文章编号：2096-1618(2025)03-0294-06

基于贝叶斯优化 BERT-BiLSTM 模型的攻击性语言识别与分类方法

刘雪明，杜之波
（成都信息工程大学网络空间安全学院/芯谷产业学院，四川 成都 610225）

摘要：当前基于 BERT 模型的攻击性语言的识别与分类方法中存在特征稀疏和上下文关联性少的问题，影响攻击性语言识别与分类的准确性，并且在参数优化方面存在人工优化费时费力、成本高、效果差等问题。为此，提出一种基于 BERT-BiLSTM 模型的攻击语言识别方法，并利用基于概率寻优的贝叶斯优化方法解决超参数优化问题。首先通过 BERT 模型训练攻击性语言数据集并提取数据集中的攻击性词特征，之后再使用 BiLSTM 模型捕获深层次的上下文关联性，最后将获得的特征向量输入到回归模型中进行分类。经过对 CLODataset 中文数据集的测试，并将 BERT 模型和 BiLSTM 模型进行对比实验，证明该方法有效地捕获序列特征和上下文信息，从而提升文本分类性能，使模型在测试集上的 F1 值提升了0.11。

关　键　词：BERT 模型；BiLSTM 模型；贝叶斯优化
中图分类号：TP309.2　　**文献标志码**：A
doi：10.16836/j.cnki.jcuit.2025.03.006

0　引言

随着微博、推特和知乎等社交媒体平台上攻击性语言的急速增加，社会亟需解决这一问题。攻击性语言[1]涵盖了对个人或群体进行各种形式的有针对性攻击，其中包括涉及种族、宗教、性别等方面的粗鲁、不尊重、侮辱、威胁和亵渎的言辞[2]，无论是含蓄还是直接的表达。手动过滤这类内容不仅耗时烦琐，还会给审查者带来类似创伤后应激障碍的心理压力。因此，研究者开始将这一过程自动化。然而，要实现自动化处理，首先需要将其视作一个分类问题[3]，并训练相应模型以便检测攻击性语言的存在。因此，研究高效准确的模型对攻击性语言的识别与分类具有重要意义。

近年来，国内外对攻击性语言检测进行了大量研究。文献[4]介绍了关于社交媒体中的攻击性语言识别和分类的 SemEval-2019 任务 6 的结果和主要发现。该任务在英文数据集上测试模型，使用 BERT 模型取得了最佳效果，F1 值达到82.9%。成为当年最受欢迎的任务之一。文献[5]利用迁移学习将英文数据集模型应用于多种其他语言，使用 BERT 模型取得准确率80.9%的最佳结果。文献[6]对比分析在机器翻译后的英文数据集上使用的 LSTM、BILSTM 和 BERT 模型，结果显示 BERT 模型的正确率为0.79，BILSTM 模型为0.68，LSTM 模型为0.78，并证明 BERT 模型在原始数据集和机器翻译数据集上检测结果变化不大。文献[7]介绍使用 SVM、KNN 和 LSTM 混合的神经网络模型对孟加拉语数据集进行检测，取得 80% 的准确率。文献[8]利用 BERT 模型结合机器学习中的 SVM 和逻辑回归对英文数据集进行检测，得到84.3%的准确率。文献[9]利用 BERT 模型结合针对阿拉伯语的 BERT 模型的改进版本进行阿拉伯语数据集检测，得到 90% 的准确率。文献[10]与[11]介绍 BERT 模型在文本分类中的应用并在 11 种语言测试效果最佳。文献[12]介绍 RNN 的改进版 LSTM 模型对于长距离的上下文依赖关系的提取更有优势。文献[13]介绍 BiLSTM 模型对于获取上下文关系更有优势，并且可以更好地执行文本分类任务。文献[14]中文攻击性语言数据集 CLODataset，并使用 BERT 模型进行检测分类，但准确率仍有待提高。

尽管对文本分类任务做了大量实验，但未实际有效提高攻击性语言分类的准确性并且都是使用人工参数寻优，无法确认实验的参数是否最优。因此，本文建立 BERT-BiLSTM 模型针对已标注好的攻击语言数据集进行分类，解决文本较短和因特殊攻击词带来的特征稀疏和上下文关联性少的问题。鉴于这些模型调参耗时费力，找到优化参数的方法至关重要。前期研究表明，贝叶斯超参数寻优方法在机器学习任务上已取得一定成功[16-18]。本文采用贝叶斯超参数寻优方法[15]来提升 BERT-BiLSTM 模型的分类性能。

收稿日期：2023-12-27
基金项目：四川省科技计划资助项目(2021ZYD0011)
通信作者：杜之波. E-mail：duzb@cuit.edu.cn

1　模型介绍

1.1　BERT 模型

BERT 模型是基于多头注意机制的掩码语言模型。在训练初期,它会随机地对输入进行掩码,然后利用攻击词的上下文信息来预测被掩盖的词,以更准确地理解单词的语义。好处是保留单词的语义信息,避免完全掩盖导致信息丧失的情况。因此,BERT 模型能够根据上下文中罕见词的出现情况进行预测,从而解决攻击词的复杂性问题。与传统的 word2vec 和 fastText 模型不同,BERT 可以根据特定的下游任务进行微调,以在特定任务上取得更优异的结果,使 BERT 在应用领域上更加灵活多样。BERT 模型结构如图 1 所示。其中,$E_1,E_2,\cdots,E_k$ 表示输入的数据,T_m 表示多层双向 Transformer 的编码层,$T_1,T_2,\cdots,T_k$ 表示输出的数据。

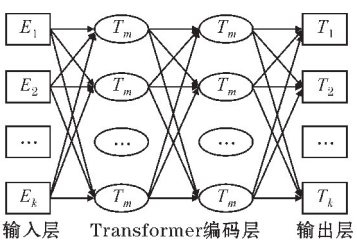

图 1　BERT 模型结构示意图

1.2　BiLSTM 模型

BiLSTM 模型由双向 LSTM 组成。LSTM 是对 RNN 进行改进的模型,解决了 RNN 中导数爆炸和导数损失的问题,并可以处理不同长度的序列数据,从而克服了循环神经元中信息快速丢失的问题。BiLSTM 是一种双向长短期记忆神经网络,通过捕获上下文依赖关系,解决了基于长短期记忆(LSTM)不能从后到前编码信息、只能单向性的问题,通过计算遗忘门和记忆门,分别选择被遗忘的信息和被记忆的信息。它们的输入都是前一时刻的隐式状态 h_{t-1} 和当前时刻的输入字 X_t,遗忘门的输出值为遗忘门的值 f_t,记忆门的输出值为记忆门的值和临时单元状态的值。遗忘门的计算公式如下:

$$f_t=\sigma(W_f\cdot[h_{t-1},X_t]+b_f)$$

式中,W 为权重参数,b 为偏差参数。

存储器门的计算公式如下:

$$f_i=\sigma(W_f\cdot[h_{t-1},X_t]+b_i)$$

$$\tilde{C}_t=\tanh(W_c\cdot[h_{t-1},X_t]+b_c)$$

然后,通过输入存储门它的值和遗忘门 f_t,临时单元状态 $\tilde{C}_t$ 和前一时刻单元状态 C_{t-1} 的值来计算单元的当前状态,输出的是当前时刻的单元格状态 C_t,其计算公式如下:

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tilde{C}_t$$

最后,计算当前矩的输出门和隐藏状态,输入值为前一矩的隐藏状态 h_{t-1},当前矩的输入字 X_t 和当前矩的单元状态 C_t,其计算公式如下:

$$o_t=\sigma(W_o\cdot[h_{t-1},X_t]+b_o)$$

$$h_t=o_t\cdot\tanh(C_t)$$

BiLSTM 模型如图 2 所示。其中,$X_1,X_2,\cdots,X_t$ 表示输入的数据,前向 LSTM 层按顺序读入数据,后向 LSTM 层按照反向顺序读入数据,得到前向隐层和后向隐层的特征量 $h_1,h_2,\cdots,h_t$。

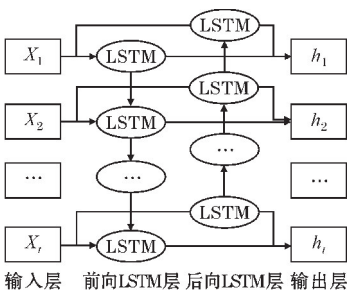

图 2　BiLSTM 模型结构示意图

2　贝叶斯优化

贝叶斯优化是一种用于优化目标函数的方法,适用于目标函数不能直接访问梯度信息或者代价较高的情况。与传统的优化方法相比,贝叶斯寻优使用概率替代模型来建模目标函数,从而在每次迭代中选择最有希望的下一个点进行评估,在合理减少评估次数的同时提高搜索效率。贝叶斯优化的框架包含:概率替代模型和采集函数。只有选择恰当的概率替代模型和采集函数,才能获得更加优越的优化效果。

2.1　概率替代模型

概率替代模型包括一个捕获目标函数概率的先验分布和一个描述数据生成机制的观察模型。在贝叶斯寻优中,使用高斯过程作为先验模型,高斯过程是一种用于建模随机函数的强大工具,也是对无限维度的高斯分布的一种推广,用于描述一个连续的随机函数。在机器学习和统计学领域,高斯过程被广泛应用于回归、分类、优化等问题。

高斯过程可以由两个部分完全描述:均值函数 $m(x)$ 和协方差函数(或核函数) $k(x,x')$。均值函数描述了在每个输入点上随机函数的期望值,而协方差函数描述了输入点之间的相关性。

具体来说,对于任意输入点 x 和 x',高斯过程的联合分布可以表示为

$$f(x)\sim N(m(x),k(x,x'))$$

其中,$m(x)$ 是均值函数,$k(x,x')$ 是协方差函数。

2.2 采集函数

在高斯过程中,期望改进即 EI 是一种常用的采样策略,用于在贝叶斯优化中选择下一个评估点。其基本思想是在当前模型的基础上,寻找使目标函数值比当前已知的最佳值更优的潜在点,并计算该潜在点的期望改进。具体步骤如下。

首先,定义目标函数,并计算最佳观测值与高斯过程模型均值之间的差异,表示为

$$Z=\frac{f(x)-\mu(x)}{\sigma(x)}$$

其中,$\mu(x)$ 表示高斯过程在点 x 的均值预测,$\sigma(x)$ 表示对应的标准差。

接着,利用 Z 值来计算标准正态分布的累积分布函数 CDF 和概率密度函数 PDF。这些函数可以通过标准的数学库来计算。得到计算 EI 的公式:

$$\mathrm{EI}(x)=(\mu(x)-f_{\mathrm{BEST}})\cdot\mathrm{CDF}(Z)+\sigma(x)\cdot\mathrm{PDF}(Z)$$

最后,选择使期望改进最大的点作为下一个评估点:

$$x_{\mathrm{next}}=\arg\max_x \mathrm{EI}(x)$$

这样得到下一个最有希望的评估点,可以在目标函数上进行评估。

通过使用期望改进作为采样策略,贝叶斯优化可以在每次迭代中选择具有最大改进期望的点,从而在相对较少的评估次数内找到近似最优解,使贝叶斯优化能有效解决复杂的模型优化和超参数寻优问题。

3 基于贝叶斯优化的 BERT-BiLSTM 模型

在攻击语言文本中,由于攻击词的特殊性带来特征稀疏和上下文联系少等问题,导致模型识别精度低。为解决这一问题,本文将 BERT 模型与 BiLSTM 模型进行结合检测攻击性语言。

首先基于中文 BERT 预训练模型对中文攻击性语言数据集进行训练,利用词粒度进行切片,词粒度的切片有助于生成更加连续的特征表示,更好地捕捉词之间的语义关系,降低特征的稀疏性。其次,利用 BiLSTM 模型输出的隐藏量 h_t 可以捕获攻击语言文本中的上下文信息。

3.1 BERT-BiLSTM 模型

BERT-BiLSTM 攻击语言识别分类模型架构如图 3 所示,网络模型共分为 4 层:BERT 层、BiLSTM 层、softmax 层和分类输出层。输入数据首先进入 BERT 模型从而得到词向量,将其输出值作为隐藏层的输入,计算后输出值作为 softmax 层输入值,输出样本数据分类至每种攻击语言分类的概率分布,最后由分类输出层输出样本对应的攻击语言分类。

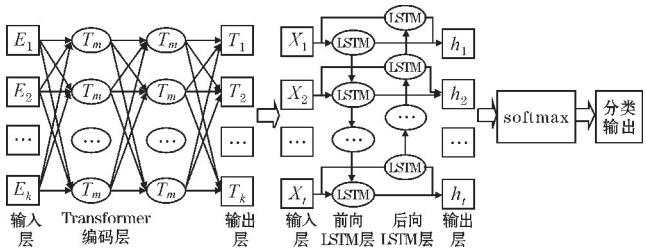

图 3 BERT-BiLSTM 模型结构

根据实际情况和 BERT-BiLSTM 模型结构,对 3 个关键超参数(初始学习率、小批量尺寸、权重)设定优化范围。而对于其他超参数,则根据调参经验进行设定。具体的参数配置如表 1 所示。

表 1 BERT-BiLSTM 模型参数设置

参数	参数含义	参数值或区间
InitalLearnRate	初始学习率	$[10^{-8},10^{-2}]$
MiniBatchSize	小批量尺寸	[4,64]
Weight	权重	$[10^{-8},10^{-2}]$

通过贝叶斯优化,对 BERT-BiLSTM 模型进行 20 次迭代,当达到计算次数后即停止优化过程,并输出最优的超参数组合。优化过程如图 4 所示。在第 20 次迭代时,模型性能达到最优水平,最佳超参数组合为:初始学习率为 1.6568991059843852e-05,小批量尺寸为 32,权重为 6.830127303922278e-05。这意味着在经过贝叶斯优化并选择不同的超参数组合后,BERT-BiLSTM 模型的整体识别准确率从 83% 提升至 94%。

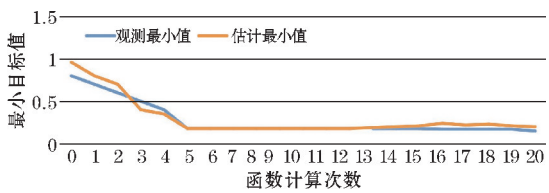

图 4 贝叶斯优化寻优过程

3.2 贝叶斯优化的超参数寻优

优化 BERT-BiLSTM 模型的超参数是一项复杂且重要的任务,因为它在很大程度上影响了算法的性能。手动调整超参数非常耗时,而网格搜索和随机搜索虽不需要人力,但也需要较长的运行时间。贝叶斯优化器迭代次数少,收敛速度快,因此采用贝叶斯优化器对 BERT-BiLSTM 模型的超参数进行优化。

本文利用高斯过程作为代理模型,建立超参数组合与 BERT-BiLSTM 模型之间的函数关系。通过基于观测数据集得到的后验分布,使用 EI 函数来选择下一个评估点,不断地修正先验信息,逐步提升代理模型的准确性,以寻找能使目标函数取得最优解的超参数组合。

本文聚焦于 BERT-BiLSTM 模型的超参数优化,考虑到其众多的超参数选项,从中重点选择 3 个关键参数进行优化,即权重、初始学习率和小批量尺寸。通过

对 BERT-BiLSTM 模型反向传播算法的了解,可以得其权重更新的公式:

$$W_{t+1}=W_t-\alpha\frac{1}{n}\frac{\partial L}{\partial W_{t+1}}$$

其中,α 是学习率,n 是小批量尺寸,W 是权重。权重、学习率和小批量尺寸是直接影响模型权重更新的关键因素,从优化角度来看,它们是影响性能收敛最为关键的参数。权重是用于控制信息流动和更新隐藏状态的重要权重参数之一,小批量尺寸决定了模型在每次参数更新时所利用的数据量,而学习率则控制了权重更新的步长。由于不同的数据和模型会对最佳学习率和小批量尺寸有不同的要求,因此在选择时并没有一套固定的准则可循。鉴于此,本文采用了贝叶斯优化的方法来进行参数选择。

3.3　基于贝叶斯优化的 BERT-BiLSTM 模型攻击语言检测

本文将贝叶斯超参数寻优应用于文本分类任务中,具体方法流程如图 5 所示。训练步骤可以总结为 7 个关键步骤:

(1)数据预处理:对原始数据进行处理,去除无效部分,如空文本或仅包含标点符号的文本,并将数据集划分为训练集、验证集和测试集。

(2)建立 BERT 模型:确定与 BERT 模型相关的超参数,并设置确定网络模型的超参数和优化范围。

(3)超参数组合选择:在当前组合下,将数据集输入 BERT 进行训练,得到初步的词向量,并计算当前超参数模型的函数值。

(4)建立 BiLSTM 模型。

(5)超参数组合选择:在当前组合下,将初步得到的词向量输入 BiLSTM 模型,并计算当前超参数模型的函数值。

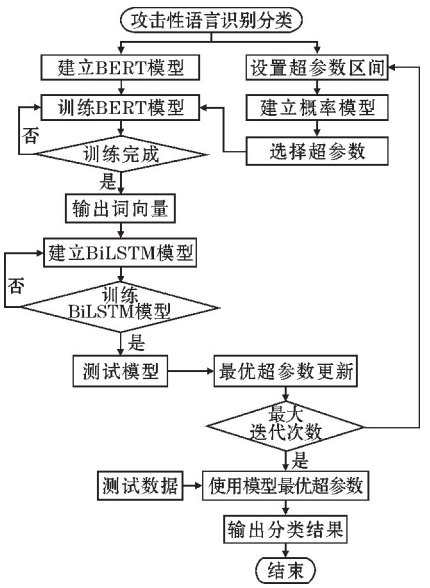

图 5　基于贝叶斯优化的 BERT-BiLSTM 模型攻击语言检测流程示意图

(6)输出优化后的最佳超参数和 BERT-BiLSTM 模型。

(7)最终分类结果:将测试集数据输入模型,获得最终的分类结果。

4　实验结果

4.1　数据集及实验环境

4.1.1　CLODataset 数据集

CLODataset 数据集是一个包含37 k个句子的中文数据集,涵盖关于种族、性别和区域偏见的主题,数据的收集过程符合 Vidgen 等[18]提供的标准。数据来源于社交媒体平台上发布的真实数据,并通过两种策略进行数据收集:关键字查询和从相关的子主题中爬取。将数据集去除重复的样本,确保数据的唯一性,之后去除对分析无关紧要的常见词汇。数据集包含有攻击性和非攻击性词语,具体的统计与分类如表 2 所示。将攻击性语言分为攻击个人和攻击组织,非攻击性语言则分为反偏见和其他非攻击性语言,细分数据如表 3 所示。

表 2　基本统计数据

数据集	攻击性	非攻击性	总数
训练集	15934	16223	32157
测试集	2107	3216	5323
总数	18041	19439	37480

表 3　细分数据

语言	宗教	性别	种族	总数
攻击个人	91	152	45	288
攻击组织	617	526	676	1819
反偏见	369	169	130	668
其他无攻击性	1010	704	834	2548
总数	2087	1551	1685	5323

4.1.2　实验环境

实验的设备配置如下:CUDA 11.7、Python 3.9 以及 RTX 3060。实验中,将模型构建为以 BERT 模型为上游数据处理模型,以 BiLSTM 为下游分类模型。数据集采用了 CLODataset 数据集。

为提高模型的稳定性并防止梯度爆炸等问题,本文采用 AdamW 优化器和交叉熵损失函数。AdamW 优化器结合了 Adam 优化器和 L2 权重衰减的方法,在优化过程中能够更有效地控制梯度更新,避免梯度爆炸问题。交叉熵损失函数是常用于分类任务的损失函数,能衡量模型预测值与真实标签之间的差异,从而指导模型参数的调整,使模型的输出更接近真实标签。

4.2　评价指标

分类模型的评价指标种类很多,本文采用准确率、召回率和 F1 值作为评价标准。

准确率是分类问题中最简单且常用的评价指标,

其计算公式如下：

$$\text{Precision}=\frac{n_c}{n_t}$$

其中，n_c 表示测试集中准确分类的样本数量，n_t 表示测试集中的样本总数。

召回率衡量了在所有阳性样本中被正确识别为真阳性的比例。召回率数值越高，表示模型可以识别出更多的真阳性样本。召回率的计算公式如下：

$$\text{Recall}=\frac{\text{TP}}{(\text{TP}+\text{FN})}$$

为更全面地评价该模型的性能，引入 F1 值。F1 值是准确率和召回率的调和平均值，能够综合考虑分类器的准确性和召回率。F1 值的计算公式如下：

$$\text{F1}=\frac{2\cdot(\text{Precision}\cdot\text{Recall})}{(\text{Precision}+\text{Recall})}$$

其中，TP、FP、TN、FN 分别表示真阳性率、假阳性率、真阴性率和假阴性率。

通过同时考虑准确率、召回率和 F1 值，能够全面评估该模型在分类任务中的表现。准确率反映模型整体的分类正确率，召回率衡量模型对于阳性样本的识别能力，而 F1 值则综合准确率和召回率，对模型的性能进行综合评估。这样的评价指标选择可以更好地了解模型的分类效果，并为模型的改进和优化提供有力的依据。

4.3 实验分析

4.3.1 基本模型对比

在 CLODataset 数据集上对 3 种不同的模型（BERT、BiLSTM 和 BERT-BiLSTM）进行测试，得到的结果如表4 所示。从表4 可以明显观察到，单独 BERT 模型的准确率为0.8，单独 BiLSTM 模型的准确率为 0.72，而结合后的 BERT-BiLSTM 模型表现最佳，准确率为0.89，其效果明显优于其他两种模型。

表4 基本模型对比表

模型	准确率	召回率	F1 值
BERT	0.8	0.82	0.81
BiLSTM	0.72	0.85	0.78
BERT-BiLSTM	0.89	0.79	0.83

结果表明，将 BERT 和 BiLSTM 两种模型结合，取得了更好的分类性能。BERT-BiLSTM 模型能够充分利用 BERT 模型对序列特征的学习以及 BiLSTM 模型对上下文信息的捕捉，从而在攻击性语言文本分类任务中取得了显著的优势。

BERT 模型与 BiLSTM 模型的优点结合，对于理解模型性能的优劣以及选择合适的模型在特定任务上进行应用都具有重要的指导意义。BERT-BiLSTM 模型的优异效果，进一步证明将不同的模型结合使用在自然语言处理任务中的有效性。

4.3.2 引入超参数寻优后的模型对比

加入贝叶斯超参数寻优后，为让模型达到更好的训练效果和更快的收敛，本文引入学习率调度器，其中参数设置为 step_size = 3，gamma = 0.1。在训练过程中自动调整学习率，使模型在适当的时候进行学习率的衰减，从而提高模型的稳定性和泛化能力。

从表5 可以看出，加入贝叶斯超参数寻优后的模型在测试集上表现明显优于之前的模型。参数寻优使模型在训练过程中选择更优的超参数组合，从而达到更好的性能。同时，学习率调度器的引入也使模型在训练过程中更加稳定，并且收敛速度更快。

表5 优化模型对比表

模型	准确率	召回率	F1 值
BERT	0.91	0.89	0.90
BiLSTM	0.87	0.90	0.89
BERT-BiLSTM	0.93	0.92	0.94

结果表明，贝叶斯超参数寻优和学习率调度器的引入，对于提升模型性能和训练效率起到了积极的作用，为本文的研究工作增添优势节省了时间成本。同时，这也为后续相关研究提供有益的借鉴和参考。

5 结束语

本文的主要工作是解决中文社交软件的中文攻击性语言分类问题。单独复现 BERT 模型和 BiLSTM 模型进行分类，通过分析 F1 值和准确率发现超参数的选择是导致准确率和 F1 值不高的原因。因此，根据社交软件中文本数据和检测分类模型的特点，提出基于贝叶斯优化的 BERT-BiLSTM 融合模型。使用 BERT 模型来提取文本的特征，并利用 BiLSTM 模型对 BERT 模型提取的特征进行分类，采用贝叶斯方法对 BERT-BiLSTM 模型的超参数进行优化。最后，通过攻击性语言的检测与分类问题的实验，证明贝叶斯优化超参数应用在自然语言处理模型中是可行的，同时也使本文提出的融合模型提高了分类的准确性。

本文的研究存在一些缺点。首先，使用的数据集是人工手动标注的，可能包含错误标记的数据，未来可以考虑采用新技术使用半自动标记的数据，以提高数据集的质量。其次，为进一步提升模型的鲁棒性，可以构建一个包含更细粒度分类、涵盖更多主题的更大数据集。这样将有助于更好地训练一个更健壮的中文攻击语言检测模型。

参考文献：

[1] Jahan M S，Oussalah M. A systematic review of Hate Speech automatic detection using Natural Language Processing[J]. Neurocomputing，2023：126232.

[2] Zampieri M，Malmasi S，Nakov P，et al. Predicting the

type and target of offensive posts in social media[J]. arXiv preprint arXiv:1902.09666,2019.
[3] Zou H,Tang X,Xie B,et al. Sentiment classification using machine learning techniques with syntax features[C]. 2015 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE,2015:175-179.
[4] Zampieri M,Malmasi S,Nakov P,et al. Semeval-2019 task 6:Identifying and categorizing offensive language in social media (offenseval)[J]. arXiv preprint arXiv:1903.08983,2019.
[5] Zhou L,Cabello L,Cao Y,et al. Cross-Cultural Transfer Learning for Chinese Offensive Language Detection[J]. arXiv preprint arXiv:2303.17927,2023.
[6] Zampieri M,Malmasi S,Nakov P,et al. Predicting the type and target of offensive posts in social media[J]. arXiv preprint arXiv:1902.09666,2019.
[7] Ahmed M F,Mahmud Z,Biash Z T,et al. Cyberbullying detection using deep neural network from social media comments in bangla language[J]. arXiv preprint arXiv:2106.04506,2021.
[8] Althobaiti M J. Bert-based approach to arabic hate speech and offensive language detection in twitter:Exploiting emojis and sentiment analysis[J]. International Journal of Advanced Computer Science and Applications,2022,13(5).
[9] El-Alami F,El Alaoui S O,Nahnahi N E. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model[J]. Journal of King Saud University-Computer and Information Sciences,2022,34(8):6048-6056.
[10] 方晓东,刘昌辉,王丽亚,等. 基于 BERT 的复合网络模型的中文文本分类[J]. 武汉工程大学学报,2020,42(6):5.
[11] 夏林中,叶剑锋,罗德安,等. 基于 BERT-BiLSTM 模型的短文本自动评分系统[J]. Journal of Shenzhen University Science & Engineering,2022,39(3).
[12] Ran X,Shan Z,Fang Y,et al. An LSTM-Based Method with Attention Mechanism for Travel Time Prediction[J]. Sensors,2019,19(4):861.
[13] Graves A,Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. NeuralNetworks,2005,18(5/6):602-610.
[14] Deng J,Zhou J,Sun H,et al. Cold:A benchmark for chinese offensive language detection[J]. arXiv preprint arXiv:2201.06025,2022.
[15] 崔佳旭,杨博. 贝叶斯优化方法和应用综述[J]. 软件学报,2018,29(10):3068-3090.
[16] Wang Z,Jegelka S,Kaelbling L P,et al. Focused model-learning and planning for non-Gaussian continuous state-action systems[C]. 2017 IEEE International conference on robotics and automation (ICRA). IEEE,2017:3754-3761.
[17] 杨欢,吴震,王燚,等. 侧信道多层感知器攻击中基于贝叶斯优化的超参数寻优[J]. 计算机应用与软件,2021,38(5):323-330.
[18] 魏佳恒,郭惠勇. 基于贝叶斯优化 BiLSTM 模型的输电塔损伤识别[J]. 振动与冲击,2023,42(1):238-248.
[19] Vidgen B,Derczynski L. Directions in abusive language training data,a systematic review:Garbage in,garbage out[J]. Plos one,2020,15(12):e0243300.

An Attack Language Recognition and Classification Method based on Bayesian Optimization BERT-BiLSTM Model

LIU Xueming,　DU Zhibo
(College of Cyber space Security Academy,Chengdu University of Information Technology,Chengdu 610225,China)

Abstract:The current recognition and classification methods for aggressive languages based on the BERT model suffer from sparse features and low contextual relevance,which affects the accuracy of aggressive language recognition and classification. In addition,there are problems such as time-consuming and laborious manual optimization,high cost,and poor performance in parameter optimization. A method for attack language recognition based on the BERT-BiLSTM model is proposed,and a Bayesian optimization method based on probability optimization is used to solve the hyperparameter optimization problem. Firstly,the aggressive language dataset is trained using the BERT model and the aggressive word features are extracted from the dataset. Then,the BiLSTM model is used to capture deep-level contextual correlations. Finally,the obtained feature vectors are input into the regression model for classification. After testing the CLODataset Chinese dataset and comparing the BERT model with the BiLSTM model,it was proven that this method effectively captures sequence features and contextual information,thereby improving text classification performance and increasing the F1 value of the model by 0.11 on the test set.

Keywords:BERT model;BiLSTM model;Bayesian optimization