基于卷积神经网络的暗网网页分类研究

洪良怡[1] 朱松林[2] 王轶骏[1] 薛 质[1]

[1](上海交通大学电子信息与电气工程学院 上海 200240)

[2](江苏省南通市公安局 江苏 南通 226001)

摘 要 在海量暗网网页中筛选敏感主题内容对执法部门具有重要意义。通过对 Freenet 等暗网网页文本特点和类别进行深入分析,提出基于 TextCNN 的暗网网页主题分类模型。模型根据暗网网页非标准化的语言特点进行数据预处理;使用预训练的词向量获得网页内容的表示,通过不同大小的卷积核进行卷积操作获得特征图像,使用最大池化函数获得最终的特征向量;对卷积网络进行正则化处理,使用 softmax 函数预测类别概率。实验结果表明,采用该方法精确率为 86.01%,召回率为 78.97%, Macro-F1 值为 82.33%,高于机器学习模型,能够有效解决暗网网页分类问题。

关键词 暗网 网页分类 卷积神经网络 机器学习

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2023.02.050

DARKNET WEBPAGE CLASSIFICATION BASED ON CONVOLUTIONAL NEURAL NETWORK

Hong Liangyi[1] Zhu Songlin[2] Wang Yijun[1] Xue Zhi[1]

[1](*School of Electric Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

[2](*Nantong Public Security Bureau, Nantong 226001, Jiangsu, China*)

Abstract It is critical for law enforcement departments to extract contents of specific topic from enormous amount of darknet webpages. After in-depth analysis on webpage texts of Freenet and other darknets, a darknet webpage topics classification model based on TextCNN is proposed. The model preprocessed the data according to the non-standardized language characteristics of darknet webpages, and then represented webpage tokens with pretrained word embeddings. The feature image was obtained by convolution operation with convolution kernels of different sizes, and the final feature vector was obtained by using the maximum pooling function. The convolution network was regularized, and the category probability was predicted by using Softmax function. The experimental results show that the model achieves precision at 86.01%, recall score at 78.97% and Macro-F1 score at 82.33%, higher than machine learning models, which can effectively solve the classification problem of darknet webpages.

Keywords Darknet Webpage classification CNN Machine learning

0 引 言

近年来随着对网络节点和身份信息的保护需求增长,大量匿名通信技术应运而生。暗网是由匿名用户产生的数据组成的,使用匿名、匿踪技术和特定软件才能访问的网络空间。通过路由中继技术以及通信中的数据加密,掩盖了用户上网地址以及暗网主机托管地址,难以追溯服务器端和客户端信息。因为不受监管、匿名和不可溯源等特点,暗网的网络空间中存在着大量非法出售、分享非法商品和数据的站点,包括伪造的证件、信用卡信息、枪支弹药、毒品以及泄露数据。例

收稿日期:2020-07-10。国家重点研发计划项目“网络空间安全”重点专项(2016QY01W0202)。**洪良怡**,硕士生,主研领域:网络空间安全。**朱松林**,工程师。**王轶骏**,讲师。**薛质**,教授。

如,2019 年 6 月至少 5 万条美国牌照数据被美国海关和边境保护局 CBP 技术分包商泄露在暗网上;7 月迈阿密和其他一些城市警方约 1 TB 执勤拍摄数据在暗网流传;12 月欺诈情报公司 Gemini Advisory 发现 850 家商店被盗的 3 000 万条支付卡数据被上传到在线网络犯罪市场 Joker's Stash[1]。

在暗网中非法论坛的行为研究方面,Alnabulsi 等[2]分析了三个暗网论坛中的犯罪类型,包括隐私、黑客、毒品、政治、革命、武器以及毒品。宋胜男[3]通过对暗网非法网站按照毒品交易、武器交易、信用交易、色情服务四种典型内容进行分类,根据法律条文按照危害程度进行排序。He 等[4]使用机器学习算法训练法律法规文本,用于暗网网络上违法内容分类。曹哲超等[5]提出了结合了页面标签特征和页面文本特征识别的重要站点筛选方法。

暗网网页文本呈现出数据量大、种类多样、分布不均、内容简略、富含非标准用语以及标注困难的特征,给面向暗网海量网页文本信息筛选需求的文本分类带来了巨大挑战。基准语料缺乏、扩展性差的问题使得近几年机器学习以及深度学习自然语言处理方面的成果难以直接应用在暗网网页文本分类问题上。

本文为解决暗网网页内容分类问题,提出一种基于卷积神经网络的模型,首先介绍暗网网页文本特征,然后介绍该模型,最后通过实验验证模型的有效性,并研究了不同实验参数对分类效果的影响。

1　暗网网页文本特征

1.1　HTML 标签

暗网网页文本具有普通网页文本同样的特征,即包括大量 HTML 标签,HTML 标签对于网页实际内容分析会造成一定干扰。

如图 1 所示,HTML 中 < style > 标签规定了前端显示的风格。但如果将这些文本误认为是网页本身的内容,就会给网页分类造成困惑,因此在分类模型的预处理阶段需要去除 HTML 标签内容。

```
    background: url("files/background.jpg");
}
#header, h1, h2 {
  text-align: center;
}
#content {
  width: 1100px;
  margin-left: auto;
  margin-right: auto;
}
```

图 1　暗网网页部分 HTML 标签

1.2　非标准用语

自然语言处理领域已有一些成熟的基准语料库以及预训练的词向量,但是直接应用在暗网网页分类中仍可能存在一些问题。

暗网网页文本在去除 HTML 标签后,一般还会包括拼写错误的单词、暗网网站名称等文本。对于拼写错误的单词可以进行纠正,对于暗网网站名称可以删除,这样可以避免影响分类的特征工程。

暗网网页上的内容和以往基准语料库有不重合的部分,尤其是非法交易网站的商品描述信息中的专有名词或特定表达难以在常规语料库中找到。此外,暗网网页上的俚语表达、拼写不规范的问题也增加了内容分析的困难。

例如在毒品交易暗网网页中,常常包含化学品的名称以及毒品名称的缩写,如果使用其他自然语言处理分类任务使用的基准语料库进行训练,将难以对其准确分类。因此需要使用合适的预训练词向量。

1.3　聚类分析

为了确定暗网网页分类的类别,本文使用简单的预处理、分词以及 K-means(K 均值)聚类的方法对暗网网页文本进行聚类分析,流程如图 2 所示。

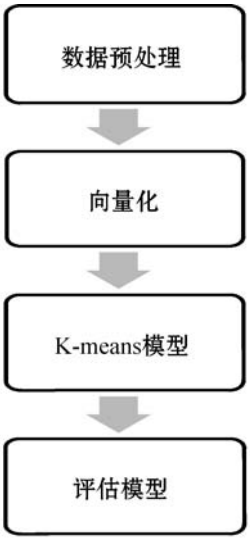

图 2　暗网网页内容聚类分析流程

数据预处理主要是以下操作:

首先清除网页文本中的 HTML 标签及其内容。将每个网页页面文本全部转换成小写,将文档词块化(tokenize)。文档词块化是把句子分割成词块(token)或有意义的字母序列的过程。使用 nltk 模块实现词干提取(stemming)和词形还原(lemmatization)。词干提取是将去除单词的前后缀得到词根。本文使用 Snowball 算法词干提取,Snowball 在 Porter 词干提取算法基础上增加优化,是目前较为推荐的词干提取方法。词

形还原是基于词典,将单词的复杂形态转变成最基础的形态。

数据预处理后,使用 TF-IDF 进行特征提取。TF(Term Frequency,词频)计算该词描述页面的能力,IDF(Inverse Term Frequency,逆文档频率)用于计算该词区分页面的能力。TF 和 IDF 的定义式如下:

$$TF = \frac{\text{这个词在该页面中出现的次数}}{\text{该页面的总词数}} \tag{1}$$

$$IDF = \log\left(\frac{\text{样本中的页面总数}}{\text{包含这个词的页面数}+1}\right) \tag{2}$$

$$TF - IDF = TF \times IDF \tag{3}$$

为防止包含这个词的页面数为零,使得分母为零的情况产生,因此对其分母加一。

一个词的 TF-IDF 的差值正比于在一个页面内出现次数,反比于在所有页面中出现的次数。TF-IDF 值越大,说明它是一个在其他页面较少出现,在当前页面出现较多的词汇,更能体现这个页面的主题。尽管该方法的精度不是很高且无法体现单词出现的位置信息,但是计算简便对于聚类分析来说足以。

K-means 是基于划分的聚类方法[6],因为在定义 K-means时,无法预先知道最优的簇群数量,即 K 值,所以本文通过遍历 K 值(从 2 到 20),将 min_df 和max_df 分别设为 10 和 0.5(忽略出现在 50% 以上网页中的词以及忽略超过 10 个网页中出现的词),设置最大特征词汇数为 1 000,并以轮廓系数(Silhouette Score)评价 K-means的分类效果。在多次实验中取得效果最好的 K 值为 15。每个类词频最高的 5 个单词如表 1 所示。

表 1 聚类结果

类别序号	词频最高的 5 个词
1	use, make, people, time, say
2	date, shoot, set, gallery, location
3	nigger, anonymous, fuck, god, right
4	use, key, site, file, link, set
5	say, like, know, make, think
6	verification, service, blog, forum, card
7	shall, state, unit, law, office
8	set, date, shoot, gallery, publish
9	model, art, studio, pose, nude
10	torrent, zip, report, list, property
11	com, net, org, free, list
12	key, use, file, download, insert
13	mar, work, net, want, think
14	set, date, shoot, publish, location
13	set, video, model, zip, key

可以看出聚类分析得到的分类中涵盖了不同话题,包括了论坛、图片音频、政治宗教等话题,但其中有一些簇的关键词基本一致,因此在实际进行文本分类时,分类类别还需要调整。

2 基于卷积神经网络的暗网网页分类

TextCNN[7] 在 2014 年被 Yoon Kim 提出,使用不同大小的卷积核来提取句子中的关键信息对句子完成文本情感二分类任务,其使用的词向量为预训练的 Word2Vec 向量。

本文针对 TextCNN 进行改进,模型主要分为 5 个层,分别是:数据预处理层、输入层、卷积层、池化层和输出层。结构如图 3 所示。

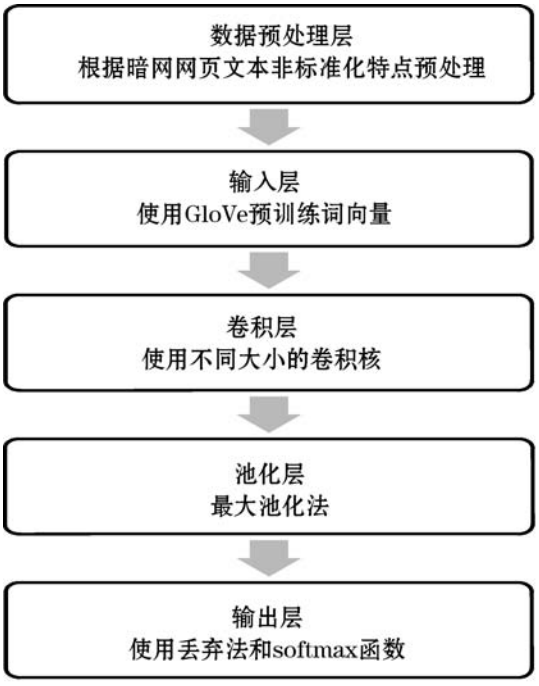

图 3 基于卷积神经网络的暗网网页分类模型结构

2.1 数据预处理层

传统的文本预处理包括分词、去除停用词以及低频词的工作。本文根据暗网网页文本数据的特征,对采集到的原始数据进行如下处理:1) 去除 HTML 标签及其内容,包括 < script > 、< style > 、< br > 、注释标签等;2) 除多余的空行;3) 去除非英文字母的字符;4) 字母转换为小写;5) 分词;6) 纠正拼写错误;7) 过滤停用词(stop-word),防止高频而无实际意义的词干扰内容分析;8) 使用 Python 第三方库 symspellpy 纠正拼写错误问题;9) 返回处理完的文本。

第 8 步是基于对大量暗网文本的观察提出的,暗网网页中经常出现拼写错误或者暗网网站名称(无意义的长文本),这给后续的特征工程带来困扰,因此有

必要在预处理阶段进行改正。

2.2　输入层

在词向量提出前,自然语言处理领域使用独热编码(one hot)的词袋模型表示单词。它的基本假设是词之间的语义和语法关系是相互独立的,因此从两个向量无法得到词的关系,不适合词汇语义的运算。另一个弊端是维度大,矩阵变得格外稀疏,会耗费大量计算资源。

词向量这个概念首先由 Hinton[8] 在 1986 年提出,使用稠密向量的表示方式,改进了独热编码维度爆炸的缺点。词向量生成方式主要有两种:静态向量和动态向量。静态向量包括 Word2Vec[9-10]、FastText[11]、GloVe[12]等,动态向量包括 ELMO[13]、BERT[14]等。

动态向量更加充分利用了上下文信息,因此能解决一词多义的问题,其中 BERT 在多个 NLP 任务中表现优异,但其对机器资源的要求较高,难以在实际业务场景中应用。因此,本文使用静态向量中的 GloVe 将文本转换为数字表达。

GloVe 是 Jeffrey Pennington 在 2014 年提出的新的词向量模型,通过构造词的共现矩阵,使用全局语料进行训练,改进了 2013 年由 Tomas Mikolov 发表的 Word2Vec 只利用窗口内语料的缺点。

本文首先获得第 i 个单词的 k 维词向量表示 $\boldsymbol{x}_i$。长度为 n 的网页在连接后被表示为:

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \cdots \oplus \boldsymbol{x}_n \tag{4}$$

2.3　卷积层

设单词个数为 n,词向量一共有 d 维,则得到 $n \times d$ 的矩阵。过滤器(filter)w 的宽度是 d(即词向量的长度),高度是超参数 h(即卷积核的大小)。本文使用的卷积核大小为 2、3、4,输出通道数都是 100。

通过对第 i 个单词到第 $i+h-1$ 个单词的窗口进行卷积操作获得特征 c_i。

$$c_i = f(w \cdot x_{i:i+h-1} + b) \tag{5}$$

式中:b 为偏差,f 是激活函数,如双曲正切函数。

将该卷积核应用在句子上的每个单词窗口上,可以得到特征图像(feature map)。

$$\boldsymbol{c} = [c_1, c_2, \cdots, c_{n-h+1}] \tag{6}$$

通过使用宽度不同的卷积核,可以得到多张特征图像,捕捉到不同个数的相邻单词的相关性,将结果送到池化层。

2.4　池化层

为了获得每个特征图像中最重要的特征,对特征图像使用最大池化函数。

$$\hat{c} = \max\{c\} \tag{7}$$

本文选择的最大池化法就是从一维的特征图像中提取最大的值,获得最重要的特征。假设有 m 个过滤器,则最终的特征向量为 $\boldsymbol{z}$。

$$\boldsymbol{z} = [\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_m] \tag{8}$$

得到最终的一维特征向量后,输入一个全连接的分类层进行分类预测。

2.5　输出层

通过 dropout(丢弃法)使卷积神经网络正则化,使用 0.5 的丢弃率随机丢弃一些神经元,迫使神经元学习有用的特征,避免过拟合。设置为 0.5 时,dropout 随机生成的网络结构最多。

因为本文研究的是多分类问题,因此在输出层设置使用 softmax 函数预测类别概率。

$$y = \text{softmax}(W^{(S)} z + b) \tag{9}$$

3　实验与分析

3.1　实验环境

本文的实验环境是:编程语言为 Python 3.7.3,深度学习框架为 PyTorch 1.4.0,内存 8 GB,处理器为 Intel Core i5。

3.2　实验数据

爬取了 Freenet、ZeroNet 和 Tor 中 11 381 个站点网页文本作为实验语料,使用 langid 库筛选英文页面,去除预处理后无文本信息的网页后,参考聚类分析的大致分类对网页进行人工标注,共分为博客、商业、暗网服务介绍、论坛、黑客技术、商店市场、政治宗教、色情、编程讨论、枪支毒品、服务托管、代笔枪手 12 个类别,共 7 665 篇网页文本。

实验数据集按照 8∶1∶1的比例分为训练集、验证集和测试集。每次实验的结果选取在验证集上表现最好的模型在测试集上的表现,最终结果是 10 次随机划分数据集获得的实验结果的平均值。

3.3　实验评价标准

在实验结果的评估方法上,本文选取了多分类任务中常用的精确率(Precision)、召回率(Recall)以及 Macro-F1 值(macro-fscore)作为指标,以此评估模型表现效果。此外,为了体现模型对于类别分布不均衡的效果,将 Macro-F1 值作为判断分类算法性能的关键指标,计算时每个类别有着同样的影响。

它们在 c_i 上的值为:

$$P_i = \frac{a_i}{b_i} \tag{10}$$

$$R_i = \frac{a_i}{d_i} \tag{11}$$

$$Macro\text{-}P = \frac{\sum_m P_i}{m} \tag{12}$$

$$Macro\text{-}R = \frac{\sum_m R_i}{m} \tag{13}$$

$$Macro\text{-}F1 = \frac{2P \cdot R}{P + R} \tag{14}$$

式中：a_i是其中被正确判断为 c_i类的文档数，b_i是被系统判断为 c_i类的所有文档数，d_i是应属于 c_i类的所有文档数，m 是分类个数。P_i是在 c_i类上的精确率，R_i是在 c_i类上的召回率，Macro-P 是总体的精确率，Macro-R 是总体的召回率，Macro-F1 是总体 F1 值。

3.4 实验设计

（1）与机器学习算法比较。在同样的数据集上，和机器学习中的 Linear SVC（线性支持向量机分类器）[15]、Multinomial Naïve Bayes（多项式朴素贝叶斯算法）[16]及 KNN（K-Nearest Neighbor，K 最近邻算法）[17]进行了对比。

线性支持向量机分类算法根据训练样本的分布，确定最佳的线性分类器。被称为支持向量的是训练数据中两个空间间隔最小的两个不同类别的数据点，可以帮助确定最优线性分类器。

朴素贝叶斯分类器基于贝叶斯理论，朴素即单独考虑模型中的各个特征，不考虑特征之间的相关性。多项式朴素贝叶斯算法的特征变量是离散变量，网页分类中特征变量体现在一个单词出现的次数或 TF-IDF 值等。但是弊端是无法将各个特征之间的联系考虑在内，因此在数据特征关联性强的分类任务上表现不佳。

KNN 算法寻找与待分类样本在特征空间中距离最近的 K 个已标记样本，则该样本属于这个类别。随着 K 的不同，分类的效果也会不同。KNN 算法的本质是对特征空间的划分，它在类别决策时只与少数相邻样本有关。

实验中均使用 GloVe 预训练的 300 维词向量。

（2）不同卷积核大小比较。为了确定卷积核对分类效果的影响，本文在其他参数一致的情况下，对卷积核大小进行对比实验。实验中卷积核大小为 2、3、4。实验中均使用 GloVe 预训练的 300 维词向量。

（3）不同词向量比较。在其他参数一致的情况下，对使用暗网文本生成的随机词向量和使用 GloVe 预训练的词向量进行对比实验。词向量维度为 50、100、200、300。

3.5 实验参数

实验数据如表 2 所示，本文使用的预训练词向量为 GloVe 的 40 万个词的 300 维词向量。Adam 算法为目前主流的优化算法，对梯度的一阶矩估计（First Moment Estimation）和二阶矩估计（Second Moment Estimation）进行综合考虑，计算出更新步长。

根据经验，本文在训练时将每批训练大小（batch size）设置为 32，迭代 50 次，丢弃率设置为 0.5，学习率设为 0.001。

此外，本文用 Focal Loss 函数来衡量模型的损失 2017 年该方法在目标识别领域被提出[18]，可以解决分类不平衡以及分类难度差异的问题。

表 2 实验参数

参数名称	参数值
预训练词向量	GloVe.6B.300d
优化器	Adam（Adaptive Moment Estimation）
每批训练大小	32
迭代次数	50
丢弃率	0.5
学习率	0.001

3.6 结果分析

（1）与机器学习算法比较。表 3 中显示了本文模型与机器学习算法对比结果。

表 3 卷积神经网络模型与机器学习算法的分类结果（%）

模型	Precison	Recall	Macro-F1
TextCNN	86.01	78.97	82.33
Linear SVC	83.78	76.52	75.61
Multinomial NB	87.30	86.14	75.41
KNN	74.98	68.57	57.97

KNN 算法的 Macro-F1 值仅为 57.97%，这是由于 KNN 对训练数据的依赖度大，对噪声数据过于敏感，如果有几处数据错误就会带来预测数据的不准确。机器学习算法中 Linear SVC 和 Multinomial Naïve Bayes 结果接近，尽管相比 KNN 速度较慢，但二者的精确率和召回率都很高。本文模型 Macro-F1 值比机器学习算法高出了至少 6%，证明了本文算法对在暗网网页分类方面效果良好。

（2）不同卷积核大小比较。表 4 显示在不同大小的卷积核下 TextCNN 分类结果的变化。

表 4　不同卷积核大小的分类结果(%)

卷积核大小	Precison	Recall	Macro-F1
4	84.63	76.93	80.58
3,4	85.21	77.78	81.32
2,3,4	86.01	78.97	82.33
3,4,5	85.55	77.99	81.58

实验结果表明卷积核个数和大小对 TextCNN 准确率影响较大,在 1% 内波动。可以看出,使用单个卷积核分类结果的精确率和召回率都是最低的,因此 Macro-F1 也是最低的。当增加卷积核个数时,分类结果得到改善。当卷积核个数设定为 3 时,根据卷积核大小分别为 2、3、4 以及 3、4、5 的两项实验结果,得出卷积核大小设置为 2、3、4 时结果最好,即本文选择提取不同 n 元组(n-gram)词组的最佳设定为 2 元组、3 元组及 4 元组的总体组合,捕捉到了上下文对单词的约束关系。

(3) 不同词向量比较。如表 5 所示,从选择的词向量角度看,使用 300 维 GloVe 预训练词向量的分类结果是最好的。

表 5　不同词向量的分类结果

词向量类型	词典大小	Precison/%	Recall/%	Macro-F1/%
随机 50 维	22 801	85.47	77.60	81.33
随机 100 维	22 801	84.73	78.86	81.69
随机 200 维	22 801	85.53	78.66	81.94
随机 300 维	22 801	86.06	78.30	81.99
预训练 50 维	40 097	85.37	78.25	81.64
预训练 100 维	40 097	85.70	78.95	82.18
预训练 200 维	40 097	86.06	78.46	82.06
预训练 300 维	40 097	86.01	78.97	82.33

使用 GloVe 预训练词向量略高于仅使用暗网文本生成的随机词向量结果,分析原因是暗网文本生成随机词向量时词典大小较大,因此 GloVe 预训练词向量的优势不明显,但仍能看出预训练词向量效果更好,Macro-F1 值高达 82.33%。

从词向量维度看,词向量维度对精确率和 Macro-F1 值有一定影响。对比随机词向量 Macro-F1 值结果,可以发现 Macro-F1 值随着维度增加而增加,300 维比 50 维的结果高出 0.66 百分点;此外 Precision 值也有递增趋势,300 维比 50 维的结果高出 0.59 百分点。在 GloVe 预训练词向量结果中,Macro-F1 值同样随着维度增加而增加,300 维的 Macro-F1 值最高,200 维的 Macro-F1 值略低于 100 维,与预期不同;Precision 最大值在 200 维取到,但仅比 300 维时高出 0.05 百分点。

本文要解决的是暗网网页文本的分类,词向量与一般文本有很大不同,例如在暗网中 set 更多指资源集的集,而非设置。除此之外,还有一些暗网中常用的缩写等预训练词向量中缺少的词,可能会影响分类结果。因此本文使用了 GloVe 预训练向量,其原始语料来自维基百科和 Gigaword。这样缓解了暗网文本数据集数据不够单独学习得到词向量矩阵的困难,同时可以应对过拟合问题。

4　结　语

传统的卷积神经网络在自然语言处理中能够取得不错的效果,但为了应用在暗网网页分类任务上,本文基于 TextCNN 提出了暗网网页分类的算法。在数据预处理层,将拼写错误的单词预先纠正,解决了暗网网页文本中出现频繁的拼写错误等问题。词向量方面比较了 GloVe 预训练词向量与文本生成随机词向量的方法,引入了预训练词向量提高分类效果。实验表明,本文模型在暗网网页数据集上分类效果优于其他模型,验证了本文方法的有效性。

实验过程中,本文所用的数据是自行爬取并标注的暗网网页文本数据集,通过聚类分析以及人工标注决定了预设的分类类别。如何提高这一过程的效率以及改进预设的分类类别将是下一步研究的重点。未来可以探索的方向主要有以下几个:

1) 池化层改进:使用 K-max pooling(K-最大池化)方法,避免最大池化丢失一些重要特征的问题。

2) 模型组合:通过组合 Text-RNN 和 Text-CNN,捕获网页更全局的特征,且不会产生性能问题。

参　考　文　献

[1] Kevin C, Sergio H. At least 50,000 license plates leaked in hack of border contractor not authorized to retain them[EB/OL]. [2023 - 01 - 01]. https://edition. cnn. com/2019/06/17/politics/customs-and-border-protection-data-breach-license-plates-leaked/index. html.

[2] Alnabulsi H, Islam R. Identification of illegal forum activities inside the dark net[C]//International Conference on Machine Learning & Data Engineering,2018.

[3] 宋胜男. 暗网域名收集与内容分析方法研究[D]. 北京:北京交通大学, 2019.

[4] He S, He Y, Li M. Classification of illegal activities on the dark web[C]//Proceedings of the 2019 2nd International Conference on Information Science and Systems. ACM,2019: 73 - 78.

(下转第 330 页)

3.2 SM4 加密算法保护数据安全

SM4 是我国商用分组密码算法。在商用密码体系中,SM4 是一种对称加密算法,其算法公开,分组长度与密钥长度均为 128 bit,加密算法与密钥扩展算法都采用 32 轮非线性迭代结构,S 盒为固定的 8 bit 输入 8 bit 输出[7-8]。

安全传输系统采用 HTTPS 进行随机数交换,交换后的随机数异或后采用 SM4 加密后得到会话密钥,系统采用 ECB 模式进行传输数据的 SM4 加密。

4 结 语

针对某政务系统数据上报、下发的需求,采用多层架构设计及 TCP/IP 交互协议,设计开发了基于 SSL 及国密算法的数据安全传输系统。通过研究 SSL、SM3、SM4、防中间人攻击、防重放攻击等关键技术,设计一套通过 SSL 通道交换预分配密钥加密的随机数,完成系统间双向握手,建立会话过程的系统实现机制。采用 ActiveMQ、MyBatis、Durid、Redis、业务集群等技术,有效保证了系统的稳定高效传输。系统采用国密算法 SM4,既保证数据的安全,又因为采用对称加密算法,保证了数据加解密传输效率。通过时间戳有效防止应用数据重放攻击,通过杂凑值有效防止应用数据中间人攻击,增强了系统的安全性,达到了数据安全传输的设计目的。

参 考 文 献

[1] 韦俊琳,段海新,万涛. HTTPS/TLS 协议设计和实现中的安全缺陷综述[J]. 信息安全学报,2018,3(2):1 - 15.

[2] 康荣保,张玲,兰昆. SSL 中间人攻击分析与防范[J]. 信息安全与通信保密,2010(3):85 - 87,90.

[3] 金敏捷,秦飞龙. 基于中间人攻击的 SSL 防范对策探究[J]. 船舶,2017,28(4):92 - 94.

[4] 陈宇琦. 一种基于时间戳的无线射频重放攻击抵御方案[J]. 现代计算机,2012(6):24 - 25,29.

[5] 肖斌斌,徐雨明. 基于双重验证的抗重放攻击方案[J]. 计算机工程,2017,43(5):115 - 120,128.

[6] 徐静,常朝稳. SSL 协议的安全性分析[J]. 微计算机信息,2006,22(9):19 - 21.

[7] 杨润东,李子臣. 基于国密算法的新型电子邮件加密系统研究与实现[J]. 信息安全研究,2018,4(11):1046 - 1051.

[8] 伍娟. 基于国密 SM4 和 SM2 的混合密码算法研究与实现[J]. 软件导刊,2013(8):127 - 130.

(上接第 325 页)

[5] 曹哲超,王轶骏,薛质. 基于页面标签和文本特征的暗网重要站点识别[J]. 通信技术,2019,52(12):3021 - 3026.

[6] MacQueen, J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of Berkeley Symposium on Mathematical Statistics & Probability,1965.

[7] Kim Y. Convolutional neural networks for sentence classification[EB/OL]. [2023 - 01 - 01]. https://arxiv. org/abs/1408. 5882.

[8] Hinton G E. Distributed Representations[M]. Cambridge: MIT Press, 1986.

[9] Mikolov T, Chen K, Corrado G S, et al. Efficient estimation of word representations in vector space[EB/OL]. [2023 - 01 - 01]. https://arxiv. org/abs/1301. 3781? ref = hackernoon. com.

[10] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013.

[11] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,2017.

[12] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),2014.

[13] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[EB/OL]. [2023 - 01 - 01]. https://arxiv. org/abs/1802. 05365.

[14] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2023 - 01 - 01]. https://arxiv. org/abs/1810. 04805v1.

[15] Dumais S T, Chen H. Hierarchical classification of web content[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval,2000:256 - 263.

[16] Frank E, Bouckaert R R. Naive bayes for text classification with unbalanced classes[C]//European Conference on Principles of Data Mining and Knowledge Discovery, 2006: 503 - 510.

[17] Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based Framework for Text Categorizatio [J]. Procedia Engineering, 2014,69: 1356 - 1364.

[18] Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//International Conference on Computer Vision, 2017: 2999 - 3007.