以人为中心的自然语言处理事实核查技术现状

Anubrata Das*，Houjiang Liu，Venelin Kovatchev 和MatthewLease

[a]德克萨斯大学奥斯汀分校信息学院，美国德克萨斯州奥斯汀

文章信息

关键词：自然语言处理 错误信息 虚假信息 可解释性 人机协作

摘 要

错误信息通过削弱人们对科学的信任、改变公共卫生领域的叙事、加剧社会极化，以及扰乱民主选举和金融市场等方式，威胁着现代社会，并造成诸多其他社会危害。为应对这一问题，越来越多的专业事实核查员和记者对声称属实的信息开展高质量调查。然而，这些主要依赖人工的工作难以跟上问题的巨大规模。为此，人们提出了越来越多的自然语言处理（NLP）技术，以实现更大规模的事实核查。尽管相关研究迅速增长，如今自然语言处理技术在事实核查中的实际应用仍处于起步阶段。

本文回顾当前自然语言处理技术在事实核查方面的能力与局限。我们尤其致力于进一步梳理这些技术的设计空间，探讨如何运用和改进它们，以更好地满足人工事实核查员的需求。为此，我们回顾了基于自然语言处理的事实核查的关键方面：任务形式化、数据集构建、建模，以及以人为中心的策略，例如可解释模型和人在回路方法。接下来，我们回顾将基于自然语言处理的事实核查工具用于协助人工事实核查员的成效。我们建议，未来的自然语言处理研究应尽早与事实核查领域的利益相关方合作，并在模型开发中纳入以人为中心的设计实践，从而进一步引导面向人类使用和实际应用的技术开发。最后，我们主张开展更多研究，推动基准测试开发，以支持对以人为中心的事实核查技术进行外部评估。

arXiv:2301.03056v1 [cs.CL] 8 Jan 2023

1. 引言

错误信息及相关问题（虚假信息、欺骗性新闻、标题党、谣言和信息可信度）正日益威胁社会。尽管自书面文本出现之初，人们就一直关注错误信息（Marcus，1992年），但随着社交媒体的近期发展，创建和传播内容的门槛已降至前所未有的低点。此外，网络极化助长了错误信息的传播，而错误信息反过来又加剧了极化（Cinelli、Pelicon、Mozetič、Quattrociocchi、Novak 和 Zollo，2021a、b；Vicario、Quattrociocchi、Scala 和 Zollo，2019年）。要打破这一恶性循环，就必须从根源上解决错误信息问题。

新闻业（Graves，2018b；Graves 和 Amazeen，2019年；Neely-Sardon 和 Tignor，2018年）和档案学（LeBeau，2017年）等领域对错误信息进行了广泛研究，近年来，旨在应对这一问题的事实核查倡议也显著增多。如今，各类组织都专注于事实核查（例如 PolitiFact、Snopes、FactCheck、First Draft 和 Full Fact），而国际事实核查网络（国际事实核查网络）等组织则为独立事实核查者和记者提供培训与资源，以进一步扩大专家事实核查的规模。

尽管专业事实核查员和记者会对传闻中的事实进行高质量调查，以向公众提供信息，但仅靠人力难以应对这一问题在全球互联网范围内的规模。为此，越来越多的研究开始探索自然语言处理（NLP），以实现事实核查的全部或部分自动化（Guo、Schlichtkrull 和 Vlachos，2022年；Nakov、Corney、Hasanain、Alam、Elsayed、Barrón-Cedeño、Papotti、Shaar 和 Da San Martino，2021a；Zhou 和 Zafarani，2020年；Zeng、Abumansour 和 Zubiaga，2021年；Graves，2018a）。然而，即使是最新技术水平的自然语言处理技术，在许多方面仍无法比肩人类能力，因此不足以在实践中实现事实核查自动化。专家指出（Arnold，2020年；Nakov 等人，2021a），事实核查是一个复杂的过程，需要主观判断和专业知识。尽管当前的自然语言处理系统在处理简单事实核查任务方面日益出色，但识别具有语境依赖性且超出简单陈述范畴的虚假主张

*通讯作者 开放研究者与贡献者身份识别码：0000-0002-5412-6149 (A. Das)；0000-0003-0983-6202 (H. Liu)；0000-0003-1259-1541 (V. Kovatchev)；0000-0002-0056-2834 (M. Lease)[1]https://www.politifact.com/、https://www.snopes.com/、https://www.factcheck.org/、https://firstdraftnews. org/、https://fullfact.org/和https://www.poynter.org/ifcn/，分别。

仍超出全自动系统的能力范围（Chen、Sriram、Choi 和 Durrett，2022年；Fan、Piktus、Petroni、Wenzek、Saeidi、Vlachos、Bordes 和 Riedel，2020年）。例如，隐藏在对话系统中的主张、社交媒体社区的评论串中的主张，以及多媒体内容中的主张，对自动化系统而言尤其具有挑战性。此外，大多数事实核查从业者希望自然语言处理工具能够集成到现有事实核查工作流程中，并降低延迟（Nakov 等人，2021a；Graves，2018b；Alam、Shaar、Dalvi、Sajjad、Nikolov、Mubarak、Da San Martino、Abdelali、Durrani、Darwish、Al-Homaid、Zaghouani、Caselli、Danoe、Stolk、Bruntink 和 Nakov，2021b）。

在这篇文献综述中，我们全面、系统地概述了当前最先进的技术，并探讨了面临的挑战与机遇，以便更有效地利用自然语言处理技术进行事实核查。本文的目标有两个。首先，我们涵盖事实核查自然语言处理流程的各个方面：任务形式化、数据集构建和建模方法。其次，我们强调以人为本的方法，旨在增强并加速人工事实核查，而非取而代之。与此不同，先前的文献综述（Zeng 等人，2021年；Guo 等人，2022年；Oshikawa、Qian 和 Wang，2020年）要么概述现有方法，要么仅介绍事实核查流程某个特定环节的细节（Kotonya 和 Toni，2020a；Hardalov、Arora、Nakov 和 Augenstein，2021年；Hanselowski、AvineshP.V.、Schiller、Caspelherr、Chaudhuri、Meyer 和 Gurevych，2018年；Demartini、Mizzaro 和 Spina，2020年）。

此外，我们认为，应将事实核查自然语言处理技术的综述从模型开发扩展到人机交互领域，因为技术设计应反映用户需求，从而使技术开发更好地融入现实使用情境（Graves，2018a；Lease，2020年；Kovatchev、Smith、Lee、Traynor、Aguilera 和 Devine，2020年；Nakov 等人，2021a；Micallef、Armacost、Memon 和 Patil，2022年；Juneja 和 Mitra，2022年）。具体而言，我们建议读者参阅第7 节，其中我们提出了未来研究的具体方向。

当前挑战在很大程度上源于自动化事实核查技术尚处于相对早期的发展阶段。具体而言，当前研究往往对事实核查流程中的各个组件进行内在评估，而非对整个事实核查任务进行端到端外在评估。此外，各组件的准确率可能仍低于实际应用所需的门槛。尽管研究界聚焦于预测准确率并因此取得了值得称赞的进步，但人为因素（例如可用性、可理解性和信任）受到的关注或取得的进展要少得多，而这些因素对于实际应用至关重要。

这些局限对未来研究具有重要影响。首先，自然语言处理技术在事实核查中的实际应用，很可能会采用混合式、人在回路方法，而非完全自动化。其次，随着技术日趋成熟，端到端评估变得愈发重要，以确保开发的解决方案能够应对现实世界中的实际使用场景。为此，新的基准数据集若能促进对自动化事实核查应用进行外部评估，并适用于实际场景，或许有助于推动开发能够在真实环境中采用的解决方案。最后，为了构建有效的人在回路系统，加强自然语言处理与人机交互之间的跨领域整合，有助于改进事实核查工具的设计，使其在实践中准确、可扩展且易于使用。为此，在自然语言处理研究早期与利益相关方加强合作，并在开发模型时融入以人为中心的设计实践，或许会有所裨益。

撰写本文时，我们考虑了不同的读者群体。对于刚接触自动化事实核查的研究人员和事实核查人员，本文全面概述了这一问题。我们探讨了该领域面临的挑战、最新技术水平以及机遇，并强调如何利用机器学习和自然语言处理来打击虚假信息。我们建议刚接触这一主题的研究人员按照各章节的逻辑结构通读全文。其他在该领域经验更丰富的读者，可能已经熟悉我们讨论的某些概念。对于他们而言，本文从一个新颖的以人为中心的视角审视自动化事实核查，并探讨这一视角如何影响系统设计、实现和评估。为了方便经验更丰富的读者阅读本文，我们简要概述了各章节的内容。

- 第2 节介绍自动化事实核查流程。我们概述人工事实核查员和自动化解决方案的工作流程。
- 第3 节讨论任务形式化：事实核查中各项子任务的目标和形式化定义。
- 第 4 节介绍数据集构建流程，展示用于自动化事实核查的热门语料库，并概述数据的一些局限性。

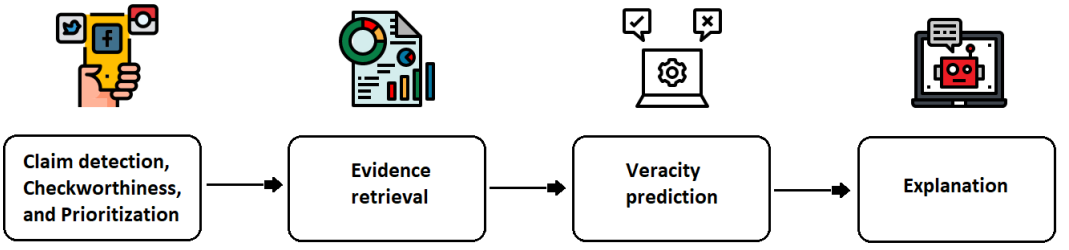

图 1： 事实核查流程

- 第 5 节回顾自动化事实核查的方法。我们将讨论自然语言处理通用能力（第 5.1节）、可解释方法（第 5.2 节）以及用于事实核查的5.3人在回路方法（第 5.3节）。
- 第 6节综述了在实际场景中将自然语言处理应用于事实核查的现有工具。我们认为，以人为中心的视角对于自动化解决方案的实际应用至关重要。
- 第 7节在以人为中心的事实核查语境下提出了未来研究方向。我们在第7.1节讨论混合主动式事实核查中人类与人工智能之间的工作分工。在第7.2 节，我们提出一种衡量信任的新概念，以及一种以人为中心的自然语言处理评估方法，以协助事实核查人员。
- 我们在第8节结束文献综述。

2. 事实核查流程

自动化事实核查的核心理念，是让人工智能能够基于现有信息进行推理，从而判断某项主张的真实性。要成功实现自动化，首先必须了解新闻事实核查这一复杂流程：它不仅需要人类专业知识，还需要投入专业精力来收集和综合证据。处理异构来源（例如来自各种数字与非数字来源的信息）也增加了复杂性。数据还分布在不同模态中，例如图像、视频、表格、图表等。此外，目前缺乏能够支持高效事实核查工作流程的工具（Graves，2018a；Nakov 等人，2021a；Arnold，2020年）。

Graves （2017年）将人工事实核查员的实际事实核查机制划分为多个步骤，例如：a）识别需要核查的主张，b）追踪虚假主张，c）咨询专家，以及 d）分享最终的事实核查结果。越来越多的人工智能研究文献——尤其是自然语言处理领域——致力于自动化事实核查流程。我们综合了多篇相关综述（Guo 等人，2022年；Nakov 等人，2021a；Graves，2018a；Zeng 等人，2021年；Micallef 等人，2022年），并区分出构成自动化事实核查技术流水线的四个典型阶段（见图1）。请注意，以下描述的流水线紧密遵循Guo 等人 （2022年）的结构，同时也将更广泛的文献纳入这四个阶段：

- **主张检测、核查价值与优先级排序：** 主张检测包括监测新闻和/或在线信息，以发现可能需要进行事实核查的虚假内容。必须识别可能可证伪的主张（例如，声称的事实，而非主观观点）（Guo 等人，2022年；Zeng 等人，2021年）。此外，由于事实核查资源（人工或自动化）有限，不可能对所有在线信息进行事实核查，因此事实核查人员必须确定核查事项的优先顺序（Arnold，2020年）。自然语言处理研究者尝试通过自动预测主张的“核查价值”来为此类优先级排序提供依据（Nakov 等人，2021a）。此外，为避免重复劳动，事实核查人员在判断新主张的真实性之前，可能会查阅现有的事实核查数据库（主张匹配（*Zeng* 等人，*2021*年））。我们将主张匹配视为主张优先级排序的一部分，因为事实核查人员会优先考虑不再核查这类主张。

- **证据检索：** 确定要对哪些主张进行事实核查后，下一步便是收集相关且可信的证据，以评估这些主张（Guo 等人，2022年；Zeng 等人，2021年）。
- **真实性预测：** 给定证据后，必须对其进行评估，以确定主张的真实性（Guo 等人，2022年；Zeng 等人，2021年）。

- **解释：** 最后，为了便于人们使用，必须通过人类可理解的理由，解释模型作出判断的事实核查结果（Graves，2018a；Kotonya 和 Toni，2020a；Guo 等人，2022年）。

在接下来的章节中，我们将结合自动化事实核查领域现有的自然语言处理研究，讨论上述各项任务。其他一些步骤（例如，文本中的宣传检测、标题党检测）也与事实核查相关，但并不直接属于上述阶段。我们将在第 3.5 节简要讨论这些步骤。

3. 自动化事实核查的任务形式化

现代自然语言处理在很大程度上由数据驱动。在本文中，我们区分任务形式化（概念层面）与数据集构建（在给定任务定义的情况下开展的实现活动）。尽管如此，合适的数据集是否可用，或构建新数据集是否可行，也会影响任务的形式化方式。

3.1. 主张检测、核查价值与优先级排序

事实核查人员和新闻机构会监控信息来源，例如社交媒体（Facebook、Twitter、Reddit 等）、政治竞选活动和演讲，以及政府官员就重要议题发表的公开讲话（Arnold，2020年；Nakov 等人，2021a）。其他来源还包括端到端加密平台上的线索举报渠道（如 WhatsApp、Telegram 和 Signal）（Kazemi、Garimella、Shahi、Gaffney 和 Hale，2021b）。各个平台上的信息量庞大，使得高效监控所有来源以发现错误信息变得困难。Zeng 等人（2021年）将主张检测这一步骤定义为识别、筛选并确定主张的优先级。

为识别主张，人们通常会监测社交媒体信息流中的谣言（Guo 等人，2022年）。传闻可定义为尚未得到证实、且正在网上传播的主张（Zubiaga、Aker、Bontcheva、Liakata 和 Procter，2018年）。传闻的特点是语言具有主观性，且内容能够触达用户（Qazvinian、Rosengren、Radev 和 Mei，2011年）。此外，在判断一条帖子是否为传闻时，也会考虑与传播力相关的元数据，例如分享次数（或转推和转发次数）、点赞数或评论数（Zhang、Cao、Li、Sheng、Zhong 和 Shu，2021a；Gorrell、Kochkina、Liakata、Aker、Zubiaga、Bontcheva 和 Derczynski，2019年）。然而，仅检测传闻不足以判断某项主张是否需要进行事实核查。

对于每段需要关注的文本，事实核查系统需要回答的关键问题包括：

1. 是否存在需要核查的主张？2. 该主张是否包含可验证信息？3. 该主张是否值得核查？4. 是否已有可信来源核查过该主张？

关于第一个标准——是否存在主张——可以问问该主张包含的是据称的事实还是观点（Hassan、Arslan、Li 和 Tremayne，2017a）。例如，像“雷鬼乐是最有灵魂的音乐流派”这样的陈述代表个人偏好，无法核查。相比之下，“*<NAME>* 在奥运会上赢得了金牌”则可以通过将 <NAME> 与所有金牌得主名单进行匹配来核查。

判断主张是否包含可验证信息更具挑战性。例如，如果某项主张只能通过私人知识或个人经历来验证，而这些信息并非广泛可得，那么就无法核查（Konstantinovskiy、Price、Babakar 和 Zubiaga，2021年）。例如，如果有人声称自己昨天吃了某种食物，那么除了其个人陈述之外，恐怕不可能对这一说法进行验证。

正如这个例子所示，主张是否包含可验证信息，很大程度上取决于有哪些证据可用于核查。反过来，在完成证据检索之前，这一点可能并不明确。实际工作中，事实核查人员可能会进行一些初步调查，但通常只根据主张本身来评估其核查价值。

此外，这只是决定是否核查某项主张时需要考虑的众多因素之一。即使某项主张看起来无法验证，只要它具有重大的公众利益，仍可能值得进行事实核查。此外，即使进行了事实核查，最终仍无法得出结论（也就是说，没有证据可以证实或驳斥该主张），仅仅表明该主张的真实性无法确定，也可能是有价值的结果。

如果一项主张具有重要的公众利益，或可能造成危害，则该主张被视为值得核查（Nakov、Da San Martino、Elsayed、Barrón-Cedeño、Míguez、Shaar、Alam、Haouari、Hasanain、Babulkov 等人，2021b；Hassan、Li 和 Tremayne，2015）。例如，与疫苗对新冠肺炎感染率的影响相关的主张，比关于某位哲学家最喜欢吃什么的主张更关乎公众利益，因此也更值得核查。

主张和模因一样，往往会多次出现，和/或出现在多个平台上（形式相同或略有改动）（Leskovec、Backstrom 和 Kleinberg，2009年；Nakov 等人，2021a；Arnold，2020年）。事实核查组织维护着一个不断扩大的数据库，其中收录了已完成事实核查的主张。因此，可信组织会将检测到的主张与已核查主张的数据库进行比对（Shaar、Martino、Babulkov 和 Nakov，2020年；Shaar、Alam、DaSan Martino 和 Nakov，2021a）。将新主张与此类数据库进行比对，有助于避免重复核查已核查过的主张。这一步也称为主张匹配（Zeng 等人，2021年）。

从业者的报告指出，如果主张未能在最初几小时内得到核查，延迟进行的事实核查对于改变正在持续的错误信息叙事不会产生太大影响（Nakov 等人，2021a；Arnold，2020年）。此外，事实核查资源有限，因此组织必须对待核查的主张进行优先排序（Borel，2016年）。可以根据主张的核查价值对其进行优先排序（Nakov、Da San Martino、Barrón-Cedeño、Zaghouani、Míguez、Alam、Caselli、Kutlu、Strub 和 Mandl，2022年；Nakov 等人，2021b）。Nakov 等人（2022年）指出，核查价值取决于诸如以下因素：

1. 核查一项主张的紧迫程度如何？2. 一项主张可能造成多大危害（Alam 等人，2021b；Alam、Dalvi、Shaar、Durrani、Mubarak、Nikolov、Da San Martino、Abdelali、Sajjad、Darwish 和 Nakov，2021a；Shaar、Hasanain、Hamdan、Ali、Haouari、Nikolov、Kutlu、Kartal、Alam、Da San Martino、Barrón-Cedeño、Míguez、Beltrán、Elsayed 和 Nakov，2021b）？3. 为解决根本问题，该主张是否需要政策制定者的关注?

请注意，估算危害颇具挑战性，尤其是在尚未充分理解错误信息造成的危害并制定相应衡量方法之前。Neumann、De-Arteaga 和 Fazelpour (2022年)。

社交媒体上主张的传播情况，是识别公众利益的另一种潜在信号（Arnold，2020年）。本着“让尽可能多的人受益”的原则，病毒式传播的主张可能会被优先处理，因为其中的任何虚假信息都可能对大量人群造成负面影响。另一方面，由于公平性考量要求为所有人提供同等保护，我们不能以牺牲少数群体为代价，只服务多数人（Ekstrand、Das、Burke、Diaz 等人，2022年；Neumann 等人，2022年）。此外，这些少数群体可能更为弱势，因此更需要保护，也可能受到错误信息与虚假信息的不成比例影响（Guo 等人，2022年）。更多讨论请参见第3.6节。

3.2. 证据检索

自动化事实核查中的一些子任务可以在没有明确证据的情况下完成。例如，可以利用文本的语言学特征判断其是否由机器生成（Wang，2017年；Rashkin、Choi、Jang、Volkova 和 Choi，2017年）。然而，没有证据就评估主张的真实性显然更具挑战性（Schuster、Schuster、Shah 和 Barzilay，2020年）。

主张的来源也能反映信息质量；已知的不可靠来源或传播渠道往往屡次散播虚假信息[2]。当内容通过多个来源（Twitter 错误信息机器人）系统性传播时，对来源的分析可能会进一步复杂化（Jones，2019年）。

通常认为，事实核查需要收集可靠且可信的证据，以提供有关主张的信息，供人们据此进行推理（Graves，2018a；Li、Gao、Meng、Li、Su、Zhao、Fan 和 Han，2016年）。在某些情况下，需要核查主张的多个方面。此时，事实核查者会将该主张分解为若干不同的问题，并为这些问题收集相关证据（Borel，2016年；Chen 等人，2022年）。从信息检索（IR）的角度来看，我们可以将这些问题分别视为“信息需求”，事实核查者必须针对这些需求向搜索引擎提出一个或多个查询（Bendersky、Metzler 和 Croft，2012年），以检索必要的证据。

证据可以存在于多种模态中，包括文本、表格、知识图谱、图像和视频。各种元数据也可以提供证据，有时是评估主张所必需的。例如，消歧主张术语所需的语境，或主张来源者（个人或组织）的背景。R

etrievi ng re levant ev id ence a lso depend s on t he followi ng quest ions (Singh, Das, Li and L 便捷, 2021年):

1. 是否有足够的证据与某项主张相关？2. 这些证据是否可访问，或是否属于公共领域？3. 这些证据是否采用可读取和处理的格式?

[2] https://disinformationindex.org/

如前文第 3.1 节所述，前面的主张检测任务需要评估某项主张是否包含可验证信息；这在一定程度上取决于有哪些证据可供检索，而在实际进行证据检索之前，我们并不知道这一点。现在进入证据检索这一步，我们确实能够发现是否存在足够的证据来支持或反驳该主张。

此外，证据应当值得信赖、具有公信力（Nguyen、Kharosekar、Lease 和 Wallace，2018b；Lease，2018年），且不带偏见（Chen、Khashabi、Yin、Callison-Burch 和 Roth，2019a）。

检索到证据后，立场检测会评估证据在多大程度上支持或反驳该主张（Nguyen、Kharosekar、Krishnan、Krishnan、Tate、Wallace 和 Lease，2018a；Ferreira 和 Vlachos，2016年；Popat、Mukherjee、Yates 和 Weikum，2018年）。立场检测通常被表述为针对每条检索到的证据进行分类（或序数回归）的分类任务。请注意，有些研究将立场检测表述为一项独立任务（Hanselowski 等人，2018年；Hardalov 等人，2021年）。

3.3. 真实性预测

给定一项主张和收集到的证据，真实性预测涉及对收集到的证据和主张进行推理。真实性预测可以表述为二元分类任务（即真或假）（Popat 等人，2018年；Nakashole 和 Mitchell，2014年；Potthast、KIESELJ 等人，2018年），也可以根据新闻事实核查实践表述为细粒度的多分类任务（Augenstein、Lioma、Wang、Lima、Hansen、Hansen 和 Simonsen，2019年；Shu、Mahudeswaran、Wang、Lee 和 Liu，2020年；Wang，2017年）。在某些情况下，现有信息可能不足以判断一项主张的真实性（Thorne、Vlachos、Christodoulopoulos 和 Mittal，2018年）。

请注意，事实核查可能是一个递归过程，因为检索到的证据本身可能也需要经过事实核查，才能被信任并据此采取行动（Graves，2018a）。这也与信息素养领域更广泛的教育实践相一致[3]，在这些实践中，读者同样会受到鼓励，去评估自己所接触信息的质量。对信息可靠性的评估可以自然地融入真实性预测任务，同时考量证据的可靠性及其立场（Nguyen 等人，2018b；Guo 等人，2022年）。

3.4. 解释真实性预测

社交媒体平台或许会利用自动化真实性预测来决定是否自动屏蔽或降低内容的推荐权重，但事实核查技术的使用往往涉及人在回路，无论这个人是平台审核员、记者还是最终用户。当我们考虑以人为中心的事实核查技术应用时，如果只提供自动化真实性预测，却不说明判断依据，就可能导致系统被忽视或不受信任，甚至强化人们对虚假主张的错误信念（即“反噬效应”（Lewandowsky、Ecker、Seifert、Schwarz 和 Cook，2012年））。鉴于最先进的自然语言处理系统在面对对抗样本时性能会显著下降，解释和论证依据尤为重要（Kovatchev、Chatterjee、Govindarajan、Chen、Choi、Chronis、Das、Erk、Lease、Li 等人，2022年）。因此，面向人类用户的自动事实核查系统，应以类似于现有新闻事实核查实践的方式解释其真实性预测（Uscinski，2015）。简要来说，许多解释研究都聚焦于为参与系统开发的研究人员和工程师提供解释（包括解释的类型、生成方法和评估流程）。相比之下，我们在此强调面向系统用户的解释。

可以通过多种方式提供解释，例如
1. 证据归因 2. 解释事实核查的决策过程 3. 总结证据
4. 基于案例的解释

证据归因是识别证据或证据中特定方面（例如段落、句子，甚至感兴趣的词元）的过程（Thorne 等人，2018年；Popat 等人，2018年；Shu、Cui、Wang、Lee 和 Liu，2019年；Lu 和 Li，2020年）。此外，证据的相对重要性也可以为事实核查结果提供依据（Nguyen 等人，2018a）。另外，将决策过程中的各个部分拆解开来的一组规则或交互，也可以作为解释（Gad-Elrab、Stepanova、Urbani 和 Weikum，2019年；Nguyen 等人，2018b）。这种表述更侧重于说明如何处理证据以得出决策。对真实性的解释也可以表述为基于收集到的证据进行摘要，以解释事实核查结果（Atanasova、Simonsen、Lioma 和 Augenstein，2020a；Kotonya 和 Toni，2020b）。最后，基于案例的解释可以向用户提供相似的人工标注实例（Das、Gupta、Kovatchev、Lease 和 Li，2022年）。

[3] https://en.wikipedia.org/wiki/Information_素养

3.5. 相关任务

除了被视为自动化事实核查流程核心的任务之外，还有一些相关且互补的任务值得一提。此类任务包括宣传检测（Da San Martino、Cresci、Barrón-Cedeño、Yu、Pietro 和 Nakov，2020年）、标题党检测（Potthast、Köpsel、Stein 和 Hagen，2016年）和论证挖掘（Lawrence 和 Reed，2020年）。此外，有些任务可以独立于事实核查流程进行设计，并在之后用于改进各项事实核查子任务。例如，预测社交媒体内容的传播力（Jain、Garg 和 Jain，2021年）有助于改进主张检测和主张核查价值。同样，对假新闻传播进行网络分析（Shao、Ciampaglia、Flammini 和 Menczer，2016年）有助于分析来源。

为了构建更加以人为中心的人工智能方法，也有一些任务可用于帮助自动化事实核查流程中的部分环节。例如，可通过为可能需要事实核查的内容提供网址推荐引擎，来改进主张检测（Vo 和 Lee，2018年）。此外，事实核查人员也可从预测主张难度的评分中获益（Singh 等人，2021年）。在证据检索和真实性预测方面，可以生成事实核查简报，帮助缺乏经验的事实核查人员（Fan 等人，2020年）。与一般性地总结证据（第3.4节）不同，也可以专门以决策支持为目标进行总结（Hsu 和 Tan，2021年）。

3.6. 主要挑战

自动事实核查领域的大多数工作都集中在真实性预测上，其次是解释生成。近年来，人们对主张检测和核查价值的关注日益增加。相比之下，证据检索方面的研究仍不够成熟。

主张检测 Guo 等人 (2022年)指出了主张核查价值任务中的若干偏差来源。不同社会群体可能对主张产生不同程度的兴趣。此外，与一般人群相比，可能对边缘化群体造成更大危害的主张，可能无法获得足够的关注。理想情况下，识别核查价值的模型需要克服任何可能的差别影响。

Full Fact 的报告中也出现了类似的担忧（Arnold，2020年）。在多个组织用于选择主张进行事实核查的标准中，有一项是：“该主张是否可能威胁民主进程或少数群体？”然而，这一标准可能与对传播力的关注相冲突。事实核查组织通常会监测传播度指标，以决定对哪些主张进行事实核查（Arnold，2020年；Nakov 等人，2021a）。不过，如果虚假主张针对少数族裔，此类主张可能无法达到病毒式传播阈值。

确定哪些主张应优先进行事实核查，需要关注各种人口统计特征：内容创作者、读者和主题。因此，主张核查价值数据集的设计可以从考虑人口统计特征中获益。

证据检索 在自动化事实核查的自然语言处理文献中，证据检索在很大程度上遭到了忽视。人们往往假定证据已经可用，或者将主张输入搜索引擎，以此收集粗粒度证据（Popat 等人，2018年；Nguyen 等人，2018a）。然而，Hasanain 和 Elsayed （2022年）的研究表明，针对相关性进行优化的搜索引擎很少能检索到对真实性预测最有用的证据。尽管在信息检索领域，可信信息检索已得到充分研究（Clarke、Rizvi、Smucker、Maistro 和 Zuccon，2020a），但仍需开展更多专注于检索真实性评估证据的研究（Lease，2018年；Clarke、Rizvi、Smucker、Maistro 和 Zuccon，2020b）。

真实性预测与解释 自动化系统面临的一项关键挑战是：既要基于多个来源的证据进行推理，又要考虑来源信誉。此外，解释复杂的推理过程并非易事。模型解释这一概念本身就有多重含义，且总体而言仍在不断发展，更不用说在事实核查语境中的含义了。随着可解释自然语言处理的发展，自动化事实核查任务也需要不断演进，并提供人类利益相关者能够理解、同时又忠实于底层模型的解释。例如，尽管已有用于宣传检测的系统被提出（Das 等人，2022年），但基于案例的解释在自动化事实核查中仍鲜有探索。

在许多自然语言处理任务中，例如机器翻译或自然语言推理，目标是构建完全自动化的端到端解决方案。然而，在事实核查语境下，最新技术水平的局限性表明，在可预见的未来仍需要人在环中参与。鉴于此，支持人工事实核查员的自动化工具至关重要。然而，

对事实核查者需求的理解，以及在任务形式化中纳入这些需求，在很大程度上仍未出现在自动化事实核查文献中，只有少数值得注意的例外（Nakov 等人、2021a；Demartini 等人，2020年）。未来的研究可以通过让事实核查人员更多地参与自然语言处理研究过程，并将目标从完全自动化转向支持人类而受益。

4. 数据集构建

与任务形式化（第 3 节）相对应，我们也围绕主张、证据、真实性预测和解释来组织对事实核查数据集的介绍。请注意，并非所有数据集都包含这些组成部分。

4.1. 主张检测和主张核查价值

主张检测数据集通常包含主张及其来源（文档、社交媒体信息流、政治演讲的转录文本）（Guo等人，2022年）。主张检测的一种形式是在社交媒体上识别谣言，这类数据集主要使用来自 Twitter 的文本构建（Zubiaga 等人，2018年；Qazvinian 等人，2011年）以及 Reddit（Gorrell 等人，2019年；Lillie、Middelboe 和 Derczynski，2019年）。有些研究提供了主张在社交媒体上出现时所处的语境（Zhang 等人，2021a；Ma、Gao、Mitra、Kwon、Jansen、Wong 和 Cha，2016年）。然而，多项研究指出，大多数主张检测数据集所包含的语境不足。正如第 3 节关于元数据的讨论所示，更广泛的语境可能包括：社交媒体传播范围、传播度指标、主张的来源以及相关用户数据（即谁发布了某项主张、他们在线上的影响力如何等）（Arnold，2020年；Nakov 等人，2021a）。

主张核查价值数据集（Nakov 等人，2021b；Shaar 等人，2021b；Barrón-Cedeño、Elsayed、Nakov、Da San Martino、Hasanain、Suwaileh、Haouari、Babulkov、Hamdan、Nikolov 等人，2020年；Atanasova、Barrón-Cedeño、Elsayed、Suwaileh、Zaghouani、Kyuchukov、Da San Martino 和 Nakov，2018年；Konstantinovskiy 等人，2021年；Hassan 等人，2015）通过依据核查价值标准（见第 3.1 节）对主张进行标注，从来源中筛选主张（与主张检测类似，来源包括社交媒体动态和政治辩论文字稿等）。每条主张都会获得一个核查价值评分，以生成一个排序列表。请注意，主张检测和核查价值数据集可以由专家标注（Hassan 等人，2015），也可以由众包标注（Nakov 等人，2021b；Shaar 等人，2021b；Barrón-Cedeño 等人，2020年；Atanasova 等人，2018年；Konstantinovskiy 等人，2021年）[4]。

上述讨论的数据集并未涵盖多模态数据集，而涵盖此类数据集的也寥寥无几。其中一个此类数据集是 r/Fakeddit（Nakamura、Levy 和 Wang，2020年）。该数据集包含来自 Reddit 的图像及相关文本内容，并将其作为主张。错误信息也可能通过多模态模因传播，而 Facebook（现为 Meta）的仇恨言论任务——仇恨模因挑战赛（Kiela、Firooz、Mohan、Goswami、Singh、Ringshia 和 Testuggine，2020年）——则表明，错误信息检测也可以采用类似的做法。

4.2. 证据

早期的事实核查数据集提供的元数据仅将主张作为证据。此类元数据包括社交媒体帖子属性、用户信息、发布日期和来源信息（Wang，2017年；Potthast 等人，2018年）。如第3.2节前文所述，此类元数据不包含对复杂主张进行推理所必需的世界知识。为解决上述局限，近期的数据集开始考虑外部证据（Guo 等人，2022年）。

证据的收集方式因问题设置而异。对于人工生成的主张，证据通常从单一来源检索，例如维基百科文章（Thorne 等人，2018年；Jiang、Bordia、Zhong、Dognin、Singh 和 Bansal，2020年；Schuster、Fisch 和 Barzilay，2021年）。现实世界主张的领域限定证据取自特定问题相关的来源，例如用于科学主张的学术文章（Kotonya 和 Toni，2020b；Wadden、Lin、Lo、Wang、van Zuylen、Cohan 和 Hajishirzi，2020年），或事实核查网站上列出的特定证据（Vlachos 和 Riedel，2014年；Hanselowski、Stab、Schulz、Li 和 Gurevych，2019年）。现实世界主张的开放领域证据通常通过搜索引擎从网络收集（Popat 等人，2018年；Augenstein 等人，2019年）。

近年来，越来越多的研究开始关注自由文本之外的证据。这些形式包括结构化或半结构化的证据。结构化证据的来源包括知识库（Shi 和 Weninger，2016年）以及来自半结构化知识库的半结构化证据（Vlachos 和 Riedel，2015）、表格数据（Chen 等人，2019a；Gupta、Mehta、Nokhiz 和 Srikumar，2020年），以及文档中的表格（Aly、Guo、Schlichtkrull、Thorne、Vlachos、Christodoulopoulos、Cocarascu 和 Mittal，2021年）。

此外，还有一些检索专用数据集，旨在从搜索引擎中检索可信信息（Clarke等人，2020b）。然而，此类任务并未将主张核查作为一项明确任务纳入其中。

[4]其中一些数据集（如 CheckThat! 数据集）的标注部分由众包完成，部分由专家完成 d.

4.3. 真实性预测

证据检索和真实性预测数据集通常是联合构建的。需要注意的是，在某些情况下，数据集中可能不存在证据。真实性预测数据集通常不会分别处理主张检测或主张核查价值判断任务。相反，此类数据集包含一组主张，这些主张要么是人为构建的，要么是从互联网收集的。

真实性预测数据集中的人工构造主张往往范围有限，是为自然语言推理研究而构建的（Aly 等人，2021年；Thorne 等人，2018年；Schuster 等人，2021年；Jiang 等人，2020年；Chen、Wang、Chen、Zhang、Wang、Li、Zhou 和 Wang，2019b）。例如，FEVER（Thorne 等人，2018年）和 HoVer（Jacovi 和 Goldberg，2021年）从维基百科页面获取主张。一些数据集还采用主谓宾三元组，以便与知识库进行事实核查（Kim 和 Choi，2020年；Shi 和 Weninger，2016年）。

事实核查网站是创建基于真实主张的真实性预测数据集的热门来源。一些数据集从单个网站获取主张，另一些则从多个此类网站收集主张并汇总（Wang，2017年；Hanselowski 等人，2018年；Augenstein 等人，2019年；Vlachos 和 Riedel，2014年）。请注意，这些主张本质上均由专家标注。其他主张来源包括社交媒体（Potthast 等人，2018年；Shu、Sliva、Wang、Tang 和 Liu，2017年）、新闻媒体（Horne、Khedr 和 Adali，2018年；Gruppi、Horne 和 Adalı，2021年；Nørregaard、Horne 和 Adalı，2019年）、博客、问答论坛中的讨论，或类似的用户生成内容发布平台（Mihaylova、Nakov、Màrquez、Barrón-Cedeño、Mohtarami、Karadzhov 和 Glass，2018年）。

此外，一些事实核查数据集针对科学文献（Wadden 等人，2020年）、气候变化（Diggelmann、Boyd-Graber、Bulian、Ciaramita 和 Leippold，2020年）以及公共卫生（Kotonya 和 Toni，2020b）等特定领域的问题。大多数数据集是单语的，但近期已开始将多语主张纳入其中（Gupta 和 Srikumar，2021年；Barnabò、Siciliano、Castillo、Leonardi、Nakov、Da San Martino 和 Silvestri，2022年）。

早期数据集侧重于二元真实性预测——真或假（Mihalcea 和 Strapparava，2009年）。近期数据集通常采用模仿事实核查网站的有序真实性标注方案（Vlachos 和 Riedel，2014年；Wang，2017年；Augenstein 等人，2019年）。然而，每个事实核查网站的真实性等级各不相同，因此跨越多个网站的数据集会带来归一化问题。一些数据集不对标签进行归一化（Augenstein 等人，2019年），而另一些则在事后进行归一化（Kotonya 和 Toni，2020a；Gupta 和 Srikumar，2021年；Hanselowski 等人，2019年）。

4.4. 解释

虽然解释与真实性预测密切相关，但只有少数数据集明确针对可解释的真实性预测这一问题（Atanasova 等人，2020a；Kotonya 和 Toni，2020b；Alhindi、Petridis 和 Muresan，2018年）。在自然语言处理领域，通常会突出显示输入中的某些部分，以解释预测结果。这种解释形式称为抽取式理由（Zaidan、Eisner 和 Piatko，2007年；Kutlu、McDonnell、Elsayed 和 Lease，2020年）。借鉴抽取式理由的理念，一些数据集会在标签旁附上一条证据句（Thorne 等人，2018年；Hanselowski 等人，2018年；Wadden 等人，2020年；Schuster 等人，2021年）。尽管这类数据集并未明确将证据界定为此类情况下的一种解释，但如果证据本身就是解释，那么证据检索与解释之间的界限就会变得模糊。不过，解释与证据在几个方面有所不同。特别是，解释需要简洁，便于用户阅读；而证据则可以由多份文档或篇幅较长的文档构成。解释还需考虑用户差异。因此，仅凭证据作为解释，可能隐含着对用户的一些假设，而不同用户群体（例如专家与非专家）未必都能理解这些假设。

4.5. 挑战

主张核查价值数据集高度不平衡，即，与不值得核查的主张相比，值得核查的主张数量相对较少（Williams、Rodrigues 和 Novak，2020年）。由于特定领域语境有限，数据集也难以推广（Guo 等人，2022年）。此外，尽管现有数据集涵盖英语、阿拉伯语、西班牙语、保加利亚语和荷兰语等多种语言，但它们主要是单语数据集。因此，构建多语言核查价值预测器仍然具有挑战性。核查价值数据集中的许多数据最初并非用于分类。不同组织在选择要核查哪些主张时所采用的标准往往带有主观性，且可能无法推广到特定组织之外。

一些标注实践可能会在数据集中产生伪影。例如，人工构造的虚假主张（如 FEVER 中基于否定的虚假主张）可能会在模型中导致伪影（Schuster 等人，2021年）。由于模型可能会过拟合标注规范，因此无法在数据集之外进行良好泛化（Bansal、Nushi、Kamar、Lasecki、

Weldand Horvitz，2019年）。识别此类盲点的一种方法是使用对抗性数据集进行事实核查。这种设置已被纳入FEVER 2.0（Thorneand Vlachos，2019年）。

为研究构建的数据集未必总能反映事实核查人员在实际工作中的做法。这会导致基于这些数据集构建的算法存在局限性。例如，对事实核查人员的访谈表明，他们在进行事实核查时，往往会综合考虑帖文内容及相关的传播度指标（用于衡量传播范围）（Arnold，2020年）。然而，大多数事实核查数据集都不包含传播度指标。

证据检索 有些数据集通过将主张原文用作查询，并将排名靠前的搜索结果作为证据来构建。然而，有些查询比其他查询更能检索到所需信息。因此，或许应更加谨慎地设计有效查询，或以其他方式改进证据检索，使最终构建的数据集更有可能包含用于真实性预测的高质量证据。否则，低质量证据会成为事实核查流程后续阶段所训练模型质量的瓶颈（Singh 等人，2021年）。

真实性预测 真实性预测数据集面临的一项关键挑战是，不同事实核查网站的标签并不统一，进行标准化可能会引入噪声。

解释 一些数据集将整篇事实核查文章作为证据，并将其摘要作为解释形式（Atanasova 等人，2020a；Kotonya 和 Toni，2020b）。在这种情况下，“解释”部分默认已有现成的事实核查文章。相比之下，提供抽象式摘要，并说明基于证据的推理过程，会更有价值。

数据生成 近年来，人们对将数据生成和数据增强用于各种自然语言处理任务的兴趣日益浓厚（Liu、Swayamdipta、Smith 和 Choi，2022年；Hartvigsen、Gabriel、Palangi、Sap、Ray 和 Kamar，2022年；Dhole、Gangal、Gehrmann、Gupta、Li、Mahamood、Mahendiran、Mille、Srivastava、Tan 等人，2021年；Kovatchev、Smith、Lee 和 Devine，2021年）。在事实核查领域，合成数据的应用尚未得到充分探索。

5. 自动化事实核查

自动化事实核查领域的自然语言处理研究，主要聚焦于利用现有数据集构建模型，以完成不同的自动化事实核查任务。在下文中，我们将概述文献中采用的各种建模策略，并更详细地讨论用于自动化事实核查的可解释方法。

5.1. 自然语言处理通用能力

主张检测与核查价值主张检测通常仅作为分类任务来实现，而主张核查价值通常既通过排序（Nakov 等人，2021a）实现，也通过分类任务（Zeng 等人，2021年）实现。正如任务形式化一节（3.1）前文所述，主张检测这一广义任务可细分为若干子任务：识别主张、筛除重复主张，以及根据主张的核查价值确定优先级。识别主张的另一种情形，是检测社交媒体信息流中的谣言。

早期的一些谣言检测研究侧重于从可用元数据和文本本身中进行特征工程（Enayet 和 El-Beltagy，2017年；Aker、Derczynski 和 Bontcheva，2017年；Zhou、Jain、Phoha 和 Zafarani，2020年）。更先进的主张检测方法采用长短期记忆网络及其他序列模型（Kochkina、Liakata 和 Augenstein，2017年）。这类模型更善于捕捉文本的语境（Zubiaga、Liakata、Procter、Wong Sak Hoi 和 Tolmie，2016年）。树形长短期记忆网络（Ma、Gao 和 Wong，2018年）和层次注意力网络（Guo、Cao、Zhang、Guo 和 Li，2018年）能够捕捉主张的内部结构，或主张出现时所处的语境。此外，图神经网络方法能够在捕捉文本的同时，捕捉相关的社交媒体活动（Monti、Frasca、Eynard、Mannion 和 Bronstein，2019年）。

同样，早期关于主张核查价值的研究采用基于文本特征的支持向量机，并根据优先级对主张进行排序（Hassan 等人，2017a）。例如，Konstantinovskiy 等人（2021年）通过将标签合并为可核查与不可核查主张，构建了一个用于评估核查价值的分类模型。他们构建了一个逻辑回归模型，将词嵌入与基于句法的特征（词性标签和命名实体）结合使用。在早期的核查价值共享任务中，长短期记忆网络等神经网络方法表现良好（Elsayed、Nakov、Barrón-Cedeño、Hasanain、Suwaileh、Da San Martino 和 Atanasova，2019年）。此外，Atanasova、Nakov，

Màrquez、Barrón-Cedeño、Karadzhov、Mihaylova、Mohtarami 和 Glass （2019b）也表明，捕捉语境有助于完成可核查性任务。RoBERTa 等模型在后续一届*CheckThat!* 共享任务中取得了更好的表现（Williams 等人，2020年；Martinez-Rico、Martinez-Romo 和 Araujo，2021年），该任务针对英语声明。如今，在其他语言中，将此类模型微调用于声明检测任务，也日益成为评估主张核查价值的常见做法（Hasanain 和 Elsayed，2020年；Williams 等人，2020年）。

筛选先前已核查的声明是该领域一项相对较新的任务。Shaar 等人（2020年）提出了一种方法，使用 BERT 和 BM-25 将主张与事实核查数据库进行匹配，从而查找数据库中已有的主张。此外，在多个事实核查数据集上对 RoBERTa 进行微调后，模型在识别重复主张方面表现出色（Bouziane、Perrin、Cluzeau、Mardas 和 Sadeq，2020年）。此外，将预训练模型 Sentence-BERT 与 LambdaMART 重新排序相结合，在检测先前已核查的主张方面表现良好（Nakov 等人，2021b）。

证据检索与真实性预测 流水线中的证据检索与真实性预测可以按顺序建模，也可以联合建模。与主张检测和核查价值模型类似，早期研究使用文体特征和元数据，在不借助外部证据的情况下进行真实性预测（Wang，2017年；Rashkin 等人，2017年）。纳入证据检索的模型通常使用商业搜索 API，或采用 TF-IDF、BM25 等检索方法（Thorne 等人，2018年）。与问答模型类似，一些研究采用两步法：首先大规模使用较简单的模型（TF-IDF 或 BM-25），随后在初步筛选后使用更复杂的模型进行重新排序（Thorne 等人，2018年；Nie、Wang 和 Bansal，2019年；Hanselowski 等人，2019年）。此外，也采用文档检索与段落检索相结合的方式，或采用两阶段的“伸缩式”方法：第一阶段检索相关文档，第二阶段检索相关段落。由于第一阶段比第二阶段效率更高，两阶段方法有助于扩展应用规模。对于特定领域的证据检索，使用领域专用的词嵌入已被证明有效（Zeng 等人，2021年）。

信息检索任务并不总是流程的一部分。相反，人们往往假定可靠的证据已经唾手可得。虽然这简化了事实核查任务，使研究人员能够专注于真实性预测，但在实践中，证据检索必不可少，不容忽视。此外，在推理过程中，实际应用还必须应对嘈杂（无关）、质量低劣且带有偏见的搜索结果。

如第 3.3 节前文所述，可能有必要评估所收集证据的可靠性。如果假定证据可信，那么只需检测每条证据的立场，再以某种方式汇总，以推断真实性（例如，可以假定所有证据都同等重要且可信）。然而，实际情况下，人们往往必须面对可靠性存疑的“野外”证据；此时，在将证据用于真实性预测之前，评估其质量（及偏见）就十分重要。

真实性预测利用文本蕴含，基于单个文档或多个文档作为证据来推断真实性。Dagan、Dolan、Magnini 和 Roth （ 2010 年）将文本蕴含定义为 “ 判断给定两个文本片段时，一个文本的含义是否蕴含（可从中推断出）另一个文本的含义”。现实世界的应用通常需要对多个文档进行推理（Augenstein 等人，2019年；Kotonya 和 Toni，2020b；Schuster 等人，2021年）。对多个文档进行推理，可以采用拼接（Nie 等人，2019年）或加权聚合（Nguyen 等人，2018b）。加权聚合实际上会对所考虑的证据重新排序，以筛除不可靠的证据（Ma、Gao、Joty 和 Wong，2019年；Pradeep、Ma、Nogueira 和 Lin，2021年）。有些方法还将知识库用作所有证据的中央存储库（Shi 和 Weninger，2016年）。然而，证据仅限于知识库中已有的内容（Guo 等，2022；Zeng 等，2021）。此外，知识库的一个根本局限在于，并非所有知识都能轻易纳入结构化关系中。

大型语言模型的最新进展有助于拓展知识库方法。事实核查模型可以依托预训练模型，为真实性预测提供证据（Lee、Li、Wang、Yih、Ma 和 Khabsa，2020年）。不过，这种方法可能会编码语言模型中存在的偏差（Lee、Bang、Madotto 和 Fung，2021年）。

另一种方法是通过处理证据，帮助事实核查人员完成下游任务。例如，利用 BERT 根据现有证据生成摘要（Fan 等人，2020年）。

局限性 随着大型预训练语言模型以及用于自然语言处理的深度学习技术近期不断发展，我们看到整个事实核查流程的性能得到了显著提升。随着 FEVER（Thorne 等人，2018年；Thorne、Vlachos、Cocarascu、Christodoulopoulos 和 Mittal，2019年；Aly 等人，2021年）和 *CheckThat!* （Nakov 等人，2021b）的推出，我们拥有了用于人工生成和现实世界主张检测与验证模型的基准数据集。然而，即使是最新技术水平的自然语言处理模型，在上述基准数据集上的表现仍然不佳。例如，在

FEVER 2018年共享任务（Thorne 等人，2018年）报告的准确性为 0。67[5]。模型在多模态共享任务 FEVEROUS（Aly 等人，2021年）上的表现更差：表现最佳的模型报告的准确性得分为 0.56[6]。同样，在 *CheckThat!* 2021年共享任务中，用于识别推文中可核查性的主张的最佳可核查性模型，对阿拉伯语主张和英语主张分别仅取得了 0.65 和 0.224 的平均精确率（Nakov 等人，2021b）。另一方面，用于识别辩论中值得核查的主张的最佳模型报告的平均精确率为 0.42。令人惊讶的是，Barrón-Cedeño 等人（2020年）提出的可核查性检测最佳模型，报告的平均精确率为 0.806（Williams 等人，2020年）。对于 CheckThat! 2021年（Nakov 等人，2021b）的事实核查任务，表现最佳的模型报告了 0.83 的宏平均 F1 分数。然而，排名第二的模型报告的 F1 分数仅为 0.50。鉴于顶尖系统与其他系统之间存在如此显著的性能差距，未来开展研究时，在更多数据集上对系统进行基准测试，将有助于更好地评估这些发现的普适性。

要直接比较在不同设置和不同数据集上评估的不同方法并非易事（Zeng 等人，2021年）。此外，自动化事实核查的流水线设计会造成潜在瓶颈。例如，在大多数数据集上，真实性预测任务的表现取决于主张检测任务的表现或检索到的证据质量。要系统地纳入事实核查流程中所有先前的子任务，还需要大规模基准测试（Zeng 等人，2021年）。

许多人工智能研究都面临一个根本性的权衡：是采用多样的问题表述，还是使用标准化基准测试来衡量进展。自动化事实核查研究也受到这一权衡的影响。尽管已有 FEVER 和 *CheckThat!* 等基准数据集，但基于这些基准数据集构建的大多数模型可能无法很好地推广到实际应用场景中。抽象且易于处理的问题表述或许有助于我们开发能够促进实际应用的技术。然而，实际应用需要投入大量超出研究范畴的工程工作。理想情况下，我们希望自动化事实核查研究在纳入多样问题表述的同时，继续朝着越来越贴近现实的基准数据集迈进。

5.2. 可解释方法

尽管可解释性和可解释性这两个术语经常互换使用，有时甚至被定义为同义词（Molnar，2020年），我们对可解释性与可解释性的区分与（Kotonya和Toni，2020a）类似。具体而言，可解释性指能够直接洞察人工智能系统各组成部分（如特征和变量）的方法，这通常需要对特定算法有所了解，而且往往是为模型调试等专家用途而设计的。可解释性则指无需参考系统的实际组成部分即可理解人工智能模型的方法。请注意，在任务形式化一节中，我们也讨论了如何解释真实性预测。这类解释的目标源于事实核查者的需求，即帮助读者理解事实核查结论。因此，相比可解释性，解释真实性预测更符合可解释性的范畴。当可解释性与可解释性之间的区别无关紧要时，我们遵循Vaughan和Wallach（2020年）的做法，采用可理解性（Vaughan和Wallach，2020年作为涵盖这两个概念的总称。

Sokol 和 Flach (2019年) 提出了机器学习应用用户体验设计的要求。Kotonya 和 Toni (2020a) 将这些要求扩展到事实核查语境中，并提出了可理解性的八项属性：可操作性、因果性、连贯性、语境完整性、交互性、无偏或公正、简约性和时间顺序性。

此外，自然语言处理中针对可解释方法还有三个专门维度（Jacovi 和 Goldberg，2020年）：

1. **可读性：**解释是否清晰？ 2. **合理性：**解释是否令人信服或具有说服力？
3. **忠实度：**解释是否忠实于模型的实际推理过程？

与自然语言处理中现有的可理解性方法相比（Wiegreffe 和 Marasovic，2021年），只有少数方法应用于现有的事实核查研究。下文仅介绍常见的可解释事实核查方法（Kotonya 和 Toni (2020a) 也提及了这些方法）。

基于注意力的可理解性尽管关于注意力是否是可靠的可理解性方法仍存在争论（Jain 和 Wallace，2019年；Wiegreffe 和 Pinter，2019年；Serrano 和 Smith，2019年；Bibal、Cardon、Alfter、Wilkens、Wang、François 和 Watrin，2022年），注意力仍然是现有事实核查深度神经网络方法中一种广受欢迎的方法。基于注意力的解释有多种形式：

1. 突出文章中的词元（Popat 等人，2018年）

[5]https://fever.ai/2018/task.html[6]
https://fever.ai/task.html

2. 利用与帖子相关的评论，突出证据中的显著摘录（Shu 等人，2019年） 3. 使用自注意力进行 n 元语法提取（Yang、Pentyala、Mohseni、Du、Yuan、Linder、Ragan、Ji 和 Hu，2019年） 4. 利用主张文本本身以外不同来源的注意力，例如推文的来源、转推传播和转推者属性（Lu 和 Li，2020年）

将规则发现作为解释 规则挖掘是知识库系统中常见的一种解释形式（Gad- Elrab 等人，2019年；Ahmadi、Lee、Papotti 和 Saeed，2019年）。这类解释可能更加全面，但正如上一节所述，由于底层知识库本身存在局限，并非所有陈述都能通过基于知识的方法进行事实核查。一些方法尝试提供通用规则挖掘，以解决这一局限（Ahmadi、Truong、Dao、Ortona 和 Papotti，2020年）。

将摘要作为解释 抽取式摘要和抽象式摘要都可以为事实核查提供解释。Atanasova 等人（2020a）提供自然语言摘要，以解释事实核查决策。他们探索了两种不同的方法：将解释生成和真实性预测作为独立任务，以及对二者进行联合训练。联合训练的效果不如单独训练。Kotonya 和 Toni（2020b）结合抽象式和抽取式方法，提出了一种新颖的摘要方法。Brand、Roitero、Soprano、Rahimi 和 Demartini（2018年）表明，联合训练预测和解释生成，并使用 BART（Lewis、Liu、Goyal、Ghazvininejad、Mohamed、Levy、Stoyanov 和 Zettlemoyer，2020年）等编码器—解码器模型，能够生成有助于大众更好地进行真实性评估的解释。

反事实与对抗方法 针对不透明模型的对抗攻击有助于识别模型中的盲点、偏差，并发现数据伪影（Ribeiro、Wu、Guestrin 和 Singh，2020年）。共享任务 FEVER 2.0（Thorne 等人，2019年）要求参与者设计生成对抗性主张的方法，以识别事实核查方法中的弱点。GPT-3 等自然语言生成模型可以协助制定对抗性主张。通过操控自然语言生成方法的输入，并将生成的文本限制在原始词汇范围内，可以更好地控制生成过程（Niewiński、Pszona 和 Janicka，2019年）。Atanasova、Wright 和 Augenstein （2020b）通过在输入文本中插入 n 元语法来生成主张。Thorne 和 Vlachos（2019年）尝试了多种对抗方法，例如基于规则的对抗方法、语义等价对抗规则（或 SEARS）（Ribeiro、Singh 和 Guestrin，2018年）、否定和基于释义的对抗方法。对抗攻击根据样例的效力 （正确性）以及系统性能的下降程度进行评估。虽然 SEARS 和释义等方法会损害系统性能，但人工构造的对抗样例具有更高的效力分数。

可解释方法（非黑箱） 一些事实核查研究采用白箱或本质上可解释的模型进行事实核查。Nguyen 等人（2018b，a）使用概率图模型，构建了一种用于事实核查的交互式可解释模型，允许用户直接覆盖模型决策。Kotonya、Spooner、Magazzeni 和 Toni（2021年）提出了一种可解释的图神经网络，用于在 FEVEROUS 数据集上进行可解释的事实核查（Aly 等人，2021年）。

局限性自然语言处理领域，尤其是事实核查中的可理解方法仍处于起步阶段。对Kotonya 和 Toni （2020a）的分析表明，大多数方法都未能满足本节前文提到的期望。具体而言，他们发现现有模型无一符合可操作性、因果性和时间顺序性的标准。他们还指出，目前没有方法明确分析解释是否公正。某些形式的解释（例如基于规则的三元组）不含完整句子，或仅包含信息片段，因此不带偏见（Kotonya 和 Toni，2020a）。

一些可解释方法针对任务的某种特定简化形式。例如，Kotonya 和 Toni(2020b); Atanasova 等人 (2020a) 均假设专家撰写的事实核查文章已经存在，并将这些事实核查文章的摘要作为解释。然而，在实际应用中，事实核查系统无法获取针对未知主张的此类文章。

在自动化事实核查中，大多数可理解方法侧重于解释结果，而不是描述如何得出结果的过程（Kotonya 和 Toni，2020a）。此外，事实核查流程中的所有任务并未得到同等程度的可解释方法关注。Kotonya 和 Toni （2020a）还认为，自动事实核查可以受益于能够阐明事实核查流程中较早子任务的结果如何影响后续子任务结果的可解释方法。

大多数可解释自然语言处理研究评估的是解释质量，而不是解释效用或忠实度。Jacovi 和 Goldberg （2020年）主张对可解释模型进行全面的忠实度评估。例如，尽管基于注意力的解释可能提供高质量的解释，但它们未必忠实。此外，还需要通过单独评估解释是否同时改善以下两方面来衡量解释效用：i）人类对模型的理解（Hase 和 Bansal，2020年）；以及 ii）人类在下游任务中的有效性（Nakov 等人，2021a）。此外，大多数可理解方法仅建立了从模型到人类的单向沟通。相反，解释可以通过建立双向反馈循环来提升模型和人类的表现。

5.3. 人在回路方法

人在回路（HITL）方法能够在利用人类智能处理复杂任务的同时，扩大自动化解决方案的应用规模。应用人在回路方法的方式多种多样，例如，将子任务分派给众包工作者（Demartini、Trushkowsky、Kraska、Franklin 和 Berkeley，2013年；Demartini、Difallah 和 Cudré-Mauroux，2012年；Sarasua、Simperl 和 Noy，2012年）、主动学习（Settles，2009年；Zhang、Lease 和 Wallace，2017年）、交互式机器学习（Amershi、Cakmak、Knox 和 Kulesza，2014年；Joachims 和 Radlinski，2007年），以及决策支持系统：人类根据模型输出和解释作出最终决策（Zanzotto，2019年）。

尽管人工智能领域广泛采用人在回路方法，但近期只有少数研究在事实核查中运用了此类方法。与事实核查流程的其他环节相比，在回路方法主要应用于真实性预测任务。例如，Demartini 等人（2020年）提出了一个用于打击网络错误信息的在回路框架。不过，他们只考虑了事实核查流程中两个子任务的混合方法：a）主张是否值得核查，以及 b）真实性判断（与真实性预测相同）。下文将根据系统在事实核查流程的各个子任务中如何利用人力，讨论现有的在回路方法。

主张检测、核查价值与优先级排序 作为主张检测任务的一部分，人们通常会监测社交媒体信息流中的谣言（Guo 等人，2022年）。Farinneya、Abdollah Pour、Hamidian 和 Diab (2021年) 在主张检测阶段采用基于主动学习的方法，以识别社交媒体上的谣言。对于传统监督方法而言，域内数据对于实现良好的谣言检测效果至关重要（Ahsan、Kumari 和 Sharma，2019年），但在现实场景中，开发初期可能无法获得充足的域内标注数据。主动学习等半监督方法有助于在使用较少数据点的情况下实现较高性能。实证结果表明，Tweet-BERT 与基于最小置信度的样本选择方法相结合时，使用的标注数据远少于监督方法，性能却与之相当（Farinneya等人，2021年）。

同样，Tschiatschek、Singla、Gomez Rodriguez、Merchant 和 Krause(2018年) 提出了一种人在回路方法，旨在自动汇总用户举报，并从被举报内容中筛选出少量内容推荐给专家进行事实核查。该方法基于贝叶斯推断，能够同时学习检测假新闻，并识别用户举报准确性随时间的变化。该方法的一项优势是，算法能够随着时间推移，更准确地识别用户举报的准确性。因此，该算法的性能也会随时间不断提升。该方法还能够抵御垃圾信息发送者。研究人员在公开可用的 Facebook 数据上运行模型；尽管大多数用户都具有对抗性，实验仍表明，该算法表现良好。

杜克大学的 Tech & Check 团队在主张核查价值这一环节引入了人在回路（Adair 和 Stencel，2020年）。为避免标记出错误的、看似值得核查的主张，人工专家会对 ClaimBuster 检测出的主张进行排序（Hassan、Zhang、Arslan、Caraballo、Jimenez、Gawsane、Hasan、Joseph、Kulkarni、Nayak 等人，2017b），筛选出事实核查人员认为更重要的主张，并通过电子邮件发送给多个组织。从本质上说，这种方法通过增加一道筛选环节，帮助事实核查人员确定要核查的主张的优先级。目前，PolitiFact 上已有多篇事实核查报道最初是由 Tech & Check 的邮件提醒的。

请注意，*CheckThat!* (Nakov 等人，2021b；Shaar 等人，2021b；Barrón-Cedeño 等人，2020年；Atanasova、Nakov、Karadzhov、Mohtarami 和 Da San Martino，2019a）是一个广受欢迎的共享任务，涉及主张检测、核查价值判断和优先级排序。然而，这类共享任务往往没有采用人在回路方法的参赛方案。针对人在回路方法设置共享任务，有助于鼓励更多解决方案，从而弥补仅依赖模型的方法的局限性。

证据检索与真实性预测 人在回路事实核查领域的大多数工作都着眼于真实性预测，只有少数工作将证据检索作为一项独立任务加以考虑。尽管关于信息检索中人在回路方法的文献已有一定积累（Chen 和 Jain，2013年；Demartini，2015），但据我们所知，尚无研究将这一方向应用于事实核查。

Shabani、Charlesworth、Sokhn 和 Schuldt （2021年）利用人在回路方法，就声明来源、作者、信息和拼写（SAMS）提供反馈。标注员需要回答四个是非问题，以判断文章是否有来源、作者、清晰且不带偏见的信息，以及是否存在拼写错误。此外，该研究还将人类提供的特征整合到机器学习流程中，使准确性提高了 7.1%。然而，该研究是在一个规模较小的数据集上进行评估的，其中的主张与新冠肺炎有关。目前尚不清楚这种方法能否推广到该领域之外。此外，在这项工作中，还可以通过自动化拼写检查和语法检查来减少额外的人力投入。在现实生活情境中，SAMS 可能相当有限，因为大多数精心编造的错误信息往往看起来像真实新闻。模型生成的假新闻可以成功欺骗标注员（Zellers、Holtzman、Rashkin、Bisk、Farhadi、Roesner 和 Choi，2020年），因此 SAMS 也可能无法标记此类假新闻。

Qu、Barbera、Roitero、Mizzaro、Spina 和 Demartini(2021a) 以及Qu、Roitero、Mizzaro、Spina 和 Demartini(2021b) 探讨了如何利用人类与机器置信度分数，构建用于事实核查的人在回路方法。他们考察了标注员明确自我报告的置信度，并根据十名众包工作者之间的标准差计算隐式置信度。模型置信度则通过自助法获得：对模型进行十次不同版本的抽样（Efron andTibshirani，1985），再计算 Softmax 层返回分数的标准差。评估结果表明，众包工作者和模型明确表达的置信度都不是准确分类决策的有效指标。尽管众包工作者和模型会犯不同的错误，但没有明确迹象表明置信度与准确性有关。不过，他们发现，众包工作者的隐式置信度可作为有用信号，用于判断何时应请专家收集标签。较新的一项研究表明，由十人组成、政治立场均衡的众包群体，其结果与三名事实核查人员的平均评分相关（Allen、Arechar、Pennycook 和 Rand，2020年）。Gold、Kovatchev 和 Zesch (2019年) 还发现，十人众包群体的标注结果与三名标注员对文本蕴含的判断相关，而文本蕴含被真实性预测模型所采用。

一系列研究表明，众包工作者能够可靠地识别错误信息（Roitero、Soprano、Fan、Spina、Mizzaro 和 Demartini，2020a；Roitero、Soprano、Portelli、Spina、Della Mea、Serra、Mizzaro 和 Demartini，2020b；Soprano、Roitero、La Barbera、Ceolin、Spina、Mizzaro 和 Demartini，2021年）。此外，Roitero 等人（2020b）表明，众包工作者不仅能够识别虚假主张，还能检索恰当的证据来佐证其标注。这项研究的一个不足之处是，只要求用户提供一个网址作为证据。然而在实践中，事实核查可能需要对多个信息来源进行推理。尽管这些研究并未提出新颖的人在回路解决方案，但它们提供了充分的实证证据和见解，说明可以在事实核查流程的哪些环节可靠地引入众包工作者。

Nguyen 等人（2018b）提出了一种结合众包标注与机器学习的联合建模方法，用于检测文本声明的真实性。该模型的主要优势在于，它假设所有标注员都可能出错，因为事实核查是一项艰巨的任务。另一项优势是，该模型允许用户将自己的知识导入系统。此外，这种人在回路方法可以按需收集众包立场标签，并将其纳入真实性预测。实证评估表明，该方法具有出色的预测性能。一项后续研究提供了一种用于事实核查的交互式人在回路工具（Nguyen 等人，2018a）。

Nguyen、Weidlich、Yin、Zheng、Nguyen 和 Nguyen（2020年）提出了一种人在回路系统，以最大限度地减少用户的工作量和成本。用户会验证算法预测，但只需以最低成本验证最有助于改进系统的预测。该系统为用户提供引导式交互，并随着用户的参与而逐步改进。

需要指出的是，关于通过众包进行真实性判断的研究仍处于早期阶段。人口统计特征、政治倾向、评估者专业知识的判定标准（Bhuiyan、Zhang、Sehat 和 Mitra，2020年）、认知因素（Kaufman、Haupt 和 Dow，2022年）乃至评分量表（La Barbera、Roitero、Demartini、Mizzaro 和 Spina，2020年）等不同因素，都会导致其与专家评分的一致程度有所不同。Bhuiyan 等人（2020年）概述了研究方向，旨在针对不同类型的错误信息设计更完善的众包流程，从而成功利用众包工作者。

解释真实性预测事实核查中的人在回路系统通常会使用真实性解释来纠正模型错误。如前所述，Nguyen 等人(2018a)提出了一种可解释模型，使用户能够在模型出错时传授自己的知识。实证评估表明，用户能够将自己的知识传授给系统。同样，Zhang、Rudra 和 Anand (2021b)提出了一种从解释中收集用户反馈的方法。请注意，该方法根据检索到的证据及其立场来解释真实性预测结果。用户针对检索到的证据的立场和相关性提供反馈。该方法采用终身

学习，使系统能够随着时间推移不断改进。目前尚无对该系统的实证评估，无法确定这种方法的有效性。

尽管自然语言生成模型正变得越来越好（Radford、Wu、Child、Luan、Amodei、Sutskever等人，2019年），但生成抽象式事实核查解释仍处于起步阶段（Kotonya 和 Toni，2020b）。可以利用人在回路方法撰写报告，为事实核查解释提供依据。

局限性 在审视了不同事实核查任务中现有的人在回路方法后，我们还列出了以下几项局限。首先，一些人在回路方法采用若干可解释模型来整合人类输入，但由此得到的模型性能不如最新的深度学习模型（Nguyen 等人，2018b，a）。

Farinneya等人（2021年）采用人在回路方法，利用有限的标注数据扩大谣言检测的规模。尽管该方法能以少样本方式将算法泛化到新主题，但其弱点之一是，来自其他领域或主题的数据会导致模型性能出现较大波动。因此，在模型训练中引入域外数据时，域内模型性能可能会下降。这一问题可能会妨碍模型在实践中的泛化能力，尤其是在无法明确划分主题领域的情况下。

更重要的是，目前缺少关于如何应用人在回路的事实核查方法以推动实际应用的实证研究。尽管人在回路方法提供了一种让人类参与模型开发过程的机制，但在现实世界应用场景中采用这一方法时，必须考虑可用性、可理解性和信任等多种人为因素。事实核查是一项时间敏感型任务，需要专业知识来处理来自多个来源的复杂信息（Graves，2017年）。事实核查人员和政策制定者往往对任何自动化或半自动化解决方案持怀疑态度，因为这类研究需要人类的创造力和专业知识（Arnold，2020年；Micallef 等人，2022年）。因此，需要更多实证证据来评估将不同的人在回路方法应用于自动化事实核查的有效性。

6. 现有事实核查工具

上一节中，我们回顾了当前用于事实核查的自然语言处理技术的详情。随后，我们将对自动化事实核查的回顾扩展至人机交互文献，并讨论将事实核查应用于现实世界工具、以协助人工事实核查员的现有实践。简而言之，目前缺少从以人为中心的视角对事实核查工具进行的全面综述。此外，我们发现，在这一领域，人力劳动与人工智能工具之间的分工仍不明晰。研究问题包括但不限于：1）自然语言处理工具如何促进人类在不同事实核查任务中的工作？2）我们如何纳入用户需求并发挥人类专业知识的作用，以指导自动化事实核查的设计？

本节将考察当前应用于事实核查不同阶段的现实工具，并阐明这些工具的主要使用场景。我们认为，应开展更多关于构建自动化事实核查的人为因素研究，例如用户研究、以人为中心的设计和可用性研究，以提高自动化事实核查的实际应用。这些研究有助于我们确定在现实自然语言处理技术中应用可解释方法和人在回路方法的设计空间。

6.1. 主张检测与优先级排序

主张检测的第一步是获取可能需要核查的内容。在 WhatsApp、Telegram 和 Signal 等端到端加密平台上，基于众包的举报热线在识别其他途径无法获取的可疑内容方面发挥着重要作用（Kazemi 等人，2021b）。另一个例子是 Meedan 推出的 *Check* [7]，这是一种举报热线服务工具，也有助于事实核查人员监测社交媒体上的假新闻。用户在 Facebook 等社交媒体平台上标记可疑内容，也是一项有价值的信号，可用于识别此类内容；而 Twitter 的 BirdWatch 等众包项目还能进一步帮助筛选主张，并确定后续调查的优先级。

在寻找并选择待核查主张的阶段，事实核查人员会评估主张与事实核查相关的质量，并决定是否对其进行事实核查（Graves，2017年；Micallef 等人，2022年）。用于主张检测、主张匹配和核查价值判断的自然语言处理模型，有助于辅助上述决策过程。然而，将这些模型整合到帮助事实核查人员确定核查优先级和核查对象的现实工具中，还需要投入更多个性化工作。Graves(2018a)指出，设计上述模型时，必须考虑事实核查人员的组织利益、利益相关方需求以及不断变化的新闻趋势。

[7] https://meedan.com/check

作为主张的一项事实核查属性，可核查性可以通过客观方式进行分析：判断主张是否包含一个或多个可验证的所谓事实（第3.1节）。事实核查人员发现，将用于识别可核查主张的模型应用于现有工作流程很有帮助，因为在选择待核查主张时，这些模型可以帮助他们过滤无关内容和不可核查的主张（Arnold，2020年）。ClaimBuster 是知名的主张检测工具之一，旨在从大规模文本内容中找出可核查主张（Hassan 等人，2017a）。主张检测也可以整合到语音识别工具中，以便从实时语音中识别主张（Adair，2020年）。

此外，如果某个主张已经经过事实核查，事实核查人员便可跳过该主张，优先处理尚未核查的主张。作为一项相对较新的自然语言处理任务，主张匹配已集成到一些现有的现成搜索引擎或事实核查工具中，以帮助事实核查人员查找已核查事实主张。例如，Google Fact Check Explorer[8] 可以通过将类似的事实核查内容与用户输入的查询进行匹配，检索已核查事实主张。同样，借助 Meedan 的 *Check*，如果用户提交的提示中包含已核查过的假新闻，该工具会进一步帮助事实核查人员检索此前的事实核查结果并发送给用户。

是否对某个主张进行事实核查，取决于某个组织的目标和利益。为主张检测构建的工具需要考虑这些利益。例如，Full Fact 开发了一种主张检测系统，可将主张划分为不同类别，例如数量、预测、相关性或因果关系、个人经历，以及法律或操作规则（Konstantinovskiy 等人，2021年）。这些主张类别由其事实核查人员设计，旨在满足他们在实时事实核查情境下核查英国政治新闻的需求。识别某些主张（例如涉及数量、相关性或因果关系的主张）可能特别有助于事实核查人员评估政治人物声明和主张的可信度。该系统还帮助事实核查人员定制下游任务，例如事实核查任务分配以及针对统计主张的自动化核验（Nakov 等人，2021a）。

事实核查人员也会使用社交媒体监测工具来寻找待核查的主张，例如 CrowdTangle、TweetDeck 和 Facebook（未命名）的事实核查工具，但这些工具在检测可核查主张方面并不十分有效。一些事实核查人员报告称，Facebook 事实核查工具标记的主张中，实际上只有约 30% 是可核查的（Arnold，2020年）。一个容易实现的改进方向是，将主张检测模型集成到这些社交媒体监测工具中，让事实核查人员更容易识别既具有传播力又可核查的主张。此外，这些工具应能根据事实核查人员的兴趣和利益相关方的需求，帮助他们找到特定人物、机构或部门，从而更好地识别并优先处理真正值得核查的主张。在实现这类系统时，一个重要问题是如何衡量主张的传播力及其随时间的变化。

将真实性预测集成到先前提到的事实核查工具中也很有用，因为事实核查人员可能最关注那些[9] 可疑且真假难辨的主张（显而易见为真或为假的主张可能无需事实核查）。不过，也应向事实核查人员提供用于生成此类预测的信息或数据点。如果模型用于预测主张真实性的来源、证据、传播模式或其他语境信息能够向事实核查人员清晰说明，他们就能综合评估这些指标，更全面地确定主张的优先级。

6.2 证据检索工具

确定要核查的主张并确定优先级后，事实核查人员会通过三项主要活动调查主张：1）拆解主张；2）寻找证据；3）追溯主张的来源及其传播过程。需要注意的是，在事实核查流程中，这三项活动借助不同的信息检索工具彼此交织。事实核查人员会将主张拆解为子问题，以此搜索证据。在调查主张时找到的证据可能会进一步修改或补充这些子问题（Singh 等人，2021年）。事实核查人员通过在线搜索反复调查主张，重建主张的形成和传播过程，从而评估其真伪（Graves，2017年）。本节将探讨现有信息检索工具的实用性，包括现成的搜索引擎和领域专用数据库，这些工具可在各项活动中为事实核查人员提供帮助。

声明分解并非定性研究人员在事实核查研究中报告或分析过的一项具体活动，但我们可以从事实核查组织对其方法的描述[10]，以及事实核查人员在事实核查中处理复杂主张的方式[11]中了解更多细节。声明分解是指事实核查人员如何解读主张中的模糊术语，并界定事实核查范围以寻找证据。分解

[8] https://toolbox.google.com/factcheck/explorer [9] https://www.factcheck.org/our-process/ [10] https://leadstories.com/how-we-work.html [11] https://www.factcheck.org/2021/10/oecd-data-conflict-with-bidens-educational-attainment-claim/ 在这次事实核查中，事实核查人员拆解了拜登总统所说的“发达经济体”和“年轻人”分别指什么。对这两个术语的定义方式会直接影响事实核查结果。

主张需要事实核查人员具备敏锐的好奇心和新闻判断力，而这些能力需要经过多年的实践才能培养。不幸的是，我们尚未发现任何能够促进这一过程的现有工具。

分解主张的传统方法是提出子问题。近期的自然语言处理研究将这一过程构建为问答任务，以此进行模拟（Fan 等人，2020年；Chen 等人，2022年）。研究人员从现有的事实核查中提取论证依据，并通过众包征集子问题，以分解主张。对于自动事实核查，这项自然语言处理任务或许能通过自动将主张分解为更小、可核查的查询，大幅提升证据检索的性能（Chen 等人，2022年）。尽管自然语言处理难以匹敌专业事实核查员的能力，但它或许有助于扩大传统人工事实核查流程的规模。它也能帮助公众、新晋事实核查人员或记者更有效地调查复杂主张并查找证据。

事实核查人员寻找证据的方式通常遵循领域特定的报道流程，例如联系专家，或从可靠来源查找特定文档（Graves，2017年；Micallef 等人，2022年）。大多数事实核查人员不会在网上随意搜索，而是会列出一系列用于查找证据的可靠来源。用于搜索领域数据集的工具也能帮助事实核查人员查找证据。例如，Li、Fang、Lou、Li 和 Zhang(2021年)构建了一个分析型搜索引擎，用于检索新冠肺炎新闻数据，并以易于理解的表格形式对其进行汇总。该系统能够将分析查询分解为结构化实体，并从新闻数据中提取定量事实。此外，如果系统在域内数据集上的证据检索足够准确，便可进一步自动核查与领域相关的主张。第6.3节将介绍更多真实性预测的应用案例。

事实核查人员主要使用现成的搜索引擎，如谷歌、必应等，通过公开可访问的文档追溯主张的来源（Beers、Haughey、Melinda、Arif 和 Starbird，2020年；Arnold，2020年）。其他数字数据集，如*LexisNexis* 和 *InternetArchive*，也有助于事实核查人员追溯主张来源。为了捕捉主张的形成与变化，搜索引擎不应只过滤无关内容，还应检索主题相关且具有证据价值的内容。 Hasanain 和 Elsayed (2021年)报告称，谷歌检索到的大多数主题相关网页并不包含证据性信息，例如统计数据、引语、实体或其他类型的事实。此外，社交媒体平台内置的大多数搜索引擎，如 Twitter 和 Facebook，只会筛选“易于传播”的内容，而非“可信”的内容（Alsmadi、Alazzam 和 AlRamahi，2021年）。

此外，这些现成的搜索引擎不支持多语言搜索，因此，如果主张是从其他语言翻译而来，事实核查人员就很难追踪这些主张（Graves，2017年；Nakov 等人，2021a）。自然语言处理研究人员已开始使用多语言嵌入模型来表示不同语言中与主张相关的文本，并将其与现有的事实核查相匹配（Kazemi、Gaffney、Garimella和Hale，2021a）。这项工作不仅有助于事实核查人员更轻松地从其他语言中找到已核查事实主张，也有助于考察媒体如何在不同语言和社会政治语境中转化和重塑主张。

6.3. 主张核验的领域专用工具

如第3.2节和第3.3节所述，大多数核验预测模型都以收集到的证据和主张为基础。要构建端到端的主张核验系统，自然语言处理开发人员需要构建包含主张和证据的领域专用数据集。与包含多个论据且需要分解的复杂主张不同，具有简单语言结构并附有证据，或包含统计事实的主张可以自动核验（Nakov 等人，2021a）。

Karagiannis、Saeed、Papotti和Trummer (2020年) 构建了 CoronaCheck，这是一款搜索引擎，可通过检索由专家整理的官方数据，直接核验与新冠肺炎相关的统计主张（Dong、Du和Gardner，2020年）。Full Fact（波因特研究所，2021年）也采用了类似方法，通过从英国议会报告和国家统计数据中检索证据，核验统计性宏观经济声明。此外，Wadden 等人(2020年) 利用包含证据的摘要构建了科学声明核验流程，以核验给定的科学主张。

然而，事实核查人员在实际使用这些特定领域的核验工具时，仍会遇到一些陷阱。例如，CoronaCheck 工具无法核查“德尔塔变异株导致的死亡人数多于阿尔法变异株”这一主张，单纯是因为数据库中没有新冠变异株的细粒度死亡统计数据。此外，核查统计或科学主张可能只是核查更复杂主张流程中的一部分，而这要求事实核查人员结合语境，判断先前统计或科学核查的真实性。总的来说，在有条件时，特定领域的工具显然很有价值；但在实践中，这些工具往往并不完善，单靠它们不足以核查复杂主张。

7. 讨论

在这篇综述中，我们 1）从任务形式化开始到工具采用结束，横向梳理了将自然语言处理技术应用于事实核查的研究；同时 2）纵向讨论了自然语言处理技术在事实核查流程各个步骤中的能力与局限。我们发现，缺乏能够兼顾这两个方面、以协助事实核查人员的研究。可解释方法和人在回路方法从以人为中心的视角出发，利用人类与计算智能，但仍需提供可操作的指南，帮助人们运用这两种方法来设计实用的事实核查工具。本节将提出若干研究方向，以探索运用自然语言处理技术协助事实核查人员的设计空间。

7.1. 混合主动式事实核查中的人机任务分配

事实核查实践已成为一种复杂且分布式的计算机中介工作（Graves，2018a）。尽管 Graves（2017年）将传统的新闻工作者事实核查流程分解为五个步骤，但实际核查某项主张的情况要复杂得多（Juneja 和 Mitra，2022年）。事实核查人员会动态且多样地采用各种人工智能工具，以完成不同的事实核查任务（Arnold，2020年；Beers 等人，2020年；Micallef 等人，2022年）。

研究人员和从业者日益认为，未来的事实核查应采取混合主动式实践，由人类执行特定任务，同时由机器接手其他任务（Nguyen 等人，2018a；Lease，2020年；Nakov 等人，2021a）。要将这种混合且动态的人机协作融入现有的事实核查工作流程，就需要明确人类与人工智能之间的任务安排，并了解每项任务的预期成果和评判标准。此外，针对不同事实核查任务设计混合主动式工具，需要对人类和人工智能的任务定义得更加细致（Lease，2018年，2020年）。在第5.3节中，我们讨论了若干研究，这些研究强调了人类在事实核查工作流程中的作用，例如：a）人类专家从主张检测工具中筛选出值得核查的主张（Hassan 等人，2017b），并将其交给事实核查人员；b）请众包工作者判断哪些是可靠的主张来源（Shabani 等人，2021年）；或 c）标记潜在的错误信息（Roitero 等人，2020b），以改进真实性预测。以上所有人类活动都是混合主动式事实核查流程中的微任务。

此前关于众包的研究表明，可以有效地拆解学术研究流程，并利用众包工作者参与规模较小的研究任务（Vaish、Davis 和 Bernstein，2015；Vaish、Gaikwad、Kovacs、Veit、Krishna、Arrieta Ibarra、Simoiu、Wilber、Belongie、Goel 等人，2017年）。鉴于这些证据，我们也可以将传统事实核查流程中的子任务拆分为更细致的任务。因此，关键研究问题包括：a）如何设计这些微任务以支持事实核查的各项子任务；b）在不同微任务中，人类和人工智能分别适合承担哪些角色？

为了有效协调人类与人工智能的工作，研究人员需要了解人类和人工智能各自的角色，以及二者将如何互动，因为这会直接影响人类是否决定采纳人工智能建议（CimolinoandGraham，2022年）。通常，如果人工智能旨在辅助高风险决策任务，例如再犯预测（Veale, Van Kleek and Binns，2018年）和医学治疗（Cai,Winter,Steiner, Wilcox and Terry，2019年），那么风险和信任将是人们采用此类人工智能助手时需要考虑的重要因素（Lai, Chen, Liao,Smith-Renner and Tan，2021年）。在事实核查的语境中，如果人工智能直接预测主张的判定结果，事实核查人员可能会自然地对人工智能如何做出此类预测持怀疑态度（Arnold，2020年）。另一方面，如果人工智能只帮助筛除不可核查的主张，例如观点和个人经历，事实核查人员可能更愿意使用此类自动化功能，而较少关心人工智能是如何实现这一功能的。判断主张真伪是事实核查人员面临的一项高风险决策任务，而筛除不可核查的主张则是一项不那么重要但十分繁琐的任务，事实核查人员希望通过自动化来减轻这项工作。因此，人类对人工智能的接受度会因人类对分配给人工智能的任务的评估方式而异，从而产生不同的人为因素，例如信任、透明度和公平性。研究人员需要明确说明或将这些人为因素分解为不同的关键变量，以便在模型开发过程中进行测量。深入理解人类与人工智能之间的任务关系后，研究人员便可以进一步探讨如何运用可解释方法，或采用人在回路系统与自动化解决方案来开展事实核查。在此，我们列出若干涉及混合主动任务的具体研究主题，包括：a）借助众包工作者评估主张难度；b）将一项主张拆解为多跳推理任务，并让众包工作者查找与子声明相关的信息；以及c）设计微任务，对网络搜索检索到的大量文档进行解析，以识别包含真实性预测所需证据的来源。

7.2. 面向事实核查者的自然语言处理技术以人为中心的评估

本节首先提出用于评估系统的人为因素关键指标（即测量什么以及如何测量）：准确性、时间、模型理解和信任（第7.2.1节）。随后，我们进一步提出一份用于事实核查以人为中心评估的实验方案模板（第7.2.2节）。

7.2.1. 指标

准确性 大多数事实核查用户研究都将任务准确性视为用户的首要目标（Nguyen 等人，2018a；Mohseni、Yang、Pentyala、Du、Liu、Lupfer、Hu、Ji 和 Ragan，2021年）。非专业用户（即社交媒体用户或其他形式的内容消费者）可能最关注真实性判定结果及其依据，而事实核查者往往希望在工作流程中交替使用自动化和人工投入（Arnold，2020年；Nakov 等人，2021a）。因此，我们需要采用更细致的方法来衡量准确性，而不应只关注最终的真实性预测准确率。要进行细粒度的准确性评估，同样至关重要的是衡量事实核查员的准确性，尤其是在他们使用事实核查工具完成的子任务中。

假设事实核查流程中的所有子任务都有“真实标签”，就可以通过比较用户的答案与真实标签来计算准确性。请注意，衡量子任务层面的准确性比衡量端到端事实核查的准确性更复杂。可以针对每个子任务分别开展实验，以评估子任务层面的准确性。假设我们关注的是了解用户检测主张核查价值的表现。在这种情况下，我们需要收集专门针对主张核查价值任务的额外数据。

在某些情况下，为了评估，可以将多个子任务合并。例如，Miranda、Nogueira、Mendes、Vlachos、Secker、Garrett、Mitchel 和 Marinho (2019年) 通过衡量以下两个关键变量来评估其工具对记者的效用：a）检索到的证据的相关性；b）预测立场的准确性。这种方法能为证据检索、立场检测以及最终的事实核查任务提供重要见解。具体指标细节需要根据工具的可供性作相应调整。

请注意，时间和准确性指标都需要控制声明属性。例如，如果某项主张之前已经过事实核查，那么对这类主张进行事实核查所需的时间会更少。另一方面，一项较难评估的新主张则需要更多时间。

模型理解 事实核查人员希望了解他们所使用的工具。例如，Arnold (2020年) 指出，事实核查人员表示有必要了解 CrowdTangle 用于检测各个社交媒体平台上病毒式传播内容的算法。同样，Nakov 等人(2021a) 观察到，不同组织所使用的事实核查工具需要提高系统透明度。Lease (2018年) 认为，透明度对于非专业用户理解底层系统并作出知情判断同样重要。尽管这不是与用户表现相关的关键变量，但它对于实际应用十分重要。

为了衡量理解程度，可以要求用户在李克特量表上自我报告其理解水平。然而，仅询问参与者是否理解该算法并不是一个充分的指标。例如，这并不能说明参与者是否能够模拟工具行为（Hase 和 Bansal，2020年）。我们根据先前的研究，提出以下衡量模型理解程度的步骤（Cheng、Wang、Zhang、O’Connell、Gray、Harper 和 Zhu，2019年）。

1. **决策预测**。 为了全面了解用户对工具的理解，可以向用户提供主张，并询问：“该工具会为这条主张分配什么标签？”
2. **替代预测**。 衡量理解程度的另一种方法是了解输入的变化如何影响输出，例如询问用户：当输入的某些部分发生变化时，工具会为一条主张分配什么标签。设想有一种工具，会向用户展示它在得出真实性结论时所考虑的证据。那么，如果替换其中某些证据，这会如何反映在模型预测中？

信任 对于实际应用而言，事实核查工具能否获得所有用户群体的信任至关重要。虽然模型理解通常与信任呈正相关，但仅有理解可能不足以建立信任。在这一领域，事实核查人员和记者可能较少信任算法工具（Arnold，2020年）。另一方面，也存在过度信任的风险，即用户盲目遵循模型预测（Nguyen 等人，2018a；Mohseni 等人，2021年）。为了最大限度地提高工具的有效性，我们希望用户既不全盘否定模型预测（完全怀疑），也不盲目遵循所有模型预测（完全信任）。相反，校准用户信任对于最有效地使用工具至关重要。我们建议衡量一种校准信任 （Lee 和 See，2004）：用户遵循正确模型决策、推翻错误模型决策的频率。

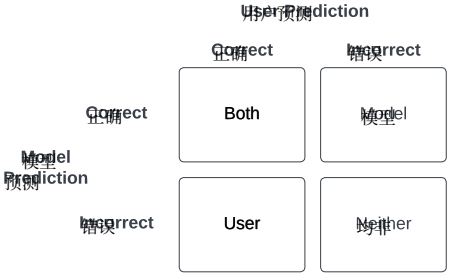

图2： 用户预测与模型预测相对于真实标签（Gold）的混淆矩阵。我们假设模型预测会提供给用户，由用户决定接受还是拒绝模型预测。左上象限（两者）涵盖用户正确遵循模型预测的情况。右上象限（模型）表示模型判断正确、但用户错误地拒绝模型决策的情况。左下象限 用户 表示用户正确拒绝错误模型预测的情况。右下象限 两者皆非 表示用户错误地接受错误模型预测的情况。以这种方式量化用户与模型的预测，可以衡量校准信任：用户遵循正确模型决策并推翻错误模型决策的频率。

为了衡量校准信任，我们设想了图2所示的混淆矩阵。行表示正确与错误的模型预测，列表示正确与错误的用户预测。盲目遵循所有模型预测的用户，其行为将完全落在主（主要）对角线上；而怀疑一切并拒绝所有模型预测的用户，其行为则将完全落在次对角线上。理想用户的行为将完全落在第一列：接受所有正确的模型预测，并拒绝所有错误的模型预测。为促进有效的人机协作，人工智能工具应帮助人类用户建立稳固的校准信任，以便根据具体情况恰当地信任或不信任模型预测。

除了校准信任之外，还可以借鉴人机信任研究领域的方法来衡量定量信任（Lee 和 Moray，1992年）。例如，Cheng 等人（2019年）将先前的研究方法改编为 7 点李克特量表。类似的量表也可用于评估用户对事实核查工具的信任。例如，我们可以设计五个不同的李克特量表，以衡量用户对以下陈述的认同（或反对）程度：

- 我理解这个事实核查工具。
- 我可以预测这个工具会如何运行。
- 我相信这个工具能够应对不同的事实核查任务 k.
- 我信任这个工具作出的决策。
- 我可以指望这个工具提供可靠的事实核查决策。

其他因素 用户之间的个体差异可能导致实验结果出现显著差异。例如，技术素养（Cheng 等人，2019年）、对主张的既有了解，以及用户的政治倾向（Thornhill、Meeus、Peperkamp 和 Berendt，2019年）都可能影响用户在使用事实核查工具时的任务表现。因此，在研究设计中考量这些因素很有价值。例如：

1. **技术素养：**用户对常见技术工具（如推荐引擎、垃圾邮件检测器）的熟悉程度，以及其编程经验（Cheng 等人，2019年），以及对现有事实核查工具的熟悉程度。2.**媒体素养：**用户对以下内容的熟悉程度：1）事实核查流程；2）PolitiFact 和 FactCheck.Org 等知名组织发布的事实核查。3. **人口统计特征：**用户的教育水平、性别、年龄和政治倾向。

仅靠量化指标并不足够，因为它们无法捕捉工具融入事实核查者工作流程的实际效果中的某些细微差别。例如，即使用户理解并信任某个

工具的工作原理，我们仍不清楚他们为何如此。因此，可以在研究结束时向用户提出几个开放式问题，以收集定性见解。此类问题可以包括：

1. 描述你对该工具的理解。它的设计是否有任何特定方面有助于或妨碍你理解其工作方式？2. 你为什么信任或不信任该工具？3. 除了本研究之外，你会使用该工具吗？如果会，你会如何使用？

7.2.2. 实验方案

衡量上述指标的一种策略是设计一项混合方法研究。下面我们概述此类研究的模板。假设研究目标是衡量用户使用一种新工具（我们称之为 **工具*A***）进行事实核查时的表现，并将其与现有工具（**工具*B***）进行比较。现实世界中的事实核查任务可能会受到用户对待核查主张的先验认知影响。因此，为考虑这些先验认知，被试内研究方案或许更为合适（Shi、Bhattacharya、Das、Lease 和 Gwizdka，2022年）。

1. **任务前阶段**：首先会要求用户对一组主张进行事实核查。为此，首先会要求用户使用 e
一个现有的**工具 *B***。在此阶段，可以根据具体 r
使用场景，将工具 B 替换为不同的基线，从非专业用户进行简单的网络搜索，到事实核查人员使用的专有工具，d
以及记者使用的工具。此阶段会要求用户进行出声思考。2. **学习阶段**：在此阶段，用户将熟悉新工具（**工具*A***）。用户需要 o
对一组与第一组不同的主张进行事实核查。用户还可以看到真实标签，以形成 a
关于工具可能会犯哪类错误的先验认识。此处的主张将随机选取，以反映多样性 l
功能。此外，还会向用户提供工具性能指标作为补充信息。用户 s
在此阶段可以就该工具提问。3. **预测：** 现在会要求用户对上文第 1 步中的相同主张进行事实核查，但这次他们将 e
被要求借助**工具*A***。用户需要在此阶段边思考边说出自己的想法。用户也可以简单地猜 s
出答案并取得较高的准确性分数。因此，第（1）和（3）阶段所选的主张将 d
包含主张的集合，其中真阳性、真阴性、假阳性和假阴性样本的数量相同。此想法借鉴了以往的研究（Hase 和 Bansal，2020年）。4. **任务后调查**：现在会要求用户填写一份简短的调查问卷，以了解他们的信任、理解、技术素养、媒体素养和人口统计信息。5. **任务后访谈**：完成上述步骤后，将通过开放式问题访谈用户，以深入了解他们对该系统的理解和信任。

这些衡量指标和研究方案可用于评估任何新事实核查系统相较于现有系统或做法的表现。具体内容可能因目标用户群体和工具的预期用途而异。上文以整个事实核查流程为例，说明了我们的实验方案。不过，只要我们掌握该子任务结果的真实标签，这种方法也可应用于自动化事实核查的其他子任务。例如，假设有人提出了一种新的主张检测工具，可从线索举报渠道获取主张（Kazemi 等人，2021b）。目前，事实核查人员使用现有的声明匹配算法，筛除已经过事实核查的主张。现在，如果我们将上文的**工具*B*** 替换为现有的声明匹配算法，并将**工具*A*** 替换为所提出的主张检测工具，就可以采用上文提到的方案。总之，我们可以评估用户使用新工具完成声明检测任务时，相较于使用现有工具，在准确性、时间、理解和信任方面的表现。

虽然我们提出了一个理想且全面的评估方案，用于评估新的事实核查工具，但请注意，实际应用的方案可以根据参与者所需投入的时间以及实验开展成本进行调整。

8. 结论

本文回顾了使用自然语言处理进行自动化事实核查的最新技术水平，重点介绍现有任务形式化、数据集构建和建模方法的进展与局限。我们还部分探讨了将这些自然语言处理任务应用于协助人工事实核查员的现实世界工具中的现有做法。近年来，我们看到自然语言处理在自动化事实核查方面取得了显著进展。

事实核查流程的不同环节引入了广泛的任务、数据集和建模方法。此外，随着 Transformer 和大型语言模型的最新发展，各项任务的模型准确性均有所提升。然而，现有基准数据集上的最新技术水平模型——例如 FEVER 和 *CLEF!*——可能尚未达到实际应用和部署的要求。

为解决这些局限，我们主张开发用于事实核查的混合式人在回路系统。作为起点，我们可以考虑将现有自然语言处理任务的目标从完全自动化转向决策支持。与全自动系统不同，混合系统会让人在回路中参与，并促进人机协作（Bansal 等人，2019年；Bansal, Wu, Zhou, Fok, Nushi,Kamar, Ribeiro and Weld，2021b；Bansal, Nushi, Kamar,Horvitzand Weld，2021a）。采用混合系统有助于：a）扩大人工决策的规模；b）以人工准确性增强机器学习能力；c）减轻机器错误造成的意外后果。此外，我们需要新的基准数据集和评估实践，以衡量自动化系统和混合系统如何提升下游人工准确率（Smeros, Castillo and Aberer，2021年；Fan 等人，2020年），并提高事实核查效率。

致谢

本研究部分经费由Knight基金会、美光基金会、Wipro以及Good Systems[12]，德克萨斯大学奥斯汀分校的一项旨在开发负责任的人工智能技术的重大挑战项目提供。本文所述观点仅代表作者观点，并不反映资助部门的立场。

参考文献

Adair, B., 2020. Squash 报告卡：在国情咨文期间取得的改进……以及人类将如何让我们的人工智能变得更聪明——杜克大学记者实验室。网址：https://reporterslab.org/squash-report-card-improvements-during-state-of-the-union-and-how-humans-will-make-our-ai-smarter/。 Adair, B., Stencel, M., 2020. 自动化新闻的一课：让人类回归 | 尼曼新闻实验室。https://www.niemanlab.org/2020/07/a-lesson-in-automated-journalism-bring-back-the-humans/。（访问日期：02/08/2022）。 Ahmadi, N., Lee, J., Papotti, P., Saeed, M., 2019. 基于概率答案集编程的可解释事实核查，载于：在线真相与信任会议。 Ahmadi, N., Truong, T.T.D., Dao, L.H.M., Ortona, S., Papotti, P., 2020. Rulehub：知识图谱规则公共语料库。《数据与信息质量期刊》（JDIQ）12，1–22。 Ahsan, M., Kumari, M., Sharma, T., 2019 128, 45 58。通过深度学习检测 Twitter 上语境变化的谣言。《国际先进科学与技术杂志》–。 Aker, A., Derczynski, L., Bontcheva, K., 2017. 用于谣言分析的简单开放式立场分类，载于：自然语言处理最新进展国际会议论文集，RANLP 2017，第31–39页。 Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., Nakov, P., 2021a. 应对社交媒体上的新冠疫情信息流行病：整体视角与行动倡议，载于：国际万维网与社交媒体会议。 Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., Nakov, P., 2021b. 应对新冠疫情信息流行病：对记者、事实核查人员、社交媒体平台、政策制定者和社会视角的建模，载于：计算语言学协会年会论文集：EMNLP 2021，计算语言学协会，多米尼加共和国蓬塔卡纳。第611–649页。网址：https://aclanthology. org/2021.findings-emnlp.56，DOI：10.18653/v1/2021.findings-emnlp.56。 Alhindi, T., Petridis, S., Muresan, S., 2018. 你的证据在哪里：通过论证建模改进事实核查。 Allen, J., Arechar, A.A., Pennycook, G., Rand, D.G., 2020. 运用群体智慧扩大事实核查规模。预印本，网址：https://doi.org/10.31234/osf. io/9qdza。 Alsmadi, I., Alazzam, I., AlRamahi, M.A., 2021. 在线社交网络中错误信息的本体分析。网址：http://arxiv.org/ abs/2102.11362，DOI：10.48550/arxiv.2102.11362，arXiv:2102.11362。 Aly, R., Guo, Z., Schlichtkrull, M.S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., Mittal, A., 2021. FEVEROUS：针对非结构化与结构化信息的事实提取与验证，载于：神经信息处理系统第三十五届会议数据集与基准赛道（第一轮）。 Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T., 2014. 把权力交给大众：人类在交互式机器学习中的作用。《人工智能杂志》35，105–120。 Arnold, P., 2020. 线上事实核查的挑战：技术能（和不能）如何提供帮助——Full Fact。https://fullfact.org/blog/2020/dec/the-challenges-of-online-fact-checking-how-technology-can-and-cant-help/。（访问日期：09/11/2021）。 Atanasova, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P., 2018. CLEF-2018 CheckThat! 实验室概述：自动识别和验证政治主张。任务1：核查价值。arXiv 预印本 arXiv:1808.05542。

[12] http://goodsystems.utexas.edu/

Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Da San Martino, G., 2019a. CLEF-2019 CheckThat! 实验室概述：主张的自动识别与验证。任务1：核查价值。CLEF（工作笔记）2380。 Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., Glass, J., 2019b. 利用语境和话语信息进行自动事实核查。《数据与信息质量期刊》（JDIQ）11，1–27。 Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I., 2020a. 生成事实核查解释，载于：ACL。 Atanasova, P., Wright, D., Augenstein, I., 2020b. 生成标签一致、形式规范的对抗性主张，载于：2020年自然语言处理实证方法会议（EMNLP）论文集，第3168–3177页。 Augenstein, I., Lioma, C., Wang, D., Lima, L.C., Hansen, C., Hansen, C., Simonsen, J., 2019. MultiFC：用于基于证据的主张事实核查的真实多领域数据集，载于：EMNLP。 Bansal, G., Nushi, B., Kamar, E., Horvitz, E., Weld, D.S., 2021a. 准确性最高的人工智能就是最佳队友吗？针对团队协作优化人工智能，载于：AAAI人工智能会议论文集，第11405–11414页。 Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S., Horvitz, E., 2019. 超越准确性：心理模型在人机团队绩效中的作用，载于：AAAI人类计算与众包会议论文集，第2–11页。 Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021b. 整体是否大于部分之和？人工智能解释对互补性团队绩效的影响，载于：2021年CHI计算机系统中的人为因素会议论文集，第1–16页。 Barnabò, G., Siciliano, F., Castillo, C., Leonardi, S., Nakov, P., Da San Martino, G., Silvestri, F., 2022. FBMultiLingMisinfo：用于错误信息检测的高难度大规模多语言基准，载于：2022年国际神经网络联合会议（IJCNN），IEEE。第1–8页。 Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., 等，2020. CheckThat! 2020概述：社交媒体中主张的自动识别与验证，载于：欧洲语言跨语言评测论坛国际会议，施普林格。第215–236页。 Beers, A., Haughey, Melinda, M., Arif, A., Starbird, K., 2020. 调查报道虚假信息的记者所使用的数字工具集，载于：2020年计算新闻学会议论文集 + （计算新闻学+'20）。 ACM，美国纽约州纽约市，，第5页。网址：https://cpb-us-w2.wpmucdn.com/ express.northeastern.edu/dist/d/53/files/2020/02/CJ_2020_论文_50.pdf。 Bendersky, M., Metzler, D., Croft, W.B., 2012. 利用多个信息来源有效构建查询，载于：第五届ACM网络搜索与数据挖掘国际会议论文集，第443–452页。 Bhuiyan, M.M., Zhang, A.X., Sehat, C.M., Mitra, T., 2020. 调查众包新闻可信度评估中的差异：评估者、任务与专家标准。《ACM人机交互汇刊》4，1–26。 Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., Watrin, P., 2022. 注意力是否就是解释？关于这场争论的介绍，载于：计算语言学协会第60届年会论文集（第1卷：长篇论文），计算语言学协会，爱尔兰都柏林。第3889–3900页。网址：https://aclanthology.org/2022.acl-long.269，DOI：10.18653/v1/2022.acl-long.269。 Borel, B., 2016.《芝加哥事实核查指南》。芝加哥大学出版社。 Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., Sadeq, A., 2020. Team Buster。CheckThat! 2020中的人工智能：见解与改进事实核查的建议，载于：CLEF（工作笔记）。 Brand, E., Roitero, K., Soprano, M., Rahimi, A., Demartini, G., 2018. 一种联合预测并解释陈述真实性的神经模型。《 ACM 数据与信息质量期刊》（JDIQ）。 Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M., 2019. "你好，人工智能"：揭示医疗从业者在人机协作决策方面的入门需求。《ACM人机交互汇刊》3，1–24。 Chen, E., Jain, A., 2013. 利用实时人类计算改进 Twitter 搜索。《工程博客》8，2013。 Chen, J., Sriram, A., Choi, E., Durrett, G., 2022. 生成字面与隐含子问题以核查复杂主张。网址：http: // arxiv.org/abs/2205.06938，arXiv:2205.06938。 Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., Roth, D., 2019a. 换个角度看问题：发现有关主张的多元观点，载于：计算语言学协会北美分会2019年年会：人类语言技术论文集，第1卷（长篇与短篇论文），计算语言学协会，美国明尼苏达州明尼阿波利斯。第542–557页。网址：https://aclanthology.org/N19-1053，DOI：10.18653/v1/N19-1053。 Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., Wang, W.Y., 2019b. TabFact：用于基于表格的事实核验的大规模数据集，载于：国际学习表征会议。 Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H., 2019. 通过用户界面解释决策算法：帮助非专业利益相关方的策略，载于：2019年CHI计算机系统中的人为因素会议论文集，第1–12页。 Cimolino, G., Graham, T.N., 2022. 两个脑袋胜过一个：用于统一共享控制中人类与人工智能的维度空间，载于：CHI计算机系统中的人为因素会议，计算机协会，美国纽约州纽约市。网址：https://doi.org/10.1145/3491102.3517610，DOI：10.1145/3491102.3517610。 Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P.K., Zollo, F., 2021a. 网络仇恨与错误信息的动态变化。《科学报告》11，1–12。 Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P.K., Zollo, F., 2021b. 网络仇恨：行为动态及其与错误信息的关系。arXiv 预印本 arXiv:2105.14005。 Clarke, C.L., Rizvi, S., Smucker, M.D., Maistro, M., Zuccon, G., 2020a. TREC 2020健康错误信息赛道概述，载于：TREC。 Clarke, C.L., Rizvi, S., Smucker, M.D., Maistro, M., Zuccon, G., 2020b. TREC 2020健康错误信息赛道概述，载于：TREC。 Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Pietro, R.D., Nakov, P., 2020. 计算宣传检测综述，载于：IJCAI。 Dagan, I., Dolan, B., Magnini, B., Roth, D., 2010. 识别文本蕴含：基本原理、评估与方法——勘误。《自然语言工程》16，105–105。

Das, A., Gupta, C., Kovatchev, V., Lease, M., Li, J.J., 2022年。Prototex：使用原型张量解释模型决策，载于：第60届计算语言学协会年会论文集（第1卷：长篇论文），第2986–2997页。Demartini, G., 2015年。混合式人机信息系统：挑战与机遇。《计算机网络》90，5–13。Demartini, G., Difallah, D.E., Cudré-Mauroux, P., 2012年。Zencrowd：利用概率推理和众包技术进行大规模实体链接，载于：第21届万维网国际会议论文集，第469–478页。Demartini, G., Mizzaro, S., Spina, D., 2020年。用于打击网络错误信息的人在回路中的人工智能：挑战与机遇。《数据工程技术委员会通报》43。Demartini, G., Trushkowsky, B., Kraska, T., Franklin, M.J., Berkeley, U., 2013年。Crowdq：众包查询理解，载于：CIDR。Dhole, K.D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahendiran, A., Mille, S., Srivastava, A., Tan, S., 等，2021年。Nl-augmenter：任务敏感型自然语言增强框架。arXiv预印本 arXiv:2112.02721。Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., Leippold, M., 2020年。Climate-FEVER：用于核验现实世界气候主张的数据集。Dong, E., Du, H., Gardner, L., 2020年。一个用于实时追踪新冠肺炎的交互式网络仪表板。《柳叶刀·传染病》20，533–534。Efron, B., Tibshirani, R., 1985年。用于评估统计准确性的自助法。《Behaviormetrika》12，1–35。Ekstrand, M.D., Das, A., Burke, R., Diaz, F., 等，2022年。信息获取系统中的公平性。《信息检索基础与趋势》16，1–177。Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P., 2019年。CLEF-2019 CheckThat! 实验室概述：主张的自动识别与核验，载于：欧洲语言跨语言评测论坛国际会议，施普林格。第301–321页。Enayet, O., El-Beltagy, S.R., 2017年。NileTMRG在SemEval-2017任务8中的工作：确定Twitter上谣言的真实性及其真实性支持，载于：第11届语义评测国际研讨会论文集（SemEval-2017），计算语言学协会，温哥华，加拿大。第470–474页。网址：https://aclanthology.org/S17-2082，doi:10.18653/v1/S17-2082。Fan, A., Piktus, A., Petroni, F., Wenzek, G., Saeidi, M., Vlachos, A., Bordes, A., Riedel, S., 2020年。生成事实核查简报，载于：2020年自然语言处理实证方法会议论文集（EMNLP），计算语言学协会，线上。第7147–7161页。网址：https://aclanthology.org/2020.emnlp-main.580，doi:10.18653/v1/2020.emnlp-main.580。Farinneya, P., Abdollah Pour, M.M., Hamidian, S., Diab, M., 2021年。用于社交媒体谣言识别的主动学习，载于：计算语言学协会研究成果：EMNLP 2021，计算语言学协会，蓬塔卡纳，多米尼加共和国。第4556–4565页。网址：https://aclanthology.org/2021.findings-emnlp.387，doi:10.18653/v1/2021.findings-emnlp.387。Ferreira, W., Vlachos, A., 2016年。Emergent：用于立场分类的新型数据集，载于：北美计算语言学协会年会。Gad-Elrab, M.H., Stepanova, D., Urbani, J., Weikum, G., 2019年。Exfakt：一种解释知识图谱与文本中事实的框架，载于：第十二届ACM网络搜索与数据挖掘国际会议论文集，第87–95页。Gold, D., Kovatchev, V., Zesch, T., 2019年。对语义关系之间的交互进行标注与分析，载于：第13届语言学标注研讨会论文集，第26–36页。Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., Derczynski, L., 2019年。SemEval-2019任务7：RumourEval，判定谣言真实性及其真实性支持，载于：第13届语义评测国际研讨会论文集，第845–854页。Graves, D., 2018a。理解自动化事实核查的前景与局限。Graves, L., 2017年。事实核查剖析：客观实践与事实核查中备受争议的认识论。《传播、文化与批判》10，518–537。Graves, L., 2018b。边界尚未划定：绘制全球事实核查运动的制度根源。《新闻学研究》19，613–631。Graves, L., Amazeen, M.A., 2019年。作为理念与实践的新闻业事实核查，载于：《牛津传播学研究百科全书》。Gruppi, M., Horne, B.D., Adalı, S., 2021年。NELA-GT-2020：用于研究新闻文章中错误信息的大型多标签新闻数据集。arXiv预印本 arXiv:2102.04567。Guo, H., Cao, J., Zhang, Y., Guo, J., Li, J., 2018年。使用层次化社交注意力网络进行谣言检测，载于：第27届ACM信息与知识管理国际会议论文集，第943–951页。Guo, Z., Schlichtkrull, M., Vlachos, A., 2022年。自动化事实核查综述。《计算语言学协会汇刊》10，178–206。网址：https://doi.org/10.1162/tacl_a_00454，doi:10.1162/tacl_a_00454，arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf。Gupta, A., Srikumar, V., 2021年。X-Fact：用于多语言事实核查的新型基准数据集，载于：ACL/IJCNLP。Gupta, V., Mehta, M., Nokhiz, P., Srikumar, V., 2020年。INFOTABS：作为半结构化数据的表格推理，载于：第58届计算语言学协会年会论文集，计算语言学协会，线上。第2309–2324页。网址：https://aclanthology.org/2020.acl-main.210，doi:10.18653/v1/2020.acl-main.210。Hanselowski, A., AvineshP.V., S., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C.M., Gurevych, I., 2018年。对假新闻挑战赛立场检测任务的回顾性分析，载于：COLING。Hanselowski, A., Stab, C., Schulz, C., Li, Z., Gurevych, I., 2019年。用于自动化事实核查中不同任务的丰富标注语料库，载于：第23届计算自然语言学习会议（CoNLL）论文集，第493–503页。Hardalov, M., Arora, A., Nakov, P., Augenstein, I., 2021年。用于识别错误信息与虚假信息的立场检测综述。ArXiv abs/2103.00242。Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E., 2022年。ToxiGen：用于对抗性与隐性仇恨言论检测的大型机器生成数据集，载于：第60届计算语言学协会年会论文集（第1卷：长篇论文），计算语言学协会，都柏林，爱尔兰。第3309–3326页。网址：https://aclanthology.org/2022.acl-long. 234，doi:10.18653/v1/2022.acl-long.234。Hasanain, M., Elsayed, T., 2020年。BIGIR在CheckThat! 2020中的工作：使用多语言BERT按核查价值对阿拉伯语推文进行排序，载于：CLEF（工作札记）。

Hasanain, M., Elsayed, T., 2021年。研究网络搜索在事实核查中的有效性。《美国信息科学与技术学会会刊》网址：https://onlinelibrary.wiley.com/doi/full/10.1002/asi.24577https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24577https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24577，doi:10.1002/asi.24577。Hasanain, M., Elsayed, T., 2022年。研究网络搜索在事实核查中的有效性。《美国信息科学与技术学会会刊》73，738–751。Hase, P., Bansal, M., 2020年。评估可解释人工智能：哪些算法解释能帮助用户预测模型行为？，载于：ACL。Hassan, N., Arslan, F., Li, C., Tremayne, M., 2017a。迈向自动化事实核查：通过ClaimBuster检测具有核查价值的事实主张，载于：第23届ACM SIGKDD国际知识发现与数据挖掘会议论文集，第1803–1812页。Hassan, N., Li, C., Tremayne, M., 2015年。在总统辩论中检测具有核查价值的事实主张，载于：第24届ACM信息与知识管理国际会议论文集，第1835–1838页。Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., 等，2017b。ClaimBuster：首个端到端事实核查系统。《VLDB基金会会刊》10，1945–1948。Horne, B.D., Khedr, S., Adali, S., 2018年。抽样调查新闻生产者：用于研究复杂媒体格局的大型新闻与专题数据集，载于：第十二届AAAI网络与社交媒体国际会议。Hsu, C.C., Tan, C., 2021年。决策导向摘要，载于：2021年自然语言处理实证方法会议论文集，第117–132页。Jacovi, A., Goldberg, Y., 2020年。迈向忠实且可解释的自然语言处理系统：应如何定义和评估忠实度？，载于：第58届计算语言学协会年会论文集，计算语言学协会，线上。第4198–4205页。网址：https://aclanthology.org/2020.acl-main.386，doi:10.18653/v1/2020.acl-main.386。Jacovi, A., Goldberg, Y., 2021年。让忠实解释与其社会归因相一致。《计算语言学协会汇刊》9，294–310。网址：https://doi.org/10.1162/tacl_a_00367，doi:10.1162/tacl_a_00367，arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00367/1923972/tacl_a_00367.pdf。Jain, K., Garg, A., Jain, S., 2021年。重建扩散模型以检测新闻传播网络中的病毒式传播，载于：《假新闻、政治战与打击错误信息传播研究选集》。IGI Global，第98–111页。Jain, S., Wallace, B.C., 2019年。注意力不等于解释，载于：2019年计算语言学协会北美分会年会：人类语言技术论文集，第1卷（长篇与短篇论文），第3543–3556页。Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., Bansal, M., 2020年。HoVer：用于多跳事实抽取与主张核验的数据集，载于：2020年自然语言处理实证方法会议论文集：研究成果，第3441–3460页。Joachims, T., Radlinski, F., 2007年。能够从隐式反馈中学习的搜索引擎。《计算机》40，34–40。Jones, M.O., 2019年。海湾信息战I 宣传、假新闻与虚假趋势：海湾危机中推特机器人的武器化。《国际传播杂志》13，27。Juneja, P., Mitra, T., 2022年。事实核查的人力与技术基础设施。arXiv预印本 arXiv:2205.10894。Karagiannis, G., Saeed, M., Papotti, P., Trummer, I., 2020年。Scrutinizer：一种用于大规模数据驱动主张核验的混合主动式方法。《VLDB基金会会刊》13，2508–2521。网址：https://doi.org/10.14778/3407790.3407841，doi:10.14778/3407790.3407841。Kaufman, R.A., Haupt, M.R., Dow, S.P., 2022年。人群中的个体很重要：认知因素与信念能够预测错误信息评估准确性。《ACM人机交互会刊》6，1–18。Kazemi, A., Gaffney, D., Garimella, K., Hale, S.A., 2021a。超越英语进行主张匹配，以扩大全球事实核查规模。ACL-IJCNLP 2021——第59届计算语言学协会年会与第11届自然语言处理国际联合会议论文集，第4504–4517页。网址：http://arxiv.org/abs/2106.00853, doi:10.18653/v1/2021.acl-long. 347，arXiv:2106.00853。Kazemi, A., Garimella, K., Shahi, G.K., Gaffney, D., Hale, S.A., 2021b。通过举报热线打击加密平台上的错误信息：以2019年印度大选期间的WhatsApp为例。ArXiv abs/2106.04726。Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D., 2020年。仇恨表情包挑战赛：检测多模态表情包中的仇恨言论，载于：Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H.（编），《神经信息处理系统进展》，Curran Associates, Inc.。第2611–2624页。网址：https://proceedings.neurips.cc/paper/2020/file/ 1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf。Kim, J., Choi, K.S., 2020年。利用知识图谱中正负证据路径的反向加权进行无监督事实核查，载于：第28届计算语言学国际会议论文集，第1677–1686页。Kochkina, E., Liakata, M., Augenstein, I., 2017年。图灵团队在SemEval-2017任务8中的工作：采用分支长短期记忆网络进行谣言立场分类的序列方法，载于：第11届语义评测国际研讨会论文集（SemEval-2017），计算语言学协会，温哥华，加拿大。第475–480页。网址：https://aclanthology.org/S17-2083，doi:10.18653/v1/S17-2083。Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A., 2021年。迈向自动化事实核查：制定标注规范与基准数据集，以实现一致的自动化主张检测。《数字威胁：研究与实践》2，1–16。Kotonya, N., Spooner, T., Magazzeni, D., Toni, F., 2021年。利用上下文感知线性化进行图推理，以实现可解释的事实抽取与核验，载于：第四届事实抽取与验证研讨会（FEVER）论文集，计算语言学协会，多米尼加共和国。第21–30页。网址：https://aclanthology.org/2021.fever-1.3，doi:10.18653/v1/2021.fever-1.3。Kotonya, N., Toni, F., 2020a。可解释的自动化事实核查：综述，载于：COLING。Kotonya, N., Toni, F., 2020b。针对公共卫生主张的可解释自动化事实核查，载于：EMNLP。Kovatchev, V., Chatterjee, T., Govindarajan, V.S., Chen, J., Choi, E., Chronis, G., Das, A., Erk, K., Lease, M., Li, J.J., 等，2022年。longhorns团队在DADC 2022中的工作：要愚弄一个问答模型需要多少语言学家？一种系统性的对抗攻击方法，载于：第一届动态对抗数据收集研讨会论文集，第41–52页。Kovatchev, V., Smith, P., Lee, M., Devine, R., 2021年。向量比专家更会读心吗？比较用于儿童读心能力自动评分的数据增强策略，载于：第59届计算语言学协会年会论文集

以及第11届自然语言处理国际联合会议（第1卷：长篇论文），第1196–1206页。Kovatchev, V., Smith, P., Lee, M., Traynor, I.G., Aguilera, I.L., Devine, R., 2020年。“你在想什么？”：儿童期和青春早期心智解读的自动评分，载于：第28届计算语言学国际会议论文集，第6217–6228页。Kutlu, M., McDonnell, T., Elsayed, T., Lease, M., 2020年。众包标注任务中的标注者理由。《人工智能研究期刊》69，143–189。La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., Spina, D., 2020年。众包判断真实性：判断量表和评估者偏差的影响。《信息检索进展》12036，207。Lai, V., Chen, C., Liao, Q.V., Smith-Renner, A., Tan, C., 2021年。迈向人机决策科学：实证研究综述。arXiv 预印本 arXiv:2112.11471。Lawrence, J., Reed, C., 2020年。论证挖掘：综述。《计算语言学》45，765–818。Lease, M., 2018年。事实核查与信息检索，载于：DESIRES，第97–98页。Lease, M., 2020年。设计人机协作以打击错误信息。LeBeau, C., 2017年。事实不容忽视：图书馆员在事实核查中的角色。《参考与用户服务季刊》57，76–78。Lee, J., Moray, N., 1992年。人机系统中的信任、控制策略与功能分配。《人体工程学》35，1243–1270。Lee, J.D., See, K.A., 2004年。对自动化的信任：为适当依赖而设计。《人为因素》46，50–80。Lee, N., Bang, Y., Madotto, A., Fung, P., 2021年。借助困惑度实现少样本事实核查，载于：2021年计算语言学协会北美分会年会：人类语言技术论文集，计算语言学协会，线上，第1971–1981页。网址：https://aclanthology.org/2021.naacl-main.158，doi：10.18653/v1/2021.naacl-main.158。Lee, N., Li, B.Z., Wang, S., Yih, W.t., Ma, H., Khabsa, M., 2020年。语言模型能否充当事实核查员？，载于：第三届事实抽取与验证研讨会（FEVER）论文集，计算语言学协会，线上，第36–41页。网址：https://aclanthology.org/2020.fever-1.5，doi：10.18653/v1/2020.fever-1.5。Leskovec, J., Backstrom, L., Kleinberg, J., 2009年。模因追踪与新闻周期的动态变化，载于：第15届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集，第497–506页。Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., Cook, J., 2012年。错误信息及其纠正：持续影响与成功去偏。《公共利益中的心理科学》13，106–131。Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 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：第58届计算语言学协会年会论文集，第7871–7880页。Li, T., Fang, L., Lou, J.G., Li, Z., Zhang, D., 2021年。AnaSearch：从非结构化文本中抽取、检索并可视化结构化结果，以支持分析查询。《WSDM 2021——第14届 ACM 网络搜索与数据挖掘国际会议论文集》，906–909。网址：https://doi.org/10.1145/3437963.3441694，doi：10.1145/3437963.3441694。Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J., 2016年 17，1 16。真相发现综述。《ACM SIGKDD探索通讯》–。Lillie, A.E., Middelboe, E.R., Derczynski, L., 2019年。联合预测谣言立场与真实性，载于：第22届北欧计算语言学会议论文集，第208–221页。Liu, A., Swayamdipta, S., Smith, N.A., Choi, Y., 2022年。Wanli：用于自然语言推理数据集创建的人工与人工智能协作。Lu, Y.J., Li, C.T., 2020年。GCAN：用于社交媒体可解释假新闻检测的图感知协同注意力网络，载于：第58届计算语言学协会年会论文集，第505–514页。Ma, J., Gao, W., Joty, S., Wong, K.F., 2019年。利用层次注意力网络进行主张核验的句子级证据嵌入，载于：第57届计算语言学协会年会论文集，计算语言学协会，意大利佛罗伦萨，第2561–2571页。网址：https://aclanthology.org/P19-1244，doi：10.18653/v1/P19-1244。Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M., 2016年。利用循环神经网络检测微博中的谣言，载于：IJCAI。Ma, J., Gao, W., Wong, K.F., 2018年。利用树结构递归神经网络检测 Twitter 上的谣言，载于：第56届计算语言学协会年会（第1卷：长篇论文）论文集，计算语言学协会，澳大利亚墨尔本，第1980–1989页。网址：https://aclanthology.org/P18-1184，doi：10.18653/v1/P18-1184。Marcus, J., 1992年。中部美洲书写系统：四个古代文明中的宣传、神话与历史。普林斯顿大学出版社，普林斯顿。Martinez-Rico, J., Martinez-Romo, J., Araujo, L., 2021年。L.：Nlp&ir@ uned 在 CheckThat! 2021 中的工作：使用 Transformer 模型进行核查价值估计和假新闻检测。Faggioli 等人。[33]。Micallef, N., Armacost, V., Memon, N., Patil, S., 2022年。真或假：专业事实核查人员工作实践研究。《ACM 人机交互论文集》6，1–44。网址：https://doi.org/10.1145/3512974，doi：10.1145/3512974。Mihalcea, R., Strapparava, C., 2009年。测谎器：自动识别欺骗性语言的探索，载于：ACL-IJCNLP 2009 会议短篇论文集，计算语言学协会，新加坡 Suntec，第309–312页。网址：https://aclanthology.org/P09-2078。Mihaylova, T., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Mohtarami, M., Karadzhov, G., Glass, J., 2018年。社区论坛中的事实核查，载于：第三十二届 AAAI 人工智能会议。Miranda, S., Nogueira, D., Mendes, A., Vlachos, A., Secker, A., Garrett, R., Mitchel, J., Marinho, Z., 2019年。新闻编辑室中的自动化事实核查，载于：万维网大会，第3579–3583页。Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., Ragan, E., 2021年。通过机器学习解释，防止在假新闻检测中过度信任，载于：AAAI网络与社交媒体国际会议论文集，第421–431页。Molnar, C., 2020年。可解释机器学习。Lulu.com。Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M., 2019年。利用几何深度学习检测社交媒体上的假新闻。arXiv 预印本 arXiv:1902.06673。

Nakamura, K., Levy, S., Wang, W.Y., 2020年。Fakeddit：用于细粒度假新闻检测的新型多模态基准数据集，载于：第12届语言资源与评测会议论文集，欧洲语言资源协会，法国马赛，第6149–6157页。网址：https://aclanthology.org/2020.lrec-1.755。Nakashole, N., Mitchell, T., 2014年。面向语言的事实候选项真实性评估，载于：第52届计算语言学协会年会（第1卷：长篇论文）论文集，第1009–1019页。Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., Da San Martino, G., 2021a。用于辅助人工事实核查员的自动化事实核查，载于：IJCAI。Nakov, P., Da San Martino, G., Barrón-Cedeño, A., Zaghouani, W., Míguez, R., Alam, F., Caselli, T., Kutlu, M., Strub, J.M., Mandl, T., 2022年。CheckThat! 实验室：打击新冠疫情信息流行病和假新闻检测（CLEF-2022 实验室提案）。Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., 等，2021b。CLEF-2021 CheckThat! 实验室：检测值得核查的主张、已核查事实主张和假新闻，载于：ECIR（第2卷）。Neely-Sardon, A., Tignor, M., 2018 59，108 121。聚焦事实：新闻与信息素养教学项目。《参考馆员》–。Neumann, T., De-Arteaga, M., Fazelpour, S., 2022年。错误信息检测系统中的公平正义：对算法、利益相关方和潜在危害的分析，载于：2022年 ACM 公平性、问责制与透明度会议，计算机协会，美国纽约州纽约市，第1504–1515页。网址：https://doi.org/10.1145/3531146.3533205，doi：10.1145/3531146.3533205。Nguyen, A.T., Kharosekar, A., Krishnan, S., Krishnan, S., Tate, E., Wallace, B.C., Lease, M., 2018a。信不信由你：设计用于混合主动式事实核查的人机协作，载于：第31届 ACM 用户界面软件与技术年会论文集，第189–199页。Nguyen, A.T., Kharosekar, A., Lease, M., Wallace, B.C., 2018b。用于众包事实核查的可解释联合图模型，载于：第三十二届 AAAI 人工智能会议（AAAI-18）论文集，第1511–1518页。网址：../papers/nguyen-aaai18.pdf。Nguyen, T.T., Weidlich, M., Yin, H., Zheng, B., Nguyen, Q.H., Nguyen, Q.V.H., 2020年。Factcatch：以最少用户投入实现增量式按需事实核查，载于：第43届 ACM SIGIR 国际信息检索研究与发展会议论文集，第2165–2168页。Nie, Y., Wang, S., Bansal, M., 2019年。揭示语义检索对于大规模机器阅读的重要性，载于：2019年自然语言处理实证方法会议与第9届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，计算语言学协会，中国香港，第2553–2566页。网址：https://aclanthology.org/ D19-1258，doi：10.18653/v1/D19-1258。Niewiński, P., Pszona, M., Janicka, M., 2019年。Gem：用于对抗攻击的生成式增强模型，载于：第二届事实抽取与核验研讨会（FEVER）论文集，第20–26页。Nørregaard, J., Horne, B.D., Adalı, S., 2019年。Nela-gt-2018：用于研究新闻文章中错误信息的大型多标签新闻数据集，载于：AAAI网络与社交媒体国际会议论文集，第630–638页。Oshikawa, R., Qian, J., Wang, W.Y., 2020年。假新闻检测自然语言处理综述，载于：LREC。Popat, K., Mukherjee, S., Yates, A., Weikum, G., 2018年。Declare：利用证据感知深度学习揭穿假新闻和虚假主张，载于：2018年自然语言处理实证方法会议论文集，第22–32页。Potthast, M., KIESELJ, R., 等，2018年。对超党派新闻和假新闻的文体计量研究。第56届计算语言学协会年会论文集。宾夕法尼亚州斯特劳兹堡：ACL 231240。Potthast, M., Köpsel, S., Stein, B., Hagen, M., 2016年。标题党检测，载于：欧洲信息检索会议，施普林格，第810–817页。Pradeep, R., Ma, X., Nogueira, R., Lin, J., 2021年。利用 VerT5erini 核验科学主张，载于：第12届健康文本挖掘与信息分析国际研讨会论文集，计算语言学协会，线上，第94–103页。网址：https: //aclanthology.org/2021.louhi-1.11。Qazvinian, V., Rosengren, E., Radev, D., Mei, Q., 2011年。谣言有说法：识别微博中的错误信息，载于：2011年自然语言处理实证方法会议论文集，第1589–1599页。Qu, Y., Barbera, D.L., Roitero, K., Mizzaro, S., Spina, D., Demartini, G., 2021a。结合人类与机器置信度评估真实性。《ACM数据与信息质量期刊》（JDIQ）。Qu, Y., Roitero, K., Mizzaro, S., Spina, D., Demartini, G., 2021b。用于评估真实性的人在回路系统：人类与机器置信度研究。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 等，2019年。语言模型是无监督的多任务学习者。《OpenAI 博客》1，9。Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y., 2017年。真相的不同面貌：分析假新闻与政治事实核查中的语言，载于：2017年自然语言处理实证方法会议论文集，第2931–2937页。Ribeiro, M.T., Singh, S., Guestrin, C., 2018年。用于调试自然语言处理模型的语义等价对抗规则，载于：第56届计算语言学协会年会（第1卷：长篇论文）论文集，第856–865页。Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S., 2020年。超越准确性：使用检查清单对自然语言处理模型进行行为测试，载于：第58届计算语言学协会年会论文集，第4902–4912页。Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S., Demartini, G., 2020a。众包群体能否客观识别错误信息？判断量表和评估者背景的影响，载于：第43届 ACM SIGIR 国际信息检索研究与发展会议论文集，第439–448页。Roitero, K., Soprano, M., Portelli, B., Spina, D., Della Mea, V., Serra, G., Mizzaro, S., Demartini, G., 2020b。新冠疫情信息流行病：众包群体能否客观判断近期错误信息？，载于：第29届 ACM 信息与知识管理国际会议论文集，第1305–1314页。

Sarasua, C., Simperl, E., Noy, N.F., 2012年。Crowdmap：通过微任务众包实现本体对齐，载于：国际语义网会议，施普林格。第525–541页。Schuster, T., Fisch, A., Barzilay, R., 2021年。获取维生素C！利用对比证据进行稳健的事实验证，载于：计算语言学协会北美分会2021年会议论文集：人类语言技术，计算语言学协会，线上。第624–643页。网址：https://aclanthology.org/2021.naacl-main.52，DOI：10.18653/v1/2021.naacl-main.52。Schuster, T., Schuster, R., Shah, D.J., Barzilay, R., 2020年。文体计量学在检测机器生成的假新闻方面的局限性。《计算语言学》46，499–510。Serrano, S., Smith, N.A., 2019年。注意力是否可解释？，载于：计算语言学协会第57届年会论文集，第2931–2951页。Settles, B., 2009年。主动学习文献综述。Shaar, S., Alam, F., Da San Martino, G., Nakov, P., 2021a。协助人工事实核查员：检测文档中所有已核查事实主张。arXiv预印本 arXiv:2109.07410。Shaar, S., Hasanain, M., Hamdan, B., Ali, Z.S., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y.S., Alam, F., Da San Martino, G., Barrón-Cedeño, A., Míguez, R., Beltrán, J., Elsayed, T., Nakov, P., 2021b。CLEF-2021 CheckThat!实验室任务1概述：推文和政治辩论中的核查价值估计，载于：CLEF。Shaar, S., Martino, G.D.S., Babulkov, N., Nakov, P., 2020年。这是一个已知谎言：检测此前已核查的主张，载于：ACL。Shabani, S., Charlesworth, Z., Sokhn, M., Schuldt, H., 2021年。SAMS：采用人在回路的方法打击数字虚假信息的分享，载于：CEUR研讨会论文集。Shao, C., Ciampaglia, G.L., Flammini, A., Menczer, F., 2016年。Hoaxy：追踪网络错误信息的平台，载于：第25届万维网国际会议伴随会议论文集，第745–750页。Shi, B., Weninger, T., 2016年。面向知识图谱事实核查的判别式谓词路径挖掘。《基于知识的系统》104，123–133。Shi, L., Bhattacharya, N., Das, A., Lease, M., Gwizdka, J., 2022年。交互式人工智能设计对用户行为的影响：一项关于新冠肺炎主张事实核查的眼动追踪研究，载于：第7届ACM SIGIR人类信息交互与检索会议（CHIIR）论文集。网址：https://utexas.box.com/v/shi-chiir2022。Shu, K., Cui, L., Wang, S., Lee, D., Liu, H., 2019年。defend：可解释的假新闻检测，载于：第25届ACM SIGKDD知识发现与数据挖掘国际会议论文集，第395–405页。Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H., 2020年。Fakenewsnet：一个包含新闻内容、社会语境及时空信息的数据存储库，用于研究社交媒体上的假新闻。《大数据》8卷3期，171–188。Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017年。社交媒体上的假新闻检测：数据挖掘视角。《ACM SIGKDD探索通讯》19，22–36。Singh, P., Das, A., Li, J.J., Lease, M., 2021年。自动事实核查中主张难度评估的必要性。arXiv预印本 arXiv:2109.09689。Smeros, P., Castillo, C., Aberer, K., 2021 年。Sciclops：检测并语境化科学主张，以辅助人工事实核查，载于：第30届ACM信息与知识管理国际会议（CIKM）论文集。Sokol, K., Flach, P., 2019年。可解释性的要求：用反事实解释决策树预测，载于：AAAI人工智能会议论文集，第10035–10036页。Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Mizzaro, S., Demartini, G., 2021年。真实性的诸多维度：基于多维量表众包虚假信息评估。《信息处理与管理》58，102710。波因特研究所，2021年。Global Fact 8预录片段4。网址：https://www.youtube.com/watch?v=gOhPKDaeQxI&t=770s。Thorne, J., Vlachos, A., 2019年。针对事实提取与验证的对抗攻击。arXiv预印本 arXiv:1903.05543。Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A., 2018年。FEVER：用于事实提取与验证的大规模数据集，载于：计算语言学协会北美分会2018年会议论文集：人类语言技术，第1卷（长篇论文），第809–819页。Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A., 2019年。FEVER 2.0共享任务，载于：第二届事实抽取与核验研讨会（FEVER）论文集，第1–6页。Thornhill, C., Meeus, Q., Peperkamp, J., Berendt, B., 2019年。用于抵消确认偏误的数字助推。《大数据前沿》2，11。Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., Krause, A., 2018年。通过群体信号检测社交网络中的假新闻，载于：2018年万维网大会伴随论文集，第517–524页。Uscinski, J.E., 2015年。事实核查的认识论（仍然天真）：对Amazeen的回应。《批判评论》27，243–252。Vaish, R., Davis, J., Bernstein, M., 2015年。研究过程众包。《集体智能》3。Vaish, R., Gaikwad, S.N.S., Kovacs, G., Veit, A., Krishna, R., Arrieta Ibarra, I., Simoiu, C., Wilber, M., Belongie, S., Goel, S., 等，2017年。群体研究：开放且可扩展的大学实验室，载于：第30届ACM用户界面软件与技术年度研讨会论文集，第829–843页。Vaughan, J.W., Wallach, H., 2020年。以人为中心的可理解机器学习议程。《我们信任的机器：与人工智能相处》。Veale, M., Van Kleek, M., Binns, R., 2018年。高风险公共部门决策中算法支持的公平性与问责设计需求，载于：2018年CHI计算机系统中的人为因素会议论文集，第1–14页。Vicario, M.D., Quattrociocchi, W., Scala, A., Zollo, F., 2019年。极化与假新闻：对潜在虚假信息目标的早期预警。《ACM网络汇刊》（TWEB）13，1–22。Vlachos, A., Riedel, S., 2014年。事实核查：任务定义与数据集构建，载于：ACL 2014语言技术与计算社会科学研讨会论文集，第18–22页。Vlachos, A., Riedel, S., 2015年。统计属性相关简单主张的识别与验证，载于：计算语言学协会2015年自然语言处理实证方法会议论文集，计算语言学协会，葡萄牙里斯本。第2596–2601页。网址：

https://aclanthology.org/D15-1312，DOI：10.18653/v1/D15-1312。Vo, N., Lee, K., 2018年。守护者的崛起：通过事实核查网址推荐打击假新闻，载于：第41届 ACM SIGIR 信息检索研究与发展国际会议，第275–284页。Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H., 2020年。事实还是虚构：验证科学主张，载于：2020年自然语言处理实证方法会议（EMNLP）论文集，第7534–7550页。Wang, W.Y., 2017年。“说谎，说谎，裤子着火了”：用于假新闻检测的新基准数据集，载于：ACL。Wiegreffe, S., Marasovic, A., 2021年。教我解释：可解释自然语言处理数据集综述，载于：Vanschoren, J., Yeung, S.（编），神经信息处理系统数据集与基准专题论文集。网址：
h
t
t
p
s://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/698d51a19d8a121ce581499d7b701668-Paper-round1.pdf。Wiegreffe, S., Pinter, Y., 2019年。注意力并非解释，载于：2019年自然语言处理实证方法会议与第9届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，第11–20页。Williams, E., Rodrigues, P., Novak, V., 2020年。埃森哲参加CheckThat! 2020：如你所言：使用基于 Transformer 的模型进行主张的事后事实核查。arXiv预印本 arXiv:2009.02431。Yang, F., Pentyala, S.K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E.D., Ji, S., Hu, X., 2019年。Xfake：带可视化功能的可解释假新闻检测器，载于：万维网大会论文集，第3600–3604页。Zaidan, O., Eisner, J., Piatko, C., 2007年。利用“标注者理由”改进文本分类的机器学习，载于：2007年人类语言技术：计算语言学协会北美分会会议；主会议论文集，第260–267页。Zanzotto, F.M., 2019年。人在回路中的人工智能。《人工智能研究期刊》64，243–252。Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y., 2020年。防御神经网络生成的假新闻。NeurIPS。Zeng, X., Abumansour, A.S., Zubiaga, A., 2021年。自动化事实核查：综述。《语言与语言学指南》15，e12438。Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., Shu, K., 2021a。挖掘双重情绪以检测假新闻，载于：2021年万维网大会论文集，第3465–3476页。Zhang, Y., Lease, M., Wallace, B., 2017年。主动判别式文本表示学习，载于：AAAI人工智能会议论文集。Zhang, Z., Rudra, K., Anand, A., 2021b。Faxplainac：基于可解释模型并将人工纠正纳入回路的事实核查工具，载于：第30届ACM信息与知识管理国际会议论文集，第4823–4827页。Zhou, X., Jain, A., Phoha, V.V., Zafarani, R., 2020 1, 1 25。假新闻早期检测：理论驱动模型。《数字威胁：研究与实践》–。Zhou, X., Zafarani, R., 2020年。假新闻综述：基础理论、检测方法与机遇。《ACM计算调查》（CSUR）53，1–40。Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R., 2018年。社交媒体中谣言的检测与处理：综述。《ACM计算调查》（CSUR）51，1–36。Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P., 2016年。通过对话线程分析人们如何理解和传播社交媒体上的谣言。《PLOS ONE》11，e0150989。