检测可能对对话语境敏感。*CoRR*，abs/2111。10223。（第165页）

Qizhe Xie、Zihang Dai、Yulun Du、Eduard H. Hovy 和 Graham Neubig。2017年。通过对抗性特征学习实现可控不变性。载于《神经信息处理系统进展*30*：神经信息处理系统年会*2017*年，*12*月*4-9*日，*2017*年，美国加利福尼亚州长滩，第585–596页。（第26页）

Jun-Ming Xu、Kwang-Sung Jun、Xiaojin Zhu 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会议论文集：人类语言

技术，第656–666页，蒙特利尔，加拿大。计算语言学协会。（第76、78、79、82、85、91、93、108、132和147）

Qiongkai Xu、Chenchen Xu 和 Lizhen Qu。2019年。 ALTER：用于自然语言生成的辅助文本改写工具。载于《*2019*年 自然语言处理实证方法会议暨 第*9*届国际自然语言处理联合会议（*EMNLP-IJCNLP*）：系统演示论文集，第 13–18页，中国香港。计算语言学协会。（第50页）

Ruochen Xu 和 Yiming Yang。 2017年。 用于文本分类的跨语言蒸馏。载于《第*55*届计算语言学协会年会（第 *1*卷：长篇论文），第 1415–1425页，加拿大温哥华。计算语言学协会。（第26页）

Ying Xu、Xu Zhong、AntonioJosé Jimeno-Yepes 和 Jey Han Lau。 2020年。 房间里的大象：用于评估自然语言处理中对抗样本的评估框架。*CoRR*，abs/2001。07820。（第160页）

Wei Yang、Yuqing Xie、Luchen Tan、Kun Xiong、Ming Li 和 Jimmy Lin。2019年。 用于开放域问答中 BERT 微调的数据增强。 *CoRR*，abs/1904。06652。（第135页）

Andrew Yates、Arman Cohan 和 Nazli Goharian。 2017年。 在线论坛中的抑郁与自伤风险评估。载于 《*2017*年 自然语言处理实证方法会议论文集》，第 2968–2978页，丹麦哥本哈根。计算语言学协会。（第22页）

Dawei Yin、Zhenzhen Xue、Liangjie Hong、BrianD Davison、April Kontostathis 和 Lynne Edwards。 2009。 WEB 上的骚扰检测 2。0。 *WEB* 内容分析 2。*0* （*CAW 2*。*0*）*WWW* 研讨会论文集*2009*， 2：1–7。（第82页及第 83页）

Jin Yong Yoo、John Morris、Eli Lifland 和 Yanjun Qi。2020年。寻找搜索方法：用于生成自然语言处理对抗样例的搜索算法基准测试。载于第三届自然语言处理神经网络分析与解释 *BlackboxNLP* 研讨会论文集，第323–332页，线上。计算语言学协会。（第158页）

Kang Min Yoo、Dongju Park、Jaewook Kang、Sang-WooLee 和 Woo-MyoungPark。2021年。 Gpt3mix：利用大规模语言模型进行文本增强。

发表于计算语言学协会论文集：*EMNLP 2021*年，线上活动／蓬塔卡纳，多米尼加共和国，*16-20 11*月，*2021*年，第2225–2239页。计算语言学协会。（第134 和148页）

张驰源、Samy Bengio、Moritz Hardt、Michael C. Mozer 和 Yoram Singer。2020年a. 身份危机：极端过参数化下的记忆与泛化。载于*8*届国际学习表征会议（*ICLR*）*2020*年，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26-30*，*2020*年。OpenReview.net。(第16页)

张天翼、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 Yoav Artzi。2020年b.BERTScore：使用 BERT 评估文本生成。载于*8*届国际学习表征会议（*ICLR*）*2020*年，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26-30*，*2020*年。OpenReview.net。(第61和159页)

Wei Emma Zhang、QuanZ. Sheng、Ahoud Alhazmi 和 李晨亮。2020年c. 自然语言处理中的深度学习模型对抗攻击：综述。*ACM Trans. Intell. Syst. Technol.*，11(3)：24：1–24：41。(第50和147页)

张翔、Jonathan Tong、Nishant Vishwamitra、Elizabeth Whittaker、Joseph P. Mazer、Robin M. Kowalski、胡红新、罗峰、Jamie Macbeth 和 Edward Dillon。2016年。基于发音的卷积神经网络进行网络欺凌检测。载于第*15*届*IEEE*国际机器学习与应用会议，*ICMLA 2016*年，美国加利福尼亚州阿纳海姆，*12*月*18-20*，*2016*年，第740–745页。IEEE计算机学会。（第84页）

赵睿和毛克智。2017年。基于语义增强的边缘化去噪自编码器进行网络欺凌检测。《*IEEE*情感计算汇刊》，8(3)：328–339。（第82页和84页）

赵睿、周安娜和毛克智。2016年。基于欺凌特征自动检测社交网络中的网络欺凌。载于《第*17*届国际分布式计算与网络会议论文集，新加坡，*1*月*4-7*，*2016*年，第43：1–43：6。ACM。（第84页）

赵腾飞, 葛兆成, 胡汉平, 和石定蒙. 2021年a. 通过改进的束搜索算法生成自然语言对抗样例. *CoRR,* abs/2110.08036. (第 146至 158页)

Zhixue Zhao、Ziqi Zhang 和 Frank Hopfgartner。2021年b. 使用预训练语言模型进行有毒评论分类的比较研究。载于 会议伴侣

WEB 大会伴侣 *2021*年，线上活动 / 斯洛文尼亚卢布尔雅那，*4*月 *19-23*， *2021*年，第 500–507。ACM / IW3C2。 (第166页)

周春婷、孙崇林、Zhiyuan Liu 和刘振明。2015年。一种用于文本分类的 C-LSTM 神经网络。*CoRR*，abs/1511。08630。（第84和103页）

周望春树、葛涛、徐克、魏福如和周明。2019年。基于 BERT 的词汇替换。载于《第*57*届计算语言学协会年会论文集》，第3368–3373页，意大利佛罗伦萨。计算语言学协会。（第54、55、58、65、130、131、133和148页）

Xuhui Zhou、Maarten Sap、斯瓦布哈·斯瓦扬迪普塔、Yejin Choi 和诺亚·史密斯。2021年。有害语言检测自动去偏面临的挑战。载于《第*16*届计算语言学协会欧洲分会会议论文集：主会场》，第3143–3155页，线上。计算语言学协会。（第140页）

肖莎娜·祖博夫。 2015年。 大他者：监视资本主义与信息文明的前景。 《信息技术杂志》，30(1)：75–89。 (第4页)

Kjartan Ólafsson、Sonia Livingstone 和 Leslie Haddon。2013年。 欧洲儿童对在线技术的使用：欧洲证据基础综述。伦敦政治经济学院经济学在线研究文献 50228，伦敦政治经济学院、伦敦政治经济学院图书馆。 (第75页)