doi:10.3772/j.issn.1002-0470.2024.02.005

基于深度学习的不良应用域名早期识别方法①

胡安磊②[*,**,****] 田 语[*,****] 陈 勇[**] 李振宇[*,****] 谢高岗[***,****]

([*]中国科学院计算技术研究所 北京 100190)

([**]中国互联网络信息中心 北京 100190)

([***]中国科学院计算机网络信息中心 北京 100083)

([****]中国科学院大学 北京 100049)

摘 要 不良应用网站依赖域名系统(DNS)实现不良内容传播,严重影响互联网的健康发展。尽早识别出不良应用网站对应的域名(即不良应用域名),并进行相应治理,对域名系统的管理与运行至关重要。本文从国家顶级域名(.CN)管理的角度出发,关注如何在注册阶段识别不良应用域名。分析发现不良应用域名在注册特征与文本结构2个维度,与正常域名存在显著差异。为此,提出了一种基于深度学习的不良应用域名早期识别方法。该方法首先提取域名的注册信息特征,并利用预训练语言模型基于Transformer的双向编码器(BERT)提取域名本身的文本语义特征,其次基于注意力机制融合2类特征,并最终使用全连接神经网络,构建域名分类器,实现不良应用域名的早期识别。基于真实网络数据的实验结果表明,所提方法分类准确率(*F*1分数)可达到0.99;消融实验结果也验证了所选特征的有效性和必要性。

关键词 域名系统(DNS);域名分类;深度学习;预训练语言模型

域名是用于识别和定位互联网资源的层次化、结构式字符标识,是重要的互联网基础资源,绝大多数互联网应用都以有一定意义的域名作为访问入口。然而,域名也被不良甚至恶意网站所利用,严重影响互联网的健康发展。本文从域名管理的角度出发,聚焦在承载涉黄涉赌等不良内容的网站所应用域名(简称不良应用域名)的早期(即在注册之际)识别。不良应用域名的早期识别可以更好地保护互联网用户的权益和保障互联网健康发展,对我国网络空间的治理也具有重要意义。

已有对不良应用域名识别的方法[1-4],通常是根据网页文本内容、图片内容等对域名对应的网站进行特征分析,训练模型以判断网站是否涉黄涉赌,进而判断域名是否被用于不良应用。这类检测方法准确性较好,但存在滞后性导致时效性不足。全球注册量最大的国家顶级域名,始终存在一定数量的.CN域名被用于承载涉黄涉赌等不良内容的网站。目前主要通过自动化内容检测手段发现潜在的不良应用域名,并辅以人工判定的方式,实现对不良应用域名的标注与管理。需要说明的是,不良应用域名与恶意域名尽管同属于域名滥用[5],但二者存在很大的不同。恶意域名通常不对应具体网站,所以其域名并不需要考虑易于记忆等因素,而重点考虑如何逃避检测,如使用域名生成算法(domain generation algorithm,DGA)产生大量恶意域名,而只有一个是用来与远程控制服务器连接。与此相反,不良应用域名对应具体网站,而且用户可能直接输入域名访问,所以域名的文本特性(如字符熵)等与恶意域

① 国家重点研发计划(2022YFB3103000)和国家自然科学区域联合重点基金(U20A20180,62072437)资助项目。

② 男,1979年生,博士生,正高级工程师;研究方向:网络安全,网络测量,互联网基础资源管理;E-mail:huanlei@cnnic.cn。

(收稿日期:2022-12-28)

名差异大。因此,已有针对恶意域名的识别方法并不能直接应用于不良应用域名。

本文研究不良应用域名的早期识别方法,旨在域名注册时,准确识别出将来可能被用于不良网站的域名,从而为域名管理者提前开展监测或治理提供支撑。该问题是有监督的二分类问题,即给定部分不良应用域名作为种子,对域名进行二分类,从而识别出不良应用域名和正常域名。为此,本文利用预训练语言模型基于 Transformer 的双向编码器(bi-directional encoder representation from transformers, BERT)提取域名的文本语义特征,并提出基于注意力机制融合域名的文本语义特征和注册特征(如注册商、生命周期等),实现对域名的分类。

本文的主要贡献如下。

(1)特征提取。从域名注册信息和域名文本语义 2 方面提取域名的特征,实现在域名注册阶段的特征提取,实现不良应用域名的早期检测。此外,为了准确全面提取文本语义特征,提出基于预训练语言模型 BERT 的域名文本语义特征提取方法。

(2)分类模型。设计基于注意力机制的域名分类方法。该方法使用注意力机制考虑不同特征作用差异,有效融合域名的注册信息和语义信息,并最终通过全连接神经网络实现域名的分类,识别出不良应用域名。

(3)数据与实验。从海量. CN 国家顶级域名中,提取 2021 年某时间段内发现并处置的涉黄涉赌网站域名以及正常网站域名(按工信部要求完成网站备案的域名)为实验数据集,实验结果表明所提方法分类准确率($F1$ 分数)达到 0.99,同时通过鲁棒性分析和消融实验进一步验证了所提方法的有效性。

1 相关工作

域名滥用(DN abuse),即域名的非正当使用,如域名被用于网络钓鱼欺诈、涉黄涉赌网站、垃圾邮件发送、恶意软件分发、僵尸网络控制等各种违法违规场景。近年来,互联网名称与数字地址分配机构(Internet Corporation for Assigned Names and Numbers,ICANN)所召开的历次全体大会均将域名滥用治理列为专门议题进行研讨,并开展了若干有意义的探索和尝试。此外,ICANN 还专门发起了面向全球的域名滥用活动报告项目,旨在面向各大顶级域名定期提供相应的域名滥用活动监测及评价服务。2020 年 8 月,欧盟委员会专门就域名滥用治理议题面向全社会发起公开招标,旨在评估域名滥用情况对于欧盟网络生态的影响情况及应对策略。国际安全事件应急响应小组论坛也专门设立域名滥用工作组,专门研究域名滥用治理相关议题。域名被用于涉黄涉赌类网站在我国网络管理的实践中属于域名滥用的一种,也可称之为涉黄涉赌不良应用。

传统不良应用域名检测发现主要有 2 种方式:一种是通过爬取网站的文本和图像数据,根据内容分析网站使用的域名是否涉及不良应用,网络赌博、淫秽色情等不良网站的网页内容存在高度相似性,利用图像相似性聚类和相似性搜索等技术[6],可以进行涉黄涉赌不良应用域名的检测。这种方法准确性较好但计算量大,而且域名被识别为不良应用域名时往往已经活跃了一段时间,时效性不高;另一种方法是通过网站注册域名自身特征信息进行分析识别,通过分析正常域名和不良应用域名(特别是机器产生的域名)在文本上的构成差异进行检测[7-8]。这种方法仅依赖域名文本结构特征,准确率不高且容易逃避,此外不良应用域名需要考虑便于记忆,所以通常不借助机器来产生域名。

在其他类型的域名滥用检测发现方面,如恶意域名监测,已有方法基于域名本身和域名解析数据提取特征并构建分类器。通过提取其结构特征、统计特征、语言学特征,并利用监督学习方法对域名进行分类,可以识别 DGA 产生的恶意域名[9]。通过聚类关联辨识出同一 DGA 或其变体生成的域名,然后分别提取每一个聚类集合中算法生成域名的 TTL(time-to-live)、解析 IP(Internet protocol)分布、归属、Who is 的更新、完整性及域名的活动历史特征等,利用支持向量机(support vector machine,SVM)分类器可以过滤出其中的恶意域名[10]。然而,与 DGA 域名不同,本文关注的不良应用域名需要考虑便于记忆以吸引更多的用户访问,因此其域名的字符分

布与正常域名差别并不像 DGA 域名与正常域名的差别那么大,因此 DGA 域名的识别方法并不适用于不良应用域名的识别。

近年来,深度学习方法被应用于域名的分类。文献[11]提出了一种利用长短期记忆(long-short term memory,LSTM)网络对域名进行分类的方法。文献[12]针对数据集不平衡的问题对该模型进行了改进。文献[13]在 LSTM 的基础上应用类对抗学习技术,对每一个域名实施字符级别的扰动,提升了对从未出现过的恶意域名的识别精度。仅依赖于字符特征的域名分类方法容易被攻击者逃避,为此研究者发现攻击者难以伪造域名系统(domain name system,DNS)流量中的隐藏联系,比如受相同攻击者感染的受害者倾向于查询相同或部分重叠的恶意域名集合,而未受感染的客户端几乎不查询这些域名;同时,由于经济上的限制,攻击者很难拥有大量的网络资源,因此他们对资源的重用也在恶意域名之间构建了内在关联。据此,研究人员提出通过构建域名-主机图、域名-IP 地址的异构图检测恶意域名[14-18]。但这种方法依赖于域名的请求数据,即在恶意域名活跃后才能实施检测。本文从国家顶级域管理的实际需求出发,聚焦在不良应用域名的早期发现,即在注册时检测。

在恶意域名的早期检测方面,研究人员考察了恶意域名和正常域名在注册商、注册历史、早期的 DNS 查询行为、域名结构特征等方面的差异性[19-21],但如何有效提取并融合这些特征实现对域名的准确分类,仍然具有较大的挑战。

2 方 法

本文提出了融合域名注册信息和域名文本语义信息的不良应用域名早期识别方法,具体流程如图 1所示。识别流程共分为 3 个模块,在特征提取模块,系统分别从域名注册信息和域名文本信息 2 个部分获得域名的局部特征;接着,系统结合注意力机制获得域名的向量表示;最后,系统将向量表示输入到全连接网络对域名进行分类。

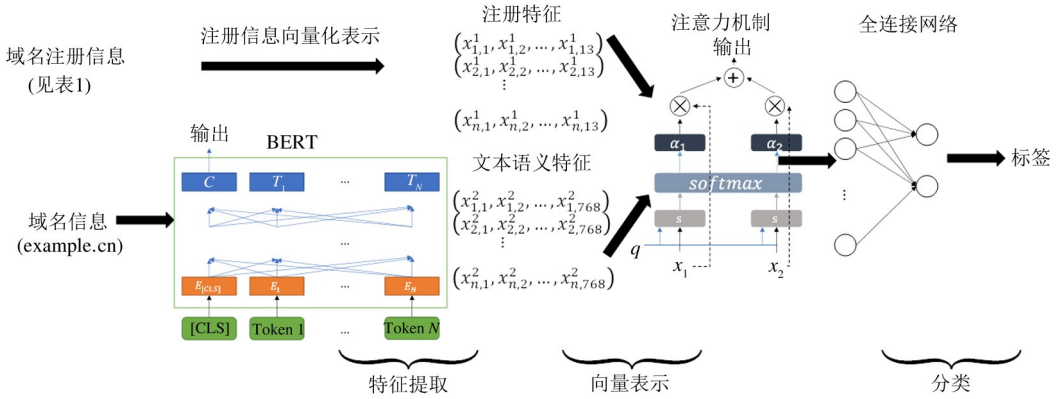


图 1 基于注册信息的不良应用域名早期识别流程

2.1 特征提取模块

本文从域名注册信息和域名文本语义 2 方面提取域名的特征,从而在域名注册阶段检测可疑的不良应用域名。

(1)注册时间。不良应用网站往往一次性注册大量不良应用域名,这一方面是因为需要足够数量的不良应用域名支撑其不良内容的承载(如定时更换所使用的域名),另一方面是出于经济成本原因(一次性注册大量域名时,单个域名成本低)。因此,注册时间是一个重要特征,可提取注册年份、注册月份、周几(day-of-week)、一个月中的第几天(day-of-month)、一年中的第几天(fay-of-year)五维特征来描述域名的注册时间。

(2)生命周期。由于注册生命周期更长的域名需要支付更高的费用,而现有检测机制下域名在用于不良应用后较短时间内往往就会被检测出来而无法继续使用,因此攻击者倾向于注册有效时长为最短年限的域名来节省成本。本文将过期时间和生命

时长作为 2 个重要特征。

(3)注册人和注册商。出于对成本和政策的考虑,不良应用网站往往会选择特定的注册商进行域名注册,因此本文将注册人和注册商也纳入域名的注册特征。

2.1.1　域名注册信息特征提取

在域名注册信息方面,提取了如表 1 所示的 13 维向量。

表 1　对域名注册信息提取的特征及其描述

特征名称	描述
reg _ year(int)	注册年份
reg _ month(int)	注册月份
reg _ day-of-month(int)	一个月中的第几天
reg _ day-of-week(int)	一周中的周几
reg _ day-of-year(int)	一年中的第几天
exp _ year(int)	过期年份
exp _ month(int)	过期月份
exp _ day-of-month(int)	一个月中的第几天
exp _ day-of-week(int)	一周中的周几
exp _ day-of-year(int)	一年中的第几天
duration	注册时间和过期时间的时间间隔
Registrant	每一个注册者对应一个整型,不同注册者对应的整型不一样
Registrar	同上

2.1.2　域名语义信息特征提取

在不良应用域名分类任务中,不良应用域名具有易记性、可读性和可解释性等特征,因此传统的语言和结构特征(如 N-gram,元音的比例、数字字符的比例等)往往无法准确区分不良应用域名和正常域名。同时,不良应用域名也会伪装成正常域名的样子,从而骗取用户点击,如 example. com 和 exanple. com 等。为此,本文采用域名的语义信息代替了域名的结构信息作为域名的文本特征。

本文选择使用 BERT 进行不良应用域名的语义表示。相较于在小规模数据集上训练网络的方法,在大型语料库上采用新定义的掩码语言模型(masked language model,MLM)进行预训练的 BERT 能够获得无偏见的单词级别的向量表示[22];相较于以往的单向语言模型(从左到右或者从右到左),BERT 采用深层的双向 Transformer[23]生成能融合左右上下文信息的深层双向语言表征。在预训练后,只需微调 BERT 模型的输出层,即可满足各种下游任务需求。由于 BERT 的输出是单词(token)级别的,因此本文将特殊标记[CLS]的 768 维输出作为域名的文本表示向量,这是因为[CLS]本身无语义信息,能更公平地融合文本中其他单词的语义信息。

2.2　向量表示模块

如图 2 所示,本文通过基于注意力机制的节点特征聚合操作来综合处理域名的注册信息和语义信息并生成域名节点的表征向量。

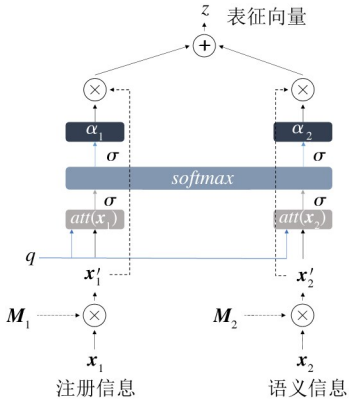


图 2　向量表示模块示意图

由于不同方面的特征维数不一样,所以本文采用线性变换矩阵 $\boldsymbol{M}_i$(该矩阵将作为神经网络的参数得到训练)将注册信息特征向量 $\boldsymbol{x}_1$ 和语义信息特征向量 $\boldsymbol{x}_2$ 映射到相同维度(各 768 维):

$$\boldsymbol{x}'_i = \boldsymbol{M}_i \cdot \boldsymbol{x}_i \tag{1}$$

此外,由于在域名向量表征中,不同源的特征对于识别和区分不同类型的域名具有不同的影响,因此本文提出了一种特征聚合方法,基于注意力机制来凸显它们之间的重要性差异。该方法将对来自不同特征源的特征向量进行加权求和,并选择性地筛选出一些重要的信息,从而使注意力集中在这些信息上。这些特征的注意力分数可以通过式(2)来进行计算。

$$\alpha_i = softmax(att(\boldsymbol{x}_i)) = \frac{\exp(att(\boldsymbol{x}'_i))}{\sum_{j=1}^{2}\exp(att(\boldsymbol{x}'_i))} \tag{2}$$

这里,引入 *softmax* 对原始计算的注意力分数进行归一化,并通过 *softmax* 的特性更加突出重要特征的权重。$att(\boldsymbol{x}_i)$ 为注意力打分函数,计算 $\boldsymbol{x}'_i$ 和 $\boldsymbol{q}$ 的相关性,常见的方法包括求两者的向量点积、向量余弦相似度等,本文采取了向量点积的方式,具体计算方式如式(3)所示。

$$att(\boldsymbol{x}_i) = \sigma(\boldsymbol{x}'_i \cdot \boldsymbol{q}) \tag{3}$$

其中,σ 代表 LeakyReLU 函数,$\boldsymbol{q}$ 代表注意力向量,是神经网络需要学习的一个参数。

最后,通过对两方面的特征加权聚合就得到了域名的 1 356 维表征向量:

$$z_i = \sigma\left(\sum_{j=1}^{2} \alpha_j \cdot \boldsymbol{x}'_i\right) \tag{4}$$

2.3 分类模块

获得域名的向量表示后,不良应用域名识别问题即转化为了二分类问题,本模块采用以线性整流函数(rectified linear unit,ReLU)为激活函数的全连接网络来进行分类,目标函数为交叉熵损失函数:

$$Loss = -\sum_{i=1}^{n} y_i \log y'_i \tag{5}$$

其中,y_i 为标签值,y'_i 为预测值。

综上域名分类器的训练过程如算法 1 所示。获得训练好的分类器后,即可对域名进行分类,从而识别出不良应用域名。

2.4 计算复杂度分析

参数量是指模型训练中需要训练的参数总数,用来衡量模型的大小,也即计算空间复杂度。

(1)全连接层。在特征映射和分类模块中,采用了全连接网络来进行线性变换,全连接层的参数包括权重矩阵和偏置矩阵,前者的参数量为 $dim_{\text{in}} \times dim_{\text{out}}$,后者的参数量为 dim_{out}。特征映射中,$dim_{\text{in}} = 13$,$dim_{\text{out}} = 768$,分类模块的全连接层中,$dim_{\text{in}} = 768$,$dim_{\text{out}} = 2$。

(2)注意力机制。在本文提出的基于注意力机制的特征聚合模块中,通过计算特征向量和注意力向量 $\boldsymbol{q}$ 的相关性给特征向量打分,$\boldsymbol{q}$ 是神经网络需要学习的参数,其大小就是特征向量的维数,在本文中维数为 768。

算法 1 神经网络参数更新流程

输入:域名特征集 batches = [{($\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$),…},…],迭代次数 epoch

输出:训练好的网络

```
for i = 1…epoch do
  for batch in batches do
    for (x_i1, x_i2) in batch do
      x'_i1 ← M_1 · x_i1
      x'_i2 ← M_2 · x_i2
      计算注意力分数
      加权聚合得到表征向量 z_i
      y'_i ← C · z_i
    end for
    计算交叉熵损失函数
    后向传播并更新参数
  end for
end for
return
```

3 实 验

3.1 实验设置

3.1.1 数据集

本文采集的数据集包括:

(1)2021 年 12 月发现并处置的被用于涉黄涉赌网站的不良应用.CN 域名;

(2)正常网站应用域名数据(按工信部要求完成网站备案的正常网站应用.CN 域名)。

实验所用数据集共包含 1 万个不良应用域名及其注册信息,以及 5 万个正常域名及其注册信息。域名信息示例如表 2 所示。需要说明的是,在数据集中,域名注册者通过去隐私化转换为字符串编号,每一个注册者对应一个唯一字符串 id,其中,对于注册者不在训练集中的新域名,该字段填充为已有最大注册者编号加 1;注册商为域名的注册服务商简称;注册日期 20210724 代表域名注册在 2021 年 7 月 24 日,到期日期 20220912 代表域名到期日期为 2022 年 9 月 12 日。

3.1.2 评价指标

评价指标如表 3 所示。考虑到在样本不均衡的场景下(不良应用域名的占比较低),仅使用准确率作为评价指标难以进行有效评估。另一方面,本文

表 2 域名信息示例

域名	注册者	注册商	注册日期	到期日期	是否为不良域名
w8680. cn	6331jna862f37y	万网	20210724	20220913	1
daeyeong. cn	hc656754346-cn	万网	20061114	20241114	0

表 3 评价指标及其描述

指标	描述
TP	被正确预测的不良应用域名数
FP	预测为不良应用域名的正常域名数
TN	被正确预测的正常域名数
FN	预测为正常的不良应用域名数
准确率	被正确预测的域名比例:(TP + TN)/(TP + FP + TN + FN)
精确率	预测为不良应用域名的样本中真实不良应用域名的比例:TP/(TP + FP)
召回率	真实不良应用域名中被正确预测为不良应用域名的样本比例:TP/(TP + FN)
*F*1 分数	2 × 精确率 × 召回率/(精确率 + 召回率)

提出的模型希望在注册阶段就尽可能地检测出可疑的不良应用域名,从而能对其后续网络活动进行更好地监测,及时发现其可能的恶意行为。因此,本文以不良应用域名为正样本,以 *F*1 分数和召回率为主,同时结合精确率和准确率作为评价指标。

3.2 实验结果对比

现有的不良应用域名检测方法主要可以分为 2 大类,一类基于域名本身和域名解析数据提取特征并构建分类器;另一类通过 DNS 流量构建域名-主机图、域名-IP 地址等异构图寻找恶意域名之间的联系。但依赖于域名解析数据的方法需要在恶意域名活跃后才能实施检测,本文旨在域名注册阶段进行不良应用域名的检测,因此本节选取了如下 4 种只依赖于域名注册信息的方法作为基线方法。

(1) FANCI (feature-based automated NXDomain classification and intelligence)[9]。它是一种经典的恶意域名检测方法,提取域名的结构特征、统计特征、语言学特征,并将 3 类 21 个特征直接连接形成 45 维的域名特征向量。FANCI 提取的结构特征包括:域名长度、子域名数目、子域名长度均值、是否拥有有效的 TLD、是否拥有 TLD 作为子域名等;提取的语言学特征包括:元音比例、不同字符的个数、重复出现的子符的比例、连续数字的比例等;提取的统计特征包括:熵值和 N-Gram。再通过支持向量机对域名进行分类。

(2)随机森林方法。它是在实际生产环境中被大量使用的经典机器学习方法,受到文献[9]方法中启发,本文选择了其他有代表性的有监督学习模型作为基线方法,输入和本文提出方法所选取的特征一致,即域名的注册信息向量和语义信息向量。

(3)决策树。与随机森林方法中所述相同,决策树的输入和本文提出方法所选取的特征一致,即域名的注册信息向量和语义信息向量。

(4)LSTM。文献[11]提出了一种利用 LSTM 网络对域名进行分类的方法,只依赖于域名的文本信息,故而也可以用以进行不良应用域名的早期识别。

本文按照 8∶2 的比例划分训练集与测试集,使用 Python3. 8 编程语言[24],并使用 Sklearn 第三方库实现上述机器学习方法。对于本文提出的方法,结合 Transformers 库获得 BERT 预训练网络,并在 PyTorch 深度学习框架上实现了后续训练的神经网络。

表 4 展示了本文提出的方法及上述基线方法的实验结果。(1)5 个方法都取得了很高的准确率,其中随机森林和本文方法的准确率都超过了 99%,同时随机森林、FANCI、LSTM 和本文方法都取得了超过 90% 的精确率。(2)对于真实的不良应用域名检测场景,考虑到不良应用域名占比极少,且危害性大,希望尽可能检测到所有的不良应用域名,因此希

望模型有更高的召回率。而在召回率方面,本文提出的方法表现远超另外 3 个方法。随机森林、决策树和 FANCI 的召回率都在 50% 左右,即仅有 50% 左右的不良应用域名被检测出来,而 LSTM 网络受到类别不平衡的影响更大,这是由于训练样本不平衡,模型更容易偏向样本量多的类别,以此来降低损失。而本文提出的方法获得了 0.999 0 的召回率,即 99.9% 的不良应用域名都可以被检出。(3)综合考虑精确率和召回率,可以看到本文方法的 *F*1 分数要远优于其他方法,这证明了本方法的有效性和可靠性。

表 4　实验结果对比

方法	准确率	精确率	召回率	*F*1 分数
随机森林	0.994 4	0.944 9	0.474 3	0.631 6
决策树	0.893 8	0.474 5	0.505 7	0.489 6
FANCI	0.947 5	0.964 6	0.466 7	0.629 0
LSTM	0.932 8	1.000 0	0.231 7	0.376 2
本文方法	0.991 5	0.997 9	0.999 0	0.994 6

3.3　鲁棒性分析

为了证明本方法在不良应用域名占比变化时的鲁棒性,本节对正常域名进行了不同程度的降采样,控制不良应用域名和正常域名的比例从 1∶1 变化到 1∶9,结果如表 5 所示。可以看到模型在所有比例的数据集下召回率都在 99% 以上,验证了本文方法的鲁棒性,即在不同比例下该方法都能检测出超过 99% 的不良应用域名。

表 5　模型在不良应用域名占比变化时的表现

不良应用域名:正常域名	准确率	精确率	召回率	*F*1 分数
1∶1	0.977 0	0.963 4	0.992 1	0.977 5
1∶2	0.984 3	0.959 5	0.995 0	0.976 9
1∶3	0.993 5	0.975 6	0.997 9	0.986 6
1∶4	0.998 6	0.997 9	0.994 9	0.996 4
1∶5	0.998 8	0.996 9	0.995 9	0.996 4
1∶6	0.999 1	0.996 9	0.996 9	0.996 9
1∶7	0.999 4	0.994 7	1.000 0	0.997 3
1∶8	0.999 3	0.997 9	0.995 7	0.996 8
1∶9	0.999 0	0.991 5	0.997 9	0.994 6

3.4　参数调节与分析

本节在不良应用域名和正常域名的比例为 1∶7 的数据集上(3.2 节中 *F*1 最高的数据集)对学习率(learning rate)和批大小(batchsize)进行了调参实验,实验结果如表 6 和表 7 所示。

(1)学习率设为 1×10^{-3} 时,会出现网络不收敛、损失(loss)震荡的情况,同时在表 5 中可以看到,学习率在 1×10^{-6} 时网络收敛慢,在相同训练轮次(epoch)时模型表现差于另外 2 种情况;在 1×10^{-4} 和 1×10^{-5} 时模型表现稳定且 *F*1 都在 0.99 以上,在后续实验中选择学习率为 1×10^{-5}。

(2)类似地,当批大小设为 16 时,会出现损失震荡的情况。批增大会使下降的方向更准确、震荡更小,减少随机性;但另一方面,批增大时,对网络参数的修正变慢,想达到同样精度所需的训练轮次会变多,因此在相同轮次的情况下,可以看到模型表现随批大小变大而变差,在后续实验中选择批大小为 64。

表 6　模型在学习率变化时的表现

学习率	准确率	精确率	召回率	*F*1 分数
1×10^{-4}	0.999 8	0.997 9	1.000 0	0.998 9
1×10^{-5}	0.999 4	0.994 7	1.000 0	0.997 3
1×10^{-6}	0.942 3	0.677 3	0.971 3	0.798 1

表 7　模型在批大小变化时的表现

批大小	准确率	精确率	召回率	*F*1 分数
32	0.996 8	1.000 0	0.999 6	0.998 4
64	0.994 7	1.000 0	0.999 4	0.997 3
96	0.869 4	0.985 1	0.980 9	0.923 7
128	0.686 9	0.940 4	0.942 6	0.793 9

3.5　消融实验

为了验证本文选取特征的有效性和必要性,本节进行了消融实验,表 8 展示了各简化版本模型的 *F*1 分数。

(1)w/o attribute(即不考虑注册信息):本实验只采用 BERT 的输出作为域名的特征向量,亦即后续神经网络的输入。

(2)w/o BERT(即不考虑域名文本语义特征):本实验只采用从域名信息中提取的 13 维属性向量

作为域名的特征向量。

(3)w/otime _ att(即不考虑注册信息中的时间信息):本实验采用从域名信息中提取的注册信息作为域名的特征向量(2 维)和 BERT 的输出作为域名的特征向量。

(4)w/oreg _ att(即不考虑注册者和注册商信息):本实验采用从域名信息中提取的时间信息作为域名的特征向量(11 维)和 BERT 的输出作为域名的特征向量。

(5)w/o registrar(即不考虑注册者信息):考虑到注册信息,尤其是注册者,是具有很强指向性的特征,本实验去掉注册者这一维特征向量,采用从域名信息中提取的时间信息和注册商信息(12 维)和 BERT 的输出作为域名的特征向量。

(6)w/o duration(即不考虑域名注册时的存活时间):考虑到在不良应用域名检测中,生命周期相较于注册时间、过期时间可能是更有区分度的特征,本实验去掉存活时间这一维特征向量,采用从域名信息中提取的注册时间信息、过期时间信息和注册信息(12 维)和 BERT 的输出作为域名的特征向量。

表 8　消融实验结果

不良应用域名:正常域名	w/o BERT	w/o attribute	w/o time _ att	w/o duration	w/o reg _ att	w/o registrar	Base
1∶1	0.530 0	0.565 8	0.954 1	0.671 1	0.942 2	0.626 7	0.977 5
1∶2	0.318 7	0.269 1	0.956 4	0.612 3	0.903 9	0.603 8	0.976 9
1∶3	0.085 5	0.133 0	0.939 0	0.539 4	0.893 9	0.772 4	0.986 6
1∶4	0.082 6	0.109 5	0.924 6	0.523 5	0.891 3	0.783 5	0.996 4
1∶5	0.062 4	0.062 2	0.905 0	0.564 3	0.861 2	0.821 8	0.996 4
1∶6	0.055 5	0.047 7	0.885 4	0.779 9	0.846 2	0.897 4	0.996 9
1∶7	0.050 7	0.062 2	0.861 6	0.809 5	0.846 5	0.963 8	0.997 3
1∶8	0.042 4	0.053 3	0.846 3	0.971 7	0.814 1	0.986 7	0.996 8
1∶9	0.033 6	0.051 0	0.832 1	0.979 9	0.797 3	0.966 5	0.994 6

通过实验(1)和(2)可以看出,在只用域名的文本特征或只用注册信息时,随着不良应用域名的占比减小,模型表现也剧烈下降,而结合两者后,模型表现很稳定,具有了很强的鲁棒性。

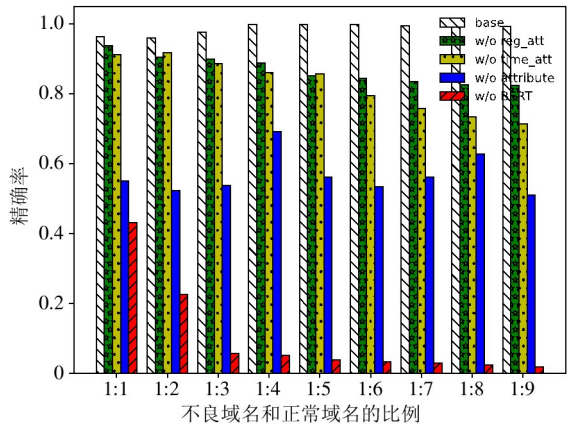
而实验(3)和(4)中,虽然模型表现仍会随着不良应用域名占比的减小而降低,但大部分都稳定在 0.80 以上,鲁棒性较前 2 组实验有大幅度提升,侧面验证了 2 种特征的结合是保障鲁棒性的来源。另一方面,第 3 组实验的模型表现普遍略高于第 4 组实验,这说明了注册人信息的贡献度高于时间信息。

实验(5)和(6)中,模型表现随不良应用域名占比减小呈现出先下降后上升的情况,这可能是因为在总数据量较小时域名存活信息和注册者信息较为重要,失去这 2 个信息会导致模型表现明显下降;但在数据量较大时模型能很好地根据如语义特征、注册时间特征等其他特征对不良应用域名进行识别。

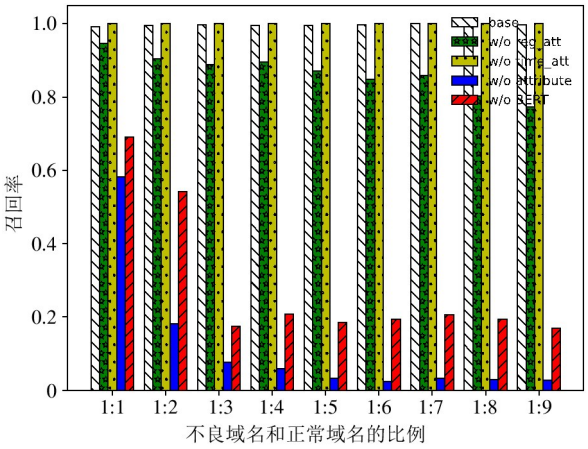
为进一步研究模型 *F*1 下降的原因,本节对前 4 组消融实验的精确率、召回率和准确率进行了比较和分析,结果如图 3 所示。

对于实验(1),可以看到其精确率和准确率随着不良应用域名占比下降而有所提高,当不良应用域名和正常域名的比例达到 1∶9 时,其模型准确率甚至超过了 90%。这是由于当样本不平衡时,模型会学习到训练集中样本比例的先验信息,而更倾向于将样本归为多数类(此时是正常域名),以至于多数类精度好而少数类精度差,召回率的剧烈下降也说明了这一点。也就是说,实验(1)能够较为精准地预测正常域名,但并不能有效地将不良应用域名识别出来。

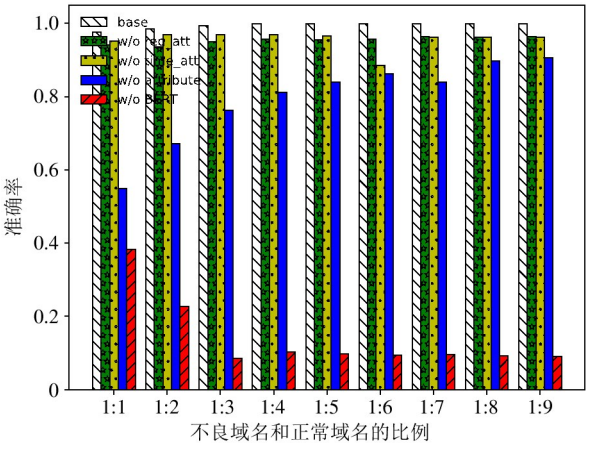
从实验(2)可以看到,其精确率、召回率、准确率都显著下降,说明此时模型已经不具有分辨不良应用域名的能力,这也进一步证明了采用文本选取特征的必要性。



(a) 模型精确率随不良应用域名占比减小的变化情况



(b) 模型召回率随不良应用域名占比减小的变化情况



(c) 模型准确率随不良应用域名占比减小的变化情况

图 3　4 组消融实验的精确率、召回率和准确率

对于实验(3)和(4),其精确率、召回率和准确率的变化趋势具有相似性,即:模型的准确率和召回率随不良应用域名占比减小变化不明显,具有一定的稳定性;而模型的精确率随不良应用域名占比减小呈明显的下降趋势,可以说模型 $F1$ 分数下降的来源正是其精确率的下降。也就是说,当正常域名占比增加时,正常域名被预测为不良应用域名的概率也会大幅增加,而不良应用域名仍然能够保持被识别。而实验(3)和(4)保持了原模型中的注意力机制和神经网络,区别只在于从域名信息中提取特征的维数,这从侧面证明了本文提出的模型的有效性。

4　结　论

本文从早期检测被用于涉黄涉赌网站的不良应用域名的实际需求出发,设计了一种基于深度学习的不良应用域名早期识别方法。该方法仅依赖注册时的信息,首先,从注册商、注册时间、有效时长等方面提取域名注册信息特征,并基于预训练模型 BERT 提取域名的语义信息;其次,使用基于注意力机制的特征聚合来综合处理域名的注册信息和语义信息并生成域名节点的表征向量;最后,通过基于全连接网络进行域名分类。本文方法在真实数据集上取得了 0.99 的 $F1$ 分数。此外,本文还对正常域名进行了不同程度的降采样,控制不良应用域名和正常域名的比例从 1∶1 变化到 1∶9,实验结果验证了本文方法的鲁棒性。综上所述,本文方法实现了高效、高准确率、高鲁棒性的不良应用域名早期检测。

参考文献

[1] WANG L, ZHANG Z, YE F. Pornographic and gambling domain recognition method based on long distance spare multi-head self-attention vision-and-language model[C]//2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications. Dalian, China: IEEE, 2022:504-508.

[2] HUANG Y, LIU D, YAN Z, et al. An abused webpage detection method based on screenshots text recognition [C]//Proceedings of the 2021 ACM International Conference on Intelligent Computing and its Emerging Applications. Jinan, China: Association for Computing Machinery, 2021:106-110.

[3] 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):3989-4009.

[4] LIU D, LEE J, WANG W, et al. Malicious websites detection via CNN based screenshot recognition[C]//2019 International Conference on Intelligent Computing and its Emerging Applications. Tainan, China: IEEE, 2019: 115-119.

[5] 樊昭杉, 王青, 刘俊荣, 等. 域名滥用行为检测技术综述[J]. 计算机研究与发展, 2022,59(11):2581-2605.

[6] 尚秋明, 王利军, 邓桂英,等. 一种不良应用域名快速核验方法的研究[J]. 电子技术应用, 2022,48(10): 72-77.

[7] YADAV S, REDDY A, REDDY A, et al. Detecting algorithmically generated malicious domain names[C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. New York, USA: Association for Computing Machinery, 2010:48-61.

[8] 张维维, 龚俭, 刘茜, 等. 基于词素特征的轻量级域名检测算法[J]. 软件学报, 2016,27(9):2348-2364.

[9] SCHUPPEN S, TEUBERT D, HERRMANN P, et al. FANCI: feature-based automated NXdomain classification and intelligence[C]//Proceedings of the 27th USENIX Conference on Security Symposium (SEC'18). Berkeley, USA: Association for Computing Machinery, 2018:1165-1181.

[10] 臧小东, 龚俭, 胡晓艳. 基于 AGD 的恶意域名检测[J]. 通信学报, 2018,39(7):15-25.

[11] WOODBRIDGE J, ANDERSONH , AHUJA A, et al. Predicting domain generation algorithms with long short-term memory networks [EB/OL]. (2016-12-02)[2023-01-15]. https://arxiv.org/pdf/1611.00791.pdf.

[12] TRAN D, MAC H, VAN T, et al. A LSTM based framework for handling multiclass imbalance in DGA botnet detection[J]. Neurocomputing, 2018,275:2401-2413.

[13] YILAMAZ I, SIRAJ A, ULYBYSHEV D. Improving DGA-based malicious domain classifiers for malware defense with adversarial machine learning[C]//2020 IEEE 4th Conference on Information & Communication Technology (CICT). Chennai, India:IEEE, 2020:1-6.

[14] KHALIL I, YU T, GUAN B. Discovering malicious domains through passive DNS data graph analysis[C]//Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. Xi'an, China : Association for Computing Machinery, 2016:663-674.

[15] 刘冰, 杨学, 杨琪,等. 一种基于图分析的不良网络应用快速发现算法[J]. 计算机应用与软件, 2022,39(11):329-336.

[16] 张斌, 廖仁杰. 基于关联信息提取的恶意域名检测方法[J]. 通信学报, 2021,42(10):162-172.

[17] SUN X, TONG M, YANG J, et al. HGDom: heterogeneous graph convolutional networks for malicious domain detection[C]//2020 Network Operations and Management Symposium. Budapest, Hungary:IEEE, 2020:1-9.

[18] SUN X, WANG Z, YANG J, et al. Deepdom: malicious domain detection with scalable and heterogeneous graph convolutional networks [J]. Computers & Security, 2020,99:102057.

[19] HAO S, KANTCHELIAN A, MILLER B, et al. PREDATOR: proactive recognition and elimination of domain abuse at time-of-registration[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: Association for Computing Machinery, 2016:1568-1579.

[20] HAO S, THOMAS M, PAXSON V, et al. Understanding the domain registration behavior of spammers[C]//Proceedings of the 2013 Conference on Internet Measurement Conference. New York, USA: Association for Computing Machinery, 2013:63-76.

[21] 田双柱, 陈勇, 延志伟,等. 基于多维度特征的不良网站检测[J]. 计算机系统应用, 2017,26(2):207-211.

[22] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis ,USA: NAACL, 2019:4171-4186.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing. Long Beach, USA: Association for Computing Machinery, 2017:6000-6010.

[24] PASZKE A, GROSS S, MASSA F, et al. Pytorch: an imperative style, high-performance deep learning library [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Association for Computing Machinery, 2019:

8026-8037.

A deep learning based approach for early detection of abused domain names

HU Anlei[*,**,****], TIAN Yu[*,****], CHEN Yong[**], LI Zhenyu[*,****], XIE Gaogang[***,****]

([*]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**]China Internet Network Information Center, Beijing 100190)

([***]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083)

([****]University of Chinese Academy of Sciences, Beijing 100049)

Abstract

The harmful websites rely on the domain name system (DNS) to achieve the dissemination of unhealthy content; these websites adversely affect the Internet's development. Therefore, it is of great importance for the DNS's operation and management to identify the domain names that correspond to the harmful website (i.e., the abused domain names) as early as possible and dispose of them accordingly. From the perspective of the country top-level domain (.CN) management, this paper focuses on the detection of the abused domain names at the registration stage. We find distinct differences between the abused domain names and normal domain names in terms of registration characteristics and text structure. Based on this observation, a deep learning based approach for early detection of abused domain names is proposed. Specifically, the proposed method first extracts the registration information features of the domain names, as well as the text semantic features of the domain names themselves using the pre-training bidirectional encoder representation from transformers (BERT). Next, the method leverages the attention mechanism to coordinate the two types of features. Finally, a fully connected neural network is used to construct the domain name classification model, where the output indicates whether a given domain is an abused one or not. Extensive experiments based on real-life network data show that the $F1$ score of the proposed method can reach as high as 0.99. The ablation results also demonstrate the effectiveness and necessity of using the selected features to construct the classification model.

Key words: domain name system(DNS), domain name classification, deep learning, pre-training model