计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：机器学习在社交媒体用户自杀意念检测中的应用研究综述
作者：蒙秀扬，王世屹，李渡渡，王春玲
网络首发日期：2024-08-29
引用格式：蒙秀扬，王世屹，李渡渡，王春玲．机器学习在社交媒体用户自杀意念检测中的应用研究综述[J/OL]．计算机科学与探索．
https://link.cnki.net/urlid/11.5602.TP.20240829.1525.017

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2405086

机器学习在社交媒体用户自杀意念检测中的应用研究综述

蒙秀扬[1,2]，王世屹[3]，李渡渡[1,2]，王春玲[1,2+]

1. 北京林业大学 信息学院，北京 100083
2. 国家林业草原林业智能信息处理工程技术研究中心，北京 100083
3. 中央民族大学 信息工程学院，北京 100081
+ 通信作者 E-mail: wangchl@bjfu.edu.cn

摘要：自杀是严峻的社会问题，是突出的全球性公共卫生挑战,也是全球死亡的重要因素之一。近年，互联网迅猛发展，社交媒体平台成为人类发布情感甚至是自杀意念、企图和行为的崭新阵地，使其成为自杀意念检测的重要数据平台和关键评估依据。随着人工智能技术的兴起，关于机器学习在社交媒体用户自杀意念的检测中的应用研究已然成为热点。但在国内，该领域相关研究较为匮乏，且尚未形成完整体系。为系统梳理其研究现状及发展脉络，对机器学习技术赋能自杀意念检测的研究进行了全面总结，是近三年来国内第一篇关于此领域的综述。首先，概述了自杀意念检测的定义、流程、常见方法及评价指标，总结了当下自杀意念检测任务中常用的数据集和现有特征工程及其技术。其次，分别从传统的机器学习和深度学习的角度对自杀意念检测进行了系统总结，对比分析了每种方法的关键技术、核心思想及优缺点。此外，归纳了当前该领域中亟待解决的问题及创新解决方法,特别介绍了 ChatGPT 等大语言模型、多模态模型在该领域的应用。最后，讨论了机器学习在社交媒体自杀意念检测中的应用研究中的局限性，并提出了未来的研究方向，以期进一步推动形成数据驱动、人机协同、跨学科融合、跨文化畛域的数智化自杀意念检测新范式，以期为相关领域的研究人员提供借鉴和参考。

关键词：自杀意念检测；社交媒体；机器学习；深度学习；特征提取

文献标志码：A **中图分类号：**TP391

Review on Application of Machine Learning in Detecting Suicidal Ideation for Social Media Users

MENG Xiuyang[1,2], WANG Shiyi[3], LI Dudu[1,2], WANG Chunling[1,2+]

1. School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China
2. Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China
3. School of Information Engineering, Minzu University of China, Beijing 100081, China

Abstract: Suicide constitutes a grave societal quandary, an eminent worldwide public health predicament, and a pivotal determinant of global mortality. In recent years, with the rapid development of the Internet, social media have

基金项目：中央高校优秀青年团队项目（QNTD202308）。
This work was supported by the Outstanding Youth Team Project of Central Universities of China (QNTD202308).

emerged as a novel domain for individuals to express their emotions, including suicidal ideation, attempts, and behaviors. Consequently, these platforms have evolved into crucial data repositories and essential assessment criteria for detecting suicidal ideation. With the advent of artificial intelligence technology, the utilization of machine learning in detecting suicidal ideation among social media users has emerged as a scintillating subject. The field in China, however, lacks sufficient research and has yet to establish a comprehensive system. To systematically review the research status and development context of suicide ideation detection, this paper presents a systematically summary of machine learning technology's application in empowering suicide ideation detection, marking the first review in this field conducted in China over the past three years. Firstly, this paper provides an overview of the definition, process, commonly employed methods, and evaluation indicators for detecting suicidal ideation. Secondly, this paper provides a comprehensive overview of suicide ideation detection techniques, encompassing both traditional machine learning and deep learning approaches. The key methodologies, fundamental concepts, merits, and limitations of each method are thoroughly compared and analyzed. Furthermore, the urgent issues and innovative solutions in this field are summarized, with a particular focus on the application of large language models such as ChatGPT and multi-modal models. Finally, the limitations of machine learning in the application research of suicide ideation detection on social media are comprehensively discussed, and future research directions are proposed, in order to further promote the formation of a new paradigm of data-driven, human-computer collaboration, interdisciplinary integration, and cross-cultural domain of suicide ideation detection, so as to provide reference and reference for researchers in related fields.

Key words: Suicidal ideation detection; Social media; Machine learning; Deep learning; Feature extraction

自杀是一项全球性的重大公共卫生危机，也是世界范围内最主要的死亡原因之一，每年有超过 70 多万人死于自杀[1-2]。由于抑郁症、焦虑症和创伤后应激障碍等精神疾病带来的困扰，这些患病人群存在明显的自杀意念，具有高度的自杀风险，导致近些年自杀案件数量持续上升[3]。因此，准确检测用户的自杀意念及自杀风险人群至关重要。早期检测和识别出自杀意念被认为是预防潜在自杀最有效的方法[4]。传统的方法主要通过临床手段与患者进行交互，例如自陈式问卷、自杀量表和临床面谈等[5-7]。然而，受个人隐私意识、疾病耻辱感和社会污名化等影响，多数患者不愿意主动向专业人士寻求帮助，治疗方式相对被动，因此难以执行长期或大规模的检测。随着医学技术水平的提升，功能磁共振成像（functional magnetic resonance imaging，fMRI）、正电子发射型计算机断层显像（positron emission computed tomography，PET）和一些可穿戴设备可以辅助检测，尽管这些方法专业性强，但价格昂贵，受众范围有限。此外，目前医疗和人力资源不足，难以做到实时检测与治疗。因此，亟需一种更主动、更经济和适用于大规模人群的实时自动化检测方法。

随着人类情感表达方式的嬗变，社交媒体平台成为用户抒发情感、痛苦甚至是自杀企图和行动的新渠道[1,4]。多项研究结果表明，产生自杀意念的用户或群体在社交媒体上发布的帖子在语言表达上具有特定形式[8-9]。因此，开始利用机器学习方法分析用户社交帖子的文本内容的研究日益增多，推动了社交媒体用户自杀意念检测的发展[10-12]。尽管如此，传统的机器学习方法对特征依赖性强，且特征构建需要大量的专业领域知识进行手工设计，模型泛化能力有限。相比之下，深度学习方法通过大规模数据和深度神经网络来解析文本中的复杂的语义和上下文信息，可以轻松实现特征的自动学习。现有的深度学习模型革新了传统机器学习方法的检测模式与框架，提供了新的范式，能够适应不同

的应用场景[13-16]。在网络监测资源缺乏的情况下，深度学习可以实现对自杀风险人群的非入侵式自动化检测，增强个体自杀意念的检测效能，对自杀预防工作具有重要意义。

随着新一代人工智能技术应用场景的不断拓宽，以及“人工智能+”行动的深入推进实施，众多机器学习赋能自杀意念检测的方法被提出，特别是在社交媒体平台上的检测已经成为该领域促进新质生产力发展的研究热点。但在这个过程中也相应面临着诸多方面的现实问题与挑战：

（1）在数据隐私和伦理标准日益严格的背景下，如何在社交媒体上构建高质量数据集？

（2）如何在社交媒体共享内容中提取关键特征，高效地检测出用户的自杀意念？

（3）如何设计和优化机器学习模型以提高自杀意念检测的准确性和效率？

（4）现有的机器学习方法在处理敏感数据时存在哪些优势和局限性？

（5）自杀意念检测领域中存在哪些前沿难题及其创新解决方案？

目前，现有的综述多以文本单模态的角度进行阐述[17-18]，或者侧重某类精神疾病，如抑郁症进行调查[19-20]，内容都有所限制。且国内关于此领域的文献仍较为稀缺，现有的研究也较为零散，尚未形成体系。因此，本文对近年来国内外运用机器学习方法在社交媒体平台上检测自杀意念的前沿研究进行系统总结和梳理，致力解决上述核心问题与挑战，以期为相关领域的研究人员提供借鉴和参考，打造数智赋能心理健康新生态。本文的贡献归纳如下：

（1）通过细致的文献回顾，从数据采集、特征提取、模型检测等角度揭示机器学习在自杀意念检测中的发展趋势、关键技术和作用；

（2）在“人工智能+”行动背景下，本文尝试结合计算机、心理学和语言学等相关学科，从传统机器学习方法和深度学习方法等角度，综合评估现有模型性能，并对比分析其优势和不足；

（3）为丰富自杀意念检测的类别，本文调查了抑郁症、焦虑症、厌食症等多种自杀风险较高的精神疾病的检测方法，以提供新的见解；

（4）本文聚焦目前先进的研究技术，尝试从多模态、大语言模型等视域阐述当前领域亟需解决的难题及应对之策。

1 应用机器学习的自杀意念检测概述

本节将详细阐述自杀意念检测的概念、目标与流程，概述目前普遍应用的机器学习检测方法，并介绍模型性能评估指标，为后续章节中归纳检测框架提供理论基础。

自杀意念是指个体产生自杀行为的念头、意图或想法，往往暗示着个体面临严重的心理健康问题[3]。自杀意念检测是指识别和评估个体有自杀想法、计划或行为的过程，还包含对精神疾病的检测例如抑郁症检测、焦虑症检测等。其目标是在悲剧发生前发现这些危险的意图或行为，并实施适当的干预和治疗措施。

在社交媒体平台上，自杀意念检测主要借助机器学习技术来分析用户的社交帖子内容，包括自杀文本分类、自杀信息推理或者自杀风险人群识别等内容。应用机器学习方法在社交媒体上进行用户自杀意念检测的一般流程如图 1 所示，主要步骤包括数据采集、数据预处理、特征工程、模型检测及模型评估。

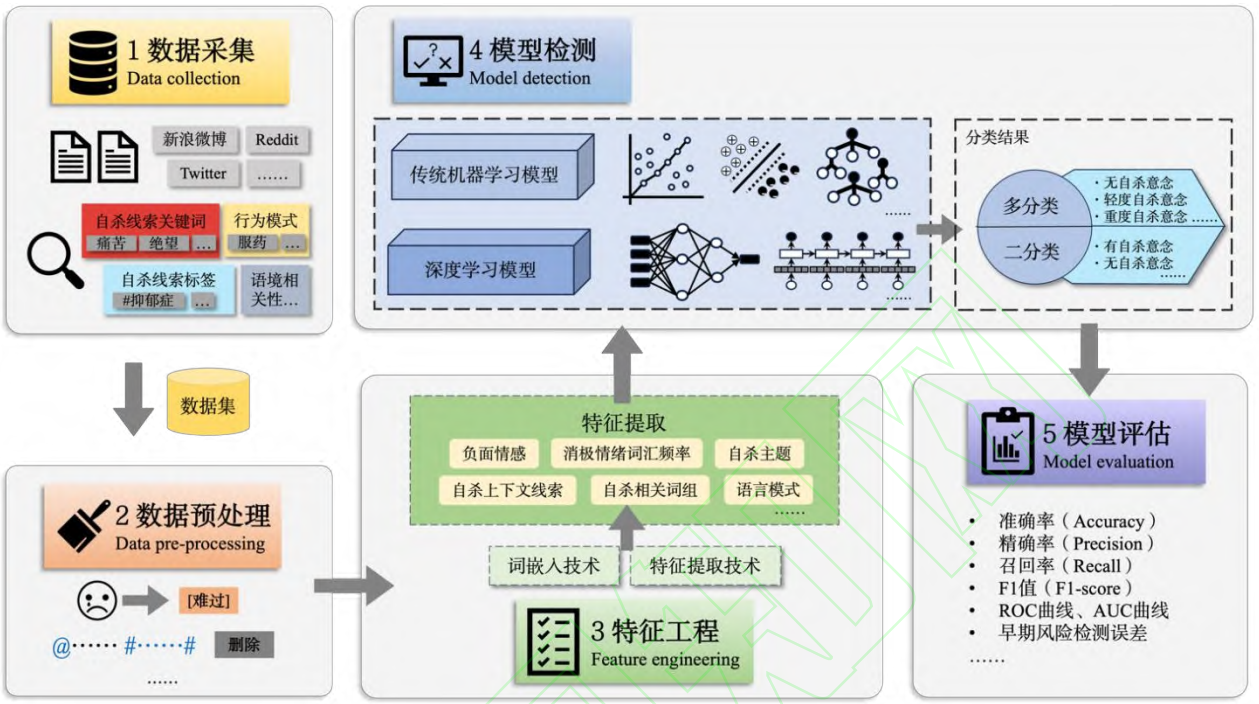

图 1　社交媒体上自杀意念检测的流程

Fig.1　Suicidal ideation detection process on social media

目前，广泛应用于社交媒体上进行自杀意念检测的传统机器学习算法多倚重于监督学习[21]，主要包括支持向量机（support vector machine，SVM）、决策树（decision tree，DT）、逻辑回归（logistic regression，LR）、朴素贝叶斯（naive bayes，NB）等，还包括一些集成方法，如随机森林（random forest，RF）和极限梯度提升算法（eXtreme gradient boosting，XGBoost）等。近年来，基于深度学习的框架，如卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）和基于 Transformers 的双向编码器表示（bidirectional encoder representation from transformers，BERT）逐步显现出检测潜力。随着技术的发展，基于生成式预训练 Transformer 模型（generative pre-trained transformer，GPT）的检测框架也崭露头角。

为了评估自杀意念模型的性能和质量，通常使用混淆矩阵来计算评价指标[22-23]，主要包括准确率（accuracy，Acc）、精确率（precision，P）、召回率（recall，R）、F1 值（F1-score，F1）和 ROC 曲线（receiver operating characteristic curve，ROC）和 AUC 曲线（area under curve，AUC）。鉴于自杀意念会随着时间变化，在此背景下，早期风险检测误差（early risk detection error，ERDE）指标也逐渐被采用[24]。

接下来，在第 2 节中，本文将介绍数据集的采集和预处理方法。第 3 节会引入特征工程及其常用方法。第 4 节会回顾对比基于传统机器学习的检测方法。第 5 节将总结分析基于深度学习的检测方法。在第 6 节里，会讨论本领域中亟待解决的问题及其创新解决方法。第 7 节将对全文进行总结，阐述其中面临的挑战，并指出未来的研究方向。本文的框架图如图 2 所示。

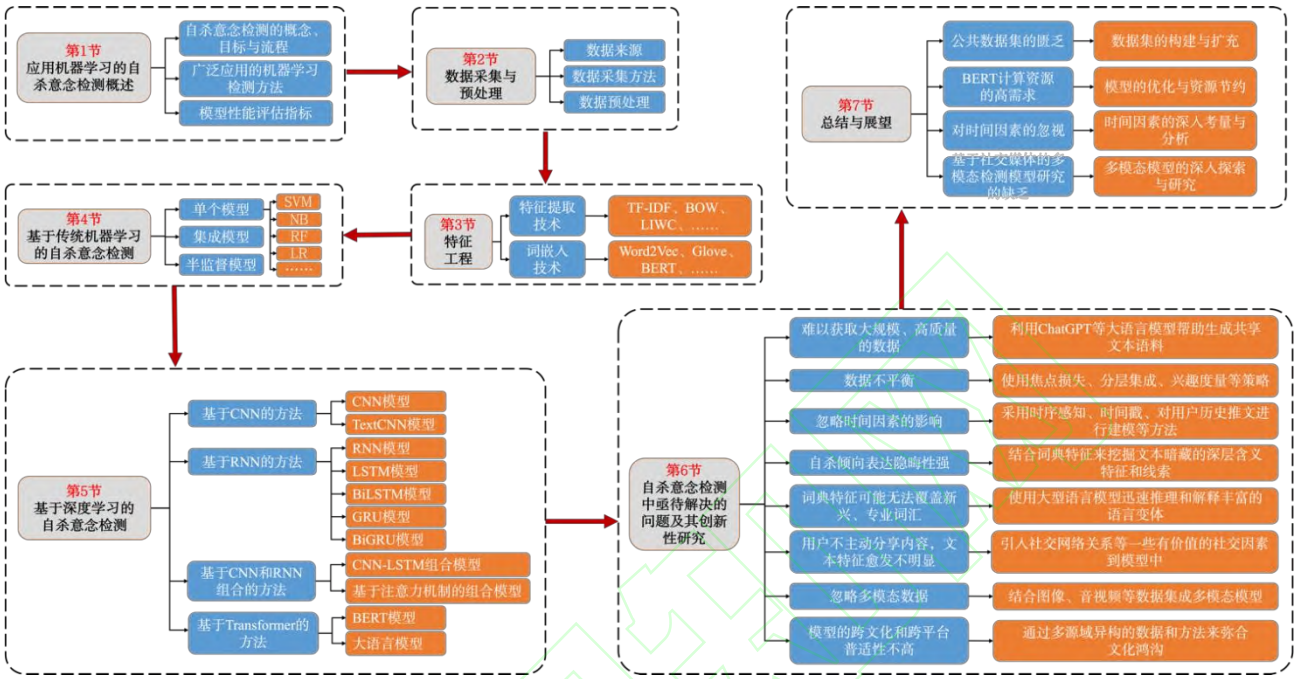

图 2　论文框架图

Fig.2　The structure of paper

2　数据采集与预处理

采用前沿的机器学习技术进行自杀意念检测会不可避免地带来隐私和伦理问题，存在数据壁垒的现象。因此，如何在保护用户隐私的前提下获取大规模、高质量、有意义的数据是本领域的一项重大挑战。本节将概述数据的来源、选择与采集方法，以及数据预处理的流程。

2.1　数据来源和采集

在进行自杀意念检测研究时，选择合适的数据来源至关重要。研究者通常利用热门的社交媒体平台，如新浪微博、Twitter、Reddit 和 ReachOut 等作为数据来源，因为这些平台汇聚了大量用户生成的文本数据，包括帖子、评论和留言等，为研究提供了丰富的数据信息和资源[25-26]。

确定数据来源后，需要选择合适的数据获取方式。多数情况下，可以通过平台的 API 接口、网络爬虫技术等手段抓取相关文本数据，在此基础上自主构建数据集。目前大多数研究构建的数据集以英文为主[9-10,15-16,27-29]，譬如，Ji 等[27]基于 Reddit 平台的 SuicideWatch 社区构建了一个英文数据集，其中包含 3549 篇有自杀意念的帖子；Nikhileswar 等[28]使用 API 在 Reddit 平台上分别收集了 116037 篇自杀相关和正常的帖子数据；Vioulès 等[29]在 Twitter 平台上成功收集了 5447 条自杀倾向推文作为数据集。

在中国，研究者主要通过新浪微博平台的虚拟社区，如微博树洞或者“抑郁症”等相关超级话题中爬取数据。在这些社区中，用户经常发布明确表明自杀想法的评论，使其成为识别有自杀意念的用户的重要资源。在遵守平台的数据使用政策和用户隐私保护的前提下，研究者利用平台提供的 API 接口，依据既定的自杀线索关键词（如绝望、孤独、无助、痛苦）、标签（如#抑郁症、#心理健康、#自杀预防、#SOS）、用户行为模式（服药、夜间发帖、互动及帖子删除频率）及语境相关性（如“沉重”在不同语境下表达不同情绪），进行自动化数据抓取。为了提高数据的相关性和覆盖度，一般会设定合理的时间范围，并采用多轮迭代的数据抓取策略，以获取不同时间段内用户表达的变化，从而得到初步的数据供后续数据预处理使用。例如，Li 等[14]从抑郁症超话中爬取了用户在 2020 年 7 月 1 日至 2021 年 4 月 30 日的共享内容，并构造了一个全新的中文自杀检测数据集，包含 4521 名正常和自杀倾向用户以及 524841 条微博；Cao 等[30]则从微博树洞里获取了 7329 名用户，并将其中 3652 名用户标注

为自杀倾向用户，此外他们还收集了这些用户在2018年5月1日至2019年4月30日所发布的74万多条纯文字微博及35万多条带图片的微博，形成研究所需的数据集。为了检验用户内心想法和情绪变化对于社交媒体自杀风险的影响，他们在2021年构建了一个包含7304名用户和500名论坛用户的数据集[31]。

还有部分研究针对阿拉伯文、俄文、西班牙文进行检测[32-35]。

除此之外，研究者还考虑利用公开可用的自杀意念检测相关数据集，如CLPsych 2015[36]（computational linguistics and clinical psychology）、CLPsych 2019[37]、Reddit C-SSRS[38-39]（columbia suicide severity rating scale）和一些用于抑郁症自杀风险识别任务相关的数据集，如eRisk 2017（early risk）[40]、eRisk 2018[41]和eRisk 2023[42]等，这些数据集提供了标注完整的自杀意念相关文本数据，可用于模型训练和评估。但目前公开可用的数据集较为匮乏，中文数据集只有Chen等[43]建立的心理咨询问答语料库efaqa（emotional first aid question answering），是心理咨询领域首个开放和最大的中文问答语料库，包2万条个体心理咨询数据，为自杀及相关症状、疾病等分类检测提供了丰富的数据支持。表1总结了用于自杀意念检测任务的数据集。

表1 用于自杀意念检测任务的数据集总结

Table 1 Summary of datasets for suicidal ideation detection task

数据集类型	数据集	社交媒体	语言	数据集内容
自主构建数据集	Ji等[27]	Reddit	英文	3549条有自杀意念的帖子
	Nikhileswar等[28]	Reddit	英文	116037条自杀相关帖子、116037条正常帖子
	Vioulès等[29]	Twitter	英文	5447条自杀倾向推文
	Li等[14]	新浪微博	中文	1606名自杀倾向用户、2915个正常用户 98680条自杀相关微博、426161条正常微博
	Cao等[30]	新浪微博	中文	3652个自杀倾向用户、3677个正常用户 252901条自杀相关微博、491130条正常微博 93461条自杀相关微博（含图片）、260667条正常微博（含图片）
	Cao等[31]	新浪微博	中文	3652个高危微博用户、3652个普通微博用户 392个高危论坛用户、108个普通论坛用户
	Baghdadi等[32]	Twitter	阿拉伯文	956条自杀推文、1071条正常推文
	Almars等[33]	Twitter	阿拉伯文	3155条抑郁症相关推文、2975条正常推文
	Narynov等[34]	VKontakte	俄文	35000条抑郁症帖子、50000条对不同主题持负面态度的帖子
	Valeriano等[35]	Twitter	西班牙文	498条自杀推文
公共数据集	CLPsych 2015[36]	Twitter	英文	SAD组：159个用户和421000条推文 PTSD组：244个用户和573000条推文 Bipolar组：394个用户和992000条推文 抑郁症组：411个用户和1000000条推文
	CLPsych 2019[37]	ReachOut	英文	11129个自杀用户、11129个自杀非用户
	Reddit C-SSRS[38]	Reddit	英文	500个用户、15755条帖子
	Reddit C-SSRS-v2[39]	Reddit	英文	448个用户、7327条帖子
	eRisk 2017[40]	——	英文	135个抑郁症用户、49557条抑郁症帖子 752个正常用户、481837条正常帖子
	eRisk 2018[41]	——	英文	214个抑郁症用户、90222条抑郁症帖子 1493个正常用户、986360条正常帖子
	eRisk 2023[42]	——	英文	3107个抑郁症用户、3807115条抑郁症帖子

efaqa[43]	——	中文	20000 条线上心理咨询数据

2.2 数据预处理

通常，从社交媒体平台收集的数据由原始的、非结构化的文本组成，特征稀疏性明显，其中包含可能对模型的检测性能产生负面影响的噪声。因此需要行数据预处理来降低数据稀疏度，在不丢失有用信息的情况下降低噪声，并将原始数据转换为机器可读的格式，以确保数据的质量和可用性。

研究者可以采用不同的技术策略进行数据预处理。文本清洗（text cleaning）是数据预处理中的关键步骤，可以提取文本中有意义的信息并去除数据中的无用信息和噪声。这一过程包括去除 HTML 标签、URL 链接、电子邮件和表情符号等。此外，在社交媒体环境中，还需特别关注移除平台特有的元素，如@提及和#话题标记等。去除停用词（stopwords removal）是常用的文本预处理步骤，旨在移除那些在语言中频繁出现但对于语义分析贡献较小的词。分词操作（tokenization）可以将文本划分为单个词语或短语，转化为计算机可以理解和处理的形式。在社交媒体的环境中，分词需要特别设计，可以结合正则表达式等方法以处理表情符号、网络用语、非标准缩写和混合语言文本。词干提取（stemming）是一种将词汇转换为词干形式的技术，可以除词汇的词缀，减少文本数据的维度。词形还原（lemmatization），是将词汇转换为在词典中列出的标准形式的过程。与词干提取不同，它考虑了社交媒体中词汇的词性、时态和语态等语法特征及新兴词汇的正确词元。

随着社交媒体蓬勃发展，使用 emoji 等表情符号成为一个流行的情感表达方式，因此，一些新兴的技术可以精细处理这些表情符号，可以更好理解用户的情绪、心理活动和自杀意图。例如，Liu[44]利用条件随机场（conditional random field，CRF）来抽取表情符号中的关键信息，将非正式的表情符号表达转换为包含情绪特征文本标准形式。Zhang 等[45]，通过开发 Emoji_text 框架将 emoji 映射为描述性文本，如以便模型更容易理解其情感内容。

通过数据预处理，能够显著提升数据的质量，以保障后续成功进行特征工程和模型检测。

3 特征工程

在自杀意念检测任务中，研究人员需要使用不同的技术从预处理后的文本数据中提取自杀意念的特征，这一过程被称为特征工程。特征工程是自杀意念检测的关键步骤，旨在将文本数据转换为机器学习算法可以处理的特征表示形式，以便为后期的机器学习模型的训练和精准检测奠定基础。本节将详尽描述特征提取技术和词嵌入技术两种常用的特征工程方法在社交媒体用户自杀意念检测中的应用，并分析它们在提取关键特征方面的潜力。

3.1 特征提取技术

特征提取技术是一种将数据选择并组合为特征的方法，能对原始数据集进行全面综合的描述，减少必须精确处理的数据量，一般与传统机器学习方法搭配使用。在特征工程的过程中，可以使用多种技术来提取特征，表 2 梳理了常用的特征提取技术及优缺点。

表 2 常用的特征提取技术总结

Table 2 Summary of commonly used feature extraction techniques

技术	作用	提取特征类型	优点	缺点
TF-IDF	衡量单词在文本中的重要性	单词重要性	可以凸显关键词的重要性，适用于文本分类和信息检索等任务	无法捕捉单词之间的顺序信息对于长文本，计算复杂度较高
BOW	衡量单词在文本中的出现频率	单词出现频率	简单有效的特征表示，适用于文本分类和聚类等任务	无法获取单词之间的顺序信息无法处理语义信息
LIWC	分析文本的情感、心理和语言特征	情感、心理和语言特征	可以提取丰富的情感和心理特征，适用于情感分析和心理研究	过于依赖预定义的词典，无法处理新兴词和上下文信息

LDA	从文本数据中发现潜在的主题结构	主题特征	可以自动从文本数据中学习主题结构，发现潜在的主题结构，无需人工干预	需要预先设定主题数目和超参数，容易忽略词语的顺序和上下文信息
N-gram	抽取单词或字符之间的顺序信息	单词或字符序列	可以理解上下文信息和语言结构，适用于文本生成和序列预测	需要选择合适的 n 值，对于长文本，特征维度较高处理能力有限
POS	学习单词在句子中的语法信息	词性标签	可以提取语法信息，适用于句法分析和语义分析等任务	对词性标注器倚靠性强，对新兴词和上下文信息的学习能力有限

大部分研究都采用多种特征提取技术组合的形式，从用户发布的非结构化文本中提取不同类型的特征，以增强检测效果。主要的特征类型包括语言特征、句法特征、上下文特征、N-gram 特征、基于知识的特征和主题特征等。Ji 等[27]使用词频逆向文件词频(term frequency inverse document frequency，TF-IDF)、POS（Part-of-Speech）、LIWC（linguistic inquiry and word count）和隐含狄利克雷分布（latent dirichlet allocation，LDA）等技术提取了句法、语言、词性、主题等特征，并输入到分类器中，取得了良好的分类效果。Rabani 等[46]将词袋（bag of words，BOW）、TF-IDF 等特征技术进行组合，提取自杀主题、消极情绪词汇频率、自杀相关词汇频率等特征，并输入到不同机器学习模型中进行实验，检测质量显著提高。Hiraga[12]、Tadesse[15]及 Huang[47]等的研究证明了 N-gram 技术在自杀意念检测中的有效性，可以挖掘出自杀相关词组、负面情感、语言模式等特征。

总体而言，早期的特征提取技术由于单一的特征的限制，导致关键信息量不足。因此更多的特征及其组合被不断地研究和引入。在综合特征的基础上，各种信息能够互相补充，使得自杀相关特征逐渐明显，但过多甚至冗余的特征会导致模型的运行效率下降。因此，在特征工程的过程中，需要谨慎选择和组合具有代表性的特征。

3.2 词嵌入技术

词嵌入技术是一种将词语映射到低维向量空间的技术，可以将自然语言文本嵌入到分布式向量空间中，其中，词语的语义和语法信息被编码为向量的位置和方向。通过运用词嵌入技术，可以将文本数据转换为更密集的向量表示，从而推断出词语之间的语义关系和上下文信息，通常被用于深度学习模型的输入。目前，Word2Vec、GloVe（global vectors for word representation）、FastText 和 BERT 是广泛应用的词嵌入技术。

Word2Vec 构造简单易用，是大多数研究使用的词嵌入技术[19,20,27,28,33]，通过训练神经网络模型来学习单词的分布式表示，包括 Skip-gram 和连续词袋（continuous bag of words，CBOW）两种模型，分别用于学习上下文单词和目标单词之间的关系。尽管 Word2Vec 简洁高效，但无法处理自杀语境中多义词和上下文中的复杂关系。由于词嵌入是基于上下文的，同一个词在不同的上下文中可能具有不同的含义，但传统的 Word2Vec 往往将其视为同一个向量表示，这容易导致语义信息混淆的问题。

GloVe[48-51]是一种基于全局词频统计的词嵌入技术，通过对全局语料库中的共现矩阵进行分解，学习单词的向量表示。它能通过关联自杀相关词语的表示与上下文线索，更好地理解自杀语境多义词的语义，尤其是稀有词，从而缓解语义信息混淆的问题。然而，在大规模自杀语料数据集上的应用时，GloVe 仍面临计算复杂度较高的挑战。

FastText[52-53]是一种基于词向量的文本分类和词嵌入模型，可以学习词语的字符级 N-gram 表示。它使用层次 Softmax 和负采样等技术，以及基于子词的词嵌入方法，通过将单词分解为子词来减少词汇表的大小，能够快速进行训练、自杀线索推理和读取词语内部结构，在处理大规模文本数据时具有较高的处理效率。但 FastText 的词向量表示是基于词频的统计信息，对复杂语义理解有限，处理长文本时效果差强人意。

BERT 是近年广泛使用的一种基于

Transformer 模型的词嵌入技术[32,54-57]，通过预训练模型来学习自杀领域语料的上下文表示，具有强大的上下文理解和解决歧义、多义问题的能力以及支持多语言、多任务学习的特性。BERT 可以通过一个额外的输出层进行灵活微调，以创建适用于广泛任务的模型，如问答和语言推理，而无需对特定任务的架构进行大量修改，从而提高计算效率，能更好地适用于大规模数据计算。尽管应用前景开阔，但其参数量过多，训练的成本较高。

相较于特征提取技术，词嵌入技术可以实现特征向量的自动生成和提取，但仍存在一些限制。例如，词嵌入技术对于低频词和未登录词的处理效果不佳。由于这些词在训练数据中出现的频率较低，模型很难学习到它们的准确表示，进而影响模型整体的语义表达质量。词嵌入技术还存在数据偏见的问题，即模型在训练过程中可能受到数据集中类别和数量的不平衡性影响，导致对自杀风险及相关特定群体（抑郁症、焦虑症、厌食症等患者）或主题的表达能力不足。因此，在选择和应用词嵌入技术时，需要综合考虑这些问题。不同的词嵌入技术具有各自的特点和适用场景，需要根据具体任务需求和数据特点来选择合适的技术，方便后续输入到深度学习模型中进行训练。

表 3 总结概括了自杀意念检测中常用的词嵌入技术。

表 3 常用的词嵌入技术总结

Table 3 Summary of commonly used feature extraction techniques

技术	作用	数据表示	优点	缺点
Word2Vec	基于上下文预测词语，将词语映射到低维向量空间	分布式表示	模型构造简单，提取语义和上下文信息能力强，计算效率高	无法处理自杀语境中多义词、上下文中的复杂关系，存在语义混淆问题
GloVe	基于共现矩阵分解，学习全局词语共现信息	共现矩阵表示	可以解决语义信息混淆的问题，对稀有词解释性强	计算复杂度及成本高，难以在大规模数据上应用
FastText	基于子词信息预测词语，学习词语的字符级 N-gram 表示	字符级 N-gram 表示	能快速进行训练、推理和读取词语内部结构，处理大规模数据效率高	对词汇复杂语义关系的理解能力有限，长词的处理能力较差
BERT	基于 Transformer 的双向预训练，学习上下文相关的词语表示	上下文相关表示	具有强大的上下文理解能力和预训练和微调的灵活性，适应多语言、多任务学习，能解决多义歧义的问题	参数量大，训练成本高，难以复现

4 基于传统机器学习的自杀意念检测

在完成特征工程后，提取的特征被输入到传统机器学习的自杀意念检测模型中进行训练。当下，监督学习是应用最广泛的传统机器学习方法，通过使用已标注的数据集来训练模型，使其能够掌握输入特征与输出标签之间的关系。训练完成后，模型可以用于检测新的文本数据是否存在自杀意念。本节将总结不同的传统机器学习检测方法，对比其优缺点，以及深入探讨它们的构建及优化过程。

Herath 等[58]使用 TF-IDF 和 N-gram 技术提取出丰富的文本情感特征来计算与自杀意念相关的句子的正极性和负极性，并分别应用了逻辑回归、朴素贝叶斯和支持向量机三个单一分类模型检测 Facebook 平台上僧伽罗语帖子的自杀倾向，贝叶斯模型 F1 值达 82%，对低资源语言检测效果显著。Chadha 等[59]利用 TF-IDF 技术，并应用于支持向量机和逻辑回归模型上，准确率均超过 80%。Ji 等[27]在构建 LDA 等四类特征的基础上，比较了 XGBoost、逻辑回归、随机森林和梯度提升决策树 4 种分类方法，准确率均超过 90%，证明了传统机器学习在自杀倾向分类任务上赋能的有效性。Rabani 等[60]提出特征工程增强机制，利用 TF-IDF 特征提取技术并

与支持向量机、逻辑回归和 XGBoost 模型结合，可以抽取自杀意念相关的潜在语义和主题等特征。该模型可以在 Twitter 和 Reddit 文章中检测自杀倾向的用户并进行分类，其中 XGBoost 最优，准确率达到 96.33%。Saifullah 等[61]结合选用支持向量机、决策树和随机森林等方法对 YouTube 上关于新型冠状病毒话题的印尼语视频评论中进行焦虑症检测，其中随机森林模型达到最优的准确率，为 98.4%。

尽管上述模型表现出色，但都属于单一模型，仍存在欠拟合或过拟合、无法理解复杂的语义关系、对数据分布的假设过于简化、对噪声和异常值过于敏感等缺陷。集成模型的引入可以缓解这些症候，通过耦合多个单一模型，结合多个模型的优势来提高检测性能和模型泛化能力。Liu 等[62]通过实验表明，一个由支持向量机、朴素贝叶斯和正则化逻辑回归构建的集成模型比单个基线模型的效果更显著，准确率提升了 5.36%。Lekkas 等[63]在研究中融合了 XGBoost、优化的逻辑回归树以及一个三层的前馈神经网络等模型，形成 Consensus 急性自杀意念预测模型，研究结果显示，该模型的 AUC 值和 F1 值比单个最差模型攀升了 10%以上。

集成模型虽然在检测任务中表现出色，但在处理大规模的社交媒体数据时，其计算复杂度会明显增加，导致训练和检测的效率下降。为了解决这些问题，聚焦半监督式的方法开始出现。有监督的方法需要标注完整的数据进行训练，而这些数据通常很少，而半监督方法可以在较少的已标注数据的情况下进行训练。例如，Farruque 等[64]提出了一个半监督学习模型 SSL，该模型结合了初始的监督学习模型和零样本学习模型以收集与抑郁症状相关的样本，创建了最大的临床注释数据集。此外，该模型利用当前先进的大型心理健康论坛文本预训练语言模型，并在此基础上进一步微调，以提高对抑郁症状检测，Macro-F1 值达到 56%。Sharmeen 等[65]提出了一种基于半监督学习的自训练分类器来检测自杀意念，可以同时利用有标记和无标记数据进行训练，在标记数据不足时，可以提高模型的性能，准确率和 AUC 值均达到 93%。

总体而言，上述方法在应用传统机器学习进行自杀意念检测方面取得了良好的进展。但仍掣肘于特征选择和提取困难、数据不平衡以及高计算复杂度等问题，尤其是当涉及到复杂的文本数据和语义关系时，传统机器学习方法的建模能力会受到限制，仍有进一步提升的空间。

自杀意念检测中的传统机器学习算法归摄于表 4 中。

表 4　基于传统机器学习的自杀意念检测算法总结

Table 4　Summary of algorithms for detecting suicide ideation based on traditional machine learning techniques

算法类别	文献	年份	算法	社交媒体	语言	特征提取技术	评价指标	优点	缺点
单一模型	Herath 等[58]	2024	NB* SVM	Facebook	僧伽罗文	TF-IDF N-gram	Acc=0.79 F1=0.82 P=0.7	结合丰富的情感特征和低资源语言特征，提高检测性能	数据地区代表性强，模型可移植性差。数据预处理过程要把僧伽罗语转成英语，存在信息表达不准确的风险
	Chadha 等[59]	2022	SVM* LR	Reddit	英文	TF-IDF	Acc=0.807	模型简单易解释，计算运行速度快	数据集限于 Reddit，对数据分布的假设过于简化
	Ji 等[27]	2018	RF*	Reddit	英文	TF-IDF	Acc=0.96	提供了丰富的特	过多的特征之

			SVM GBDT XGBoost	Twitter		LDA POS LIWC	AUC=0.986 F1=0.964	征和知识，有助于理解自杀意念	间产生的相互作用会影响模型性能，对噪声和异常值较为敏感
	Rabani 等[60]	2023	SVM RF XGBoost*	Reddit Twitter	英文	TF-IDF LDA	Acc=0.963 F1=0.962 P=0.982 R=0.954	提出增强型特征工程框架，用于从社交媒体帖子中提取特征，并用于多类分类	数据规模较小，存在欠拟合或过拟合的现象
	Saifullah 等[61]	2023	RF* DT NB SVM XGBoost	YouTube	印尼文	TF-IDF	Acc=0.984	结合多种机器学习算法和特征提取方法，提高了检测焦虑的准确性	模型专注于印尼语，对其他文化的解释性差
集成模型	Liu 等[62]	2022	正则化 LR、NB 和 SVM 集成模型	新浪微博	中文	TF-IDF	Acc=0.903	在中文数据集上表现优良	模型过多耦合带来了结构上的负担
	Lekkas 等[63]	2021	Consensus	Instagram	德文	LIWC	Acc=0.70 AUC = 0.78 F1=0.74	利用集成学习方法提高了预测的准确性	数据样本量较小，模型泛化能力较差
半监督模型	Farruque 等[64]	2022	SSL	Reddit	英文	N/A	Acc=0.92 F1=0.92	扩大训练数据集能力强，创建了较大的临床注释数据集	缺乏持续的人工注释或人工参与策略
	Sharmeen 等[65]	2023	SGD	Reddit	英文	TF-IDF	Acc=0.93 AUC=0.93 F1=0.93	利用少量标记数据训练未标记数据，适用于标记数据不足的情况	集成模型过多，无法确定各模型和特征的权重

*代表最优模型

5 基于深度学习的自杀意念检测

在传统的机器学习方法中，数据标注和特征工程往往需要手动选择和提取特征，需要耗费大量的人力和时间成本，且复杂语义理解能力有限。相比之下，深度学习方法表现出更出色的性能，它能够在大规模的数据中自动学习自杀相关的特征表示，无需手动提取特征，并且能够有效获取到文本数据间逻辑复杂的语义关系。在实际应用中，可以利用基于 CNN 和 RNN 模型对文本数据进行建模，还可以使用 BERT 等基于 Transformer 的预训练语言模型，通过学习上下文相关的词嵌入表示来增强模型的检测效果。在本节中，将重点讨论不同深度学习方法在自杀意念检测领域中的应用，分析对比各自优势及劣势，以及介绍检测模型的搭设流程，为模型的深入优化和创新提供关键视角。

5.1 基于 CNN 的自杀意念检测

CNN 因其强大的特征提取能力，在社交媒体上的自杀意念检测中发挥着重要的作用。利用 CNN 进行自杀意念检测的基本框架如图 3 所示。首先，使用预训练的词嵌入模型将文本数据中的单词映射到低维向量空间；接着，通过卷积层提取文本数据中的局部特征，并使用池化层进行降维和特征选择；最后，通过全连接层将特征连接并映射到预定义的类别，进行自杀意念的检测与分类。

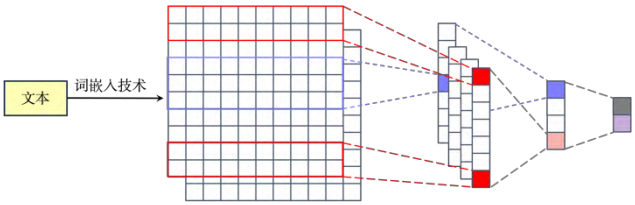

图 3 基于 CNN 的自杀意念检测框架

Fig.3 CNN-based framework for suicidal ideation detection

许多研究对 CNN 模型进行了深入探索。Beniwal 等[66]尝试使用一种 BERT 集成的 CNN 模型分析包含文本、表情符合和图像的多模态推文数据，进而检测抑郁症用户，准确率高达 99%。Gorai 等[67]发挥了 CNN 参数共享的优势，采用并行 CNN 的架构，来减少模型的参数量，提高了基于句子分析的自动自杀识别的性能，准确率超过 99%。

Kim 等[68]提出了 TextCNN 模型，开启了自杀意念文本识别的全新范式。Yao 等[69]在 Kim 的研究基础上设计了一个新的 CNN 框架来检测服用阿片类药物用户的自杀行为，在兼顾运行速度和识别精度时表现卓越，F1 值高达 96%。考虑到大多数 CNN 模型存在信息缺失的问题，Li 等[14]在 TextCNN 的基础上引入 Word2Vec 词嵌入技术，提出了一种多特征标签关联的文本卷积神经网络模型，能抽取并综合词典、用户发布时间和社交信息三个关键特征，准确率超过 88%，在面向微博自杀用户和中文文本识别上取得了满意的效果。

5.2 基于 RNN 的自杀意念检测

由于 CNN 的局部特征提取机制，它无法捕捉到文本数据中的长距离依赖关系。在一些需要考虑上下文关系的自杀倾向分类任务中，可能会导致性能下降。与之相对，RNN 模型通过其循环机制处理文本数据，将当前时刻的输入与前一时刻的隐藏状态相结合，以此关联文本数据中的上下文信息。在应用上，Alabdulkreem 等[70]通过 RNN 模型来分析阿拉伯妇女的推文，以确定她们是否患有抑郁症，并据此检测自杀风险，模型准确率为 72%。Apoorva 等[71]采用 RNN 模型对 Twitter 平台推文进行建模，提取指向抑郁症的语言线索特征，来识别抑郁症群体，准确率高达 96%，但在处理不同语境中上下文信息时存在限制，模型的性能还有待增强。

鉴于 RNN 模型普遍存在的梯度消失问题，导致模型性能受限，因此衍生出 RNN 的变体模型，如 LSTM（long short-term memory）和 GRU（gated recurrent unit）等。LSTM 的顶部包含一个存储单元，使其更容易、有效地将信息从一个时间实例传输到下一个时间实例。因此，与 RNN 相比，LSTM 能够从更早的状态中召回更多数据，减轻消失梯度问题。Kancharapu[49]和 Deepa J 等[72]的研究都证实了 LSTM 在检测、预测和识别自杀相关推文的有效性。由于传统的 LSTM 只能从前向的顺序处理序列数据，这限制了其对序列数据中后续信息的感知能力。BiLSTM（bidirectional long short-term memory）的出现打破了这一限制，它引入了后向处理机制，使得模型能够同时从前向和后向处理序列数据，从而更全面地理解并整合序列数据中的上下文信息。Almars[33]提出了一种结合 BiLSTM 和注意力机制的阿拉伯语抑郁症文本分类模型。在分类效果上，该模型在准确率方面相比基础的 BiLSTM 提升了 3%。Kancharapu 等[73]使用三个基于不同语义的 BiLSTM 框架集成的模型来预测疫情期间的自杀频率。研究结果显示，相较于 CNN、LSTM、RNN 等神经网络模型，Bi-LSTM 模型的准确率最高，达到了 86.47%，能够提供更为丰富的自杀特征表示。然而，上述研究仍受缚于处理序列较长的上下文信息较难的问题上。基于此，Kumar A 等[51]开发了一种能够通过在门控循环单元中捕获局部上下文信息与表征的模型，以鉴别出社交媒体用户的自杀风险因素，缓解了较长序列所衍生的问题。

5.3 基于 CNN 和 RNN 组合的自杀意念检测

随着深度学习的迅猛发展，在自杀意念检测任务中，对文本序列数据的处理需求、特征表达能力和模型性能的要求日益提高，基于 CNN 和 RNN 的组合模型逐渐成为研究焦点。组合模型通过充分吸收 CNN 的特征提取和 RNN 序列建模的优点，能够更有效地处理序列数据，读取长文本中蕴含的复杂

逻辑结构，精细提取局部特征，进一步扩大提高模型的特征表达能力。这为各个领域的研究和应用提供了崭新的思路和方法。Mumeni 等[74]提出了基于 CNN 和 GRU 的组合模型，并通过实验与单独的 LSTM、CNN 和 GRU 模型进行比较，组合模型的准确率达到了 97.27%，均高于单个模型，证实了组合模型在文本自杀意念检测中的有效性。Kour[21]、Oyewale 等[75]分别构建了 CNN-BiLSTM 模型，F1 值均超过 94%，进一步揭示了组合模型在增强检测结果方面的成功应用。但在上述研究中，单一 CNN 的提取能力有限，Priyamvada 等[76]采用一种栈式堆叠的 CNN 和双层 LSTM 混合模型来评估自杀风险，识别准确率比单个 CNN 的模型提升了近 5%。

在社交媒体自杀意念检测中，大部分组合模型面临长期依赖和局部感知等方面的挑战，在处理长文本和序列数据中的关键特征时可能存在一定的局限性。为了克服这些问题，一些研究开始将注意力机制引入到组合模型中。注意力机制能够根据输入的序列数据自动学习权重分配，使模型能够从众多信息中更加关注并选择与任务相关的关键信息。Renjith 等[77]在卷积层之前加入注意力层，这有助于模型考虑加权输入值的不同部分之间存在的关系，更充分地提取到自杀文本中的隐含特征，尽管准确率和 F1 值均超过 90%，但是未考虑数据类别不平衡的问题。Chadha 等[16]提出了 ACL 模型（attention convolution long short-term memory），该模型结合了注意力机制、CNN 和 LSTM 模型提取的最佳特征，通过注意力层在预测时重点关注所需数据的细节和特定单词，不仅关注一个单词，而且综合考虑不同长度的混合单词来识别文本特征，对长文本中的潜在自杀信息理解得更加透彻。然而，ACL 模型更多地是对语言、语义等相关特征的提取，并未充分考虑社交媒体中文本之外的其他有效特征。为解决这一问题，Zogan 等[50]提出了一种基于多层次注意力网络的抑郁症检测模型 MDHAN（multi-aspect depression detection hierarchical attention network）。该模型采用基于推文和单词级别的两级注意力机制对用户推文进行编码，计算每条推文和单词的重要性得分，并结合社交行为、时间和主题等非文本特征与 GloVe 技术，用于抑郁症患者检测的工作，即使在文本特征不明显的情况下，该模型的可解释性也得到强化，检测准确率达 89.5%。

5.4 基于 Transformer 的自杀意念检测

基于 Transformer 的自杀意念检测模型已成为当前研究的热点，与传统的 CNN 和 RNN 模型相比，它具备并行计算、长期依赖建模和全局特征表达等显著优势。该类模型在自注意力机制的基础上引入了位置编码来处理序列数据，对输入序列中的每个位置元素进行编码标记，使得模型能够更好地理解序列数据中的顺序关系，从而更有效地关联文本的语义和情感信息。BERT 作为一种基于两层双向 Transformers 的预训练语言模型，具有强大的上下文感知、预训练和微调、多层特征提取、零样本学习和多任务学习等能力及特点，是当下自杀意念检测领域的流行模型。BERT 模型可以直接在数据集上进行预训练和微调，并接入分类器完成具体的下游任务，其结构如图 4 所示。

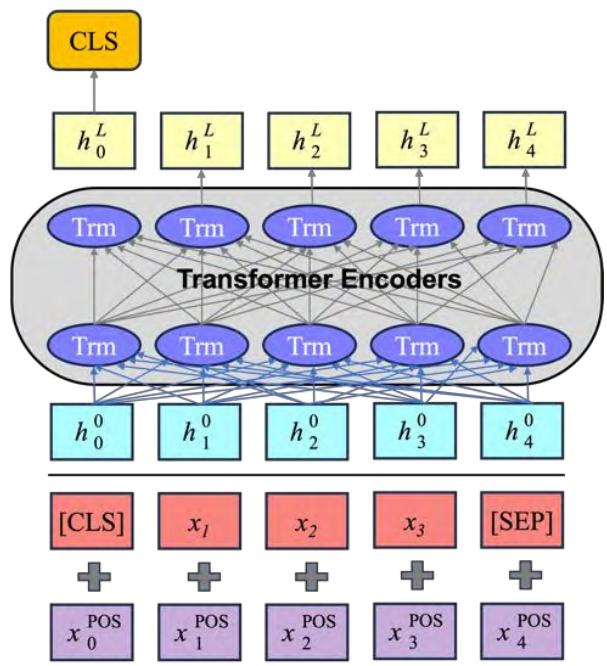

图 4 BERT 模型结构

Fig.4 Structure of BERT model

针对传统词嵌入模型在词向量表示中无法解

决词语多义性的问题，谌志群等[78]采用 BERT 模型在 5 万条微博的数据集上提取出词语中不同的语义特征，并输入到 BiLSTM 模型中进行倾向性分类，各项评估指标均超过 92%。随着技术的进步，更多的研究开始关注改进原始 BERT 模型。Baghdadi 等[32]在 BERT 基础上融入通用句子编码器，设计了五种 BERT 改进模型，评估指标均超过 92%，验证了其在阿拉伯文自杀意念检测上的有效性。Tavchioski 等[57]对 BERT 模型进行微调，用集成的方法构建了 RoBERTa（robustly optimized BERT approach）、BERTweet 和 mentalBERT 模型，准确率相较于原始的 BERT 基线模型均有提升，最高达到 86%，但该研究忽略了情感分析的重要性。为了解决这个问题，Lin 等[79]在 RoBERTa 模型基础上增加情感特征，提出了一种全新的自杀检测模型 RoBERTa-CNN，能从更抽象的情感层面对用户情绪进行分析，提取多种有效的情感特征，检测效果得到强化，准确率高达 98%。随着网络新潮语言和主题的多元化，Wang 等[80]量身定制了一个 BERT 模型，通过对文本进行多维度的自杀主题特征抽取和情感评估，运用 TF-IDF 技术和卡方检验确定关键词，并缩小范围进行频率分析，最大化激发了 BERT 模型对新兴自杀文本分类潜力，准确率高达 98%。为了精准识别各类精神疾病，Aragón 等[81]提出了 DisorBERT 模型，通过对词汇资源进行编码来指导语言模型的掩码过程，从而帮助模型更聚焦抽取与精神障碍相关的单词。该模型可以检测厌食症、抑郁症和自残三种精神障碍的迹象，F1 值最高可到 83%。

上述研究从特征提取方法、神经网络模型的集成和预训练模型的优化等不同角度对基于深度学习的自杀意念检测模型进行了一系列深入的实验与探索并取得了显著的成效。不同于传统的机器学习方法，深度学习模型在处理社交媒体多样化和非结构的文本上展现出更强的泛化能力；能够通过大规模的训练数据和复杂的网络结构，自动学习和解析更复杂和抽象的模式与规律，无需手动设计特征或规则，实现模型高效化、简洁化的训练和推理过程。值得注意的是，深度学习模型数据需求量大，对标注数据的依赖性强，对特征提取能力要求更高。尽管在神经网络的基础上加入 BERT 模型和注意力机制等可以一定程度上缓解以上问题，但往往需要更高的计算资源、牺牲更多的运行效率和搭设更复杂的模型结构，导致模型参数量巨大，模型间的匹配度要求高，不利于进行重新训练，这在一定程度上会影响模型性能。

自杀意念检测中的深度学习算法总结如表 5 所示。

表 5 基于深度学习的自杀意念检测算法总结

Table 5 Summary of algorithms for detecting suicide ideation based on deep learning techniques

算法类别	文献	年份	算法	社交媒体	语言	词嵌入技术	评价指标	优点	缺点
基于CNN	Beniwal 等[66]	2024	BERT-CNN	Instagram	多语言	BERT	Acc=0.99	使用多模态数据提高抑郁检测的准确性	涉及语言太多，模型容易导致语言识别混乱的现象
	Gorai 等[67]	2024	BERT-parallel CNN	Reddit Twitter	英文	BERT	Acc=0.995	采用并行结构，减少了模型需要的总体参数数量	数据样本不足存在数据偏差
	Yao 等[69]	2020	CNN	Reddit	英文	FastText GloVe	F1 = 0.966	利用 Reddit 的元数据进行个性化分析，提高了模型性能	模型可能对某些极端情况过于敏感

	Li 等[14]	2023	TCNN-MF-LA	Weibo	中文	Word2Vec	Acc=0.888 P=0.906	提出标签关联机制，考虑信息缺失问题，能综合不同特征	模型训练成本高
基于RNN	Alabdulkreem 等[70]	2021	RNN	Twitter	阿拉伯文	Word2Vec GloVe	Acc=0.72	考虑了阿拉伯语的特殊性，序列处理能力强	数据区域性太强，模型的泛化能力
	Apoorva 等[53]	2023	RNN	Twitter	英文	Golve	Acc=0.96	在处理上下文结构和分析句子任务方面较强	处理不同语境中上下文信息时存在限制，存在梯度消失问题
	Kancharapu 等[49]	2022	LSTM	Twitter	英文	N/A	Acc=0.87 F1=0.83	能发挥 LSTM 模型的处理长序列的优势	训练时间较长，存在数据偏差
	Deepa J 等[72]	2023	LSTM	Twitter	英文	N/A	Acc=0.908	序列处理能力强，能缓解梯度消失问题	对句子后续信息获取能力有限
	Almars[33]	2022	Bi-LSTM-Attention	Twitter	阿拉伯文	Word2Ve	F1=0.85	结合 Bi-LSTM 模型，前后向处理序列数据能力强	参数量大，调节困难
	Kancharapu 等[73]	2023	Bi-LSTM-ensemble	Twitter	英文	Word2Vec GloVe	Acc=0.86	可以结合 Bi-LSTM 模型和情感分析来检测与自杀相关的推文	不适用大规模数据，对异常数据敏感
	Kumar A 等[51]	2021	BiGRU_Mattn	Reddit	英文	GloVe	Micro F1=0.30	模型能较好捕捉输入序列中的局部上下文，特征多维度融合能力强	在多语言和长序列数据上的表现不佳
基于CNN和RNN组合	Mumenin 等[74]	2023	CNN-GRU	Reddit	英文	Word2Vec	Acc=0.973	可以有效协同 CNN 和 GRU 的特征提取能力，并应用了 XAI 技术简要了解预测的结果并检查模型的正确性	文本信息容易丢失丢失
	Kour 等[21]	2022	CNN-BiLSTM	Twitter	英文	自定义生成式词嵌入	Acc=0.94 AUC=0.95 F1=0.95	能够同时提取上下文语义信息和情感语义信息	句子序列长度的限制了模型性能
	Oyewale 等[75]	2024	CNN-BiLSTM	Reddit	英文	FastText Golve	F1=0.94	结合两种模型的优势，避免过拟合	只适用于小规模数据
	Priyamvada 等[76]	2023	栈式堆叠的 CNN 和双层 LSTM 混合模型	Twitter	英文	Word2Vec	Acc=0.94	扩大了模型的特征提取能力	参数量大，计算资源高

	Renjith 等[77]	2022	LSTM-Attention-CNN	Reddit	英文	Word2Vec	Acc=0.90 F1=0.93	能考虑加权输入值的不同部分之间存在的关系，可以学习自杀文本中的隐含特征	未考虑数据类别不平衡问题
	Chadha 等[16]	2022	ACL	Reddit	英文	Word2vec GloVe	Acc=0.88 F1=0.91 P=0.87	能关注所需数据的细节和特定单词	未考虑到社交媒体文本之外的特征
	Zogan 等[50]	2022	MDHAN	Twitter	英文	GloVe	Acc=0.895 F1=0.89	使用多层次注意力网络结构，可以对用户推文进行编码	未考虑话题和情感来分析用户的推
基于 Trans-former	谌志群等[78]	2020	BERT-Bi-LSTM	微博	中文	BERT	Acc= 0.92 F1=0.91	可以处理不同语境下多义词	计算复杂度较高，过于依赖预训练模型
	Baghdadi 等[32]	2022	AraElectra BERT-BV2 BERT-LV2 BERT-BV02T BERT-LV02T*	Twitter	阿拉伯文	BERT	最佳加权和指标 WSM=0.95	能够同时进行多个自然语言处理任务，模型的通用性强	数据集有地区性的限制，可移植性待验证
	Tavchioski 等[57]	2023	BERT BERTweet* mentalBERT RoBERTa	Reddit Twitter	英文	BERT	Acc=0.86 F1=0.86	用集成的方法增强原始 BERT 的能力	没有考虑情感分析，存在信息的丢失的问题
	Lin 等[79]	2024	RoBERTa-CNN	Reddit	英文	BERT	Acc=0.98 AUC=0.976 F1=0.968	加入情感分析，能提取多种有效的情感特征	结构复杂，计算陈本高
	Wang 等[80]	2023	改进的 BERT 模型	Reddit	英文	BERT	Acc=0.98	可以进行多维度情感和主题建模，对新兴自杀文本分类能力较优	BERT 模型复现和改造比较困难
	Aragón 等[81]	2023	DisorBERT	Reddit	英文	BERT	F1=0.83	可以有效区分不同的精神疾病	各类的数据量太小，模型可解释性有待加强

*代表最优模型

6 自杀意念检测中亟待解决的问题及其创新性研究

在实际应用中，仍存在一些亟待解决的问题，一系列创新性研究尝试克服这些问题，并取得了有效的检测效果。本节将归纳和讨论目前存在的八个主要问题及其创新应对方法，为未来的研究方向提供深入的洞见和指导。

首先，自杀意念检测是一个高度敏感的研究领域，难以获取大规模且高质量的全标注数据。由于自杀意念的复杂性和个体差异性，手动标注这类数据不仅成本高昂、耗时长，而且可能涉及伦理和隐私问题。且社交媒体的匿名性、动态变化性，特别是时间公开性，即设置了内容公开的时间范围，如仅一个月可见，也使得数据收集和持续更新变得更加困难。随着诸如 ChatGPT 等大

语言训练模型的出现，一些研究开始利用这些模型帮助生成共享文本语料，提供了解决此问题的新的研究途径。大语言模型可以通过给定的上下文或主题，如自杀意念、精神疾病服药等相关的讨论或者新潮的网络用语、emoji 表达，生成连贯且逼真的文本序列，从而扩充现有的数据集。亦可训练现有文本的进行重新表述或扩展，生成新的文本样本增强数据，以此增加数据集的多样性和覆盖度。在标注的过程中，大语言模型能够根据文本的上下文信息预测语料的情感倾向和潜在的自杀风险，为半自动化的标注流程提供初步的标签。除此之外，可以通过执行情感分析任务，辅助识别文本中的情绪状态，为标注过程提供参考。在初步标注后，随后由专业人员进行审核和微调，可大量节省人力资源。在实际应用上，研究人员探索利用这些模型来辅助生成或标注数据，为数据稀缺问题提供潜在的解决方案。Chen 等[82]通过引用 ChatGPT 模型将简短的推文评论信息进行自动扩充，生成高质量实验语料，规避短文本信息提供不足导致的模型分类能力差的风险。该研究使用 SVM 和 NB 两种模型进行测试，实验结果表明，在使用 ChatGPT 生成的数据进行抑郁症预测时，模型的性能均得到提升，特别是使用 NB 模型时，平均准确率从 53.34%提升到 69.34%，F1 值从 41.34%显著提升到 57.18%。Ghanadian 等[83]提出一种创新策略，利用 ChatGPT、Flan-T5 和 Llama 等先进的语言生成模型来创建用于自杀意念检测的合成数据。该研究采用了生成式对话模型，可以根据输入的上下文进行逻辑推理，并生成连贯的回复，而不仅仅是进行简单的分类判断进行数据合成。该模型还可以通过预训练和微调的方式，自适应学习大规模的对话数据中的特定语言模式和连贯的上下文信息，实现文本的自动分类和标注，而无需耗费大量时间和人力进行标注。在大语言模型带来便利的同时，会产生一些风险。一是大语言模型生成的数据的真实性有待加强。这些模型可能无法完全感知真实世界数据的复杂性和多样性，导致模型训练效果与实际应用场景存在偏差。二是存在数据偏见的问题。在训练的过程中，大语言模型可能会继承了其训练数据中的偏见，这导致生成的数据带有性别、种族或文化上的偏见，影响模型的公正性和准确性，更会引发文化冲突。三是会触发伦理道德危机。生成涉及敏感话题的数据，如自杀意念、抑郁用药等，这触及的用户个人隐私，需要确保遵守社交媒体伦理标准及规则，避免造成心理伤害或误导。

其次，数据不平衡问题一直是自杀意念检测领域的一项关键挑战，这限制了当前广泛使用的模型性能。在数据中，自杀风险人群的样本数量远少于普通或积极情绪的样本数量，且在自杀类别中某些类型的文本可能会被过度代表，又存在少数样本间的不平衡。因此，模型可能偏向于选择对多数类有区分度的特征，忽略了对少数类重要的特征，在多数类上表现良好，但在少数类上准确度不足，导致整体性能下降。在模型训练的过程中，会增加模型收敛的难度和不稳定性。重采样技术、STMOE 算法（synthetic minority oversampling technique）是常用的解决策略，但可能会导致过采样、过拟合、信息丢失等问题。因此亟需一些创新的方法来缓和。Gao 等[84]设计了焦点损失（focal loss）方法，通过重塑标准交叉熵损失，降低分配给分类良好样本的损失的权重，从而增加对难以分类样本的关注。同时采用不同的机器学习模型在类别不平衡的 YouTube 的粤语评论中进行实验。结果表明，LSTM 模型表现最好，G-mean 达 84.3%，可以有效缓解数据不平衡带来的负面影响。Li 等[85]提出了一种基于分层集成策略的深度分层集成自杀检测模型 DHE-SD（deep hierarchical ensemble model for suicide detection），通过将超过 55 万条微博的不平衡数据集划分为多个平衡的子数据集，并用这些子数据集训练基础分类器，并集成不同分类器的结果。此外，模型使用句子级屏蔽机制，来删除包含明显自杀倾向的帖子，以使模型在删除了微博树洞的帖子后仍能有效检测用户的自杀倾向，准确率超过 95%。Ben Hassine 等[86]实施了一种全新的兴趣度量（interestingness

measure）策略，通过引入五种兴趣度量标准，使用相关属性评估、增益比评估和信息增益评估三种特征选择方法来改进传统的关联分类算法（classification based on associations，CBA）。该方法可以同时关注多数类别（非自杀类别）和少数类别（自杀类别），从中选择出高度有趣的规则来增强分类过程，无需考虑数据的分布是否平衡，全局准确率达 76%，分类效果较好。但上述方法可能存在难以解释和调试、对数据分布的微小变化异常敏感、模型的复杂性可能会掩盖决策过程中的关键因素等风险，仍需定期评估和调整模型参数，持续监控模型在实际应用中的表现，并根据需要进行更新应对策略。

再者，现有研究大多基于社交媒体用户发布的所有帖子进行自杀检测任务，时间范畴较广，但在实际情况中，用户的表达和情绪状态是流动多变的，而现有研究通常忽略了这种动态性，将用户的帖子视为静态数据进行分析。自杀意念是一个情感累积的过程，也可能会因为战争、家庭变故演变成应激过程。这种静态的处理方式可能无法准确感知用户情绪的即时变化和短期波动，会导致重要信息的丢失，无法及时识别用户高风险状态。基于此，应用时间序列分析方法，考虑帖子发布的时间顺序和频率，以识别情绪变化的趋势和长短期追踪用户行为，分析其情绪变化的连贯性和一致性是重要的解决途径。Sawhney 等[87]考虑了时间对用户整体自杀意念变化的影响，提出了一种基于时序感知网络的自杀评估框架 STATENet（suicidality assessment time-aware temporal network），通过对用户历史推文和上下推文联系进行建模，综合评估时间对用户情感变化的影响，准确率达 85.1%，但没有考虑用户的阶段性自杀意念变化。基于此，Sawhney 等[88]设计了阶段性自杀意识及情绪发展检测模型 PHASE（phase aware suicidality identification emotion progression），采用基于上下文的历史情感表示、基于时间敏感的情感 LSTM 模型和相位自适应卷积技术，对用户每个阶段进行建模，深入分析用户的历史推文，推理出用户的历史情感图谱，可以提取出用户跨度十年的阶段性情感特征，准确率达 85.6%，但仍难以区分偶尔有自杀倾向的用户，该类自杀意念是用户在短时间内形成的。Yohapriyaa 等[89]发现了时间序列数据中的事件段时间间隔变化对个体情绪变化的影响，尝试利用 LSTM-CNN 组合模型，基于文本时间戳等信息来分析数据以检测抑郁症程度，准确率超过 78%。自杀意念是一个抽象的概念，即使是有诸如抑郁症这类可以通过临床诊断的疾病辅助判断，但并不意味着自杀风险状态在每时每刻都存在。虽然目前该研究方向取得进展，但仍存在提取和建模时间及其对情绪的影响进行代表性的特征程度不高的现象，还需进一步增强研究。除此之外，社交媒体数据中存在的的噪声和缺失值会导致数据质量降低、模型误判、数据可用性不高、处理难度大等问题，这会对时间序列分析的准确性产生会产生影响。

此外，自杀倾向的表达具有隐晦性，语义存在歧义性。这种表达的特征往往不明显，用户可能会使用隐喻、比喻或模糊的语言来传达内心的痛苦和绝望。此外，语义上的歧义性使得同样的词汇或短语在不同的上下文或文化背景下可能有不同的解释。用户的表达方式受到用户性格、文化背景、语言习惯、宗教信仰和社交环境等多重因素的影响，这些因素共同作用于文本特征，增加了自杀检测的复杂性。如果不挖掘文本暗藏的深层含义特征，可能导致识别率降低。引入词典特征的一个流行的解决方案。Ji 等[52]采用了一种基于注意力机制的关系网络模型 RN（attentive Relation Networks），可以结合基于情感词典的文本表示和潜在主题，并通过关系网络进行编码，有效推理自杀风险线索及提取，增强文本表示，各项评估指标均超过 83%。Kodati 等[90]构建了两种深度学习模型，采用基于词典特征的多头注意力机制，将输入的特征与长序列的文本词汇特征结合来捕获多种类型的情感以此甄别自杀相关文本中关键的负面情感。两个模型的准确率均突破 98%，识别效果显著。词典特征通常基于预

先定义的规则和已知的心理语言学知识构建，允许用户探索不同词典条目对模型输出的影响，可以为模型的决策过程提供一定程度的解释透明度。

值得关注的是，在网络发展日趋成熟的进程中，新兴的网络用语和特定社群的表达方式不断涌现，传统词典更新速度滞后，词典特征可能无法覆盖专业领域术语和新颖术语，无法充分理解术语背后的深层含义和情感色彩。最新的大型语言模型，如ChatGPT，可以大规模数据上进行训练，能迅速推理和解释丰富的语言变体和新兴词汇，在处理更宽泛的词汇和表达上更接近人类的解释，且拥有定期更新词典机制，可以追踪社交媒体动态，实时纳入新兴术语和网络用语，以保持词典的时效性和覆盖度。在应用中，Yang 等[91]在探索了将情感增强思考链与 ChatGPT 进行耦合，以提高模型的对心理健康相关的语言及其上下文结构的解析能力。与其他基线大语言模型相比，该模型的性能最优，准确率和 F1 值均达到 84%以上，但存在推理不准确的问题。Lan 等[92]首次将专业医学知识与先进的大型语言模型 ChatGPT3.5-Turbo 相结合提出了抑郁症检测框架 DORIS。该框架在医学领域知识的决策与支撑下，从用户的情绪记录中提取和总结情绪强度高和关键信息的文本来形成情绪历程，从而为每个用户构建专业特征，实现了抑郁症检测结果的高可解释性，AUC 值高达 97.1%。但目前大语言模型仍存在不准确的推理问题，尤其是对上下文的关键信息解析。尽管大语言模型在处理社交媒体语言变体和新兴词汇方面表现出色，但可能在解析关键上下文信息时存在局限，特别是在理解长篇幅文本和多层次语义时，将专业心理健康、医学知识与大型语言模型结合可能面临知识融合和上下文适配的挑战，影响模型的准确性和可靠性。

同时，随着社会的发展，在社交媒体平台上用户的行为模式正在发生变化。相比于过去积极发帖和评论，越来越多的用户选择成为“网络缄默者”，他们主要浏览和阅读内容而不主动分享自己的想法和信息。这种趋势导致文本数据仓库逐渐萎缩，更新速度放缓，文本特征愈发不明显。因此，可以考虑将一些有价值的社交因素加入到模型中，例如社交网络关系。社交网络关系在心理健康分析领域的重要性日益凸显，尤其是在社交媒体文本数据稀缺的情况下。社交网络关系不仅包括用户的直接联系，如关注、点赞，还涵盖了由用户互动形成的复杂网络结构。这些关系提供了一种理解用户行为和情绪状态的新途径。例如，通过分析用户在社交网络中的位置和连接，我们可以评估他们获得的社交影响力和支持水平。此外，这种网络结构一定程度下反应社交亲密度和互动模式，可以揭示用户的情绪变化和潜在的心理状态，有助于及时发现和干预心理健康问题。例如，苗红闪[93]引入了微博社交网络关系特征，采用改进的 PageRank 模型得到的节点值，迭代计算出微博社交网络上的用户影响力来识别潜在的抑郁症用户，但只考虑了微博用户之间的关注关系，提取出的社交网络关系特征尚显不足。Meng 等[94]提出了 TCNN-SN 模型(text convolutional neural network based on social network relationships)，利用社交网络关系特征，在加权线性融合框架下引入修正因子，增强了特征的提取，可以从检测关键的用户个体从而扩展识别更多隐藏的自杀用户人群。该模型准确率、F1-score 和 AUC 指标分别达到了 88.6%，88.8%和 94%，展示了卓越的检测能力。随着时间的推移，用户的社交关系和互动模式可能发生变化，导致旧数据过时，不再反映当前状态，且用户往往与观点相似的人建立联系，这可能导致“信息泡沫”现象，限制了数据的多样性和代表性。这要求分析模型能够适应快速变化的网络拓扑。目前该研究方向文献较少，是一个值得深入探讨的研究方向。

另一方面，在当前的自杀意念检测研究中，尽管文本信息是主要的关注点，但大多数推文往往包含丰富的图像、音视频等多模态数据，可以提供丰富的自杀线索。图像和视频数据，例如用户清晰的面部表情和发布图片的颜色组合、亮度、灰度和饱和度，都能一定程度上表明用户的心理状态[95]。音

频数据，包括音频片段和语音记录可以提供听觉的线索，例如音量、音调、语速、流畅度和颤音等。然而，将不同模态的数据有效融合，以获得更准确的分析结果是一个技术挑战，需要结合计算机视觉、语音处理和自然语言处理等多个领域的技术。部分学者开始探索多模态内容的检测方法。Cao 等[30]采用了 SDM 框架，通过利用面向自杀的词嵌入和分层注意力机制来检测包含图片微博的潜在自杀风险，准确率高达 91%，证明了多模态数据的有效性。Gui 等[96]尝试采用结合文本和视觉图像两个维度的强化学习方法 COMMA（cooperative misoperation multi-agent），可以同时联合两个特征进行分析学习，各项指标均超过 90%，可以有效检测抑郁症。Ramírez-Cifuentes 等[97]提出了 SNPSY 模型（social networks and psychological），通过从多个社交平台提取的文本、图像、行为和关系等多模态数据来检测西班牙用户的自杀风险，特别是对推文图片的分析，可以有效提取自杀迹象特征，AUC 值攀升至 92%。尽管目前对音频、视频的自杀意念检测的研究相对丰富，但在社交媒体环境中，该类内容发布相对罕见，导致相关数据的收集面临重大挑战，限制了模型的训练和验证。且多模态数据的分析需要集成多种技术，如何有效地融合来自不同模态的特征，以提高检测的准确性，也是一个技术难题。目前相关研究较少，是一项未来值得填补的研究空白。

最后，自杀意念检测模型往往针对特定平台的用户行为和语言表达进行优化，这可能导致模型对特定文化背景下的语言模式和社交习惯过于敏感。当尝试将这些模型应用于其他社交媒体平台或不同文化环境时，由于语言使用、社交规范和用户行为的差异，将模型移植到其他社交媒体平台可能会面临挑战，模型跨文化和跨平台的普适性有待加强。通过耦合多个社交媒体平台，增加数据来源的多样性，研究不同文化在自杀意念表达上的差异与共性，探讨文化对自杀意念表达的影响是有必要的。例如，Shen 等[98]考虑到国家背景和文化的多样性，提出了一种基于特征自适应转换的跨领域深度神经网络模型 DNN-FATC（deep neural network model with feature adaptive transformation and combination strategy），采用多源域异构的文本数据，通过特征归一化与对齐、特征的组合与转换的方式来弥合不同文化特征的异质性，可以实现跨文化地域的抑郁症自杀倾向检测，F1 值达 78.5%。Mbarek 等[99]考虑到一个用户可能存在多个社交媒体平台账号，通过获取他们在不同平台上可用的共享文本内容作为数据集，并使用决策树模型进行检测，准确率达 85.4%，证明了从多个社交媒体数据源来描绘同一用户全面的画像，建设完整的用户资料库，进而精准检测自杀风险的可行性。但由于一些潜在的自杀倾向用户可能不会在其他社交网络上提及自己的身份，或者只在单一社交网络的账户上活动，这使得寻找这些用户的匹配档案变得困难，导致导致数据集的覆盖度不足。且对于低资源语言，缺乏成熟的语言、文化特征提取方法，使得基本的文本处理任务，如分词、词性标注和句法分析，都变得具有挑战性，产生了文化和技术隔阂。

自杀意念检测任务中亟待解决的问题及其创新性解决算法归纳在表 6 中。

表 6　目前亟待解决的问题及其创新性解决算法总结

Table 6　Summary of current urgent problems and innovative solution algorithms

问题	文献	年份	算法	社交媒体	评价指标	创新点	缺点
难以获取大规模、高质量、标注好的数据	Chen 等[82]	2023	ChatGPT+SVM* ChatGPT+NB	Twitter	Acc=0.893 F1=0.907	利用 ChatGPT 模型扩充推文短文评论信息，自动生成高质量实验语料，模型性能得到显著提升	增加样本量较小，模型泛化能力有限
	Ghanadian 等[83]	2024	ChatGPT	Reddit	F1=0.85	采用了生成式对话模型的方法，可以根据输入的上下文生成连贯的回复，产生大规模数据，不需要耗费	合成数据可能会与真实世界数据存在差异，需要专家和专

						大量人力资源	业领域知识的支撑
数据不平衡	Gao 等[84]	2019	LSTM	YouTube	G-mean[1]=0.843	提出焦点损失方法，通过减少对已经被模型正确分类的样本的关注来增加对难以分类样本的关注，从而提高模型对少数类的识别能力，可以有效解决数据不平衡的问题	仅测试了对粤语分类的有效性，模型对特定语言和语境的适应性不强
	Li 等[85]	2022	DHE-SD	新浪微博	Acc=0.958 F1=0.939	提出基于层级集成策略的模型，能将不同数据的样本进行分层划分并集成结果，更好地提取、补充少数类别的特征	数据标注过程中存在表达不规范、符号缺失、主观性等问题，使得提出的模型难以在实践中应用和推广
	Ben Hassine 等[86]	2022	改进的 CBA	Twitter	GAcc[2]=0.76 F1=0.71	通过引入五个新的兴趣度量标准，优化传统的关联分类算法，无需考虑数据分布不平衡的情况，避免使用人工过采样或欠采样技术，如 SMOTE 等带来的风险	提出的方法主要集中在解决数据不平衡的问题上，文本数据的分析、分类方面有待加强
忽略用户的自杀意念随着时间动态变化，容易影响检测结果	Sawhney 等[87]	2020	STATENet	Twitter	Acc=0.851 R=0.81 F1=0.799	提出了一种基于时序感知网络的自杀评估框架 STATENet,可以有效评估时间对用户整体自杀意念变化的影响	没有考虑用户的阶段性自杀意念变化
	Sawhney 等[88]	2021	PHASE	Twitter	Acc=0.856 Macro F1=0.805 R=0.812	根据历史推文对用户每一个情感阶段进行建模，开发阶段性自杀意识及情绪发展检测模型，自适应地学习基于阶段的情感上下文特征，精准考虑用户阶段性的自杀意念变化	提出方法假设连续推文之间的时间间隔是均匀的，但实际情况下，用户发布推文的时间间隔可能是不规则的，可能影响对用户自杀性进展的评估
	Yohapriyaa 等[89]	2022	LSTM-CNN	Twitter	Acc=0.78	提出了一种基于文本时间戳的文本分析模型，可以通过分析时间序列数据中的事件段时间间隔变化对个体情绪变化的影响检测抑郁程度	不适用于大规模的数据集，模型计算能力有限，可解释性有待提升
自杀倾向表达隐晦性强，语义具有歧义性，仅依靠浅层文本特征识别率较低	Ji 等[52]	2021	RN	Reddit Twitter	Acc=0.839 F1=0.838 P=0.838 R=0.839	提出了一种基于注意力机制的关系网络模型 RN,通过引入基于情感词典的文本表示和潜在主题并通过关系网络进行编码，提取与自杀风险相关的特征，增强文本表示	使用情感词典和主题模型的预处理过程可能会引起错误传播
	Kodati 等[90]	2022	C-BiGRU-MHA-CNN* L-BiLSTM-MHA-CNN	Reddit	Acc=0.981 F1=0.959 P=0.969 R=0.948	提出基于词典特征的多头注意力机制，采用两种深度学习模型，通过将输入的特征与长序列的文本词汇特征耦合来识别自杀相关文本中关键的负面情感	模型参数太多，计算需求大
词典特征可能无法覆盖专业	Yang 等[91]	2023	基于情感增强思考链的	Reddit Twitter	F1=0.842 R=0.857	尝试将情感增强思考链与 ChatGPT 进行耦合，以提高模型对心理健康	存在不准确的推理问题，尤其是在长

领域术语和新颖术语，特别是新兴的网络用语和特定社群的表达			ChatGPT 模型			相关的语言及其上下文结构的解析能力，在心理健康分析任务上的表现优于其他测试的大语言模型	文本的上下文中可能会忽略关键信息
	Lan 等[92]	2024	DORIS	新浪微博	AUC=0.971 F1=0.76 P=0.76 R=0.79	以专业的医学知识赋能大语言模型，从用户的情绪记录中提取和总结情绪强度高和关键信息的文本，形成情绪历程，构建用户专属的专业特征，生成准确且可解释的抑郁症诊断	实验基于单个数据集，不具有代表性
用户不主动分享内容，文本数据仓库逐渐萎缩，文本特征愈发不明显	苗红闪[93]	2020	Ptr-Rank	新浪微博	Acc=0.895	通过考虑用户社交网络关系，改进 PageRank 算法，迭代计算用户在社交网络上的影响力，可以识别出微博文本自杀特征表达不明显的微博用户	只考虑了用户之间的关注关系，假设每个用户之间都是平等的，模型提取出的特征不明显
	Meng 等[95]	2024	TCNN-SN	新浪微博	Acc=0.886 AUC=0.94 F1=0.888	在用户社交网络关系的基础上，加入一系列有价值的社交行为因素，如网络活跃度、博客夜间发布频率、网络置信度等，并将它们归纳为修正因子引入模型中，增强社交网络关系特征，提高模型性能	模型主要聚焦静态网络结构，需要考虑动态网络结构来增强网络关系特征
大部分研究主要使用文本特征，忽略了其他可能影响自杀意念检测的多模态数据，如图像、视频、音频等	Cao 等[30]	2019	SDM	新浪微博	Acc=0.913 F1=0.909	采用基于面向自杀的词嵌入和分层注意力机制的模型，可以有效处理包含图片的微博，以检测微博用户的潜在自杀风险	模型对图像中的内容理解需要高级的语义分析能力，模型分析能力有待提升
	Gui 等[96]	2019	COMMA	Twitter	Acc=0.9 F1=0.9 P=0.9 R=0.901	首次提出结合文本和视觉信息的多模态强化学习方法，可以同时联合两个特征进行分析学习来检测抑郁症	模型可能对特定类型的文本和图像过于敏感，导致误报或漏报
	Ramírez-Cifuentes 等[97]	2020	SNPSY	Reddit Twitter Instagram	AUC=0.92 P=0.88	提出了一个综合的社交媒体文本、图像、行为和关系等多模态的自杀风险检测模型，可以提取用户的多模态表征，有效提取自杀迹象特征	模型的决策过程可能不够透明，尤其是在涉及图像内容时，模型的可解释性较差
模型的跨文化和跨平台普适性不高	Shen 等[98]	2018	DNN-FATC	新浪微博 Twitter	F1=0.785	系统地分析了跨文化领域抑郁症相关特征模式，从文化异构性和发散性的角度出发，可以自适应学习不同文化的语言结构和特征	由于文化差异，同一特征在不同语言领域可能对模型有不同的影响，甚至可能在不同平台上有负面影响
	Mbarek 等[99]	2022	DT	Twitter Tumblr YouTube	F1=0.854	考虑同一个用户拥有多个社交媒体平台的个人账号，并分析他们在多个平台数据源上可用的共享内容，提取不同类型的特征，描绘完整的用户自杀画像	研究使用的自杀用户数据集有限，影响了模型的泛化能力和检测准确性

*代表最优模型

[1]G-mean 是一种常用的不平衡数据性能评估方法

[2] 全局精度（Global Accuracy, GAcc）是一种常用的度量分类器整体准确率的指标。

7 总结与展望

迄今为止，自杀仍然是全球的主要死亡原因之一。在这一背景下，将机器学习作为技术支撑、社交媒体作为数据基础，形成数据驱动、人机协同、跨学科融合、跨文化畛域的数智化自杀意念检测新范式，已然成为一个至关重要的研究领域。本文全面回顾了在社交媒体平台上应用机器学习技术来检测自杀意念用户的相关研究脉络与发展趋势；总结了现有的数据集，从传统机器学习和深度学习两个角度讨论了不同社交媒体平台上的机器学习算法及优缺点，并归纳了目前领域中亟待解决的问题及创新性解决方法。这是近三年来国内首篇从机器学习的视角对自杀意念检测的方法、模式和应用进行全面总结的综述。

下面将在现有的研究基础上进行分析，梳理出的目前研究的局限性并将其作为未来的研究方向。

7.1 现有研究的局限性

如第 6 节所述，尽管该领域内存在亟待解决的问题已有对应的解决策略，但目前大部分研究都普遍存在以下几个方面的局限性。

（1）公共数据集的匮乏。自杀领域因其敏感性和隐私性，长期缺乏具有公信力和权威性的数据集。目前大多数研究的数据集以英文为主。由于汉语和汉字的语言文字系统的复杂性，针对中国社交媒体的自杀监测数据集更为罕见。若使用英文数据集对中国人进行检测，可能会导致检测准确率下降。

（2）BERT 计算资源的高需求。BERT 作为一个庞大的模型，拥有数亿个参数，其训练和推理过程需要大量的计算资源和时间。目前，大部分研究只能在已公开的或者训练完成的模型基础上进行微调。在资源受限的环境下，使用 BERT 可能会面临重大挑战。

（3）对时间因素的忽视。现有方法主要关注用户在特定时间节点的自杀意念，无法有效区分长期、短期和偶尔产生自杀意念的用户。现实研究表明，用户的自杀意念会随着时间动态变化，且在每个阶段展现出不同的情感特征。这是一个重要且复杂的问题。

（4）基于社交媒体的多模态检测模型研究的缺乏。虽然目前结合图像、音频和视频等多模态模型不胜枚举。但在社交媒体领域，目前只有涉及图像的研究，对音频、视频的研究仍处尚未充分开发的阶段，因为涉及个人隐私，用户和平台不会发布相关的语音和面部视频语料，数据的收集是一项重要的挑战。此外，图像和其他视觉数据通常需要专业的标注，这既耗时又昂贵。

7.2 未来研究方向

为了克服现有研究的局限性，未来研究方向可以考虑以下几个方面。

（1）数据集的构建与扩充。在遵循伦理和隐私保护原则下，可以考虑使用 ChatGPT 等大语言模型模拟用户在社交媒体上的对话，生成大规模数据。同时，通过数据增强和迁移学习等技术来合成或转换中文数据，扩充训练数据的规模和多样性，以构建专门的中文自杀检测数据集。此外，除了文本数据，还可以考虑整合其他模态的数据，譬如图像、视频、音频等，构建多模态的自杀检测数据集来解决自杀领域数据资源匮乏的问题。

（2）模型的优化与资源节约。针对 BERT 计算需求高的问题，未来的研究可以通过模型压缩和优化、分布式训练和推理、预训练模型的迁移学习、云计算和 GPU 加速以及模型量化和部署优化等方式，提高 BERT 模型的可行性和实用性，从而节省大量的计算资源和时间。

（3）时间因素的深入考量与分析。未来的研究可以通过长期监测和跟踪，对每个用户的历史数据进行深入考量、分析、存储和建模。通过建立时间序列模型和历史情感知识图谱，分别对不同阶段的自杀意念进行检测和分类，更精准地描绘、衡量用户的自杀意念的动态变化。除了文本数据，还可

以考虑结合其他相关信息源，如社交媒体上的帖子发布时间戳、用户的社交行为数据等，更全面理解和分析时间对用户自杀意念的影响。

（4）多模态模型的深入探索与研究。由于多模态数据有限，可以尝试研究多模态数据增强的方法，结合不同模态的特点，例如，使用图像内容来指导文本描述的生成，或者基于文本内容来生成相关的图像。数据增强不仅可以应用于单一模态，如图像或文本，还可以跨模态进行，以增强数据的多样性和丰富性。还可以利用 ChatGPT 等大语言模型根据用户对话来帮助生成和标注相关的图音像数据，扩充训练数据仓库。此外，条件生成对抗网络（generative adversarial networks, GAN）也是一个值得考虑的解决方案，可以尝试根据给定的文本描述生成相应的图像，或者根据音频样本生成具有特定特征的面部表情图像。

参考文献：

[1] ZHANG T, YANG K, JI S, et al. Emotion fusion for mental illness detection from social media: A survey[J]. Information Fusion, 2023, 92: 231-246.

[2] World Health Organization. World Health Statistics 2024: Monitoring Health for the SDGs, Sustainable Development Goals[M]. Switzerland: World Health Organization, 2024.

[3] FRANKLIN J C, RIBEIRO J D, FOX K R, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research[J]. Psychological bulletin, 2017, 143(2): 187.

[4] ALDHYANI T H H, ALSUBARI S N, ALSHEBAMI A S, et al. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models[J]. International journal of environmental research and public health, 2022, 19(19): 12635.

[5] XUE Y, LI Q, WU T, et al. Incorporating Stress Status in Suicide Detection through Microblog[J]. Computer Systems Science and Engineering, 2019, 34(2):65-78.

[6] LINDH Å U, BECKMAN K, CARLBROG A, et al. Predicting suicide: A comparison between clinical suicide risk assessment and the Suicide Intent Scale[J]. Journal of affective disorders, 2020, 263: 445-449.

[7] CALEAR A L, BATTERHAM P J. Suicidal ideation disclosure: Patterns, correlates and outcome[J]. Psychiatry research, 2019, 278: 1-6.

[8] 王呈珊, 宋新明, 朱廷劭, 等. 一位自杀博主遗言评论留言的主题分析[J]. 中国心理卫生杂志, 2021, 35(2): 121-126.
WANG C S, SONG X M, ZHU T S, et al. An analysis of the theme of a suicide blogger's comment[J]. Chinese Mental Health Journal, 2021, 35(2): 121-126.

[9] SIERRA G, ANDRADE-PALOS P, BEL-ENGUIX G, et al. Suicide risk factors: a language analysis approach in social media[J]. Journal of language and social psychology, 2022, 41(3): 312-330.

[10] SHAH S, KADAM S, PANDHARE S, et al. Suicidal Thoughts Prediction from Social Media Posts using Machine Learning and Deep Learning[J]. Quest Journal of Electronics and Communication Engineering Research, 2022, 8(5): 64-71.

[11] DEWANGAN D, SELOT S, PANICKER S. The Accuracy Analysis of Different Machine Learning Classifiers for Detecting Suicidal Ideation and Content[J]. Asian Journal of Electrical Sciences, 2023, 12(1): 46-56.

[12] HIRAGA M. Predicting depression for japanese blog text[C]//Proceedings of ACL 2017, student research workshop, Vancouver, Canada, July 30 - August 4, 2017. Stroudsburg: ACL, 2017:107-113.

[13] KUMAR E R, VENKATRAM N. Predicting and analyzing suicidal risk behavior using rule-based approach in Twitter data[J]. Soft computing, 2023: 1-9.

[14] LI Z, CHENG W, ZHOU J, et al. Deep learning model with multi-feature fusion and label association for suicide detection[J]. Multimedia systems, 2023, 29(4): 2193-2203.

[15] TADESSE M M, LIN H, XU B, et al. Detection of suicide ideation in social media forums using deep learning[J]. Algorithms, 2019, 13(1): 7.

[16] CHADHA A, KAUSHIK B. A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data[J]. New Generation Computing, 2022, 40(4): 889-914.

[17] HECKLER W F, DE CARVALHO J V, BARBOSA J L V. Machine learning for suicidal ideation identification: A systematic literature review[J]. Computers in Human Behavior, 2022, 128: 107095.

[18] ABDULSALAM A, ALHOTHALI A. Suicidal ideation detection on social media: A review of machine learning methods[J]. arXiv:2201.10515, 2022.

[19] HASIB K M, ISLAM M R, SAKIB S, et al. Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey[J]. IEEE Transactions on Computational Social Systems, 2023, 10(4): 1568-1586.

[20] LIU D, FENG X L, AHMED F, et al. Detecting and measuring depression on social media using a machine learning approach: systematic review[J]. JMIR Mental Health, 2022, 9(3): e27244.

[21] KOUR H, GUPTA M K. An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM[J]. Multimedia Tools and Applications, 2022, 81(17): 23649-23685.

[22] KABIR M, AHMED T, HASAN M B, et al. DEPTWEET: A typology for social media texts to detect depression severities[J]. Computers in Human Behavior, 2023, 139: 107503.

[23] NOVA K. Machine learning approaches for automated mental disorder classification based on social media textual data[J]. Contemporary Issues in Behavioral and Social Sciences, 2023, 7(1): 70-83.

[24] SURYAWANSHI C, TAMBOLI T, TAYADE S, et al. Detection of Depression or Sentiment Analysis[J]. International Journal of Scientific Research in Science and Technology, 2020, 5(8), 162-169.

[25] TALAAT F M, EL-GENDY E M, SAAFAN M M, et al. Utilizing social media and machine learning for personality and emotion recognition using PERS[J]. Neural Computing and Applications, 2023, 35(33): 23927-23941.

[26] GHOSH T, AL BANNA M H, AL NAHIAN M J, et al. An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla[J]. Expert Systems with Applications, 2023, 213: 119007.

[27] JI S, Yu C P, FUNG S, et al. Supervised learning for suicidal ideation detection in online user content[J]. Complexity, 2018, 2018: 1-10.

[28] NIKHILESWAR K, VISHAL D, SPHOORTHI L, et al. Suicide ideation detection in social media forums[C]//2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, October 7-9, 2021. Piscataway: IEEE, 2021: 1741-1747.

[29] CHATTERJEE M, SAMANTA P, KUMAR P, et al. Suicide ideation detection using multiple feature analysis from Twitter data[C]//2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, February 11-13, 2022. Piscataway: IEEE, 2022: 1-6.

[30] CAO L, ZHANG H, FENG L, et al. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention[J]. arXiv:1910.12038, 2019.

[31] CAO L, ZHANG H, WANG X, et al. Learning users inner thoughts and emotion changes for social media based suicide risk detection[J]. IEEE Transactions on Affective Computing, 2021, 14(2): 1280-1296.

[32] BAGHDADI N A, MALKI A, BALAHA H M, et al. An optimized deep learning approach for suicide detection through Arabic tweets[J]. PeerJ Computer Science, 2022, 8: e1070.

[33] ALMARS A M. Attention-Based Bi-LSTM Model for Arabic Depression Classification[J]. Computers, Materials & Continua, 2022, 71(2): 3091-3106.

[34] NARYNOV S, MUKHTARKHANULY D, KERIMOV I, et al. Comparative analysis of supervised and unsupervised learning algorithms for online user content suicidal ideation detection[J]. Journal of Theoretical and Applied Information Technology, 2019, 97(22): 3304-3317.

[35] VALERIANO K, CONDORI-LARICO A, SULLA-TORRES J. Detection of suicidal intent in Spanish language social networks using machine learning[J]. International Journal of Advanced Computer Science and Applications, 2020, 11(4): 688-695.

[36] COPPERSMITH G, DREDZE M, HARMAN C, et al. CLPsych 2015 shared task: Depression and PTSD on Twitter[C]//Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, Denver, Colorado, June 5, 2015. Stroudsburg: ACL, 2015: 31-39.

[37] ZIRIKLY A, RESNIK P, UZUNER O, et al. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts[C]//Proceedings of the sixth workshop on computational linguistics and clinical psychology, Minneapolis, June 6, 2019. Stroudsburg: ACL, 2019: 24-33.

[38] GAUR M, ARIBANDI V, ALAMBO A, et al. Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS[J]. PloS one, 2021, 16(5): e0250448.

[39] GAUR M, ALAMBO A, SAIN J P, et al. Knowledge-aware assessment of severity of suicide risk for early intervention[C]//The world wide web conference San Francisco, May 13-17, 2019. New York: ACM, 2019: 514-525.

[40] LOSADA D E, CRESTANI F, PARAPAR J. eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations[C]//Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017. Cham: Springer, 2017: 346-360.

[41] LOSADA D E, CRESTANI F, PARAPAR J. Overview of eRisk: early risk prediction on the internet[C]// Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018. Cham: Springer, 2018: 343-361.

[42] PARAPAR J, MARTÍN-RODILLA P, LOSADA D E, et al. Overview of erisk 2023: Early risk prediction on the internet[C]//International Conference of the Cross- Language Evaluation Forum for European Languages, Thessaloniki, Greece, September 18–21. Cham: Springer, 2023: 294-315.

[43] CHEN Y, LI Y, WEN M. Chinese Psychological QA Database and its Research Problems[C]//2022 9th International Conference on Dependable Systems and Their Applications (DSA), Wulumuqi, China, August 4-5, 2022. Piscataway: IEEE, 2022: 786-792.

[44] LIU D, FU Q, WAN C, et al. Suicidal ideation cause extraction from social texts[J]. IEEE Access, 2020, 8: 169333-169351.

[45] ZHANG T, YANG K, JI S, et al. SuicidEmoji: Derived Emoji Dataset and Tasks for Suicide-Related Social Content[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, USA, July 14-18, 2024.New York: ACM, 2024: 1136-1141.

[46] RABANI S T, KHAN Q R, KHANDAY AMUD. Detection of suicidal ideation on Twitter using machine learning & ensemble approaches[J]. Baghdad science journal, 2020, 17(4): 1328-1328.

[47] HUANG Y, LIU X, ZHU T. Suicidal ideation detection via social media analytics[C]//Human Centered Computing: 5th International Conference, HCC 2019, Čačak, Serbia, August 5-7, 2019. Cham: Springe, 2019: 166-174.

[48] ZHANG D, ZHOU L, TAO J, et al. KETCH: A Knowledge-Enhanced Transformer-Based Approach to Suicidal Ideation Detection from Social Media Content[J]. Information Systems Research, 2024.

[49] KANCHARAPU R, SRINAGESH A, BHANUSRIDHAR M. Prediction of Human Suicidal Tendency based on Social Media using Recurrent Neural Networks through LSTM[C]//2022 International Conference on Computing, Communication and Power Technology (IC3P), Visakhapatnam, India, January 7-8, 2022. Piscataway: IEEE, 2022: 123-128.

[50] ZOGAN H, RAZZAK I, WANG X, et al. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media[J]. World Wide Web, 2022, 25(1): 281-304.

[51] KUMAR A, TRUEMAN T E, ABINESH A K. Suicidal risk identification in social media[J]. Procedia Computer Science, 2021, 189: 368-373.

[52] JI S, LI X, HUANG Z, et al. Suicidal ideation and mental disorder detection with attentive relation networks[J]. Neural Computing and Applications, 2022, 34(13): 10309- 10319.

[53] GHOSAL S, JAIN A. Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier[J]. Procedia Computer Science, 2023, 218: 1631-1639.

[54] BOONYARAT P, LIEW D J, CHANG Y C. Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19[J]. Information Processing & Management, 2024, 61(4): 103706.

[55] METZLER H, BAGINSKI H, NIEDERKROTENTHALER T, et al. Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach[J]. Journal of medical internet research, 2022, 24(8): e34705.

[56] YANG Q, ZHOU J, WEI Z. Time Perspective-Enhanced Suicidal Ideation Detection Using Multi-Task Learning[J]. International Journal of Network Dynamics and Intelligence, 2024: 100011-100011.

[57] TAVCHIOSKI I, ROBNIK-ŠIKONJA M, POLLAK S. Detection of depression on social networks using transformers and ensembles[J]. arXiv:2305.05325, 2023.

[58] HERATH S, WIJAYASIRIWARDHANE T K. A Social Media Intelligence Approach to Predict Suicidal Ideation from Sinhala Facebook Posts[C]//2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)., Colombo, Sri Lanka, April 4, 2024. Piscataway: IEEE, 2024, 7: 1-6.

[59] CHADHA A, GUPTA A, KUMAR Y. Suicidal ideation detection on social media: a machine learning approach[C]//2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, October 10-12, 2022. Piscataway: IEEE, 2022: 685-688.

[60] RABANI S T, KHANDAY A M U D, KHAN Q R, et al. Detecting suicidality on social media: Machine learning at rescue[J]. Egyptian Informatics Journal, 2023, 24(2): 291-302.

[61] SAIFULLAH S, DREŻEWSKI R, DWIYANTO F A, et al. Sentiment analysis using machine learning approach based on feature extraction for anxiety detection[C]//International Conference on Computational Science, Prague, Czech Republic, July 3-5, 2023. Cham: Springer, 2023: 365-372.

[62] LIU J, SHI M. A hybrid feature selection and ensemble approach to identify depressed users in online social media[J]. Frontiers in Psychology, 2022, 12: 802821.

[63] LEKKAS D, KLEIN R J, JACOBSON N C. Predicting acute suicidal ideation on Instagram using ensemble machine learning models[J]. Internet interventions, 2021, 25: 100424.

[64] FARRUQUE N, GOEBEL R, SIVAPALAN S, et al. Depression symptoms modelling from social media text: A semi-supervised learning approach[J]. arXiv:2209.02765, 2022.

[65] SHARMEEN R, KHAN S, SANJANA T, et al. Suicidal Ideation Detection Using Semi-Supervised Learning Technique: Self-Training Classifier[C]//2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI). Dhaka, Bangladesh, December 9-10, 2023. Piscataway: IEEE, 2023: 1-6.

[66] BENIWAL R, SARASWAT P. A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data[J]. The Computer Journal, 2024: bxae018.

[67] GORAI J, SHAW D K. A BERT-encoded ensembled CNN model for suicide risk identification in social media posts[J]. Neural Computing and Applications, 2024, 36(18): 10955-10970.

[68] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv:1408.5882, 2014.

[69] YAO H, RASHIDIAN S, DONG X, et al. Detection of suicidality among opioid users on reddit: machine learning–based approach[J]. Journal of medical internet research, 2020, 22(11): e15293.

[70] ALABDULKREEM E. Prediction of depressed Arab women using their tweets[J]. Journal of Decision Systems, 2021, 30(2-3): 102-117.

[71] APOORVA A, GOYAL V, KUMAR A, et al. Depression detection on twitter using RNN and LSTM models[C]// International Conference on Advanced Network Techno- logies and Intelligent Computing, Varanasi, India, De- cember 22-24, 2022. Cham: Springer, 2022: 305-319.

[72] DEEPA J, SHRIRAAMAN S, SHRUTI V V, et al. Detecting and Determining Degree of Suicidal Ideation on Tweets Using LSTM and Machine Learning Models[J]. Journal of Survey in Fisheries Sciences, 2023, 10(2S): 3217-3224.

[73] KANCHARAPU R, A AYYAGARI S N. A comparative study on word embedding techniques for suicide prediction on COVID-19 tweets using deep learning models[J]. International Journal of Information Technology, 2023, 15(6): 3293-3306.

[74] MUMENIN N, BASAR M R, HOSSAIN A B M K, et al. Suicidal Ideation Detection from Social Media Texts Using an Interpretable Hybrid Model[C]//2023 6th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, February 13, 2024. Piscataway: IEEE, 2023: 1-6.

[75] OYEWALE C T, IBITOYE A O J, AKINYEMI J D, et al. Suicide Ideation Prediction Through Deep Learning: An Integration of CNN and Bidirectional LSTM with Word Embeddings[C]//Science and Information Conference. Cham: Springer, 2024: 271-283.

[76] PRIYAMVADA B, SINGHAL S, NAYYAR A, et al. Stacked CNN-LSTM approach for prediction of suicidal ideation on social media[J]. Multimedia Tools and Applications, 2023, 82(18): 27883-27904.

[77] RENJITH S, ABRAHAM A, JYOTHI S B, et al. An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(10): 9564-9575.

[78] 谌志群, 鞠婷. 基于 BERT 和双向 LSTM 的微博评论倾向性分析研究[J]. 情报理论与实践, 2020, 43(8): 173-177.
CHEN Z, JU T. Research on Tendency Analysis of Microblog Comments Based on BERT and BLSTM[J]. Information studies: Theory & Application, 2020, 43(8):173-177.

[79] LIN E, SUN J, CHEN H, et al. Data Quality Matters: Suicide Intention Detection on Social Media Posts Using a RoBERTa-CNN Model[J]. arXiv:2402.02262, 2024.

[80] WANG Z, JIN M, LU Y. High-Precision Detection of Suicidal Ideation on Social Media Using Bi-LSTM and BERT Models[C]//International Conference on Cognitive Computing, Shenzhen, China, December 17-18, 2023. Cham: Springer, 2023: 3-18.

[81] ARAGÓN M, MONROY A P L, GONZALEZ L, et al. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, July 9-14, 2023. Stroudsburg: ACL, 2023: 15305- 15318.

[82] CHEN L S, LUO Z J, NALLURI V. Constructing Depression Prediction Model using ChatGPT and Machine Learning Algorithms[C]//2023 12th International Con- ference on Awareness Science and Technology (iCAST), Taichung, China, December 21, 2023. Piscataway: IEEE, 2023: 233-236.

[83] GHANADIAN H, NEJADGHOLI I, Al OSMAN H. Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models[J]. IEEE Access, 2024.

[84] GAO J, ChENG Q, YU P L H. Detecting comments showing risk for suicide in YouTube[C]//Proceedings of the Future Technologies Conference (FTC) 2018, Vancouver, Canada, November 15-16, 2018. Cham: Springer, 2019: 385-400.

[85] LI Z, ZHOU J, AN Z, et al. Deep hierarchical ensemble model for suicide detection on imbalanced social media data[J]. Entropy, 2022, 24(4): 442.

[86] BEN HASSINE M A, ABDELLATIF S, BEN YAHIA S. A novel imbalanced data classification approach for suicidal ideation detection on social media[J]. Computing, 2022, 104(4): 741-765.

[87] SAWHNEY R, JOSHI H, GANDHI S, et al. A time-aware transformer based model for suicide ideation detection on social media[C]//Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), November 16-20, 2020. Stroudsburg: ACL, 2020: 7685-7697.

[88] SAWHNEY R, JOSHI H, FLEK L, et al. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media[C]//Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics, April 19-23, 2021. Stroudsburg: ACL, 2021: 2415-2428.

[89] YOHAPRIYAA M, UMA M. Multi-variant classification of depression severity using social media networks based on time stamp[C]//Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021, Coimbatore, India, August 27-28, 2021. Singapore: Springer Nature Singapore, 2022: 553-564.

[90] KODATI D, TENE R. Identifying suicidal emotions on social media through transformer-based deep learning[J]. Applied Intelligence, 2023, 53(10): 11885-11917.

[91] YANG K, JI S, ZHANG T, et al. Towards interpretable mental health analysis with large language models[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, December 6-10, 2023. Stroudsburg: ACL, 2023: 6065-6077.

[92] LAN X, CHENG Y, SHENG L, et al. Depression Detection

蒙秀扬（1998—），男，海南定安人，硕士研究生，CCF 学生会员，主要研究方向为大数据技术与人工智能、机器学习、自杀意念检测。

MENG Xiuyang, born in 1998, M.S. candidate, CCF student membership. His research interests include big data technology and artificial intelligence, machine learning, suicide ideation detection.

王世屹（2004—），男，云南石屏人,本科生,主要研究方向为机器学习、图像识别。

WANG Shiyi, born in 2004, undergraduate. His research interests include machine learning, image recognition.

李渡渡（2000—），女，河南洛阳人，硕士研究生，主要研究方向为自然语言处理、机器学习。

LI Dudu, born in 2000, M.S. candidate. Her research interests include natural language processing, machine learning.

王春玲（1975—），女，山东莱州人,博士,副教授，主要研究方向为大数据技术与人工智能、智能信息处理。

WANG Chunling, born in 1975, Ph.D., associate professor. Her research interests include big data technology and artificial intelligence, intelligent information processing.

on Social Media with Large Language Models[J]. arXiv: 2403.10750, 2024.

[93] 苗红闪. 基于微博抑郁症识别方法研究[D]. 北京: 北京工业大学, 2020.
MIAO H, Research on depression recognition method based on Micro-blog[D]. Beijing: Beijing University of Technology, 2020.

[94] MENG X, WANG C, YANG J, et al. 2024. Predicting Users' Latent Suicidal Risk in Social Media: An Ensemble Model Based on Social Network Relationships. Com- puters, Materials & Continua [J], 79: 4259-4281.

[95] YAZDAVAR A H, MAHDAVINEJAD M S , BAJAJ G ,et al.Multimodal mental health analysis in social media[J]. PLoS ONE, 2020, 15(4):e0226248.

[96] GUI T, ZHU L, ZHANG Q, et al. Cooperative multimodal approach to depression detection in twitter[C]//Proceedings of the AAAI conference on artificial intelligence, Honolulu Hawaii, USA, January 27-February 1, 2019. Palo Alto: AAAI, 2019: 33(01): 110-117.

[97] RAMÍREZ-CIFUENTES D, FREIRE A, BAEZA-YATES R, et al. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis[J]. Journal of medical internet research, 2020, 22(7): e17758.

[98] SHEN T, JIA J, SHEN G, et al. Cross-domain depression detection via harvesting social media[C]//27th International Joint Conference on Artificial Intelligence, Stockholm, 13-19 July, 2018. San Francisco: Morgan Kaufmann, 2018: 1611-1617.

[99] MBAREK A, JAMOUSSI S, HAMADOU A B. An across online social networks profile building approach: Application to suicidal ideation detection[J]. Future Generation Computer Systems, 2022, 133: 171-183.