arXiv:2305.01011v1 [cs.CL] 1 May 2023

通过软域迁移进行特征增强的欺骗检测

Sadat Shahriar*、Arjun Mukherjee 和 Omprakash Gnawali

美国德克萨斯州休斯敦市休斯敦大学，邮编 77004，
sshahria@cougarnet.uh.edu，arjun@cs.uh.edu，odgnawal@central.uh.edu

摘要。 在信息爆炸的时代，欺骗者利用不同的信息领域或媒介来欺骗用户，例如新闻、电子邮件和推文。尽管已有大量研究致力于检测这些领域中的欺骗行为，但新事件中的信息匮乏，使得这些领域必须相互关联，以共同对抗欺骗。为建立这种关联，我们提出了一种特征增强方法，利用神经模型的中间层表示。与自身领域基线模型相比，我们的方法最多可提升 6.60%。我们发现，推文对于假新闻和钓鱼邮件检测最有帮助，而新闻则最有助于推文谣言检测。我们的分析为领域知识迁移提供了有益见解，有助于构建比现有研究成果更强大的欺骗检测系统。

关键词： 欺骗 · BERT · LSTM · 网络钓鱼 · 假新闻 · 谣言

1 引言

文本中的欺骗意味着发送者故意歪曲事实或营造虚假印象 [2]。文本中的欺骗可能出现在新闻、推文、电子邮件等多个领域，已有研究致力于检测特定领域中的欺骗行为 [13, 16, 19]。尽管欺骗者在各个领域采用的欺骗手法各有风格，但各种欺骗行为的目的都是欺骗人们。因此，在一个领域中检测欺骗的经验可以用于辅助检测其他领域中的欺骗。本文通过探讨如何利用领域 A 中欺骗检测的优势来检测领域 B 中的欺骗，开展软域迁移研究。我们还探讨了在源领域不具有欺骗性的情况下，领域迁移的有效性。

研究人员从整体角度研究欺骗，希望捕捉欺骗风格中的细微差异 [12]。然而，由于不同领域中的欺骗差异很大，目前尚不清楚是否存在这样清晰的模式。此外，还需要进一步研究，以量化一个领域从另一个领域获得的“帮助”。为此，在实现领域知识迁移方面仍存在显著的研究空白。我们将欺骗领域定义为

* 通讯作者

欺骗发生时所涉及的信息来源，我们将假新闻、钓鱼邮件和谣言作为不同领域中的欺骗实例。作为非欺骗领域，我们采用新闻组主题、情感检测和维基百科本体检测。因此，我们提出第一个研究问题（**RQ1**）：从不同领域迁移知识能否帮助改进欺骗检测？此外，我们提出第二个研究问题（**RQ2**）：在欺骗领域与非欺骗领域之间，哪些领域组合对检测欺骗最有帮助？为回答这些问题，我们针对三个欺骗领域和三个非欺骗领域，训练了六种不同的 BERT 和 LSTM 模型 [3,6] 。我们收集目标领域的中间层信息，并通过结合中间层信息来利用外部领域的优势，再训练一个全连接深度神经网络（FC-DNN）来检测欺骗。通过这种特征增强方式，我们将其他领域的知识注入输入信息，从而在 FC-DNN 模型中利用这些知识。

这项研究意义重大，体现在多个方面。首先，在许多领域中，欺骗性数据十分稀缺。例如，个人和企业不愿分享自己收到的钓鱼邮件，以免陷入尴尬境地 [1]。其次，随着社交媒体的兴起，一旦出现新事件，信息便会在不同领域间传播。例如，COVID-19 的出现导致新闻、推文和 Facebook 帖子中的错误信息大幅激增；因此，仅依赖单一领域来学习欺骗模式，会遗漏其他领域的信息。最后，本研究可指导研究人员制定从不同领域选择性迁移知识的方案，并找出欺骗行为的通用模式。据我们所知，我们工作的创新之处在于，这是首项探索领域迁移在欺骗检测中有效性的研究，并为后续有望取得成果的研究开辟了新方向。

2 相关研究

Hern´andez-Casta˜neda 等人提出了一种使用 SVM 进行跨领域欺骗检测的方法，其数据集包含针对不同主题的观点 [5]。他们旨在不同的实验训练-测试设置中，找出一组通用特征。 [14] 和 [9]也开展了类似研究。在假新闻检测任务中，P`erez-Rosas 等人针对两个不同的数据集进行了跨领域实验，并展示了模型泛化能力面临的挑战 [11]。Gautam 和 Jerripothula 使用 Spinbot、Grammarly 和基于 GloVe 的方法进行跨领域假新闻检测[4]。然而，这些研究旨在使欺骗检测系统不受主题限制，而非不受欺骗媒介限制。因此，在跨领域设置下发挥欺骗检测能力，仍是一个尚未探索的领域。

3 数据集

对于电子邮件领域，我们使用来自 ACM IWSPA 反网络钓鱼试点项目的钓鱼邮件数据集 2018 [17]。训练集包含 5092 封合法邮件和 629

钓鱼邮件，测试集大小为 4300，其中 3825 封为合法邮件，475 封为钓鱼邮件。我们将钓鱼邮件标记为欺骗性，将非钓鱼邮件标记为非欺骗性。对于新闻域，我们使用 LIAR 数据集 [18], ，该数据集包含六种新闻标签，即：真实、大多真实、半真、大多虚假、虚假、彻头彻尾的虚假。我们将前两类文本视为非欺骗性文本，将后四类视为欺骗性文本，遵循文献 [15]的做法。对于推文域，我们使用 PHEME 数据集，其中包含 2402 条谣言文本和 4023 条非谣言文本 [20]。我们将谣言推文标记为欺骗性，将非谣言推文标记为非欺骗性。

对于非欺骗性任务，我们使用了三个数据集。IMDB 电影评论数据集包含 50,000 条评论，标注为正面或负面 [10]。20 新闻组数据集包含约 18000 个样本，标签对应于关于 20 个主题的新闻组帖子 [7]。维基百科主题分类数据集包含 342,782 篇文章，涵盖 9 个主题类别 [8]。我们从每个非欺骗性领域中随机抽取 10,000 条文本，并按 80:20 的比例划分训练集和测试集。

4 方法

我们使用两种神经网络模型作为基线方法：来自变换器的双向编码器表示（BERT）模型和长短期记忆（LSTM）模型 [3, 6]。BERT 模型由具有自注意力能力的变换器层构成，这些层包含编码器和解码器。我们对基线自身领域 BERT 模型进行微调，提取模型最后的 [CLS] 层，并使用全连接层和 Softmax 来完成下游分类任务。LSTM 是一种高效的循环神经网络（RNN）变体，并加入了解决长期依赖问题的机制。我们将词序列作为双层 LSTM 模型的输入，并在其顶部使用全连接层对文本进行分类。

在特征增强过程中，我们执行中间层拼接（**ILC**），如图 1 所示。对于 BERT 模型，我们提取不同领域经过自身领域训练的 [CLS] 层并将它们拼接起来，作为目标领域的增强特征集。同样，对于 LSTM 模型，我们提取不同领域的最终 LSTM 层表示的输出并将它们拼接起来。随后，将增强特征集输入一个 2 层全连接（FC）模型来检测欺骗行为。最后，使用从训练集中抽取 20% 数据生成的验证集来设置所有网络超参数。

5 结果与讨论

表 1 展示了使用基于 BERT 的 ILC 模型，在不同欺骗领域进行特征增强的效果。对于电子邮件领域，我们观察到，新闻域使钓鱼检测的 F1 分数提高了 2.31%，推文域则使性能提高了 4.89%。将推文域和新闻域结合后，F1 分数提高了 6.60%。对于新闻域和推文域，我们也观察到，欺骗性特征增强提升了性能。电子邮件有助于假新闻

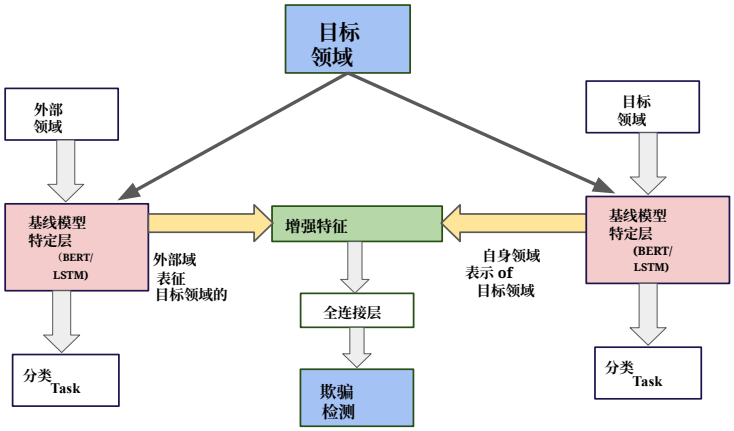


图1. 通过软域迁移进行特征增强，以提高欺骗检测性能。我们将目标域和外部域的基线模型的中间层表示进行拼接，以增强欺骗性特征，并将增强后的特征输入全连接网络进行欺骗检测。

检测，性能提高了 0.75%；推文则有助于提升 1.69%。推文谣言检测的性能从新闻域获得了 1.36% 的提升，从电子邮件领域获得了 1.14% 的提升。不过，与新闻域和推文域相比，电子邮件领域的性能提升更大。作为预训练模型，BERT 更有可能在新闻和推文等公开文本上表现出色，因此基线模型能够更好地理解这两个领域中的欺骗行为。因此，其他欺骗领域带来的增强，对钓鱼邮件检测的提升幅度大于对其他领域欺骗检测的提升幅度。

表 1. 基于 BERT 模型的跨领域欺骗检测。**E**、**T** 和 **N** 分别代表电子邮件、推文和新闻。例如，“ILC-TN”代表结合推文域和新闻域的 ILC 模型。

	基线 – BERT		**ILC – EN**		**ILC – TN**		**ILC – ET**		**ILC – ETN**	
领域	**F1 分数**	**ACC**	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**
电子邮件	80.99	95.41	83.31	96.03	–	–	85.88	96.79	**87.59**	**97.39**
News	76.88	63.55	77.63	63.93	78.57	**67.80**	–	–	**78.80**	67.56
推文	80.34	84.79	–	–	81.70	86.23	81.48	85.99	**82.07**	**86.77**

我们进一步通过使用奇异值分解（SVD）方法将数据投影到二维子空间，研究跨领域特征增强的有效性。图2清楚地显示，当将推文域和新闻域与电子邮件域相结合时，特征分离得到改善，距离也随之增大。

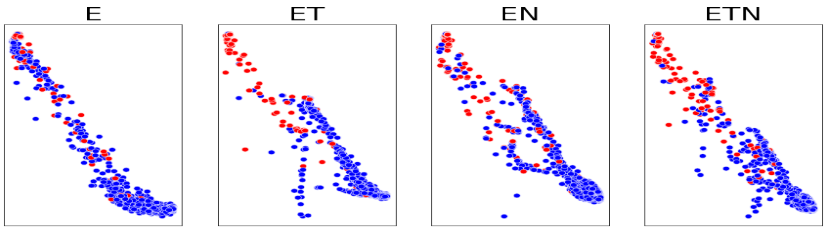


图 2. 电子邮件领域的 SVD 降维表示（BERT 模型）：其自身领域特征与不同欺骗领域的中间层拼接特征。蓝点表示非欺骗，红点表示欺骗

当三个领域全部拼接在一起时，欺骗性样本和非欺骗性样本的云中心距离缩短了50.23%。

采用基于 LSTM 的特征增强技术后，我们牺牲了整体性能，但与 BERT 不同，我们没有使用预训练模型。因此，我们观察到三个欺骗领域的性能均得到稳定提升（表 2）。在电子邮件领域，与基于 BERT 的 ILC 模型一样，推文域带来的帮助最大；结合增强后，整体性能提升最高达 3.97%。推文域对电子邮件和新闻领域的帮助最大。不过，将三个领域结合时性能最佳，新闻域和推文域的性能分别提升了 4.82% 和 3.39%。作为单独的领域，新闻域对推文域的帮助最大，使 F1 分数提升了 1.76%。

表 2. 基于 LSTM 模型的跨领域欺骗检测。

	基线 – LSTM		**ILC – EN**		**ILC – TN**		**ILC – ET**		**ILC – ETN**	
领域	**F1 分数**	**ACC**	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**
电子邮件	71.23	92.23	72.04	92.45	–	–	74.96	94.68	**75.20**	**94.87**
News	72.71	62.03	74.69	62.43	75.45	63.18	–	–	**77.53**	**63.59**
推文	73.11	77.16	–	–	74.87	79.29	74.18	79.00	**76.50**	**82.13**

表 3. 基于 BERT 的欺骗检测：通过非欺骗领域进行特征增强。

	情感		**新闻组**		**维基百科**		**合并**	
领域	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**
电子邮件	81.24	96.09	80.95	96.14	81.04	96.13	81.26	96.09
News	77.43	63.10	77.58	63.19	77.89	63.51	78.05	63.79
推文	81.41	86.07	81.32	85.91	81.24	85.80	81.89	86.30

接下来，我们研究使用 BERT 模型通过非欺骗领域增强后的欺骗检测性能。从表 3 可以看出，情感和维基百科略微提升了钓鱼邮件检测的性能；结合这些领域后，F1 分数提升了 0.27%。对于新闻域，维基百科带来的帮助最大；结合所有领域后，F1 分数整体提升了 1.16%。情感对检测推文中的谣言最有帮助，使 F1 分数提升了 1.07%；结合各领域后，性能提升了 1.55%。我们还发现，基于 LSTM 的 ILC 模型表现相似，在结合各领域时性能最佳，电子邮件、新闻和推文域的欺骗检测性能分别提升了 2.60%、5.04% 和 3.10%（表 4）。

通过上述讨论，我们发现，来自不同领域的特征增强有助于提升欺骗检测任务的效果。然而，当外部域具有欺骗性时，性能提升幅度大于外部域不具有欺骗性的情况，因此会发生软域迁移。

表 4. 通过来自非欺骗性领域的特征增强进行基于 LSTM 的欺骗检测。

	情感		**新闻组**		**维基百科**		**组合**	
领域	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**	**F1 分数**	**准确率**
电子邮件	73.18	93.41	71.49	92.18	72.17	92.77	73.83	94.11
News	72.80	61.97	72.96	62.11	75.50	63.37	76.48	63.41
推文	73.93	78.32	73.00	77.91	75.17	81.56	76.21	82.02

6 结论与未来工作

尽管现有研究已在许多领域开展了欺骗检测，但如何将一个领域中获得的知识迁移到另一个领域，以实现跨领域欺骗检测，仍存在研究空白。本文采用神经模型的中间层拼接方法填补了这一空白。未来可从以下几个方向继续研究。首先，我们的分析仅限于三个领域。还可以探索其他领域，例如评论、Facebook 帖子和 WhatsApp 消息转发，以开展跨领域欺骗检测。此外，我们在每个领域仅使用了一个数据集。未来可在这些领域使用更多数据集开展研究，从而进一步验证我们的假设。

致谢 本研究部分由 NSF 1838147 和 ARO W911NF-20-1-0254 项目资助。本文所述观点和结论仅代表作者意见，不代表资助方立场。尽管本文包含任何版权声明，美国政府仍有权为政府用途复制和分发再版。

参考文献

1. Aassal, A.E., Moraes, L., Baki, S., Das, A., Verma, R.：在 ACM IWSPA 2018 上开展反网络钓鱼试点，使用适用于不平衡数据集的新指标评估性能。第 2–10 页。http://ceur-ws.org/Vol-2124/#anti-phishing-pilot 2. Burgoon, J.K., Buller, D.B.：人际欺骗：III. 欺骗行为对感知到的交流和非言语行为动态的影响。《非言语行为杂志》**18**(2), 155–184（1994年6月）。https://doi.org/10.1007/BF02170076, https://doi.org/10.1007/BF02170076 3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深层双向 Transformer 预训练。ArXiv **abs/1810.04805** (2019) 4. Gautam, A., Jerripothula, K.R.：SGG：基于 Spinbot、Grammarly 和 GloVe 的假新闻检测。2020 IEEE第六届多媒体大数据国际会议（BigMM），第 174–182 页（2020） 5. Hern´andez-Casta˜neda, ´A., Calvo, H., Gelbukh, A., Flores, J.J. G.：使用支持向量网络进行跨领域欺骗检测。《软计算》**21**(3), 585–595（2017年2月）。https://doi.org/10.1007/s00500-016-2409-2, https://doi.org/10.1007/s00500-016-2409-2 6. Hochreiter, S., Schmidhuber, J.：长短期记忆。《神经计算》**9**(8), 1735–1780 (1997) 7. Lang, K.：Newsweeder：学习过滤网络新闻。载于：第十二届国际机器学习会议论文集。第 331–339 页（1995） 8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.：DBpedia——从维基百科提取的大型多语言知识库。《语义网》**6**, 167–195 (2015) 9. Li, J., Ott, M., Cardie, C., Hovy, E.：探索识别欺骗性观点垃圾信息的一般规则。载于：计算语言学协会第52届年会论文集（第1卷：长篇论文）。第 1566–1576 页。计算语言学协会，巴尔的摩，马里兰州（2014年6月）。https://doi.org/10 1147 10 3115 1 14 1147, 14 . /v /P - https://aclanthology.org/P - . Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.：为情感分析学习词向量。载于：计算语言学协会第49届年会：人类语言技术论文集。第 142–150 页。计算语言学协会，美国俄勒冈州波特兰（2011 11 1015 年6月），http://www.aclweb.org/anthology/P - . Perez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.：自动检测假新闻。载于：COLING（2018）12. Shahriar, S., Mukherjee, A., Gnawali, O.：一种领域无关的欺骗检测整体方法。载于：自然语言处理近期进展国际会议论文集（RANLP 2021）。第 1308–1317 页（2021）13. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.：社交媒体上的假新闻检测：数据挖掘视角。《ACM SIGKDD探索通讯》**19**(1), 22–36 (2017) 14. S´anchez-Junquera, J., Villase˜nor-Pineda, L., y G´omez, M.M., Rosso, P., Stamatatos, E.：通过屏蔽领域特定信息进行跨领域欺骗检测。《模式识别快报》**135**, 122–130 (2020)。https://doi.org/https://doi.org/10.1016/j.patrec.2020.04.020, https://www.sciencedirect.com/science/article/pii/S0167865520301422

15. Upadhayay, B., Behzadan, V.：情感说谎者：用于虚假声明分类的扩展语料库和深度学习模型（2020） 16. Varshney, G., Misra, M., Atrey, P.K.：网络钓鱼检测方案的综述与分类。《安全与通信网络》**9**(18), 6266–6284（2016） 17. Verma, R.M., Zeng, V., Faridi, H.：安全挑战中的数据质量：网络钓鱼、恶意软件和入侵检测数据集的案例研究。载于：2019年ACM SIGSAC计算机与通信安全会议论文集。第2605–2607页。CCS '19, 计算机协会，纽约，美国（2019）。https://doi.org/10.1145/3319535.3363267, https://doi.org/ 10.1145/3319535.3363267 18. Wang, W.Y.：“骗子，骗子，裤子着火了”：一个用于假新闻检测的新基准数据集。载于：第55届计算语言学协会年会论文集（第2卷：短篇论文）。第422–426页。计算语言学协会，加拿大温哥华（2017年7月）。https://doi.org/10.18653/v1/P17-2067, https://www.aclweb.org/ anthology/P17 2067 19 - . Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.：社交媒体中的谣言检测与处理：综述。《ACM计算调查》（CSUR）**51**(2), 1–36（2018） 20. Zubiaga, A., Hoi, G.W.S., Liakata, M., Procter, R.：PHEME 谣言与非谣言数据集（2016年10月）。https://doi.org/10.6084/m9.figshare.4010619.v1