计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):48−58 [doi: 10.15888/j.cnki.csa.009888] [CSTR: 32024.14.csa.009888]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

置信学习驱动下融合心理语言学特征的人格检测①

王春东[1,2], 杨宇涵[1,2], 林　浩[1,2], 黄思源[3]

[1](天津理工大学 计算机科学与工程学院, 天津 300384)
[2](天津理工大学 计算机病毒防治技术国家工程实验室, 天津 300384)
[3](河北工业大学 人工智能与数据科学学院, 天津 300401)
通信作者: 王春东, E-mail: michael3769@163.com

摘　要: 随着互联网的普及, 越来越多用户倾向于在社交平台公开表达自己的个人细节和情感内容, 这些网络文本数据往往体现着不同场景下的真实表达, 反映了用户内在的心理特质及人格倾向. 近年来, 基于社交文本的人格检测研究取得了显著进展, 然而, 研究者们大多直接使用未经处理的公开数据集, 这些数据集因其收集过程导致不可避免地存在噪声, 此外, 大多过分依赖预训练模型提取的文本语义特征, 而缺乏对心理语言学特征的引入. 为了解决以上问题, 提出一种新型的人格检测研究方法. 该方法首先基于置信学习完成噪声数据清洗, 提高数据集质量. 其次, 提取多层次心理语言学特征来填补单一文本语义特征的不足. 最后通过动态深度图卷积网络来优化特征表达, 完成最终的人格检测任务. 在公开的 Kaggle MBTI 数据集上对该方法进行性能评估, 结果表明, 与目前先进的方法相比, 该方法在准确率和 *F*1 值上分别提升了 5.48% 和 4.22%.

关键词: 人格检测; 社交媒体; 数据清洗; 迈尔斯布里格斯类型指标

引用格式: 王春东,杨宇涵,林浩,黄思源.置信学习驱动下融合心理语言学特征的人格检测.计算机系统应用,2025,34(7):48–58. http://www.c-s-a.org.cn/1003-3254/9888.html

Personality Detection with Psycholinguistic Feature Driven by Confident Learning

WANG Chun-Dong[1,2], YANG Yu-Han[1,2], LIN Hao[1,2], HUANG Si-Yuan[3]

[1](School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China)
[2](National Engineering Laboratory for Computer Virus Prevention and Control Technology, Tianjin University of Technology, Tianjin 300384, China)
[3](School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China)

Abstract: With the widespread use of Internet, an increasing number of users are inclined to share personal details and emotion on social platforms. These online text data often capture genuine expressions in various contexts, reflecting the users' internal psychological traits and personality tendencies. In recent years, research on personality detection based on social media text has made significant progress. However, most researchers rely on unprocessed public datasets, which inevitably contain noise due to their collection process. In addition, there is an over-reliance on semantic features extracted by pre-trained models, with insufficient attention to psycholinguistic features. To address these issues, this study proposes a novel method for personality detection. First, a plug-and-play data cleaning module based on confident learning is used to remove noisy data and improve dataset quality. Second, multi-level psycholinguistic features are extracted to complement the semantic features of the text. The proposed method is evaluated on the public Kaggle MBTI dataset, with results showing that, compared to existing advanced methods, it achieves improvements of 5.48% in accuracy and 4.22% in *F*1-*score*.

Key words: personality detection; social media; data cleaning; Myers-Briggs type indicator (MBTI)

① 基金项目: 国家重点研发计划 (2023YFB2703900); 天津市科委重大专项 (15ZXDSGX00030)
收稿时间: 2024-12-02; 修改时间: 2025-01-02; 采用时间: 2025-01-21; csa 在线出版时间: 2025-04-25
CNKI 网络首发时间: 2025-04-27

人格被定义为行为模式、认知模式、情感模式和思维模式的特征集合[1]. 因此, 任何涉及对人类行为的理解、分析、预测的技术都不可或缺地需要考虑人格. 由于个性化推荐系统[2]、人机交互、网络空间安全等众多应用场景的需求, 人格检测这一新兴任务引起了计算心理语言学和自然语言处理研究人员的极大兴趣. 现有的人格检测使用文本[3]、图像[4]、音频、视频[5]甚至脑电图[6]来完成, 因为文本获取相对容易, 所以基于文本的检测方法成为最普遍且实用的方法, 也成为其他复杂检测方法的基石.

近年来, 全球范围内的社交平台用户数量已达到数十亿级别, 大量用户倾向于在社交平台公开表达自己的个人细节和情感内容, 这些内容与用户的人格特质显著相关, 成为自动化人格检测研究的优质数据来源[7–10]. 早期的人格检测研究主要集中在使用统计方法提取文本的语言特征, 并基于这些特征完成后续的人格检测任务. 后来, 该领域的研究专注于基于深度学习的人格检测方法, 通过使用更有效的嵌入生成方法[11–13], 在人格检测方面取得了相当大的进步.

人格检测的研究虽然取得了显著的成果, 但仍面临着一些普遍的问题. 当前研究者们大多使用公开的数据集来完成人格检测任务, 但是他们却忽略了这些数据集中不可避免地存在着噪声数据. 带有噪声的数据集十分普遍, 研究者们使用以模型为中心的方法通常掩盖了训练数据中存在的噪声问题[14]. 当面对大量噪声数据时, 对于有监督的人格检测任务, 模型的表现可能不尽如人意, 或是存在偏差. 同时, 一些人格检测方法过于依赖预训练语言模型进行迁移学习而忽略了关键的心理语言学特征. Vinciarelli 等人[15]指出, 在心理语言学理论中, 人对词语的选择不仅由词语本身的意义决定, 还受到情绪、态度、人格特质等心理现象的影响, 适当地加入心理语言学特征可以提升模型的性能以及可解释性.

为解决上述问题, 本文提出了一种置信学习驱动下融合心理语言学特征的人格检测方法. 该模型针对目前人格检测数据集不可避免地存在数据噪声问题, 提出了基于置信学习的数据清洗方法, 通过识别并去除噪声标签数据来完成数据清洗, 优化了人格检测数据集的质量, 同时提取了多种新的心理语言学特征, 补充了单一文本语义特征的不足. 最终, 为有效构建社交媒体中不同帖子之间的互补关系, 充分利用用户帖子, 模型分别为每个用户构建基于社交文本特征的动态图, 并使用深度图卷积网络进行迭代学习, 最终通过聚合后的节点特征完成人格检测任务.

1 相关工作

1.1 人格理论

人格理论不仅提升了心理学对个体差异的研究精度, 也为其与计算机领域的结合提供了理论依据. 人格的分类通常从不同的维度来定义, 迄今为止, 对个体差异的心理学研究已经衍生出许多人格特质模型. 其中, MBTI 作为世界上最常用的人格特质模型之一, 广泛应用于各种场景[16].

MBTI 人格理论[17]是在心理学家卡尔·荣格先生对于心理类型划分的基础上研究加以发展, 形成的共 4 个维度、8 个极的 16 种描述人格的类型, 如表 1 所示. 用户的 MBTI 人格类型可由 MBTI 性格量表对用户人格进行评估测量得到, 这些类型通过 4 个维度的二元分类组合在一起.

表 1 MBTI 人格特质

维度	类型	
精力支配	E (外倾)	I (内倾)
认识世界	S (实感)	N (直觉)
判断事物	T (思维)	F (情感)
生活态度	J (判断)	P (知觉)

研究 MBTI 人格特质的检测方法为心理学家提供了广泛的研究样本、技术和实际应用, 因此, 本文针对 MBTI 人格检测进行研究.

1.2 人格检测

人格是由环境和生物因素随着时间的推移而形成的认知、行为和情感的集合. 同时, 大多数理论认为人格是相对稳定的[18].

传统的人格检测依赖于通过问卷调查和实验室研究收集的自我报告, 但这种方法需要对评估人员进行人工评估和培训, 耗时费力. 同时, 参加测试的人可能会受到参照群体效应的影响, 并因存在反应偏差而以一种更有利的方式回答问题[19]. 随后, 研究人员开始寻求新的方法, 尝试应用语言和自然语言处理技术从文本数据中推断出人格.

Mairesse 等人[20]首次应用统计学来从文本和对话的语言线索中识别人格, Zhu 等人[21]提出了一种新的词汇心理语言学知识导向图神经模型, 用于可解释人

格检测. 大量的研究展示了语言和人格之间更详细的联系[22,23]. 与此同时, 其他研究也表明[24,25], 将情感、用户兴趣、观点和品牌偏好等与人格相关的变量作为特征, 可以提高人格检测的效果.

随着预训练语言模型的出现, 越来越多的学者将其应用于人格检测任务. Jiang 等人[26]提出了一种基于预训练语言模型 RoBERT 和注意神经网络的自动人格检测方法. Wang 等人[27]提出了一种结合胶囊网络和预训练语言模型 XLNet 的人格检测框架. 近年来, 一些研究者将心理语言学特征与预训练的语言特征相结合来完成人格检测任务. Kazameini 等人[28]将 BERT 提取的特征与 Mairesse 特征进行拼接, 并将拼接后的特征输入到多个 SVM 中进行集成学习. Ren 等人[11]将 BERT 提取的特征与 SenticNet 特征相结合, 提高了人格检测模型的性能.

大量研究表明, 心理语言学特征在人格检测中发挥着重要作用, 将心理语言学特征与预训练语言模型提取的特征相结合提升了人格检测模型的准确性和可解释性.

1.3 数据清洗

随着数据规模的扩大, 标签不一致性和噪声问题成为人格检测模型面临的主要挑战. 以往的研究发现, 不同数据集采集方法中普遍存在着标签不一致问题, 这不仅显著降低了模型的预测性能, 还导致了评价偏差[29]. 即使在黄金标准的基准数据集中, 测试集错误也很常见, 这可能会误导研究者选择劣质模型进行部署.

近年来, Northcutt 等人[14]首次提出了置信学习 (confident learning) 的概念, 为解决标签噪声问题提供了有效工具. 不同于直接修改模型或损失函数来适应噪声标签, 置信学习通过分析模型预测的置信度和数据标签之间的不一致性, 能够推断出标签错误的可能性, 从而极大地提高监督学习任务的鲁棒性和准确性. 近年来, 基于置信学习的数据清洗因其出色的表现得到了研究者们的广泛应用. 赵宏伟等人[30]结合置信学习和协同训练修正低质量标注数据, 使得模型效果得到有效改善. Wang 等人[31]提出了基于置信学习的 CGEC 组件对数据进行去噪, 以提高代码漏洞检测数据集的质量.

在人格检测任务中, 数据集的质量直接影响最终的模型性能, 因此数据清洗成为优化模型效果的重要环节. 目前, 在人格检测领域, 数据清洗问题仍未得到足够重视, 针对噪声数据的处理, 将成为提升人格检测模型性能的关键方向之一.

2 模型架构

本文提出的融合置信学习与心理语言学特征的人格检测方法整体架构如图 1 所示, 该模型基于置信学习完成数据清洗, 同时通过动态的深度图卷积网络来学习每个用户的文本语义和心理语言学特征表示, 最终通过聚合后的特征完成人格检测任务. 下面对模型进行详细介绍.

2.1 数据清洗

数据清洗主要过程如图 2 所示, 其主要分为 3 个步骤.

(1) 置信联合矩阵的构建

令数据集表示为 $X:=(x,\tilde{y})^n\in(R^d,[m])^n$, 其中, 样本为 x, 共 n 个, 其中含有噪声的初始标签为 $\tilde{y}$, 共 m 种类别. 首先, 数据集通过在参数为 θ 的初始模型 model 下进行交叉验证得到第 i 个样本在第 j 个类别下的样本预测概率 $\hat{p}[i][j]$, 并计算每个类别的平均概率作为置信阈值, 同时在处理这些计算时考虑了分类样本的数量. 第 j 个类别的置信阈值 t_j 计算公式如下:

$$t_j=\frac{\sum\limits_{x\in X_{\tilde{y}=j}}\hat{p}(\tilde{y}=j;x,\theta)}{\left|X_{\tilde{y}=j}\right|} \tag{1}$$

其中, $X_{\tilde{y}=j}$ 表示所有真实标签为 j 的样本集合, $\left|X_{\tilde{y}=j}\right|$ 表示真实标签为 j 的样本数量.

这个阈值反映了模型对特定类别的平均置信度, 用于确定一个样本是否被可信地认为属于其预测的类别, 并帮助识别和清洗错误标签. 如果某个样本的预测概率大于或等于这个阈值, 则可以更自信地认为其标签是正确的; 反之, 则可能需要进一步的检查或修正.

随后, 定义置信联合矩阵 $C_{\tilde{y},y^*}$ 来计算预测类别和给定类别之间的统计矩阵:

$$C_{\tilde{y},y^*}[i][j]:=\left|\hat{X}_{\tilde{y}=i,y^*=j}\right| \tag{2}$$

其中,

$$X_{\tilde{y}=i,y^*=j}:=\Big\{x\in X_{\tilde{y}=i}:\hat{p}(\tilde{y}=j;x;\theta)\geqslant t_j,$$

$$j=\underset{l\in[m]:\hat{p}(\tilde{y}=l;x;\theta)\geqslant t_l}{\arg\max}\ \hat{p}(\tilde{y}=l;x,\theta)\Big\} \tag{3}$$

其中, y^* 是在参数为 θ 的初始模型 model 下预测的标签.

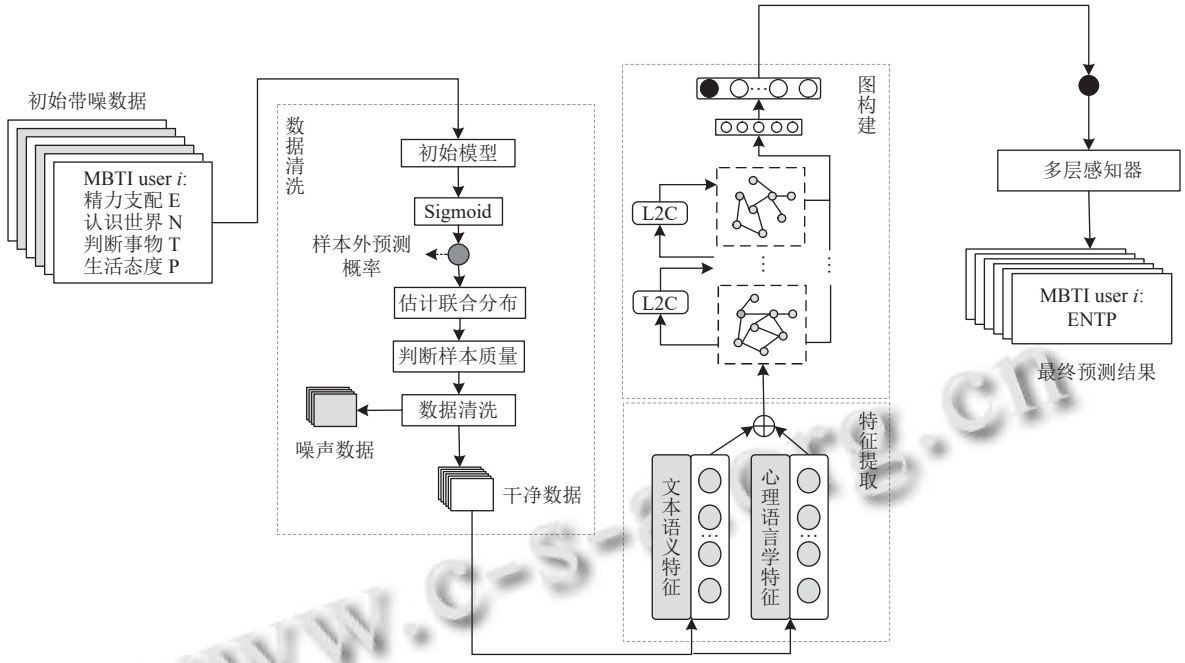

图 1 模型架构

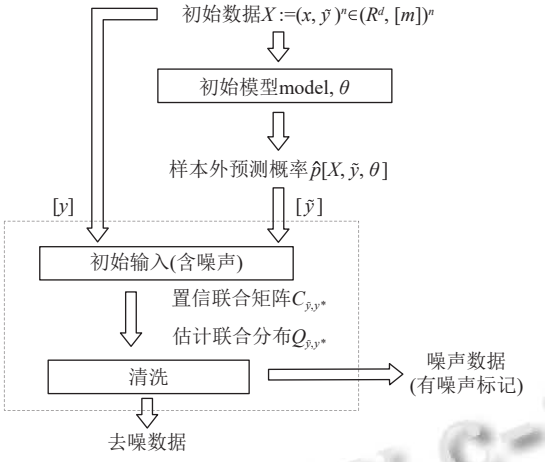

图 2 基于置信学习的数据清洗

(2) 估计联合分布

噪声标签和真实标签的联合分布$Q_{\tilde{y},y*}$通过置信联合矩阵来进行估计, 并表征数据集X中的初始标签中的噪声. 最后可以计算得到噪声标签和真实标签的联合分布:

$$Q_{\tilde{y}=i,y^*=j}=\frac{\dfrac{C_{\tilde{y}=i,y^*=j}}{\sum_{j\in[m]}C_{\tilde{y}=i,y^*=j}}\cdot\left|X_{\tilde{y}=i}\right|}{\sum_{i\in[m],j\in[m]}\left(\dfrac{C_{\tilde{y}=i,y^*=j}}{\sum_{j'\in[m]}C_{\tilde{y}=i,y^*=j'}}\cdot\left|X_{\tilde{y}=i}\right|\right)} \tag{4}$$

其中, $\left|X_{\tilde{y}=i}\right|$为$\tilde{y}$取到标签$i$的样本总数.

这一步骤估计了每个标签类别上的噪声率, 从而为后续的样本过滤提供依据. 其中$Q_{\tilde{y},y^*}$的对角元素表示样本标签被正确标记的概率, 非对角元素表示错误标记的概率.

(3) 查找并过滤错误样本

根据计算得到的噪声率, 识别并过滤掉标签错误的样本, 如果$Q_{\tilde{y}=i,y^*=j}$超过j类的平均阈值, 则判断样本标记错误.

在人格检测领域, 某些人格维度分布极其不平衡, 动态阈值方法根据类别的实际预测概率调整阈值, 能够更精准地反映模型对各类别的信心, 与其他传统的方法相比, 使用动态阈值对类失衡问题具有更好的鲁棒性. 该步骤通过对联合分布的估计, 可以有效地找出噪声样本, 大大减少人工检查的工作量, 降低数据集清洗的成本.

2.2 特征提取

BERT 学习的上下文嵌入相对于传统词嵌入具有理论和经验优势, 在文本分类任务中, 其 CLS 向量常常被用作整个输入序列的表示, 因此本文提取 BERT 预训练模型的 CLS 向量作为文本语义特征. 针对心理语言学特征, LIWC 常用于自然语言处理中的人格检测任务, 但由于它不是开放获取的, 所以本文选取了 SenticNet7 (https://sentic.net/downloads/)、TextbBlob (https://pypi.org/project/textblob/) 和 VADER (https://github.com/

vaderSentiment/vaderSentiment) 词典进行特征提取, 特征相关介绍如表 2 所示.

对于最多可以接受 512 个 token 序列的 BERT 模型而言社交媒体平台中用户的样本数据量太大, 无法进行完整的文本语义特征提取. 因此为了充分利用用户发布的每一条帖子, 模型对用户的每个帖子分别进行文本语义特征、心理语言学特征的提取, 其具体结构如图 3 所示.

表 2 心理语言学特征介绍

特征	维度	介绍
SenticNet7	10	SenticNet7词典提供了多层次的情感表示, 为了更准确地理解文本中的人格, 提取了包括情感极性、内省、脾气、态度、敏感性等在内的细粒度特征
VADER	3	VADER词典被设计用于处理社交媒体上的文本数据, 考虑了情感表达的非正式性, 因此基于该词典提取了文本的积极情绪、消极情绪以及中性情绪特征
TextBlob	2	TextBlob常用于自然语言处理和文本分析, 基于此提取了文本的情感极性以及主观性特征

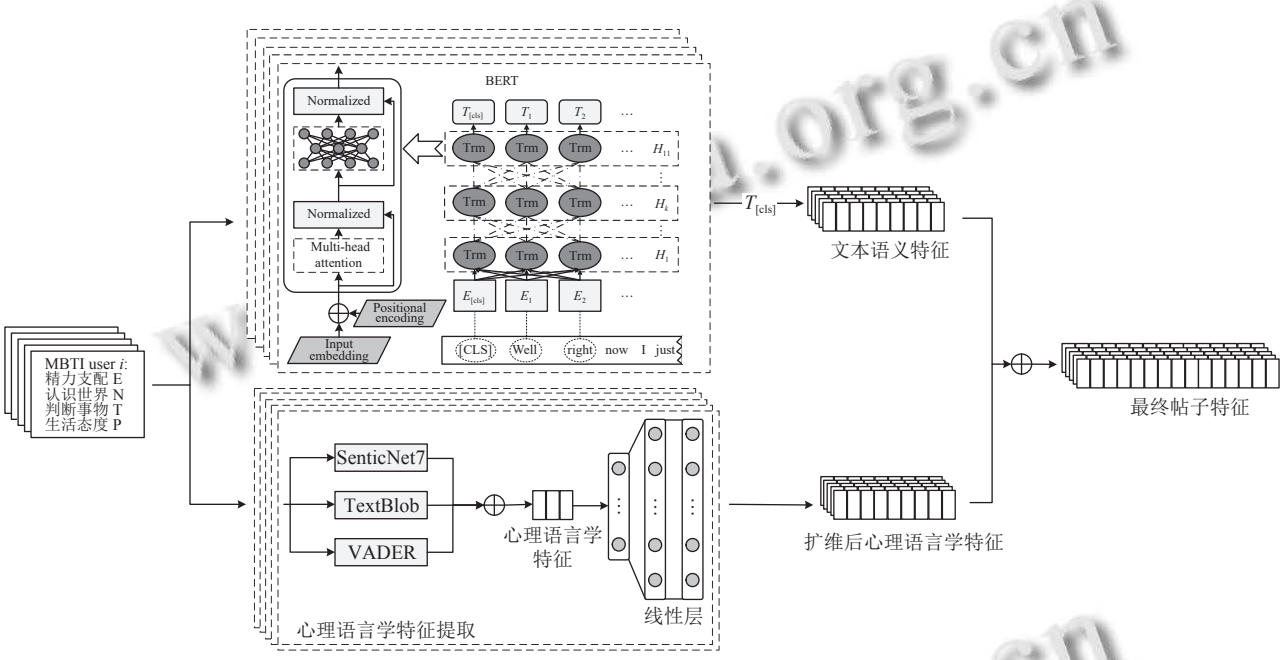

图 3 帖子特征提取

首先将用户的多个帖子分别输入到 BERT 模型中, 用提取到的每个 CLS 向量来表示每个用户的社交文本的语义特征.

对于每个用户的第 i 个帖子有其文本语义特征为:

$$cls_i = BERT(p_i) \tag{5}$$

对于每个用户的第 i 个帖子有其提取的初始心理语言学特征为:

$$psy_i = SenticNet7(p_i) \oplus TextBlob(p_i) \oplus VADER(p_i) \tag{6}$$

将提取到的心理语言学特征附加到文本语义特征上, 但与 768 维度的文本语义特征相比, 15 维度的心理语言学特征对于神经网络来说维度过低. 为了二者的充分融合, 模型将心理语言学特征通过线性层进行扩维变换后映射到更多维度再与文本语义特征进行融合, 以进行后续的图构建工作. 对于每个用户的第 i 个帖子, 得到其最终的特征 f_i 为:

$$psy_i' = Linear(psy_i) \tag{7}$$

$$f_i = cls_i \oplus psy_i' \tag{8}$$

2.3 图构建

模型为每个用户构建一个图结构, 并通过深度图卷积网络进行节点更新, 最终通过更新后的节点特征完成人格检测任务, 其具体结构如图 4 所示.

首先将用户的每个社交媒体帖子通过先前描述的方法提取出文本语义和心理语言学特征, 二者融合后组成图中的节点. 考虑到每个节点可能只反映了用户在特定时间和情境下的行为和状态, 创建一个可以代表用户整体人格倾向的全局节点 f_0, 其特征向量通过计算所有个别节点的特征向量的算术平均值得到:

$$f_0 = \frac{1}{N}\sum_{i=1}^{N} f_i \tag{9}$$

对于每位用户, 其图中所有的节点表示为:

$$F=\{f_1,f_2,\cdots,f_N,f_0\} \tag{10}$$

随后将上述节点直接构建成一个全连接的图, 并参考 Yang 等人[32]的方法, 通过 L2C 学习连接模块动态调整并优化图的结构.

每一层的邻接矩阵A_k通过以下方式进行计算:

$$A_k=L2C(F_{k-1}) \tag{11}$$

其中, F_{k-1}表示深度图卷积网络在$k-1$层的节点表示, $L2C(\cdot)$是确定两个节点之间是否存在边的函数.

在计算出动态调整的邻接矩阵后, 用户的节点被输入到深度图卷积网络中进行编码, 节点表示的更新方式为:

$$F_{k+1}=\hat{A}_{k+1}F_k \tag{12}$$

其中, $\hat{A}=D^{-\frac{1}{2}}(A+I)D^{-\frac{1}{2}}$是标准化对称邻接矩阵.

得到所有L层的节点表示矩阵$\mathbf{F}=\left[F^0,F^1,\cdots,F^L\right]$后, 为避免浅层或深层信息的丢失, 模型通过一个可训练的投影向量$\vec{c}$与节点表示的点积计算权重, 利用 Sigmoid 函数获得各层的权重, 经过加权求和后得到最终的节点表示F^{out}:

$$F^{\text{out}}=Reshape(\sigma(\mathbf{F}\cdot\vec{c}))\otimes\mathbf{F} \tag{13}$$

其中, $\otimes$表示矩阵乘法, $Reshape(\cdot)$通过重塑操作来确保二者形状匹配.

在上述深度图卷积网络中, 信息在节点之间流动, 促进了特征间的相互作用和融合.

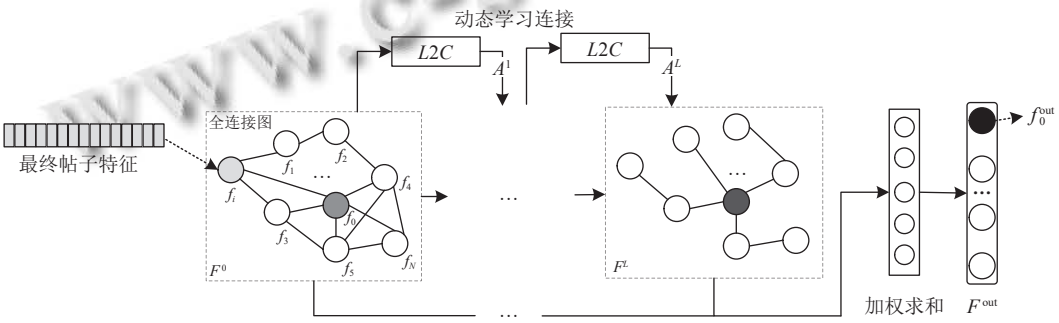

图 4 图构建

2.4 人格分类器

在第 2.3 节, 模型为每个用户的图增加了一个特殊的全局节点f_0, 用于汇总用户帖子信息. 在最终的人格分类的任务中, 将F^{out}节点中的f_0^{out}输入到一个包含两个隐藏层的多层感知器 (MLP), 采用的激活函数为 ReLU, MLP 通过对这些特征进行深入的非线性转换, 预测用户的人格特质.

3 实验分析

3.1 数据集

具有标准人格标签的人格数据集很少, 而且由于隐私问题以及寻找专业心理学家的成本, 它们的收集非常困难. 因此实验采用了被广泛使用的人格基准数据集: Kaggle MBTI 数据集 (https://github.com/TaoYang225/TrigNet). 图 5 统计了 Kaggle MBTI 数据集中不同人格类型的分布.

该数据集收集了 8 675 名用户从非正式在线社交媒体 Personality Cafe 论坛上发布的内容, 所有用户都表明了自己的 MBTI 人格类型. 用户的平均文本长度是 1 288 个单词, 每个文本都由 50 个社交媒体帖子组成, 整个数据集的总大小约为 1 120 万词.

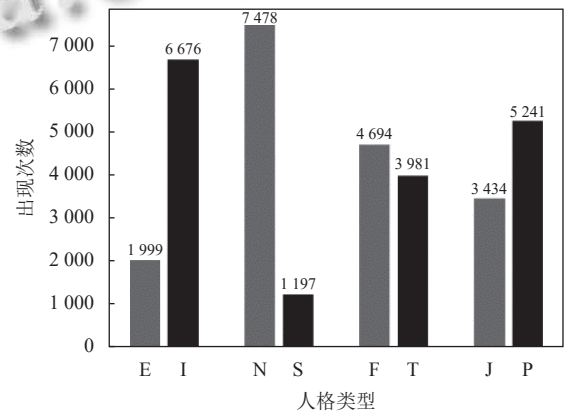

图 5 数据集人格类型统计

3.2 实验指标

本文采用准确率 (accuracy, *Acc*) 和$F1$值 ($F1$ score) 作为模型性能的评价标准, 准确率和$F1$值的定义如下:

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$P=\frac{TP}{TP+FP} \tag{15}$$

$$R=\frac{TP}{TP+FN} \tag{16}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{17}$$

其中, TP为正样本预测正确数量; TN为正样本预测错误数量; FP为负样本预测正确数量; FN为负样本预测错误数量; P为精准率, 表示正确预测的正样本占实际预测正样本的比例; R为召回率, 表示正确预测的正样本占总正样本的比例. 在数据集样本失衡的情况下, $F1$值比准确率更好地度量了性能.

3.3 实验设置

本文所有实验均使用 Python 3.8 进行编写运行, 使用的深度学习框架为 PyTorch 2.0.1, CPU 为 Intel Core i5 12490F, GPU 为 NVIDIA A40.

模型使用 BERT-base-cased 预训练模型, 该模型网络层数为 12, 隐藏层维度为 768, 注意力头为 12, 模型参数总数超过 1 亿. 模型训练时, 损失函数选用交叉熵和 $L2C$ 模块的 L0 范数联合, 并添加 dropout 和 layer normalization 来增强模型泛化能力. 实验使用的关键参数设置如表 3 所示.

表 3 实验参数设置

参数	值
文本语义特征维度	768
心理语言学特征维度	128
MLP隐藏单元	128
Batch size	8
学习率	1×10^{-5}
Epoch	15
Dropout	0.2
优化函数	Adam

3.4 数据清洗实验

为获得数据的样本外预测概率, 首先将数据集中的数据划分为 5 个不相交子集, 用于进行五折交叉验证. 每一折中, 轮流选择一个子集作为测试集, 其余 4 个子集作为训练集. 为避免数据集划分导致的样本不均匀问题, 采用 StratifiedKFold 分层抽样方法进行数据集划分. 模型在每一折的训练集上进行训练, 并在对应的测试集上进行预测, 获得样本外预测概率.

随后, 利用样本外预测概率和原始标签估计标签和预测之间的联合分布, 来理解模型预测和实际标签之间的关系, 识别出可能的噪声标签. 这一步基于预测概率和标签之间的矛盾程度, 特别是那些预测概率高但标签错误的样本, 将识别为有错误标签的样本从训练集中移除来清洗数据集, 减少噪声影响, 从而提高模型的训练质量和稳健性. 例如, 针对 E/I 人格维度计算得到的部分数据集质量情况如表 4 所示.

表 4 数据集质量情况

文本	是否噪音标签	标签质量分数	原始标签	预测标签
Then you didn’t read my OP. Go back and …	False	0.375 732 63	1	0
I’ve been wondering why I do that too …	True	0.330 285 9	0	1
'https://www.youtube.com/watch?v=3sT569-ftd …	True	0.000 878 889 87	1	0
'I love you, I loved you Don’t open my …	False	0.999 995 8	0	0
'When s/he seems afraid to tell you that …	False	0.533 702 73	1	0
…	…	…	…	…

注: 英文文本前的“”为数据集中文本自带内容

清洗后得到每个人格维度的干净数据、噪声数据占比如图 6 所示. 每个人格维度具有不同的噪声率.

最后, 将噪声数据进行过滤, 并使用清洗后的数据重新训练模型, 使得模型在一个更清洁、更少噪声的数据集上进行训练, 模型因此能够学习到更准确的特征表示和决策边界. 同时, 为了使得模型能够得到更优异的表现, 进一步调整模型参数, 不断优化.

3.5 对比实验

为了进行综合评价和比较, 采用以下人格检测模型作为基准进行比较:

Transformer-MD[33]: 该模型通过 BERT 对用户帖子进行顺序编码并存储在内存中, 使得帖子可以访问前一个帖子的信息.

TrigNet[34]: 该模型在 LIWC 中使用一种改进的 GAT 来融合帖子, 并使用心理语言学知识构建图来完成任务.

Novel-PD[35]: 该模型通过数据增强技术来帮助训练更准确和健壮的模型, 同时引入心理语言学特征用

于人格检测.

Longformer-ML[36]: 该模型结合情感分类与人格检测任务, 并通过外部情感标注模型对文档数据进行增强, 从而提高检测效果.

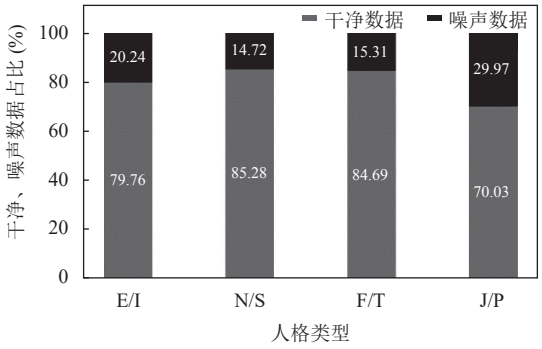

图 6 数据清洗结果

在表 5、表 6 中, 本文的研究结果与已有的研究结果进行了比较. 实验结果表明, 相较于其他模型, 本文提出的方法在各人格维度上表现出色. 尽管在 N/S 维度的 $F1$ 值未达到最优, 但本文方法在其他维度的表现更为均衡, 整体效果更加突出. 与当前最优的 Longformer-ML 模型相比, 本文方法的准确率平均提高了 5.48%, $F1$ 值平均提高了 4.22%.

表 5 对比实验准确率结果 (%)

方法	E/I	N/S	F/T	J/P	均值
Transformer-MD	76.69	86.45	78.21	67.98	77.33
TrigNet	77.43	86.37	78.07	68.06	77.48
Novel-PD	79.51	87.09	77.24	71.87	78.93
Longformer-ML	85.93	92.22	84.38	80.69	85.81
本文方法	**91.77**	**94.06**	**93.06**	**86.27**	**91.29**

表 6 对比实验 $F1$ 值结果 (%)

方法	E/I	N/S	F/T	J/P	均值
Transformer-MD	66.08	69.10	79.19	67.50	70.47
TrigNet	69.54	67.17	79.06	67.69	70.86
Novel-PD	78.68	87.39	77.34	70.36	78.44
Longformer-ML	79.57	81.69	84.74	79.33	81.33
本文方法	**85.54**	**79.26**	**93.04**	**84.38**	**85.55**

仅依赖预训练语言模型的 Transformer-MD, 由于仅提取了单一的文本语义特征, 其各项指标均低于其他模型. 这一结果恰好表明, 整合多层次的心理语言学特征能够捕捉到用户帖子中更全面和细致的信息. 相比于未使用图结构的 Novel-PD 和采用静态图结构的 TrigNet, 本文方法通过动态深度图卷积网络适配社交媒体数据, 能够更好地理解数据内部的关系和依赖性, 从而进一步提升了模型在人格检测任务中的表现, 获得了优越的性能表现. 对于依赖外部情感分析模型标注质量的 Longformer-ML, 本文的方法基于心理语言学词典能够捕捉到更细粒度且更精准的情感信息, 有利于深入理解复杂的人格特征, 推动模型性能的进一步提升. 同时, 本文提出的基于置信学习的数据清洗方法有效减少了数据集中噪声的影响, 确保模型在更高质量的数据上进行训练, 降低了训练过程中的干扰, 从而提升了准确率和 $F1$ 值.

3.6 消融实验

为了验证各模块的有效性, 模型在 Kaggle MBTI 数据集上进行消融实验, 并以准确率和 $F1$ 值作为评价指标.

首先, 基准模型设定为不包含心理语言学特征且未进行数据清洗的初始人格检测模型. 随后, 基准模型与以下 5 种模型进行了对比.

+心理语言学特征: 在基准模型的基础上加入人工提取的心理语言学特征.

+数据清洗 (传统阈值): 添加基于置信学习的数据清洗模块, 该方法使用传统的阈值计算方式, 即每个类别预测概率的均值作为阈值, 但未充分考虑类失衡问题.

+数据清洗 (特征向量): 添加基于置信学习的数据清洗模块, 其中使用交叉验证提取的图特征通过逻辑回归分类器单独计算样本外预测概率.

+数据清洗: 在基准模型中, 通过交叉验证提取样本外预测概率, 并采用动态阈值进行基于置信学习的数据清洗. 清洗后的数据再次用于模型的微调与训练, 得到最终模型.

+心理语言学特征+数据清洗: 将手工提取的心理语言学特征与本文提出的数据清洗模块相结合, 构成最终模型.

消融实验的结果如表 7、表 8 所示. 实验结果表明, 无论是加入心理语言学特征, 还是进行基于置信学习的数据清洗, 基准模型的准确率和 $F1$ 值都有不同程度的提升. 基准模型在各人格维度的表现差异较大, 且整体平均性能不够理想. 加入心理语言学特征后, 虽然 E/I 维度的准确率略有下降, 但在其他 3 个维度以及平均准确率上都有不同程度的提高, $F1$ 值也全面提高, 这表明心理语言学特征对文本语义信息有补充作用, 能够增强模型效果. 对于基于置信学习的数据清洗, 处理噪声样本的效果十分显著. 然而, 基于传统阈值的数据清洗由于未充分考虑类别分布不均衡问题, 清洗效果

有所下降, 尤其是在类别分布最不均匀的 N/S 维度上表现最为明显. 而基于特征向量的数据清洗效果也有所下降, 原因在于未使用训练后的分类器进行优化. 最终, 将数据清洗与心理语言学特征结合的模型取得了最佳的性能, 显著提升了模型的准确率和$F1$值, 进一步证明了数据清洗的必要性和心理语言学特征的加入对模型性能提升具有重要意义.

表 7 消融实验准确率结果 (%)

方法	E/I	N/S	F/T	J/P	均值
基准模型	79.42	83.40	80.12	68.01	77.74
+心理语言学特征	79.31	83.52	80.40	68.65	77.97
+数据清洗 (传统阈值)	84.74	91.37	88.53	79.50	86.04
+数据清洗 (特征向量)	86.04	91.92	91.69	82.59	88.06
+数据清洗	91.11	93.73	92.11	83.33	90.07
+心理语言学特征+数据清洗	**91.77**	**94.06**	**93.06**	**86.27**	**91.29**

表 8 消融实验 $F1$ 值结果 (%)

方法	E/I	N/S	F/T	J/P	均值
基准模型	67.95	63.37	79.90	66.69	69.48
+心理语言学特征	68.70	65.55	80.32	67.31	70.47
+数据清洗 (传统阈值)	70.15	61.34	88.52	77.03	74.26
+数据清洗 (特征向量)	77.33	74.96	91.65	80.60	81.13
+数据清洗	84.65	78.17	92.06	82.14	84.25
+心理语言学特征+数据清洗	**85.54**	**79.26**	**93.04**	**84.38**	**85.55**

3.7 跨模型泛化性实验

为了验证数据清洗实验的泛化性, 使用清洗后的数据在开源模型 TrigNet (https://github.com/TaoYang225/TrigNet) 上进行了训练与测试.

表 9、表 10 展示了采用原始数据和清洗后数据的实验结果.

表 9 泛化性实验准确率结果 (%)

数据集	E/I	N/S	F/T	J/P	均值
原始数据	77.43	86.37	78.07	68.06	77.48
清洗后数据	**90.70**	**93.40**	**92.18**	**80.05**	**89.08**

表 10 泛化性实验 $F1$ 值结果 (%)

数据集	E/I	N/S	F/T	J/P	均值
原始数据	69.54	67.17	79.06	67.69	70.86
清洗后数据	**81.52**	**74.44**	**92.09**	**78.09**	**81.53**

实验结果表明, 清洗后的数据同样显著改善了其他模型的性能, 证明了本文提出的数据清洗方法的广泛适应性和普适性. 在准确率方面, 使用清洗过后的数据, TrigNet 模型在人格的 4 个维度上均有不同程度的提高, 整体准确率均值提高了 11.6%. 在$F1$值方面, TrigNet 模型同样展现出较为明显的改善. 在 4 个维度的$F1$值上, 使用清洗后的数据是的模型在每个维度上均有不同程度的提高, 整体$F1$值均值提高了 10.67%. 这表明, 本文提出的基于置信度驱动的数据清洗方法并非针对特定模型, 而是具有跨模型的优化效果.

4 结论与展望

本文提出了一种置信学习驱动下融合心理语言学特征的人格检测模型, 该模型通过去除数据噪声, 并结合文本语义特征与心理语言学特征, 共同完成了人格检测任务. 实验结果表明, 清洗掉数据中的噪声后, 模型能够在更干净、更准确的数据上进行训练, 从而提升泛化能力和预测性能. 同时, 心理语言学特征的引入也帮助模型在人格分析中做出更准确的预测, 这些信息是传统文本语义特征难以全面捕捉到的. 总体而言, 本文为研究者提供了一个独特的视角, 不仅提出了新的有效特征组合, 也引入了数据清洗的思路, 进一步拓宽了人格检测领域的研究路径. 基于此, 未来的研究者可以进一步探索更多影响人格检测的关键因素, 推动该领域的深入发展.

然而, 本文的方法也存在一定的局限性. 模型仅基于社交媒体中的帖子文本来进行人格分析, 而诸如点赞、转发等其他互动信息也同样可能为人格检测提供有价值的线索. 置信学习结合交叉验证的方式在处理大规模数据集时可能会带来计算成本的增加, 如何提高清洗过程的效率仍值得进一步优化. 此外, 人格特质之间的相关性可能会有助于提升检测效果, 未来的研究可以尝试引入先进的多任务建模方法, 以进一步提高模型的准确性和表现.

参考文献

1 Phan LV, Rauthmann JF. Personality computing: New frontiers in personality assessment. Social and Personality Psychology Compass, 2021, 15(7): e12624. [doi: 10.1111/spc3.12624]

2 沈铁孙龙, 付晓东, 岳昆, 等. 融合人格特征的概率推荐模型. 计算机科学与探索, 2023, 17(1): 251–262.

3 Metha Y, Fatehi S, Kazameini A, *et al*. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM). Sorrento: IEEE, 2020. 1184–1189.

4 Zhu HC, Li LD, Jiang HY. Inferring personality traits from

user liked images via weakly supervised dual convolutional network. Proceedings of the 2018 Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and 1st Multi-modal Affective Computing of Large-scale Multimedia Data. Seoul: ACM, 2018. 63–69.

5 Zen G, Lepri E, Ricci E, *et al*. Space speaks: Towards socially and personality aware visual surveillance. Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis. Firenze: ACM, 2010. 37–42.

6 Li WY, Hu X, Long XF, *et al*. EEG responses to emotional videos can quantitatively predict big-five personality traits. Neurocomputing, 2020, 415: 368–381. [doi: 10.1016/j.neucom.2020.07.123]

7 Cui B, Qi C. Survey analysis of machine learning methods for natural language processing for MBTI personality type prediction [Technical Report]. Palo Alto: Stanford University, 2017. https://cs229.stanford.edu/proj2017/final-reports/5242471.pdf

8 费定舟, 赵雅婷. 社交媒体中的人格计算研究综述. 计算机工程与应用, 2019, 55(20): 34–42. [doi: 10.3778/j.issn.1002-8331.1904-0084]

9 Amirhosseini MH, Kazemian H. Machine learning approach to personality type prediction based on the Myers-Briggs type indicator®. Multimodal Technologies and Interaction, 2020, 4(1): 9. [doi: 10.3390/mti4010009]

10 Lynn V, Balasubramanian N, Schwartz HA. Hierarchical modeling for user personality prediction: The role of message-level attention. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5306–5316.

11 Ren ZC, Shen Q, Diao XL, *et al*. A sentiment-aware deep learning approach for personality detection from text. Information Processing & Management, 2021, 58(3): 102532.

12 Jeremy NH, Suhartono D. Automatic personality prediction from Indonesian user on Twitter using word embedding and neural networks. Procedia Computer Science, 2021, 179: 416–422. [doi: 10.1016/j.procs.2021.01.024]

13 Christian H, Suhartono D, Chowanda A, *et al*. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 2021, 8(1): 68. [doi: 10.1186/s40537-021-00459-1]

14 Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 2021, 70: 1373–1411. [doi: 10.1613/jair.1.12125]

15 Vinciarelli A, Mohammadi G. A survey of personality computing. IEEE Transactions on Affective Computing, 2014, 5(3): 273–291. [doi: 10.1109/TAFFC.2014.2330816]

16 林浩, 王春东, 孙永杰. 面向社交媒体数据的人格识别研究进展. 计算机科学与探索, 2023, 17(5): 1002–1016. [doi: 10.3778/j.issn.1673-9418.2212012]

17 Myers IB. Introduction to Type: A Description of the Theory and Applications of the Myers-Briggs Type Indicator. Palo Alto: Consulting Psychologists Press, 1987.

18 Lucas RE, Donnellan MB. Personality development across the life span: Longitudinal analyses with a national sample from Germany. Journal of Personality and Social Psychology, 2011, 101(4): 847–861. [doi: 10.1037/a0024298]

19 Stajner S, Yenikent S. A survey of automatic personality detection from texts. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: ACL, 2020. 6284–6295.

20 Mairesse F, Walker MA, Mehl MR, *et al*. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research, 2007, 30: 457–500. [doi: 10.1613/jair.2349]

21 Zhu YF, Hu LM, Ning NW, *et al*. A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection. Knowledge-based Systems, 2022, 249: 108952. [doi: 10.1016/j.knosys.2022.108952]

22 Hirsh JB, Peterson JB. Personality and language use in self-narratives. Journal of Research in Personality, 2009, 43(3): 524–527. [doi: 10.1016/j.jrp.2009.01.006]

23 Schwartz HA, Eichstaedt JC, Kern ML, *et al*. Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS One, 2013, 8(9): e73791. [doi: 10.1371/journal.pone.0073791]

24 Cornelisse J. Inferring neuroticism of Twitter users by utilizing their following interests. Proceedings of the 3rd Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. Barcelona: ACL, 2020. 1–10.

25 Kishima R, Matsumoto K, Yoshida M, *et al*. Construction of MBTI personality estimation model considering emotional information. Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation. Shanghai: ACL, 2021. 262–269.

26 Jiang H, Zhang XZ, Choi JD. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual

embeddings (student abstract). Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 13821–13822.

27 Wang Y, Zheng JZ, Li Q, *et al*. XLNet-Caps: Personality classification from textual posts. Electronics, 2021, 10(11): 1360. [doi: 10.3390/electronics10111360]

28 Kazameini A, Fatehi S, Mehta Y, *et al*. Personality trait detection using bagged SVM over BERT word embedding ensembles. Proceedings of the 4th Widening Natural Language Processing Workshop. Seattle: ACL, 2020. 1–4.

29 Liu SR, Guo ZQ, Li YH, *et al*. Inconsistent defect labels: Essence, causes, and influence. IEEE Transactions on Software Engineering, 2023, 49(2): 586–610. [doi: 10.1109/TSE.2022.3156787]

30 赵宏伟, 周明珠, 刘萍萍, 等. 基于置信学习和协同训练的医学图像分割方法. 吉林大学学报 (工学版): 1–8. https://link.cnki.net/urlid/22.1341.T.20231031.1518.002. (2023-10-31).

31 Wang Q, Li ZD, Liang HT, *et al*. Graph confident learning for software vulnerability detection. Engineering Applications of Artificial Intelligence, 2024, 133: 108296. [doi: 10.1016/j.engappai.2024.108296]

32 Yang T, Deng JH, Quan XJ, *et al*. Orders are unwanted: Dynamic deep graph convolutional network for personality detection. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 13896–13904.

33 Yang FF, Quan XJ, Yang YY, *et al*. Multi-document Transformer for personality detection. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 14221–14229.

34 Yang T, Yang FF, Ouyang HL, *et al*. Psycholinguistic tripartite graph network for personality detection. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). ACL, 2021. 4229–4239.

35 Lin, H, Wang CD, Hao QB. A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed gray wolf optimizer for feature selection. Information Processing & Management, 2023, 60(2): 103217.

36 黎迪. 基于情感知识增强的人格检测技术与研究 [硕士学位论文]. 北京: 北京邮电大学, 2023. [doi: 10.26969/d.cnki.gbydu.2023.002731]

(校对责编: 张重毅)