计算机科学
COMPUTER SCIENCE

文本人格检测研究综述

朱洋甫, 李美玲, 谭嘉辰, 吴斌

引用本文

朱洋甫, 李美玲, 谭嘉辰, 吴斌. 文本人格检测研究综述[J]. 计算机科学, 2024, 51(12): 209-222.

ZHU Yangfu, LI Meiling, TAN Jiachen, WU Bin. Study on Text-based Personality Detection—A Review [J]. Computer Science, 2024, 51(12): 209-222.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500071

文本人格检测研究综述

朱洋甫　李美玲　谭嘉辰　吴　斌
北京邮电大学计算机学院　北京 100876
(zhuyangfu@bupt.edu.cn)

摘　要　文本人格检测是人格计算领域一项重要的研究内容,旨在分析用户生成文本中隐含的人格特质。随着社交网络的发展,人们习惯于在线发布蕴含心理活动的内容,这为文本人格检测提供了新的机遇。准确地检测用户人格特质在心理健康诊断、舆情监控、人机交互系统设计以及大语言模型构建等方面具有重要意义。文中对文本人格检测的相关研究和新颖方法进行了深入调研和全面综述。首先介绍了人格检测相关背景知识、任务模式;其次从心理语言学统计方法、特征工程方法、深度学习方法、预训练语言模型4个方面梳理了现有方法;然后对当前广泛使用的评测数据集及模型效果进行了总结;最后从人格检测的可靠性、公平性、伦理与隐私、数据集和评价指标统一以及大语言模型与人格5个方面分析了本领域存在的问题和未来研究方向。

关键词:人格计算;社交网络;用户生成文档;大语言模型
中图分类号　TP391

Study on Text-based Personality Detection－A Review

ZHU Yangfu,LI Meiling,TAN Jiachen and WU Bin
School of Computer Science,Beijing University of Posts and Telecommunications,Beijing 100876,China

Abstract　Text-based personality detection is an important research content in the personality computing field,aiming to analyze the implicit personality traits in user-generated text. With the booming of social networks,people are accustomed to posting online content that implies their psychological activities,which provides new opportunities for text-based personality detection. Accurately detecting personality traits is important in psychological health diagnosis,public opinion monitoring,human-computer interac-tion system design,and even in the construction of large language models today. This paper provides a comprehensive review of text-based personality detection. Firstly,it introduces the background and task patterns of personality detection. Secondly,the existing detection methods are categorized into four aspects:psycholinguistic statistical methods,feature engineering methods,deep learning methods,and pre-trained language models. Then,the commonly used datasets and model performance are summarized. Finally,the issues and future research in this field are analyzed from five aspects:reliability,fairness,ethical and privacy,the unification of dataset and evaluation metrics,and the relationship between large language models and personality.

Keywords　Personality computing,Social networks,User-generated content,Large language model

1　引言

人格(Personality)指人类个体独有的思维、情感和行为方面的特征和模式。它的形成受到基因、环境等因素的影响,相对稳定并贯穿人的一生,影响着个体的认知方式、情绪反应、价值观和行为方式[1-2]。随着社交网络的普及,人们习惯于在Twitter、微博等平台发布内容,这些数字足迹为了解个体的心理活动提供了新的途径,也为人格计算(Personality Computing)的研究带了机遇[3]。人格计算是心理学、社会学和计算机科学的交叉研究领域,包括自动人格识别、感知和合成3个方面。人格识别的目标是正确预测人类个体内在的特质,人格感知则分析他人对个体的人格评价,而人格合成试图生成具有人工人格的智能体或机器人。人格计算具有广泛的应用价值,其在商业方面可以助力个性化商品推荐[4];在心理健康方面可以应用于抑郁症分析[5]、自杀早期监测等[6];在人机交互方面,拟人化机器可以提供更个性化、人性化的交互体验[7-9]。此外,在舆情管控方面,其可以赋能政治倾向分析[10]、信息传播控制[11-12]。

在心理学领域,传统的人格分析手段主要是基于心理学量表进行测试。被试需要填写量表,回答设定的问题,再由

到稿日期:2024-05-20　返修日期:2024-06-16
基金项目:国家自然科学基金(62372060);北京邮电大学优秀博士创新基金(CX2022219)
This work was supported by the National Natural Science Foundation of China(62372060) and BUPT Excellent Ph. D. Students Foundation (CX2022219).
通信作者:吴斌(wubin@bupt.edu.cn)

多位心理学家共同分析和解读。然而，这种方法存在一些限制。首先，它需要被试花费大量的时间和精力来填写大量的问题；其次，人类心理学家的解读可能存在主观性和偏见，不同的心理学家对同一份量表结果可能会有不同的解释。

随着互联网时代海量多样的语音、图像、视频、文本、社交媒体活动、人机交互等用户数据产生，利用用户数字足迹进行自动人格检测已成为心理学和计算机科学研究领域的热点之一。一些研究者对近年来关于人格计算的相关工作进行了综述，但这些综述大多从宏观角度介绍人格识别、感知、合成[13]，或者基于文本、音频、视频和用户行为对不同类型数据进行全面概括，欠缺对文本人格检测的深入调研[14-15]。此外，随着大语言模型等新技术的发展，文本人格检测的一些新的突破尚未进一步跟进[16]。

文本人格检测旨在对人类生成的文本信息中蕴含的人格特质进行识别和分析[17-18]。早期文本人格检测的方法主要利用心理学词典进行统计，因为心理学普遍的认识是，语言的使用反映了他们是谁[19]。随后特征工程结合机器学习的方法在本领域得到广泛应用。随着社交网络和深度学习技术的发展，一系列方法试图从用户生成文本数据中检测人格特质，本文将其总结梳理为深度神经网络方法、领域知识增强方法、用户群体方法、文档结构方法等。此外，随着在自然语言处理领域的预训练语言模型（Pre-trained Language Models，PLMs）的发展，从词嵌入方法到 Transformer 架构预训练语言模型，再到如今的大语言模型（Large Pre-trained Language Models，LLMs），都在文本人格分析中扮演重要角色。因此，本文较为全面地综述了人格计算中文本人格检测这一分支的发展脉络，期望能激发对人格计算领域新的思考。

本文第 2 章介绍人格检测心理学背景知识，包括人格特质理论以及心理学人格评估方法；第 3 章形式化文本人格检测任务，从心理语言学统计方法、特征工程方法、深度学习方法、预训练语言模型对文本人格检测发展脉络进行梳理并总结不同方法的优缺点；第 4 章对相关的数据资源进行整理，并对主流方法进行性能对比；第 5 章阐述了目前研究中面临的突出困难，并展望未来可能的发展方向。

2 背景知识

2.1 人格特质理论

人格特质理论是一种描述和解释个体人格特征的理论框架，它认为人格可以通过一系列相对稳定的特质来描述和区分。这些特质是个体在认知、情感和行为方面表现出的一种一致性模式。

大五人格（Big Five Personality Traits）是心理学研究中被最广泛接受的人格特质分类指标[20]。大五人格特质的理论基础是词汇假说，即人们在语言中使用的词汇涵盖了广泛的人格特质。需要注意的是，词汇假说并不意味着人格特质仅限于语言中的词汇，而是将语言作为揭示和描述人格的重要线索之一。如图 1 所示，这个理论认为人格特质包含 5 个维度，并且细分为 30 个方面，可以提供对个人行为、思想和偏好的深刻见解。这 5 个基本特质维度是：

1）外向性（Extraversion）：指个体的社交、活跃和冒险的倾向。外向性高的人通常富有社交能力，活跃好动，寻求刺激，并且喜欢与他人合作。

2）宜人性（Agreeableness）：指个体对他人的友好、合作和信任的倾向。宜人性高的人通常友善、慷慨、体贴他人，并且善于解决冲突。

3）尽责性（Conscientiousness）：指个体的目标设定、自律和组织的倾向。尽责性高的人通常有较强的自我控制能力，善于计划和追求目标，注重细节和责任。

4）神经质（Neuroticism）：指个体情绪稳定性和情绪反应的倾向。神经质高的人通常更容易体验到负面情绪，情绪波动较大，而神经质低的人更为稳定和冷静。

5）开放性（Openness）：指个体想象力、好奇心、独创性和开放思想的倾向。开放性高的人通常对新思想和体验持开放态度，喜欢探索、创造和接受挑战。

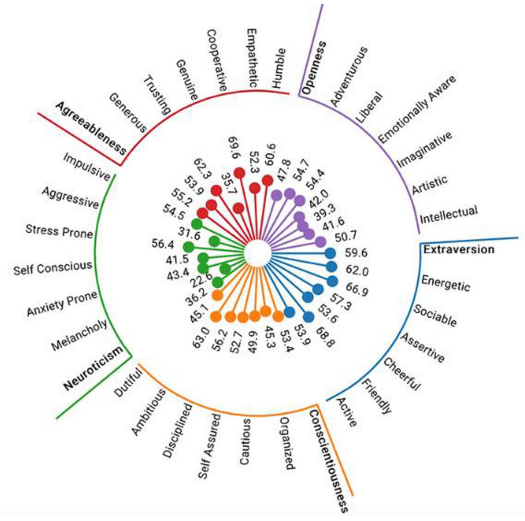

图 1 大五人格特质维度划分[1)]

Fig. 1 Division of the big five personality traits

迈尔斯布里格斯类型指标（Myers-Briggs Type Indicator，MBTI）[21]也是一种流行的人格测量指标，它通过自我报告问卷表明人们在感知世界和做出决定方面的不同心理偏好。MBTI 将人格类型划分为以下 4 对偏好：

1）在心理能量取向方面包括外向（Extraversion）和内向（Introversion）。外向倾向于关注外部世界，与他人互动，从外部环境中获取能量；内向倾向于关注内部世界，独立思考，通过内省来获取能量。

2）在信息收集方式方面包括感觉（Sensing）和直觉（Intuition）。感觉倾向于关注具体的细节和现实的信息，注重实际和具体的经验；直觉倾向于关注抽象的概念和未来的可能性，注重想象和推理。

3）在决策方式方面包括思维（Thinking）和情感（Feeling）。思维倾向于通过逻辑和客观因素做决策，注重分析、原则和一致性，情感倾向于通过个人价值观和他人需求作决策，注重情感和人际关系。

4）在生活方式方面包括判断（Judging）和知觉（Percei-

1) https://www.receptiviti.com/big-five

ving)。判断倾向于组织和计划，喜欢有结构和确定性的生活方式；知觉倾向于灵活和开放，喜欢保持灵活性和适应性的生活方式。

通过将这 4 个偏好组合起来，可以得到 16 种不同的人格类型，例如 ISTJ，ENFP，INTP 等。每种人格类型都有其独特的特征和行为倾向。MBTI 被广泛应用于个人发展、人际关系、职业规划和团队建设等领域。它提供了一种用于理解和沟通不同人格类型之间差异的框架。需要注意的是，MBTI 在学术界并非是被普遍接受的科学测量指标，一些研究对其可靠性和效度提出了质疑。因此，在使用 MBTI 时需要持谨慎的态度，并结合其他心理学理论和工具进行综合分析。

2.2 心理学人格评估

在心理学研究中，自我报告问卷是使用最广泛的人格评估方法。例如，经典的大五人格心理量表包含 240 个问卷项目[22]。类似的问卷还包括明尼苏达多项人格问卷[23]、艾森克人格问卷[24]、国际人格问卷库[1)]等。被试者需要回答一系列问题，这些问题涉及不同情况下自身的行为、感受、兴趣。问卷工具的优势在于经过多次验证，具备坚实的实证基础。然而，自我报告也存在一些缺点。首先，它们需要评估者是心理学专家或者进行过相关培训，并且对大量用户进行问卷调查是一项劳动密集型工作，需要大量人力物力支持。其次，自我报告方式容易受到用户自身环境影响。比如，面对高校中心理中心组织的心理问卷，大学生们通常会避免暴露自己的内心，这是一种社会需求偏差，即以一种使自己在他人眼中更有利的方式回答问题[25]。最后，由于涉及个人隐私问题，人们是否愿意填写问卷和参加实验室研究也是一个问题。为了克服这些缺点并寻找自动化测量方法来代替显式的自我报告，研究者们开始探索数字化的人格计算方法[26]。

3 文本人格检测方法

文本人格检测可形式化为一个多文档、多标签的分类或回归任务。给定一个用户生成的 n 篇帖子集合 $X=\{x_1, x_2, \cdots, x_n\}$，其中 $x_i=[w_1^i, w_1^i, \cdots, w_k^i]$ 表示第 i 个帖子有 k 个词，人格检测模型的目标是从特质特定标签空间 $Y=\{y_1, y_2, \cdots, y_t\}$ 中预测 t 维人格特质。例如，在 MBTI 分类中 $t=4$，在大五人格分类中 $t=5$。值得注意的是，在大五人格回归中是预测每个维度的得分。$x_i=[w_1^i, w_1^i, \cdots, w_k^i]$。图 2 给出了文本人格检测的一般流程：从社交网平台中采集若干用户的帖子构建数据集，提取推文特征，利用统计方法、机器学习或深度学习等方法训练从文本到人格特质的映射，并将训练好的模型应用到待检测的用户上，判断该用户的 MBTI 或者大五人格特质。

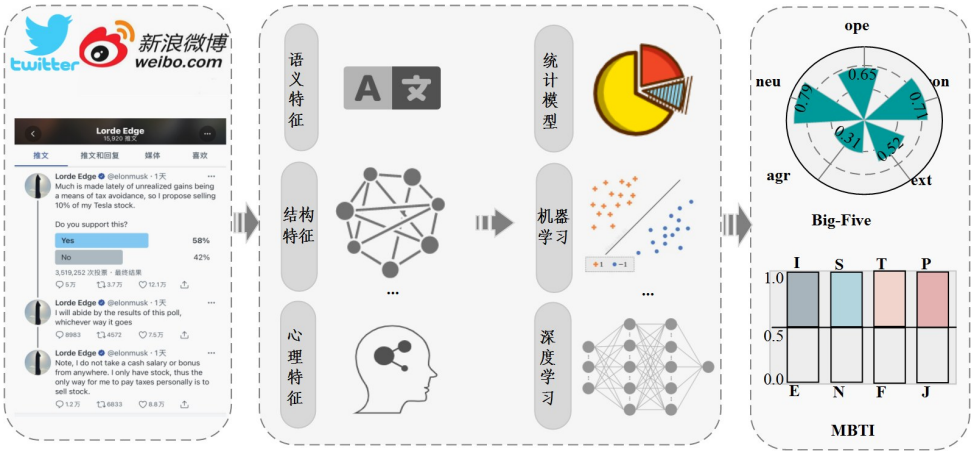

图 2 文本人格检测一般流程

Fig. 2 General process of text-based personality detection

如图 3 所示，本章将从心理语言学统计方法、特征工程方法、深度学习方法和预训练语言模型 4 个方面对文本人格检测研究进行全面梳理。在心理语言学统计方法中，主流的统计工具包括语言调查和单词计数[27]（Linguistic Inquiry and Word Count，LIWC）和心理语言学数据库[28]（Medical Research Council，MRC），这些方法被用于构建封闭词汇，以发现一些统计规律。对于特征工程方法，本章介绍早期的词嵌入和用户属性特征处理手段如何结合机器学习算法进行检测。对于深度学习方法，本章重点从经典深度神经网络模型、结合不同领域知识的方法、用户群体层面建模方法和篇章级文档结构学习方法 4 个方面进行介绍。最后，沿着自然语言处理领域中预训练语言模型的发展过程，介绍文本人格检测的进展，特别是大模型时代的一些探索。

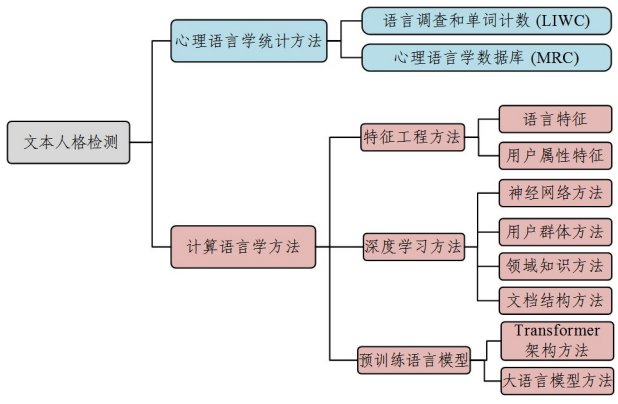

图 3 文本人格检测方法

Fig. 3 Text-based personality detection methods

1) https://ipip.ori.org/

3.1 心理语言学统计方法

广泛的研究已经证明，语言是衡量人格的客观行为标准。无意识地使用的词汇是一种理想的隐性行为衡量标准，并有助于减轻自我报告中固有的偏见。这也引发了传统问卷建模到利用心理语言学进行数字化分析的转变和发展。通过分析个体在语言中使用这些词汇的频率或倾向性来评估其人格特质称为封闭词汇法。早期的研究者 Pennebaker 和 King 发现写作风格(如词频和词性)和人格之间存在可靠的联系[29]，并提出了语言调查和单词计数(Linguistic Inquiry and Word Count，LIWC)。LIWC 将单词分为各种与心理相关的类别，如功能词类别(如冠词、连词、代词)、情感过程类别(如高兴、紧张、哭泣)和社会过程类别(如伴侣、谈话、朋友)。通过计算每个类别中单词的频率可以分析文本的人格特质。MRC 是医学研究委员会的心理语言数据库，包含 150837 个英语单词的不同心理语言属性。这些单词的不同语义、句法和拼写细节适用于心理学、语言学和人工智能领域的各种研究。

Receitiviti API[1)] 是具有代表性且得到落地实用的封闭词汇法，IBM 公司利用 LIWC 分析人们对语言的使用，几乎能从任何语言来源洞察人们的心理和认知状态。从这些封闭词汇中进行的各种实证研究可以发现，神经质患者更多地使用第一人称单数代词；随和的人表达的积极情绪更多且较少使用冠词；有责任心的人会避免使用否定消极情绪词汇和反映差异的词汇(如 should 和 would)。对于开放型的用户，他们偏爱较长的单词和表达试探性的单词(如 perhaps 和 maybe)，以及避免使用第一人称单数代词和现在时形式。外向者也比内向者更多地使用积极的情感词汇，并表现出更多的赞同和赞美。此外，在区分个体差异方面，功能词比内容词(如名词、动词和形容词)更有效。这些发现坚定了一个观点，即人们如何沟通比他们所沟通的内容提供了更多的心理线索[30-31]。

然而，这些心理语言学统计方法存在一些局限性。首先，语言的含义和解读往往与特定的上下文相关，而静态统计特征无法完全捕捉到这种复杂性，例如讽刺、隐喻等语言表达；其次，统计方法主要考察文本与人格特质之间的相关性，而不是直接从文本中自动检测人格。

3.2 基于特征工程的人格检测

随着机器学习的发展，一些研究者开始思考是否可以通过训练文本分类器来预测作者的人格。表 1 总结了机器学习时代，研究者们结合不同特征工程手段与经典机器学习算法进行文本人格分析的工作。2005 年，Argamon 等[32] 率先采用词类和虚词的相对频率作为支持向量机(Support Vector Machines，SVM)的输入，对外向和神经质两个极端的学生进行区分。Oberlander 等尝试不同的 N 元语法(N-gram)特征并结合朴素贝叶斯(Naïve Bayes，NB)对博客作者的人格特质进行二元分类和多重分类[33]。2007 年，Mairesse 等[34] 进行了更全面的探索，系统地整合和对比了心理语言特征集以及分类、回归、排序 3 类模型的效果。他们发现与基于自我报告的分数相比，基于观察者分数的预测准确度更高。此外，还有一些研究者使用机器学习技术根据用户社交网络属性[35]、时间戳[36]、俚语使用[37]、个人资料[38]来自动推断人格特质。

表 1 基于手工特征的人格检测

Table 1 Handcrafted features based personality detection

方法	发表来源	特征工程	网络架构	算法描述	使用数据
Argamon 等[32]	CSNA 2005	功能词表、连词短语、情态指示词、评价形容词	SMO	通过词汇特征进行文本人格分析	Essays
Oberlander 等[33]	COLING 2006	N-gram 特征	NB，SVM	使用不同的 N-gram 特征集来探索人格	Blog
Mairesse 等[34]	JAIR 2007	LIWC，MRC，表达方式特征	Adaboost，M5，SMO	结合不同心理特征进行人格分析	Essays，EAR
Bachrach 等[35]	WebSci 2012	个人资料特征	MLR	研究人格与 Facebook 个人属性之间的相关性	myPersonality
Farnadi 等[36]	AAAI 2013	LIWC、时间戳、社交网络属性等	SVM，KNN，NB	根据用户社交状态推断人格特质	Essay，myPersonality
Alam 等[37]	AAAI 2013	N-gram 特征、网络俚语、表情符号编码	SMO，BR，NB	结合网络俚语、表情符号编码进行人格分析	Essays，myPersonality
Volkova1 等[38]	AAAI 2015	年龄、性别、收入、教育程度、关系状态、乐观程度和生活满意度	对数线性模型	从社交媒体文本推断潜在的用户情绪和人格属性	Twitter
Zhu 等[39]	SIGCAS 2012	个人资料特征、情感属特征、时间属性特征	NB，SVM，C4.5	人人网用户的人格特质分析	人人网
Bai 等[40]	IEEE WIC 2013	个人资料特征、安全设置特征、社交网络特征、	IR	Sina 微博中文数据集上人格特质多任务回归	Sina 微博
Zhong 等[42]	中文信息学报 2017	中文版 LIWC	SPCA	分析中文聊天文本中语词与人格的关系	中文微信对话

除了英语语料库外，汉语环境下的人格检测研究也得到了发展。中科院心理研究所的 Zhu 等开展了一系列研究，分析了 209 名人人网用户[39] 和 547 名新浪微博用户[40] 的个人资料、自我表达等静态特征和微博更新、@提及、浏览记录等动态特征。研究发现外向性与社交状态再发布的比例呈正相关，神经质与用户发布的愤怒情绪博客数量成正相关。Peng

1) https://www.receptiviti.com/liwc

等[41]对 222 名 Facebook 上的中文用户的帖子进行了处理，发现利用好分词技术和特征选择技术有助于提高性能。Zhong 等[42]首次借助中文版语言查询与字词计数(SC-LIWC)[43]分析中文环境下聊天文本所表现的认知用语特征和人格的关系。

上述工作都是机器学习时代从特征工程角度出发对人格检测任务的一些早期探索。尽管手工特征能提供一定的可解释性，但特征工程耗时耗力并且在隐含语义提取方面还存在不足。

3.3 基于神经网络的人格检测

深度神经网络在句子和文档建模方面取得了显著的成绩，其中卷积神经网络(CNN)通过卷积滤波器在提取序列不同位置的特征方面表现良好。递归神经网络(RNN)可以处理任意长度的序列并捕获长期依赖关系。在文本人格检测任务中，主流的深度神经网络模型也有着广泛的应用。

在 CNN 架构上，Majumder 等[44]率先尝试利用 CNN 和 Word2vec 词嵌入技术在人格检测领域进行探索。AttRCNN[45]是一个双层卷积网络，用来学习每个用户文本帖子的深层语义特征。在 RNN 架构上，Hernandez 等[46]结合 GloVe 词嵌入方法，尝试 RNN 及一系列变种模型来建模句子之间的上下文依赖。同年，Tandera[47]和 Xue 等[48]也在大五人格数据集上验证了 RNN 及其变种的效果；而后 2CLSTM 模型[49]利用双向 LSTM 编码句子嵌入，利用 CNN 学习句子组并生成代表作者的最终特征向量。此外，C2W2S4PT[50]通过 RNN 从字符到单词再到句子进行编码。这种字符级单词编码适用于社交网络语料，因为这些数据包含大量的非官方文本。Zhou 等[51]在连接单词和表情符号嵌入后使用基于注意的双向 LSTM 提取人格特质。然而上述方法仅从文本语义角度出发，利用或者改进神经网络模型进行检测，未能有效利用人格检测这个任务相关的心理学领域知识。

3.4 基于领域知识的人格检测

一些研究发现，将心理学领域知识与文本语义相结合可以提高人格检测效果，并提供可解释性。表 2 总结了文本人格检测任务中如何结合领域知识进行研究的具体细节。Poria 等[52]利用 SenticNet[53]情感知识库和 ConceptNet[54]常识知识库从文本中提取常识性知识以及相关情感极性来进行人格分析。Han 等[55]认为现有的基于神经网络模型的人格检测模型缺乏对人格的解读能力，因此提出了一种基于人格词典的可解释人格检测模型。他们设计了一种自动构建社交媒体环境下的中文人格词典的方法，并分析了人格特质与词语类别之间的相关性。Zhu 等[56]认为依赖语义特征或基于现有工具来计算浅层的心理统计特征，无法有效利用有助于确定和解释人格特征的心理语言学知识，于是提出了一种词汇心理语言学知识引导的图神经网络 HPMN。不同于 Han 等自动构建词典的方法，该模型将心理学专家总结的人格词汇作为桥梁[57]，通过注入相关的外部知识来丰富文档的语义。他们从计算的角度为心理语言学研究中的词汇假设提供了一定程度的支持。

表 2 基于领域知识的人格检测方法

Table 2 Domain knowledge based personality detection

方法	发表来源	领域知识	算法描述	使用数据
Poria 等[52]	MICAI 2013	情感知识库 常识知识库	利用情感知识库(SenticNet)和常识知识库(ConceptNet)，将领域知识与 LIWC、MRC 心理语言特征结合进行人格检测	Essays
Han 等[55]	KBS 2020	人格词典	采用词频-逆文档频率(TF-IDF)算法从 Weibo 中提取关键词，将这些关键字分成代表不同语义类别的不同集群	Weibo
HPMN[56]	KBS 2022	心理学知识	利用心理学家总结的人格词汇作为桥梁，注入词汇相关的外部知识来丰富文档的语义	Essays，Youtube，PAN2015，MyPersonality
Ramezani 等[58]	CIN 2022	DBpedia 知识库	将输入文本的概念与 DBpedia 知识库条目进行匹配，构建互关联的概念描述知识图谱	Essays
PerKG[59]	IEEE SMC 2022	语言风格知识	人格知识图谱包含单词、人格、语言风格实体以及这些实体之间的不同关系	Youtube，PAN2015，MyPersonality
PerHGAT[60]	ICKG 2022	语言风格知识	将用户帖子中的语言风格信息聚合到词的语义学习中	Youtube，PAN2015，MyPersonality
HG-PerCon[61]	Neural Networks 2024	语言风格知识	利用基于历史语义信息和语言风格知识的用户表征进行跨视图对比学习	Youtube，PAN2015，MyPersonality

得益于知识图谱表达领域知识的能力，一些学者尝试构建人格分析领域的专业知识图谱。Ramezani 等[58]通过将文本概念与 DBpedia 知识库条目进行匹配来构建人格知识图谱，结合丰富概念的知识片段来进行人格检测。人格知识图谱 PerKG[59]利用心理语言学家发现的语言风格与人格特质之间的联系，通过思考方式、写作风格、情绪、精神状态、需求感、团队认同感等 10 余种实体，及隐含、包含、存在、需要、对立、使用和缺乏 7 种关系丰富领域知识。而后，在 PerKG 基础上，Li 等结合文本语义和语言风格来预测社交媒体用户的人格。为了对语言风格进行表征，他们提出了一种人格检测层次图注意网络 PerHGAT[60]，将用户帖子中的语言风格信息聚合到词的语义学习中，然后将词的语义信息聚合到用户表示中。最近，Li 等[61]继续提出了 HG-PerCon 模型，利用基于历史语义信息和语言风格知识的用户表征进行跨视图对比学习。

综上，利用心理学领域知识构建专业知识图谱或者将

知识作为文本语义的增强是一种有效的方式，但是在利用心理学领域知识进行文本人格检测时，仍然面临一些挑战和限制。首先，获取和构建专业知识图谱需要大量的心理学领域专家知识；其次，心理学领域的知识是不断演变和更新的，知识图谱需要进行定期更新和维护，才能确保其中的信息和关系的准确性和时效性。

3.5 基于用户群体的人格检测

现有的人格研究大多以独立同分布(Independent Identically Distribution，IID)为假设，忽略了个体之间的相互影响。一部分研究者从用户群体角度对人格检测任务进行探索，认为用户网络结构也值得被关注，因为其遵循一个通识：物以类聚，人以群分。此外，由于具有黄金标准的人格特质数据收集困难，现有的人格数据集大部分规模较小，因此研究者们也在思考是否可以摒弃之前监督范式下的检测模型，以半监督或者自监督的方式从用户群体角度来检测人格。在方法上，得益于结构特征学习能力，网络表示学习(Network Representation Learning，NRL)和图神经网络(Graph Neural Networks，GNNs)等方法被一些研究者应用到此领域。

AdaWalk[62]是第一个基于网络表示学习来检测人格的方法，对于每个人格数据集，构造一个完整的图，图中的节点代表用户，而节点之间的边通过用户生成文档相似性连接。这样就将用户人格检测问题转换成了一个图节点分类问题。而后借鉴 Node2vec[63]算法在图上随机游走的思想从群体层面学习用户节点表示。受 AdaWalk 的启发，Guan 等[64]提出了一个 Personality2vec 网络表示模型，它充分利用从文本中提取的语义特征和结构信息为每个用户生成人格向量。该模型通过学习同一组用户之间的相互影响，更全面地利用数据集。Wang 等[65]构建了用户-文档关系、文档-词关系和词共现的异质图，并且借鉴 TextGCN[66]模型思想设计了 PersonalityGCN 模型，以半监督的方式在异质图上学习用户节点的表示。

上述方法都是数据集中全部用户构建的静态图，而未见的测试用户是随时更新的。因此，这些直推式的图结构学习方法并不适用于在线社交网络人格检测。

3.6 基于文档结构的人格检测

文本人格检测旨在识别社交媒体帖子中隐含的人格特质，但用户生成文档并不是一个简单句子或者段落，而是大量帖子的集合。所以这项任务的核心是将多个分散的帖子中的信息整合在一起，以描绘每个用户的整体人格概况。近年来，一些学者开始对用户生成帖子的结构进行探索。表 3 列出了这些方法的细节。2020 年，Lynn 等[67]认为并非每个帖子的贡献都相同，因而设计了一个层次注意力网络 SN+Attn 以自下而上的方式从词粒度到帖子粒度获取用户文档表示。2021 年，Yang 等[68]认为将帖子建模为一个序列可能会引入序列偏差，因为文本人格检测任务处理的数据不是标准的篇章级上下文长文档，而是用户不同时刻生成帖子的集合，于是提出了一个 Transformer-MD 网络，借鉴长文档建模的经典模型 Transformer-XL 内存库(Menmory Bank)思想以帖子序列不可知的方式表示用户。Yang 等[69]却持不同观点，认为帖子之间存在心理语言结构，因此他们从心理学角度来聚合一个用户的所有帖子。他们将每个用户的所有帖子建模成异质三部图，3 种异构类型节点包括帖子、单词和 LIWC 类别，利用图神经网络在心理相关的单词和类别固定的结构上消息传递。2022 年，Zhu 等[70]同样从用户生成帖子的结构出发，设计了一个图对比转换网络 CGTN，通过图自监督学习提取人格检测中的辅助信号，为解决人格标签稀缺的问题提供了一个新的视角。到 2023 年，Yang 等[71]认为人为定义的帖子结构是不准确的，与之前需要预先定义确定的文档结构图的工作不同，他们设计了一个动态深度图卷积网络 D-DGCN 来自动学习帖子之间的关系。2024 年，Zhu 等[72]同样针对人格检测任务标注数据稀缺的问题，从文档结构角度构建了一种图数据增强模型 Semi-PerGCN，在有限标记用户的学习过程中，结合大规模未标记用户的一致性约束，增强人格检测模型的泛化能力。

表 3 基于文档结构的人格检测方法

Table 3 Document structure based personality detection

方法	发表来源	算法描述	使用数据
SN+Attn[67]	ACL 2020	认为并非每个帖子的贡献都相同，通过词到单个帖子、单个帖子到用户生成文档的层次注意力学习用户表示	Facebook
Transformer-MD[68]	AAAI 2021	认为将帖子建模为一个序列可能会引入序列偏置问题，Transformer-MD 在对每个帖子进行编码时，允许通过共享相同位置嵌入的记忆 tokens 访问同一用户的其他帖子	Kaggle，Pandora
TrigNet[69]	ACL 2021	认为帖子之间存在心理语言结构，TrigNet 将每个用户建模为帖子、单词和 LIWC 类别节点组成的异质三部图，通过图神经网络捕获用户文档结构信息	Kaggle，Pandora
CGTN[70]	IJCAI 2022	CGTN 从语义和心理学结构角度构建用户生成帖子的图结构，并通过图对比自监督学习捕获文档中隐含的监督信号，解决人格标签稀缺问题	Kaggle，Essays
D-DGCN[71]	AAAI 2023	认为人为定义的帖子结构具有局限性，D-DGCN 通过动态图神经网络学习每个用户固有的帖子与帖子的联系，自动化聚合用户生成帖子来学习用户表示	Kaggle，Pandora
Semi-PerGCN[72]	AAAI 2024	为每个标记和未标记的用户构建了一个异构人格图，心理语言学增强图神经网络，以半监督的方式检测人格	Youtube，PAN2015，MyPersonality

下面以层次注意力网络 SN+Attn[67]为代表，详细介绍基于文档结构的人格检测。如图 4 所示，给定一组来自用户 u 的 n 条消息，模型的第一步是为每个消息 m_i 生成编码。消息 m_i 中的每个单词都通过门控循环单元(Gated Recurrent Unit，GRU)产生隐藏状态。

$$h_j^i = GRU(w_j^i) \tag{1}$$

对隐藏状态序列$[h_1^i, h_2^i, \cdots, h_i^i]$应用注意机制。

$$d_j^i = \tanh(W_{\text{word}} h_j^i + b_{\text{word}}) \tag{2}$$

$$\alpha_j^i=\frac{\exp(d_j^{i^{\mathrm{T}}}\boldsymbol{d}_{\mathrm{word}})}{\sum_{k=0}^{l}\exp(d_k^{i^{\mathrm{T}}}\boldsymbol{d}_{\mathrm{word}})} \tag{3}$$

$$s_i=\sum_{k=0}^{l}\alpha_k^i h_k^i \tag{4}$$

其中，$\boldsymbol{d}_{\mathrm{word}}$是单词级注意的学习上下文向量，$b_{\mathrm{word}}$是偏置项，$\alpha_j^i$是$h_j^i$的标准化注意权重。得到这些消息表示后，再次层次化地堆叠GRU和消息级别的注意力机制，学习用户表示。

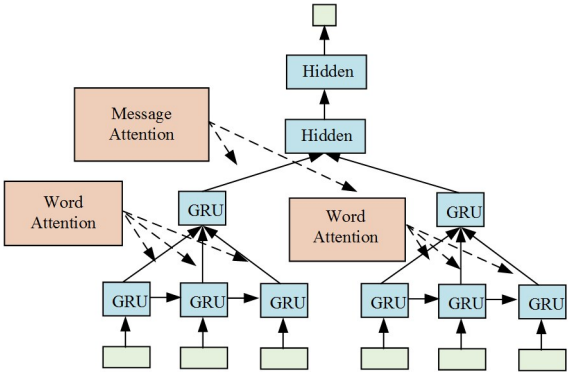

图4 层次注意力网络整体架构(图片引自文献[67])

Fig. 4 Overall architecture of hierarchical attention network[71]

3.7 基于预训练语言模型的人格检测

本节沿着预训练语言模型的发展过程，从早期Transformer架构预训练语言模型的各种微调技术，到如今大语言模型的人格特质评估与控制，梳理了文本人格检测任务的发展历程。具体如表4所列。

2018年，谷歌提出了一种Transformer架构的BERT模型，将一些掩码预测、上下文句子预测等预训练任务应用于Transformer的双向编码器来学习更复杂的单词语义。2019年，Keh等[73]第一次尝试微调BERT模型来建立了一个人格检测模型，发现借助预训练好的BERT可以有效提高文本人格检测准确率。Jiang等[74]进一步微调了变种模型RoBERTa，在Essays独白数据集上效果提升了2.49%。2020年，Mehta等[75]将心理语言学特征与BERT语言模型嵌入相结合进行人格检测，还从可解释性角度分析了个体心理语言学特征对人格特质最终预测的贡献，发现文本语义特征始终优于传统的心理语言学特征。2023年，Wen等[76]设计了一种DesPrompt微调方法，通过人格描述性提示(Prompt)来微调预训练语言模型(Pre-trained Language Models，PLMs)，以进行少样本甚至零样本人格检测而不引入额外的参数。因为在标注数据稀缺的情况下，分类头往往训练不足，导致检测性能不佳。具体来说，DesPrompt将人格检测建模为一个填字任务，输入内容首先用人格描述性提示进行封装，然后监督PLMs用描述人格特质的标签词填写提示。ABLPM[77]是一个基于标签提示的方法，该方法有效地模拟了人格标签、文本语义和写作风格之间的深度动态交互，在BERT上生成人格表示。Zhang等[78]证明了BERT只包含少量的心理信息，并提出了一种心理特征指导的预训练语言模型微调方法PsyAttention。通过微调BERT，同时设计新的目标函数，使CLS向量更接近于心理特征向量。

表4 基于预训练语言模型的人格检测

Table 4 Pre-trained language model based personality detection

方法	发表来源	语言模型	算法描述	使用数据
Jiang等[74]	AAAI 2020	BERT RoBERTa	微调预训练语言模型BERT，RoBERTa	Essays，FriendsPersona
Mehta等[75]	ICDM 2020	BERT RoBERTa AlBERT	微调预训练语言模型BERT，RoBERTa，AlBERT	Essays，Kaggle
DesPrompt[76]	IPM 2023	T5-large RoBERTa	一种生成人格描述性提示的方法，可以用有限的注释数据微调语言模型	FriendsPersona，Essays，myPersonality，PAN 2015
ABLPM[77]	IS 2023	BERT	人格特质标签提示方法，生成每个人格特质和文本语义向量，同时将写作风格集成到文本语义中	MyPersonality，Iphone，Samsung
PsyAttention[78]	EMNLP 2023	BERT	微调BERT，使CLS向量更接近于心理特征向量	Essays，Kaggle
Ganesan等[79]	ACL 2023	GPT 3	以GPT-3为例，研究了大模型从用户社交媒体帖子中评估大五人格特质的zero-shot能力	Facebook自建数据集
PsyCoT[81]	EMNLP 2023	GPT 3.5	利用心理问卷作为思维链，模仿人类通过与大模型多回合对话来完成人格测试	Essays，Kaggle
Caron等[82]	EMNLP 2023	BERT GPT2	使用心理测量问卷评估和操纵大语言模型的人格	Reddit自建数据集
Jiang等[83]	NIPS 2024	BART，GPT-Neo 2.7B，GPT-NeoX 20B，T0++ 11B，Alpaca，GPT 3.677B	引入了机器人格量表(MPI)对LLMs进行标准化和量化的人格评估	MPI自建数据集

3.8 基于大语言模型的人格检测

ChatGPT等大语言模型(Large Pre-trained Language Models，LLMs)借助指令微调(Instruction Tuning)和人类反馈强化学习(Reinforcement Learning from Human Feedback，RLHF)等技术，在遵循人类指令时表现出了卓越的对齐能力，并且在各种自然语言处理任务中展示了非凡的零样本(Zero-shot)学习能力。为了在下游任务中充分利用LLMs的潜力，一些研究工作手动或自动精心设计提示。人格检测领域也不例外，2023年，纽约州立大学的Ganesan等[79]系统地研究了GPT-3在零样本设定下进行人格评估的能力。他们使用了3类知识设计提示模板：1)人格特质的定义；2)不同特质的人经常使用和不经常使用的词语；3)借助心理问卷作为明确的辅助推理工具。实验发现，使用简单的提示并不能获得强大的性能，但注入特质相关知识会显著提高性能。长期

以来，问卷调查量表一直是用于测量人类人格特质的工具。Yang 等[80]将心理问卷作为额外的指导，利用语言序列生成模型从上下文表征中获取问卷项目相关信息。但它们依赖于利用数据驱动的方法来训练模型以捕获内隐人格线索。在此基础上，他们继续提出了 PsyCoT 模型[81]，从心理学家精心设计的心理问卷中汲取灵感，认为这些问卷项目可以被视为结构良好的思维链(Chain-of-Thought，CoT)过程的集合。他们将大语言模型作为文本分析的人工智能助理，提示助手在每个回合对单个问卷项目进行评级，并利用历史评级结果分析人格偏好。

另一方面，随着 LLMs 在各种自然语言处理任务中的颠覆式发展，研究者们开始探索 LLMs 是否具有人格特质，是否可以人为操纵 LLMs 的人格特质？北卡罗莱纳大学的 Caron 等[82]首次结合心理测量问卷来探索大语言模型的人格，并且通过实验证明了公共语言模型的人格特质可以通过文本上下文进行控制。随后，北京大学的 Jiang 等[83]扩展大语言模型验证范围，包括 GPT-NeoX，T0＋＋，Alpaca，GPT 3.5 等，并且引入机器人格量表(MPI)工具来研究机器人格，设计了一种人格提示方法，以可控的方式诱导具有特定人格的 LLMs，使其能够产生多样化和可验证的行为。

下面以心理问卷思维链大模型微调模型 PsyCoT[81]为代表，详细介绍基于大语言模型的人格检测。如图 5 所示，人格问卷中的项目被用作回答最终人格调查的思维链。根据文本提示 LLM 评分项目，通过多回合对话模拟人类完成人格测试的过程。具体地，大模型标准提示过程是设计合适的提示符 $P=\{D,I\}$来实现这一目标。其中 D 代表任务描述提示符，它告知 LLM 任务的定义；I 代表推理提示符，它推动 LLM 直接从给定的输入文本 X 推断出人格特质。

$$\hat{y}^i = LLM(D,X,I) \tag{5}$$

与标准提示相比，PsyCoT 在以下几个关键方面加强了标准提示：1)修改任务描述 D，指导 LLM 对问卷中的每个陈述(项目)进行评分；2)在 D 中包含了对问卷评分规则的描述，包括评分系统(如，“1＝非常不同意，2＝稍微不同意，3＝中立，4＝稍微同意，5＝非常同意”)，并强调了颠倒陈述的重要性；3)引入了 k 个推理步骤 $R=\{r_k\}_{k=1}^{K}$，然后通过多对话访问人格特质 $\hat{y}^i$，引导 LLM 以更合理的方式推断人格。

综上，大语言模型与人格研究工作还处于初步阶段，仍然存在许多挑战。例如，LLMs 人格的形成与哪些因素有关？LLMs 的人格会像人类一样影响下游任务吗？是否可以用不同人格诱导的 LLMs 作为代理来研究人类的社会行为？

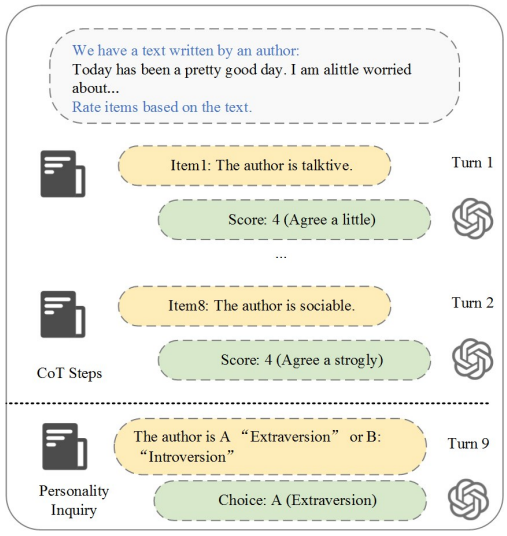

图 5 心理问卷思维链微调的大语言模型整体架构[81]

Fig. 5 Overall architecture of psychological questionnaire based chain-of-thought fine-tuned large language model[81]

4 数据集和性能评估

4.1 数据集

本研究以私有数据集为主，也有一部分开源的人格文本数据集，如表 1 所列。这些具有代表性的数据集描述如下：

myPersonality 是一款 Facebook 应用程序[1)]，它允许用户通过填写人格问卷来参与心理学研究。该数据集是现有最大人格文本数据集，大约有 600 万人参与了问卷调查。但遗憾的是，该数据集由于数据隐私问题在 2018 年被关闭。

Essays[29]是一个著名的意识流文本数据集，包含 2 468 个匿名学生，每个学生记录了他们的日常写作。学生的人格是通过回答大五人格量表来评估的。

PAN 2015[2)]数据集来自数据科学竞赛，包含英语、西班牙语、意大利语和荷兰语的 Twitter 文本和作者的大五人格评分，其得分分布介于－0.5～0.5 之间。

YouTube[84]是一个视频转录文本数据集，该数据集的标签是从众包标注任务中收集的。注释者观看每个视频博客，然后用问卷对五大人格评分，得分范围为 1～7。

FriendsPersona[3)]是在老友记电视节目上构建的人格数据集[85]，包含 711 个提取的对话。每段对话都由 3 名注释者评估主讲人的大五人格特质二元类别。

WASSA[4)]是 ACL2022 年举办的预测对新闻故事反应的同理心、情绪和人格共享任务挑战赛的相关数据集，数据集包括新闻文章、个人层面的人口统计信息(如年龄、性别)、人格信息、共情反应和 6 种基本情绪。

Kaggle[1)]数据集来源于 PersonalityCafe 论坛[2)]，人们在这里分享自己的人格类型，讨论健康、行为等问题，数据集中共

1) myPersonality.org

2) https://pan.webis.de/clef15/pan15-web/author-profiling.html

3) https://github.com/emorynlp/personality-detection

4) https://openreview.net/group?id=aclweb.org/ACL/2022/Workshop/WASSA

有 8675 名用户并且他们都通过问卷自测了 MBTI 人格倾向，每个用户发表 45～50 篇帖子。

Pandora[3] 数据集是由 Reddit 平台上 10 000 名用户评论组成的数据集。该数据集部分标记为大五人格和 MBTI 两种人格指标。值得注意的是，其中提供了一部分人口统计学信息，包括用户的年龄、性别、收入和地理位置[86]。

表 5 代表性文本人格数据集

Table 5 Representative text personality datasets

数据集	年份	大小	人格特质指标	文本类型	语言
Essays	1999	2 479	Big-5	随笔	英语
YouTube	2013	442	Big-5	YouTube 转录	英语
PAN2015	2015	300	Big-5	Twitter 帖子	英语、西班牙语、意大利语、荷兰语
myPersonality	2015	115 864	Big-5	Facebook 帖子	英语
Kaggle	2017	8 600	MBTI	PersonalityCafe	英语
FriendsPersona	2020	711	Big-5	电视节目转录	英语
Pandora	2021	10 288	Big-5，MBTI	Reddit 帖子	英语、西班牙语、意大利语、法语、德语、葡萄牙语
WASSA 2022	2022	2 655	Big-5	新闻和随笔	英语

4.2 性能评估

文本人格检测面向不同的检测场景，使用的评价数据集不尽相同。研究者通常将人格检测任务建模为分类和回归两类任务，评价指标也有所差异。在分类任务中，学者们希望预测 MBTI 类别或者是将大五人格特质分为高低维度。精准率（Precision）、召回率（Recall）、准确率（Accuracy）和 F1 值是常用的人格特质分类评价指标。

$$Precision=\frac{TP}{TP+FP} \tag{6}$$

$$Recall=\frac{TP}{TP+FN} \tag{7}$$

$$F1=2\frac{Precision\times Recall}{Precision+Recall} \tag{8}$$

其中 TP，TN，FP，FN 分别表示真阳性、真阴性、假阳性、假阴性。F1 值转化为精准率和召回率（召回率的调和平均值，大多数研究报告 F1 分数而不是报告精度和召回率）。在回归任务中，学者们希望直接预测大五人格特质得分。评估指标包括均方根误差（Root Mean Square Error，RMSE）、平均绝对误差（Mean Absolute Error，MAE）和皮尔逊相关系（Pearson Correlation Coefficient，PCC）。

$$RMSE=\sqrt{\frac{\sum_{t=0}^{N-1}(y_t-\hat{y}_t)^2}{N}} \tag{9}$$

$$MAE=\sum_{t=0}^{N-1}|(y_t-\hat{y}_t)| \tag{10}$$

$$PCC=\frac{cov(y_t,\hat{y}_t)}{\sigma_{y_t}\sigma_{\hat{y}_t}} \tag{11}$$

其中 y_t 和 $\hat{y}_t$ 表示人格特质的真实值和预测值，N 表示用户个数，cov 为协方差，σ_{y_t} 和 $\sigma_{\hat{y}_t}$ 分别为 y_t 和 $\hat{y}_t$ 的标准差。

评价指标的多样性使得模型性能横向对比变得困难。因此，本节以大五人格指标下的 Essays 数据集和 MBTI 指标下的 Kaggle 数据集为基础，以分类任务中 F1 指标为基准对比了手工特征方法、神经网络方法、文档结构方法、预训练语言模型方法、大语言模型方法的代表性工作，结果如表 6 所列。整体来说，深度学习方法相比手工特征结合机器学习方法有很大的提升。深度学习方法的优势在于其对数据的自动学习和表征能力，通过深层神经网络的结构，可以从原始数据中提取更高层次的特征表示。相比之下，手工特征结合机器学习方法需要依赖领域专家的知识和经验来设计特征，这往往是一个耗时且具有挑战性的过程。文档结构方法得益于捕获用户生成多个帖子内容，这类长文档建模方式在处理长文本时表现出色，在 Essays 数据集上获得了最高的人格检测性能。大语言模型的应用为人格检测任务带来了新的可能性。大语言模型可以通过预训练和微调的方式，学习丰富的语言知识和解锁语境理解。大语言模型方法正在逐步适配到人格检测任务中，我们可以期待人格检测能力的进一步提升，并为个性化推荐、情感分析等领域带来更多应用和创新。

表 6 代表性模型性能对比

Table 6 Performance comparison of representative models

方法类型	方法	Big-5（Essays）					MBTI（Kaggle）			
		AGR	ON	EXT	NEU	OPN	I/E	S/N	T/F	J/P
手工特征方法	LIWC+SVM[87]	47.50	52.00	49.20	50.90	50.90				
	LIWC+SMO[34]	55.78	55.29	54.93	57.35	62.11				
神经网络方法	AttRCNN[45]	71.92	63.46	71.50	62.36	67.84	59.74	64.08	78.77	66.44
	SN+Attn[67]	70.82	64.19	72.25	68.10	68.50	65.43	62.15	78.05	63.92
文档结构方法	CGTN[70]	77.12	76.21	78.78	70.87	72.17	71.12	70.44	80.22	72.64
	DDGCN[71]	—	—	—	—	—	70.26	60.66	78.91	71.73
预训练语言模型方法	BERT-base[75]	58.80	59.20	60.00	60.50	64.60	78.30	86.40	74.40	64.40
	DesPrompt[76]	63.80	47.30	59.40	47.80	64.90	—	—	—	—
大语言模型方法	PsyAttention[78]	66.75	64.21	64.43	64.27	68.62	87.94	91.47	85.24	80.53
	PsyCoT[80]	61.13	61.13	59.74	59.74	59.74	66.56	61.70	74.80	57.83

1) https://kaggle.com/datasnaek/mbti-type

2) https://personalitycafe.com/forum

3) https://psy.takelab.fer.hr/datasets/

5 挑战与展望

作为一个新兴的跨学科研究领域，人格计算已经引起了心理学和计算机学者的广泛关注。尽管在这个领域取得了许多成果，但现有研究仍然存在一些局限和挑战。

5.1 人格检测的可靠性问题

人格检测的可靠性问题涉及人格特质测量误差和数据噪声两个层面。人格特质是潜在的理论变量，无法直接或客观地观察到。心理学中的人格工具如自我报告问卷和他人观测通常用来估计真实的人格特质。由于这种近似过程具有不确定性，因此观测到的测量值与真实值之间会存在差异，即测量误差。在自动人格检测研究中，将这些测量结果作为训练和验证的黄金标准时，测量误差可能会传播到预测结果中，导致模型的检测结果不可靠。Akrami 等[16]的研究支持了这个假设，他们发现人格检测模型在测量误差较小的小型数据集上表现更好，而在测量误差较大的大型数据集上表现较差。遗憾的是，调查发现尚未有研究针对人格检测中测量误差问题进行探索。另一个方面是数据噪声问题，文本人格检测的数据大多来源于社交网络，社交网络用户生成的数据具有多源异构、动态性、交互性等特点，但也存在数据质量参差不齐的问题。尽管社交网络数据具有样本量大的特点，但收集的用户生成文档并不总是反映用户真实的心理状况。因此，数据集中蕴含着一些与人格特质相关的假阳性文本内容，会混淆模型学习到文档与特质之间的虚假相关。

未来研究：针对测量误差问题，需认识到测量误差的存在，并采取相应的措施来最小化其对模型训练和应用的影响，这可能包括改进问卷设计、标注一致性检验等。针对数据噪声问题，通过特征过滤方法选择复杂场景中有信息量和区分度的特征，鲁棒性训练技术，提升模型容忍度可能是一种方式。

5.2 人格检测的公平性问题

机器学习中的公平性研究涉及识别和减轻系统中可能存在的偏见，特别是对特定群体的偏见[88]。有偏差的训练数据是导致算法偏差的一个重要因素。一个公平的人格检测模型应该在不同的人口统计群体中表现出相同的预测性能。以人种为例，在以美国白人为主的训练集中训练人格检测模型，当检测具有相同人格特质的非裔美国人时，模型不可避免地捕获甚至放大白人语言风格与特质的关系，从而错误推断人格特质。扩展来说，除了人种，用户年龄、性别、教育程度、文化背景、甚至发表文本时的环境都可被视为影响人格检测公平性的因素。从数据质量的角度来看，如果选择的人格测量工具在所有相关人口群体中缺乏同等的有效性和可靠性，同样会导致不同群体之间人格测量质量的差异。因此，如何通过技术手段去除这些人格检测系统内的固有存在的偏置是一个挑战。

据调研，只有少数研究将人口统计信息作为附带特征输入预测模型，或者从理论角度研究人口统计信息与人格特质的相关性，并未从公平性模型构建角度进行探索。未来，利用公平性机器学习方法实现去偏置的文本人格检测是值得探索的方向。

5.3 数据集和评价指标统一的问题

调查发现现有的数据集以私有数据集为主，也存在多个公开人格数据集可供研究人员访问。可共享的数据集是推进人格检测研究的关键，这就引出了下一个挑战：这些公开数据集基于不同的来源，比如学生的日常写作、各种社交媒体，数据集间的差异使得比较和综合分析变得困难。此外，不同的研究者和研究团队可能使用不同的标注标准，其标签来源不够准确，尤其涉及从 Twitter 或 Reddit 获得个性标签的数据集，这些标签是基于帖子或用户资料中自带的 MBTI 或大五人格信息。另外，在评价指标方面，一部分研究将人格检测建模为分类任务，使用了一组多样化的指标，即宏观、微观、加权或非加权形式的准确率、召回率和 F1。也有研究以回归任务的形式直接预测人格特质得分，评价指标包括 MSE、RMSE、皮尔逊相关系数(Pearson Correlation Coefficient)等。这是一类更精准的评价指标，对特质的标注要求更高。综上，不统一的评价指标使得综合比较和分析变得困难。

未来研究：鼓励更多的数据集的公开共享和合作研究，这有助于结果的可重复性。此外，还需要制定统一的评价指标和标准，以保证评价集的一致性和可比性。虽然流行的 MBTI 人格数据更容易采集，但是从科学研究角度来说，大五人格是现有的最准确的人格分类法。与 MBTI 的二分法相反，大五人格在每个维度是具体的人格得分。每个维度还包括一些细分方面，可以更精准地描述个体间的差异。因此，以大五人格得分回归建模可能是更合适的方式。

5.4 伦理与隐私问题

人格识别涉及隐私保护、滥用风险和个体权利等一系列伦理问题。自动化人格特质检测的是一把双刃剑，这项技术有助于对有重大影响的决策制定，比如对网络舆情控制或法律判决等，也可以应用于其他符合伦理的场景，如精神健康诊断，从而减少医疗资源的浪费。当然，这类心理定位系统可能存在被滥用的风险，尤其是对危险行为中的弱势群体有害，例如用在线赌博广告瞄准成瘾者。此外，通过对用户心理分析进行网络欺诈、恶意操纵也具有重大危害。因此，如何采取措施确保人格特质自动识别结果的正确使用，防止其被滥用或歪曲是需要学者和业界关注的问题。

未来研究：在构建公开数据集时，可采用数据匿名化和去标识化的方法，以最大程度地保护个人隐私，确保个人身份无法被直接或间接识别。在开发和使用人格检测系统时，我们主张检测系统以一种排除检索人格结果的方式单独提供。现阶段联邦学习等技术已经展现出保护数据隐私安全的可行性，未来联邦化的人格计算是值得探索的。除了技术上的改进外，道德伦理审查制度也需要完善，并确保用户的根本利益不被损害。

5.5 大语言模型与人格

大型语言模型在学术界和工业界越来越受欢迎，因为它们能在各种任务中模仿人类的智力和表现。随之而来的问题是，这些模型是否已经发展出了人格特质，能控制机器在现实世界中表现出来的行为模式吗？提出这个问题的原因是心理学研究普遍认为理解个体人格是理解其行为倾向的核心。最新的一些工作尝试利用为人类设计的问卷来测量和量化

LLMs 的人格，其潜在假设是，用于评估人类人格的工具也适用于机器。但是其可靠性还无法保证，这突出了开发工具来分析和测量语言模型人格的必要性。

未来研究：LLMs 的人格评估并不是最终目标，如何利用人格评估体系提供基准结果和发现的影响因素，实现人格可编辑的大模型进而控制其行为是值得探索的。随着 LLMs 的学习能力快速提升，LLMs 也会从海量训练语料中学习到人类异常的心理行为，如高神经质、边缘人格障碍等。如何在潜在风险的情况下正确训练 LLMs，可能是实现人类的价值、真实意图和伦理原则对齐，确保人工智能安全的一种方式。

结束语 本文对文本人格检测研究进行了系统化调研。首先，从心理学背景知识、人格检测任务模式两方面进行了归纳总结；然后从心理语言学统计方法、特征工程方法、深度学习方法和预训练语言模型 4 个方面对文本人格检测方法进行了全面梳理，并且指出了这些方法的优点和局限性，同时对现有数据集进行了全面的总结并对代表性方法的检测性能进行对比；最后，探讨了文本人格检测研究目前面临的挑战以及未来可发展的研究方向。人格计算的最终目标不仅仅是通过技术手段从数字化数据中获取和分析人的个性特征和人格特点。目前的人格计算仍然处于数据驱动的初级阶段，而心理认知与行为模式密切相关。因此，如何将人格认知原理应用于机器智能仍需要心理学、计算机科学等跨学科领域的持续探索。当前，大语言模型人工智能技术在理解和生成自然语言方面取得了显著进展，人们不由得会将人与人工智能进行比较。然而，机器还无法像人类一样通过认知来理解世界。因此，借助人格计算研究实现机器从感知到认知的能力转变也是人工智能领域的一个核心问题。

参 考 文 献

[1] KAUSHAL V，PATWARDHAN M. Emerging trends in personality identification using online social networks—a literature survey[J]. ACM Transactions on Knowledge Discovery from Data，2018，12(2)：1-30.

[2] MEHTA Y，MAJUMDER N，GELBUKH A，et al. Recent trends in deep learning based personality detection[J]. Artificial Intelligence Review，2020，53(4)：2313-2339.

[3] VINCIARELLI A，MOHAMMADI G. A Survey of Personality Computing[J]. IEEE Transactions on Affective Computing，2014，5(3)：273-291.

[4] YANG Q，NIKOLENKO S，HUANG A，et al. Personality-Driven Social Multimedia Content Recommendation[C]// Proceedings of the 30th ACM International Conference on Multimedia. 2022：7290-7299.

[5] JAISWAL S，SONG S，VALSTAR M. Automatic prediction of depression and anxiety from behaviour and personality attributes[C]// Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction. IEEE，2019：1-7.

[6] GHOSH S，MAURYA D K，EKBAL A，et al. EM-PERSONA：emotion-assisted deep neural framework for personality subtyping from suicide notes[C]// Proceedings of the 29th International Conference on Computational Linguistics. 2022：1098-1105.

[7] CHAWLA K，WU I，RONG Y，et al. Be Selfish，But Wisely：Investigating the Impact of Agent Personality in Mixed-Motive Human-Agent Interactions[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023：13078-13092.

[8] CHIEN S Y，CHEN C L，CHAN Y C. The Influence of Personality Traits in Human-Humanoid Robot Interaction[J]. The Association for Information Science and Technology，2022，59(1)：415-419.

[9] LANG Y，LIANG W，WANG Y，et al. 3d face synthesis driven by personality impression[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019：1707-1714.

[10] WU T，ZHENG K F，WU C H，et al. A Survey on Personality in Cyberspace Security[J]. Journal of Electronics & Information Technology，2020，42(12)：2827-2840.

[11] YAN D，CAO J，XIE W，et al. PersonalityGate：A general plug-and-play GNN gate to enhance cascade prediction with personality recognition task[J]. Expert Systems with Applications，2022，203：117381.

[12] YIN C，ZHANG X，LIU L. Reposting negative information on microblogs：Do personality traits matter? [J]. Information Processing & Management，2020，57(1)：102106.

[13] YANG L，LI S，LUO X，et al. Computational personality：a survey[J]. Soft Computing，2022，26(18)：9587-9605.

[14] JUNIOR J C S J，GÜÇLÜTÜRK Y，PÉREZ M，et al. First impressions：A survey on vision-based apparent personality trait analysis[J]. IEEE Transactions on Affective Computing，2019，13(1)：75-95.

[15] ZHAO X M，TANG Z W，ZHANG S Q. Research advance of multimodal personality recognition based on audio and visual cues[J]. CAAI Transactions on Intelligent Systems，2021，16(2)：189-201.

[16] AKRAMI N，FERNQUIST J，ISBISTER T，et al. Automatic extraction of personality from text：Challenges and opportunities[C]// Proceedings of IEEE International Conference on Big Data. IEEE，2019：3156-3164.

[17] ŠTAJNER S，YENIKENT S. A survey of automatic personality detection from texts[C] //Proceedings of the 28th International Conference on Computational Linguistics. 2020：6284-6295.

[18] ZHANG L，CHEN Z X，YANG B. Personality Analysis and Prediction of Social Network Users[J]. Chinese Journal of Computers，2014，37(8)：1877-1894.

[19] LEE C H，KIM K，SEO Y S，et al. The relations between personality and language use[J]. The Journal of General Psychology，2007，134(4)：405-413.

[20] DIGMAN J M. Personality structure：Emergence of the five-factor model[J]. Annual review of psychology，1990，41(1)：417-440.

[21] JUNG C G. Personality types[J]. The portable Jung，1971：178-272.

[22] COSTA J, PAUL T, MCCRAE R. Neo P ersonality Inventory [EB/OL]. https://doi. org/10. 1002/9780470479216. corpsy0590.

[23] BUTCHER J N, WILLIAMS C L, GRAHAM J R, et al. Minnesota multiphasic personality inventory-adolescent[M]. University of Minnesota Press, 1992.

[24] SATO T. The Eysenck personality questionnaire brief version: Factor structure and reliability[J]. The Journal of Psychology 2005, 139(6): 545-552.

[25] KRUMPAL I. Determinants of social desirability bias in sensitive surveys: a literature review[J]. Quality & Quantity, 2013, 47(4): 2025-2047.

[26] STACHL C, PARGENT F, HILBERT S, et al. Personality research and assessment in the era of machine learning[J]. European Journal of Personality, 2020, 34(5): 613-631.

[27] PENNEBAKER J W, FRANCIS M E. BOOTH R J. Linguistic inquiry and word count: LIWC 2001[J/OL]. https://www. researchgate. net/publication/246699633_Linguistic_inquiry_and_word_count_LIWC.

[28] COLTHEART M. The MRC psycholinguistic database[J]. The Quarterly Journal of Experimental Psychology Section A, 1981, 33(4): 497-505.

[29] PENNEBAKER J W, KING L A. Linguistic styles: language use as an individual difference[J]. Journal of Personality and Social Psychology, 1999, 77(6): 1296.

[30] PENNEBAKER J W, MEHL M R, NIEDERHOFFER K G. Psychological aspects of natural language use: Our words, our selves[J]. Annual Review of Psychology, 2003, 54(1): 547-577.

[31] TAUSCZIK Y R, PENNEBAKER J W. The psychological meaning of words: LIWC and computerized text analysis methods[J]. Journal of Language and Social Psychology, 2010, 29(1): 24-54.

[32] ARGAMON S, DHAWLE S, KOPPEL M, et al. Lexical predictors of personality type[C]// Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America. 2005: 1-16.

[33] OBERLANDER J, NOWSON S. Whose thumb is it anyway? Classifying author personality from weblogtext[C]// Proceedings of The International Conference on Computational Linguistics. 2006: 627-634.

[34] MAIRESSE F, WALKER M A, MEHL M R, et al. Using linguistic cues for the automatic recognition of personality in conversation and text[J]. Journal of Artificial Intelligence Research, 2007, 30(1): 457-500.

[35] BACHRACH Y, KOSINSKI M, GRAEPEL T, et al. Personality and patterns of Facebook usage[C]// Proceedings of the 4th Annual ACM Web Science Conference. 2012: 24-32.

[36] FARNADI G, ZOGHBI S, MOENS M F, et al. Recognising personality traits using facebook status updates[C]// Proceedings of the International AAAI Conference on Web and Social Media. 2013: 14-18.

[37] ALAM F, STEPANOV E A, RICCARDI G. Personality traits recognition on social network-facebook[C]// Proceedings of the International AAAI Conference on Web and Social Media. 2013: 6-9.

[38] VOLKOVA S, BACHRACH Y, ARMSTRONG M, et al. Inferring latent user properties from texts published in social media [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015: 4296-4297.

[39] BAI S, ZHU T, CHENG L. Big-five personality prediction based on user behaviors at social network sites[J]. arXiv: 1204. 4809, 2012.

[40] BAI S, HAO B, LI A, et al. Predicting big five personality traits of microblog users[C]// Proceedings of International Joint Conferences on Web Intelligence and Intelligent Agent Technologies. 2013: 501-508.

[41] PENG K H, LIOU L H, CHANG C S, et al. Predicting personality traits of Chinese users based on Facebook wall posts[C]// Proceedings of 24th Wireless and Optical Communication Conference. 2015: 9-14.

[42] ZHONG Y, FEI D Z. Judging Personality by Informal Words: a Sparse PCA Approach[J]. Journal of Chinese Information Processing, 2017, 31(1): 192-204.

[43] HUANG C L, CHUNG C K, HUI N, et al. The Development of the Chinese Linguistic Inquiry and Word Count Dictionary [J]. Chinese Journal of Psychology, 2012, 54(2): 185-201.

[44] MAJUMDER N, PORIA S, GELBUKH A, et al. Deep learning-based document modeling for personality detection from text [J]. IEEE Intelligent Systems, 2017, 32(2): 74-79.

[45] XUE D, WU L, HONG Z, et al. Deep learning-based personality recognition from text posts of online social networks[J]. Applied Intelligence, 2018, 48(11): 4232-4246.

[46] HERNANDEZ R K, SCOTT I. Predicting Myers-Briggs type indicator with text[C]// Proceedings of 31st Conference on Neural Information Processing Systems. 2017.

[47] TANDERA T, SUHARTONO D, WONGSO R, et al. Personality prediction system from facebook users[J]. Procedia Computer Science, 2017, 116: 604-611.

[48] XUE X, FENG J, SUN X. Semantic-enhanced sequential modeling for personality trait recognition from texts[J]. Applied Intelligence, 2021, 51(11): 1-13.

[49] SUN X, LIU B, CAO J, et al. Who am I? Personality detection based on deep learning for texts[C]// Proceedings of IEEE International Conference on Communications. 2018: 1-6.

[50] LIU F, PEREZ J, NOWSON S. A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017: 754-764.

[51] ZHOU L, ZHANG Z, ZHAO L, et al. Attention-based BiLSTM models for personality recognition from user-generated content [J]. Information Sciences, 2022, 596: 460-471.

[52] PORIA S, GELBUKH A, AGARWAL B, et al. Common sense knowledge based personality recognition from text[C]// Proceedings of Advances in Soft Computing and Its Applications: 12th Mexican International Conference on Artificial Intelli-

gence，2013：484-496.

[53] CAMBRIA E，HAVASI C，HUSSAIN A. Senticnet 2：A semantic and affective resource for opinion mining and sentiment analysis[C]// Proceedings of Twenty-Fifth International FLAIRS Conference. 2012：1-6.

[54] HAVASI C，SPEER R，ALONSO J. ConceptNet 3：a flexible，multilingual semantic network for common sense knowledge [C]// Proceedings of Recent Advances in Natural Language Processing. 2007：27-29.

[55] HAN S，HUANG H，TANG Y. Knowledge of words：An interpretable approach for personality recognition from social media [J]. Knowledge-Based Systems，2020，194：105550.

[56] ZHU Y，HU L，NING N，et al. A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection[J]. Knowledge-Based Systems，2022，249：108952.

[57] ARKONI T. Personality in 100，000 words：A large-scale analysis of personality and word use among bloggers[J]. Journal of Research in Personality，2010，44(3)：363-373.

[58] RAMEZANIM，FEIZI-DERAKHSHIMR，BALAFARMA. Knowledge graph-enabled text-based automatic personality prediction [J]. Computational Intelligence and Neuroscience. 2022，1：3732351.

[59] ZHU Y，GUAN Z，WEI S，et al. PerKG：A Personality Knowledge Graph for Personality Analysis[C]// Proceedings of IEEE International Conference on Systems，Man，and Cybernetics. 2022：580-585.

[60] LI M，LIU H，WU B，et al. Language Style Matters：Personality Prediction from Textual Styles Learning[C]// Proceedings of IEEE International Conference on Knowledge Graph. 2022：141-148.

[61] LI M，ZHU Y，LI S，et al. HG-PerCon：Cross-view contrastive learning for personality prediction[J]. Neural Networks，2024，169：542-554.

[62] SUN X，LIU B，MENG Q，et al. Group-level personality detection based on text generated networks[J]. World Wide Web，2020，23(3)：1887-1906.

[63] ROVER A，LESKOVEC J. node2vec：Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016：855-864.

[64] GUAN Z，WU B，WANG B，et al. Personality2vec：Network representation learning for personality [C]// Proceedings of IEEE Fifth International Conference on Data Science in Cyberspace. 2020：30-37.

[65] WANG Z，WU C H，LI Q B，et al. Encoding text information with graph convolutional networks for personality recognition [J]. Applied Sciences，2020，10(12)：4081.

[66] YAO L，MAO C，LUO Y. Graph convolutional networks fortext classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019：7370-7377.

[67] LYNN V，BALASUBRAMANIAN N，SCHWARTZ H A. Hierarchical modeling for user personality prediction：The role of message-level attention[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：5306-5316.

[68] YANG F，QUAN X，YANG Y，et al. Multi-document transformer for personality detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021：14221-14229.

[69] YANG T，YANG F，OUYANG H，et al. Psycholinguistic Tripartite Graph Network for Personality Detection[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021：4229-4239.

[70] ZHU Y，HU L，GE X，et al. Contrastive Graph Transformer Network for Personality Detection [C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. 2022：4559-4565.

[71] YANG T，DENG J，QUAN X，et al. Orders are unwanted：dynamic deep graph convolutional network for personality detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023：13896-13904.

[72] ZHU Y，XIA Y，LI M，et al. Data Augmented Graph Neural Networks for Personality Detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：664-672.

[73] KEH S，CHENG I. Myers-Briggs personality classification and personality-specific language generation using pre-trained language models[J]. arXiv：1907. 06333，2019.

[74] JIANG H，ZHANG X，CHOI J D. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：13821-13822.

[75] MEHTA Y，FATEHI S，KAZAMEINI A，et al. Bottom-up and top-down：Predicting personality with psycholinguistic and language model features[C]// Proceedings of IEEE International Conference on Data Mining. IEEE，2020：1184-1189.

[76] WEN Z，CAO J，YANG Y，et al. DesPrompt：Personality-descriptive prompt tuning for few-shot personality recognition[J]. Information Processing & Management，2023，60(5)：103422.

[77] CHEN L，WU Y，LI Q，et al. Mining the User's Personality with an Attention-based Label Prompt Method[J]. IEEE Intelligent Systems，2023，39(2)：31-39.

[78] ZHANG B，HUANG Y，CUI W，et al. PsyAttention：Psychological Attention Model for Personality Detection[C]// Findings of the Association for Computational Linguistics：EMNLP 2023. 2023：3398-3411.

[79] GANESANA V，LAL Y K，NILSSON A H，et al. Systematic Evaluation of GPT-3 for Zero-Shot Personality Estimation[C]// Proceedings of the 13th Workshop on Computational Approaches to Subjectivity，Sentiment，and Social Media Analysis. 2023：390-400.

[80] YANG F，YANG T，QUAN X，et al. Learning to answer psychological questionnaire for personality detection[C]// Findings of the Association for Computational Linguistics. 2021：1131-1142.

[81] YANG T，SHI T，WAN F，et al. PsyCoT：Psychological Ques-

tionnaire as Powerful Chain-of-Thought for Personality Detection[C]// Findings of the Association for Computational Linguistics. 2023:3305-3320.

[82] CARON G, SRIVASTAVA S. Manipulating the Perceived Personality Traits of Language Models[C]// Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 2370-2386.

[83] JIANG G, XU M, ZHU S C, et al. Evaluating and inducing personality in pre-trained language models[C]// Proceedings of Advances in Neural Information Processing Systems. 2023.

[84] BIEL J I, TSIMINAKI V, DINES J, et al. Hi YouTube! Personality impressions and verbal content in social video[C]// Proceedings of the 15th ACM on International Conference on Multimodal Interaction. 2013:119-126.

[85] JIANG H, ZHANG X, CHOI J D. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 13821-13822.

[86] GJURKOVIĆ M, KARAN M, VUKOJEVIĆ I, et al. PANDORA Talks: Personality and Demographics on Reddit[C]// Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. 2021:138-152.

[87] IGHE E P, URETA J C, POLLO B A L, et al. Personality Trait Classification of Essays with the Application of Feature Reduction[C]// Proceedings of SAAIP@ IJCAI. 2016:22-28.

[88] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys, 2021, 54(6):1-35.

ZHU Yangfu, born in 1992, postgraduate. His main research interests include data mining and personality computing.

WU Bin, born in 1969, Ph.D, professor, Ph.D supervisor. His main research interests include social computing and complex network.

（责任编辑：何杨）