计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):74−83 [doi: 10.15888/j.cnki.csa.009554]

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

多任务学习在不良言论与个体特征检测中的应用①

肖博健[1], 曹霑懋[2], 许莉芬[2]

[1](华南师范大学 人工智能学院, 佛山 528225)
[2](华南师范大学 计算机学院, 广州 510631)
通信作者: 曹霑懋, E-mail: caozhanmao@scnu.edu.cn

摘　要: 多任务学习在自然语言处理领域有广泛应用, 但多任务模型往往对任务间的相关性比较敏感. 如果任务相关性较低或信息传递不合理, 可能会严重影响任务性能. 本文提出了一种新的共享-私有结构的多任务学习模型BB-MTL (BERT-BiLSTM multi-task learning model), 并借助元学习的思想为其设计了一种特殊的参数优化方式MLL-TM (meta-learning-like train methods). 进一步引入一个新的信息融合门 SoWLG (*Softmax* weighted linear gate), 用于选择性地融合每项任务的共享特征与私有特征. 实验验证所提出的多任务学习方法, 考虑到用户在网络上的行为与其个体特征密切相关, 文中结合了不良言论检测、人格检测和情绪检测任务进行了一系列实验. 实验结果表明, BB-MTL 能够有效学习相关任务中的特征信息, 在 3 项任务上的准确率分别达到了 81.56%、77.09% 和 70.82%.

关键词: 多任务学习; 信息融合; 不良言论检测; 人格检测; 情绪检测

引用格式: 肖博健,曹霑懋,许莉芬.多任务学习在不良言论与个体特征检测中的应用.计算机系统应用,2024,33(7):74–83. http://www.c-s-a.org.cn/1003-3254/9554.html

Application of Multi-task Learning in Hate-speech and Individual Characteristics Detection

XIAO Bo-Jian[1], CAO Zhan-Mao[2], XU Li-Fen[2]

[1](School of Artificial Intelligence, South China Normal University, Foshan 528225, China)
[2](School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Multi-task learning is widely used in the field of natural language processing, but multi-task models tend to be sensitive to the relevance between tasks. If the task relevance is low or the information transfer is unreasonable, the task performance may be seriously affected. This study proposes a new shared-private structure multi-task learning model, BERT-BiLSTM multi-task learning (BB-MTL). It designs a special parameter optimization method, meta-learning-like train methods (MLL-TM) for the model with the help of meta-learning ideas. Further, a new information fusion gate, *Softmax* weighted linear gate (SoWLG), is introduced for selectively fusing the shared and private features of each task. To validate the proposed multi-task learning method, a series of experiments are conducted by combining the tasks of hate-speech detection, personality detection, and emotion detection, taking into account the fact that user behavior on the Internet is closely related to individual characteristics. The experimental results show that BB-MTL can effectively learn feature information in relevant tasks, and the accuracy rates reach 81.56%, 77.09%, and 70.82% in the three tasks, respectively.

Key words: multi-task learning; information fusion; hate-speech detection; personality detection; emotion detection

相对稳定的思维、认知和行为模式的特征集合称之为个体特征, 通常体现在一个人的人格类型、情绪状态、偏好和欲望等方面[1]. 其中, 人格检测与情绪检测已经成为心理学与认知科学中的一项基本任务, 被

① 收稿时间: 2024-01-08; 修改时间: 2024-02-04; 采用时间: 2024-02-26; csa 在线出版时间: 2024-05-31
CNKI 网络首发时间: 2024-06-04

广泛应用于推荐系统、谣言检测、精神疾病诊断等领域[2]. 基于传统调查问卷的测验方法人力成本高, 采集大量样本数据难. 近年来, 随着社交媒体数据的海量增加, 基于深度学习的个体特征检测逐渐成为研究热点. 然而, 社交媒体在带来大量可用数据的同时, 也产生了许多负面影响, 如用户发表的不良言论, 浏览信息时产生的负面情绪等, 任由其发展可能会产生严重的社会问题.

有研究表明[3], 言论作为一种情感和认知表达的载体, 与个体的人格类型以及言论发表时的情绪状态等个体特征密切相关. 例如, 在网络上发布不良言论的用户可能潜在具有神经质、激进型人格, 且发表言论时通常处于愤怒、厌恶等负面情绪状态中; 而温和、保守型人格则很少会发布不良言论, 且通常会处于开心、惊喜等积极情绪状态中.

许多学者已经在不良言论检测领域和个体特征检测领域 (如人格检测和情绪检测) 打下了坚实的理论基础. 但他们通常专注于单任务研究, 从而忽略了一些相关任务中可能提升目标任务性能的潜在信息. 多任务学习可以通过任务间的知识共享, 提升模型的多维理解能力, 进而同时提升多个任务的性能[4]. 同时, 多任务学习也是解决小数据集局限的有效方法之一.

然而, 还没有研究将多任务学习方法同时应用于不良言论检测、人格检测以及情绪检测这 3 项任务中, 并探究它们之间的关系. 如果建立有效模型来探索 3 项任务相结合的学习, 预期会有效提高检测的性能. 针对上述待开展的热点, 本文做出了独特的探索.

本文的主要贡献如下.

(1) 提出了一个多任务学习模型 BB-MTL (BERT-BiLSTM multi-task learning), 给出了参数优化方法 MLL-TM (meta-learning-like train methods), 该模型能够在充分学习每个任务私有特征的同时, 整合来自其他任务的共享特征.

(2) 在现有信息融合机制的基础上, 提出了新的信息融合门方法 SoWLG (*Softmax* weighted linear gate), 用于提高不同特征融合的有效性.

(3) 将不良言论检测、人格检测以及情绪检测任务进行了多任务学习, 并在相关数据集上进行了基线实验和消融实验. 实验结果表明, 3 项任务之间存在一定的关联性, 且本文方法能够有效提升任务性能.

本文第 1 节介绍不良言论检测、人格检测、情绪检测以及多任务学习的相关工作. 第 2 节详细介绍本文提出的 BB-MTL 等方法. 第 3 节展示一系列的实验与分析, 说明本文方法的有效性; 最后, 进行总结与展望.

1 相关工作

在不良言论检测、人格检测和情绪检测领域, 主要从 3 个方面来设计检测模型.

基于手工特征提取的方法设计检测模型. 如王嘉伟等人提出一种基于遗传算法和粒子群算法的分类模型[5], 用于识别新浪微博平台上的不良言论信息; Choong 等人[6]利用文本关键词统计特征, 以及 LIWC 等心理语言学特征来检测人格; 王洁等人[7]利用情感词加权和词性分类等方法, 对中文音乐进行了作者情感分析. 手工特征提取的方法通常可解释性较高, 对计算资源需求小, 但过于依赖特征的选取, 且难以设计跨平台的通用特征以及处理不同类型的文本[8].

基于传统神经网络的方法设计检测模型. 如 Wang 等人[9]使用多种深度学习方法检测社交媒体平台上的不良政治言论; Maulidah 等人[10]利用不同的循环神经网络来构建人格分类模型, 并发现 LSTM 相较于传统的 RNN 和 GRU 表现更好; Widarmanti 等人[11]分别使用卷积神经网络、多项式朴素贝叶斯等方法对某款洗发水的评论进行情绪检测, 揭示了产品销量与产品评论情绪之间的关联性. 传统神经网络的方法通常在特定领域效果较好, 但往往对数据要求高、模型难以泛化, 不适用于小样本问题[8].

基于预训练语言模型的方法设计检测模型. 如闫尚义等人[12]结合 ALBERT 模型, 提出了一种融合字词特征的双通道分类模型, 用于检测互联网敏感言论; Mehta 等人[13]使用多种方法进行了人格检测实验, 并认为预训练语言模型提取的特征始终优于传统心理语言学特征; 丁美荣等人[14]通过构建预训练模型, 对酒店评论进行了有效的情绪检测. 预训练语言模型的方法可以有效提升任务性能, 增加模型的可迁移性, 但仍需较多的数据来进行微调或领域自适应的过程[15].

上述研究大多专注于单任务方法, 其特点表现为: 模型可以在特定任务上表现出卓越的性能, 但泛化能力却不佳, 对于样本数据的要求也较高. 针对此类局限, 多任务方法可以从一些相关任务中学习能够提升目标任务性能的潜在知识, 提升模型泛化能力, 同时减少对特定任务数据量的要求[4].

本文关注的 3 项任务中, 也有学者针对其中一项或两项任务, 或将其中某项任务与其他领域任务相结合, 以探索多任务方法. 例如, Li 等人[16]采用多任务学习的策略, 将人格检测与图像美学评估任务进行联合训练, 有效提升了两者的性能. Elourajini 等人[17]将人格检测任务中的回归和分类视为两个独立却相互关联的任务来进行多任务训练, 为未来的研究提供了新的研究思路. Nelatoori 等人[18]将不良言论检测和文本逻辑提取任务进行多任务学习, 相较于单任务模型, 这两个任务的准确率分别提升了 4% 和 2%. Plaza-Del-Arco 等人[19]将情绪状态识别作为辅助任务, 成功提升了不良言论检测任务的性能. Liu 等人[20]将情绪检测和认知投入检测任务进行联合训练, 证明了学生在学习过程中的情绪状态和认知投入, 均与学习成绩存在一定的关联性.

多任务学习方法在上述检测任务中的诸多研究, 对于任务性能的提升均是有效的, 但美中不足的是, 它们大多都基于传统的硬参数共享结构或软参数共享结构. 硬参数共享结构通过构建共享底层结构来服务各个任务, 可能会由于任务间的差异导致优化冲突; 软参数共享结构为每项任务都构建一个独立结构, 通过特定于任务的参数来共享信息, 但难以抽象出每项任务的共同特征. 与上述方法不同, 本文尝试采用一种新的共享-私有结构的多任务学习方法来综合两种共享结构的优势: BB-MTL 为每项任务都构建独立结构的同时, 再构建一个并行共享结构, 用于学习各个任务间的共同特征. 同时, 设计一个特殊的信息融合门, 用于私有特征与共享特征的融合.

2 BB-MTL 模型

本文涵盖的任务, 包括不良言论检测、人格检测和情绪检测, 本质上均可视为文本分类任务. 因此, 可以采用传统文本分类任务的表述方式来定义这些任务: 假设存在 K 个任务, 其中第 i 个任务的某个输入序列表示为 $S_i=\left\{s_i^1,s_i^2,\cdots,s_i^n\right\}$, 其中 $i\in\{1,2,\cdots,K\}$, n 表示为文本的长度. 对于第 i 个任务, 其目标是将输入文本正确分类为 $C_i=\left[c_i^1,c_i^2,\cdots,c_i^m\right]$, 其中 m 代表任务 i 中的类别数. 例如, 对于不良言论检测和情绪检测任务, 这些任务属于多分类问题, C_i 每个维度表示是否具有特定的标签属性. 而对于人格检测任务, 涉及多标签分类问题, 因此 m 取决于所采用的人格模型维度数, 每个维度对应一个人格属性.

2.1 模型框架

针对每项任务, 都有一个私有网络结构用于学习特定任务的私有特征. 同时, 共享网络结构用于学习与特定任务无关的共享特征, 并进行不同任务间的信息存储与传递. 为了有效融合这两种不同类型的特征, 进一步在私有网络和共享网络之间引入一个特征交流通道——信息融合门, 以促进信息的共享和协同学习.

BB-MTL 模型结构如图 1, 每项任务的输入会同时进入共享网络结构和各自的私有网络结构, 两种结构的详细描述在第 2.2 节和第 2.3 节给出. 特征信息通过信息融合门进行控制与传递, 以确保两个网络之间的特征共享和传递没有明显的冲突. 融合后的特征信息会被送入分类层, 最终得到分类结果.

2.2 任务私有网络结构

对于每项任务, 都会建立一个私有网络结构用于学习特定任务的私有特征. 为了充分利用每个单词的上下文信息, 采用双向长短期记忆网络 (BiLSTM) 作为私有网络的主要结构, 并使用 GloVe 词嵌入方法对任务输入进行嵌入操作. 任务私有网络结构设计如图 2.

任务的原始输入是一系列单词组成的序列, 通过 GloVe 嵌入层处理后, 每个词都会被表示为一个 300 维的嵌入向量 e_i^g. 因此, 每个输入序列都可以被表示为一组嵌入向量. 随后, 将这些嵌入向量送入 BiLSTM 层以学习语句序列的信息. BiLSTM 层通过正向和逆向操作对输入序列进行特征提取, 组合这些特征从而得到一个最终的隐藏状态集合, 作为从私有网络结构中学习到的隐藏特征. 整个计算过程如下:

$$E_n^{\mathrm{GloVe}}=\mathrm{GloVe}(S_i) \tag{1}$$

$$H_n=\mathrm{BiLSTM}\left(E_n^{\mathrm{GloVe}}\right) \tag{2}$$

其中, S_i 表示任务 i 的某个原始输入序列, GloVe 函数表示词嵌入的处理过程, 其输出为一组嵌入向量 $E_n^{\mathrm{GloVe}}=\left[e_1^g,e_2^g,\cdots,e_n^g\right]$, 其中 $e_n^g\in R^{300}$, n 表示单词数量. BiLSTM 函数则表示 BiLSTM 网络的处理过程, 其输出可以得到一个最终的隐藏特征状态集合 $H_n=[h_1,h_2,\cdots,h_n]$.

2.3 任务共享网络结构

除了为每项任务建立私有网络结构外, 还会针对所有任务建立一个共享网络结构, 用于学习与特定任务无关的共享特征, 并进行任务间的信息存储与传递.

鉴于共享网络主要用于学习与特定任务无关的特征, 采用基于预训练语言模型、具有良好泛化能力的 BERT 模型作为共享网络的主要结构.

对于每项任务的原始输入序列, 共享网络首先使用 BERT 模型的默认分词器 *Tokenizer* 对其进行标记, 然后将其转换为 BERT 嵌入向量, 这些嵌入信息随后会被输入进预训练好的 BERT 模型中学习序列特征信息. 使用[CLS]标记的特征向量作为共享网络的最终特征表示, 整个计算过程如下:

$$E_n^{\mathrm{BERT}} = Tokenizer(S_i) \tag{3}$$

$$C_n = \mathrm{BERT}\left(E_n^{\mathrm{BERT}}\right) \tag{4}$$

其中, S_i 表示每项任务的原始输入序列, *Tokenizer* 函数表示 BERT 的默认分词与嵌入过程, 其输出为一组嵌入向量 $E_n^{\mathrm{BERT}} = \left[e_1^b, e_2^b, \cdots, e_n^b\right]$, 其中 $e_n^b = R^{768}$, n 表示单词数量. BERT 函数则表示 BERT 模型的处理过程, 输出可以得到一个含有整个输入特征状态的[CLS]向量集合 $C_n = [c_1, c_2, \cdots, c_n]$.

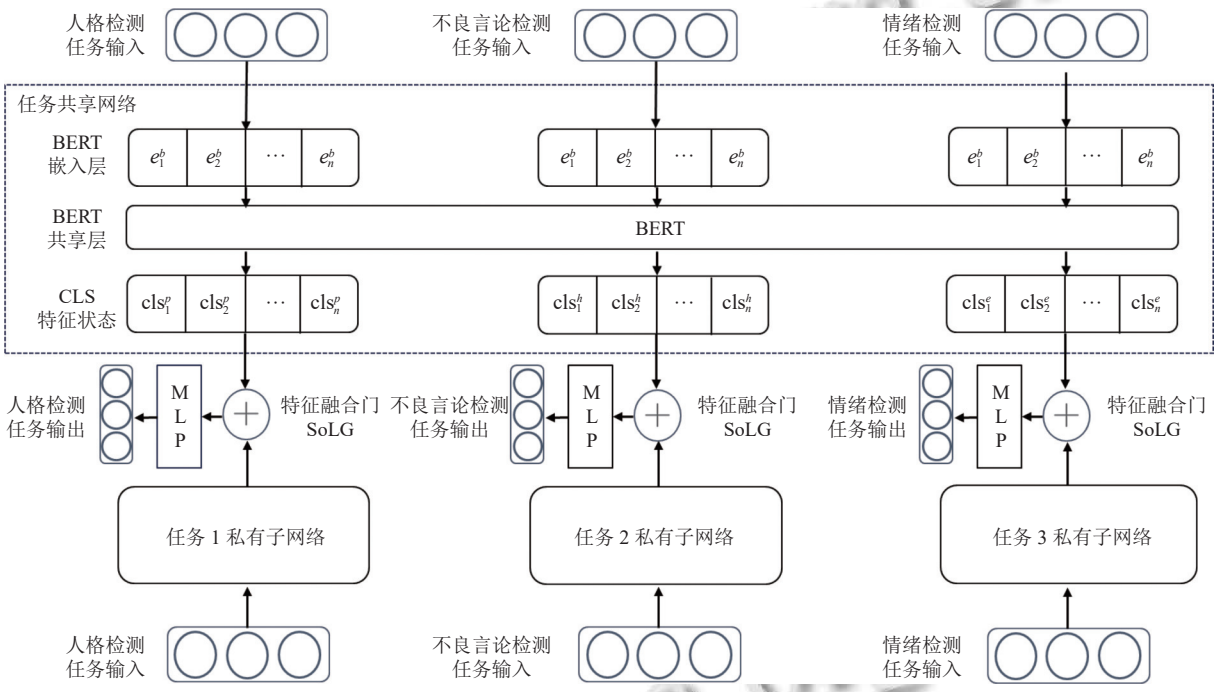

图 1 BB-MTL 模型结构

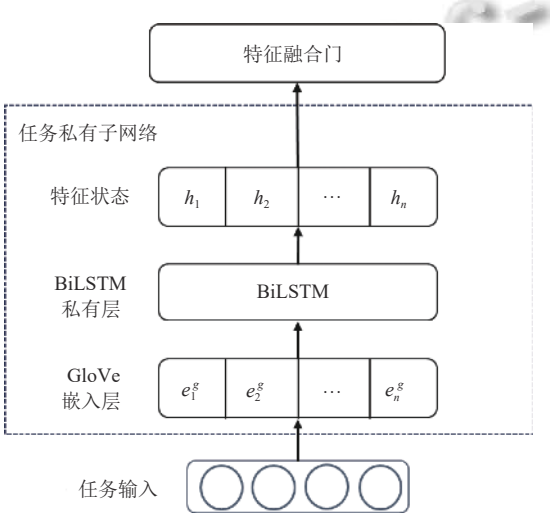

图 2 任务私有网络结构

2.4 信息融合门

在信息流经过私有结构和共享结构后, BB-MTL 已分别提取了每项任务的私有特征 H_n 以及所有任务的共享特征 C_n. 此时需要一个有效的信息共享机制, 将这些特征融合起来, 否则可能会严重干扰任务结果. 而门融合机制可以有效解决这个问题, 它并不直接为每个特征分配权重, 而是允许特征向量的每个维度对预测产生不同的贡献.

目前, 有多种方法可用于设计两个网络之间的信息融合门, 其中广泛使用的一种方法是由 Elfwing 等人[21]在强化学习领域提出的 SiLU (Sigmoid linear units), 该方法利用一个 Sigmoid 函数来计算多种信息状态的流动权重, 以实现信息融合的效果.

本文在 SiLU 的基础上, 将 Sigmoid 函数替换为 *Softmax* 函数, 并在另一个特征信息前引入一个权重因子 1–*Softmax*, 将其命名为 SoWLG (*Softmax* weighted linear gate), 其信息融合过程如图 3. 这些修改旨在更好地归一化两个网络的特征权重, 从而实现更好的特征选择效果, 也更适应于两个网络融合的情况. SoWLG 的具体计算方式如下:

$$m_i^p = Softmax\left(c_i^p\right)\cdot c_i^p \tag{5}$$

$$h_i = \left(1 - Softmax\left(c_i^p\right)\right)\cdot c_i^s + m_i^p \tag{6}$$

其中, c_i^p 与 c_i^s 分别表示私有网络与共享网络的第 i 个特征, m_i^p 表示成功通过了 SoWLG 的私有特征, 并即将与同样通过了门机制的共享特征进行融合. $Softmax()$ 及 1–$Softmax()$ 可视为一个注意力权重向量, 用于控制私有和共享特征中可以通过融合门的比例. h_i 则表示最终融合后所得的特征. 本文将在第 3.5 节详细讨论与其他几种信息融合门的对比, 以证明 SoWLG 方法的有效性.

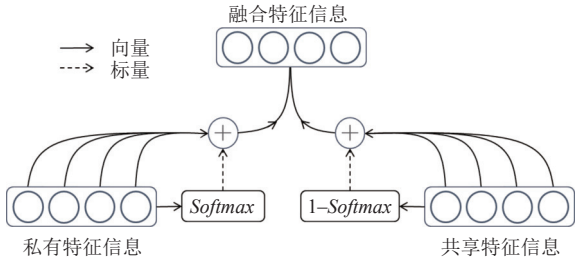

图 3 SoWLG 的信息融合过程

2.5 分类输出层

对于每项任务, 通过信息融合门后的联合特征向量将输入给对应的分类层. 该分类层由一个多层感知器 (multi-layer perception, MLP) 和一个 *Softmax* 函数组成. 在两个线性投影层之间采用非线性激活函数和权值归一化方法, 并引入 Dropout 方法避免出现过拟合现象. 分类层计算方式如下:

$$f = ReLU\left(W_f\cdot h_i + b_f\right) \tag{7}$$

$$p = Softmax\left(W_e\cdot f + b_e\right) \tag{8}$$

其中, W_f、W_e、b_f 以及 b_e 均是可以学习的权重和偏差. p 为最终的预测结果列表, 它的每个位置表示分类任务中每个类别的置信度, 可用于每个输入文本的分类.

2.6 基于元学习思想的参数优化方法

多任务学习中, 每个任务通常都拥有自己的训练目标. 对于不良言论检测和情绪检测任务, 通常每个句子仅涉及一个标签, 而每个标签则可能包含多个类别, 因此可视为多分类任务, 采用交叉熵损失 (cross-entropy loss) 作为目标函数, 计算方式为:

$$L_{\text{Hate}} = L_{\text{Emo}} = -\sum_{i=1}^{C} y_i\cdot\log\left(p_i\right) \tag{9}$$

其中, y_i 表示真实的类别标签, p_i 表示模型对类别的预测概率, C 表示类别的数量.

而对于人格特征的分析, 往往涉及一个句子对应多个互不排斥的人格标签的情况, 这使得人格检测成为一项较为复杂的多标签分类任务. 因此可以采用多标签分类任务中常用的二元交叉熵损失 (binary cross-entropy loss) 来定义人格分类任务的目标函数, 具体计算方式为:

$$L_{\text{Per}} = -\sum_{i=1}^{N}\sum_{j=1}^{L} y_{ij}\log\left(p_{ij}\right) + \left(1-y_{ij}\right)\log\left(1-p_{ij}\right) \tag{10}$$

其中, y_{ij} 表示第 i 个样本的第 j 个人格标签的真实值 (取值为 0 或 1), p_{ij} 则表示模型对第 i 个样本的第 j 个人格标签的预测值. N 表示样本的总数量, L 表示标签的总数量, 这里特指人格类别的数量 (例如 big five 人格理论中, L=5).

为了在训练过程中综合考虑多个任务的目标函数, 一种简单的方法是直接将每项任务的损失值求和. 但考虑到不同任务可能具有不同的收敛速度, 本文采用了一种加权平均算法来计算 L_{Multi}, 计算方式如式 (11). 它允许不同任务对模型参数更新的重要性有所不同, 缓解了不同任务收敛速度不一致的问题. 需要注意的是, 所有任务都以相同的次数进行训练, 以确保平衡训练的效果.

$$L_{\text{Multi}} = \sum_{i=1}^{K}\alpha_i L_i \tag{11}$$

其中, K 代表任务的数量, α_i 是可以学习的权重, 用于表示第 i 个任务对于整个多任务损失函数的重要性, 而 L_i 则表示第 i 个任务的损失值.

本文采用一种参数分离训练的策略: 模型的参数被划分为两个部分: 即共享网络的参数 bert_param 和非共享网络的参数 other_param. 由于本研究中不同任务的数据集来源各不相同, 为构建不同任务之间的关联, 本研究借助元学习[22]的思想, 引入了一种特殊的参数优化算法 MLL-TM, 该方法可以使得模型在一个任务上学习如何快速适应另一个任务, 从而实现不同来

源任务数据集的关联.

元学习是一种能够根据以往的学习经验来改进其未来学习和适应过程的方法, 旨在学会如何更有效地学习. 元学习不仅是学习任务的特定知识, 更重要的是学习“学习”的过程, 从而即使新任务的数据集来源不同, 也能够快速调整自己的参数. 具体来说, MLL-TM 借助了元学习中的 k-shot 思想, 在训练期间选择 k-batch-shot 方式来形成训练数据对. 即每一轮训练都从任务列表中随机选取一个任务进行一步或多步的梯度下降, 以优化该任务的损失函数, 然后模型的初始参数通过考虑所有任务上的性能改进来进行更新, 从而实现多任务间的交替学习. 这一策略使得模型能够随机学习来自多个异构任务的信息, 从而更有效地优化模型参数 (尤其是共享网络参数), 有助于构建更加鲁棒和全面的网络模型.

MLL-TM 第 1 步为随机初始化模型参数; 第 2 步则是整个循环训练过程, 每轮训练都涵盖了所有的任务, 其中子步骤 2.2 是随机抽取任务的训练过程, 在每一轮训练内部采用随机抽样的方式从任务列表中选择一个任务进行训练, 计算当前任务的损失值并累积到相应任务的损失列表中, 在每个任务的训练迭代次数达到预设最大值时, 将该任务从任务列表中移除. 该参数更新策略有助于确保每个任务都在适当的程度上得到训练, 同时也提高了整个训练过程的效率.

MLL-TM 的参数更新策略具体步骤如算法 1.

算法 1. MLL-TM 参数优化算法

1) 随机值初始化模型参数;
2) 对每轮训练进行参数更新;
2.1) 创建列表 Θ 和 Φ 用于记录每个任务的损失值和前任务已迭代次数;
2.2) 获取当前剩余的任务列表并随机选择一个任务;
2.3) 获取当前任务下一个 batch 的数据, 根据模型进行前向传播, 得到任务预测结果 p;
2.4) 计算当前任务损失值 L_{Single} 并将其累加到对应任务损失值列表 Θ 中;
2.5) 反向传播计算模型参数梯度, 并对其进行裁剪, 以防止梯度爆炸;
2.6) 采用 AdamW 算法同时更新两种网络参数, 同时更新学习率 r 并清除梯度;
2.7) 当前任务的迭代次数++;
2.8) 如果当前任务迭代次数达到最大值, 则从任务列表中移除该任务;
3) 计算总损失值 L_{Multi}.

3 实验分析

在本节中, 介绍了本文多任务模型的实验设置. 为了验证此模型的有效性, 在相同数据集下比较了其他的基线模型, 并设计了一系列消融实验验证了本文方法的合理性.

3.1 实验数据集

实验所使用的数据集包括:

(1) Kaggle 网站公开的基于迈尔斯-布里格斯类型指标的 MBTI 人格数据集: 其标签从 4 个人格特征角度进行排列组合, 即外倾与内倾 (E/I)、实感与直觉 (S/N)、思维与情感 (T/F)、判断与知觉 (J/P), 因此共可组合成 ENFJ 等 16 种人格类型以描述人格.

(2) 基于大五人格理论 (big five) 的 Essays 人格数据集[23]: 该理论从外向性 (EXT)、神经质 (NEU)、宜人性 (AGR)、尽责性 (CON) 和开放性 (OPN) 这 5 个维度来刻画用户的人格.

(3) Founta 等人提出的 Founta 不良言论数据集[24]: 包含 89 990 个带有不良言论标签的单标签句子, 标签包括正常 (normal)、辱骂 (abusive)、厌恶 (hate) 和垃圾邮件 (spam, 通常指广告或诈骗) 这 4 种.

(4) Scherer 等人提出的 ISERA 情绪数据集[25]: 包含 7 666 条带有愤怒 (anger)、厌恶 (disgust) 等情绪标签的单标签句子. 本文所使用数据集的详细信息如表 1.

表 1 数据集统计信息

数据集	数据量	标签分类
MBTI	8 675	{ENFJ, ENFP, ENTJ, ENTP, ESFJ, ESFP, ESTJ, ESTP, INFJ, INFP, INTJ, INTP, ISFJ, ISFP, ISTJ, ISTP}
Essays	2 467	{EXT, NEU, AGR, CON, OPN}
Founta	89 990	{Normal, abusive, hate, spam}
ISEAR	7 666	{anger, disgust, fear, joy, sadness, shame, guilt}

3.2 实验参数设置

实验采用五折交叉验证的方法, 将所有数据集按 4:1 的比例随机划分为训练数据和测试数据. 使用 PyTorch 深度学习框架进行编码. 除了共享网络结构的 BERT 模型直接使用预训练好的嵌入值外, 其他私有模型均使用 GloVe 方法进行嵌入操作, 且嵌入维度设置为 300. 当句子嵌入长度小于固定嵌入长度时, 使用“UNK”标记填充到固定长度. 采用 AdamW 优化器进行模型的参数优化, 其中 *bert_param* 前 100 个参数的初始学习率设置为 2E–5, 其他参数的初始学习率设置为 5E–5; *other_param* 的初始学习率均设置为 1E–3. 设置 batch size 为 32. BiLSTM 和 BERT 网络的 dropout 设置为 0.1, MLP 网络的 dropout 设置为 0.2. 为了评估模型效果, 采用准确度 (accuracy) 作为评估指标来衡量

本文方法的性能.

3.3 基线模型对比实验

为验证本文方法的有效性, 本文将 BB-MTL 与当前一些主流的基线模型进行了准确率对比实验. 具体来说, 将 MBTI 和 Essays 人格数据集分别与 Founta 言论数据集和 ISERA 情绪数据集组合, 然后将这些数据输入不同的基线模型中进行训练. 实验结果如表 2 和表 3 所示. 用于对比实验的基线模型如下介绍.

表 2 MBTI + FOUNTA + ISEAR 数据集上 BB-MTL 与不同网络结构的对比实验 (%)

方法	MBTI					FOUNTA	ISEAR
	E/I	S/N	T/F	J/P	AVG		
CNN	76.12	85.38	71.65	61.43	73.64	79.44	59.53
BiLSTM	77.44	86.85	73.03	62.99	75.07	80.31	55.26
BERT	78.38	86.76	74.68	**66.45**	76.59	79.89	59.34
SoGMTL-M	76.42	85.73	72.53	61.62	74.08	79.71	60.23
$MTL_{GatedDEncoder}$	78.17	**87.82**	73.66	63.56	75.80	80.34	59.78
MT-DNN	78.59	87.19	74.90	65.31	76.49	80.47	69.36
BB-MTL	**78.90**	87.74	**75.62**	66.08	**77.09**	**81.56**	**70.82**

表 3 Essays + FOUNTA + ISEAR 数据集上 BB-MTL 与不同网络结构的对比实验 (%)

方法	Essays						FOUNTA	ISEAR
	EXT	NEU	AGR	CON	OPN	AVG		
CNN	55.43	55.03	54.57	54.21	61.08	56.06	79.48	59.59
BiLSTM	58.72	58.26	56.94	58.15	63.86	59.19	80.35	55.24
BERT	60.03	60.52	58.82	59.29	64.68	60.67	79.85	59.31
SoGMTL-M	56.18	55.87	55.92	56.18	62.72	57.37	79.24	59.82
$MTL_{GatedDEncoder}$	59.84	59.63	58.88	**60.57**	64.62	60.71	81.79	59.93
MT-DNN	**60.31**	60.97	59.79	59.92	65.16	61.23	80.22	69.16
BB-MTL	59.42	**61.76**	**59.94**	60.41	**66.51**	**61.69**	**81.95**	**70.32**

(1) CNN[26]: 使用 CNN 模型进行单任务训练.

(2) BiLSTM[27]: 使用 BiLSTM 模型进行单任务训练.

(3) BERT[28]: 使用 BERT 预训练语言模型训练, 并在下游进行单任务微调.

(4) SoGMTL-M[29]: 基于 CNN 构建的一种层次信息共享的多任务学习方法, 该方法还采用了一种名为 MAML-like 的参数优化方法.

(5) $MTL_{GatedDEncoder}$[30]: 使用 BiLSTM 编码器构建的一种信息共享模式的多任务学习方法.

(6) MT-DNN[31]: 一种将 BERT 模型与多任务学习相结合以构建语言表征的多任务学习方法.

其中, CNN、BiLSTM 和 BERT 为单任务基线模型, 训练过程中会独立训练 3 项任务; 而 SoGMTL-M、$MTL_{GatedDEncoder}$ 以及 MT-DNN 为多任务基线模型, 训练过程中会同时训练 3 项任务. 选择部分单任务模型作为基线的主要原因, 在于它们在文本处理任务中的广泛应用和被验证过的有效性, 更重要的是, 它们还常常被用作构建更复杂多任务模型的基本组件. 对比单任务模型基线, 可以在一定程度上评估不同类型的基本组件对于多任务模型的适用性和效果增益, 直观地展示出多任务模型设计所带来的性能提升, 也为 BB-MTL 的组件选择提供了理论基础.

实验结果表明, 多任务方法整体上会优于单任务方法, 但性能提升并不大. 此外, 将单任务模型与其由相同组件构成的多任务模型进行比较 (例如 CNN 与 SoGMTL-M), 发现多任务模型在性能上总是会有所提升, 进一步验证了本研究中 3 个任务之间确实进行了有效的信息传递, 从而使得每项任务都得以受益.

在单任务模型方面, 可以观察到 CNN 在 ISEAR 任务上表现较为出色, 而在 Founta 任务上表现欠佳; 与此不同, BiLSTM 在 Founta 任务上表现较好, 在 ISEAR 任务上则稍显不足; 而 BERT 在各个任务中均保持了相对稳定的整体性能, 这一现象归因于预训练语言模型的强大泛化能力, 而对于 BiLSTM 这类非预训练语言模型则通常在某些特定任务中表现出色. 这一发现为 BB-MTL 选择模型组件提供了参考依据.

在多任务模型方面, BB-MTL 结合了 BiLSTM 和 BERT 两者的优势, 在几乎所有数据集上均取得了最佳的性能: 通过使用预先训练好的 BERT 作为共享网络, 可以充分发挥其泛化优势; 而将非预训练的 BiLSTM 用作私有层, 则有助于学习特定任务的特征, 从而提高每个任务的准确性.

3.4 多任务组合训练情景

为探讨多任务学习的有效性, 以及不同任务组合训练的效果, 本实验将人格检测任务视为主任务, 在 BB-MTL 模型中与其他不同任务组合进行训练, 组合方式如表 4 所示. 其中, Essays 任务与 MBTI 任务视为两个独立任务, 并将每个人格维度的平均准确度分数作为任务的性能衡量指标.

图 4 和图 5 分别为 MBTI 任务和 Essays 任务在不同任务组合情景下的实验结果. 可以发现, 随着组合训练任务数量的增加, 两项人格检测任务的性能会有所提升. 然而, 并不是所有情况下任务性能都会随着任务数量的增加而提高. 例如 Essays 任务与 Founta 任务组合训练后, 性能反而略有下降. 这些现象表明多任务学

习中任务选择的复杂性: 如果选择了相关度较低的任务进行多任务训练, 或设计了不当的信息融合机制, 很可能会引发优化冲突, 从而干扰任务训练效果.

表 4 多任务组合训练案例

数据集	多任务组合案例
Case1	MBTI
Case2	MBTI + Essays
Case3	MBTI + Founta
Case4	MBTI + ISEAR
Case5	MBTI + Essays + Founta
Case6	MBTI + Essays + ISEAR
Case7	MBTI + Founta + ISEAR
Case8	MBTI + Essays + Founta + ISEAR
Case9	Essays
Case10	Essays + MBTI
Case11	Essays + Founta
Case12	Essays + ISEAR
Case13	Essays + MBTI + Founta
Case14	Essays + MBTI + ISEAR
Case15	Essays + Founta + ISEAR
Case16	Essays + MBTI + Founta + ISEAR

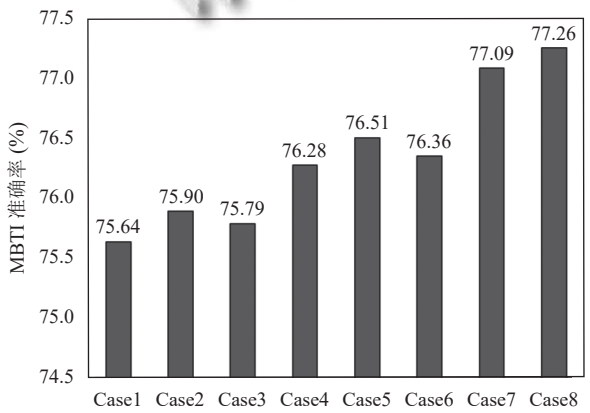

图 4 MBTI 与不同任务组合训练结果

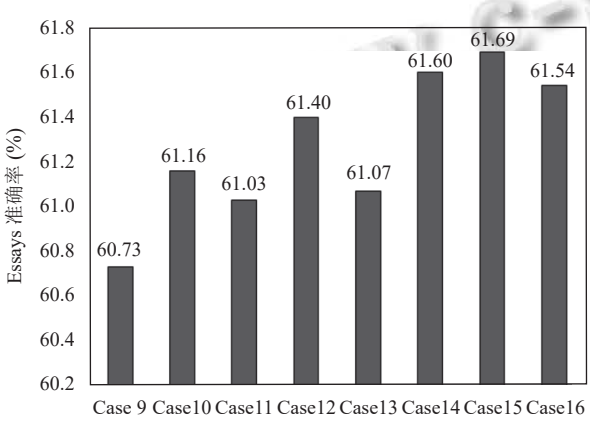

图 5 Essays 与不同任务组合训练结果

3.5 信息融合门对比实验

当前存在多种不同的方法用于设计两个网络之间的信息融合门. 为验证 SoWLG 在信息融合过程中的有效性, 本节将对几种不同的信息融合门进行对比实验, 并讨论这些机制各自的优缺点. 具体地, 在 BB-MTL 中采用不同的信息融合门进行实验, 并同时训练 4 个任务 (人格检测任务中 Essays 和 MBTI 视为两个独立任务). 对比的几种信息融合门如下.

SiWLG (Sigmoid-weighted linear gate): 将本文 SoWLG 中的 *Softmax* 函数替换为 Sigmoid 函数.

ReWLG (*ReLU*-weighted linear gate): 将 SoWLG 中的 *Softmax* 函数替换为 *ReLU* 函数.

TanWLG (tanh-weighted linear gate): 将 SoWLG 中的 *Softmax* 函数替换为 tanh 函数.

SoG (*Softmax* gate): 去除了 SoWLG 中的权重机制, 直接通过 *Softmax* 函数值将信息从一个网络传递到另一个网络, 计算过程如式 (12)、式 (13). 其优点在于它非常简单, 可以有效提升特征融合效率, 但它仅将任务的特征信息作为一个函数值传递, 可能会丢失大量有效的特征信息, 并引发优化冲突.

$$m_i^{t1} = Softmax\left(c_i^{t1}\right) \tag{12}$$

$$h_i^{t2} = c_i^{t2} + m_i^{t1} \tag{13}$$

SoLG (*Softmax* linear gate): 与 SoG 不同, SoLG 并不直接使用 *Softmax* 的函数值作为信息传递给另一个任务, 而是将其作为信息权重间接传递, 这样可以保留更多的原始特征信息, 计算过程如式 (14)、式 (15). 实际上, SoLG 是 SoWLG 中 c_i^{t2} 参数恒为 1 的特殊情况.

$$m_i^{t1} = Softmax\left(c_i^{t1}\right)c_i^{t1} \tag{14}$$

$$h_i^{t2} = c_i^{t2} + m_i^{t1} \tag{15}$$

不同信息融合门的对比实验结果如表 5 所示. 实验结果表明, BB-MTL 在几乎所有任务中, 使用 SoWLG 作为信息融合门的表现均为最佳. 具体来说, 通过与 SiWLG、ReWLG 以及 TanWLG 的实验结果进行对比, 可以观察到如果将 SoWLG 中的 *Softmax* 函数替换为 Sigmoid、*ReLU* 等其他激活函数, 模型性能均会下降, 这可能是因为 *Softmax* 函数可以更有效地将两个任务的特征权重归一化, 从而实现更好的特征选择效果; 而通过与 SoG 和 SoLG 实验结果的对比可知, 将 *Softmax* 函数值作为一个信息权重间接传递任务特征, 要比直接将函数值作为信息进行传递效果要好得多, 这也证明了 SoWLG 中权重信息的有效性.

表 5 SoWLG 与其他不同信息融合方法的对比实验 (%)

数据集 (任务)	MBTI	Essays	Founta	ISEAR
SiWLG	76.93	61.08	81.40	71.48
ReWLG	77.01	61.31	**81.72**	70.94
TanWLG	76.83	60.90	81.63	71.15
SoG	76.02	59.88	79.66	69.87
SoLG	76.87	60.92	81.36	71.35
SoWLG	**77.26**	**61.54**	81.56	**71.55**

3.6 参数优化方法的消融实验

在本文提出的 BB-MTL 模型中, 我们采用了一种名为 MLL-TM 的参数优化方法, 该方法兼顾了多任务学习中的知识共享和任务差异问题. 为验证该方法的有效性, 本文也进行了把 MLL-TM 应用于 SoGMTL-M, MTLGDEncoder 和 MT-DNN 模型的对比实验, 并与未使用 MLL-TM 方法情景下的模型进行了对比. 具体地, 我们仍然在每个模型上同时训练 4 个任务 (人格检测任务中 Essays 和 MBTI 任务视为两个独立任务), 并选取不良言论检测任务的准确率作为实验指标, 实验结果如图 6 所示.

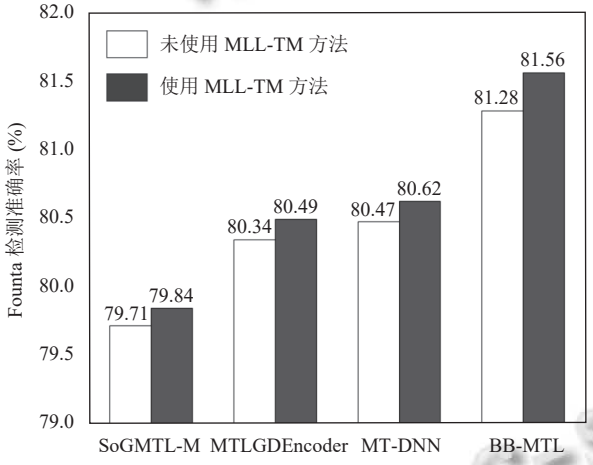

图 6 MLL-TM 应用在不同模型中的结果

实验结果可知, 相对于未使用 MLL-TM 参数优化方法, 使用 MLL-TM 方法后模型性能均有所提升, 其中 BB-MTL 模型的性能提升效果最为显著. 这一结果表明, MLL-TM 方法是一种有效的参数优化策略, 在提高多任务学习模型性能方面具有一定优势. 此外, MLL-TM 在非 BB-MTL 模型中的表现 (前 3 组实验), 也表明该方法可以被推广到其他的多任务学习模型中, 以进一步提高多任务模型的性能和鲁棒性.

4 结论与展望

本文提出了 BB-MTL 模型, 用于实现不良言论与个体特征的多任务检测. 不同于以往的多任务方法, BB-MTL 结合了硬参数共享和软参数共享结构的优势, 是一种新的共享-私有结构的多任务方法. 此外, 针对 BB-MTL 的特点, 提出了一种参数优化方法 MLL-TM, 以及一个信息融合门 SoWLG, 用于提升任务间信息传递的有效性. 在不同任务数据集上进行了一系列实验, 表明了本文方法的有效性.

本研究还有不足之处. 比如, 仅在 3 个任务上进行了实验, 仅选取了人格特征和情绪特征作为个体特征进行探索, 且不良言论和情绪检测任务中仅使用一个数据集, 这些都可能会导致模型泛化不足. 同时, 在实验中仍观察到了任务干扰的现象, 例如, 在多任务组合训练情景实验中, Essays 任务与 Founta 任务组合训练后, 性能反而有所下降, 表明了相关度较低任务可能存在的优化冲突问题. 未来, 将尝试把更多任务应用于 BB-MTL, 并使用不同的数据集进行验证, 以及探究最佳任务组合的方法.

参考文献

1 Mehta Y, Majumder N, Gelbukh A, *et al*. Recent trends in deep learning based personality detection. Artificial Intelligence Review, 2020, 53(4): 2313–2339. [doi: 10.1007/s10462-019-09770-z]

2 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]

3 Markov I, Ljubešić N, Fišer D, *et al*. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. ACL, 2021. 149–159.

4 Caruana R. Multitask learning. Machine Learning, 1997, 28(1): 41–75. [doi: 10.1023/A:1007379606734]

5 王嘉伟, 胡曦, 丁子怡, 等. 基于 GA-IPSO-BSVM 算法的新浪微博评论信息分类. 计算机系统应用, 2022, 31(8): 169–175. [doi: 10.15888/j.cnki.csa.008602]

6 Choong EJ, Varathan KD. Predicting judging-perceiving of Myers-Briggs type indicator (MBTI) in online social forum. PeerJ, 2021, 9: 11382. [doi: 10.7717/peerj.11382]

7 王洁, 朱贝贝. 面向中文歌词的音乐情感分类方法. 计算机系统应用, 2019, 28(8): 24–29. [doi: 10.15888/j.cnki.csa.006959]

8 Humeau-Heurtier A. Texture feature extraction methods: A survey. IEEE Access, 2019, 7: 8975–9000. [doi: 10.1109/ACCESS.2018.2890743]

9 Wang CC, Day MY, Wu CL. Political hate speech detection

and lexicon building: A study in Taiwan. IEEE Access, 2022, 10: 44337–44346. [doi: 10.1109/ACCESS.2022.3160712]

10 Maulidah M, Pardede HF. Prediction of Myers-Briggs type indicator personality using long short-term memory. Jurnal Elektronika dan Telekomunikasi, 2021, 21(2): 104–111. [doi: 10.14203/jet.v21.104-111]

11 Widarmanti T, Widodo MP, Ramadhani DP, *et al*. Text emotion detection: Discover the meaning behind YouTube comments using indo RoBERTa. Proceedings of the 2022 International Conference on Advanced Creative Networks and Intelligent Systems. Bandung: IEEE, 2022. 1–6.

12 闫尚义, 王靖亚, 朱少武, 等. 融合字词特征的互联网敏感言论识别研究. 计算机工程与应用, 2023, 59(13): 129–138. [doi: 10.3778/j.issn.1002-8331.2203-0301]

13 Mehta 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. Sorrento: IEEE, 2020. 1184–1189.

14 丁美荣, 冯伟森, 黄荣翔, 等. 基于预训练模型和基础词典扩展的酒店评论情感分析. 计算机系统应用, 2022, 31(11): 296–308. [doi: 10.15888/j.cnki.csa.008779]

15 Ding N, Qin YJ, Yang G, *et al*. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 2023, 5(3): 220–235. [doi: 10.1038/s42256-023-00626-4]

16 Li LD, Zhu HC, Zhao SC, *et al*. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. IEEE Transactions on Image Processing, 2020, 29: 3898–3910. [doi: 10.1109/TIP.2020.2968285]

17 Elourajini F, Aïmeur E. AWS-EP: A multi-task prediction approach for MBTI/Big5 personality tests. Proceedings of the 2022 IEEE International Conference on Data Mining Workshops. Orlando: IEEE, 2022. 1–8.

18 Nelatoori KB, Kommanti HB. Multi-task learning for toxic comment classification and rationale extraction. Journal of Intelligent Information Systems, 2023, 60(2): 495–519. [doi: 10.1007/s10844-022-00726-4]

19 Plaza-Del-Arco FM, Molina-González M, Ureña-López LA, *et al*. A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access, 2021, 9: 112478–112489. [doi: 10.1109/ACCESS.2021.3103697]

20 Liu S, Liu SQ, Liu Z, *et al*. Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. Computers & Education, 2022, 181: 104461.

21 Elfwing S, Uchibe E, Doya K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks, 2018, 107: 3–11. [doi: 10.1016/j.neunet.2017.12.012]

22 Hospedales T, Antoniou A, Micaelli P, *et al*. Meta-learning in neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5149–5169.

23 Pennebaker JW, King LA. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 1999, 77(6): 1296–1312. [doi: 10.1037/0022-3514.77.6.1296]

24 Founta A, Djouvas C, Chatzakou D, *et al*. Large scale crowdsourcing and characterization of Twitter abusive behavior. Proceedings of the 12th International AAAI Conference on Web and Social Media. Stanford: AAAI, 2018. 491–500.

25 Scherer KR, Wallbott HG. Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 1994, 66(2): 310–328. [doi: 10.1037/0022-3514.66.2.310]

26 Majumder N, Poria S, Gelbukh A, *et al*. Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, 2017, 32(2): 74–79. [doi: 10.1109/MIS.2017.23]

27 Zhou P, Shi W, Tian J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 207–212.

28 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.

29 Li Y, Kazemeini A, Mehta Y, *et al*. Multitask learning for emotion and personality traits detection. Neurocomputing, 2022, 493: 340–350. [doi: 10.1016/j.neucom.2022.04.049]

30 Rajamanickam S, Mishra P, Yannakoudakis H, *et al*. Joint modelling of emotion and abusive language detection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 4270–4279.

31 Liu XD, He PC, Chen WZ, *et al*. Multi-task deep neural networks for natural language understanding. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 4487–4496.

(校对责编: 张重毅)