（四）自然语言处理在医疗健康领域中应用

近年来，自然语言处理（NLP）正与现代医学深度融合，推动医疗智能从事后回溯迈向事前预判与实时干预。NLP 不仅能使机器理解病历中的细微情绪、识别医嘱中的潜在歧义、捕捉患者未言明的痛苦，而且还能从海量非结构化文本中挖掘风险模式与个性化预警信号。这不仅为解析语言背后的深层语义提供了有效方法，也深化了对疾病表达多样性的认知——同一病症在不同文化背景、年龄群体和心理状态下的语言表现各异，如今可被精准映射、系统归因并动态建模。NLP 已成为连接患者语言表达与医疗质量改进的重要桥梁。由此构建的智慧医疗系统，具备实时响应与持续进化能力，为医疗质量全景监测、个体化诊疗及公共卫生风险前瞻防控奠定坚实技术基础。下以宗教因素影响抑郁患者分析和心血管诊断错误识别两项研究为例来具体说明 NLP 的实际应用。

第一篇文章主要是利用智能体生成和自然语言处理技术分析宗教信仰对抑郁的影响。抑郁症是全球范围内导致残疾的主要原因之一，对患者心理健康造成深远影响。研究表明，宗教信仰对心理健康具有双重作用：一方面，宗教团体能提供社会支持和意义感，缓解抑郁症状；另一方面，部分宗教社区对抑郁症存在污名化，可能阻碍患者寻求帮助。在此背景下，社交媒体平台（如 Reddit）因其匿名性，成为个人自我监控和寻求信息的重要渠道，也为抑郁症研究提供了新的数据来源。本研究主要分析了四类人群的社交媒体帖子：抑郁基督徒、非抑郁基督徒、一般抑郁个体以及一般非抑郁个体。研究人员首先手工收集并标注了 93 篇来自基督教论坛的帖子，基于用户是否自我报告被医生诊断为抑郁症进行分类。然后，采用自然语言处理技术训练了一个分类器，该分类器在测试集上达到了 86%的 F1 评分，并用于预测未标注帖子的类别，最终从 255 篇帖子中识别出 205 篇抑郁相关文本。最后，利用 Text2Onto 工具自动生成了四类人群的本体论，并通过概念重叠度计算了本体间的相似性。研究结果显示，一般抑郁组与非抑郁组之间的本体相似度最高（约 0.37），而基督徒与非基督徒之间的相似度较低（约 0.18）。抑郁基督徒的帖子中，“抑郁”“焦虑”等概念更为突出，宗教术语如“上帝”“信仰”在两类基督徒中均占据重

要地位。与基于文献构建的本体相比，用户生成内容对抑郁症医学方面的关注较少。该研究首次将智能体生成应用于宗教群体的抑郁分析，为理解宗教信仰如何影响抑郁症表达提供了新视角，未来可整合进心理咨询平台，辅助抑郁症的早期识别与干预。

第二篇文章则将自然语言处理（NLP）和机器学习（ML）技术应用于识别患者安全事件（PSE）报告中心血管诊断错误。研究对象为美国中大西洋地区一家多医院医疗系统在 2016 年 1 月至 2021 年 8 月期间上报的 PSE 报告。为获得 PSE 报告中真实的心血管诊断错误标签，每份 PSE 报告均经人工审查，以识别临床医生依据当前心血管诊断错误定义所撰写的叙述性内容。包含心血管诊断错误相关叙述的 PSE 报告被标记为 1，否则标记为 0。研究采用四种二元分类机器学习模型，用于识别心血管诊断错误相关叙述，并挖掘 PSE 报告文本中的共性特征：逻辑回归、弹性网络、XGBoost 和深度神经网络。研究结果显示：XGBoost 在识别心血管诊断错误相关报告方面表现优于其他模型，在测试集上取得了优异的性能指标（AUROC = 0.914，特异度 = 0.982，阳性预测值（PPV）= 0.866，准确率 = 0.929，F1 分数 = 0.738，AUPRC = 0.783）。其中，“起搏器”一词在 PSE 报告中成为识别心血管诊断错误的重要信号词。进一步分析发现，在心血管诊断错误相关的 PSE 报告中，“起搏器”常与“需行 MRI 检查”等表述共现，提示起搏器患者接受 MRI 检查这一临床场景存在安全隐患。此外，“顺序”“心电图”“心脏”“胸部”及“导联”是识别心血管诊断错误事件的五个最重要文本特征——这些术语多用于描述心脏疾病、相关诊疗过程及潜在安全事件。研究结果表明，机器学习与自然语言处理技术具备从现有 PSE 数据中有效识别心血管诊断错误相关报告的可行性；但其泛化能力仍需在外部独立医疗系统中进一步验证。

（徐健编辑）

译文一：

基于智能体生成和自然语言处理的抑郁症分类分析

Nicolae Goga , Andrei Vasilăteanu, Ramona Cristina Popa, Alexandru-Filip Popovici, Ramona Popovici, Maria Goga, Diana Todea, Alexandra Eni

来源：Front Psychol.

时间：2026 年 4 月

链接：https://doi.org/doi: 10.3389/fpsyg.2026.1780802.

1. 引言

抑郁症是最常见的精神障碍之一，可能引发众多心理和身体问题，也是全球导致残疾的主要原因。在此背景下，研究表明，属于宗教团体对心理健康有益。宗教可以通过提供与共享共同价值观和信仰的社区的接触，以及通过它所传播的信息，作为一种应对机制。大量研究表明，宗教参与可以通过提供来自当地社区的意义和支持来减轻抑郁症状。

然而，在某些宗教社区中，可能存在与多种疾病相关的污名化，如抑郁症。缺乏信仰或参与被视为罪恶的行为，这些行为与社区规范不符，可能导致更强烈的抑郁感，并可能延迟寻求帮助。不幸的是，大多数抑郁症患者获得专业帮助的机会有限。从这个意义上说，一种选择是转向在线社区，在那里保持匿名，个人可以自我监控和寻求信息。其中一个平台是 Reddit，这是一个允许用户匿名发布话题的社交平台。近年来，该平台作为公共数据来源越来越受欢迎，部分原因是其独特的研究促进方式。

在该平台上，用户可以保持匿名，这对某些人来说可能是至关重要的方面。与传统的自我评估方法相比，社交平台的数据提供了对个人感受和思考的实时测量的可能性。近期研究表明，通过分析用户帖子，可能能够识别某些抑郁症的迹象或症状。在此背景下，利用本体比较作为调查工具的兴趣日益增长。本体论提供了一种组织由领域内关键概念及其关系组成的知识的方式，从而捕捉对分类目的相关的深层意义。基于本体论的分析特别适合研究心理构念，如抑郁症，因为它能够结构化地表示症状、情绪状态、认知模式和体验概念。抑郁症是一种复杂且多维的状态，通过

语言表现出来，如情绪、自我认知、社会关系和行为倾向的表达。本体论允许系统地提取、组织和分析这些语义元素。

以往的研究已证明基于本体的方法在心理健康应用中的有效性，如抑郁症诊断支持系统、抑郁症状的社交媒体分析以及基于语义的情感与情绪分析。这些方法表明本体论能够捕捉与心理状态相关的有意义语义模式，使其成为分析抑郁相关文本数据的合适工具。

近年来，研究人员利用本体论作为存储和共享信息的便捷方式。本体可以用多种知识表示语言编写，如 OWL/RDF，这些语言可以被计算机解析，同时保持研究人员的理解性。多项研究探讨了本体的自动生成，提出了不同的方法和工具。

大多数现有研究，如上述，依赖半自动生成过程，需要额外的人工审核，或难以跨领域应用，而非采用全自动且用户友好的方法。在我们的研究中，我们重点关注使用 Text2Onto 工具实现的智能体生成，该工具已在多种研究环境中成功应用。

此外，近期研究重点是将自然语言处理（NLP）技术应用于更广泛的任务，文本处理和理解可作为自动诊断、疾病预测或潜在危险行为识别的工具。

本文扩展了这一研究方向，将这一首批方法应用于特定基督徒群体内抑郁症的分类。我们研究的另一个方面是利用智能体生成，利用社交媒体帖子语料库，提取并比较抑郁基督徒、非抑郁基督徒以及不一定是基督徒的抑郁个体和非抑郁个体（不一定是基督徒）所写文本的概念，采用定性和定量方法。

尽管有大量关于普通人群社交媒体使用的研究，并得出了关于抑郁症个体社会支持的若干结论，但几乎没有研究聚焦于具有特定特征的群体，这些特质可能影响抑郁的体验。我们研究的目标之一是分析抑郁在社交网络互动中的表现，特别关注宗教群体。

由于本体论难以创建，更可行的方法是从非结构化文本中自动生成本体。迄今为止，自动化本体生成尚未应用于抑郁症宗教人士的文本，也未对宗教抑郁者与非宗教抑郁症者文本中提取的概念进行比较分析。由于本体生成需要大量文本，另一个支持目标是创建一个工具，将帖子归类为抑郁组或对照组。

研究表明，区分抑郁症用户的一种方法是基于他们明确表示自己已被医生诊断的文本。由于并非所有基督教用户都提供了此类明确陈述，因此开发了一个基于自然语言处理（NLP）的人工智能引擎来分类这些文本。

Engine 培训基于论坛帖子，用户明确表示自己被医生诊断为抑郁症。培训后，该引擎被应用于论坛帖子中用户未提及诊断内容，以区分基督徒和非基督徒抑郁症。

所得的分类文本被用来利用现有的智能体生成引擎（Text2Onto）生成独立的本体论。随后，所得本体与文献中关于抑郁症患者的现有本体进行分析和比较。比较进行了定性和定量的比较，定量比较采用了相似度比。

2. 材料与方法

2.1. 文本选择与手工分类

我们的研究旨在分析社交网络帖子在四类用户中的差异（抑郁基督徒、普通基督徒、不一定是基督徒的普通抑郁个体、不一定是基督徒的非抑郁个体）。首先，我们基于在 Reddit 及其子版块上的搜索，以及专注于抑郁相关话题的公共论坛和社交媒体群组，创建了文本语料库。所选文本包含了直接的第一人称对抑郁症的引用。用户自认属于基督教教派（东正教、天主教、新教或新新教）。

选择文本主要基于两个标准：用户报告由医疗专业人员诊断为抑郁症，且属于基督教教派（正统、天主教、新教、新新教）;或一般认同为基督徒（例如：“我是基督徒，被诊断为抑郁症”，我是天主教/东正教/新教徒，被诊断为抑郁症“）。此外，用户必须年满 18 岁并在论坛上拥有活跃账户。报告其他精神或身体疾病（如精神分裂症、PTSD、强迫症和癌症）的参与者被排除在研究之外。子版块和其他论坛的讨论主题主要集中在基督教和抑郁症上。文本搜索使用了关键词，如*抑郁*症、抑*郁症*、*诊断抑郁*症，以及*我被诊断为抑郁*症。

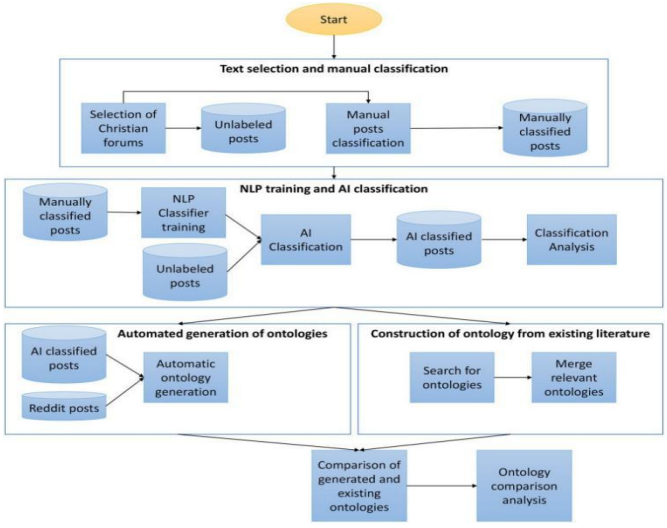

图 1 研究工作流程

研究团队心理学家共挑选并手动标注了 93 篇帖子。其中 36 篇帖子根据文献支持的标准和明确自我报告接受医疗专业人员诊断抑郁症，被归类为患有抑郁症的基督徒。其余 57 篇帖子为对照组（普通基督徒），未包含自我报告抑郁症诊断。这些帖子平均字数为 313 字，手工收集自与不同基督教教派（天主教、东正教和新教）相关的公开论坛。图 1 展示了研究工作流程。

用于生成各自本体论的两类——一般抑郁个体（不一定是基督徒）和一般非抑郁个体（不一定是基督徒）——的文本来自 Reddit 帖子。 这些数据通过美国乔治城大学与罗马尼亚布加勒斯特理工大学签署的协议获得。抑郁个体帖子总词数约为 195,538,000 字，而一般非抑郁个体的总词数约为 7.33 亿 2 万字。由于生成本体论的工具存在一些大小限制，我们需要缩小文本大小（更多细节见本体生成部分）。该网站（见上文）报告称，这些数据已被用于其他多项科学研究。

2.2. 自然语言处理训练与人工智能分类

如引言所述，为了分析文本并生成本体论，我们创建了一个 NLP 分类器，用于分类基督教论坛中的文本。该分类器基于我们团队心理学家手工标记的文本训练，这些文本由抑郁或非抑郁个体作者，基于用户是否报告被医生诊断。数据集包含 93 篇帖子，其中 36 篇属于抑郁组，57 篇属于对照组。

在预处理过程中，我们使用 Python 的正则表达式去除含数字或未知字符的单词，并剔除常见的英语停卡词。帖子中的每个单词随后通过 NLTK 软件包中的 WordNetLemmatizer 进行词汇化。帖子集随后用 tf-idf 向量化，因为机器学习算法需要数字输入。数据集被随机分为训练集和测试集，比例为 80–20。

首先，我们利用交叉验证和 F1 作为评分标准，评估了训练数据上的多个基线分类器。结果见表 1。

表 1 基线分类器分数

分类器	F1 评分	SD
带有 SGD 训练的线性分类器（SGD 分类器）	85%	0.11
随机森林分类器	82%	0.09
逻辑回归分类器	61%	0.18
多项模型的朴素贝叶斯分类器	12%	0.15

F1 评分被选为主要评估指标，因为它提供了结合准确性和回忆性的平衡衡量标准。该指标特别适用于类别不平衡的数据集，单靠准确性可能产生误导性结果。在抑郁症检测的背景下，精度（避免假阳性）和回忆（正确识别抑郁病例）都很重要，使 F1 评分成为一个合适且广泛使用的绩效指标。

使用 SGD 分类器获得最佳结果。接下来，我们使用随机搜索选择了 SGD 的最佳超参数。我们设置了 2000 次迭代，分别选择损失、惩罚、正则化率、学习率、班级权重和初始学习率。基于 F1 评分的最佳得分为 0.94，采用以下超参数获得：

- - L2 惩罚（标准 SVM 正则化器）
- - 铰链损耗（线性 SVM）
- - 学习率的缩放。
- - 预计学习率（ETA0）为 1
- - 班级体重 60–40
- - 学习率为 0.01

随后，使用这些超参数的分类器在测试数据上进行了测试，结果见表 2。

表 2 分类结果

分类器	精度	召回	F1 评分
SGD	83%	93%	86%

随后，训练好的分类器被用来预测基督教论坛中未标记帖子的标签。在共 255 条帖子中，有 205 条被归类为抑郁症患者。

2.3. 本体的自动生成

本研究的目的是利用自动化方法构建、分析并比较四个关注领域的本体论：抑郁基督徒、非抑郁基督徒、一般非抑郁个体（不一定是基督徒）和一般抑郁个体（不一定是基督徒）。

我们本体论中使用的概念生成分为三个步骤：

第一步：首先，如上所述构建文本语料库。

步骤 2：第二，为了从语料库中提取相关概念，我们使用了 Text2Onto 工具。这种自动化方法比手动提取文本概念更省时。Text2Onto 的另一个优势是能够为从语料库中提取的每个概念分配相关性评分。该工具能够从大量非结构化文本语料中提取概念。

在使用不同语料库测试概念提取过程时，我们发现了 Text2Onto 的几个局限性。最重要的限制与语料库大小有关：该工具最多接受 90,000 个字符，约为 15,000 个单词。超过这个数的语料库无法被 Text2Onto 处理。因此，对于每个领域，我们必须确保语料库大小低于这一限制。

对于“一般”案例，我们能够构建比“基督教”案例更大的语料库。由于这两个较大的语料库超过了容量限制，我们将它们划分为更小的子语料库，并在每个子部分运行工具。然后将结果合并，创建了每个领域的统一本体论（如步骤 3 所述）。

通过运行 Text2Onto 工具，我们自动生成了每个感兴趣领域的概念列表。在我们的研究中，我们只考虑了相关性评分最高的概念。相关性是一个正值，范围从 0 到 1，Text2Onto 会根据该分数自动排序生成的概念。在我们的本体中，我们只包含了相关性分数高于动态确定阈值的概念。该阈值是根据语料库的大小和具体的本体领域计算的。为了评估阈值的影响，我们构建并分析了使用不同概念限制的领域本体：100、250 和 500。

第三步：为了构建“一般案例”的最终本体，需要一个解析器。我们在 Java 中实现了一个解析器，处理在第二步生成的子语料库中运行 Text2Onto 时获得的概念列表。解析器比较这些列表中的概念，对于每个相同的概念，计算出最终的相关性得分，作为每个列表相关性分数的总和。

可以分析每个本体所用概念的示例（见图 2）。

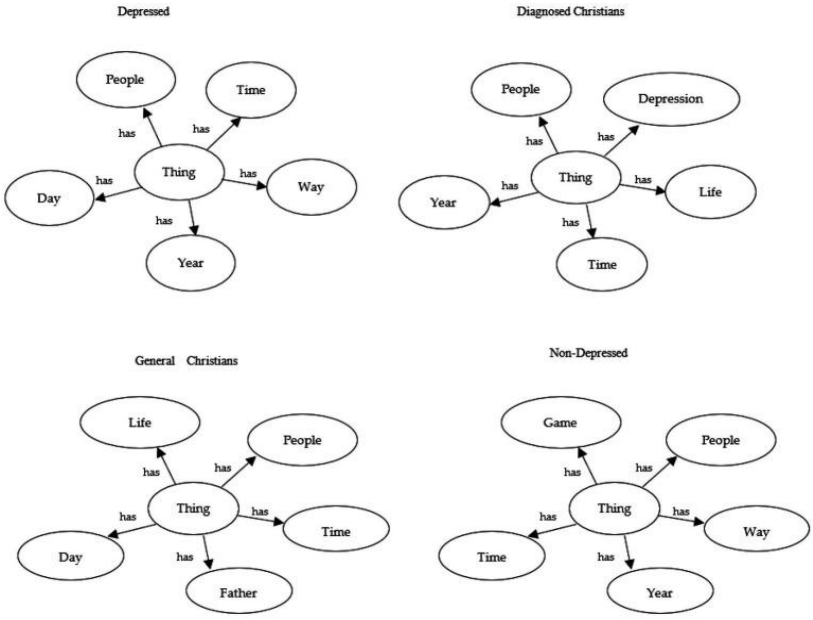

图 2 生成的本体论示例主要概念

2.4. 从现有文献构建本体论

除了生成的四类本体论外，我们还构建了一个基于文献的本体论，如下描述（见图 3）。我们使用以下数据库对科学文献进行了系统检索：Web of Science、Scopus 和 Google Scholar。用于搜索的字符串搜索是“本体论”和“抑郁*”。识别出 5 篇 2015 年至 2019 年间发表的相关同行评审文章。其中三篇本体论描述了青少年抑郁症，其余两篇聚焦于普通人群的抑郁症。其中两种本体旨在为诊断或协助治疗抑郁症的应用提供框架，而另外三个本体则旨在为分析社交媒体数据提供语义基础。

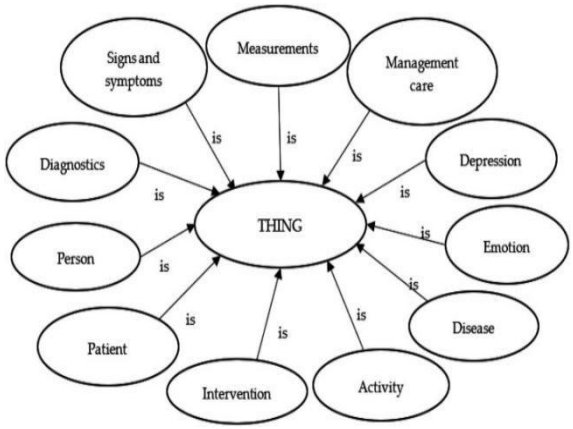

图 3 文献研究中的本体论根本概念

最终，五个本体论中只有四个被认为适合本研究目的，且仅这些本体被合并。一个本体论被排除，因为它是另一项研究的延续，且包含若干与框架不符的术语。所选本体通过 Protégé Web 本体软件合并。

我们结合了四个本体论中的主要概念，如抑郁症的风险因素、症状和干预，以及其他概念，如测量、电子邮件档案和在线发布的内容。由于每个本体论使用不同的词来分类这些类别，我们在将本体导入 Protégé Web Ontology 后识别出同义类。最终合并的本体包括 21 个类、43 个子类和 58 个体。

3. 结果

为了验证生成本体的相似性，首先我们按照公式计算它们的相似性：

相似度（Ont1， Ont2） = Nr-com-项/（Nr-项-Ont1 + Nr-项-Ont2）。

本体相似度可以通过多个互补组成部分来评估，如概念相似度、层级相似性和关系相似度。概念重叠是本体比较中一个基本且广泛使用的衡量标准。例如，Maedche 和 Staab（2002）部分基于概念集之间的重叠来定义本体相似性，而 Dellschaft 和

Staab（2006）则通过基于共享词的精确度和回忆度量比较概念来评估自动学习的本体，并将这种比较扩展到基于概念匹配的分类层级。

由于本体是通过 Text2Onto 自动生成的，该方法可靠地提取领域相关概念，但产生有限且不一致的关系结构，基于概念的相似度提供了最稳健且可比的本体相似度衡量标准。这种方法允许对从分析语料库中提取的语义词汇进行一致比较。

我们将这些公式应用于本体论，并根据给定的相关性对概念进行了排序。由于生成的本体包含数百项项,我们决定将该公式应用于本体论的多个维度,分别分为 800、500 和 200。结果见表 3–5。

表 3　本体论间的相似性限制在 800 个概念内

	非抑郁	**被确诊的基督徒**	**普通基督徒**	**抑郁**
非抑郁个体	1.000	0.180	0.186	0.373
被确诊的基督徒	0.180	1.000	0.185	0.212
普通基督徒	0.186	0.185	1.000	0.206
抑郁个体	0.373	0.212	0.206	1.000

表 4 本体论之间的相似性限制在 500 个概念内

	非抑郁	**被确诊的基督徒**	**普通基督徒**	**抑郁**
非抑郁个体	1.000	0.185	0.184	0.364
被确诊的基督徒	0.185	1.000	0.186	0.218
普通基督徒	0.184	0.186	1.000	0.215
抑郁个体	0.364	0.218	0.215	1.000

表 5 本体论间的相似性限制在 200 个概念内

	非抑郁	**被确诊的基督徒**	**普通基督徒**	**抑郁**
非抑郁个体	1.000	0.172	0.187	0.375
被确诊的基督徒	0.172	1.000	0.225	0.222
普通基督徒	0.187	0.225	1.000	0.222
抑郁个体	0.375	0.222	0.222	1.000

可以观察到，在不同长度的相似度之间没有相关差异。因此，我们可以得出结论，报告的相似度测量是稳健的。从表格中，我们观察到，一般抑郁组与非抑郁组个体之间的相似度最高，约为 0.37;而基督徒与非基督徒（抑郁和非抑郁）之间的相

似度最低，非抑郁个体约为 0.18，一般抑郁个体为 0.21。在这两类基督徒之间，相似度约为 0.18。

两类非基督徒之间的最高相似度（0.37）可能表明普通人群共享悲伤、压力和抑郁的共同经历。相比之下，宗教人士之间的相似度较低，这可能反映了他们信仰的影响，可能导致两类基督教群体在抑郁体验上存在差异。

在下一节中，我们将比较生成本体论中最相关的术语与基于文献构建的术语。

在一般抑郁和非抑郁个体中，尽管它们属于不同的类别，但前 20 个概念中的许多相似，包括“人”、“游戏”、“年份”、“屎”和“某人”等词汇。从这些观察中可以得出几个结论：游戏对两组人都很重要，反映了它们在当代生活中的相关性;关系（例如，“人”和“某人”）很重要;时间上的称呼（例如，“年”和“天”）可能反映了日常压力源;脏话（例如，“屎”）的使用相对常见。

在为抑郁者生成的本体论和基于文学构建的本体论之间，相似的概念并不多。在基于文献的本体论中，抑郁、情绪和诊断等概念高度相关，而在文本生成的本体论中则较不突出。这种差异可能归因于写帖的个体可能抑郁程度较低（因为他们有精力和动力在线参与），且对抑郁症医学方面关注度可能低于专家预期。相似性分析和对最相关概念的分析都表明，抑郁和非抑郁用户产生的内容共享许多共同概念，但与专家观点的一致性较小。

关于基督徒的两类，宗教术语在最相关的概念中占有显著地位，如“上帝”、“信仰”、“基督徒”和“教会”，反映了他们的宗教信仰。在抑郁的基督徒中，“抑郁”和“焦虑”等概念比非抑郁的基督徒更为突出，表明抑郁无论宗教信仰如何都会影响个体。对于这两个基督教类别，社会关系（如“人”）和时间指称（如“年份”）也是重要概念。相比之下，与咒骂或游戏相关的词汇较少，表明这些并非该群体的主要关注点。

两位心理学家审查了该分类器对抑郁症和非抑郁症基督徒的分类，并验证了其结果。在被归类为抑郁基督徒的帖子中，发现了多种抑郁相关症状，如悲伤感、决策受损、身体症状以及感知到需要帮助。所选帖子还强调了识别该病症并寻求专业支持的重要性。

有些帖子讨论抑郁的潜在原因，涉及心理、情感或身体因素，并描述以麻木或无价值感为特征的个人经历。在此背景下，用户还反思了自己在心理健康挑战上的经历、药物的作用、改变不良思维模式的重要性，以及信仰在康复过程中的作用。

此外，一些帖子还回应了基督徒不经历抑郁症的误解，以及抑郁症患者可能从宗教群体中收到的有害言论，例如被告知抑郁是由于缺乏信仰所致。帖子还反映了抑郁症是软弱表现的观点，导致对受影响个体的评判和污名化。心理学家在将帖子归类为抑郁类别时考虑了这些特征。

那些不抑郁的基督徒，报告过悲伤或孤独感，通常与抑郁相关的情绪，甚至提到最近经历过抑郁症状（但未确诊），他们表现出一种独特的话语模式，将痛苦视为人类经历的一部分。尽管存在孤立感，但仍有一种神圣的存在感，减轻了这些情绪的强度。

在这种语境下，抑郁或痛苦常被理解为与个人选择相关，个体将自己定位为能够影响其情绪状态的主动主体，这种视角与临床抑郁症通常相关的无助感形成鲜明对比。此外，即使在描述难以忍受的痛苦时，这些人也明确表示自杀从未被视为选项。

心理学家在将帖子归类为非抑郁类别时，考虑了这些特征。

4. 讨论与结论

本研究的目的是利用自动化方法构建、分析和比较四类本体论，分别指抑郁的基督徒、非抑郁的基督徒、一般非抑郁个体（不一定是基督徒）和一般抑郁个体（不一定是基督徒）。据我们所知，这是首个在宗教群体中探讨该主题的研究。

结果显示，这四个类别之间的相似性无显著差异。研究的另一个目标是对社交媒体帖子语料库应用智能体生成，并对四类帖子进行定性和定量比较。关于抑郁患者发布帖子中出现频率最高的词汇，我们的发现与 Mustafa 等人（2020）报告的一致，后者发现了相似的关键词权重用于抑郁分类。这种趋同支持了我们结果的潜在普遍性。

关于被诊断为抑郁症的宗教人士与一般抑郁群体之间的差异，基于用词的分类表明宗教维度对该宗教群体尤为突出。这一发现与先前研究探讨宗教作为应对机制作用的研究一致，暗示宗教信仰可能成为管理抑郁症的资源。

研究的一个局限是样本量较小，尤其是从基督教参与者中选取的文本，这些文本来自有限的帖子数量，而非基督徒抑郁症患者则如此。另一个局限是 NLP 分类器的推广性。该分类器是在 Reddit 和基督教论坛帖子上训练的，这些帖子共享类似特征，如非正式语言、个人叙事和心理健康讨论。然而，语言模式、用户人口统计和沟通风格可能因平台或人群而异。因此，分类器的表现可能无法完全泛化到不同语境，

如临床文本或其他社交媒体平台。未来研究应利用更多样化的数据集验证和完善分类器，以提高其泛化性。

另一个方法论上的限制是我们自己使用的分类器进行文本到本体的分析。尽管该工具在识别概念间关系方面存在局限性，但本研究的比较主要基于提取的概念，而非它们之间的联系。因此，这一限制不太可能显著影响结果。

作为未来发展方向，此类系统可整合进咨询平台，促进抑郁症的早期发现，因为早期诊断有助于及时实施有效干预策略。

译文二：

将机器学习和自然语言处理应用于患者安全事件报告：识别心血管诊断错误的模式

Azade Tabaie, Alberta K. Tran, Codrin Parau, Sonita S. Bennett, Sadaf Kazi,Kelly Smith, John Yosaitis, Kristen E. Miller

来源：PLoS One.

时间：2026 年 4 月

链接：https://doi.org/10.1371/journal.pone.0345693.

1. 简介

心血管疾病是美国（U.S.）发病率和死亡率的主要原因。心血管疾病的症状可能很微妙，表现形式多样，尤其在女性和老年人群体中，从而增加了心血管诊断错误的风险。诊断错误的识别，包括心血管诊断错误，是安全监测和质量改进的重要组成部分。美国国家科学院、工程院和医学院（NASEM）建议医疗机构“监控诊断过程，及时识别、学习并减少诊断错误和近距离事故”。然而，许多旨在检测诊断错误的安全项目目前依赖于对自由文本叙述性资料的人工审查——如患者安全事件（PSE）报告或病例回顾——这既劳动密集又耗时。

多项研究发现机器学习（ML）技术在预测诊断错误方面有效，包括延迟和漏诊。例如，Bhasuran 等人利用机器学习模型结合电子健康记录（EHR）数据，识别出患有罕见但可治疗疾病的未诊断患者。另一项研究中，机器学习用于分类通过会诊联络精神病学诊断的患者中谵妄诊断准确与误诊的案例。此外，机器学习模型在识别延迟癌症诊断标志物方面具有广泛应用。然而，尽管机器学习模型已被应用于心血管疾病检测，据我们所知，机器学习技术尚未被评估用于识别心血管诊断错误风险较高的患者。虽然识别心血管疾病风险是患者护理计划和临床轨迹的重要组成部分，但进一步了解心血管诊断错误的成因有助于制定措施和干预措施，以避免这些漏诊或延迟诊断可能带来的严重患者健康后果。

许多医疗机构拥有使用 PSE 报告的安全系统，允许医疗提供者匿名提交关于临床环境中的近距离事故或安全问题的自由文本描述，以增强安全监测和学习效果。

自然语言处理（NLP）和机器学习方法已被越来越成功地用于将 PSE 中丰富的信息分析为易于理解的类别，以激励解决方案。在一项最新研究中，NLP 和机器学习技术被应用于 PSE 报告系统，将记录的用药错误分类为更常用的类别（例如，错误的药物、错误的时间、错误的强度或浓度等）。在另一项研究中，Chen 等人研究了机器学习模型在自动分类不同事件类型（如护理协调或沟通、实验室检测、药物相关等）方面的有效性，利用母体接触和母婴单元的 PSE 报告。Fong 等人将机器学习技术应用于由十家医院组成的 PSE 报告健康系统中的自由文本格式数据，以识别与健康信息技术相关的安全事件。据我们所知，尚无已发表的研究将机器学习技术和 PSE 报告结合用于识别心血管诊断错误。然而，卫生系统提升诊断能力及其他质量和安全问题的努力通常包括对 PSE 及其他类似安全报告进行逐案手动审查和分析。例如，由医疗研究与质量局开发的公开资源 Measure Dx，旨在为医疗机构识别、分析和学习诊断安全事件提供指导，发现其项目中的大多数组织重新审查了已识别安全事件的现有质量与安全数据;然而，样本中超过一半的组织采用了电子健康记录增强的病历审查策略，以识别高风险诊断或暗示错失机会的护理模式，和/或加强人工病历审查策略。因此，进一步研究先进方法以帮助筛选 PSE 报告及其他类似现有数据，有助于医疗机构更高效地识别和响应心血管疾病及其他疾病的诊断错误，提升护理服务和质量。

虽然 PSE 报告是丰富且在大多数医疗系统中常用的数据源，但缺乏实际一致的诊断错误标签限制了机器学习和预测模型在识别经历诊断错误患者时的应用，尤其是在特定疾病方面。例如，以下 PSE 报告叙述被认定为对延迟心血管诊断事件的解释。然而，要将它与心血管诊断错误联系起来，需要手动阅读每份 PSE 报告：

“患者总是被送回去时没有 MRI 的遥测监测仪，如果患者有心律失常，由于患者往返运输延迟，长时间未接到监护仪，也不会被察觉。”

本研究中，我们与临床医生、研究人员和诊断安全专家团队合作，对大量 PSE 报告进行了注释，并识别出带有心血管诊断错误叙述的报告。然后使用带有真实标签（即心血管诊断错误与非心血管诊断错误）的注释数据集，我们训练了二进制机器学习模型，以分类包含相似叙述的 PSE 报告。研究了有心血管诊断错误患者与否患者人口统计特征之间的统计学显著差异。检测到 PSE 报告中与心血管诊断错误相关的重要术语。虽然我们采用了成熟的自然语言处理和机器学习方法，但我们的贡

献在于将这些技术新颖应用于检测 PSE 报告系统数据中的疾病特异性诊断错误，而 PSE 报告系统数据是传统上未被用于诊断安全监测的数据源。这种转化方法旨在连接数据科学的方法论进展与患者安全和质量提升的运营需求。

2. 材料与方法

2.1 数据来源

数据来源为 2016 年 1 月至 2021 年 8 月间，美国中大西洋地区一家多医院医疗系统根据一般事件类型诊断/*治疗*或*诊断影像*分类的 PSE 报告。一般事件类型是 PSE 报告的一部分，用于识别临床环境中近距离事故或安全问题的总体主题（如跌倒、血液制品、医疗信息技术等）。我们重点关注包含一般事件类型*诊断/治疗或可能*包含心血管诊断错误信息的 PSE 报告。在人工复查前，删除了空缺或简短事实描述不足的报告（例如“右肱二头肌区域浸润 100cc omni 350”）。还从每份 PSE 报告中提取了病历编号（MRN）、出生日期（DOB）、严重程度级别、事件日期、简短事实描述及报告的促成因素，以进行进一步的描述分析。患者的人口统计信息通常不在 PSE 报告中被记录。因此，PSE 报告通过 MRN 和出生日期与电子健康记录数据进行交叉比对，以捕捉患者人口统计学，如性别、种族、族裔和主要语言。

2.2 心血管诊断错误标签

我们根据美国国家科学研究院（NASEM）《*改善医疗诊断*报告》中的诊断错误定义，将心血管诊断错误定义为心血管相关疾病的诊断/治疗延迟或误诊。为确认监督机器学习算法学习数据模式的真实性（即真实的心血管诊断错误标签），我们对 PSE 报告中的自由文本描述进行了注释，并手动审查了 7,467 份报告，以识别心血管诊断错误的“真实”迹象。

PSE 报告/自由文本数据集的数据注释通常涉及对自由文本数据的人工审查，以调查特定叙述的存在。例如，本研究中，数据注释旨在检测 PSE 报告自由文本部分的诊断错误叙述。为获得 PSE 报告的真实心血管诊断错误标签，一位专注于人因工程且熟悉 PSE 报告的研究助理根据 NASEM 报告的关键定义手动审查了本研究中包含的 7,467 份 PSE 报告。本研究将心血管事件定义为任何可能对心肌或周围血管结构造成损害的事件，这一定义源自一般医学文献中的心脏和心血管事件。在临床专家团队的指导下，研究助理审查了所有被归类为诊断相关事件的 PSE，判断其是否 1）

符合 NASEM 的诊断错误定义，2）是否包含与心血管疾病相关的关键词或过程。在 PSE 报告中被识别为心血管诊断错误的关键词和短语包括但不限于胸部、深静脉血栓、心脏超声、双相压和起搏器。任何需要更深入临床审查或理解的报告均被标记为“不明确”，由两位临床主题专家审查。例如，深静脉血栓和肺栓塞等血管问题最初未被归类为心血管事件，但在与临床团队协商后，有意纳入我们的审查，以符合临床医生对潜在心血管后果的临床框架和概念化。表 1 总结了在 PSE 报告人工审查中包含的诊断错误类别。其他类别如“良好捕获”（即可能造成重大伤害但及时避免的事件）和“治疗延迟”（即与治疗相关而非诊断错误的事件）被纳入 ML 模型训练，但本研究未将其视为诊断错误。

表 1 心血管诊断错误的类别

心血管诊断错误的类别	定义	示例
漏诊	当某个诊断被忽视或未被考虑时。	患者因下消化道出血、贫血和粪便血从急诊入院。患者应被安排非心脏心电望，但未提供任何相关指示。患者被分配到内外科病床并被送至病房。抵达不久后，下达了将患者转入内科医师科的命令。这是因为患者有广泛的心脏病史，急诊室出现心动过速。
延迟诊断	当诊断被确认但未及时沟通或采取行动时。	会诊确认患者仅需静脉通路，实际上不需要 PICC 导管。医生称患者前一天失去了静脉通路，决定等待介入放射科放置 PICC 导管。令人担忧的是，这名患者需要 ICU 监护和护理，却长时间没有静脉通路。这是不合适的 PICC 静脉导管会诊。
错误诊断	当错误诊断时，	我们的数据中没有任何例子。
未识别并发症	当诊断正确，但未发现并发症，导致患者#39;病情恶化时。	A 医生正准备和外科住院医师一起给病人翻转。血压从 90 多/60 多开始，然后是 61/27。翻转时血压是 120/74。A 医生断开麻醉机，切换到便携式监护仪。B 医生进来，看到病人#39;心率在 80 多，考虑到患者#39，心率算是心动过缓;心率在 140 多，随后降到 100 和十几。他摸不到脉搏，但看到电活动，要求开始胸部按压，并给予 1 毫克肾上腺素。共进行了 3 次胸部按压和 1 毫克肾上腺素，脉搏恢复，患者稳定下来。
未诊断出无关疾病	当临床医生只关注初级诊断而漏诊无关疾病时。	患者当天早些时候从 CVICU 返回。患者按 EHR 指示转院。注册护士发现血糖偏低，转院报告中未提及。注册护士致电 CVICU 注册护士询问。CVICU 注册护士表示给患者喝了苹果汁。此事件发生在班次变更时间附近。

注：在识别带有此类叙述的 PSE 报告时，纳入了不同的心血管诊断错误，包含去标识的 PSE 报告示例。

包含表 1 定义中心血管诊断错误叙述的 PSE 报告被标记为 1（否则标记为零）。图 1 总结了我们识别和标记显示心血管诊断错误的 PSE 报告的方法。

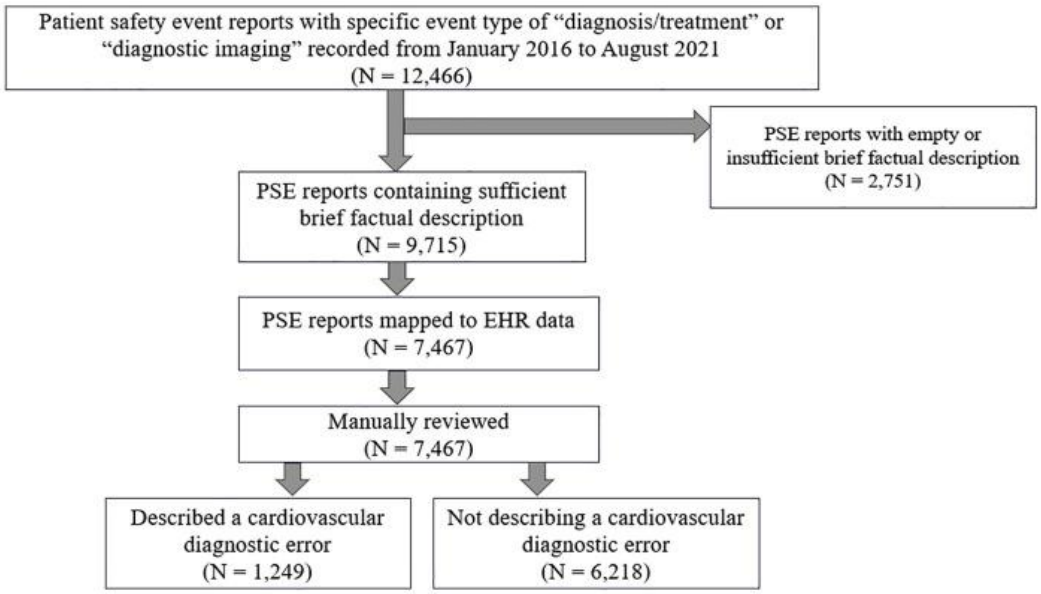

图 1　定量数据分析方法的图形表示，用于捕捉患者安全事件中的心血管诊断错误

为评估人工注释过程的检测者间信度（IRR），第二位具备临床数据注释经验的评审者对随机 1%样本（n = 75）进行了独立注释，该评审者对第一位评审者分配的原始标签进行了盲注。两位评审者之间的内部收益率为 90%。在抽样报告中，第一位评审者标记为描述心血管诊断错误的报告，第二评审者标记全部 10 份相同。第二位评审者发现另外两份报告最初标注为非心血管诊断错误，描述心血管诊断错误;一份报告描述了心脏外科团队间的沟通漏洞导致手术延迟，另一份报告描述了 BiPAP 机器尺寸不匹配。这些发现表明评审者之间的高度一致性，支持注释过程的有效性。

2.3 统计检验

采用 Wilcoxon 秩和检验（数值特征）和卡方检验（分类特征）评估有无心血管诊断错误患者组间患者特征（如年龄、性别、种族）的统计学显著差异。

2.4 机器学习（ML）模型

2.4.1 从自由文本数据中提取特征

临床医生提交的 PSE 报告的非结构化自由文本描述作为分类模型的输入。自由文本数据被转换为小写字母。标点符号、额外空白、停止词和数字被去除，词语被标记化。我们从这些转换后的自由文本数据中计算了词频逆文档频率（TF-IDF）特征。选择 TF-IDF 特征提取因其简洁、计算效率和易于解释。本研究未采用更先进的 NLP

方法，如上下文嵌入（如 BERT、ClinicalBERT），因为它们需要显著增加的计算资源并引入额外复杂度。我们的目标是确定在未来工作中扩展到基于变换器的模型之前，是否能利用 TF-IDF 从 PSE 报告中提取有意义信号。

2.4.2 机器学习模型的训练

PSE 报告在报告层面分析，而非患者层面，因为患者可能有多个 PSE 提交，且患者标识符并非所有报告中都一致可用。每个 PSE 记录都有唯一的报告 ID，跨患者聚合不可行。

训练和测试数据的分配通过随机分层抽样完成。80%的数据用作训练集，其余用于测试分类器。鉴于我们数据集中属于中度类别不平衡（约 17%的 PSE 报告含有心血管诊断错误），我们评估了多种策略以减轻偏倚，包括对少数群体进行过度抽样、对多数群体进行不足抽样以及类别权重的应用。虽然过度抽样和样提高了训练敏感度，但这些方法降低了模型在测试数据上的普遍性。因此，我们选择将模型敏感度阈值设为 0.8，以确保模型间的可比性并保持外部效度。我们测试了四种类型的机器学习分类模型，用于分类显示心血管诊断错误的 PSE 报告：（1）简单逻辑回归，（2）带有 L1L2 或弹性净正则化的逻辑回归模型，（3）极端梯度增强（XGBoost），以及（4）具有二元交叉熵损失函数（DNN）的前馈深度神经网络。文献中采用简单逻辑回归模型作为基线模型，替代类似的机器学习研究。

分类模型的超参数通过不同的优化方法（贝叶斯优化、交叉验证）进行了优化。附录 A 包含超参数列表、检索范围和停止标准。为评估训练模型的性能，我们计算了操作特征曲线下的面积（AUROC）、敏感度、特异性、阳性预测值（PPV）、阴性预测值（NPV）、准确性、F-1 评分以及精确-回忆曲线下的面积（AUPRC）。

2.4.3 功能重要性。

SHapley 加法解释算法（SHAP）是一种基于博弈论解释预测模型的方法，用于捕捉影响最佳模型决策的特征。SHAP 值展示了每个特征对模型决策过程的贡献及其对预测结果（如心血管诊断错误）的影响大小。模型决策过程中价值较高的特征，平均绝对 SHAP 值也更高。

2.5 伦理考量

本研究于 2019 年 8 月 26 日获得我们的机构审查委员会（IRB #00001245）批准，自此开始有效。本研究于 2019 年 8 月 26 日至 2024 年 12 月 31 日进行。在此期间，我们的团队访问并分析了 PSE 报告。本研究未查询或使用患者或提供者标识。

3. 结果

3.1 描述性分析

共查询了 12,466 份 PSE 报告。其中 2,751 份（占 12,466 份，占 22.1%）的 PSE 报告内容空无一物或简短事实描述不足，且在数据注释前被删除。共有 9,715 份（占 12,466 份，占 77.9%）的 PSE 报告记录了简短事实描述。其中，7,467 份（共 9,715 份，76.9%）与 6,999 名独立患者相关的 PSE 报告成功映射到 EHR 数据，随后进行人工审核。我们的分析使用了 7,467 份可映射到 EHR 数据的 PSE 报告。共有 1,249 份 PSE 报告（共 7,467 例，16.7%），涉及 1,124 名独特患者（6,999 例，16.1%），包含心血管诊断错误相关叙述。表 2 显示研究队列及 PSE 报告特征。

表 2　研究样本特征

患者特征				
特征	**所有患者（n = 6,999）**	**已识别心血管诊断误差 a（n = 1,124）**	**无心血管诊断误差 a（n = 5,875）**	**p 值**
独立患者数量	6,999	1,124（16.06%）	5,875（83.94%）	
平均年龄（标准差）	58.26（20.74）	62.34（19.41）	57.44（20.98）	<0.001*
性别 N（%）				
女性	3,767（53.82%）	590（52.49%）	3,177（54.08%）	0.33
男性	3,232（46.18%）	534（47.51%）	2,698（45.92%）	
种族 N（%）				
白种	3,217（45.96%）	484（43.06%）	2,732（46.5%）	0.008*
黑种	3,182（45.46%）	557（49.56%）	2,624（44.66%）	
其他	600（8.58%）	83（7.38%）	517（8.84%）	
族裔 N（%）				
西班牙裔/拉丁裔	120（1.71%）	12（1.07%）	108（1.84%）	0.19
非西班牙裔/拉丁裔	6,530（93.3%）	1,055（93.86%）	5,475（93.19%）	
未知	349（4.99%）	57（5.07%）	292（4.97%）	
主要语言 N（%）				

患者特征				
特征	**所有患者（n = 6,999）**	**已识别心血管诊断误差 a（n = 1,124）**	**无心血管诊断误差 a（n = 5,875）**	**p 值**
英语	6,783（96.91%）	1,099（97.78）	5,684（96.75%）	0.08
西班牙语	100（1.43%）	10（0.89%）	90（1.53%）	
手语	9（0.13%）	3（0.27%）	6（0.1%）	
其他	107（1.53%）	12（1.06%）	95（1.62%）	
PSE 报告特点				
特征	**所有 PSE 报告**	**已识别心血管诊断错误 a**	**未识别心血管诊断错误 a**	**p 值**
PSE 报告数量	7,467	1,249（16.73%）	6,218（83.27%）	
严重等级 N（%）				
没关系	686（9.8%）	96（8.54%）	590（10.04%）	0.27
伤害	6,307（90.11%）	1,032（91.81%）	5,272（89.74%）	
逝世	64（0.91%）	10（0.89%）	54（0.92%）	

注：计算百分比的分母是每组中独特患者数（即所有患者=6,999;心血管诊断误差患者=1,124;无心血管诊断误差患者=5,875）。严重程度百分比的总和不高达 100%，因为一名患者可能有多个 PSE 报告，每份报告分别有不同严重程度类别。单一星号显示有无心血管诊断错误患者之间的统计学显著差异（即 p 值<0.05）。统计比较未经调整，仅供探索性解释。

[a] 由评审者决定。

3.2 机器学习模型的性能

附录 B 包含 TF-IDF 特征列表。四个机器学习模型的性能见表 3。XGBoost 在测试数据中表现优于其他模型，AUROC 为 0.914，特异度为 0.982，PPV 为 0.866，准确率为 0.929，F-1 评分为 0.738，AUPRC 为 0.783。弹性网模型更为灵敏，在所有训练模型中灵敏度为 0.711。简单逻辑回归模型在测试集中实现了 100%的敏感性和 0%特异性，表明其将所有心血管诊断误差报告归类为阳性。该退化解法凸显了未正则化线性模型在应用于不平衡高维文本数据时的局限性。虽然作为基线纳入，但该模型不可解释且临床上无效。图 2 展示了 XGBoost 模型在 AUROC 和 AUPRC 方面表现优异的表现。附录 C 包含 XGBoost 模型的优化超参数值。

表 3　使用 TF-IDF 特征矩阵分类心血管诊断错误 PSE 报告的表现

指标	简单逻辑回归		弹性网		XGBoost		DNN	
	训练	测试	训练	测试	训练	测试	训练	测试
曲线下面积	0.828 [0.812，0.843]	0.79 [0.79, 0.79]	0.929 [0.921, 0.938]	0.902 [0.903, 0.903]	0.981 [0.978, 0.985]	**0.914** [0.914, 0.914]	0.958 [0.951, 0.965]	0.911 [0.911, 0.911]
灵敏度	1 [1,1]	1 [1,1]	0.8 [0.8, 0.806]	0.711 [0.694, 0.733]	0.8 [0.8, 0.806]	0.642 [0.634, 0.66]	0.802 [0.8, 0.805]	0.677 [0.647, 0.703]
特异性	0 [0, 0]	0 [0, 0]	0.907 [0.881, 0.926]	0.898 [0.872, 0.913]	0.989 [0.983, 0.991]	**0.982** [0.975，0.983]	0.964 [0.954, 0.973]	0.95 [0.936, 0.964]
付费观看	0.155 [0.155, 0.155]	0.155 [0.155, 0.155]	0.613 [0.552, 0.664]	0.561 [0.514, 0.594]	0.927 [0.897, 0.944]	**0.866** [0.828，0.871]	0.802 [0.76, 0.847]	0.714 [0.668, 0.765]
净现值	0 [0, 0]	0 [0, 0]	0.961 [0.96, 0.962]	**0.944** [0.941，0.947]	0.964 [0.964, 0.965]	0.937 [0.936, 0.94]	0.964 [0.963, 0.964]	0.941 [0.937, 0.945]
准确性	0.155 [0.155, 0.155]	0.155 [0.155, 0.155]	0.891 [0.868, 0.906]	0.869 [0.851, 0.879]	0.959 [0.955, 0.962]	**0.929** [0.926，0.93]	0.939 [0.93, 0.947]	0.908 [0.9, 0.916]
F-1 评分	0.268 [0.268, 0.268]	0.269 [0.269, 0.269]	0.694 [0.653, 0.726]	0.627 [0.607, 0.64]	0.859 [0.846, 0.867]	**0.738** [0.733，0.74]	0.802 [0.78, 0.824]	0.695 [0.685, 0.711]
平均精度-召回曲线	0.797 [0.778, 0.817]	0.726 [0.726, 0.726]	0.762 [0.739, 0.788]	0.726 [0.726, 0.726]	0.926 [0.913, 0.94]	**0.783** [0.783，0.783]	0.862 [0.843, 0.882]	0.774 [0.774, 0.774]

注：加粗字体显示各绩效指标的最佳值。括号内数字代表估计的 95%置信区间。

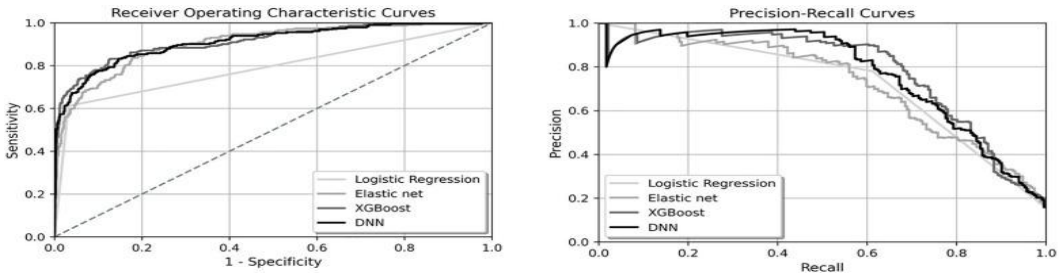

图 2 采用 TF-IDF 特征矩阵分类心血管诊断错误 PSE 报告的受试者操作特征（ROC）曲线和精确回忆曲线（PRC）

为评估训练机器学习模型特征数量的减少，我们还对 TF-IDF 特征进行了卡方特征选择，并与其他模型进行比较。然而，我们发现该方法并未提升分类表现（附录 D）。此外，将患者人口统计与 TF-IDF 特征整合并未提升 XGBoost 模型的分类性能（附录 E）。

3.3 最重要的特征

图 3 展示了在 PSE 报告样本中识别心血管诊断错误报告时最具影响力的特征。起搏器（SHAP = 0.17）、次数（SHAP = 0.14）、心电图（SHAP = 0.12）、心脏（SHAP = 0.12）和胸部（SHAP = 0.12）分别是 XGBoost 模型决策中最重要的五个特征。

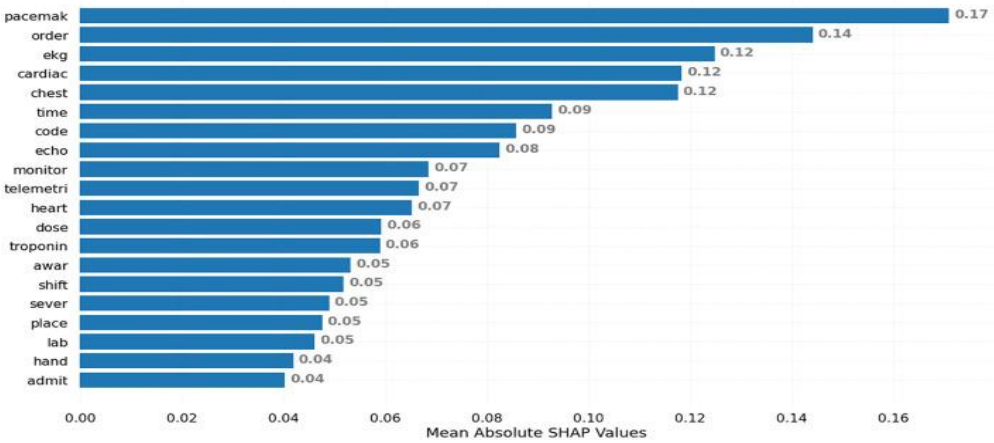

图 3　平均 SHAP 值图

对于每个特征，我们计算了所有观测值的平均 SHAP 值。具体来说，我们取绝对值的平均值，因为我们不希望正负值相互抵消。贡献较大的正负特征平均 SHAP 值较大。换句话说，这些特征对模型预测产生了显著影响。词干是干的。

4. 讨论

在 16.1%的 PSE 报告中发现了心血管诊断错误。据我们所知，本研究是首个从 PSE 报告中评估心血管诊断错误的项目。为了替代文献中心血管诊断错误的估计患病率，我们将数据中的心血管诊断错误率与 CRICO 策略比较基准系统（CBS）进行比较，后者是全球最大的详细编码医疗事故索赔数据库，包含>35 万起门诊普通医学医疗事故索赔案件。CBS 数据库中心血管诊断错误的患病率为 10%，低于 PSE 报告中的水平。然而，我们的数据涵盖了门诊和医院环境，这可能导致我们研究中心血管诊断错误率较高。

我们在 PSE 报告样本中观察到年龄、种族与心血管诊断错误相关叙事存在的统计学上显著关联，且在年长黑人患者中，已识别的诊断错误叙述比例更高。我们的

观察结果与文献一致，最近的系统综述描述了急诊科环境中黑人患者漏诊心肌梗死和急性冠状动脉综合征的风险持续增加。其中一个可能的解释是，历史上边缘化的患者亚群体中观察到的医生评估等待时间更长。例如，一项针对 18 至 55 岁成年人急诊就诊的最新研究——国家医院门诊医疗护理调查发现，女性和有色人种在医生评估前等待时间较长，且与其临床特征无关。总体而言，这些人口差异应被解读为联想模式，而非因果关系，尤其是在临床医生报告的数据使用下。临床医生报告的数据最好被理解为捕捉诊断错误信号，更适合用于监测、学习和假设生成，而非估计真实诊断错误率。因此，我们的发现可能反映了医疗系统内的系统性、情境性或报告相关因素，而非诊断准确性固有差异。此外，表 2 中心血管诊断错误组与无错误组之间种族分布的差异凸显了评估算法公平性的重要性。不同人口统计组间患者安全事件的记录和报告方式差异可能影响模型特征和表现。虽然我们当前的评估侧重于整体准确性，但未来工作将纳入子组公平性分析，如分层敏感性和特异性估计以及公平机器学习指标（如机会均值），以确保各人口统计学群体的表现公平，自动识别方法不会无意中强化现有的诊断安全性差异。

我们的研究结果表明，医疗系统有潜力从现有 PSE 数据源中捕捉疾病特异性诊断错误，进一步加速诊断安全学习和改进工作。尽管我们的研究仅聚焦于心血管疾病，未来研究仍可利用类似方法识别与其他疾病相关的诊断错误，并利用现有 PSE 数据。我们发现 XGBoost 模型在识别心血管诊断错误风险较高患者方面表现优于其他模型。类似机器学习模型的研究也发现，该算法优于逻辑回归等线性模型，且在分类表格数据（如本研究中的 TF-IDF 特征）中通常优于 DNN 模型。此外，DNN 模型需要大型训练数据集，而 XGBoost 则可以轻松地在中小数据上训练。XGBoost 高度灵活且可定制，非常适合高维特征空间和类不平衡数据的分类任务。

XGBoost 模型的灵敏度达到了 64.2%，表明约三分之一的心血管诊断错误相关报告未被识别。这一漏接率可能过高，不适合仅用于患者安全监测，因为敏感性通常被优先考虑，以避免忽视潜在的诊断问题。然而，作为概念验证的分诊工具，该模型仍可通过大幅减少需要人工审核的报告量发挥价值，前提是它补充而非替代人工审核。可接受的敏感阈值最终取决于系统的预期用例。调整概率阈值可以提高灵敏度，但这将以特异性和审查员工作量增加为代价。未来工作中，进一步的阈值优化和临床验证将至关重要。

我们观察到训练集和测试集之间的性能差异（例如，XGBoost 敏感度从 0.80 降至 0.642，PPV 从 0.927 降至 0.866）。鉴于 TF-IDF 特征空间的高维度及叙事安全报告固有的语言变异性，这些差距很可能反映了预期的过度拟合。尽管采用了交叉验证、正则化和贝叶斯优化来减轻过拟合，但这些结果强调了外部验证的重要性，并表明模型可能正在学习系统特定的术语。未来工作将探索更多正则化策略，并评估其他医疗系统 PSE 报告数据的性能。

值得注意的是，本研究的方法学组成部分（TF-IDF 特征提取和标准监督学习分类器）是有意选择的，基于其可解释性和可重复性，而非算法创新性。本研究的创新在于证明这些既有技术可应用于医疗安全基础设施中，自动识别心血管诊断错误叙事，这是一项此前未曾测试的应用，支持可扩展的诊断安全监测。

根据平均绝对 SHAP 值，起*搏器*在我们的 PSE 报告中成为心血管诊断错误的重要信号。对包含"*起搏器*"一词的报告自由文本描述进一步分析显示，一个常见主题是关于为有起搏器或被错误怀疑使用起搏器的患者安排或完成 MRI 检查的问题。例如，PSE 报告中报告了患者完成 MRI 扫描的延迟，描述了关键 MRI 筛查信息不准确或未传达，导致检测结果和诊断过程延误的情况。例如，"*患者两天前被安排做 MRI。MRI 仍未完成。今天早上叫来 MRI 时，他们误传了患者装有起搏器的信息，而患者实际上并没有。这是患者护理的延迟，因为该检查是在两天前下达*的。"其他具有高平均绝对 SHAP 值的特征包括*心率*、*心电图*、*心脏和胸部*。大多数具有高 SHAP 值的词汇，可以被医疗系统用来进一步探讨患者的心脏状况、相关护理流程及相关安全事件。需要强调的是，文本特征（如起搏器、命令、心电图）与心血管诊断错误之间的关系，反映的是模型捕捉的联想模式，而非因果机制。观察到的特征重要性值表示 PSE 报告中与诊断错误叙事统计相关的术语，但不应被解释为因果通路或促成因素的证据。

尽管我们的探索性综述发现许多与起搏器相关的 PSE 报告涉及 MRI 订购和设备兼容性的延迟或混淆，但这只是一个示例，而非全面的子分析。起搏器相关报告可能代表多种不同的失效模式，包括不准确或不完整的 MRI 筛查信息、设备状态的文档错误、设备兼容性的不确定性以及临床团队间沟通失误。对这些报告进行系统性定性和定量分析（例如量化归因于 MRI 相关问题的比例与其他过程失效）将提供更深入的见解，但超出了本概念验证研究的范围。未来工作将开发结构化分类法，以

分类这些失效模式并评估有针对性的干预措施，如标准化 MRI 筛查工作流程、改进植入设备文档以及加强心脏设备管理的沟通路径。

在现有形式下，该模型应被视为一个概念验证筛查工具，旨在支持而非取代人工 PSE 审查流程。在许多医疗系统中，PSE 报告由临床医生、患者安全官员、质量人员或人因专家手动审核，鉴于报告量大且诊断错误相关事件相对罕见，这一过程可能非常耗时。自动化分类器可作为初步的分诊步骤，优先处理与心血管诊断错误相关的语言模式报告。模型标记的报告随后将由临床和安全专家进行传统人工审核，确保人类判断始终处于核心地位。这种分诊方法有潜力减轻审查员的工作量，聚焦于潜在高风险叙事，并加快组织学习周期。然而，任何运营部署都需要外部验证、工作流程整合评估以及潜在时间或资源节省的评估。

我们的研究存在一些关键局限性。首先，尽管 PSE 报告来自多家不同医院，但所有医院均属于同一医疗系统，因此可能反映该医疗系统独有的特定语言、格式或共享文化价值观，因此可能无法推广到其他医院和医疗机构。开发的机器学习分类器应在多家医院数据上进行验证，以验证其可推广性。因此，本研究应被视为一个概念验证，需在临床或运营使用前进行外部验证。未来工作将重点利用其他医疗系统的 PSE 报告进行外部验证，以评估模型在不同报告环境中的可重复性和稳健性。事件分类法、叙述风格和安全报告流程的差异可能影响模型性能，在异构数据集中验证该方法对于确保可推广性并促进诊断安全监控工作流程中的广泛应用至关重要。其次，由于 PSE 报告数量众多且审查过程劳动强度大，我们聘请了全职非临床研究助理手动审查每份 PSE 报告并分配心血管诊断错误标签。虽然该研究助理专长于人因工程并熟悉健康系统安全分析，但评审人偏见可能贯穿整个评审过程。为减少偏倚，评审人员接受培训识别“不明确”报告，并与临床专家协同进一步审查。为进一步提升标签过程的可信度，我们对随机子集报告进行了 IRR 评估。两位评审者 90% 的 Cohen's Kappa 支持模型训练和评估中手动注释的可靠性。不过，未来研究可考虑由第二位评审者审阅更大样本的报告样本，以减少编码模式中潜在的评审偏差。第三，确实有大量分类器可以针对该问题进行训练;但为了保持清晰和聚焦，我们有意选择了一组简洁的模型，最适合基于文本、不平衡的数据，并代表不同的方法家族（线性、正则化、集合和神经网络）。未来工作应扩展机器学习分类器，以寻找其他潜在适合的模型。最后，本研究采用了 TF-IDF 方法从自由文本数据中提取特征。

该方法被选为探索叙述报告中嵌入信息内容的有效初始步骤。作为文本预处理的一部分，去除了数字和停字，以减少 TF-IDF 特征空间中的噪声。这一步骤适用于 TF-IDF，因为此类模型依赖原始词频，不捕捉语义或语境关系。我们承认，在更先进的 NLP 方法（如 BERT 变体）中，数值、多词医学短语和上下文线索可能携带临床意义的信息，因此会被保留在这些模型中。尽管中间方法如 n-gram 或上下文嵌入可能提升模型性能，但本研究的目标是通过 interpretab 演示概念验证 LE，计算效率高的特征表示。未来采用基于变换器的方法将纳入这些元素。更先进的技术，如基于 BERT 的模型和大型语言模型（LLM），为未来探索揭示 PSE 叙事中更深层的上下文模式提供了有前景的方向。目前，支持其在 PSE 报告中表现的证据有限，需进一步研究以评估其在该领域的可行性、准确性和可解释性，才能在临床或运营实施前进行。此外，未来研究可通过纳入 PSE 报告之外的数据来源，扩大分析范围，包括问题列表、ICD-10 代码和出院摘要。整合这些数据源可能有助于识别特定漏诊或延迟心血管诊断。

最后，由于不同运营场景需要不同的性能权衡，未来实现应探索阈值调优，以使模型输出与患者安全团队的需求保持一致。在将该模型整合到实际操作前，需要对敏感性与特异性权衡进行全面评估。

5. 结论

本研究展示了利用机器学习和自然语言处理识别包含心血管诊断错误叙事的 PSE 报告的概念验证方法。虽然结果显示了单一医疗系统内的可行性，但该方法在其他机构的推广性尚不明确。在考虑更广泛实施之前，跨不同报告环境进行外部验证将至关重要。