当社交媒体上的网络攻击性预测遇上 BERT

周振坤

中国北京首都经济贸易大学统计学院

电子邮箱：zhenkun@cueb.edu.cn

于梦莉（通讯作者）

中国天津南开大学新闻与传播学院；中国天津南开大学融合媒体研究中心。电子邮箱：mengliyu@nankai.edu.cn

何雨欣

中国北京首都经济贸易大学统计学院 电子邮箱：heyuxin@cueb.edu.cn

彭星宇

中国北京航空航天大学软件开发环境国家重点实验室 电子邮箱：xypeng@buaa.edu.cn

摘要：

网络攻击行为日益成为侵蚀社交媒体环境的普遍现象。然而，由于主观性强且成本高昂，传统自我报告问卷难以应用于当前的网络环境。本研究基于前沿深度学习算法，提出了网络攻击性预测模型。我们基于 320 名活跃微博用户的社交媒体活动，构建了基础特征、动态特征和内容特征。我们从三个维度阐释网络攻击行为：社会排斥、恶意幽默和诱发内疚。随后，我们结合预训练 BERT 模型构建预测模型。实证结果表明，该模型表现优异；与缺乏额外预训练信息的传统机器学习模型相比，BERT 模型具有更强的预测能力。本研究为网络攻击性预测提供了坚实的理论模型。此外，本研究也有助于探究网络攻击行为，并促进社交媒体平台的组织管理。

关键词：

社交媒体、网络攻击行为、行为预测、深度学习、BERT

引言

传统上，攻击行为被视为在面对面情境中，个体处于生理唤起状态时，故意伤害他人的行为。[1] 在当前的网络时代，信息技术和社交媒体的迅猛发展改变了人际交往的性质与方式。社交媒体的使用已成为全球人们日常生活中的普遍现象。借助感知匿名机制，用户能够更加自发地表达情感，也更倾向于在社交网络中这样做。因此，社交媒体为用户实施攻击行为开辟了新的渠道。网络攻击行为是指通过手机、计算机及其他电子设备故意伤害他人的行为。[2,3] 其中包括社会排斥、恶意幽默和诱发内疚等敌对行为。[3,4] 以往研究探讨了热门社交网络平台（如 Facebook 和 Twitter）上的攻击行为，并证实用户特征（如年龄和性别）会对其产生影响。[5] 例如，研究人员发现，男性比女性更可能在 Facebook 上实施攻击行为，例如发送侮辱性信息和发布攻击性评论。[6] 与此同时，先前研究关注了青少年和儿童等特定群体。Pabian 等人[7] 探讨了青少年在 Facebook 上的黑暗三人格特质与攻击行为之间的关系，并发现 Facebook 使用强度与青少年的攻击行为显著相关。

显然，网络攻击行为会造成负面后果，包括关系破裂、物质使用、违规行为，甚至重大犯罪活动。[3,8] 由于其负面影响，研究人员一直在尝试识别和检测网络攻击行为。[8,9] 作为一种稳定的个体心理行为，传统上，网络攻击行为是通过自我报告直接测量的。[3] 然而，受成本、主观性和灵活性低等多种局限因素影响，自我报告会降低数据质量和结论的效度。[10] 社交媒体的普及为研究用户的人格和心理行为提供了新途径，可借助大量数据进行探索。[11] 例如，Sharif 和 Hoque[9] 提出了一种攻击性文本分类系统，将攻击性孟加拉语文本划分为宗教、性别、言语和政治攻击类别。Sadiq 等人[8] 从社交媒体提取评论，并将评论文本分为三种不同的攻击类别。然而，关于

结合深度学习方法，根据社交媒体活动和个人特征预测用户网络攻击行为的研究仍然十分匮乏。

目前只有少量研究通过文本分析探讨网络攻击行为。[12,13] 其中，Chavan 和 Shylaja[12] 使用 TF-ID 和 N-gram 等传统特征提取技术，检测社交网络上的网络欺凌评论。Al-Garadi 等人[13] 详细阐述了特征选择，并测试了多种机器学习算法，以预测网络欺凌行为。然而，传统机器学习技术在处理未经加工的自然数据时受到限制，需要精心设计特征并具备领域专业知识。[14] 随着深度神经网络的发展，深度学习在解决上述问题方面取得了重大进展。具体而言，事实证明，深度学习中的语言模型预训练在自然语言处理方面表现出色。[15] 因此，本研究采用先进的 BERT（基于变换器的双向编码器表示）模型来预测用户的网络攻击行为。

网络攻击行为在社交媒体平台的运营和互动环境中发挥着关键作用。海量社交媒体数据与计算社会科学为描绘个体的性格和行为提供了先进途径。通过收集社交媒体用户的自我报告和在线数据，我们提出了一种新方法，即利用深度学习模型客观地预测个体的网络攻击行为。本研究有助于检测这一现象并优化平台组织。

方法

参与者与数据收集

本研究采用在线调查方法论收集数据。我们邀请中国北京地区的学生参与这项调查，共收集到456份有效问卷。所有参与者均提供了自己的微博账号，因此我们收集了他们的微博活动和个人资料数据。为确保用户数据的质量，我们仅选取发布过20条以上微博帖子的活跃用户。此外，入选用户自2020年以来曾发布过帖子，并且至少连续两个月在微博上保持明显活跃。本研究选取了320名活跃对象，研究其社交媒体上的网络攻击行为，其中男性74人，女性246人。我们采用间接攻击量表（攻击者版）

版本[4,5] 测量参与者的网络攻击行为，包括三种类型：社会排斥（10个条目）、恶意幽默（9个条目）和诱发内疚（6个条目）。具体而言，社会排斥指 通过在社交层面排斥受害者而起作用的行为，例如隐瞒信息、将某人排除在活动之外，以及挑拨他人反对某人／社会操纵；恶意幽默指 主要通过使用幽默伤害受害者的行为，例如以讽刺进行侮辱、故意使人难堪和恶作剧； 诱发内疚指故意使他人产生内疚感的行为，例如情感勒索、施加不当压力和胁迫。研究人员向参与者说明：“请根据您在微博上与联系人的互动情况回答以下问题。”各条目采用7点李克特量表测量，范围从“1=从不这样做”到“7=总是这样做”。三个分量表的α系数介于0.76至0.92之间。社会排斥（μ=2.44， σ=1.11）、恶意幽默（μ=1.69， σ=0.77）和诱发内疚（μ=2.03， σ=1.02）的得分通过计算相关条目的平均值获得（μ 为均值，σ 为标准差）。

用户分类

受相关研究启发，[10] 根据均值和标准差，将研究对象划分为高、中性和低三个组，以满足心理预测的目标是合理的。例如，我们将高社会排斥标记为类别 1，中性社会排斥标记为类别 0，低社会排斥标记为类别 -1。具体而言，高社会排斥组由得分大于 3.00（$\mu + \sigma/2$）的受试者组成，低组由得分低于 1.88（$\mu - \sigma/2$）的受试者组成，中性组则由得分介于 1.88 至 3.00 之间的受试者组成。同时，阈值（$\mu \pm \sigma/2$）有助于平衡三个类别的规模，避免机器学习模型产生偏差。根据上述方法，恶意幽默和诱发内疚也分别划分为三个组。将 320 名活跃用户分为三类后，我们可以得到一个训练集，并据此建立不再需要自我报告的机器学习模型。

表 1. 用于预测的类别划分界限及各类别的用户数量。

	$\mu - \sigma/2$	$\mu + \sigma/2$	数量 高组	数量 神经群体	数量 低群体
社会排斥	1.88	3.00	82	135	103
恶意幽默	1.30	2.08	86	98	136
诱发内疚	1.52	2.54	84	143	93

特征

针对社会排斥、恶意幽默和诱发内疚这三个预测目标，我们提取并分析三组特征，用以刻画用户特征，同时也将其作为模型输入。第一组和第二组特征分别包含用户的基本特征和动态特征。第三组特征刻画用户推文的内容，主要描述用户推文中的文本。根据我们先前的研究，[11] 我们提取**（1）基本特征**，以反映用户的人口统计特征、静态信息以及在社交媒体中的基础互动。此外，我们将带图片的帖子的数量和比例加入互动特征；**（2）动态特征**旨在反映微博中不同时间段的社交互动模式。此处的社交互动包括发帖、提及和转发。我们计算了由一周中不同小时或日期的互动平均发生频率构成的向量；**（3）社交媒体中的内容特征**可以有效反映心理特质。[16,17] 我们采用预训练的 300 维词嵌入来表示微博中的每个词，这些词嵌入来自基于正点互信息（PPMI）的预训练中文词向量。[18]我们对词嵌入取平均，得到微博级内容向量，再对微博级向量取平均，得到用户级的 300 维内容向量。 用户发帖中的 情绪可以据此计算。[19,20] 我们使用 MoodLens 模型进行情绪分类，[21] 将短文本分为五种情绪之一，包括愤怒、厌恶、快乐、悲伤和恐惧。对于每位活跃用户，我们计算五项特征，分别表示相应情绪文本所占的比例。请注意，用户个人简介中的文本也被纳入内容特征。在这项研究中，我们还使用预训练 BERT 模型从文本内容中提取特征，相关介绍见“模型”一节。提取后得到一个 134 维的基本特征与动态特征向量，以及一个 300 维的内容（词嵌入）特征向量。

模型

我们构建了三种传统机器学习模型来预测用户的网络攻击行为：逻辑回归（LR, C=1）、支持向量机（SVM，径向基函数核，C=1）和神经网络（NN，包含128、64和32个神经元的3个隐藏层，共9万个参数）。此外，自然语言处理领域的进展，例如 BERT（基于变换器的双向编码器表示），有助于推动相关系统的开发

，从而有望实现文本编码的自动化，进而显著推动心理学研究。我们使用 BERT 为每位社交媒体用户生成文本表示。BERT 旨在通过在所有层中联合考虑左右文脉，从未标注文本中预训练深度双向表示。这种简单而强大的方法在十一项自然语言处理任务上取得了新的最先进成果。因此，在本研究中，我们使用预训练的 BERT 神经网络来预测用户的心理特征，尤其是网络攻击行为。需要指出的是，在本研究中，BERT 是指采用预训练 BERT 的神经网络。本研究尝试使用 BERT架构的一种先进改进版本——文心一言3.0，从微博内容中提取文本内容特征（嵌入）。该模型结合了自回归网络和自编码网络，使预训练模型能够通过少样本学习实现自然语言理解。[22] 我们通过该预训练模型对推文内容进行编码，并据此计算 BERT 生成的平均词嵌入，作为512维内容嵌入。在输出层，我们还加入134维的基础+ 动态特征，以用于训练和微调，最终得到 BERT 模型。匿名数据集和代码已发布在 GitHub 上（参见“数据和代码可用性”部分）。

结果

通过准确率（准确率）、F1分数（F1值）和曲线下面积来验证上述模型的预测性能。通过计算每个标签 精确率 和召回率（召回率）可得：F1 $=1\,3\sum 2(Pri\cdot ReiPri+Rei\,i=-1,$ $0,1$. 此外，还可以测量每个类别相对于其余类别的受试者工作特征曲线下面积（曲线下面积），即一对多曲线下面积。[22] 即使测试数据集并非不平衡，F1值和曲线下面积也可作为模型验证的替代指标。

表2展示了模型对三个预测目标的预测性能。对于社会排斥，采用基础 + 动态 + 内容特征的 BERT 模型表现出色，准确率为=62.50%，F 值为1=53.55%，曲线下面积为=67.69%。BERT 的表现显著优于神经网络和其他机器学习模型。比较不同特征组（即是否包含内容特征）的结果时，并没有

显著差异。除 BERT 模型外，我们还比较了采用不同特征的模型的曲线下面积。采用基础 + 动态特征的逻辑回归模型的曲线下面积为 58.55%，甚至略优于采用 基础 + 动态 + 内容特征（58.36%）的神经网络模型。这表明，采用简单的传统词嵌入时，微博内容可能无助于预测人们遭受社会排斥的情况。对于另外两个目标——恶意幽默和诱发内疚，我们也得到了相同的结果：BERT 始终表现最佳。与次优模型相比，BERT 的准确率分别提高了 19.69%、12.81% 和 15.00%。

我们认为，利用 BERT 等先进的深度学习预训练模型，有望提升网络攻击性预测的效果，并进一步增强自动化文本编码的潜力。当然，模型仅基于少量对象构建这一情况也不容忽视。

表2。三分类机器学习模型的性能（%）。

预测目标	特征	模型	ACC	F1	AUC
社会排斥	基本 + 动态	LR	41.25	32.75	58.55
		SVM	42.19	35.95	55.72
		NN	40.94	34.88	56.40
	基本 + 动态 + 内容	LR	42.50	40.48	57.42
		SVM	40.31	37.90	54.36
		NN	42.81	40.12	58.36
		BERT	**62.50**	**53.55**	**67.69**
恶意幽默	基本 + 动态	LR	41.88	31.74	57.87
		SVM	42.19	35.43	56.35
		NN	37.81	38.90	53.94
	基本 + 动态 + 内容	LR	42.50	38.05	59.47
		SVM	43.13	41.41	58.37
		NN	43.44	39.97	56.98
		BERT	**56.25**	**52.02**	**62.43**
诱发内疚	基础 + 动态	LR	40.94	28.02	49.07
		SVM	44.38	32.83	52.93
		NN	37.81	36.55	51.22
	基础 + 动态 + 内容	LR	41.25	32.14	53.21
		SVM	41.56	34.09	53.55
		NN	39.69	35.74	53.64
		BERT	**59.38**	**53.76**	**64.53**

讨论与结论

本研究采用预训练 BERT 模型，通过社交媒体用户的活动准确预测人类网络攻击行为，并预测用户在三个行为维度上的表现：社会排斥、恶意幽默和诱发内疚。本研究对理论和实践均具有重要启示。从理论角度看，本文首先为网络攻击行为研究作出了贡献。除传统的自我报告外，本研究响应了通过深度学习算法开展网络攻击性预测研究的呼吁。其次，本研究为结合基础特征、动态特征和内容特征的 BERT 模型预测提供了明确的指导。具体而言，我们发现，不同特征在预测社会排斥、恶意幽默和诱发内疚方面发挥着不同作用。第三，本研究拓展了预测模型研究，指出 BERT 的表现优于缺乏外部预训练信息的传统机器学习模型（即逻辑回归、支持向量机和神经网络）。预训练模型确实显著提升了社交媒体用户的行为预测效果，尤其适用于心理学研究中样本量较小的情境。此外，本研究也拓展了社交媒体情境下的理论预测。从实践角度看，本研究初步展示了一种用于预测网络攻击行为的技术方案。该方案突破了传统自我报告问卷的局限，并为人格研究提供了启示。此外，社交媒体平台可以监测并识别用户的网络攻击行为，以避免用户实施有害行为。这有助于优化平台组织。

然而，本研究仍存在一定局限。首先，越来越多的研究者开始关注深度学习的可解释性，包括采用注意力机制的 BERT 系列模型。然而，我们研究的是 BERT 模型的构建和网络攻击性预测，而非用于行为预测的模型可解释性；后者将是一个很有前景的研究方向。其次，尽管本研究采用深度学习 BERT 进行人类行为预测，但编码行为的主要来源仍是用户的文本内容。未来，社交媒体用户的综合行为，即基本特征、动态行为和推文内容，可以作为统一的代理信息输入行为编码器。该行为编码器将在行为预测与深度学习之间搭建桥梁。

作者贡献

研究设计、研究构思、模型构建、数据分析及稿件撰写：ZKZ。研究设计、数据收集、数据分析及稿件撰写：MLY。模型构建及数据分析：YXH 和 XYP。批准提交的版本以供发表：ZKZ、MLY、YXH 和 XYP。

数据与代码可用性声明

匿名数据集和代码可在 GitHub 上获取。
https://github.com/kayzhou/When-Cyber-Aggression-Prediction-Meets-BERT-on-Social-Media。

参考资料

1. Brain P F, Susman E J. 攻击与暴力的激素因素。理解和预防暴力 1994; 2: 173-244. 2. Mishna F, Regehr C, Lacombe-Duncan A, 等. 大学校园中的社交媒体、网络攻击行为与学生心理健康。心理健康杂志 2018; 27(3): 222-229. 3. Mladenović M, Ošmjanski V, Stanković S V. 网络攻击行为、网络欺凌与网络诱骗：一项调查及研究挑战。*ACM*计算调查（*CSUR*）2021; 54(1): 1-42. 4. Forrest S, Eatough V, Shevlin M. 成人间接攻击行为的测量：间接攻击量表的开发与心理测量评估。攻击行为：国际攻击行为研究学会官方期刊 2005; 31(1): 84-97. 5. Meter D J, Ehrenreich S E, Underwood M K. 父母心理控制与父母及青少年社交攻击行为之间的关系。儿童与家庭研究杂志2019; 28(1): 140-151. 6. Bogolyubova O, Panicheva P, Tikhonov R, 等. Facebook上的黑暗人格：有害的在线行为与语言。人类行为中的计算机 2018; 78: 151- 159.

7. Pabian S, De Backer C J S, Vandebosch H. 黑暗三人格特质与青少年网络攻击行为。人格与个体差异 2015; 75: 41-46. 8. Sadiq S, Mehmood A, Ullah S, 等. 基于深度神经模型检测Twitter上的攻击行为。未来一代计算机系统 2021, 114: 120-129. 9. Sharif O, Hoque M M. 社交媒体中文本攻击性的识别与分类：资源创建与评估。紧急情况下区域语言在线敌意帖文治理国际研讨会。施普林格，查姆；2021: 9-20. 10. Zhou Z, Xu K, Zhao J. 外向者和内向者在微博上的发帖方式不同。《*EPJ*数据科学》 2018, 7: 1-22. 11. Yu M, Zhou Z. 预测社交媒体上内在与外在目标内容的追求。《网络心理学、行为与社交网络》 2022, 25(8): 540-545. 12. Chavan V S, Shylaja S S. 运用机器学习方法检测社交媒体网络中同龄人发布的网络攻击性评论。*2015*年计算、通信与信息学进展国际会议（*ICACCI*）。 IEEE; 2015: 2354-2358. 13. Al-Garadi M A, Hussain M R, Khan N, 等. 大数据时代使用机器学习算法预测社交媒体上的网络欺凌：文献综述与未决挑战。《*IEEE Access*》 2019; 7: 70701-70718. 14. LeCun Y, Bengio Y, Hinton G. 深度学习。《自然》 2015; 521(7553): 436-444. 15. Devlin J, Chang M W, Lee K, 等. BERT：用于语言理解的深度双向变换器预训练。*arXiv*预印本 *arXiv* 2018; 1810.04805. 16. De Choudhury M, Gamon M, Counts S, 等. 通过社交媒体预测抑郁症。第七届*AAAI*网络日志与社交媒体国际会议 2013. 17. Panicheva P, Mararitsa L, Sorokin S, 等. 预测高风险俄罗斯心理健康应用用户的主观幸福感。《*EPJ*数据科学》 2022; 11(1): 21. 18. Li S, Zhao Z, Hu R, 等 . 中文形态与语义关系的类比推理。*arXiv*预印本 *arXiv* 2018; 1805.06504. 19. Lerman K, Marin L G, Arora M, 等. 在线社交网络的语言、人口统计特征、情绪与结构。《计算社会科学杂志》 2018; 1(1): 209-225.

20. Li S、Wang Y、Xue J 等。新冠肺炎疫情宣告对心理后果的影响：对活跃微博用户的一项研究。《国际环境研究与公共卫生杂志》2020年；17(6)：2032。21 Zhao J、Dong L、Wu J 等。Moodlens：一种基于表情符号的中文推文情感分析系统。《第18届ACM SIGKDD知识发现与数据挖掘国际会议论文集》2012年；1528-1531。22. Sun Y、Wang S、Feng S 等。文心一言3.0：用于语言理解与生成的大规模知识增强预训练。《arXiv预印本 arXiv》2107；02137，2021。