https://doi.org/10.1038/s44271-024-00097-2

基于问题的计算语言方法在量化情绪状态方面优于评级量表

斯维尔克・西克斯特罗姆, 伊娃・瓦拉维丘特, 伊纳里・库塞拉 和妮可・埃沃斯

心理结构通常用封闭式评分量表来量化。然而，自然语言处理（NLP）的最新进展使得开放式语言反应的量化成为可能。在这里，我们证明，与传统的评分量表相比，使用 NLP 分析的描述性词语响应在情绪状态分类方面显示出更高的准确性。一组参与者（N = 297）生成与抑郁、焦虑、满足或和谐相关的叙述，用五个描述性词语对其进行总结，并使用评级量表对它们进行评级。另一组（N = 434）从作者的角度评估这些叙述（使用描述性词语和评分量表）。使用 NLP 对描述性词语进行量化，并使用机器学习将反应分类为相应的情绪状态。结果显示，基于描述性词语（64%）的准确分类数量明显高于基于评分量表（44%）的准确分类，这对评分量表在测量情绪状态方面比基于语言的测量更准确的观点提出了质疑。

虽然人工智能 (AI) 在心理健康领域的应用是一个很有前景的研究领域，但必须确保正在开发的算法可靠、准确和透明。基于人工智能的心理健康相关技术的准确性可以通过将基于人工智能的语言分析方法与真实情况进行比较来估计，其中常见的方法是标准化评分量表。在这项工作中，我们采用了另一种方法，引入了一种范例，其中第一阶段的参与者被指示写一篇关于情绪状态的自传体叙述，由第二阶段的参与者阅读。这种方法很有趣，因为验证是由参与者的自我体验的情绪。我们向参与者询问一个与叙事情感相关的开放式语言问题，他们用五个描述性词语进行回答，这些词语通过人工智能方法进行分析，并将其与标准化评分量表进行比较。

语言是人们交流心理状态的自然方式。尽管如此，标准化数字评分量表是行为科学家测量心理状态的主要方式，因为它们被认为具有更高的有效性。例如，在《人格与社会心理学杂志》上发表的一篇典型研究文章中，87% 用于得出结论的数据都是基于封闭式评分量表。然而，这些量表具有局限性，因为它们是一维的，通常范围从“强烈不同意”到“强烈同意”，它们容易产生集中趋势错误，具有光环效应，或者受到自我观察能力的限制，仅举几例

学者们对评级量表的有效性越来越持怀疑态度，因为它们可能会过度简化人类体验。

或者，开放式语言反应可以传达一个人的心理状态的更加以人为中心和整体的表征。例如，可能有多种方式可以表达特定疾病的指标，以解决 DSM-V 规定之外的症状。在重度抑郁症的情况下，可能会出现躯体症状（例如头痛或消化问题）等其他症状，这些症状通常与抑郁症无关。DSM-V 因其过分强调根据特定标准对症状进行分类而受到普遍批评，可能忽视了精神障碍的多样化表现和变化。它还因过于严格而受到批评，可能会排除具有不符合指定标准的临床显着症状的个体。这表明采用开放式响应格式的方法可能会在诊断过程中提供更大的灵活性，从而适应个体差异和背景因素。开放式回答可以对心理健康状况提供更全面的评估。 这种方法可以补充诊断手册，提高对心理健康状况的准确性和个性化理解。通过考虑 DSM-V 规定之外的更广泛的指标，研究人员和临床医生可以对心理健康状况有更细致的了解，

隆德大学心理学系，隆德 SE-221 00，瑞典。 电子邮件：sverker.sikstrom@psy.lu.se

可以更好地捕捉每个人的复杂性并允许更个性化的治疗计划。

对开放式回答进行手动分析有缺点：这是一个耗时且费力的过程，容易受到个人偏见的影响。因此，在计算机技术繁荣之前，评级量表的广泛采用成为一种受欢迎的替代方案，缓解了这些挑战。最近的研究结果表明，捕获精神疾病的传统方式（例如，针对抑郁症的 PHQ-9 或针对焦虑症的 GAD-7）可能会忽略与特定精神疾病相关的其他同等重要的症状。幸运的是，自然语言处理 (NLP) 的最新进展提供了一种潜在的解决方案，可以有效解释和量化语言反应，同时保持测试的可靠性。

技术进步，尤其是人工智能和机器学习 (ML) 领域的技术进步，通过分析大型数据集并识别传统方法（例如基本统计建模或简单回归分析）可能不易察觉的模式，极大地提高了预测结果的准确性和效率。 ）。这些过程实现了跨不同领域决策的自动化，因为系统无需外部干预或监督即可自主学习任务。人工智能的最新发展简化了营销专业人员、临床医生、统计学家和各种分析师的工作，并且在网络搜索、定向营销和金融等领域观察到了有希望的结果。这些机器学习应用程序促进了情感分析器、文本分类器、聊天机器人和虚拟助手的发展，这些应用程序也可能能够改变心理健康诊断领域。

机器学习模型在基于广泛不同的数据集、遗传学、磁共振成像、脑电图和临床数据（例如 2 型糖尿病）的评估和预测精神障碍方面显示出了巨大的前景。在临床环境中，ML 的一个流行应用是 NLP，它特别用于电子健康记录、医疗诊断和社交媒体文本数据挖掘。然而，使用此类输入数据需要访问并能够提供大量最新医疗记录。为了克服这些限制，我们专注于可以通过提示参与者回答与心理健康相关的单一开放式问题来轻松评估的数据。此类回答可以通过几个描述性词语在几秒钟内得到回答，并与常用的评分量表进行比较，每个评分量表由多个项目和多个回答选项组成。

人工智能在精神健康领域加速应用，有助于诊断和治疗方法的改进。亚马逊等私营企业已经在这一领域取得了长足的进步。例如，亚马逊开发了一项专利，允许其 Alexa 设备识别抑郁症和自杀倾向，并计划将该技术与其医疗保健和制药业务结合起来，创造新的盈利机会。然而，私营公司开发的心理健康算法可能并不总是透明的。人们担心这些算法的潜在偏见和道德影响，特别是当它们在没有适当监督和监管的情况下开发时。在这里，研究对于确保算法的可靠性和有效性至关重要。例如，NLP 算法取得了进步，可以分析患者对话并识别表明潜在临床问题的模式。该领域的许多工作都集中在开发可以根据社交媒体数据预测心理健康结果的计算模型。例如，研究表明，可以通过分析社交媒体帖子来识别有抑郁或焦虑风险的人。尽管社交媒体研究显示出巨大的前景，但评估当前情绪需要当前的相关社交媒体数据，而这些数据并不总是可访问的。

近年来，NLP 模型取得了重大进展。 Transformer 是一种强大的机器学习技术，它的出现带来了显着的性能改进。其中，BERT（来自 Transformers 的双向编码器表示）脱颖而出，成为最常被引用的基于 Transformer 的语言模型。

Transformer 是一种灵活且庞大的统计模型，以其在上下文中捕获单词含义的能力而闻名。评估一致表明，与早期模型相比，BERT 显着减少了错误。凭借其大尺寸和灵活性，这些模型擅长表示不同上下文中的不同单词含义，从而增强研究人员掌握演讲者和作者微妙意图的能力。虽然 BERT 确实有其局限性，包括可能反映诊断和社会刻板印象的主观性和系统偏见，但不可否认的是，它在 NLP 领域取得了重大突破。 BERT 的开创性贡献在于引入了无监督的预训练模型，使它们能够从大量未标记的文本数据中获取见解。

研究表明，通过计算方法分析的基于文本的答案确实可以预测相应的封闭式评分量表，例如 PANAS、Ryff 心理健康量表、生活满意度量表 (SWLS)、抑郁、焦虑和压力量表等。最近引入了单词标记的概念，用于通过情绪回忆任务检索情绪状态，参与者列出了上个月经历的 10 种情绪状态。该研究证明了情绪回忆任务和 PANAS 之间存在显着相关性，揭示了不同的心理搜索策略，特别是对于具有不同积极或消极情绪水平的个体。在此基础上，研究人员开发了 DASentimental，这是一种基于情绪回忆任务的半监督机器学习模型。这个创新系统学会了将焦虑、压力和抑郁的心理测量结果映射到用户生成的单词序列，采用植根于认知科学和关联知识模型的搜索策略。基于问题的计算语言评估 (QCLA) 涉及通过要求参与者回答开放式问题来生成文本，这些问题可以使用 NLP 转换为可量化的向量。该方法捕获心理健康的严重程度和心理状态的详细描述。一项相关研究还指出了 QCLA 的有效性，其中使用 NLP 分析的计算语言评估比使用传统的评分量表更准确地区分人类合作。总的来说，这些研究强调了认知信息的重要性，这些信息可以使用人工智能从基于单词的反应中得出。 这对于使用非结构化临床记录的临床研究人员尤其重要，这些临床记录本来就没有记录，也没有额外的好处。

虽然 NLP 方法具有优势，但过去基于语言的方法的研究尚未达到与评级量表相同的准确性水平。其原因是，之前基于语言的方法的验证主要使用评分量表作为结果衡量标准，其中有效性是通过与评分量表的相关性来衡量的，因此不允许测试评分量表或语言衡量是否具有最高有效性。为了解决这个问题，本研究使用独立于评分量表的结果标准来比较基于文字的描述性反应和评分量表。我们特别关注一个结果标准，即指导参与者生成自我体验的叙述，其中包含与抑郁、焦虑、满足或和谐（也称为情绪状态）等关键心理结构相关的特定情感内容。然后，其他人使用评级量表和描述性词语对参与者制作的叙述进行评估。我们研究了这些评估如何将叙事分类为用于生成叙事的情感。

这项研究的主要和原创贡献在于，它检查了使用人生成的文本响应与评级量表相比评估情绪状态的准确性。虽然之前在其他作品中已经介绍过用于分析语言数据的计算方法，但本研究特别关注这两种评估方法的比较评估。目前的工作旨在表明，与专门用于测量这些状态的常用评级量表相比，通过 NLP 分析的基于语言的情绪状态测量在对情绪状态进行分类方面具有更高的准确性。特别是，我们向参与者提出了一个开放式问题，他们用五个描述性词语进行回答（即“写下五个最能捕捉您的/

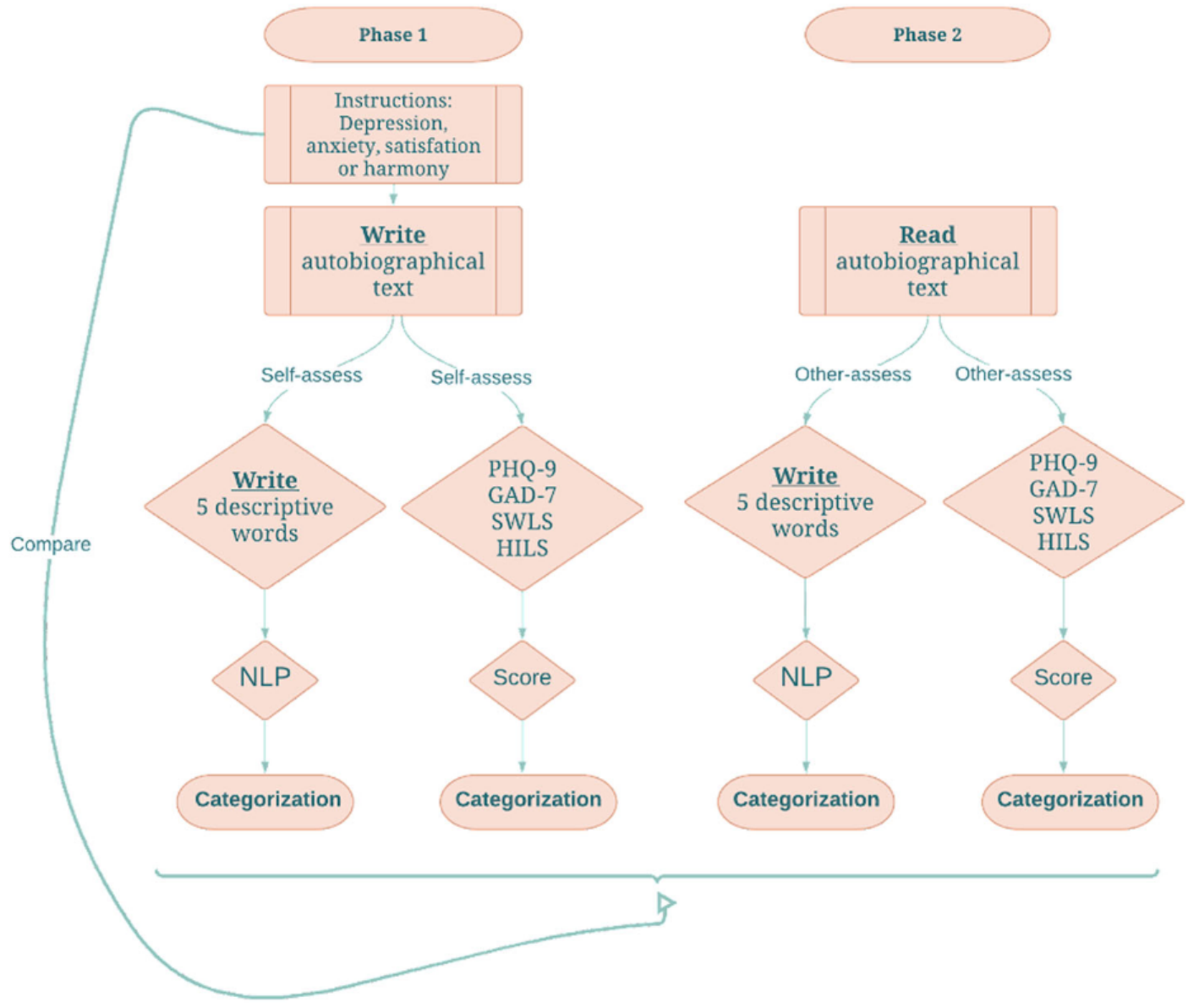

图1|范式设计。该图显示了研究中使用的范式。在第一阶段，在左侧，参与者被要求写一篇自传体叙述，讲述他们经历过抑郁、焦虑、满足或和谐的经历。接下来，他们用五个描述性词语总结叙述中的情感，并填写与四种情感相对应的评分量表（分别为 PHQ-9、GAD-7、SWILS 和 HILS）。这五个单词被大型语言模型（BERT）量化为向量。机器学习算法（多重逻辑回归）用于对评分量表的向量或总分进行分类。这种分类与参与者在撰写叙述时被指示使用的情感进行比较。第 2 阶段与第 1 阶段相同，不同之处在于参与者阅读第 1 阶段中生成的故事（而不是写它们）。

作者的情绪状态”）。我们选择了五个描述性单词作为响应格式，因为之前的工作表明，与自由文本数据相比，这种数据类型显示出更准确的预测，并且添加更多单词并不会显着提高预测准确性。编写一些描述性词语可以在相当短的时间内（例如几秒钟）完成，而编写自由文本叙述通常需要更长的时间（例如几分钟）。我们还选择了一些与情绪状态相关的最常见的心理学概念，即抑郁、焦虑、满足和生活和谐，因为它们代表了积极心理学和临床心理学元素的混合，每个元素都具有特定的情绪效价，可以通过使用标准化评级量表进行量化，并且通常具有高度相关性。这四种心理结构的选择也与我们研究小组之前的工作相一致，确保了一致性并促进了研究结果的可比性。

在这里，我们引入了一个由两个阶段组成的新范式，参与者在第一阶段产生情感叙述，并由第二阶段的其他参与者进行评估（见图 1）。在研究的第一阶段，参与者被要求写一篇自传体文本，内容涉及抑郁、焦虑、满足或和谐的特定自我经历事件。在第二阶段，一组不同的参与者（由对照组和医疗保健专业人员组成）被要求阅读由第一阶段某人编写的文本。在这两个阶段中，参与者用五个描述性词语描述了叙述的情感内容，并完成了评分对应于四种情绪的量表（即患者健康问卷（PHQ9）、广泛性焦虑症量表（GAD-7）、SWLS 和生活和谐量表（HILS））。

我们采用 NLP 技术将五个单词的响应转换为嵌入，特别是语义向量。这涉及利用基于我们早期研究工作的计算方法。嵌入和评分量表的分类基于多项回归，将每种回归分配给四种情绪状态之一。作为 NLP 模型，我们选择了最先进的基于 Transformer 的语言模型（即 BERT）。分类准确性的评估基于十倍交叉验证程序。 预先注册的假设

预计在评估预定情绪状态时，在进行 NLP 分析时，与相应的评分量表相比，描述性词语反应将表现出卓越的预测能力（H1）。此外，据假设，对不同情绪状态的描述性词语反应和评级量表评估的联合考虑将产生比单独评估更高的预测准确性（H2）。我们还假设，与相应的评分量表相比，与叙述相匹配的描述性词语反应在第一阶段和第二阶段之间表现出更高的人际可靠性（H3）。最后，由于预计医疗保健专业人员的数量相当少，因此没有针对该群体做出具体假设。因此，我们测试了与对照组相比，专业人士对情绪状态的评估是否与评分量表更紧密相关，表明预测能力更高（H4）。

方法

该研究获得瑞典隆德地区伦理委员会的伦理批准，遵守瑞典法律（Dnr 2021 – 04627），数据收集于2022年5月完成。所有方法均按照相关指南和规定进行，遵循赫尔辛基宣言。在开始数据收集之前，已获得所有受试者的书面知情同意书。该研究的设计、假设和分析计划在研究完成前已在开放科学框架 (OSF) (2022-03-15) 上预先注册，可通过 https://osf.io/6fx72 访问。关于 OSF 的第三个预先注册假设的准确措辞是：“在参与者之间进行测量时，语义尺度的相关性高于相应的估计尺度。”

参加者

这里使用的数据是通过方便抽样收集的。参与者是通过 Prolific 招募的，Prolific 是一个用于收集行为科学数据的在线招聘平台。入选标准为年满 18 岁且母语为英语。该研究分为两个阶段，参与者不同，但纳入标准相同。参与者在第一阶段获得 2 英镑补偿，在第二阶段获得 1.5 英镑补偿

表 1 |参与者人口统计数据

阶段	总氮	N 排除	Age	性别	国籍	Time	教育
1	350 (297)	53	19–76, 29.37 (13.62)	女性-165 男122 未公开-10	US-115 英国-108 其他-74	14.25 (17.51)	107 62 46 6 76
2	465 (434)	31	18–79, 34.57 (11.79)	女性-263 男159 未公开-12	US-235 英国-110 其他-89	9.84 (11.78)	127 192 63 8 44
1 & 2	815 (731)	84	18–79, 31.97 (12.71)	女性-428 男281 未公开-22	US-350 英国-218 其他-163	12.05 (14.65)	234 254 109 14 120

[a] 排除之前（括号内）的参与者总数。

[b] 年龄以最小和最大年龄范围、平均值和标准差（以岁为单位）表示；

[c] 完成研究所花费的时间用平均时间（SD）表示，以分钟和秒为单位；

[d] 学历（从上到下）：高中、本科、研究生、博士、其他。

参与研究。共有 350 名参与者完成了第一阶段的研究，其中 53 名参与者因未能正确回答控制问题而被剔除，最终样本量为 297。 共有 465 名参与者完成了第二阶段的研究，其中 34 名参与者是通过额外筛选招募的医疗保健专业人员。同一平台还用于招募医疗保健专业人员，根据参与者在医疗保健系统内是否拥有专业职业的迹象来筛选参与者；医生、紧急医疗人员、护士、护理人员、药剂师、心理学家或社会工作者。总共 31 名参与者因没有正确回答控制问题而被剔除，最终样本量为 434 人。因此，两个阶段的最终样本由 731 名参与者组成（女性 = 428；男性 = 281；受访者未公开）性别 = 22），年龄范围为 18-79 岁（M = 31.97，SD = 12.71）（更多详细信息请参见表 1）。

措施

自传体叙事。该研究包括情感自传文本反应、描述性词语反应、四种评分量表和人口统计问题。第一阶段的参与者被要求写下他们感到抑郁、焦虑、满足或和谐时的亲身经历的事件。供参与者书写的心理结构的分配是随机进行的，确保每个结构收到大致相同数量的回复。说明如下：“请写一篇关于您在生命中经历过[抑郁/焦虑/满足/和谐]的一段时间（几天到几个月）的文字。请至少写一个段落（大约五句话）来回答问题。写下对您来说最重要和最有意义的那些方面。笔记。请不要在您的文字中使用“[抑郁/焦虑/满意/和谐]”一词。”

虽然这两个成对的心理概念（抑郁-焦虑和满足-和谐）在理论上是不同的，但在通过评级量表进行评估时，它们往往表现出高度的相关性。鉴于这种概念和基于标准的区别，我们研究中的语义测量被建议作为比传统评级量表更明显地区分它们的手段。

描述性词语。参与者还被要求用五个描述性单词捕捉叙述的情感方面（不允许包含超过一个单词的短语，如果在文本字段中输入超过一个单词，则不允许参与者继续进行调查）。第一阶段的说明如下：“写下五个最能体现您之前所写的情绪状态的关键词。笔记请不要在文字中使用[抑郁/焦虑/满足/和谐]这个词。”第二阶段的参与者只被要求描述他们在文本中读到的心理状态，而不知道文本具体指的是哪种情绪。参与者以随机顺序呈现文本，确保每篇文本至少被阅读一次，但不超过两次。第二阶段的措辞如下：“阅读这篇关于某人生命中某个时期的文字，并写下五个最能体现作者情绪状态的关键词：[文字已插入此处]。”

评级量表。本研究使用了以下四种标准化评分量表。 PHQ-9 是一种广泛使用的评估抑郁症严重程度的工具。它由九个问题组成，对应于 DSM-IV 中重度抑郁症的诊断标准。 PHQ-9 上的九个项目涵盖了抑郁症的各种症状，例如情绪、睡眠、食欲和能量水平。每个项目的评分范围为 0 到 3。然后将参与者对九个项目的回答相加，得到总分，范围为 0 到 27。总分表明抑郁症的严重程度：轻微、轻度、中度、中重度抑郁症和重度抑郁症。广泛性焦虑症（GAD-7）量表是衡量广泛性焦虑症严重程度的常用心理健康指标。 GAD-7 上的七个项目的评分范围为 0 到 3。然后将参与者对这七个项目的回答相加，得到总分，范围为 0 到 21。总分表明广泛性焦虑的严重程度障碍：轻度、轻度、中度和重度焦虑。 SWILS 是一种广泛使用的衡量个人整体生活满意度的工具。 HILS 衡量一个人对生活的整体满意度和满意度。 HILS 和 SWILS 量表均由 5 个陈述组成，参与者按 1 到 7 的等级进行回应。然后将参与者对这些项目的回应相加，得到总分，总分范围为 5 到 35。分数越高表示水平越高生活满意度/生活和谐度，而较低的分数表明生活满意度/和谐度较低。

参与者被要求使用上述量表来捕捉叙述的情感状态。换句话说，标准化评分量表的说明在第一阶段进行了修改，以便它们指的是自我经历的事件中的情绪，而不是当前的状态，而在第二阶段，它们被修改为与所阅读的叙述中的情绪相对应（详情请参阅 OSF 的预注册报告）。在第一阶段，评级量表说明如下：“在生命的那段时期，您多久被以下问题困扰过：……”。在第二阶段，说明被修改为与阅读的叙述相关，例如：“考虑作者的情绪状态：在他们生命的那段时期，通过点击适当的框来表明他们对每个项目的同意。”。

程序

Prolific（一个招募在线研究参与者的平台）将参与者引导至 Qualtrics 问卷，其中记录了他们的回答。 Qualtrics 调查开始后，使用安全技术 reCAPTCHA 来区分人类用户和自动化机器人。通过该阶段后，参与者被告知研究的目的、他们随时退出的权利以及他们的匿名和自愿的回答。他们还被告知，不会收集任何个人或可识别信息，如果对调查有任何疑问，他们可以联系研究人员（有关详细信息，请参阅 OSF 的预注册报告）。

在第一阶段，参与者被要求写一篇自传体文字，讲述他们生活中经历过以下情绪状态之一的一段时期：抑郁、焦虑、和谐或满足。因此，每个参与者都写了关于一种情绪的叙述，但不知道其他三种情绪。所描述的情绪状态在参与者之间均匀分布。然后，参与者被要求写下五个单词来描述叙述中的情绪状态，完成四个标准化评分量表，并回答包括年龄、性别、出生国家和完成教育水平在内的人口统计问题。

第二阶段的程序与第一阶段相同，只是参与者被要求阅读第一阶段某人写的文本，而不是自己写的文本。阅读完文本后，他们对所读的叙述写下五个描述性词语，并完成与叙述中的情感相关的评分量表，即，不是第一阶段中他们自己经历的情感，而是作者的情感。这是通过在语义问题和评分量表的措辞中用“作者”和“他们”替换“你”来完成的。每个参与者只阅读并回答一个叙述。然而，由于第二阶段的参与者比第一阶段更多，第一阶段的一些叙述在第二阶段被阅读和评估了不止一次（但不超过两次）。

在这两个阶段，参与者都可以自由地撰写叙述并完成问卷的其余部分，只要他们愿意。这些叙述经过了人工审查，以确保排除任何无意义的数据，因此不会删除任何叙述。对于叙述，没有最小或最大字数。平均而言，参与者在叙述中写下 N = 81.32 个单词（SD = 40.66）。最后，为了确保参与者认真遵循说明，在两个阶段中都嵌入了四个控制问题，每个评级量表问题中都有一个。参与者被要求用预先给定的选项回答问题，例如：“回答‘同意’这个问题”。如果四个控制问题中的任何一个没有正确回答，则该参与者的数据将从最终样本中删除。

数据分析

使用 BERT 对参与者生成的描述性词语进行量化。四种情绪的预测分类是通过多项逻辑回归生成的。 分析是在 SemanticExcel.com 上进行的，这是一个用于语义数据统计分析的在线工具，主要作者的底层代码是用 MATLAB 编写的。

语义数据的预处理。作为本研究的一部分收集的描述性词语和自传体叙事反应根据 Kjell 及其同事提供的程序进行了修改。首先使用手动程序清理参与者生成的描述性词语响应。拼写错误的单词会通过 Microsoft Word 中的拼写工具进行更正，而且只有在作者的意图明确的情况下才会进行纠正；否则，它们的原始形式将被保留。错误回答控制问题、连续重复单词或参与者写下“N/A”的情况被排除在外。对自传体叙述进行了最小程度的修改，主要是解决拼写错误和重复单词（例如，the - hte；the the）。由于改动较小没有对原始叙述进行明确的分析。我们还明确指示参与者不要在文本或描述性词语响应中使用感兴趣的词语。没有删除任何完整的自传体叙述。

量化描述性词语和评级量表。通过 BERT 模型（即“bert-base-uncased”）分析参与者在第 1 阶段和第 2 阶段生成的描述性单词，其中我们使用最后一层（即第 12 层）中的嵌入。 BERT 嵌入有 768 个维度。分类基于以下组成的向量：仅从单词数据生成的嵌入（768 维）、仅从四个评分量表（4 维）生成的嵌入，或评分量表和嵌入的组合（4 + 768 维）。由于嵌入的维数 (768) 相对于数据集的大小相当大，因此使用称为奇异值分解的数据压缩算法来压缩该向量，以便第一维包含有关原始向量的最重要信息。为了使三种分析之间的比较公平或具有可比性，该数据压缩算法也应用于仅评分量表分析，尽管这在计算上不是必需的，因为维数已经很低（即 4）。假设数据分布是正常的，但这没有经过正式测试。

使用机器学习对响应进行分类。使用多项逻辑回归将参与者的反应分类为四种情绪之一，我们确保测试和训练始终在不同的数据子集中进行。分类通过 10 倍遗漏交叉验证程序进行评估，其中 90% 的数据用于训练多项式模型，并将生成的模型应用于 10% 遗漏数据点。该省略程序重复十次，以便所有数据点都获得预测值。每个折叠中的组都是随机生成的，其限制是在测试数据集中找不到训练数据集中的叙述。在每个折叠中，根据训练数据集，通过尝试前 1, 2, 3, 5, 7, 10, 14, 19, 26, 35, 46, 61, 80, 105, 137 来优化维度数、179、234、305、397、517 和 768 维，其中我们选择了维数错误分类数量最少。然后将训练数据集中找到的优化维数应用于保留的测试数据集。对于仅文字分析，十倍的优化维度平均数为 70，标准差为 14。对于仅评级量表分析，使用的平均维度数为 3.64，标准差为 0.48（即，在大多数情况下使用所有尺寸）。

为了检验假设 3，我们还进行了多元线性回归来预测四个评分量表中每个评分量表的连续值。这些回归模型以与上述多项回归模型相同的方式进行和评估。

词云。词云是由语义 t-test 生成的，该方法允许进行比较，无需逻辑回归所需的参数拟合，即可进行二元分类，其中我们使用二元区分来区分一个词与其他三个词是否属于情感状态州。语义t检验是一种统计方法，通过计算语义相似度分数来计算两组语义表示是否彼此不同。第一步是将一种情绪生成的单词响应的语义表示总结为一种语义表示，然后将该向量归一化为长度。其他三种情绪的话语也是如此概括。然后通过将这两个向量相减并将该向量的长度归一化来计算差异向量。然后，通过取这两个向量之间的点积（即，数学上相当于向量之间角度的余弦）来计算该差异向量与描述每个唯一单词的向量之间的语义相似性。在计算差异向量时实施了 10 倍省略程序，以便该向量不包括

待测词。在进行多重比较的 Bonferroni 校正后，使用 t 检验来确定语义相似性分数是否显着大于零。对四种情绪重复相同的过程，以产生四种不同的词云。该图显示了每种情绪 t 值最高的 25 个单词，其中所有单词都具有统计显着性。

报告摘要

有关研究设计的更多信息，请参阅本文链接的《自然投资组合报告摘要》。

结果

预测分类

与我们的第一个假设一致，结果表明，与第 2 阶段的四个评分量表的总分 (44%) 相比，基于文字响应的情感叙述正确分类的百分比 (64%) 显着更高。非专业组（X(1, 400) = 16.10, p = 0.0001, ϕ = 0.20 [0.10, 0.29]）（见图2）。基于四个评级量表的各个项目（总共 26 个项目，即 PHQ-9 9 个项目、GAD-7 7 个项目、SWLS 5 个项目和 HILS 5 个项目）进行分类的分类准确性较低 (30%) (X (1, 400) = 8.41，p = 0.0037，ϕ = 0.14 [0.05, 0.24]）。

根据我们的第二个假设

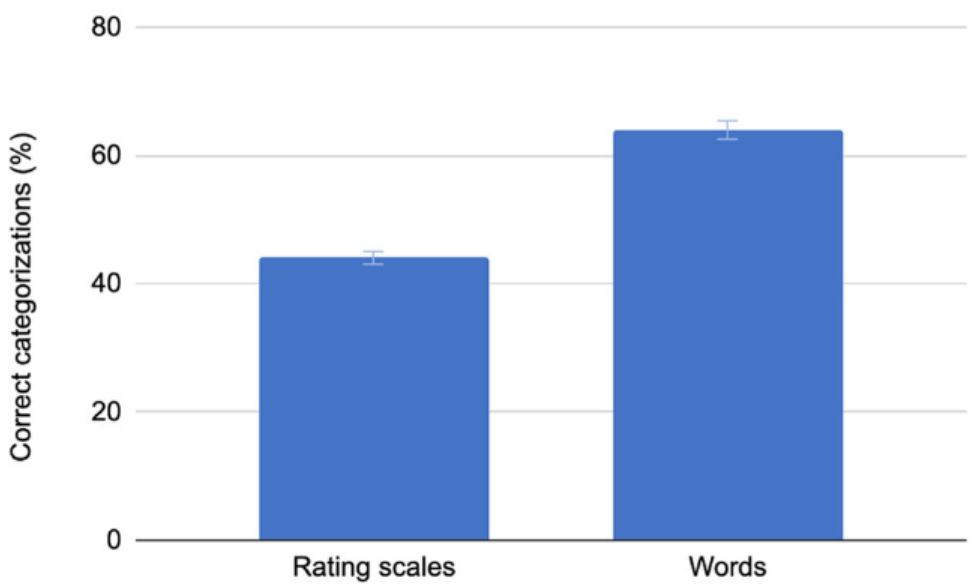

图2|根据评级量表和描述性词语响应进行正确分类。误差线显示平均值的标准差 (N = 348)。

对于非专业组，在第 2 阶段，仅单词反应 (63%) 与评分量表和单词反应组合 (64%) 之间的分类准确性没有统计学上的显着差异 (X(1, 400) = 0.04，p = 0.8355, ϕ = 0.01 [−0.09, 0.11])。与满意度 (X(1, 100) = 5.88, p = 0.0154, ϕ = 0.25 [0.06, 0.42]）和焦虑 (X(1, 100) 情况下的文字反应相比，评级量表的分类准确性较低= 45.63，p < 0.01，ϕ = 0.68 [0.55, 0.77]）（表 2）。评级量表分析始终显示焦虑的准确性较低（即在所有情况下均降低 15%），而语义分析显示所有情绪状态的准确性相当高（第二阶段专业人士的小群体的满意度除外）。这表明基于评级量表的分析在识别焦虑方面特别差。

关于我们的第四个探索性假设，在医疗保健专业组中，单词反应的情绪状态正确分类百分比 (56%) 和评级量表反应 (50%) 之间没有统计学上的显着差异 (X(1, 34) = 0.12，p = 0.7260，ϕ = 0.09 [−0.25，0.41]）。表 2 提供了与所有数据点的评分量表和语言测量的准确分类百分比相关的结果的详细分类和综合视图。

准确度和精密度

与评级量表相比，当对单词响应进行分类时，准确度测量（即正确的正分类数量加上正确的负分类数量除以分类总数）更高 (X(1, 434) = 4.04， p = 0.0445, ϕ = 0.10 [0.00, 0.19]）（见表 3）。类似地，与评级量表相比，对单词响应进行的分类的精度测量（即正确的正面分类的数量除以正面预测的数量）也更高 (X(1, 434) = 20.88, p < 0.001, ϕ = 0.22 [0.13, 0.31]）。表 3 列出了每种情绪状态的精确度和准确度值的细目，详细描述了每个类别特定的绩效指标。

混淆矩阵和相关矩阵

图 3 显示了混淆矩阵，即模型在第一阶段的情绪状态下进行预测的次数。对于基于评分量表（表格上部）的预测，大多数错误发生在模型预测抑郁症时，但正确答案是焦虑（N = 64）。根据词语，错误分布更加均匀，当模型预测满意度时错误数量最多，但正确状态是和谐（N = 24）。

表 4 显示了多项估计系数之间的 Pearson 相关性。与基于单词的模型 (0.07 < r < 0.47) 相比，评级量表模型 (0.68 < r < 0.92) 的所有绝对相关值都较大，表明评级量表模型存在更大的混乱。

表 2 |正确分类分为阶段、模式、情绪状态

相位（-P/P）	模型	全部 （%）	和谐度（%）	满意 （%）	沮丧 （%）	焦虑 （%）	N
1	RS	39	46	21	65	15	297
1	字	70	71	58	76	75	297
1	单词+RS	67	62	60	71	73	297
2(-P)	RS	44	38	31	88	02	400
2(-P)	字	63	56	55	74	66	400
2(-P)	单词+RS	64	57	58	75	65	400
2(P)	RS	50	60	20	82	12	34
2(P)	字	56	70	20	45	75	34
2(P)	单词+RS	59	70	20	55	75	34
1 & 2(-P)	RS	42	43	27	78	08	731
1 & 2(-P)	字	66	63	55	73	70	731
1 & 2(-P)	单词+RS	65	60	58	72	69	731

该表显示了非专业人士 (-P) 和专业人士 (P) 的正确分类百分比 (N(正确)/N (总计)*100，分为阶段 1 和阶段 2，以及预测的分类是否基于评分量表 (RS) 或文字反应 (Words)。

评估者间协议

表 5 显示了基于评分量表或语言内容的评分者间协议，包括第二阶段非专业人士的数据以及两个阶段非专业人士的综合数据。评分量表的分类一致性并不显着高于第二阶段非专业数据中的单词 (X(1, 263) = 3.50, p = 0.062, ϕ = 0.12 [−0.00, 0.24])，也不对于整个数据集（X(1, 721) = 0.12, p = 0.724, ϕ = 0.01 [−0.06, 0.09]）。

通过匹配阶段 1 和阶段 2 中的叙述并计算 HILS、SWILS、PHQ-9 和 GAD-7 测量之间的皮尔逊相关性来评估第三个假设。在这里，我们将这些度量的评分量表得分与这些评分量表的相应描述性单词预测进行了比较。描述性单词预测基于方法部分中所述的多元线性回归。表 6 中的结果表明，描述性词语响应与相应的评分量表测量相比具有显着更高的相关性（使用 Fisher 的 r 到 z 变换），支持假设 3。

分类词云

生成词云来展示具有统计显着性 t 值的单词，阐明哪些单词最能代表四种心理状态（见图 4）。表示抑郁的词语包括悲伤和沮丧，表示焦虑的词语包括焦虑、担心和紧张。表示和谐和满足的词语包括“快乐”和“满足”，而“平静”对于和谐更重要，而“希望”对于满足感更重要。请注意，该图主要展示基于情感的词语，其中受访者主要使用此类术语，而词云反映了这种数据驱动的方法。然而，鉴于参与者收到提示，也可以使用 QCLA 方法来调查症状。例如，请参阅 Kjell 及其同事如何研究躯体症状。本研究评估了 QCLA 如何捕获与重度抑郁症和广泛性焦虑症相关的主要和次要症状相关的项目的功效。

讨论

这项研究表明，平均而言，与使用四种标准化评分量表相比，采用基于单个开放式问题和五词响应格式的计算方法，在对参与者生成的描述情绪状态的叙述进行分类方面可以产生更高的准确性。这一发现具有重大意义，因为它表明开放式描述性的基于单词的反应可能比通常用于心理健康评估的评级量表具有更高的有效性。此外，效果大小具有统计显着性，其中单词响应总体正确分类的百分比为 64%，而评级量表的正确分类百分比为 44%，也就是说，正确分类的情绪状态百分比差异为 20%。

这项研究的结果与其他研究一致，表明计算语言评估与和谐和满意度评分量表产生非常强的相关性（例如，r = 0.84），可以与重测可靠性和项目间相关性的理论极限相媲美。除了语言评估之外，来自相关领域（例如面部表情和合作行为）的证据表明，与传统的评分量表相比，语言反应可能具有更高的有效性。例如，Kjell 及其同事探索了使用词语反应和评分量表来识别面部表情，结果显示词语反应具有一定的优势（4%），尽管明显小于我们当前研究中观察到的优势。同样，在另一项关于合作行为的研究中，与评级量表相反，词语反应预测了参与者在一次给予一些困境游戏中的合作行为，其中参与者完成了评级量表（HILS 和 SWLS）或和谐与满意度的词语反应测量在执行 GSDG 之前（有关详细信息，请参阅 Van Lange & Kuhlman）。在本研究中，我们认为单词反应的分类优于评级量表的论点源于结果变量（叙述的真实事实）与评级量表的独立性，这与早期的定量内容分析（QCLA）研究有所不同。通常通过与评级量表的相关性来验证。这强调了我们当前的研究结果在现有文献的更广泛背景下的相关性，强调了采用具有开放式问题的计算方法来评估情绪状态的优势。

当分类准确性在每个单独的情绪状态之间细分时，发现了一些差异。在第二阶段（但不是第一阶段），焦虑的评定量表分类始终表现出较低的准确度。这意味着评级量表模型存在潜在偏差，因为它倾向于为焦虑分配低分，这可能会以在分类过程中更好地捕捉其他情绪为代价。

表 3 |准确度和精确度测量

模型	措施	和谐	满意	沮丧	焦虑
RS	准确性	76	66	68	79
RS	精确	43	35	49	29
字	准确性	82	79	80	86
字	精确	60	62	64	65

该表显示了第 1 阶段验证的第 2 阶段 (N = 434) 的和谐、满意度、抑郁和焦虑的评分量表 (RS) 和词语响应 (Words) 的准确性和精确度测量（以百分比表示）。

图3|混淆矩阵。热图显示与第 2 阶段 (N = 369) 中多项模型的情绪状态预测数量（以列表示）同时出现的经验情绪状态数量（以行表示）。深绿色区域表示共现率最高，而深红色区域表示共现率最低。使用以下缩写：和谐 (Har)、满意度 (Satis)、抑郁 (Dep) 和焦虑 (Anx)。

表 4 |相关矩阵

	评定量表			字		
	Har.	满足。	Dep.	Har.	满足。	Dep.
萨蒂斯	0.87 [0.85, 0.88]			−0.07[−0.14, 0.01]		
Dep	−0.92[−0.93,−0.91]	−0.95 [−0.96, −0.95]		−0.47[−0.53,−0.41]	−0.42[−0.48,−0.36]	
Anx	−0.88[−0.89,−0.86]	−0.75 [−0.78, −0.71]	0.68[0.64,0.72]	−0.41[−0.47,−0.35]	−0.42[−0.47,−0.35]	−0.21[−0.27,−0.13]

该表显示了第 2 阶段多项系数估计值的 Pearson 相关分数 (N = 434)。方括号表示 95% 置信区间。使用以下缩写：和谐 (Har)、满意度 (Satis)、抑郁 (Dep) 和焦虑 (Anx)。

表 5 |评估者间协议

阶段	模型	协议	正确的	N
2(-P)	RS	90	79	263
2(-P)	字	82	65	263
1 & 2 (-P)	RS	83	69	721
1 & 2 (-P)	字	82	66	721

该表显示了不同评估者模型和阶段的相同叙述的所有成对分类之间的一致性百分比。各列显示阶段、模型、一致性（即相同分类的百分比）、正确性（即正确分类的一致性百分比）以及成对比较的数量。使用以下缩写：非专业人士 (-P) 和评级量表 (RS)。

表 6 |第一阶段和第二阶段措施与叙述相匹配的相关性

	评定量表	字	p
HILS	0.19 [0.09, 0.28]	0.76 [0.09, 0.28]	<0.001
SWLS	0.47 [0.39, 0.54]	0.76 [0.39, 0.54]	<0.001
PHQ-9	0.48 [0.41, 0.55]	0.77 [0.41, 0.55]	<0.001
GAD-7	0.17 [0.07, 0.26]	0.55[0.07,0.26]	<0.001

该表显示了与叙述身份匹配的 HILS、SWLS、PHQ-9 和 GAD-7 的第 1 阶段和第 2 阶段测量之间的 Pearson 相关性 (N = 434)。方括号表示 95% 置信区间。第二列显示这些评级量表 (RS) 的值，第三列显示这些评级量表的描述性单词预测。p 值表示使用 Fisher 的 r 到 z 变换时两个相关性是否不同。

通过查看混淆矩阵发现了更多信息，其中与单词响应相比，第二阶段评级量表的错误预测数量更高。与基于单词的模型相比，评级量表模型的估计系数之间的皮尔逊相关性要高得多，这表明单词反应可以更好地区分情绪状态。这与之前的研究一致，表明词图比相关概念之间的评分量表更好地区分。

第二阶段数据的一部分包括目前在医疗保健相关行业就业的个人。该亚组被纳入研究是因为预期他们对抑郁、焦虑、和谐和满意度的定义和评估有更深入的了解。尽管如此，该组并没有使用评分量表获得更高的分类准确性，并且他们的评分量表正确分类的标称值小于单词响应的标称值。名义上，这些数据看起来与更大的非专业对照参与者数据相似。 然而，专业受试者的数量太少（N = 34），无法得出任何明确的结论。

除了与评级量表相比具有更高的有效性之外，基于语言的心理健康评估的计算方法还有其他几个优点。首先，语言是人们交流心理状态的自然方式。人们更喜欢用语言而不是评级量表来传达心理健康，因为他们发现语言更精确和详细，并且他们更喜欢在与临床医生沟通时使用语言，尽管评级量表被认为更容易和更快。其次，开放式语言反应可以对参与者的心理健康状况进行特殊描述，从而为以人为本的医疗保健提供机会。这与衡量研究定义的固定结构的评级量表非常不同，患者无法添加以人为本的观点。第三，拟议的语言措施时间较短，因此实施起来很快。就目前情况而言，可以通过要求个人提供五个描述其情绪状态的单词来进行简短的对话。相比之下，在同一时间范围内完成单一评级量表将构成挑战。

图4|显示表示四种情绪状态的单词的词云。词云从左到右依次显示和谐（黑色）、满意（红色）、抑郁（深蓝色）和焦虑（浅蓝色）。数据集中的单词数为N = 3787。图中的单词数为N = 100。单词的大小与词频成正比，最具指示性的单词位于图的中心。请注意，有些词（例如，快乐）表示不止一种情绪。另外，请注意，单词是由阶段 1 和阶段 2 中的参与者生成的，因此参与者在阶段 1 的特定条件下不允许使用的单词（例如，抑郁状态下的“抑郁”）可以由阶段 2 生成。 2 名参与者，或处于其他状况的第 1 阶段参与者（即，在第一阶段）的“焦虑”状态中允许使用“抑郁”。该图像是使用 SemanticExcel 创建的。

我们的研究表明，在使用标准化评分量表的环境中使用 QCLA 等语义测量的潜在好处。使用文字而不是数字可以提供一种更加以人为本的方法，这可以帮助患者感到更容易被理解，减少人格解体。例如，结构化电子健康记录中很少提供非结构化临床记录，而允许患者用自己的语言而不是一维封闭格式回答健康相关问题的响应格式提供了许多机会。所提出的方法可以提高诊断准确性和治疗计划，最终改善治疗结果。 与传统的评分量表方法不同，开放式问题可能不太可能强加社会所期望的默认反应或暗示可能的症状，例如“您放松有困难吗？”(GAD-7)并且它们可以说是一种更自然的表达形式。数字化的情感支持工具正在激增，学术界最近发现社交媒体文本挖掘研究有所增加。 NLP 方法可以同时有效评估数百个预测因素，并提出经济敏感的解决方案，可以预测未来的结果，例如自杀的实现、尝试或构思。社交媒体文本挖掘以书面自传体叙述个人心理状态为主要沟通方式，为预防性筛查和检测人群中的精神疾病（特别是在前驱阶段）以及评估不同人群的风险提供了另一种选择。整个心理健康问题。

医疗保健相关数据可以很好地洞察我们社区的健康状况。然而，在文本分析中使用 NLP 的主要缺点之一是出于心理健康目的而扫描整个人群的隐私和道德观念。虽然 NLP 算法确实有潜力分析大量文本数据并检测心理健康问题的模式，但如果没有适当的道德考虑，这样做可能会被视为侵入性的，并可能会给个人和整个社会带来负面后果。为了确保此类算法的使用合乎道德，制定严格的准则和法规非常重要。例如，任何使用 NLP 扫描文本以了解心理健康状况的计划或举措都应该有关于数据收集、共享、存储和使用的明确协议，以保护个人隐私。因此，在临床环境中实施文本分析将有办法确保尊重隐私和道德考虑。随着《通用数据保护条例》的出台以及减少社会偏见的努力，例如

为了使词嵌入更加中立，临床环境中隐私和公正的智力进步之间的平衡正在进行中。因此，我们相信 QCLA 作为一种纳入临床环境的工具具有巨大的潜力。

局限性

术语“焦虑”和“抑郁”在临床环境中的使用与外行不同。临床医生使用 DSM-V 定义来评估精神健康障碍，而撰写有关抑郁或焦虑事件的参与者可能会使用对该术语的更广泛理解，这可能适合也可能不适合临床医生的评估。因此，从临床问卷中获得的预测与描述性反应的对比的公平性是一个有待讨论的问题。然而，与本研究一致的是，接受临床抑郁症诊断评估的个体经常用自己的话表达自己的想法。将我们的样本与诊断为临床抑郁症的参与者进行比较，可以在未来的研究中提供见解，并有可能详细解决这个问题。用于心理健康评估的语义方法或评级量表的选择取决于多种因素，包括评估的背景和目的。如果目标是获得可以与以前的文献进行比较的特定心理健康方面的简单标准化数值评估，则封闭式评级量表可能是合适的。然而，开放式问题也可以标准化，以比较来自不同样本的个体。当旨在涵盖更广泛的情感和经历时，特别是当受访者对情感有不同的解释时，开放式回答可以提供有价值的定性观点。其目的不是用一种方法替代另一种方法，而是用语义测量来补充传统量表，例如，协助初步诊断。

为了进一步审视个人对抑郁和焦虑等病症的重要性的看法与临床医生认为相关的内容之间的差异，重要的是要利用个人的主观感受与所需的更全面、客观的评估之间的区别在临床环境中。如果将 NLP 应用到临床实践中，考虑根据临床评估的特定目的（例如预防性筛查、诊断或跟踪治疗进展）调整模型可能会有所帮助。本研究没有按性别、国籍或教育程度对 NLP 预测进行分层，以调查结果模型中可能存在的差异。认识到需要进行定制以解决人口因素和心理健康状态的不同维度（包括认知、生理和行为），可以增强评估工具在现实世界应用中的相关性和有效性。通过评定量表观察到的误诊和差异的持续存在，在黑人患者的抑郁症或精神分裂症诊断等情况下尤其明显，绝不能转移到使用 NLP 分析的开放式描述性词语反应，因此需要彻底关注。

另一个担忧是，参与者被要求报告过去的情感事件，以便他们当时的情感体验可能与他们在撰写叙述时的感受有所不同。如果个人反复回忆相同的经历，使他们在记忆重新巩固过程中容易受到记忆扭曲和改变，这种担忧可能会变得特别值得注意。准确地回忆和描述情感经历对于个人来说可能会变得具有挑战性，因为可以合理地假设个人可能在不同的时间点有不同的情绪并且处于不同的环境中，这可能会引入数据质量问题。在第一阶段，一组独立的参与者对与四种情绪之一相关的自我经历事件进行了叙述；随后，第二阶段的参与者阅读了这些内容，用五个词进行了描述，并使用通常用于测量相应情绪的评级量表进行了解释。因此，第二阶段数据分类的成功或失败取决于参与者如何解释第一阶段数据，而第一阶段的参与者可能对如何理解这些情绪有不同的看法或解释，与评级量表的一般构建方式相比。尽管我们发现这种可能性不太可能，因为抑郁、焦虑、满足和和谐是常用的概念，但仅仅依靠自我报告的情绪测量可能无法提供完整或准确的情绪体验图景。

我们承认依赖过去的记忆会带来潜在的挑战，例如回忆偏差，但它符合我们将描述性词语响应的预测能力与相应的评分量表进行比较的目标。要求参与者分享最近的生活经历并根据他们当前的感受使用问卷调查将是另一种有效的方法。然而，我们的具体研究设计旨在建立一个真实数据集，参考自传语境中过去的情绪状态回忆。它优先考虑建立历史基线，以便对一段时间内的情感体验进行更全面的分析。尽管如此，这一方面为前瞻性研究提供了一条有趣的途径，可以探索情绪的操纵，并且可以训练专门的模型来识别和分类特定的情绪等因素。此外，受试者内研究可能是解释重复测量的另一种可能性。

最后，参与者只能提供单个单词的回答，这一方法在之前的研究中得到了验证。重要的是，这可能会导致对相关症状的不完全理解。例如，当个体被提示使用描述性词语表达抑郁发作期间的情绪状态时，存在忽视相关信息的风险。描述性的使用可能会限制所表达的各种状态或症状，有些需要多个词才能准确描述，例如“吃得少”、“无法集中注意力”或“思考”等短语。关于伤害我自己。”考虑到 BERT 模型的适应性，可能有可能重新评估这一约束，而该模型不一定受此限制的约束。 BERT 通过理解这两个短语甚至容纳非单词和拼写错误来展示灵活性。探索比较单个单词和短语的有效性可能会引起未来研究的兴趣，因为比较这两种方法的研究有限。

未来的工作

本研究的主要重点是情绪的分类。然而，从临床的角度来看，一个可能引起额外兴趣的研究问题是探索应用类似方法将详细描述参与者心理健康的叙述分类为特定心理健康诊断的可行性。进行这样的研究对心理健康状况的评估和诊断具有直接影响，为更细致和更全面地了解个人的心理健康提供了潜力。这种研究途径可以为心理健康领域有效诊断工具和干预措施的开发提供宝贵的见解。

选择 BERT 作为我们研究的语言模型是基于它在 NLP 中的广泛使用和已建立的标准化。此外，这一选择还受到与研究小组内其他研究工作的一致性的影响。虽然 MentalBERT 等其他语言模型可能是合理的替代方案，但在本研究中使用 BERT 的决定植根于其在各种 NLP 任务中经过验证的性能和多功能性。未来的研究可以从探索不同的语言模型中受益，以增强我们分析的全面性。

结论

QCLA 作为增强识别的工具具有巨大的前景，可能会改善各种精神健康疾病的治疗。后续涉及更多、更多样化人群的研究有可能更深入地研究 QCLA 的功效及其更广泛的适用性。本研究表明，描述性词语反应为对心理结构进行更精确和自然的评估提供了机会，我们的研究结果显示了前所未有的准确性

与一维数字评分量表相比，对情绪状态的叙述进行分类。这在心理健康领域具有重要意义，因为定量评估中通常使用标准化评分量表，而不是开放式语言反应。因此，语义测量可能构成情绪状态补充评估方法的基石方法。我们的工作表明，当使用 NLP 的计算方法进行分析时，对开放式问题进行回答的定向问题可用于评估心理结构，其有效性和可靠性比标准化数字评分量表更高。

数据可用性

最初的 Qualtrics 调查、标准化评级量表的确切措辞描述以及匿名参与者数据已在 https://osf.io/gdkcb 上公开。

代码可用性

通讯作者可以根据要求提供定制的 MATLAB 计算机代码，使读者能够复制已发布的结果。没有使用代码来预处理数据。使用 https://osf.io/gdkcb 中的数据，无需代码即可在语义Excel.com 中重现所做的分析。数据可以复制粘贴到软件中。

收稿日期：2023 年 7 月 31 日；接受日期：2024 年 5 月 3 日；
Published online: 23 May 2024

参考

1. Flake, J. K.、Pek, J. 和 Hehman, E. 在社会和人格研究中构建验证：当前实践和建议。苏克。

   心理。个人的。科学。 8, 370–378 (2017)。

2. Diener, E.、Emmons, R.、Larsen, R. 和 Griffin, S. 生活满意度量表。 J.佩尔斯。评估。 49, 71–75 (1985)。

3. Newmann, F. 研究新闻和评论：关于“语义学、心理测量学和评估改革：仔细审视‘真实’评估”的意见交换。教育。资源。 27, 19–22 (1998)。

4. 美国精神病学协会。精神疾病诊断和统计手册 (DSM-5®) 第 5 版。 xliv，947（美国精神病学出版公司，美国弗吉尼亚州阿灵顿，2013 年）。 https://doi.org/10.1176/appi.books.9780890425596。

5. Allsopp, K.、Read, J.、Corcoran, R. 和 Kinderman, P. 精神病学诊断分类的异质性。精神病学研究 279, 15–22 (2019)。

6. Clark, L. A.、Cuthbert, B.、Lewis-Fernández, R.、Narrow, W. E. 和 Reed, G. M. 理解和分类精神障碍的三种方法：ICD-11、DSM-5 和国家心理健康研究所的研究领域标准（RDoC）。心理。科学。公共利益 18, 72–145 (2017)。

7. Levitt, H. M. 特别部分简介：质疑既定的定性方法和假设。合格。心理。 8, 359–364 (2021)。

8. 勒格拉兹，A.等人。心理健康中的机器学习和自然语言处理：系统评价。 J. Med。互联网资源。 23、e15708 (2021)。

9. 辛南伯格，L.等人。 Twitter 作为健康研究工具：系统评价。是。 J. 公共卫生 107，e1-e8 (2017)。

10. Skaik, R. 和 Inkpen, D. 使用社交媒体进行心理健康监测：综述。 ACM 计算。幸存者。 CSUR 53, 1–31 (2020)。

11. Liu, X., Shin, H. 和 Burns, A. C. 检查奢侈品牌社交媒体营销对客户参与度的影响：使用大数据分析和自然语言处理。 J.巴士。资源。 125, 815–826 (2021)。

12. Fisher, I. E.、Garnsey, M. R. 和 Hughes, M. E. 会计、审计和金融中的自然语言处理：文献综合与未来研究路线图。英特尔。系统。帐户。

    财务管理。 23, 157–214 (2016)。

13. Allesøe，R.L. 等人。通过生成深度学习模型发现 2 型糖尿病的药物组学关联。纳特。生物技术。 41, 399–408 (2023)。

14. 卡斯特罗，V.M.等人。双相情感障碍病例和对照的电子健康记录表型验证。是。 J. 精神病学 172, 363–372 (2015)。

15. 纳瓦罗，M.C. 等人。对预测青春期或成年早期自杀企图的早期生活因素的机器学习评估。 JAMA 网络。打开 4，e211450–e211450 (2021)。

16. Kolanu, N.、Brown, A. S.、Beech, A.、Center, J. 和 White, C. P. OR29-02 放射学报告的自然语言处理可改善骨折患者的识别。 J.内分泌。苏克。 4、OR29–02（2020）。

17. Levanti，D.等人。美国 7 个主要城市新冠疫情居家期间推特上的抑郁和焦虑情况。 AJPM 焦点 2, 100062 (2023)。

18. Jin, H. & Wang, S. 基于语音的用户身体和情感特征的确定。美国专利 US 10,096,319 (2018)。

19. Gaonkar, B.、Cook, K. 和 Macyszyn, L. 由于训练人工智能的偏见而产生的道德问题。医疗保健和数据共享中的算法作为潜在的解决方案。人工智能伦理学 J. 1, 1–9 (2020)。

20. Sidey-Gibbons, J. A. M. 和 Sidey-Gibbons, C. J. 医学中的机器学习：实用介绍。 BMC 医学。资源。方法。 19, 64 (2019)。

21. 艾希施塔特，J.C. 等人。 Facebook 语言预测医疗记录中的抑郁症。过程。国家。阿卡德。科学。 115, 11203–11208 (2018)。

22. Guntuku, S. C.、Yaden, D. B.、Kern, M. L.、Ungar, L. H. 和 Eichstaedt, J. C. 在社交媒体上检测抑郁症和精神疾病：综合审查。电流。意见。行为。科学。 18, 43–49 (2017)。

23. Seabrook, E. M.、Kern, M. L. 和 Rickard, N.S. 社交网站、抑郁和焦虑：系统评价。 JMIR 门特。健康 3，e5842 (2016)。

24. Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向 Transformer 的预训练。在

    过程。 2019 年计算语言学协会北美分会会议：人类语言技术，1（长论文和短论文）4171–4186（计算语言学协会，明尼苏达州明尼阿波利斯，2019 年）。 https://doi.org/10.18653/v1/N19-1423。

25. 张 H.、卢 A. X.、阿卜杜拉 M.、麦克德莫特 M. 和 Ghassemi, M. 伤人的话：量化临床语境词中的偏差

    嵌入。摘自 ACM 健康、推理和学习会议 (CHIL '20) 会议记录，110–120（计算机协会，美国纽约州纽约市，2020 年）。

26. Rogers, A.、Kovaleva, O. 和 Rumshisky, A. BERT 学入门：我们对 BERT 工作原理的了解。跨。副教授。计算。语言学家。 8, 842–866 (2021)。

27. Li, Y., Masitah, A. & Hills, T. T. 情感回忆任务：并置回忆和基于识别的情感量表。 J.Exp。心理。学习。

    内存。认知。 46, 1782–1794 (2020)。

28. Fatima, A.、Li, Y.、Hills, T. T. 和 Stella, M. DASentimental：通过情绪回忆、认知网络和机器学习检测文本中的抑郁、焦虑和压力。大数据认知。计算 5, 77 (2021)。

29. Kjell, O. N.、Kjell, K.、Garcia, D. 和 Sikström, S. 语义测量：使用自然语言处理来测量、区分和描述心理构造。心理。方法 24, 92 (2019)。

30. Kjell, O.、Daukantaitė, D. 和 Sikström, S. 对生活和谐的计算语言评估（而不是对生活或评级量表的满意度）与合作行为相关。正面。心理。 12、601679 (2021)。

31. Kjell, O. N.、Sikström, S.、Kjell, K. 和 Schwartz, H. A. 使用基于人工智能的转换器分析自然语言，预测传统主观幸福感测量的准确性接近理论上限。科学。报告 12, 1–9 (2022)。

32.Stochl，J.等人。关于维度、测量不变性和
PHQ-9 和 GAD-7 总分的适用性。评估 29, 355–366 (2022)。

33. Kjell, K.、Johnsson, P. 和 Sikström, S. 使用人工智能分析的自由生成的单词响应可预测自我报告的抑郁、焦虑和担忧症状。正面。心理。 12、602581（2021）。

34. Sikström, S.、Pålsson Höök, A. 和 Kjell, O。精确的语言反应与简单的评分量表——将受访者的观点与临床医生对受访者观点的信念进行比较。 PLOS ONE 18，e0267995 (2023)。

35. Kroenke, K.、Spitzer, R. L. 和 Williams, J. B. PHQ-9：简短抑郁症严重程度测量的有效性。 J. Gen. 实习生。医学。 16, 606–613 (2001)。

36. Spitzer, R. L.、Kroenke, K.、Williams, J. B. W. 和 Löwe, B. 评估广泛性焦虑症的简要措施：GAD-7。拱。
实习生。医学。 166, 1092–1097 (2006)。

37. Kjell, O.、Daukantaitė, D.、Hefferon, K. 和 Sikström, S。生活量表的和谐补充了生活量表的满意度：扩展了主观幸福感认知成分的概念化。苏克。印度语。资源。 126, 893–919 (2016)。

38. Sikström, S., Kjell, O. N. E. & Kjell, K. SemanticExcel.com：基于自然的文本数据统计分析在线软件
语言处理。统计语义学博士：方法和
应用（Sikström, S. 和 Garcia, D. 编辑）87–103（Springer International Publishing，Cham，2020）。 https://doi.org/10.1007/9783-030-37250-7_6。

39. Stone, M. 交叉验证和多项式预测。生物计量学 61, 509–515 (1974)。

40. Kjell, O. N.、Kjell, K.、Garcia, D. 和 Sikström, S. 预测和语义训练量表：检查对抑郁和担忧的语义反应与相应的评分量表之间的关系。统计语义学 73–86（Springer，2020）。

41. Van Lange, P. A. 和 Kuhlman, D. M. 社会价值取向以及伴侣诚实和智慧的印象：力量与道德效应的检验。 J.佩尔斯。苏克。心理。 67, 126 (1994)。

42. 汉，S.等人。使用基于深度学习的自然语言处理从非结构化电子健康记录中对健康的社会决定因素进行分类。 J.生物医学。通知。 127, 103984 (2022)。

43. Ford, E.、Shepherd, S.、Jones, K. 和 Hassan, L. 针对健康研究社交媒体文本挖掘的道德框架：系统评价。正面。数字。健康 2, 592237 (2021)。

44. 卡拉菲拉基斯，E.等人。与疫苗接种相关的社交媒体监测方法：系统范围界定审查。 JMIR 公共卫生调查 7，e17149 (2021)。

45. Ahn, J. 和 Oh, A. 减轻 BERT 中语言相关的种族偏见。
2021 年自然语言处理经验方法会议论文集，533-549（计算语言学协会，在线和多米尼加共和国蓬塔卡纳，2021 年）。

46. Bartl, M.、Nissim, M. 和 Gatt, A。揭露情境刻板印象：衡量和减轻 BERT 的性别偏见。第二届自然语言处理中的性别偏见研讨会论文集，1-16（计算语言学协会，西班牙巴塞罗那，2020 年）。

47. Mozafari, M.、Farahbakhsh, R. 和 Crespi, N. 一种基于 BERT 的迁移学习方法，用于在线社交媒体中的仇恨言论检测。复杂网络及其应用 VIII（编：Cherifi, H.、Gaito, S.、Mendes, J. F.、Moro, E. 和 Rocha, L. M.）928–940（Springer）
国际出版，Cham，2020）。 https://doi.org/10.1007/9783-030-36687-2_77。

致谢

这项工作得到了瑞典研究委员会 Grant VR 2021 – 04627 的支持。资助者在研究设计、数据收集、分析、出版决定或手稿准备中没有任何作用。我们感谢所有参与这项研究的参与者。

作者贡献

S.S.：概念化、方法论、软件、数据解释、形式分析、写作审查和编辑、项目管理、监督和资金获取。 I.V.：概念化、方法论、调查、数据解释、形式分析、撰写初稿准备、审查和编辑、可视化。 I.K.：方法论、调查、写作——原始草案准备。 N.E.：方法论、调查、写作——初稿准备。

资金

由隆德大学提供的开放获取资金。

利益竞争

第一作者声明了以下相互竞争的利益：S.S. 是一家名为 Ablemind 的初创公司的联合创始人和股东，该公司使用计算语言评估来诊断心理健康问题。其他作者声明没有竞争利益。本文中提到的任何公司（例如 Amazon 或 Alexa）与作者或 Ablemind 之间没有任何联系。

附加信息

补充信息 在线版本包含补充材料，网址为 https://doi.org/10.1038/s44271-024-00097-2。

信函和材料请求应发送至 Sverker Sikström。

同行评审信息通信心理学感谢玛塔
Maslej 和其他匿名审稿人对这项工作的同行评审做出的贡献。主要处理编辑：Jonna Vuoskoski 和 Jennifer Bellingtier。同行评审文件可用。