在心理健康记录文本中识别疼痛提及：一种自然语言处理方法

Jaya Chaturvedi（精神病学、心理学与神经科学研究所，伦敦国王学院），Sumithra Velupillai（精神病学、心理学与神经科学研究所，伦敦国王学院），Robert Stewart（精神病学、心理学与神经科学研究所，伦敦国王学院，英国健康数据研究中心，南伦敦与莫兹利生物医学研究中心，伦敦，英国Angus Roberts（精神病学、心理学与神经科学研究所，伦敦国王学院，英国健康数据研究中心）

疼痛是患者寻求医疗资源的常见原因，也是一个日益受到关注的研究领域，尤其是疼痛与心理健康的交集。心理健康电子健康记录是研究这种交集的良好数据来源。然而，有关疼痛的大量信息都存储在这些记录的自由文本中；由于疼痛提及具有歧义性，因此对其进行自然语言处理是一项独特的挑战。本项目使用一个匿名化心理健康电子健康记录数据库中的数据，训练一种基于机器学习的分类算法，将句子分类为是否讨论患者疼痛。这样有助于从大型数据库中提取相关疼痛信息，并利用这些结果进一步研究疼痛与心理健康。研究人员对1,985份文档进行了人工三重标注，以创建金标准训练数据，并用这些数据训练了三种常用的分类算法。表现最佳的模型取得了0.98的F1 分数（95%置信区间：0.98–0.99）。

关键词。 自然语言处理、电子健康记录、疼痛、心理健康、Transformer。

1. 引言

疼痛被定义为一种令人不快的感觉和情绪体验，并受到多种生物、心理和社会因素的影响 [1]。疼痛是人们寻求医疗服务的常见原因，因此电子健康记录（EHR）可能成为疼痛相关信息的来源 [2]。

电子健康记录（EHR）是关于个人病史或医疗保健情况的纵向电子数据汇编[3]。由于它们为研究患者症状以及从结构化和非结构化字段中获取的发现提供了探索机会，因此在研究中的应用日益广泛。由于这些结构化字段对疼痛的记录并不充分，使用来自非结构化临床文本的数据补充这些信息可能会有所帮助 [4]。

一种常用的基于机器学习的自然语言处理方法是文本分类，即为文本单元（句子/段落/文档）分配标签 [5]。常用的分类算法包括支持向量机 [6–8] 和K近邻 [9–11]。近期最先进的方法采用嵌入模型和基于Transformer的神经网络架构 [12]，，例如BERT的双向编码器表示 [13]。认识到通用文本与生物医学文本之间存在语言差异，因此需要专门的模型；此后，出现了许多面向医疗领域的模型，例如PubMedBERT[14], BioBERT [15], ClinicalBERT [16], UmlsBERT [17] 和SAPBERT [18] [19]。

本文介绍了开发自然语言处理应用的方法，该应用用于对临床文本中提及躯体疼痛的内容进行句子级分类。我们训练了两个BERT模型——bert_base和SAPBERT，并将其与两个传统模型——支持向量机（SVM）和K近邻（KNN）进行了比较。据我们所知，此前尚未有人使用自然语言处理从精神健康临床文本中提取有关疼痛的信息。

2. 方法

2.1. 数据来源

南伦敦和莫兹利国家医疗服务体系基金会信托（SLaM）是欧洲规模最大的心理健康服务机构之一，其电子健康记录数据的匿名化版本存储在临床记录交互式检索

（CRIS）数据库 [20]。CRIS 的基础设施已有详细描述，并附有队列概况综述 [21]。CRIS 包含超过 3000 万份文档，平均每位患者有 90 份文档 [22]。共有 23 种不同的文本来源（如附件、事件记录、护士评估信等）。大多数文本包含在附件和事件记录中，因此本项目将它们用作数据来源。

2.2. 伦理与数据访问

CRIS 已获得伦理批准（牛津C研究伦理委员会，参考编号 18/SC/0372）。使用 CRIS 数据库的研究项目均由患者主导的监督委员会审查并批准（详见 [23]）。服务使用者可选择退出，相关退出机制会在所有宣传材料和活动中予以说明。数据由第三方莫兹利生物医学研究中心（BRC）所有，该中心负责运行 CRIS 工具，并提供匿名化数据的访问权限。这些数据只能由获准人员在安全防火墙内访问。

2.3. 数据提取

疼痛可以通过多种术语以多种方式描述。为了帮助识别 CRIS 中可能讨论疼痛的文档，我们综合了从文献和生物医学本体中提取的疼痛相关术语，并补充了词嵌入模型提供的其他相似术语，据此编制了一部疼痛术语词典。该词典及其编制过程已在 [24] 中作了更详细的介绍。我们使用这部疼痛词典中的术语，识别 CRIS 中可能讨论疼痛的文档。随后使用 SQL 提取包含疼痛术语的文档，以供进一步处理。提取时未应用时间或诊断筛选条件。

2.4. 标注任务

通过对提取的文档进行标记（即标注与疼痛相关或无关的文本片段），构建了一个讨论患者疼痛的文本语料库。每个片段包含疼痛相关术语前后各200个字符。首先，制定了一套标注指南，明确何种情况下应将一个句子视为讨论疼痛。接着，在提取的文档中标出疼痛词典中的术语。三名医学生标注员阅读提取的文档，并审阅包含先前标出疼痛术语的这些文本片段。标注员为每个片段标注以下三种标签之一：相关，即指患者经历的字面疼痛；不相关，即与疼痛无关、与患者无关的提及，或假设性及隐喻性提及；以及否定，即表示没有疼痛。在多轮标注后计算标注者间一致性，每轮均由三名标注员对200份文档进行标注。本研究使用的标注工具为 MedCAT [25]。对于任何分歧，标注员都会进行讨论，并据此更新标注指南。这个迭代过程持续进行，直至标注者间一致性超过0.80；随后，每名标注员获分配一组单独的文档。在迭代过程中由三名标注员共同标注的文档，则依据一套与最新版标注指南一致的裁定指南进行裁定。标注指南和裁定指南可在线查阅[1]。

2.5. 自然语言处理应用

这些标注分别按 80/10/10 的比例划分为训练集、测试集和验证集。如表 1 所示，共训练了四种不同的模型。参数根据文献 (21) 中的建议进行选择，并检查模型是否存在过拟合。

表 1. 模型规格

模型	分词器	预处理	其他参数
1. 支持向量机 2. K近邻邻居	NLTK	小写化、停用词，空白字符和标点符号去除，词形还原和分词	TF-IDF 向量化器默认参数来自 sklearn
3. BERT	bert_base_uncased	分词 在句子前添加特殊标记 [CLS] 和 在末尾添加特殊标记 [SEP]	训练轮数：3 批次大小：16 优化器：AdamW，学习率 3e-5
4. SAPBERT	cambridgeltl/SapBE		训练轮数：4

[1] https://GitHub.com/jayac haturved i/pain_in_mental_health/blob/main/ 标注%20Gu idelines%20-%20Pa in%20-%20for%20github.pdf

RT-from- Pad 填充并截断句子 批次大小：16 PubMedBERT- 全文最大长度设为 105（默认值为 511）优化器：AdamW，学习率 2e-5

3. 结果

3.1. 数据提取

从 723 名患者的文档中随机选取并提取了共 1,985 份包含词典中疼痛相关关键词的文档。这些患者最常见的诊断编码为心境障碍（ICD-10 第 F30-39 章）（占患者的 33%）。平均每位患者有 8 条标注。

3.2. 标注

经过四轮三重标注（每轮包含 200 份文档）后，标注者间一致性达到 90%（科恩卡帕系数为 0.88）。共获得 5,644 条标注。其中，72% 被标记为相关，15% 被标记为不相关，13% 被标记为否定。相关标注被标记为 1。不相关和否定标注合并后标记为 0。训练数据和测试数据中，71% 的标注被标记为 1 类（相关）。

3.3. 自然语言处理应用的评估

训练模型时使用了一块图形处理器（Tesla T4）。模型评估采用 K 折交叉验证，并计算了 95% 置信区间。各算法的结果见表 2。BERT模型的表现优于支持向量机和 K近邻模型。

表 2。评估指标，包括 95% 置信区间

模型	精确率	召回率	F1 分数（10 次的平均值，折叠交叉验证）
支持向量机	0.86（0.83-0.88）	0.98（0.97-0.99）	0.91（0.90-0.93）
K近邻	0.84（0.81-0.87）	0.91（0.89-0.93）	0.87（0.85-0.89）
BERT	0.96 (0.94-0.97)	0.98 (0.97-0.99)	0.97（0.96-0.98）
SAPBERT	0.98（0.97-0.99）	0.99（0.98-0.99）	0.98（0.98-0.99）

3.4. 错误分析

在标注过程中，常见的分歧包括某个实例可被解读为字面含义还是隐喻含义，例如“……使他感到疼痛”，以及假设性提及，例如“……她担心疼痛”和“？偏头痛”。

模型训练完成后，在测试数据的错误分析中发现了一些假阳性。对于 BERT¯base 模型，出现了诸如“……希望将自己的疼痛投射到他人身上”和“父亲的髋部疼痛”这样的情况。还发现了一些假阴性，例如“否认除胃痛以外的症状”和“……如果疼痛加剧”。

SAPBERT 模型在文本中包含无法辨认的符号时出现了假阴性。这些符号可能是在文本匿名化过程中产生的；此外，拼写错误或连写词也会导致此类情况，例如“dabdominal pain”和“achespainodd sensations”。假阳性则包括“有发生压疮的潜在风险”这样的情况。

4. 讨论

本项目凸显了疼痛概念的模糊性，尤其是在标注过程中，三位受过临床训练的标注员经过多轮讨论才对疼痛提及的含义和解读达成一致。考虑到这一点，分类模型难以处理假设性和隐喻性表述也就不足为奇。这凸显了语境的重要性，以及自然语言处理模型在分类任务中纳入并考量语境的必要性。这正是 BERT 等基于 Transformer 的模型的一项优势，也可能是它们表现优于支持向量机和 K近邻的原因。

在训练的两个BERT模型中，使用生物医学本体统一医学语言系统（UMLS）进行预训练的SAPBERT，表现略优于 bert¯base。本项目所用的两个BERT模型在分词方式上有所不同，其中SAPBERT能够更准确地对临床概念进行分词

。分词方面的这一改进可能影响并提升了模型的整体性能。

5. 结论

本项目旨在开发一种基于机器学习的自然语言处理应用，用于将临床文本中提及疼痛的内容分类为相关或不相关。采用基于Transformer的机器学习技术和上下文嵌入的BERT模型，表现优于其他算法。这是一种从心理健康记录中提取疼痛相关信息的新方法，借助非结构化临床笔记识别提及相关疼痛的患者。随后，这些患者队列可用于流行病学及其他疼痛相关研究，从而更有把握地判断文本中提及的疼痛是否实际存在。

致谢

本研究使用了患者提供并由英国国家医疗服务体系在提供护理和支持过程中收集的数据。作者还感谢Aurelie Mascio博士提供部分Python脚本的访问权限。

参考文献

1. 术语 | 国际疼痛研究协会 [互联网]。国际疼痛研究协会（IASP）。[引用于2022年10月31]。可从以下网址获取：https://www.iasp-pain.org/resources/terminology/ 2. Motov SM, Khan AN。急诊科疼痛管理中的问题与障碍：我们能否取得进展？疼痛研究杂志[互联网]。2008年12月9 [引用于2022年10月31]；2:5–11。可从以下网址获取：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3004630/ 3. 美国医学研究所患者安全委员会、Aspden P、Corrigan JM、Wolcott J、Erickson SM。《电子健康记录系统的关键能力：致函报告》[互联网]。《患者安全：实现新的护理标准》。美国国家科学院出版社；2004 [引用于2022年10月31]。可从以下网址获取：https://www.ncbi.nlm.nih.gov/books/NBK216105/ 4. Carlson LA、Jeffery MM、Fu S、He H、McCoy RG、Wang Y、Hooten WM、Sauver JS、Liu H、Fan J。两个医疗卫生系统临床文本中慢性疼痛发作的特征分析：全面标注与语料库分析。JMIR医学信息学[互联网]。2020年11月16 [引用于2022年8月22]；8(11):e18659。可从以下网址获取：https://medinform.jmir.org/2020/11/e18659 5. Minaee S、Kalchbrenner N、Cambria E、Nikzad N、Chenaghlu M、Gao J。基于深度学习的文本分类：综合综述。arXiv:200403705 [cs, stat] [互联网]。2021年1月4 [引用于2021年9月13]；可从以下网址获取：http://arxiv.org/abs/2004.03705 6. Wright A、McCoy AB、Henkin S、Kale A、Sittig DF。使用支持向量机对自由文本笔记进行分类：评估其在两个机构中的准确性。美国医学信息学会杂志[互联网]。2013年9月1 [引用于2022年11月30]；20(5):887–90。可从以下网址获取：https://doi.org/10.1136/amiajnl-2012-001576 7. Yao L、Mao C、Luo Y。结合基于规则的特征和知识引导的卷积神经网络进行临床文本分类。《BMC医学信息学与决策制定》[互联网]。2019年4月4 [引用于2022年11月30]；19(3):71。可从以下网址获取：https://doi.org/10.1186/s12911-019- 0781-4 8. Garla V、Taylor C、Brandt C。采用拉普拉斯支持向量机进行半监督临床文本分类：在癌症病例管理中的应用。生物医学信息学杂志[互联网]。2013年10月[引用于2022年11月30]；46(5):869–75。可从以下网址获取：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3806632/ 9. Trstenjak B、Mikac S、Donko D。基于词频-逆文档频率的文本分类K近邻框架。《Procedia工程》[互联网]。2014年1月1日[引用于2022年11月30]；69:1356–64。可从以下网址获取：https://www.sciencedirect.com/science/article/pii/S1877705814003750 10. Jindal R、Taneja S。医学文档文本分类的词汇方法。《Procedia计算机科学》[互联网]。2015年1月1日[引用于2022年11月30]；46:314–20。可从以下网址获取：https://www.sciencedirect.com/science/article/pii/S1877050915000903 11. Xing W、Bei Y。基于K近邻分类算法的医疗健康大数据分类。IEEE Access。2020；8:28808–19。12. Vaswani A、Shazeer N、Parmar N、Uszkoreit J、Jones L、Gomez AN、Kaiser L、Polosukhin I。Attention Is All You Need[互联网]。arXiv；2017[引用于2022年10月31]。可从以下网址获取：http://arxiv.org/abs/1706.03762 13. Devlin J、Chang MW、Lee K、Toutanova K。BERT：面向语言理解的深度双向Transformer预训练。载于：《2019年计算语言学协会北美分会人类语言技术会议论文集》第1卷（长篇与短篇论文）[互联网]。明尼阿波利斯，明尼苏达州：计算语言学协会；2019 [引用于2021年4月21]。第4171–86页。可从以下网址获取：https://www.aclweb.org/anthology/N19-1423

14. Gu Y、Tinn R、Cheng H、Lucas M、Usuyama N、Liu X、Naumann T、Gao J、Poon H。面向生物医学自然语言处理的领域特定语言模型预训练。ACM《计算医疗保健汇刊》[互联网]。2022年1月31 [引用于2022年10月31]；3(1):1–23。可从以下网址获取：http://arxiv.org/abs/2007.15779 15. Lee J、Yoon W、Kim S、Kim D、Kim S、So CH、Kang J。BioBERT：用于生物医学文本挖掘的预训练生物医学语言表征模型。《生物信息学》[互联网]。2019年9月10 [引用于2022年10月31]；btz682。可从以下网址获取：http://arxiv.org/abs/1901.08746 16. Huang K、Altosaar J、Ranganath R。ClinicalBERT：对临床病历建模并预测再入院[互联网]。arXiv；2020 [引用于2022年10月31]。可从以下网址获取：http://arxiv.org/abs/1904.05342 17. Michalopoulos G、Wang Y、Kaka H、Chen H、Wong A。UmlsBERT：利用统一医学语言系统元词库增强上下文嵌入的临床领域知识。arXiv:201010391 [cs] [互联网]。2021年4月12 [引用于2021年4月21]；可从以下网址获取：http://arxiv.org/abs/2010.10391 18. Liu F、Shareghi E、Meng Z、Basaldella M、Collier N。面向生物医学实体表征的自对齐预训练。arXiv:

201011784 [ ] cs 互

[ ] 2021 7 [2021 21] 2010 11784

联网 。4月 引用于5月 ； 可从以下网址获取：http://arxiv.org/abs/ 。19. Alsentzer E、Murphy J、Boag W、Weng WH、Jindi D、Naumann T、McDermott M。公开可用的临床BERT嵌入。载于：第二届临床自然语言处理研讨会论文集 [互联网]。明尼阿波利斯，明尼苏达州，美国：计算语言学协会；2019 [引用于2021年4月21]。第72–8页。可从以下网址获取：https://www.aclweb.org/anthology/W19-1909 20. Stewart R、Soremekun M、Perera G、Broadbent M、Callard F、Denis M、Hotopf M、Thornicroft G、Lovestone S。南伦敦和莫兹利国家医疗服务体系基金会信托生物医学研究中心（SLAM BRC）病例登记库：建立与描述性数据。《BMC精神病学》。2009年8月12日；9:51。21. Perera G、Broadbent M、Callard F、Chang CK、Downs J、Dutta R、Fernandes A、Hayes RD、Henderson M、Jackson R、Jewell A、Kadra G、Little R、Pritchard M、Shetty H、Tulloch A、Stewart R。南伦敦和莫兹利国家医疗服务体系基金会信托生物医学研究中心（SLaM BRC）病例登记库队列概况：现状及近期对源自电子精神健康病历的数据资源的增强。《BMJ开放版》。2016年3月1日；6(3):e008721。22. Velupillai S、Suominen H、Liakata M、Roberts A、Shah AD、Morley K、Osborn D、Hayes J、Stewart R、Downs J、Chapman W、Dutta R。利用临床自然语言处理开展健康结局研究：概述及未来发展的可行建议。《生物医学信息学杂志》[互联网]。2018年12月1 [引用于2021年3月31]；88:11–9。可从以下网址获取：https://www.sciencedirect.com/science/article/pii/S1532046418302016 23. Fernandes AC、Cloete D、Broadbent MTM、Hayes RD、Chang CK、Jackson RG、Roberts A、Tsang J、Soncul M、Liebscher J、Stewart R、Callard F。为源自精神健康电子病历的病例登记库制定并评估去标识化流程。《BMC医学信息学与决策制定》。2013年7月11日；13:71。24. Chaturvedi J、Mascio A、Velupillai SU、Roberts A。疼痛词典的开发。《数字健康前沿》[互联网]。2021 [引用于2021年12月14]；3:193。可从以下网址获取：https://www.frontiersin.org/article/10.3389/fdgth.2021.778305 25. Kraljevic Z、Searle T、Shek A、Roguski L、Noor K、Bean D、Mascio A、Zhu L、Folarin AA、Roberts A、Bendayan R、Richardson MP、Stewart R、Shah AD、Wong WK、Ibrahim Z、Teo JT、Dobson RJ。使用MedCAT（医学概念标注工具包）开展多领域临床自然语言处理。arXiv:201001165 [cs] [互联网]。2021年3月25 [引用于2022年2月26]；可从以下网址获取：http://arxiv.org/abs/2010.01165