模板：https://www.overleaf.com/project/67b7dcdfb9ae12009060dbdc

建议

周一 11:59 前**到期**

第 5 **点**

本课程的一个主要组成部分是一个关于自然语言处理主题的期末项目（涉及 1 到 3 名学生），重点是使用 NLP 来支持一个实证研究问题。

建议书（2 月 24 日（周一）11:59 前提交）应

- 概述你要开展的工作
- 将其理由作为一个值得提出的有趣问题加以激励
- 通过将其置于科学界相关文献中，评估其贡献新知识潜力。(引用 5 个相关资料来源）
- 明确团队中的成员、各自的职责以及完成这些任务的内部期限。您应该提供足够详细和具体的信息，以便让您的团队成员负责履行他们的职责。

上交本作业时，请为每组上传一个 pdf 文件。所有报告都应使用 ACL 样式文件，该文件可作为 Overleaf 模板使用。

您的**建议书**长度应为 **1000 字 [+/- 200 字]**，**不包括结尾的参考文献列表。**当您阅读其他研究论文时，请务必避免不经意地抄袭任何内容--请查看艾米丽-迈尔斯（Emily Myers）制作的有关抄袭的有用信息图表，以进行复习。您必须是所提交文本的作者，而不是人工智能助理。

中期报告

应于 3 月 31 日晚上 11:59 之前**提交**

10 **分**

中期报告（3 月 31 日（周一）晚上 11:59 交）应详细说明您在期末项目上的进展情况。在这个阶段，每个项目都应该有个人反馈，并对中期报告应包括的内容期望。一般来说，在这一阶段，内容包括

-- 你对相关工作的文献综述应该完整；引用至少 10 篇与你的工作直接相关的资料，并说明你所提出的方法如何与已有工作相适应。根据粗略的经验，您可能需要阅读/浏览 25-30 篇论文，才能找到这 10 篇重要的资料。(本阶段评估的一个主要部分是审查的彻底性）。

-- 最好能收集到所有数据。

-- 进行一些初步实验，以检验你提出的方法是否可行或是否需要调整。

-- 详细说明评估方法绩效的评估策略。

-- 与提案一样，列出您将的具体任务、团队中的哪些人将执行这些任务，以及他们任务的最后期限。请提供足够具体的信息，以便您能让他们承担责任。

上交本作业时，请为每组上传一个 pdf 文件。所有报告都应使用 ACL 样式文件，该文件可作为附页模板提供

.中期报告应为 **2000 字 [+/- 200 字]**，**不包括结尾的参考文献列表。**当您阅读其他研究论文时，请务必避免不经意地抄袭任何内容--请查看 Emily Myers 制作的有关抄袭的有用信息图表，以进行复习。您必须是所提交文本的作者，而不是人工智能助手。

介绍

> 4 月 28 日 11:59 前**到期**
> 第 5 **点**

最终项目展示将于 4 月 29 日（星期二）和 5 月 1 日（星期四）**在课堂上**进行。在这节课上，请准备一张介绍项目的幻灯片，并准备好回答听众（包括指导教师）的问题。此时您的作品应相对完整。(在提交最终项目报告之前，您将有时间吸收收到的任何反馈意见，但您的实验和分析在此阶段应该已经完成）。

一般准则：

- 此时，您的工作应该完成了大约 85%（足以终点，但不一定完成）。
- 您的幻灯片应概述有关项目目标、方法或发现的基本信息。
- 除演讲本身外，本次作业的交付成果还包括一张 **PDF 格式的幻灯片**；请于 4 月 28 日（周一）晚上 11:59 之前在 bCourses 上提交该幻灯片（无论您是在周二还是周四演讲）。

最终报告

> **应于** 5 月 12 日晚上 11:59 前**提交**
> 25 **分**

最终项目报告，5 月 12 日星期一（11:59pm）到期

您的最终项目应为 **3000 字[+/- 200 字**，**不包括最后的参考文献列表**，采用 ACL 格式，单倍行距。除论文外，**您还需提交为本项目编写的代码，以便我们能够重现结果**。

您将根据以下标准接受评估：

- 可重复性。提交的代码能在多大程度上复制论文中的结果?

- 清晰度。对于准备充分的读者来说，是否清楚做了什么以及为什么要这样做？论文是否写得好，结构是否合理？
- 原创性。提出的方法或问题的原创性如何？论文在主题、方法或内容上是否有新的突破？论文描述的研究有多令人兴奋和创新？
- 合理性。技术方法是否合理、选择是否恰当？，论文的主张是否可信--是否有适当的实验、证明或其他论证？
- 实质内容。这篇论文是否有足够的实质内容，或者是否可以从更多的观点或结果中获益？作者是否指出了其工作的潜在局限性？
- 评估。在多大程度上对应用程序或工具进行了测试和评估？
- 有意义的比较。作者是否明确说明了所提出的系统与现有文献的关系？参考文献是否充分？系统/应用的优点是否得到充分支持？
- 影响。所述工作的意义有多大？该系统的新颖之处是否会其他研究人员在自己的工作中采用该方法？

现在，您已经读过几篇以这种格式撰写的论文了；更多例子，请参阅：

https://aclanthology.org/events/acl-2024/。

您会注意到，许多论文都遵循标准格式，包括 "引言"、"相关工作"、"数据"、"方法"、"分析 "和 "结论"。

要上交此作业，请上传：

- 每组一份包含项目报告的 pdf 文件。
- **压缩文件（.zip 或 .tar.gz），其中包含**重现作品所需的**代码和数据**。**请勿**提交任何私人信息（包括您可能使用过的 LLM 服务的任何 API 密钥--请确保这些密钥未被硬编码到您的脚本中）。

所有报告都应使用 ACL 样式文件，该文件作为附页模板提供

.在您撰写其他研究论文时，请务必避免不经意地抄袭任何内容--请查看 Emily Myers 制作的有关抄袭的有用信息图表，以进行复习。您必须是所提交文本的作者，而不是人工智能助手。

主题	**数据来源**	**功能**	**参考文献**	
虚假信息检测	-新闻 -假新（作为图案或xx）	给定信息（可以是社交媒体帖子），检查其是否为错误信息	聚焦	Prem, Yunka Mimpi
剑龙关键词搜索器	职位描述	给定职位描述文本，返回关键字	供应商：简化、工作扫描	Prem Yunkai
逻辑推理 - 烹饪书 食谱生成				Mimp Prem
财经新闻摘要/内容	社交媒体，新闻	- 给定股票名称，我们该不该投资 - 股价将上涨/下跌	https://www.researchgate.net/profile/Rodrigue-Andrawos/publication/361164679_NLP_in_Stock_Market_Pred ediction_A_Review/links/62a0885255273755ebdc1bd0/NLP-in-Stock-Market-Prediction-A-Review.pdf	Prem Mimp
可持续发展报告		1. 查看评级良好的公司在其报告中是怎么写的 2. 如果他们真的这样做了		Mimp

项目概述

标题社区笔记在信息核查中的作用和效果

摘要："社区注释"（Community Notes）是一个众包的事实核查系统，旨在为在线讨论提供额外的背景信息和验证。该系统目前在 Twitter 上使用，允许用户或撰稿人在推文中添加说明性注释，同时允许社区就注释是否有用进行投票，从而在打击错误信息方面发挥了重要作用。然而，建立一个开放的、参与式的系统也面临着挑战，包括抵制操纵、确保观点的多样性以及保持高质量的贡献。本文探讨了社区笔记的三个关键方面：**分析社区笔记数据、所提供信息的有效性和准确性**，以及**其在增强 LLM 事实性方面的潜在作用**。通过分析错误信息趋势的数据、交叉引用注释的准确性以及探索人工智能的整合，本研究强调了社区注释的优势和局限性。随着 Facebook 等大型平台计划在 2025 年实施类似系统，了解社区笔记的影响和可扩展性变得越来越重要。本文旨在为优化社区驱动的事实核对提供见解，以促进公共讨论和人工智能应用。*<需要添加我们将在本项目中使用的数据以及本项目的范围> <需要添加我们将在本项目中使用的数据以及本项目的范围> <需要添加我们将在本项目中使用的数据以及本项目的范围> <需要添加我们将在本项目中使用的数据以及本项目的范围>*。

- **一般主题参考**
  - 社区笔记能否取代专业事实核查员？

    2025 年 2 月 19 日 https://arxiv.org/html/2502.14132v1
  - 社区笔记增加了人们对社交媒体事实核查的信任 2024 年 5 月 31 日 https://pmc.ncbi.nlm.nih.gov/articles/PMC11212665/
  - 2023 年 7 月

    https://publications.iadb.org/en/reducing-misinformation-role-confirmation-frames- 事实核查-干预
- **方面 1 分析社区说明数据**
  - 社区笔记与窥探：群众如何在社交媒体上选择事实核查目标 *https://ojs.aaai.org/index.php/ICWSM/article/view/31387/33547*
- **第 2 方面 所提供信息的有效性和准确性**
  - *政治事实范围，政治相关*
  - 谁来检查跳棋？探索 Twitter 社区笔记中的来源可信度 https://arxiv.org/html/2406.12444v1
  - 报告称 X 上的众包事实核查无法解决美国大选错误信息泛滥的问题 https://apnews.com/article/x-musk-twitter-misinformation-ccdh-0fa4fec0f703369b93be248461e8005d
  -
  -

- *方面 3* **在提高法律硕士实事求是精神方面的潜在作用**

- 解读 DPO 和 PPO：区分从偏好反馈中学习的最佳做法
  https://arxiv.org/abs/2406.09279
    - 报告提到，在进行偏好反馈评估时，事实准确性得分并没有提高很多。
- 利用大规模知识图谱评估大型语言模型的真实性
  https://arxiv.org/pdf/2404.00942
    - 某些模型的事实准确度得分有所提高，特别是在 LLaMA 系列中，但也有一些取舍
- https://medium.com/@techsachin/openfactcheck-a-unified-framework-for-factualit y-evaluation-of-llms-d88f2946ca94
-

背景与动机

事实核查对于验证信息的准确性至关重要，尤其是在政治、公共卫生和气候变化等问题上。在 Twitter/X 和 Facebook 等社交媒体平台上，由于匿名发布、人们更有可能转发虚假或新奇信息[1]、算法优先考虑参与度而非准确性以及缺乏验证机制等因素，错误信息会迅速传播。传统的事实核查依赖于人类专家，他们
与可靠信息来源相互参照，以评估其有效性。PolitiFact、Snopes 和 FactCheck.org 等组织在减少错误信息方面发挥了重要作用。然而，人工事实核查耗时长、规模有限，而且可能受到个人偏见的影响。为了解决这些局限性，许多工具（如 GPTzero 或 Originality.AI）都采用了人工智能技术。
人工智能驱动的事实核查已经出现，它利用机器学习和自然语言处理（NLP）来更有效地分析和验证信息。
然而，基于人工智能的工具也有其挑战，包括训练数据的偏差、易受敌对错误信息的影响、难以理解细微的语境，以及如果根据不可靠的信息来源进行训练，则有可能强化错误的叙述。

社区笔记的崛起

社区笔记最初是 Twitter 于 2021 年 1 月推出的 "鸟类观察"（Birdwatch），它是一个基

于社区的事实核查系统，可帮助用户添加上下文。

对他们认为包含虚假或误导信息的帖子进行评论。这些说明不会立即发布，而是要经过一个审核过程，由贡献者投票决定其是否有用。只有当投票来自不同的角度时，说明才会公开，从而提供更多的背景信息，帮助用户评估原始内容。研究表明，这种方法可以达到很高的准确率--对 COVID 相关笔记的研究显示，准确率高达 97%，而且可以显著减少错误信息的传播[2]，但研究也表明，审核过程可能比较缓慢，有时会将笔记的发布推迟到错误信息已经传播之后。2024 年，扎克伯格指出："我们已经到了错误太多、审查太多的地步。"到 2025 年，Facebook 和 Instagram 等平台将从第三方事实核查过渡到使用社区笔记。这一转变非常重要，因为它标志着网络内容的审核将变得更加透明，社区将更多地参与其中，这将带来一个更负责任、反应更迅速的信息环境。

因此，鉴于 "社区笔记 "在打击错误信息和提高透明度方面发挥着至关重要的作用，我们的工作重点是利用其潜力来加强事实核查流程。我们研究了三个关键方面：分析 "社区笔记 "数据以识别错误信息趋势、评估所提供上下文的有效性和准确性，以及探索如何将这些见解与人工智能相结合以提高大型语言模型的真实性。随着 Facebook 等大型平台也将采用类似的系统，了解社区笔记的性能和扩展性对我们加强基于社区的事实核查至关重要，这将确保我们的公共讨论保持知情，并确保人工智能工具能够使用准确的数据。

[1] 研究：在 Twitter 上，虚假新闻比真实新闻传播得更快，
https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-030 8

[2] 社区笔记有用吗?

https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/14/do-community-notes-w ork/

[3] Meta 正在淘汰事实核查员。扎克伯格承认，现在平台上会出现更多有害内容

https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html

相关工作

方法

方法:

1) 收集和处理数据
    a) 社区说明
    b) PolitiFact （2024 年数据）关注选举
2) 通过分析数据了解社区说明
    a) 主题分析：使用类似 [xx] 的主题建模
    b) xxx
3) 社区笔记的准确性
    a) 如果我们使用政治事实
        i) 使用 50% 的虚假信息，50%
    b) 法学硕士
4)

数据来源

- 推特社区说明
- 现有数据库，如 FakeNewsNet
  https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UEM MHS, https://github.com/KaiDMML/FakeNewsNet?tab=readme-ov-file , Politifact

技术与工具

- 指标 1 分析

评估指标：

- Aspect1 - 不需要?
- 指标角度 2 - 已核对数据的准确性

Dataset	Input	#Inputs	Evidence	Verdict	Sources	Lang
CrimeVeri (Bachenko et al., 2008)	Statement	275	✗	2 Classes	Crime	En
Politifact (Vlachos and Riedel, 2014)	Statement	106	Text/Meta	5 Classes	Fact Check	En
StatsProperties (Vlachos and Riedel, 2015)	Statement	7,092	KG	Numeric	Internet	En
Emergent (Ferreira and Vlachos, 2016)	Statement	300	Text	3 Classes	Emergent	En
CreditAssess (Popat et al., 2016)	Statement	5,013	Text	2 Classes	Fact Check/Wiki	En
PunditFact (Rashkin et al., 2017)	Statement	4,361	✗	2/6 Classes	Fact Check	En
Liar (Wang, 2017)	Statement	12,836	Meta	6 Classes	Fact Check	En
Verify (Baly et al., 2018)	Statement	422	Text	2 Classes	Fact Check	Ar/En
CheckThat18-T2 (Barrón-Cedeño et al., 2018)	Statement	150	✗	3 Classes	Transcript	En
Snopes (Hanselowski et al., 2019)	Statement	6,422	Text	3 Classes	Fact Check	En
MultiFC (Augenstein et al., 2019)	Statement	36,534	Text/Meta	2–27 Classes	Fact Check	En
Climate-FEVER (Diggelmann et al., 2020)	Statement	1,535	Text	4 Classes	Climate	En
SciFact (Wadden et al., 2020)	Statement	1,409	Text	3 Classes	Science	En
PUBHEALTH (Kotonya and Toni, 2020b)	Statement	11,832	Text	4 Classes	Fact Check	En
COVID-Fact (Saakyan et al., 2021)	Statement	4,086	Text	2 Classes	Forum	En
X-Fact (Gupta and Srikumar, 2021)	Statement	31,189	Text	7 Classes	Fact Check	Many
cQA (Mihaylova et al., 2018)	Answer	422	Meta	2 Classes	Forum	En
AnswerFact (Zhang et al., 2020)	Answer	60,864	Text	5 Classes	Amazon	En
NELA (Horne et al., 2018)	Article	136,000	✗	2 Classes	News	En
BuzzfeedNews (Potthast et al., 2018)	Article	1,627	Meta	4 Classes	Facebook	En
BuzzFace (Santia and Williams, 2018)	Article	2,263	Meta	4 Classes	Facebook	En
FA-KES (Salem et al., 2019)	Article	804	✗	2 Classes	VDC	En
FakeNewsNet (Shu et al., 2020)	Article	23,196	Meta	2 Classes	Fact Check	En
FakeCovid (Shahi and Nandini, 2020)	Article	5,182	✗	2 Classes	Fact Check	Many

Table 2: Summary of factual verification datasets with natural inputs. KG denotes knowledge graphs. ChectThat18 has been extended later (Hasanain et al., 2019; Barrón-Cedeño et al., 2020; Nakov et al., 2021b). NELA has been updated by adding more data from more diverse sources (Nørregaard et al., 2019; Gruppi et al., 2020, 2021).

- 指标 3 - 基准精度

团队与职责 Bhirajaya

Bhumimars Supakarn

Bunlert

李云恺

时间表

参考资料

社区说明的作用

元/ 文章

Meta 将放弃事实核查员，转而采用 X 风格的社区说明。它们能起作用吗？2025 年 1 月 25 日
https://www.bbc.com/news/articles/c4g93nvrdz7o

事实核查与社区笔记：媒体素养的无意之举 2025 年 1 月 23 日
https://www.linkedin.com/pulse/fact-checking-vs-community-notes-unintentional-case-media-dej ong-rtdgf/

社区笔记及其对虚假信息的狭隘理解 2025 年 2 月 3 日
https://www.techpolicy.press/community-notes-and-its-narrow-understanding-of-disinformation/

研究显示，事实核查者是 X's Community Notes 的主要来源之一 2025 年 2 月 20 日
https://www.poynter.org/ifcn/2025/fact-checkers-contribute-improve-community-notes-x/

论文

社区笔记能否取代专业事实核查员？2025 年 2 月 19 日
https://arxiv.org/html/2502.14132v1

社区笔记增加了人们对社交媒体事实核查的信任 2024 年 5 月 31 日
https://pmc.ncbi.nlm.nih.gov/articles/PMC11212665/

谁来检查跳棋？探索 Twitter 社区笔记中的来源可信度 https://arxiv.org/html/2406.12444v1