

基于深层双向LSTM与MultiHeadAttention的敏感文本识别

朱定局, 邓任丰, 吴超

(华南师范大学人工智能学院, 广东 佛山 528000)

摘要: 随着互联网的发展和网民规模的扩大, 线上信息传播更加便捷, 但其中夹杂的敏感信息会危害网络社会和谐, 尤其是影响低龄网民的价值观。因此, 对敏感信息的自动化监测和识别非常重要。传统的自动化识别方法依赖于敏感词库, 有一定的局限性, 而深度学习技术的发展为敏感信息识别带来了新契机。基于LSTM模型, 参考BiLSTM模型逻辑, 结合注意力机制提出基于深层双向LSTM与MultiHeadAttention的敏感文本识别方法DBLSTM_MultiHeadAttention, 并以新浪微博内容为数据来源, 经主题模型筛选、回译法增强后, 得到33 452条高质量数据用于训练、验证和测试。实验结果显示, 该模型准确率达到90.79%, 精确率达到93.37%, 召回率达到94.21%, F1值为0.937 9, 综合性能优于其他模型。该研究为互联网信息监管提供了有力的技术支持。

关键词: LSTM; MultiHeadAttention; 敏感文本识别; 深度学习; 信息监管

DOI: 10.11907/rjdk.251081

扫描二维码阅读全文:



中图分类号: TP391.1

文献标识码: A

文章编号: 1672-7800(2026)004-0057-08

Sensitive Text Recognition Based on Deep Bidirectional LSTM and MultiHeadAttention

ZHU Dingju, DENG Renfeng, WU Chao

(School of Artificial Intelligence, South China Normal University, Foshan 528000, China)

Abstract: With the development of the Internet and the expansion of the scale of netizens, online information is easily disseminated, but the sensitive information mixed with it will harm the harmony of the network society, especially affect the values of younger netizens. Therefore, it is very important to automatically monitor and identify sensitive information. Based on the LSTM model, the BiLSTM model logic and attention mechanism, this paper proposes a sensitive text recognition method DBLSTM_MultiHeadAttention based on deep bidirectional LSTM and MultiHeadAttention, and takes Sina Weibo as the data source. After LDA model screening and translation enhancement, 33 452 high-quality data were obtained for training, verification and testing. The experimental results show that the accuracy rate of the model is 90.79%, the accuracy rate is 93.37%, the recall rate is 94.21%, the F1 value is 0.937 9, and the comprehensive performance is better than other models. This research provides strong technical support for Internet information supervision.

Key Words: LSTM; MultiHeadAttention; sensitive text recognition; deep learning; information supervision

0 引言

随着互联网的快速发展和普及, 网民群体的规模越来越庞大, 得益于手机、手提电脑、平板等移动设备以及微信、微博、抖音等软件社交平台的发展, 线上信息传播交流

变得更加简单、方便。互联网在给人们生活带来便利的同时, 也携带着危害网络社会和谐发展的内容, 如色情、暴恐、歧视等敏感信息^[1]。如果对这些敏感信息的监管不到位, 其会污染网络社交环境, 对社会生产生活产生较大危害^[2]。尤其是低龄网民, 心智尚未成熟, 容易受到网络信息的影响。因此, 对敏感信息的监测是十分必要的。

收稿日期: 2025-03-25

基金项目: 广东省特支计划项目(2022)

作者简介: 朱定局(1978-), 男, 博士, CCF杰出会员, 华南师范大学人工智能学院教授、博士生导师, 研究方向为人工智能、大数据; 邓任丰(1997-), 男, 华南师范大学人工智能学院硕士研究生, 研究方向为敏感文本识别; 吴超(2000-), 男, 华南师范大学人工智能学院硕士研究生, 研究方向为敏感文本识别。本文通讯作者: 朱定局。

随着网民基数的激增,以微信、微博、短视频为代表的社交媒体中产生的文本数据呈指数级增长。显然,如此庞大的数据无法以人工监测的方式进行处理。传统自动识别敏感信息的方法大多是基于敏感词库,以敏感关键词为模式串进行字符串匹配^[3]。该方法对数据库的依赖性较强,如果数据库不够完善,检测结果将会大打折扣,且该方法有一定的局限性。因此,需要一种能考虑上下文语义关系的自动化分类检测技术。

近年来,随着机器学习、深度学习等技术的发展,以及服务器硬件设备的升级,越来越多学者基于机器学习或深度学习方法,致力于互联网文本信息中敏感文本的识别和分类研究,助力于构建健康、文明、和谐的网络环境。

1 相关工作

传统的敏感信息自动化识别方法主要依赖于规则匹配,研究人员通过设定一系列规则或关键词来识别文本中的敏感信息。如秦艺文等^[4]基于敏感文本库,在获取文本敏感特征的基础上,引入类离散因子修正传统的 TF-IDF 权值确定方法,最终借助支持向量机构建分类器以识别和判断敏感文本内容;杜雪涛等^[5]提出考虑关键词上下文区间约束的策略匹配技术、跨消息关键词匹配技术和无监督分词技术,显著增强了对 5 G 消息中敏感文本的识别效果。

随着机器学习技术的发展,学者们又引入更复杂的算法,如支持向量机(Support Vector Machine, SVM)、朴素贝叶斯分类器、决策树等。这些方法通常基于手工设计的特征,如词频、TF-IDF、词性标注等,用于训练分类器来分类敏感文本,从而更好地捕捉文本中的特征,使敏感信息识别系统更准确。如李伟^[6]提出基于决策树的敏感文本分类器,其借助中文分词系统构建敏感文本向量空间模型,通过计算训练集与测试集文本的向量空间中词向量的 TF-IDF 值,得到相应的权重矩阵,最终利用决策树 C4.5 算法构建敏感文本分类器。安美芳^[7]提出基于上下文语义的敏感文本判定模型。该模型借助 word2vec 模型训练疑似敏感文本的句向量,然后利用 SVM 算法对敏感文本进行识别与判定。

随着深度学习技术的发展,特别是卷积神经网络(Convolutional Neural Network, CNN)和循环神经网络(RNN)的应用,敏感信息识别技术获得了进一步发展^[8]。这得益于深度学习能够从大规模数据中学习抽象特征,提高了系统对复杂语境和语义的理解能力。研究者使用卷积神经网络(CNN)、循环神经网络(RNN)、长短时记忆网络(Long Short-Term Memory, LSTM)、注意力机制与转换器(Transformer)等模型来捕捉文本的语义和上下文信息,以实现更准确的敏感文本分类。梁涛^[9]为了提升文本分类模型的效果,借鉴残差网络 ResNet 和 Inception-v3 网络结

构本身的优点,提出了 NRI(NLP ReNet Inception)网络模型。该模型相比单一的 CNN 网络和 RNN 网络,可以更好地提取文本数据的特征信息和语义信息;Xu 等^[10]提出用 Text-CNN 取代递归神经网络对敏感信息进行检测,从而提高了整体的效率和准确率。蔡元奇^[11]提出基于 BSMOTE(Borderline Synthetic Minority Over-sampling Technique)的敏感语义分类方法,其中设置了两级关卡,第一级为敏感关键词过滤器,第二级为文本分类器。先通过敏感关键词过滤器对文本数据进行敏感词过滤,然后使用基于 BSMOTE 算法的第二级文本分类器对文本数据进行分类。梁康^[12]为解决传统敏感信息检测算法监测含变形敏感词文本时存在的准确率下降问题,提出基于情感分析的敏感信息检测算法,一方面利用 LSTM 模型对文本进行情感分析,得到文本的正负情感倾向,另一方面利用 Text CNN 对文本进行敏感词检测,得到文本中含有的敏感词类别,最后将两个模型结果进行融合得到检测结果;Song 等^[13]通过定位和整合句子级证据进行文本色情识别,提出证据感知神经色情分类(eNPC)模型。该模型借助多实例学习模型定位文本中具有色情含义的句子,再聚合句子层面的证据,利用自注意力机制诱导文本标签。窦井苒^[14]针对长文本存在信息冗余的问题,提出一种基于 CNNBiLSTM-Attention 模型的长文本敏感信息检测框架。先利用 CNN 模型获取文本局部特征,再借助 BiLSTM 模型获取前后文语义特征,最后加入 Attention 机制,以提高模型整体检测效果;王莉娟^[15]针对敏感文本识别问题,提出基于多模块图注意力机制的短文本分类方法 MGAT-Transformer-CNN。消融实验证明,该方法可以取得较好的识别效果。胡创^[16]针对变体,尤其是同音字变体对敏感短文本识别的影响,提出基于汉字一拼音双通道 Text CNN(Text Convolutional Neural Networks)的敏感短文本识别方法和基于 Seq2Seq Attention(Sequence to Sequence Attention)的变体处理方法,使敏感短文本识别方法的识别效果得到一定提升。Gu 等^[17]针对敏感信息识别和分类问题提出一种自适应分类算法,通过对传统网络算法的聚类模型结构进行改造,引入分类信息,从而提供局部智能分类服务。张骏强等^[18]针对 Transformer 预训练语言模型参数量大、训练开销大以及推理时延长的问题,提出一种融合群助教模型的两阶段知识蒸馏文本分类方法。该方法以 F1 值的 0.73% 损失为代价,将模型参数量降低了 48.20%,推理速度提升了 56.94%。邓翠艳等^[19]针对现有电信诈骗信息识别模型中识别分类不充分、语义关联不全面、较少考虑诈骗语义信息及诈骗目的等情况,以 Transformer 架构为基础建立 encode-only 模型,并结合多任务学习模型提出一种融合 Transformer 及多任务学习的电信网络诈骗识别模型。

随着社交媒体的兴起,信息形式不再仅限于文本,还包括图像、音频和视频。因此,研究者开始探索多模态融合方法,使得敏感信息识别系统更全面和智能。如张洪

艳^[20]借助 Paddle Speech 模型对音频进行语音转文字识别,然后基于转换出的文本内容创建文本词条,再使用朴素贝叶斯分类算法对文本和音频进行敏感分类。文苗青^[21]提出一种基于多模态特征融合的网络舆情图文分类模型 Bert_VIT_Mo_GRU,主要采用 VIT 模型提取图像特征、BERT 模型的前 6 层提取文本特征,然后借助 GRU 模型对提取到图文特征进行分词级对齐和融合,再将结果与原文本特征向量结合,使信息再次融合,从而实现多模态网络舆情分类。

2 数据集与实验方法

2.1 数据集

由于敏感信息识别领域缺乏公开的数据集,学者们大多是自行爬取数据,构建数据集进行模型训练,但没有公开自己的数据集。考虑到新浪微博是一款极具影响力的社交媒体平台,用户规模庞大,内容生态极其丰富,涉及到个人生活分享、娱乐八卦、新闻资讯与热点事件等内容,其中包含了大量文本数据。因此,本文通过设置关键词的方式搜索新浪微博相关主题的微博和评论,爬取了敏感文本和非敏感文本数据,共计 25 726 条文本数据。

由于爬取的数据质量参差不齐,为进一步提高数据质量,本文采用基于主题模型的方法,在原有数据基础上筛选出高质量的数据,共计 16 726 条数据^[22]。此外,为进一步提升文本的多样性,增加数据数量,本文在筛选出的高质量数据基础上采用回译法增强数据^[23]。也即利用中英翻译间意译的特性,在丰富文本表述的同时增加文本数据量,最终得到 33 452 条文本数据。

2.2 数据预处理

数据预处理主要包括数据清洗、文本分词和文本向量化。数据清洗主要借助正则化方式,清理文本数据中无意义的字符、符号以及乱码,避免模型在训练过程中受到无

意义数据的干扰,影响数据特征的识别和提取。

文本分词采用结巴分词的精确模式,依据最大概率的组合方式将句子文本划分为词语,以确保语义的准确性,为文本向量化作准备。在文本向量化过程中使用北京师范大学中文信息处理研究所与中国人民大学 DBIR 实验室开源的预训练词向量模型 sgns.zhihu.bigram,取词汇表的前 20 万个词,每个词的词向量维度设置为 300,形成存储词向量的词向量矩阵。

2.3 实验方法

为更好地实现对敏感文本的识别,本文立足于 LSTM 模型,参考 BiLSTM 模型的逻辑原理,提出基于深层双向 LSTM 和 MultiHeadAttention 的敏感文本识别方法 DBLSTM_MultiHeadAttention^[24]。模型结构如图 1 所示,包括输入层(Input)、嵌入层(Embedding)、双向 LSTM 层、多头注意力机制(MultiHeadAttention)、拼接层(Concatenate)、LSTM 层、全连接层(Dense)、Dropout 层和输出层(Output)。

输入层接收一个形状为(batch_size, max_tokens)的输入张量,其中 batch_size 表示批大小, max_tokens 表示输入序列的最大长度,也即文本中的最大词数。实验的 batch_size 取值为 128, max_tokens 取值为 60, max_tokens 是采用正态分布对数据集计算得到的。经计算,该值取 60 时,能覆盖 95% 的数据。对于长度小于 60 的数据进行填充,对长度大于 60 的数据则进行剪切。

嵌入层(Embedding)将词汇索引转换为固定大小的向量表示。这里使用预训练的词向量矩阵作为初始化权重,并将该权重冻结,避免在训练过程中更新嵌入权重。实验的词向量矩阵取值为[200 000, 300]。

双向 LSTM 层包含 32 个 LSTM 单元,每个 LSTM 单元具有 32 个隐藏状态。实验中,每个时间步都会返回输出序列。因为是双向 LSTM,每个时间步的输出有 64 维的隐藏状态,所以输出序列的形状为(batch_size, max_tokens, 64),即每个时间步都有一个 64 维的表示。

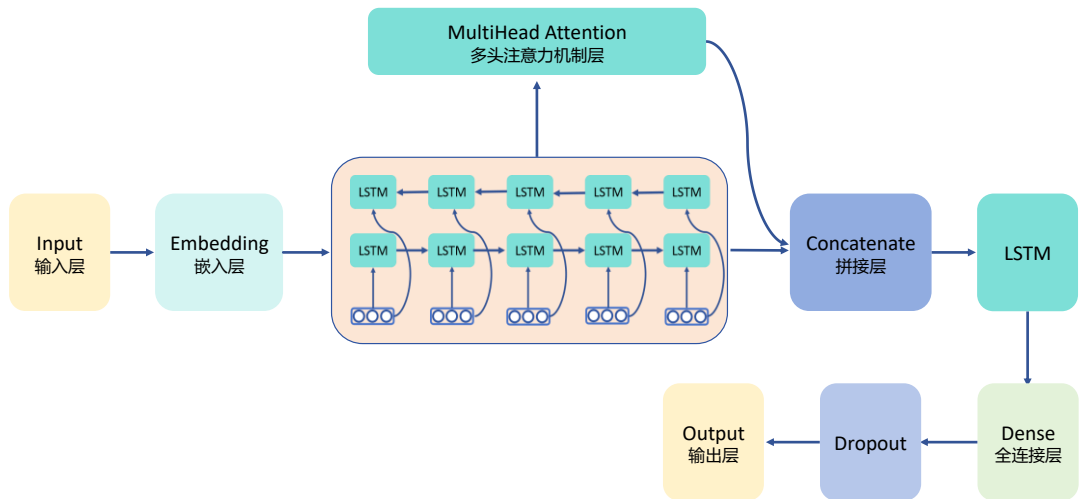


Fig. 1 DBLSTM and MultiHeadAttention model structure
图 1 深层双向 LSTM 与 MultiHeadAttention 模型结构

多头自注意力机制层(MultiHeadAttention)使用 8 个注意力头,每个注意力头的维度为 64。该层接收来自双向 LSTM 的输出作为查询(Q)、键(K)和值(V),并计算注意力权重来为每个时间步生成加权的上下文表示。注意力层的输出形状与输入相同。

拼接层(Concatenate)将双向 LSTM 层与注意力层的输出在最后一个特征维度进行拼接,拼接后的输出形状为(batch_size, max_tokens, 128)。因此,每个时间步有 128 维的表示,包含了 LSTM 的信息和注意力的上下文信息。

LSTM 层接收拼接后的输出,具有 16 个隐藏单元,表示该 LSTM 层只返回序列的最终隐藏状态,而不是整个序列的输出。输出形状为(batch_size, 16),即每个输入序列被映射到一个 16 维的向量表示。

全连接层将 LSTM 层的 16 维向量输出,映射到一个 64 维的向量空间。该层使用 relu 激活函数,并引入非线性,使模型可以学习更复杂的特征。

Dropout 层主要用于防止模型过拟合。该层随机地将

模型提取的某些特征丢弃,也即将向量中的某些元素置零,从而避免数据中噪声特征的影响。

输出层使用 Sigmoid 激活函数将 Dropout 层的输出映射到[0, 1]范围内,以表示概率值。通过对概率值设置阈值,从而判断文本是否为敏感文本。

LSTM 模型作为基础模型,在敏感文本信息识别过程中发挥着重要作用。其内部单元如图 2 所示。该模型含有一个被称为记忆单元的关键组件,通过一系列门来控制信息的流动,并具备了删除或添加信息的能力,从而有效地捕捉长期依赖关系。这些门分别是“遗忘门”、“输入门”和“输出门”,每个门由一个 Sigmoid 和一个点乘运算组成^[25]。

为捕捉到更全面的上下文信息和文本中的长期依赖关系,从而自适应不同长度的文本序列,同时更好地理解文本中的语境,本文借鉴 BiLSTM 正向和反向两个方向对序列建模,在 LSTM 的基础上采用深层双向构建模型,并引入多头自注意力机制,使得模型在敏感文本信息识别过程中更加有效。

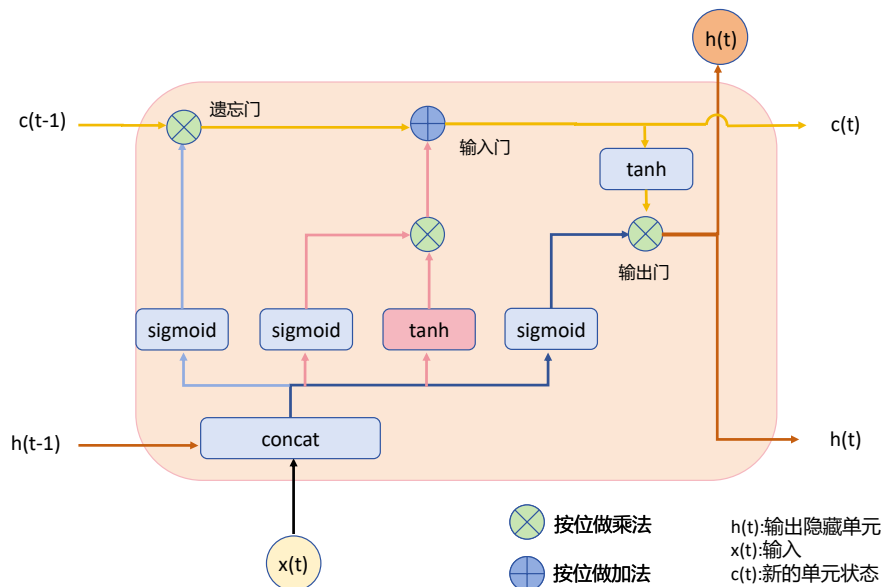


Fig. 2 LSTM internal unit

图 2 LSTM 内部单元

3 实验与结果分析

3.1 实验环境

实验在配备 Intel(R) Core(TM) i5-12400 2.50 GHz CPU、32GB RAM 和 RTX3060 12GB 显卡的计算机系统进行,以实现 DBLSTM_MultiHeadAttention 模型的训练和评估。实验过程中使用的编程语言为 Python 3.8,深度学习框架为 tensorflow,使用的 Python 库主要有 jieba0.42.1、gensim4.3.0、tensorflow2.11.0 等。

3.2 模型评估标准

为衡量模型对敏感文本的识别效果,并对结果进行展示,本文采用了以下评价指标^[26]:

(1) 损失值。损失值表示模型预测结果与真实标签之间的偏差。用于计算损失值的损失函数为二元交叉熵损失函数,公式如式(1)所示:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))) \quad (1)$$

其中, \$N\$ 表示样本总数, \$y_i\$ 表示样本的真实标签 0 或 1, \$p(y_i)\$ 表示输出为 \$y_i\$ 标签的概率。

(2) 准确率。准确率表示模型预测正确的样本数占总样本数的比例,公式如式(2)所示:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

其中, TP 表示被正确分类为敏感文本的样本数量, TN 表示被正确分类为非敏感文本的样本数量, FP 表示被错误分类为敏感文本的样本数量, FN 表示被错误分类为非

敏感文本的样本数量。

(3)精确率。精确率评估了模型对敏感文本预测的准确性,表示模型预测为敏感文本的样本中实际也属于敏感文本的比例,公式如式(3)所示:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

其中,TP表示被正确分类为敏感文本的样本数量,FP表示被错误分类为敏感文本的样本数量。

(4)召回率。召回率评估了模型对敏感文本的识别能力,即在所有实际为敏感文本的样本中模型正确预测为敏感文本的比例,公式如式(4)所示:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

其中,TP表示被正确分类为敏感文本的样本数量,FN表示被错误分类为非敏感文本的样本数量。

(5)F1值。F1值综合考虑了精确率和召回率。其是一个综合评价指标,是精确率和召回率的调和平均数,公式如式(5)所示:

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

3.3 实验参数

在模型训练、验证和测试过程中,本文将数据集的 25 726 条数据随机打乱,按 9:1 的比例划分。其中,训练集和验证集占 90%,测试集占 10%。之后,训练集和验证集的数据再按 9:1 的比例划分。模型训练时,设置的超参数如表 1 所示。

Table 1 Hyperparameters
表 1 超参数

超参数	取值
迭代次数(Epochs)	100
类别(Classes)	2
批次大小(Batch_size)	128
优化器(Optimizer)	Adam
学习率(Learning)	0.001

3.4 实验结果

3.4.1 对比实验

为评估 DBLSTM_MultiHeadAttention 模型的效果,本文训练了 LSTM、BiLSTM、TextCNN、GRU、DBLSTM_GRU、Transformer、GCN_Transformer、DBLSTM_Transformer、DBLSTM_AdditiveAttention 和 DBLSTM_HierarchicalAttention 模型,经过 100 次迭代,对比了不同模型和不同注意力机制在敏感文本识别中的效果。

不同模型在验证集中的准确率和损失值如图 3、图 4 所示。由图可知,DBLSTM_MultiHeadAttention 模型在验证集的 100 次迭代验证过程中,准确率整体呈上升趋势。前 7 次迭代上升迅速,之后趋于稳定。而损失值前期呈下降趋势,在第 6 次迭代时达到最低值,之后虽然上升,但保持在较低数值。整体而言,DBLSTM_MultiHeadAttention 模型在第 3 次迭代后,准确率明显高于其他模型,而其损失值

在达到最小值时,也明显小于其他模型的最小值。这主要得益于深层双向 LSTM 和 MultiHeadAttention 机制相互协作,使得模型能够并行地从不同角度捕捉文本中的语义信息,并且在面对变形敏感词和隐晦表达的复杂文本时,能够比较全面地挖掘上下文关系,从而识别敏感信息。其他模型在面对这些需要结合上下文多层语义理解的复杂文本时,则存在一定的局限性。

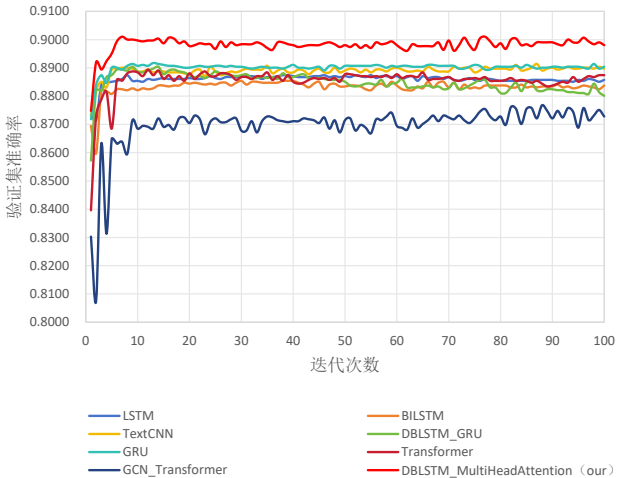


Fig. 3 Accuracy of different models in validation sets

图 3 不同模型在验证集的准确率

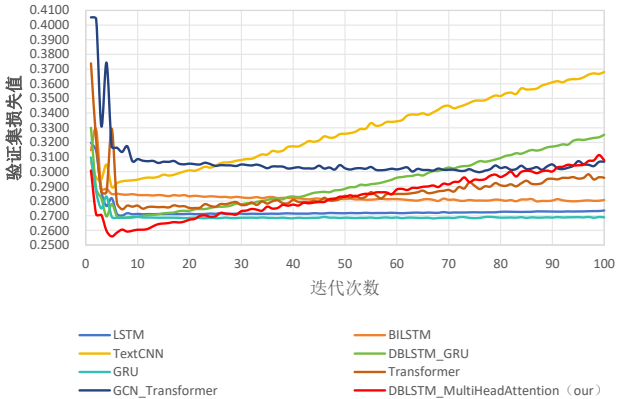


Fig. 4 The loss values of different models in validation sets

图 4 不同模型在验证集的损失值

LSTM、GRU 和 BiLSTM 由于顺序处理信息的特性,在面对包含变形敏感词、隐晦表达等复杂文本情况时,对长距离依赖信息的捕捉能力有一定局限性。而且 LSTM 和 GRU 只能单向捕捉序列信息,更加无法全面获取上下文语义。TextCNN 识别敏感文本的准确率依赖于卷积核匹配敏感信息特征的程度,而复杂文本结构复杂多变,容易遗漏关键信息。DBLSTM_GRU 在面对复杂文本时,GRU 的门控机制虽有助于处理短期依赖关系,但对于长期依赖关系的处理仍有局限性。Transformer 在敏感信息识别中,主要依靠位置编码来处理序列中的位置信息。然而,敏感信息灵活多变,可能出现在文本中的不同位置,但与其他特

定词汇或短语存在紧密关联,导致位置编码无法充分捕捉隐晦表达的敏感信息与周围文本的复杂关系。GCN_Transformer是通过将图卷积网络(GCN)与Transformer相结合来处理具有图结构和序列特征的数据,但图卷积主要关注节点之间的连接关系,敏感信息识别更多依赖于文本序列中的语义理解。因而在将图结构信息与文本序列信息融合时,未能有效整合两者的优势,导致在捕捉敏感信息的语义特征方面不如本文模型。因此,DBLSTM_MultiHeadAttention模型相比其他模型,在敏感文本识别任务中表现出更好的效果。

不同注意力机制在验证集中的准确率和损失值如图5、图6所示,DBLSTM_MultiHeadAttention组合模型在第5次迭代后,准确率明显高于DBLSTM和其他注意力机制的组合,而损失值明显低于DBLSTM和其他注意力机制的组合,说明本文提出的DBLSTM_MultiHeadAttention组合模型效果最佳。

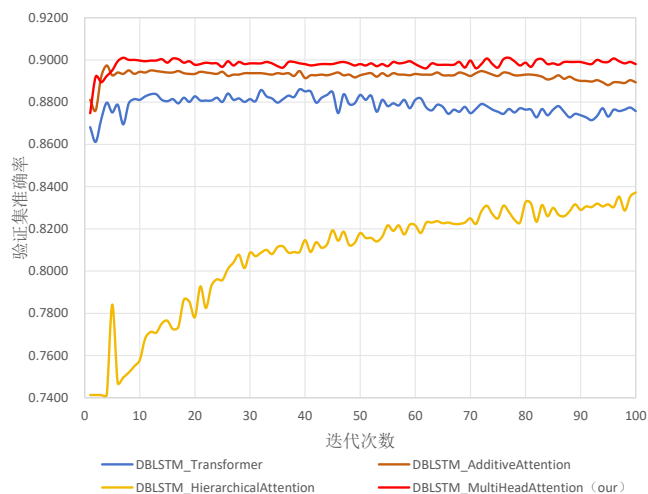


Fig. 5 Accuracy of different attention mechanisms in validation sets

图5 不同注意力机制在验证集的准确率

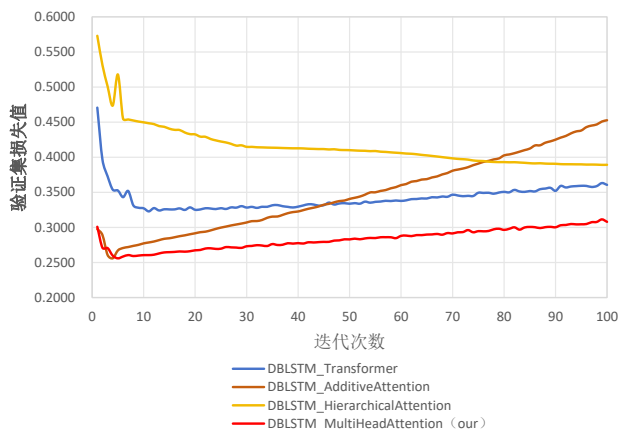


Fig. 6 The loss values of different attention mechanisms in validation sets

图6 不同注意力机制在验证集的损失值

在这些注意力机制中,Transformer具有较强的全局建模能力,但其单头注意力对信息的捕捉维度相对单一;AdditiveAttention是通过计算输入特征与可学习向量的点积来分配注意力权重,对不同位置信息的区分能力有限;HierarchicalAttention在面对大规模文本时,能逐步从局部信息过渡到全局信息,分层次地对文本进行注意力计算,但面对复杂多变的文本时,该层次结构则显得不够灵活。MultiHeadAttention是通过多个不同的线性投影将输入内容映射到多个子空间,并行地计算多个头的注意力,然后将这些头的结果拼接起来。该方式使模型能够同时关注文本的多个方面,更全面地捕捉文本中的语义依赖关系。因此,相较于其他注意力机制,其对于复杂多变内容的理解更深入,从而在敏感文本识别任务中表现出更优的性能。

为进一步评估模型在敏感文本识别任务中的效果,本文在测试集上对训练后的模型进行测试,结果如表2所示。由表2可知,DBLSTM_MultiHeadAttention模型的准确率为90.79%,精确率为93.37%,召回率为94.21%,F1值为0.9379。其中,准确率、精确率和F1值3项评估指标均优于LSTM、BiLSTM、TextCNN、GRU、DBLSTM_GRU、Transformer、GCN_Transformer、DBLSTM_Transformer、DBLSTM_AdditiveAttention和DBLSTM_HierarchicalAttention模型,而召回率仅略低于TextCNN模型,但也明显高于其他模型。该结果表明,DBLSTM_MultiHeadAttention模型在敏感文本识别任务中表现出了更好的性能、泛化能力以及稳定性,因而相比其他模型更适合该任务。

Table 2 Evaluation result of models

表2 模型评估结果

模型	准确率	精确率	召回率	F1值
LSTM	0.893 9	0.930 3	0.925 4	0.927 9
BiLSTM	0.894 8	0.930 8	0.926 3	0.928 5
TextCNN	0.902 6	0.926 0	0.943 3	0.934 6
GRU	0.893 3	0.926 8	0.928 7	0.927 7
DBLSTM_GRU	0.894 5	0.931 5	0.925	0.928 2
Transformer	0.897 2	0.931 4	0.929 1	0.930 2
GCN_Transformer	0.879 0	0.919 1	0.916 5	0.917 8
DBLSTM_Transformer	0.890 0	0.920 0	0.931 9	0.925 9
DBLSTM_AdditiveAttention	0.896 6	0.931 7	0.927 9	0.929 8
DBLSTM_HierarchicalAttention	0.837 4	0.857 4	0.935 2	0.894 6
DBLSTM_MultiHeadAttention(our)	0.907 9	0.933 7	0.942 1	0.937 9

为保证模型所使用的学习率和正则化方式为最优选择,本文以DBLSTM_MultiHeadAttention模型为基础,通过控制变量法对不同学习率和正则化方法进行了对比实验。结果分别如表3、表4所示。

由表3可知,在敏感文本识别任务中,与DBLSTM_MultiHeadAttention模型匹配的学习率不是越高越好,而是当学习率为 $1e-3$,也即0.001时,模型的学习效果最佳,准确率、精确率、召回率和F1值均达到峰值。正则化方法的作用主要在于防止过拟合,提高模型的泛化能力,改善模型的稳定性和减少训练时间,常用的正则化方法有L1、

Table 3 Comparison of different learning rate results

表 3 不同学习率结果对比

学习率	准确率	精确率	召回率	F1 值
1e-1	0.854 8	0.891 1	0.914 9	0.902 8
1e-2	0.854 8	0.891 1	0.914 9	0.902 8
1e-3(our)	0.907 9	0.933 7	0.942 1	0.937 9
1e-4	0.894 2	0.929 7	0.926 7	0.928 2
1e-5	0.890 0	0.918 3	0.934 0	0.926 1

Table 4 Comparison of different regularization method results

表 4 不同正则化方法结果对比

正则化方法	准确率	精确率	召回率	F1 值
L1	0.898 1	0.931 8	0.929 9	0.930 8
L2	0.903 5	0.936 2	0.932 7	0.934 4
L1_L2	0.904 7	0.934 5	0.936 4	0.935 4
Dropout(our)	0.907 9	0.933 7	0.942 1	0.937 9

L2、L1_L2 和 dropout。其中,L1、L2 和 L1_L2 正则化方法是通过控制权重大小实现的,Dropout 则是通过随机丢弃神经元的方式实现的。由表 4 可知,DBLSTM_MultiHeadAttention 模型结合 Dropout 方法除精确率略低外,准确率、召回率和 F1 值均高于其他正则化方法。

3.4.2 消融实验

为进一步验证 DBLSTM_MultiHeadAttention 模型中各模块的有效性,本文分别对模型的 MultiHeadAttention 模块和 LSTM 模块进行切割,得到深层双向 LSTM 模型(Deep Bidirectional Long Short-Term Memory, DBLSTM)和 BiLSTM_MultiHeadAttention 模型,并进行消融实验。结果如表 5 所示。

Table 5 Results of ablation experiments

表 5 消融实验结果

模型	准确率	精确率	召回率	F1 值
DBLSTM	0.899 0	0.937 6	0.924 6	0.931 0
BiLSTM_MultiHeadAttention	0.902 6	0.934 7	0.933 1	0.933 9
DBLSTM_MultiHeadAttention(our)	0.907 9	0.933 7	0.942 1	0.937 9

由表 5 可知,MultiHeadAttention 模块的切割导致模型缺少了多头注意力计算过程,削弱了模型捕捉文本不同特征的能力,造成模型的准确率、召回率和 F1 值明显下降。而 LSTM 模块的切割削弱了模型捕捉文本序列中长期依赖关系的能力。由于数据集的文本长度不长,绝大部分数据长度在 60 字符以内,所以模型的准确率和 F1 值仅轻微下降,但召回率下降明显。由此可见,各模块对本文模型的识别效果均发挥了一定作用。

4 结语

在互联网时代,信息传播为人们的生活带来了极大的便利。与此同时,危害社会和谐发展的敏感信息也充斥其中。若对这类敏感信息监管不力,将会给社会带来严重危害,特别是对心智尚未成熟的低龄网民,其人生观、世界观和价值观极易受到冲击。因此,对敏感信息进行有效监测显得尤为重要。

传统的敏感信息识别方法存在一定的局限性。鉴于此,本文提出一种基于深层双向 LSTM 和 MultiHeadAttention 的敏感文本识别方法 DBLSTM_MultiHeadAttention。实验结果表明,该模型在准确率、精确率、召回率和 F1 值等指标上表现优异,整体性能优于其他模型,在敏感文本信息识别方面具有显著优势,为互联网信息监管提供了强有力的技术支持。未来将致力于优化模型结构,以进一步提高模型的准确率和整体性能。

参考文献:

[1] GUAN X Y, ZHAO M, WU W C. A short text information filtering algorithm based on DFA [J]. Software Guide, 2023, 22(4): 103-108.
关兴义,赵敏,伍文昌. 一种基于 DFA 的短文本信息过滤算法[J]. 软件导刊, 2023, 22(4): 103-108.

[2] WANG L J. Sensitive text classification based on multi-module graph attention mechanism [D]. Xi'an: Xidian University, 2022.
王莉娟. 基于多模块图注意力机制的敏感文本分类[D]. 西安: 西安电子科技大学, 2022.

[3] HUANG C, ZHAO Q R. Sensitive information detection method based on language model word embedding and attention mechanism [J]. Journal of Computer Applications, 2012, 42(7): 2009-2014.
黄诚,赵倩锐. 基于语言模型词嵌入和注意力机制的敏感信息检测方法[J]. 计算机应用, 2022, 42(7): 2009-2014.

[4] QIN Y W, YANG Y. Research on document sensitive content recognition algorithm based on inter-class discreteness [C]// Shenyang: Proceedings of the 10th Annual Conference of China Communications Society, 2014: 229-234.
秦艺文,杨榆. 基于类间离散度的文档敏感内容识别算法研究[C]//沈阳: 第十届中国通信学会学术年会, 2014: 229-234.

[5] DU X T, ZHANG C, DU G, et al. Research on bad text message recognition in 5G messages [C]// Beijing: Proceedings of the 5G Network Innovation Symposium (2021), 2021: 45-49.
杜雪涛,张晨,杜刚,等. 5G 消息中不良文本消息识别技术研究[C]// 北京: 5G 网络创新研讨会(2021), 2021: 45-49.

[6] LI W. Design of web sensitive words filtering system based on decision tree [D]. Xi'an: Northwest A&F University, 2019.
李伟. 基于决策树的网页敏感词过滤系统设计[D]. 西安: 西北农林科技大学, 2019.

[7] AN M F. Research and implementation of sensitive content recognition mechanism for big data security [D]. Beijing: Beijing Jiaotong University, 2022.
安美芳. 一种面向大数据安全的敏感内容识别机制研究与实现[D]. 北京: 北京交通大学, 2022.

[8] HUANG Y, DAI X H, HUANG W J, et al. Text classification based on A-BiLSTM and CNN [J]. Computer Engineering and Design, 2024, 45(5): 1428-1434.
黄远,戴晓红,黄伟建,等. 基于 A-BiLSTM 和 CNN 的文本分类[J]. 计算机工程与设计, 2024, 45(5): 1428-1434.

[9] LIANG T. Multi-object sensitive text detection based on unbalanced data [D]. Chengdu: University of Electronic Science and Technology of China, 2020.
梁涛. 基于非平衡数据的多目标敏感文本检测[D]. 成都: 电子科技大学, 2020.

[10] XU G S, QI C H, YU H, et al. Detecting sensitive information of unstructured text using convolutional neural network [C]//2019 International

- Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 2019: 474-479.
- [11] CAI Y Q. Research and implementation of network photos sensitive detection [D]. Chongqing: Chongqing University of Posts and Telecommunications, 2021.
蔡元奇. 网络图片敏感文字的检测研究与实现[D]. 重庆:重庆邮电大学, 2021.
- [12] LIANG K. Sensitive information recognition based on sentiment analysis of user original content[D]. Xi'an: Xidian University, 2022.
梁康. 基于用户原创内容情感分析的敏感信息识别[D]. 西安:西安电子科技大学, 2022.
- [13] SONG K S, KANG Y, GAO W, et al. Evidence aware neural pornographic text identification for child protection [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14939-14947.
- [14] DOU J R. Research on detection methods for sensitive content on internet [D]. Xi'an: Xidian University, 2023.
窦井苒. 面向互联网敏感内容检测方法的研究[D]. 西安:西安电子科技大学, 2023.
- [15] WANG L J. Sensitive text classification based on multi-module graph attention mechanism[D]. Xi'an: Xidian University, 2022.
王莉娟. 基于多模块图注意力机制的敏感文本分类[D]. 西安:西安电子科技大学, 2022.
- [16] HU C. Research on the identification of undesirable short texts with variants [D]. Hangzhou: Hangzhou Dianzi University, 2022.
胡创. 带变体的不良短文本识别方法研究[D]. 杭州:杭州电子科技大学, 2022.
- [17] GU H S, LI Q, SHEN D M. Research on sensitive text classification detection and classification based on improved artificial neural network [J]. Journal of Combinatorial Optimization, 2023, 46(3): 20.
- [18] ZHANG J Q, GAO S B, SU R, et al. A two-stage knowledge distillation text classification method based on a fusion group teaching assistant model [J]. Journal of Changzhou University (Natural Science Edition), 2024, 36(6): 71-82.
张骏强, 高尚兵, 苏睿, 等. 融合群助教模型的两阶段知识蒸馏文本分类方法[J]. 常州大学学报(自然科学版), 2024, 36(6): 71-82.
- [19] DENG C Y, QI X G. Fraud identification in telecommunication networks based on transformer and multi-task learning [J]. Journal of Applied Science and Technology, 2024, 51(5): 256-262.
邓翠艳, 齐小刚. 基于Transformer及多任务学习的电信网络诈骗识别[J]. 应用科技, 2024, 51(5): 256-262.
- [20] ZHANG H Y. Sensitive information processing in audio based on speech recognition technology [D]. Chongqing: Chongqing Technology and Business University, 2023.
张洪艳. 基于语音识别技术的音频内敏感信息处理[D]. 重庆:重庆工商大学, 2023.
- [21] WEN M Q. Research on classification method of gun-related Internet public opinion information based on multi-layer information fusion [D]. Xi'an: Xi'an University of Technology, 2024.
文苗青. 基于多层信息融合的涉枪网络舆情信息分类方法研究[D]. 西安:西安理工大学, 2024.
- [22] DU M H, CUI X J, JIANG Y M, et al. Research on MOOC recommendation method based on LDA topic model [J]. Computer and Digital Engineering, 2024, 52(12): 3616-3622, 3760.
杜梦晗, 崔仙姬, 姜雨蒙, 等. 基于LDA主题模型的MOOC推荐方法研究[J]. 计算机与数字工程, 2024, 52(12): 3616-3622, 3760.
- [23] WANG S X. An analysis on the application of translation back to college English writing teaching [J]. English Abroad, 2024(12): 132-134.
汪淑娴. 回译法在大学英语写作教学中的应用探析[J]. 海外英语, 2024(12): 132-134.
- [24] ZHANG M C, PU Q M, HUANG F L. False news detection based on BERT-BiLSTM model [J]. Journal of the Chinese Academy of Electronic Sciences, 2025, 20(1): 33-40.
张敏超, 蒲秋梅, 黄方俐. 基于BERT-BiLSTM模型的虚假新闻检测[J]. 中国电子科学研究院学报, 2025, 20(1): 33-40.
- [25] WANG T J, YAN Y, GUO X R, et al. Tangka question classification model combining multi-scale CNN and bidirectional LSTM [J]. Science Technology and Engineering, 2024, 24(22): 9490-9497.
王铁君, 闫悦, 郭晓然, 等. 融合多尺度CNN与双向LSTM的唐卡问句分类模型[J]. 科学技术与工程, 2024, 24(22): 9490-9497.
- [26] GONZALEZ-HUITRON V, LEÓN-BORGES J A, RODRIGUEZ-MATA A E, et al. Disease detection in tomato leaves via CNN with lightweight architectures implemented in raspberry Pi 4 [J]. Computers and Electronics in Agriculture, 2021, 181: 105951.

(责任编辑:黄健)