ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 1050－1061
DOI:10. 16337/j. 1004-9037. 2021. 05. 020

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

海量网站中博彩类违法网站的捕获方法

刘家银[1,2,3]，印　杰[1,2,3]，牛博威[4]，诸葛程晨[1,2,3]，贺海辰[5]

(1. 江苏警官学院计算机信息与网络安全系,南京 210031;2. 江苏警官学院江苏省电子数据取证分析工程研究中心,南京 210031;3. 江苏警官学院江苏省公安厅数字取证重点实验室，南京 210031;4. 江苏省公安厅网络安全保卫总队,南京 210024;5. 南京市公安局大数据中心,南京 210005)

摘　要：针对海量网站中博彩类违法网站的检测问题,提出了一种基于BERT-BiLSTM与多分类器决策级融合的网站分类方法。该方法通过以下方式来提升分类性能:首先采用网页标签标题、超链接标题等优先的网页特征文本提取方法提升特征文本内容的丰富度;其次提出基于BERT-BiLSTM的文本分类模型,该模型具有良好的语句特征表示能力,从而提升分类性能;最后将网站标题、关键词和网页文本3种网站不同描述维度的分类结果进行决策级融合,进一步提升整个系统的性能与鲁棒性。通过采用多种策略生成疑似博彩网站的域名,提升该方法主动捕获博彩类违法网站的能力。实验结果以及在现实网络空间中的运行结果都充分验证了本文方法的有效性。

关键词：在线博彩;网站检测;自然语言处理;决策级融合;深度学习

中图分类号：TP3　　**文献标志码**:A

Capture Methods of Gambling Related Illegal Websites in Massive Websites

LIU Jiayin[1,2,3], YIN Jie[1,2,3], NIU Bowei[4], ZHUGE Chengchen[1,2,3], HE Haichen[5]

(1. Department of Computer Information and Cyber Security, Jiangsu Police Institute, Nanjing 210031, China;2. Jiangsu Electronic Data Forensics and Analysis Engineering Research Center, Jiangsu Police Institute, Nanjing 210031, China;3. Key Laboratory of Digital Forensics of Jiangsu Provincial Public Security Department, Jiangsu Police Institute, Nanjing 210031, China;4. Cyber Security Guard Corps, Jiangsu Provincial Public Security Department, Nanjing 210024, China;5. Big Data Center, Nanjing Municipal Public Security Bureau, Nanjing 210005, China)

Abstract: Aiming at the problem of detecting illegal gambling websites in massive websites, this paper proposes a classification method based on BERT-BiLSTM and multi-classifier decision-level fusion. This method improves the classification performance by adopting the following steps. Firstly, it extracts the textual information considered with high priority, i.e., meta information in HTML head and hyperlink titles on a web page, to enhance the richness of textual features. Secondly, a novel text classification model based on BERT-BiLSTM is designed, and it is proved superior in learning better sentence feature representatives and boosting performance. At last, the decision-level fusion is performed on the classification results from multiple dimensions (i.e., website title, keywords, and page text) to further improve the performance and robustness of the entire system. Moreover, a variety of strategies generating

基金项目:江苏省公安厅科技研究(2020KX008)资助项目;江苏省高等学校自然科学基金(19KJB510022)资助项目;江苏警官学院高层次引进人才科研启动基金资助项目。
收稿日期:2020-10-09;**修订日期**:2021-01-20

suspicious domain names are used to improve the ability to actively detect illegal websites. Experimental results and running results in real cyberspace demonstrate the effectiveness of the proposed method.
Key words: online gambling; website detection; natural language processing; decision level fusion; deep learning

引　言

随着互联网技术的蓬勃发展,博彩行业中一些不法分子利用互联网庞大的流量、便捷的支付和强互动性等特点,构建了组织架构严密、分工明确、公司化、制度化运作的线上博彩网站和APP等,重构了另一个网络博彩世界。网络赌博严重影响人们的身心健康,危害正常的经济秩序,败坏社会风气[1-2],因此开展网络博彩的检测与阻断,对于维护社会公序良俗,保障经济健康发展具有重要的现实意义。

违法博彩类网站的捕获问题其实质是网页分类问题。早期网页分类常采用黑名单方法[3],具有检测速度快的优点,但是不能处理完全未知的网站,且黑名单数据库的及时更新也是该方法面临的一大难点。然而博彩类违法网站为规避封堵,经常变换其域名,因此基于黑名单的方法在博彩类违法网站的检测中表现不佳。为解决黑名单方法存在的上述问题,部分研究人员通过分析统一资源定位符(Uniform resource locator, URL)[4]、网页文本以及网页图像[5]等来实现网站的分类。基于URL的方法只需从URL中提取特征,因此检测速度极快。然而URL不能完整地表达网站的特征,其应用领域极其有限。网页内容能提供丰富的信息,因此基于网页内容的网站分类方法成为当前研究的主流。该类方法通常采用机器学习方法来实现网站分类,首先从网页中提取文字特征、图像特征以及链接、标签和脚本函数等统计特征,然后训练决策树、支持向量机(Support vector machines, SVM)等分类器,最终实现对网站的分类。然而传统特征提取方法提取的特征较为单一,且大部分都只能提取到较为浅层的特征,特征抽象能力与泛化能力较弱。本文提出一种基于BERT(Bidirectional encoder representation from transformers)+BiLSTM(Bidirectional long short-term memory)模型与多分类器决策级融合的博彩类违法网站检测方法,充分利用BERT的特征抽取能力来提升文本分类精度。此外,通过对网站不同描述维度的特征分别进行训练与分类,最后进行决策级融合,进一步提升整个系统的检测性能与鲁棒性。

1　网站检测研究进展

早期研究人员经常采用黑名单来检测违法网站,该方法将可疑网站的域名与黑名单数据库中的违法域名进行匹配,如匹配成功则将该域名标记为违法域名[3]。基于黑名单的违法网站检测方法检测速度极快,但是其最大缺点在于其不能判别不在黑名单中的域名。由于网站包含的丰富信息,如链接、文本和图像等,基于网站内容的网站检测方法逐步成为研究主流。基于URL的方法利用URL字符串、URL统计信息等来提取特征进而实现网站分类[4]。此类方法因不需要访问网页里面的内容,因此检测速度极快。但由于URL能提供的信息过少,不能完整地描述违法网站的特征,因此在大部分应用场合其检测性能都较低。

相较于URL,网页文本能提供更丰富的信息,如文本、图像、层叠样式表(Cascading style sheet, CSS)和超文本标记语言(Hyper text markup language, HTML)标签等,能更好地实现网页分类。如Fa等[6]提取网页HTML中文本的词频-逆文档词频(Term frequency-inverse document frequency, TF-IDF)特征以及图像、iframe标签、ul标签和嵌入链接数量特征等,然后采用随机森林分类器实现网页的分类。Kotenko等[7]利用机器学习与数据挖掘算法来分析网页文本、HTML标签以及URL地址信息以检测含有违法信息的网站。Gaifulina等[8]通过对多个不同维度数据的子分类器结果进行融合来实现网

站的分类。

近年来,基于深度学习的图像处理技术取得了极大发展,因此部分学者研究利用网页图像来实现网站的分类。Li等[9]从网页的截图中提取基于视觉词袋模型(Bag-of-visual-word, BoVW)的加速稳健特征(Speeded-up robust features, SURF),然后采用SVM进行网站分类。Phoka等[10]则采用网页中的子图来实现钓鱼网站的检测。Mahmoud等[11]基于图像皮肤检测技术来检测网页中是否存在色情图片,进而实现色情网站检测。然而,基于视觉特征的网页分类容易受到训练集样本质量和模型泛化程度的影响,导致识别率较不高。为提高结果的鲁棒性,部分研究人员提出融合文本与图像特征来进行网页分类。Ahmadi等[12]综合分析视觉、文本和轮廓特征,提出了一种基于层次结构分类器的色情网页检测系统。Chen等[5]分别提取网页文本的Doc2vec特征并采用SVM分类器训练,以及网页截图的Spa-BoVW特征并采用随机森林分类器训练,最后再用逻辑回归(Logistic regression, LR)来融合文本与图形分类结果。也有部分研究人员利用网站的其他特征来进行检测,如Tong等[13]利用HTTP Post请示的行为模式来实现博彩网站的检测;Zeng等[14]基于网页的指纹特征来实现恶意网站的检测。

2 网络博彩网站特征

网络博彩是一种通过网络进行的新型赌博模式,目前的网络博彩类型繁多(如赌球、赌马、骰宝、轮盘和网上百家乐等)。与传统线下赌场相比,网络赌博完美地利用了互联网与生俱来的便捷性、辐射范围广等特点,只需连接网络即可随时随地完成投注、资金交割,赌资运转速度更快,赌资的数额巨大。据《在线赌博市场"规模、份额"行业报告》显示,2019年全球在线赌博市场规模估值为537亿美元,预计2020至2027年将以11.5%的复合年增长率增长[15]。从建站到引流,网络赌博已经形成技术、推广、运营和代理等构成的分工明确、制度化、团队化和国际化运作的完整黑灰产业链,基网站运营架构如图1所示。

图1 博彩网站运营架构

Fig.1 Operation structure of gambling website

为逃避监管以及吸引用户参与,网络博彩网站通常具有以下特征:

(1) 同一网站配置多个域名,以规避监管机关的封堵。这需要对网站实行实时判别,才能有效阻断此类网站。

(2) 为逃避监管,面向中国的博彩运营公司大部分位于境外,东南亚地区业已成为网络博彩团队的主要据点。这导致大量资金非法外流,也极大地增加了监管机关取证、执法的难度,因此阻碍博彩网站的访问成为防止此类案件发生的最佳选择。

(3) 通过第三方论坛、色情网站、微信群、线上代理等平台以及广告推广方式进行传播引流。

(4) 通过入侵国内网站(尤其是政府网站、学校网站和中小型企业网站)放置暗链、挂马等进行搜索引擎优化,以及域名、流量劫持等网络攻击方式进行引流,严重危害网络空间的安全。

3 本文方法

本文提出的海量网站中违法网站捕获方法主要由以下4个模块组成:网络爬虫、预处理、分类器以及分类决策。网络爬虫模块爬取指定域名的HTML文本信息;预处理模块提取网页的标题、关键词以及网页中包含的中英文特征文本;分类器模块基于网页的标题、关键词以及特征文本信息分别得出该网页为博彩网站和正常网站的概率;分类决策模块则基于分类器模块获得的软分类值,利用XGBOOST决策分类器获得最终的结果,判断出网站是否为博彩网站。本文方法流程如图2所示。

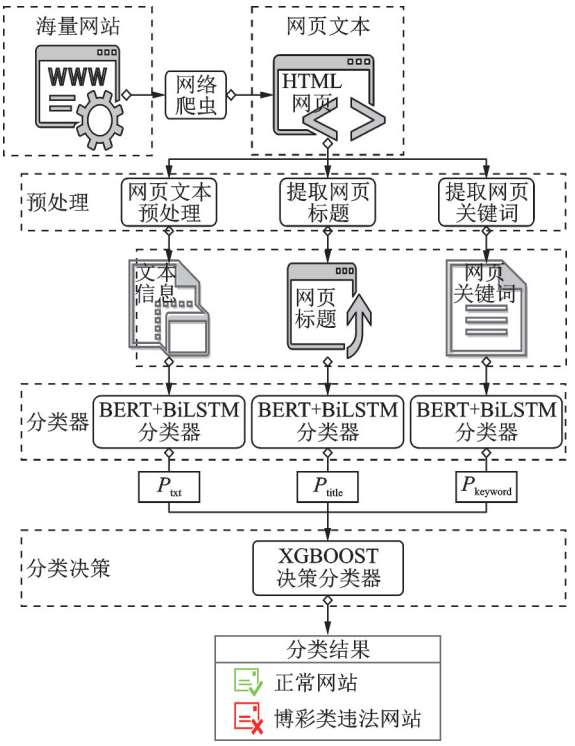


图2 本文方法流程图

Fig.2 Flow chart of the proposed method

3.1 网站内容爬取

对于博彩类网站的捕获问题,首先需要利用网络爬虫来获取待测网站的HTML内容,然后利用分类算法来判断该网站是否为博彩网站。而对于博彩网站的检测与阻断,为减少分类工作量、提高效率,通常只关注网站的主域名。因此欲从海量网站中捕获更多的博彩网站,其关键点在于获得足够多的疑似博彩网站的主域名。而常用的网页爬虫策略:深度优先策略与广度优先策略,不能很好地满足获取更多疑似博彩网站主域名的要求。经过分析发现,博彩网站通常具有如下特点:

(1) 网页中包含同一博彩网站的其他域名;

(2) 架设博彩网站的互联网数据中心(Internet data center, IDC)中可能包含其他博彩网站的域名;

(3) 同一博彩网站通常绑定多个域名,这些域名往往具有一定规律性。

基于上述发现,采用以下策略来生成疑似博彩网站主域名列表:

(1) 针对特点1,采用广度优先策略来提取其他网站主域名;

(2) 针对特点2,提取存在博彩类违法网站域名的IDC中其他网站主域名;

(3) 针对特点3,基于掌握的域名变化规律自动生成其他域名,并判断该域名是否可访问,如果可以则将其添加到待判断队列中。

利用上述方法获得网站主域名的列表后,按照如图3所示的步骤来爬取每一个网站主页的HTML

文本。具体步骤如下：

第1步 从种子URL列表中取出URL；

第2步 判断该URL是否已经被处理过，若是返回第1步，否则将该URL添加到已处理队列中，然后执行第3步；

第3步 利用爬虫工具解析URL，爬取对应网页的数据；

第4步 解码网页数据，并保存以供后续处理。

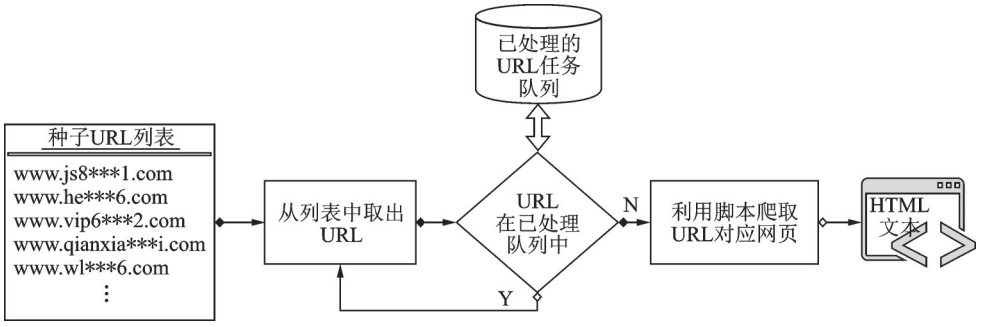


图3 网络爬虫流程图

Fig.3 Flow chart of web crawler

3.2 网页文本预处理

网站标题是对一个网页的高度概括，具有精确性与简短性，可以高效地进行网页类型的分类。部分研究人员采用清华大学的Sun等创建的THUCNews数据集[16]的中文标题来进行网页新闻类型的分类，并取得了较好的效果。然而在实际网络空间中部分网站的标题不规范，存在无标题、标题无意义以及标题与网站内容不符等现象，甚至有一些合法网站因受黑客攻击其标题被篡改为博彩、色情等非法标题。

网站的关键词通常与网站的主页内容高度相关，与网站标题相似，其也具有高概括性与简短性。因此网站关键字也可以被用于网站类型的分类，在网页标题无内容、无意义以及内容不相关时具有一定的替代作用。然而，与网站标题类似，网站关键词也存在部分网站无关键词、关键词无意义、关键词与网站内容不符以及关键词被篡改为博彩、色情等非法关键词的现象。

与标题和关键词相比，网页中的文本通常具有更丰富的信息，可以对网站类型进行更准确、全面的描述。因而采用网页文本来进行网站分类，其分类结果通常更准确。然而为吸引赌客的参与欲望，部分博彩网站通过用图片代替文字、加载Flash动画等方式来提升视觉效果，这导致网页中文字内容偏少，直接影响到基于网页文本的网站分类性能。因此，本文提出基于网站标题、关键词与网页特征文本分类结果决策级融合的方法，以提升博彩类违法网站检测的准确性与鲁棒性。

基于上述分析，本文从网站主页HTML文本中提取标题、关键词以及文本信息作为网站的分类特征。通过解析HTML文本，可以直接提取网站的标题、关键词。对于网页中的文本信息，不同网站文字数量差异巨大，从个位数到成千上万。而对于绝大部分自然语言处理算法，如Text-CNN[17]、BERT等，输入数据的长度与运算时需要的GPU显存成正相关，因而其输入的最大长度具有一定限制。

基于GPU显存容量考虑，本文提取的网页中文本的最大长度设置为256。若在解析HTML并提取网页中文本后，直接进行最大长度为256的截断，此时截取的内容可能只包含网页的某部分特定内容，文本内容丰富性低，不具备充分表达该网站类型的能力。通过对大量博彩网站的分析发现，博彩类网

站的主页中通常存在“彩票”“扎金花”“活动大厅”“优惠活动”“免费试玩”等具有典型特征的标题。因此本文提出优先提取网页中标签标题、超链接标题等来生成网页文本。相对于直接截断,该方法提取的特征文本能够从更多角度描述网页的类型。此外,为提取更多不同的标题或文本,本文对网页中文本进行最大长度为6的直接截断。

3.3 基于BERT-BiLSTM的文本分类

3.3.1 BERT模型

Devlin等在2019年提出的BERT[18]给自然语言处理预训练模型带来了突破性进展。与Word2Vec等词向量模型不同,BERT不再需要预先训练复杂的字向量和词向量,只需将语句直接输入到BERT模型中,它就会自动提取出序列的词级特征、语法结构特征和语义特征。

BERT模型通过叠加多个Transformer编码器层来实现特征的逐步抽象。编码器由自注意力机制与前向传播网络构成,并与残差网络类似,也将输入值与输出值结合在一起以解决梯度消失问题,如图4所示。

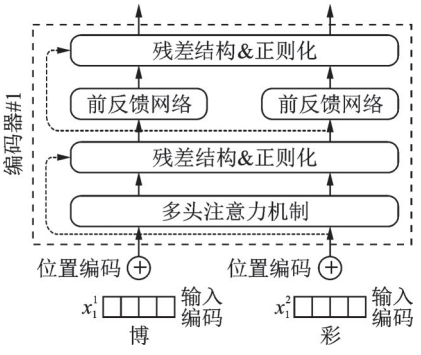


图4 Transformer编码器模型

Fig.4 Model of transformer encoder

对于第i层编码器,其输入向量为$\boldsymbol{X}_i$(第1层为输入语句的词Embedding向量,其他层为前1层的输出)。首先,将$\boldsymbol{X}_i$输入到多头自注意力模块中进行注意力权值的计算,如图5所示。

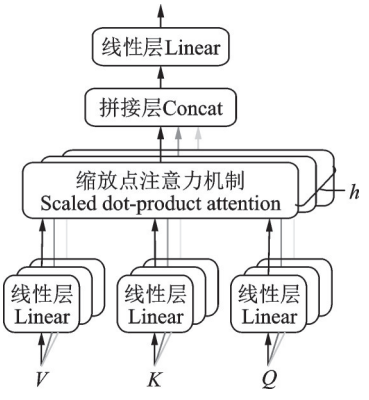


图5 多头注意力结构

Fig.5 Multi-head attention struture

自注意力权值计算方式为

$$Z_i^j=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i^j\boldsymbol{K}_i^{j\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_i^j \tag{1}$$

式中:Z_i^j表示第i层编码器的第j头自注意力权值;d_k表示Embedding维度;$\boldsymbol{Q}_i^j$、$\boldsymbol{K}_i^j$与$\boldsymbol{V}_i^j$分别表示该头自注意力模块计算而得的查询相似得分、Key相似得分和Value相似得分,其计算方式为

$$\boldsymbol{Q}_i^j=\boldsymbol{X}_i\boldsymbol{W}_i^{Qj} \tag{2}$$

$$\boldsymbol{K}_i^j=\boldsymbol{X}_i\boldsymbol{W}_i^{Kj} \tag{3}$$

$$\boldsymbol{V}_i^j=\boldsymbol{X}_i\boldsymbol{W}_i^{Vj} \tag{4}$$

式中:$\boldsymbol{W}_i^{Qj}$为查询权重矩阵;$\boldsymbol{W}_i^{Kj}$为Key权重矩阵;$\boldsymbol{W}_i^{Vj}$为Value权重矩阵。

其次,将自注意力权值Z_i^j在第2个维度上拼接起来,再做线性变换,即得多头注意力Z_i^{A}为

$$Z_i^{\mathrm{A}}=\mathrm{Concat}(Z_i^0,Z_i^1,\cdots,Z_i^h)\boldsymbol{W}^O \tag{5}$$

式中:h为注意力头的数量;$\boldsymbol{W}^O$为多头注意力的权重矩阵。

在得到多头注意力输出Z_i^{A}后,利用残差结构产生新输出为

$$Z_i^{\mathrm{L}}=\mathrm{LayerNorm}(Z_i^{\mathrm{A}}+\boldsymbol{X}_i) \tag{6}$$

然后,将Z_i^{L}传递到前向传播网络,最后再通过归一化的残差网络得到第i层编码器的输出为

$$\boldsymbol{Z}_i=\mathrm{LayerNorm}(\mathrm{Feed}(W_{\mathrm{f}}Z_i^{\mathrm{L}}+b_{\mathrm{f}})+\boldsymbol{X}_i) \tag{7}$$

式中:W_{f}为前向传播网络的权重系数;b_{f}为偏重系数。

3.3.2 BiLSTM模型

长短期记忆网络(Long short-term memory, LSTM)是一种特殊的RNN类型,它巧妙地利用门控来捕捉序列信息、达成长期记忆,并解决了RNN训练时所产生的梯度爆炸或梯度消失问题。LSTM单元结构如图6所示。

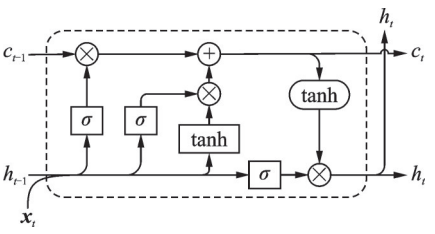


图6 LSTM单元结构

Fig.6 Cell structure of LSTM

LSTM每个单元由记忆单元c^t,输入门i^t,输出门o^t和忘记门f^t组成。$\boldsymbol{x}_t$是LSTM单元的输入,表示输入序列中一个单词的特征向量。每个LSTM单元中的3个门和记忆单元可由以下公式计算得出

$$i_t=\sigma(\boldsymbol{W}_{xi}\cdot\boldsymbol{x}_t+\boldsymbol{W}_{hi}\cdot h_{t-1}+\boldsymbol{W}_{ci}\cdot c_{t-1}+b_i) \tag{8}$$

$$f_t=\sigma(\boldsymbol{W}_{xf}\cdot\boldsymbol{x}_t+\boldsymbol{W}_{hf}\cdot h_{t-1}+\boldsymbol{W}_{cf}\cdot c_{t-1}+b_f) \tag{9}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\tanh(\boldsymbol{W}_{xc}\cdot\boldsymbol{x}_t+\boldsymbol{W}_{hc}\cdot h_{t-1}+b_c) \tag{10}$$

$$o_t=\sigma(\boldsymbol{W}_{xo}\cdot\boldsymbol{x}_t+\boldsymbol{W}_{ho}\cdot h_{t-1}+\boldsymbol{W}_{co}\cdot c_t+b_o) \tag{11}$$

$$h_t=o_t\odot\tanh(c_t) \tag{12}$$

式中:σ表示激活函数;$\boldsymbol{W}$为权重矩阵;$\odot$表示逐个点乘积;b为偏置向量;h_t表示整个LSTM单元在时刻t的输出。

LSTM模型无法同时处理上下文信息,因此Graves等[19]提出双向长短期记忆网络(BiLSTM)由两个LSTM构成,且连接着同一个输出层,为输出层的数据同时提供上下文信息。记BiLSTM中的前向LSTM和后向LSTM在时刻t的输入处理分别为

$$\vec{h}_t=\vec{F}_{\mathrm{LSTM}}(x_t,\vec{h}_{t-1}) \tag{13}$$

$$\overleftarrow{h}_t=\overleftarrow{F}_{\mathrm{LSTM}}(x_t,\overleftarrow{h}_{t-1}) \tag{14}$$

将前向输出$\vec{h}_t$和后向输出$\overleftarrow{h}_t$拼接在一起,即为BiLSTM在时刻t的输出结果,表示为

$$h_t=[\vec{h}_t,\overleftarrow{h}_t] \tag{15}$$

3.3.3 BERT-BiLSTM文本分类模型

根据Jawahar等对BERT模型内置机理的研究表明[20],BERT模型各编解码层学到的特征不尽一致,模型底层主要学到的是语句序列的短语级特征,模型中层可得到语句序列的句法结构特征,模型顶层则可提取语句序列的语义特征。BERT模型的层次越高,学到的特征越抽象,模型的特征抽取能力明显强于传统模型。基于BERT强大的特征抽取能力,本文提出了基于BERT-BiLSTM的文本分类模型。利用BERT作为编码器,将文本映射成具有更强语义描述能力的特征向量,在上面叠加BiLSTM模块进一步提取上下文序列信息,以获得更优的特征。BERT-BiLSTM模型的具体处理流程如图7所示。

首先将网页文本语句序列X经过BERT预训练语言模型进行处理,提出不同抽象能力的多层次特征信息$\{Z_1,Z_2,\cdots,Z_{12}\}$。为提高特征表征不同语境中的句法与语义信息的能力,本文将最后3层的特征信息Z_{10}、Z_{11}和Z_{12}进行拼接得到

$$Z^O=\mathrm{Concat}(Z_{10},Z_{11},Z_{12}) \tag{16}$$

然后将Z^O作为输入传入到BiLSTM模块,对组合特征做进一步训练,利用BiLSTM良好的上下文序列信息抽取能力来提取出更具有区分能力的隐藏层特征$\boldsymbol{H}=[h_1,h_2,\cdots,h_n]$,其中$h_t$为$Z^O$中第$t$个单

词的语义特征输入到式(15)中计算而得。

最后,对于BiLSTM模块提取的特征$\boldsymbol{H}$,叠加了2层全连接层以对语句的特征信息进行分类,获得文本为博彩网站与正常网站的分类结果$p(\text{illegal}|X)$和$p(\text{normal}|X)$。

本文提出的文本分类模型包括3个计算步骤,其时间复杂度分别为:BERT模块的复杂度为$O(n^2\cdot d)$;BiLSTM模块的时间复杂度为$O(n\cdot d^2)$;全连接层模块的时间复杂度为$O(n^2)$。因此本文提出的BERT-BiLSTM文本分类模型的时间复杂度为$O(nd\cdot\max(n,d))$。

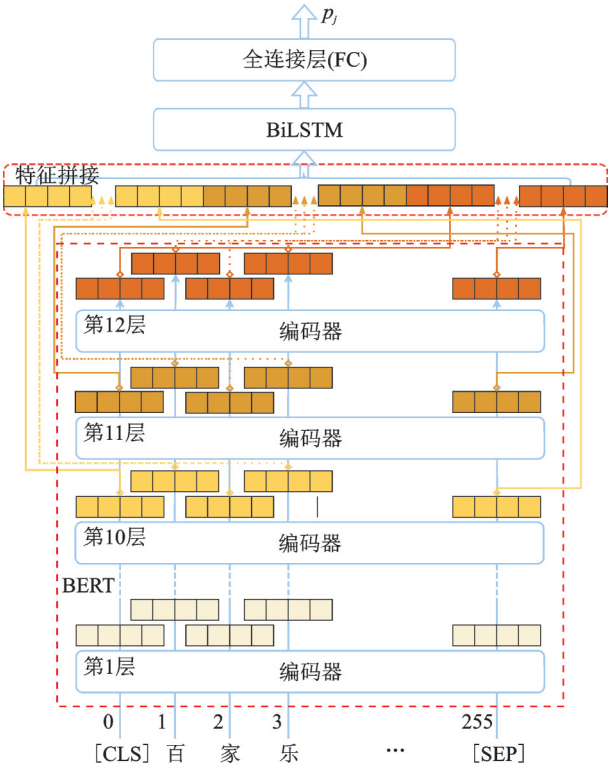


图7 BERT-BiLSTM模型

Fig.7 BERT-BiLSTM model

3.4 决策级融合

相比单分类器,融合多个分类器的结果或者集成多分类器通常能够取得更好的分类性能。因此,在BERT-BiLSTM模型分类结果的基础上,本文提出基于XGBOOST[21]的多分类器决策级融合方法,将网页标题、关键词与网页文本的分类结果进行融合,进一步提升分类性能。

在进行违法博彩类网页检测时,对于任意一个网页,通过以下步骤来对其进行多分类器决策级融合判定:

第1步 爬取网页内容P。

第2步 基于网页内容P提取3类特征文本:网页标题X_{title}、网页关键词X_{keyword}以及网页特征文本X_{text}。

第3步 将X_{title},X_{keyword},X_{text}输入到对应的BERT-BiLSTM文本分类器中,分别获得标题的分类结果:$p(\text{illegal}|X^i_{\text{title}})$、$p(\text{normal}|X^i_{\text{title}})$;关键字的分类结果:$p(\text{illegal}|X^i_{\text{keyword}})$、$p(\text{normal}|X^i_{\text{keyword}})$;网页文本的分类结果:$p(\text{illegal}|X^i_{\text{text}})$、$p(\text{normal}|X^i_{\text{text}})$。

第4步 将第3步的6个分类结果进行拼接得到一个6维的特征向量

$[\,p(\text{illegal}|X^i_{\text{title}})\ \ p(\text{normal}|X^i_{\text{title}})\ \ p(\text{illegal}|X^i_{\text{keyword}})\ \ p(\text{normal}|X^i_{\text{keyword}})\ \ p(\text{illegal}|X^i_{\text{text}})\ \ p(\text{normal}|X^i_{\text{text}})\,]$

然后将其作为输入传入XGBOOST决策分类器,获得网页P为博彩网站的概率$p(\text{illegal}|P)$与正常网站的概率$p(\text{normal}|P)$。最终根据式(17)获得最终判定结果。

$$y'=\begin{cases}\text{illegal} & p(\text{illegal}|P)>p(\text{normal}|P)\\ \text{normal} & \text{其他}\end{cases}\tag{17}$$

4 实验结果与分析

本文实验硬件平台为Intel Xeon(R)48核处理器,频率2.30 GHz,显卡为NVIDIA tesla V100。实验代码通过Python语言与深度学习框架Pytorch来实现。

4.1 实验数据集与参数配置

本文实验的数据集是在真实网络环境中采集而来，数据集中包括博彩类违法网站共135 881个，企业、政府和新闻等正常网站共140 205个。在具体的实验过程中，本文采用10折交叉验证来对算法的性能进行评估。本文实验采用“BERT-Base，Chinese”中文预训练模型，该模型的Transformer层数为12层，每个Transformer包含有12个自注意力头部，模型的隐层节点数为768。网络微调训练阶段，网页标题和关键词文本的长度设置为32，网页特征文本的长度设置为256，批次数设置为30，批处理大小为48，学习率η为$5e^{-5}$，BiLSTM的输出中使用Dropout，取值为0.2。

4.2 实验结果与分析

为验证本文提出的网页特征文本提取算法的性能，将其与直接截断网页文本的方法进行对比。对比中采用的分类算法为原始BERT算法以及本文提出的BERT-BiLSTM算法，实验结果如表1所示。表1中Text_raw指直接截断方法提取网页文本，Text_tag指本文提出的标签标题、超链接标题等优先提取的网页特征文本提取方法。从表1中可以发现，不论是采用基本的BERT还是采用本文提出的BERT＋BiLSTM模型来提取文本的特征向量，基于本文提出的文本预处理方法提取的网页特征文本，其分类性能要优于通过直接截断这一方式来提取网页特征文本的。

表1 不同文本预处理方法分类性能对比

Table 1 Classification performance comparison of different text preprocessing methods

特征文本	分类方法	准确度	精确度	召回率	*F*-值
Text_raw	BERT	0.976 9	0.974 1	0.979 0	0.976 6
	BERT＋BiLSTM	0.983 1	0.984 7	0.980 8	0.982 8
Text_tag	BERT	0.980 9	0.980 4	0.980 9	0.980 6
	BERT＋BiLSTM	0.992 9	0.992 5	0.993 0	0.992 8

为验证本文提出的BERT-BiLSTM算法的有效性，本文将其与原始BERT算法以及BERT与其他深度学习算法的组合进行对比。实验在本文提出的网页特征文本数据集上进行，实验结果如表2所示，其中BERT_MULTI指本文提出的BERT＋BiLSTM模型去掉BiLSTM模块，即将特征拼接结果直接与全连接层相连。

从表2可以发现，在BERT层后叠加一个深度学习模型，分类结果的各项性能指标基本都有明显的提升。这意味着，基于BERT模型提取语句序列的特征信息，再叠加一个深度学习模块进行分类，可以取得更好的分类性能。同时，本文提出的BERT＋BiLSTM模型的分类性能要显著优于BERT与其他深度学习算法的组合，表明本文提出的方法可以有效提高网页分类性能。

表2 BERT-BiLSTM与其他组合模型的分类性能对比

Table 2 Comparison of classification performance between BERT-BiLSTM with other combination models

组合模型	准确率	精确度	召回率	*F*-值
BERT	0.980 9	0.980 4	0.980 9	0.980 6
BERT＋CNN	0.990 1	0.989 7	0.990 1	0.989 9
BERT＋DPCNN	0.982 1	0.981 6	0.982 1	0.981 9
BERT＋RNN	0.987 5	0.987 0	0.987 6	0.987 3
BERT_MULTI	0.989 8	0.989 3	0.990 0	0.989 6
BERT＋BiLSTM	0.992 9	0.992 5	0.992 3	0.992 8

从表2中还可以发现，将BERT模型中不同层次Transformer输出的特征信息进行组合，进而用于分类，可以有效提高分类性能。这表明BERT模型中各层次的Transformer可以在不同抽象层次上提取文本的特征信息，将不同抽象层次的特征信息进行组合，可以获得更丰富的特征信息，进而提升分类性能。这也验证了本

文提出的将BERT模型中多层特征信息进行融合作为网页表示特征的有效性。

为验证本文提出的基于XGBOOST的多分类器决策级融合算法对网页分类性能的提升,本文将其与各分类器单独分类结果进行比较,对比结果如表3所示。从表3中可以发现,仅基于标题或者关键词的方法其分类性能相对偏低,这主要是因为部分网页的标题或关键词存在无内容、无意义或者描述与网站内容不符等现象导致。而基于网页文本的方法其分类性能要显著优于基于标题或者关键词的方法,这表明前文分析的网页文本能对网站类型进行更细致、准确和全面的描述。基于多分类器决策级融合的分类性能要显著优于单分类器,这也验证了本文提出的基于多分类器决策级融合方法的有效性。

表3 决策级融合方法与单一分类方法分类性能对比

Table 3 Classification performance comparison between decision level fusion and single classification method

特征项	分类方法	准确率	精确率	召回率	*F*-值
Title	BERT	0.944 5	0.943 7	0.943 6	0.943 6
	BERT+BiLSTM	0.946 9	0.945 9	0.946 3	0.946 1
Keyword	BERT	0.900 3	0.898 1	0.899 5	0.898 8
	BERT+BiLSTM	0.917 0	0.915 3	0.916 2	0.915 7
Text_tag	BERT	0.980 9	0.980 4	0.980 9	0.980 6
	BERT+BiLSTM	0.992 9	0.992 5	0.993 0	0.992 8
Fusion	BERT	0.988 7	0.988 5	0.988 6	0.988 5
	BERT+BiLSTM	0.994 8	0.995 0	0.994 4	0.994 7

最后将本文提出的方法与3类典型的网页分类算法进行对比。其中,文献[6]为基于网页文本内容的方法,文献[9]为基于网页截屏的方法,文献[5]为基于网页内容与网页截屏融合的方法。

表4 本文方法与其他方法的分类性能对比

Table 4 Comparison of classification performance between the proposed method and other methods

分类方法	准确率	精确率	召回率	*F*-值
文献[6]方法	0.942 2	0.959 6	0.921 4	0.940 1
文献[9]方法	0.951 5	0.956 8	0.943 5	0.950 1
文献[5]方法	0.994 6	0.996 8	0.992 3	0.994 5
本文方法	0.994 8	0.995 0	0.994 4	0.994 7

从表4中可以发现,文献[6]的多项分类性能指标在几种方法中均较低,说明基于传统文本特征提取方法(TF-IDF)与传统机器学习分类方法(随机森林)不能很好地利用网页文本的语义信息,导致其分类性能,尤其是召回率不高。文献[9]的分类性能略优于文献[6],这是因为博彩类网站的界面通常具有较为显著的视觉特征,充斥着醒目的图像、动画等以吸引赌徒的注意。同时网页截屏较少受文本乱码、网页脚本以及网页跳转等影响,因此视觉特征是一种较好的网页分类特征,能获得较好的分类性能。然而,文献[9]采用视觉词袋来提取视觉特征,不具备对图像内容的理解能力,容易产生性能瓶颈。例如,当面对视觉内容不够显著的博彩类网页或者视觉特征显著的普通网页时容易出现漏检和误检的现象。文献[5]的分类性能远高于文献[6]与文献[9],证明多种不同特征,尤其是文本与视觉融合,具有一定互补性。基于多种特征融合的检测方法可以显著提高违法网站的检测性能与鲁棒性。本文提出的方法与文献[5]的分类性能基本一致,在准确率、召回率与*F*-值3项指标上略高于文献[5],在精确率指标上略低于文献[5]。然而,本文提出的方法仅需要获得网页的文本信息,而文献[5]需要采用Webdriver来模拟网页访问并截屏,这需要的时间远远超过网页文本的获取时间。因此本文的博彩网站检测效率要远高于文献[5]。同时,本文仅采用网页的文本信息就达到

了文献[5]提出的文本与图像融合方法的性能,证明本文提出的分类算法可以充分利用网页文本的语义信息,实现博彩类网站的有效检测。此外,将本文提出的博彩类违法网站检测系统部署于实际网络空间中,一个多月以来已检测出新的博彩类违法网站域名近20万个。综上所述,从理论与实际运行结果两方面都可以验证本文提出的基于BERT+BiLSTM模型的多分类器决策级融合算法可以有效检测出博彩类违法网站。本文的源代码可以参见如下网址:https://github.com/smiton/IllegalWebsiteClassifier。

5 结束语

本文总结了博彩类违法网站的一些特征,并在此基础上提出了一种基于BERT+BiLSTM模型与多分类器决策级融合的博彩类违法网站检测方法。通过实验以及在网络空间中的实际运行结果都验证了本文提出算法的性能。然而在实际运行中也发现本文提出方法捕获到的新博彩类违法网站域名大部分都与已有网站相关,捕获与之前无关的、由网络博彩团队新创建的博彩网站相对较少,这主要是由本文提出的疑似博彩网站主域名生成方法导致,因此如何高效发现真新博彩网站主域名是下一步工作的要点。本文提出的方法相对依赖网页中的文本信息,容易受到拆字、文字替换等因素的影响,因此研究如何利用博彩网站中的图像、网站风格等,对于提高博彩类违法网站捕获的鲁棒性具有重要意义,也是下一步工作的要点。

参考文献:

[1] LI E, BROWNE M, RAWAT V, et al. Breaking bad: Comparing gambling harms among gamblers and affected others[J]. Journal of Gambling Studies, 2017, 33(1): 223-248.

[2] 裴广一, 黄光于. 国内博彩业研究述评与未来展望[J]. 特区经济, 2017, 9: 29-32.
PEI Guangyi, HUANG Guangyu. A research review and the future prospects of the gambling industry in China[J]. Special Zone Economy, 2017, 9: 29-32.

[3] AGBEFU R E, HORI Y, SAKURAI K. Domain information based blacklisting method for the detection of malicious webpages[J]. International Journal of Cyber Security and Digital Forensics, 2013, 2(2): 36-48.

[4] SAHOO D, LIU C, HOI S C H. Malicious URL detection using machine learning: A survey[EB/OL]. (2019-08-21)[2020-07-26].https://arxiv.org/abs/1701.07179.

[5] CHEN Y, ZHENG R, ZHOU A, et al. Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism[J]. Sensors, 2020, 20(14): 3989.

[6] FA Z, GENG G G, YAN Z W, et al. A robust internet abuse detection method[C]//Proceedings of 2017 IEEE International Conference on Big Data (Big Data). Boston, USA: IEEE, 2017: 1712-1715.

[7] KOTENKO I, CHECHULIN A, KOMASKINSKY D. Categorisation of web pages for protection against inappropriate content in the Internet[J]. International Journal of Internet Protocol Technology, 2017, 10(1): 61-71.

[8] GAIFULINA D, CHECHULIN A. Development of the complex algorithm for web pages classification to detection inappropriate information on the Internet[C]//Proceedings of International Symposium on Intelligent and Distributed Computing. Saint-Petersburg, Russia: [s.n.], 2019: 278-284.

[9] LI L, GOU G, XIONG G, et al. Identifying gambling and porn websites with image recognition[C]//Proceedings of Pacific Rim Conference on Multimedia. Harbin, China: [s.n.], 2017: 488-497.

[10] PHOKA T, SUTHAPHAN P. Image based phishing detection using transfer learning[C]//Proceedings of 2019 11th International Conference on Knowledge and Smart Technology (KST). Phuket, Thailand: [s.n.], 2019: 232-237.

[11] MAHMOUD T M, ABD-EL-HAFEEZ T, OMAR A. An efficient system for blocking pornography websites[C]//Proceedings of Computer Vision and Image Processing in Intelligent Systems and Multimedia Technologies. [S.l.]: IGI Global, 2014: 161-176.

[12] AHMADI A, FOTOUHI M, KHALEGHI M. Intelligent classification of web pages using contextual and visual features[J].

Applied Soft Computing, 2011, 11(2): 1638-1647.

[13] TONG S, ZHANG H, SHEN B, et al. Detecting gambling sites from post behaviors[C]//Proceedings of 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). Hefei, China: IEEE, 2016: 2495-2500.

[14] ZENG X, KANG C, SHI J, et al. A novel website fingerprinting method for malicious websites detection[C]//Proceedings of Information and Communication Technology for Intelligent Systems. Singapore: Springer, 2019: 723-730.

[15] RESESEARCH G D. Online gambling market size, share|industry report[EB/OL]. (2020-04-01)[2020-07-07]. https://www.grandviewresearch.com/industry-analysis/online-gambling-market.

[16] SUN M, LI J, GUO Z, et al. Thuctc: An efficient Chinese text classifier[EB/OL]. (2016-05-17)[2020-07-26]. http://github.com/thunlp/THUCTC.

[17] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. (2014-09-03)[2020-07-26].https://arxiv.org/abs/1408.5882.

[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24)[2020-07-26].https://arxiv.org/abs/1810.04805.

[19] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.

[20] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: [s.n.], 2019: 3651-3657.

[21] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016: 785-789.

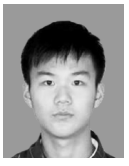
作者简介:

刘家银(1986-),男,博士,讲师,研究方向:网络安全、机器学习、数据融合技术,E-mail:liujiayin@jspi.cn。

印杰(1977-),**通信作者**,男,高级工程师,研究方向:机器学习、大数据、网络空间安全,E-mail: yinjie@jspi.cn。

牛博威(1983-),男,硕士研究生,研究方向:计算机科学与技术、软件工程和网络安全技术。

诸葛程晨(1986-),男,博士,讲师,研究方向:网络安全、机器学习。



贺海辰(1995-),男,硕士研究生,研究方向:大数据与数据挖掘。

(编辑:刘彦东)