arXiv:2301.05863v1 [cs.CL] 14 Jan 2023

阿拉伯语推文中权威机构对谣言的立场检测：一项初步研究

Fatima Haouari 和 Tamer Elsayed

卡塔尔大学计算机科学与工程系
{200159617,telsayed}@qu.edu.qa

摘要。已有大量研究探讨了 Twitter 上的谣言核验问题，这些研究或利用传播网络中的证据，或利用来自网络的外部证据。然而，这些研究均未利用来自可信权威人士的证据。本文定义了推文中权威机构对谣言的立场检测任务，即判断来自权威机构的推文是同意、不同意谣言，还是与谣言无关。我们认为，这项任务有助于扩充现有谣言核验系统所利用的证据来源。我们构建并发布了首个权威人士对谣言的立场（AuSTR）数据集，其中的证据取自阿拉伯语 Twitter 上权威人士的时间线。鉴于该数据集规模相对有限，我们研究了现有数据集在这一立场检测任务中的实用性。我们表明，现有数据集对该任务有一定帮助，但显然仍不充分，因此有必要通过添加经标注的数据来扩充这些数据集，其中包含来自 Twitter 的权威机构立场信息。

关键词：证据 · 主张 · 社交媒体

1 引言

现有社交媒体谣言核验研究利用传播网络作为证据来源，重点考察回复的立场 [33,24,13,34,8,29]，回复结构 [27,12,14,32,19,9]，以及转推者的个人资料特征 [26]。近期，Dougrez-Lewis 等人 [17] 提出将网络证据纳入传播网络。据我们所知，此前尚无研究探讨如何利用可信权威人士的时间线中的证据来核验社交媒体上的谣言；权威机构是指具备核实或否认特定谣言所需真实知识或权力的实体 [11]。我们认为，检测相关权威机构对谣言的立场，能够为现有谣言核验系统所采用的证据来源提供重要补充。它还可以成为事实核查员的宝贵工具，帮助自动化检查权威机构推文、核实特定谣言的流程。值得一提的是，权威机构的立场只是能够补充其他来源的一种证据来源（但很重要），而且在某些情况下，单凭这一来源可能不足以完全判断谣言的真实性。

本文对阿拉伯世界 Twitter 上传播的谣言，开展了检测权威机构立场的初步研究。利用 Twitter 上的证据来源进行阿拉伯语谣言核查仍未得到充分研究；现有研究仅关注推文文本来进行核查 [18,20,28,2,31,5]。一个值得注意的例外是 Haouari 等人 [19] 的研究，他们利用回复、回复结构以及回复者的个人资料特征来核验阿拉伯语 COVID-19 谣言。已有若干研究探讨了 Twitter 上的阿拉伯语立场检测；然而，其目标是特定主题，而非谣言 [15,22,6]。近期发布了少量用于阿拉伯语声明核实的立场检测数据集，其中证据要么是新闻文章 [10,3] ，要么是人工编写的句子 [23]。然而，无论是阿拉伯语还是其他语言，目前都没有以推文为谣言、并从权威机构时间线中检索证据的数据集。为弥补这一空白，我们的工作有四项贡献：(1) 定义推文中检测权威机构对谣言立场的任务；(2) 构建并发布首个谣言权威立场（AuSTR）数据集[1] ；(3) 首次研究现有立场检测数据集对我们任务的适用性；(4) 开展失败分析，为该任务的后续研究提供见解。本文旨在回答的研究问题是：现有用于声明核实的阿拉伯语立场检测数据集，是否适用于检测阿拉伯语推文中权威机构的立场。

本文其余部分安排如下：我们在第2节概述 AuSTR 的构建方法，在第3节介绍实验设置，在第4节讨论并分析结果，最后在第5节总结全文并提出一些未来研究方向。

2 构建 AuSTR 数据集

为了构建一个谣言和证据均为推文的 AuSTR 数据集，我们利用事实核查文章以及不同变体的权威推特账号。

利用事实核查文章。试图核实谣言的事实核查员通常会在其事实核查文章中提供一些传播特定谣言的社交媒体帖子（例如推文），以及来自可信权威人士、可作为证据支持其核查结论的其他帖子。为了构建 AuSTR，我们利用了事实核查员提供的两类推文示例：陈述谣言的推文，以及展示来自权威机构的证据的推文。具体而言，我们使用了 AraFacts [4], ，这是一个从 5 个事实核查网站收集的阿拉伯语谣言大型数据集。在这些谣言中，我们仅选取以推文形式表达且也有推文证据的谣言。[2] 随后，我们按如下方式提取谣言—证据对。对于真实和虚假的谣言，我们选取一条推文示例以及所有提供的证据推文，并分别将它们标记为具有赞成立场和反对立场。[3] 如果事实核查员提供了

[1] https://github.com/Fatima-Haouari/AuSTR [2] 由于 AraFacts 未公开此信息，我们联系了其作者以获取相关信息。[3] 我们仅保留以 $text$ 形式表达的证据，而不保留以图像或视频形式表达的证据。

权威机构账号，但表示没有找到支持或否定该谣言的证据，我们便从该权威机构的时间线上选取一到两条在谣言出现前不久发布的推文，并将这些配对标记为无关。

利用权威机构账号。鉴于事实核查员更关注虚假谣言而非真实谣言，按照上述步骤完成后，我们最终只得到 4 对同意样本，而不同意样本有 118 对。为进一步增加同意样本，我们采用了与之前方法相反的做法：先收集证据。具体而言，我们从一组覆盖大多数阿拉伯国家及多个领域（如政治、健康和体育）的权威机构 Twitter 账号（例如部长、总统、大使馆和组织账号等）入手，并从其时间线中选取近期发布的、陈述主张的推文。对于每项主张，我们使用 Twitter 搜索界面查找普通用户表达该主张的推文，同时尽量避免完全重复。最后，为了更贴近真实情形——在权威机构的时间线中，无关推文的比例通常高于同意和不同意推文的比例——我们针对每一对同意和不同意样本，进一步从权威机构时间线中选取一至两条在谣言出现之前发布的近期无关推文，以扩充无关样本。

总体而言，我们最终得到409对样本，涵盖171个独特的主张，其中41个为真，130个为假。在这些样本对中，118对表示不同意（29%），62对表示同意（15%），229对无关（56%）。

3 实验设置

数据集。为了研究现有面向声明核实、以立场为目标的阿拉伯语数据集的实用性，我们采用了以下数据集进行训练：

1. ANS [23] 包含3,786对（主张，句子）样本，其中主张提取自可信来源的新闻文章标题，随后要求标注员分别通过改写和矛盾的方式，针对这些主张生成真实和虚假的句子。这些句子被标注为对主张同意、不同意或其他。2. ArabicFC [10] 包含3,042对（主张，文章）样本，其中主张提取自一个核查有关叙利亚战争的政治声明的事实核查网站，文章则通过在谷歌中使用相关主张进行搜索收集。文章被标注为与主张同意、不同意、讨论或无关。3. AraStance [3]：3,676对（主张，文章）样本，其中主张提取自3个涵盖多个领域和阿拉伯国家的阿拉伯语事实核查网站。文章的收集和标注方式与ArabicFC类似。

为了训练模型，我们仅考虑三个标签，即同意、不同意或无关。对于 ANS 和 AraStance，我们使用了作者提供的相同数据划分；不过，我们将 ArabicFC 的主张分别按 70%、10% 和 20% 划分为训练集、开发集和测试集[4]。在划分

[4] 我们发布 ArabicFC 数据集划分，以便复现研究结果。

数据时，我们将所有具有相同主张的配对分配到同一个数据划分中。表1展示了这三个数据集不同数据划分的规模。由于 AuSTR 的规模有限，本研究仅将其用作测试集，同时使用上述数据集进行训练，以展示它们在我们任务中的效用。

表1. 用于训练的阿拉伯语立场数据集的数据划分。

标签	**ANS**			ArabicFC			AraStance		
	训练集	开发集	测试集	训练集	开发集	测试集	训练集	开发集	测试集
同意	903	268	130	323	32	119	739	129	154
不同意	1686	471	242	66	8	13	309	76	64
无关	63	16	7	1464	198	410	1553	294	358
总计	2652	755	379	1853	238	542	2601	499	576

立场模型。为了训练我们的立场模型，我们对 BERT [16] 进行了微调，以判断证据句/文章是同意、不同意还是与主张无关。我们将主张文本作为句子 A、证据作为句子 B（必要时截断）输入 BERT，并用 [SEP] 标记分隔。最后，我们将 [CLS] 标记的上下文表示作为输入，传入一个添加在 BERT 架构之上的单层分类器。该分类器有三个输出节点，用于计算每种立场类别的概率。

近期发布了多种基于 BERT 的阿拉伯语模型 [7,30,25,21,1]；我们选择了 ARBERT [1]，因为研究表明，它在我们工作中采用的立场数据集上表现更好 [3]。我们采用了作者的设置 [3]，训练模型最多 25 个轮次，将提前停止设为 5，序列长度设为 512。我们在消融研究中使用前文提到的立场数据集的不同组合，训练了 7 个不同的模型。

4 结果与讨论

本初步研究探讨的问题是：现有的立场检测数据集是否适用于我们的任务。为回答这一问题，我们将现有数据集的不同组合用于训练，并使用 AuSTR 进行测试。我们还展示了在这些组合数据上训练的模型在各自对应的域内测试集上的表现。由于这些域内测试集各不相同，其结果无法直接比较，但可作为性能上限的估计。为了评估模型，我们报告了各类别的 F_1 和宏平均-F_1 得分。表 2 展示了所有实验的性能结果，并揭示了若干有趣的观察。

首先，我们注意到，几乎所有模型（少数模型除外）在各自的域内测试集上的表现都优于在 AuSTR 上的表现。这表明领域适应的效果并不显著（因此，训练模型需要使用适用于我们任务的域内数据）。

其次，使用单个立场数据集进行训练时，在 AuSTR 上测试，基于 AraStance 训练的模型在所有指标上的表现都明显优于其他模型。我们注意到，ArabicFC 的类别分布严重不均衡，其中反对类仅占数据的 3.3%，因此即使在其自身的域内测试集上测试，该类别的表现也很差。以往研究也得出了类似结论 [10,3]。至于 ANS，其证据由人工编写，不如来自权威机构的推文真实。相比之下，AraStance 的主张取自三个事实核查网站，[5] 涵盖多个领域和阿拉伯国家，与 AuSTR 类似；其证据则来自记者撰写的文章，而非人工编写。

第三，在 AuSTR 上进行测试时，使用所有数据集的合并数据训练的模型在反对类上的表现最佳；然而，与 AraStance 模型相比，其在赞同类上的表现大幅下降。这确实需要进一步研究。

此外，我们观察到，仅使用 AraStance 进行训练以及将其与其他数据集结合使用时，均取得了最高的 F_1(D)。为探究这一现象，我们人工检查了随机抽取的 10% 反对立场训练文章。我们发现，这些文章包含一些常见词语，例如谣言、不属实、否认和虚假；AuSTR 的一些反对立场推文中也出现了类似的关键词。

最后，我们观察到不同类别之间的表现存在明显差异。以使用所有数据集训练的模型为例，F_1(A) 为 0.74，而 F_1(D) 为 0.65。此外，很明显，检测反对立场是最具挑战性的子任务，我们预计域内训练会对此有所帮助。总体而言，我们认为在推文上进行训练和测试非常不同，因为推文非常简短且非正式，因此需要特殊的预处理。

表 2. 在域内测试集和 AuSTR 上的性能，以逐类 F_1（A：同意，D：不同意，U：无关）和宏观-F_1 衡量。在 AuSTR 上，粗体和下划线值分别表示最佳和次佳性能。

	在域内测试集上测试				在 AuSTR 上测试			
训练集	F_1(A)	F_1(D)	F_1(U)	m-F_1	F_1(A)	F_1(D)	F_1(U)	m-F_1
ANS	0.824	0.901	0.923	0.882	0.653	0.578	0.709	0.647
ArabicFC	0.770	0.090	0.915	0.591	0.641	0.434	0.799	0.625
AraStance	0.898	0.833	0.95	0.894	0.837	0.613	0.865	0.772
ANS+ArabicFC	0.807	0.866	0.899	0.857	0.678	0.587	0.862	0.709
ANS+AraStance	0.893	0.909	0.955	0.919	0.743	0.629	0.847	0.740
ArabicFC+AraStance	0.765	0.555	0.897	0.739	0.754	0.635	0.862	0.750
三个数据集	0.778	0.742	0.889	0.803	0.741	0.646	0.866	0.751

失败分析。我们对 AuSTR 中的 17 个样例进行了失败分析，这些样例均未能被我们的 7 个训练模型正确预测。我们发现

[5] 主张收集自不同于我们用于构建 AuSTR 的来源。

这些失败主要可归因于两个原因：（1）写作风格，即权威机构以第一人称否认关于其本人的谣言。这占所分析失败案例的 64.7%。我们认为，这是因为我们用于训练的立场数据集中，没有任何证据是由权威机构自行撰写的；证据来源要么是记者撰写的新闻文章，要么是标注人员手动编写的新闻标题释义或反驳。（2）间接反对／同意，即权威机构间接否认／支持谣言。表3列出了这两类失败案例。研究结果表明，有必要利用来自 Twitter 的谣言—证据对扩充现有立场数据集，以进一步提升根据推文检测权威机构对谣言立场的效果。

表3。 所有模型都未能正确预测的样例 。这些样例的标准标签为同意或不同意。这些样例依次对应的错误类型为写作风格、间接反对和间接赞同。

谣言推文 [发布时间]	证据推文 [发布时间]
穆尔塔达·曼苏尔最近去世 死于心脏病 袭击。[29-10-2021]	@MortadaMansour：我正在 吃晚餐，几分钟后我会 发布一段语音和视频，让你们放心， 并回复那些打扰了 我在村里的家人并造成了 给我所有粉丝带来的焦虑。[29-10-2021]
埃及不向 其公民提供疫苗 ，海湾 国家赞助了他们: 沙特阿拉伯 / 苏丹国的 阿曼／卡塔尔拒绝他们的 干预，因此没有 除科威特外的其他国家，这个 充满人道精神的国家接纳 他们并供养他们。究竟 有什么神秘的秘密？科威特 对埃及给予特别 待遇。[07-05-2021]	@mohpegypt：有关信息 新冠疫苗。预约 疫苗，请访问网站 http://egcovac.mohp.gov.eg 或前往 最近的卫生站（适用于 难以在线注册的公民）。如需 了解更多信息，请拨打热线： 15335 #一起放心。[10-05-2021]
紧急：第四频道总部 遭到冲击 萨德尔派民兵 运动在首都 巴格达。[04-11-2022]	@MAKadhimi：针对其中一家 伊拉克媒体机构的袭击，以及对其员工生命的威胁， 对其员工生活，是一种 这一行为令人不齿，并且代表着 对.....的最严重侵犯 法律和新闻自由，且不 属于和平、合法的 做法和抗议活动。我们已指示 追究肇事者的责任， 并加强对新闻 机构的保护。[04-11-2022]

5 结论与未来工作

本文定义了在推文中检测权威机构对谣言所持立场这一任务，并发布了首个针对阿拉伯语谣言的该任务数据集。我们研究了现有阿拉伯语数据集对于声明核实中立场检测任务的适用性。根据实验和失败分析，我们发现，尽管现有立场数据集对该任务有一定帮助，但显然仍不充分，因此需要利用 Twitter 数据补充权威机构的立场信息。除了扩充 AuSTR、为该任务提供充足的训练数据（既可单独使用，也可用于扩充现有立场数据集）之外，我们还计划探索并贡献适用于该任务的专用立场模型。

致谢

Fatima Haouari 的工作获得了卡塔尔国家研究基金 GSRA 项目资助，项目编号为 GSRA6-1-0611- 19074。Tamer Elsayed 的工作得以开展，得益于卡塔尔国家研究基金（卡塔尔基金会成员）NPRP 项目的资助，项目编号为 NPRP11S-1204-170060。本文所述观点完全由作者负责。

参考文献

1. Abdul-Mageed, M., Elmadany, A., 等：ARBERT 和 MARBERT：用于阿拉伯语的深度双向Transformer。载于：第59届计算语言学协会年会暨第11届国际自然语言处理联合会议论文集（第1卷：长论文）。第7088–7105页（2021） 2. Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D., Essam, A.：阿拉伯语虚假新闻检测：神经网络与基于Transformer的方法的比较研究。《Complexity》2021（2021） 3. Alhindi, T., Alabdulkarim, A., Alshehri, A., Abdul-Mageed, M., Nakov, P.：AraStance：用于事实核查的阿拉伯语立场检测多国家、多领域数据集。NLP4IF 2021，第57页（2021） 4. Ali, Z.S., Mansour, W., Elsayed, T., Al-Ali, A.：AraFacts：首个大规模自然产生的阿拉伯语主张数据集。载于：第六届阿拉伯语自然语言处理研讨会论文集。第231–236页（2021） 5. Alqurashi, S., Hamoui, B., Alashaikh, A., Alhindi, A., Alanazi, E.：吃大蒜可预防COVID-19感染：检测Twitter阿拉伯语内容中的错误信息。arXiv预印本 arXiv:2101.05626（2021） 6. Alqurashi, T.：利用阿拉伯语Twitter数据分析COVID-19疫情期间沙特阿拉伯王国远程教育的立场。《Sensors》22(3)，1006（2022） 7. Antoun, W., Baly, F., Hajj, H.：AraBERT：用于阿拉伯语理解的基于Transformer的模型。载于：LREC 2020研讨会，语言资源与评测会议，2020年5月11日至16日。第9页（2020）

8. Bai, N., Meng, F., Rui, X., Wang, Z.：用于立场分类和谣言真实性检测的多任务注意力树神经网络。《Applied Intelligence》第1–11页（2022） 9. Bai, N., Meng, F., Rui, X., Wang, Z.：基于源—回复对话树卷积神经网络的谣言检测。《Computing》104(5)，1155–1171（2022） 10. Baly, R., Mohtarami, M., Glass, J., M`arquez, L., Moschitti, A., Nakov, P.：在统一语料库中整合立场检测与事实核查。载于：2018年北美计算语言学协会分会年会：人类语言技术会议论文集，第2卷（短论文）。第21–27页。计算语言学协会，路易斯安那州新奥尔良（2018年6月） 11. Barr´on-Cede˜no, A., Alam, F., Caselli, T., Da San Martino, G., Elsayed, T., Galassi, A., Haouari, F., Ruggeri, F., 等：CLEF-2023 CheckThat! 实验室：新闻文章及其来源的可核查性、主观性、政治偏见、事实性与权威性。载于：第45届欧洲信息检索会议（ECIR'23）论文集（2023） 12. Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., Huang, J.：利用双向图卷积网络检测社交媒体上的谣言。载于：AAAI人工智能会议论文集。第549–556页（2020） 13. Chen, L., Wei, Z., Li, J., Zhou, B., Zhang, Q., Huang, X.J.：对消息交互演变进行建模以消解谣言。载于：第28届国际计算语言学会议论文集。第6377–6387页（2020） 14. Choi, J., Ko, T., Choi, Y., Byun, H., Kim, C.k.：结合注意力机制的动态图卷积网络，用于检测社交媒体上的谣言。《PLOS ONE》16(8)，e0256039（2021） 15. Darwish, K., Magdy, W., Zanouda, T.：在用户相似性特征空间中改进立场预测。载于：2017年IEEE/ACM社交网络分析与挖掘国际会议论文集。第145–148页（2017） 16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向Transformer预训练。arXiv预印本 arXiv:1810.04805（2018） 17. Dougrez-Lewis, J., Kochkina, E., Arana-Catania, M., Liakata, M., He, Y.：PHEMEPlus：利用外部证据丰富社交媒体谣言验证。载于：第五届事实提取与验证研讨会（FEVER）论文集。第49–58页（2022） 18. Elhadad, M.K., Li, K.F., Gebali, F.：COVID-19-FAKES：用于检测COVID-19误导性信息的Twitter（阿拉伯语/英语）数据集。载于：智能网络与协作系统国际会议论文集。第256–268页。Springer（2020） 19. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.：ArCOV19-Rumors：用于错误信息检测的阿拉伯语COVID-19 Twitter数据集。载于：第六届阿拉伯语自然语言处理研讨会论文集。第72–81页（2021） 20. Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z.S., Hamdan, B., Elsayed, T., Barr´on-Cedeno, A., Da San Martino, G., Nakov, P.：CheckThat! 2020阿拉伯语概览：社交媒体中主张的自动识别与验证。载于：CLEF（2020） 21. Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., Habash, N.：阿拉伯语预训练语言模型中变体、规模与任务类型的相互作用。载于：第六届阿拉伯语自然语言处理研讨会论文集。第92–104页（2021）

22. Jaziriyan, M.M., Akbari, A., Karbasi, H.: ExaASC：阿拉伯语通用目标式立场检测语料库。载于：2021年第11届计算机工程与知识国际会议（ICCKE）。第424–429页。IEEE（2021）23. Khouja, J.: 立场预测与声明核实：阿拉伯语视角。载于：第三届事实抽取与核验研讨会（FEVER）论文集。计算语言学协会，美国西雅图（2020）24. Kumar, S., Carley, K.: 利用带卷积单元的树形长短期记忆网络，在社交媒体对话中预测立场和谣言真实性。载于：计算语言学协会第57届年会论文集。计算语言学协会，意大利佛罗伦萨（2019年7月）25. Lan, W., Chen, Y., Xu, W., Ritter, A.: 阿拉伯语信息抽取预训练Transformer的实证研究。载于：2020年自然语言处理实证方法会议（EMNLP）论文集。第4727–4734页。计算语言学协会，线上（2020年11月）26. Liu, Y., Wu, Y.F.B.: 利用循环网络和卷积网络进行传播路径分类，以早期检测社交媒体上的假新闻。载于：第三十二届AAAI人工智能会议（2018）27. Ma, J., Gao, W., Wong, K.F.: 利用树结构递归神经网络检测Twitter上的谣言。载于：计算语言学协会第56届年会论文集（第1卷：长论文）。第1980–1989页（2018）28. Mahlous, A.R., Al-Laith, A.: COVID-19疫情期间阿拉伯语推文中的假新闻检测。《国际高级计算机科学与应用杂志》12(6)（2021）29. Roy, S., Bhanu, M., Saxena, S., Dandapat, S., Chandra, J.: gDART：利用离散注意力表示改进社交媒体谣言核验。《信息处理与管理》59(3)，102927（2022）30. Safaya, A., Abdullatif, M., Yuret, D.: KUISAIL在SemEval-2020第12项任务中的工作：用于社交媒体冒犯性言论识别的BERT-CNN。载于：第十四届语义评测研讨会论文集。第2054–2059页。计算语言学国际委员会，巴塞罗那（线上）（2020年12月）31. Sawan, A., Thaher, T., Abu-el rub, N.: 用于阿拉伯语推文假新闻识别的情感分析模型。载于：2021年IEEE第15届信息与通信技术应用国际会议（AICT）。第1–6页（2021）32. Song, C., Shu, K., Wu, B.: 用于假新闻检测的时序演化图神经网络。《信息处理与管理》58(6)，102712（2021）33. Wu, L., Rao, Y., Jin, H., Nazir, A., Sun, L.: 同一分享，不同吸收：用于假新闻检测的筛选式多任务学习。载于：2019年自然语言处理实证方法会议与第九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，香港（2019年11月）34. Yu, J., Jiang, J., Khoo, L.M.S., Chieu, H.L., Xia, R.: 用于社交媒体对话中立场感知谣言核验的耦合层级Transformer。载于：2020年自然语言处理实证方法会议（EMNLP）论文集。第1392–1401页。计算语言学协会，线上（2020年11月）