引用格式：张鹏，赵文浦，秦瑞青，张雱阳，顾严齐天．基于深度学习的恶意社交机器人识别研究[J]．中国传媒大学学报（自然科学版），2024，31(05)：23-31．
文章编号：1673-4793(2024)05-0023-09

基于深度学习的恶意社交机器人识别研究

张鹏*，赵文浦，秦瑞青，张雱阳，顾严齐天
（中国人民警察大学网络舆情治理研究中心，廊坊 065000）

摘要：基于账户特征开展社交机器人检测，构建了一种用于识别恶意社交机器人账号的深度学习分类模型。该模型由五个层组成：一个包含10个神经元的输入全连接层、两个各含128个神经元的全连接层、一个Dropout层，以及一个输出全连接层。模型训练过程中，采用了多种激活函数，并结合Adam优化器进行优化。通过与其他四种基于机器学习的模型进行对比实验，验证了所提模型的有效性。本文深度学习模型在F1值方面优于其他模型，且准确率达到了次高水平。值得一提的是，基于随机森林构建的社交机器人识别模型在准确率等指标上也优于其他主流机器学习方法，展现出良好的性能。综上所述，深度学习技术在社交机器人识别的实验中表现出卓越的性能，能够满足实际研究的需求，可应用于社交平台机器人账号检测的实际场景。

关键词：恶意社交机器人；深度学习；在线社交网络；识别检测
中图分类号：TP391　**文献标识码：**A

Research on malicious social robot recognition based on deep learning

ZHANG Peng*, ZHAO Wenpu, QIN Ruiqing, ZHANG Jiyang, GUYAN Qitian
(Research Center for Network Public Opinion Governance, China People's Police University, Langfang 065000, China)

Abstract: We carried out social bot detection based on account features and constructed a deep learning classification model for identifying malicious social bot accounts. The model consisted of five layers: an input fully connected layer containing 10 neurons, two fully connected layers containing 128 neurons each, a dropout layer, and an output fully connected layer. Multiple activation functions were used during model training and optimized in conjunction with the Adam optimizer. The effectiveness of the proposed model was verified by comparison experiments with four other machine learning based models. The deep learning model proposed in this experiment outperforms other models in terms of F1 value and reaches the next highest level of accuracy. It is worth mentioning that the social robot recognition model constructed based on random forest also outperforms other mainstream machine learning methods in terms of accuracy and other metrics, showing good performance. In summary, the deep learning technique shows excellent performance in the experiments of social robot recognition, which can meet the needs of practical research and can be applied to practical scenarios of robot account detection on social platforms.

Keywords: malicious social robots; deep learning; online social networks; identification detection

基金项目：教育部人文社会科学研究规划基金（22YJA860012）；警察大学科研重点专项课题（ZDZX202201）
作者简介（*为通讯作者）：张鹏（1981-），男，博士，副教授，硕士生导师，主要从事网络舆情与危机管理研究。Email：zhangpeng@cppu.edu.cn；赵文浦（2000-），本科生，主要从事数据警务技术研究。Email：3078115700@qq.com；秦瑞青（1999-），女，硕士生，主要从事智慧警务与大数据技术研究。Email：qinruiqing@alu.sxu.edu.cn；张雱阳（1999-），男，硕士生，主要从事智慧警务与大数据技术研究。Email：rachel0701@email.cn；顾严齐天（2004-），男，本科生，主要从事智能计算研究。Email：gyqt123456@outlook.com

1 引言

在线社交网络(OSN, online social networks)也称社交网站(SNS, social networks sites)或社交媒体网络(SMN, social media networks),是一种在线网络平台,具有帮助拥有共同爱好、习惯和生活方式的人群建立社交关系的功能特点[1]。近年来,Twitter、微博等社交网络平台上出现了大量具有攻击属性的社交机器人账号,这些机器人账号能够通过左右国际舆论场上的言论走向,对我国相关的舆论事件生成大量负面言论,进而借助推动舆论战影响我国国际形象。利用自动化程序、人工智能等相关技术模仿人类行为,传播不良信息、开展虚假宣传,甚至煽动言论、引发对立情绪[2]。最新权威统计数据,截止到2024年四月份,Twitter总计发现了大约有3000万个可疑账户,此数据相较于2023年年末显著增加了近20倍[3]。通过深入的分析表明,社交机器人所衍生出的各类虚假信息在Twitter平台占全部媒体内容的35%[4]。2019年Facebook平台上的活跃账户中机器人平均存在率达到11%,且数量与比例受政治或利益驱使呈增长趋势[5]。而当前我国正处于改革转型的攻坚期,恶意社交机器人的存在对网络空间安全和舆论安全带来了严重的危害,并成为了在线社交网络中最复杂和最高级的安全威胁之一。与此同时,伴随机器人攻击策略的不断演变,目前主流的基于人工识别和机器学习的发现方法对恶意社交机器人的识别效果不断下降。为有效防范不法分子的网络攻击及负面舆情引导,保护公民合法信息安全、减少社会恶劣影响,探索治理恶意社交机器人的识别方法成为了当下亟待解决的问题。

据此,本文提出了一种基于深度学习的恶意社交机器人识别方法,通过搭建神经网络模型,动态应对恶意社交机器人随时间推移的攻击策略演化[6]。同时实践当前主流的机器学习相关算法,并对比多个模型的训练结果,探索不同算法对恶意社交机器人识别的表现,以期验证提出的识别模型的可行性与有效性。本文研究结果一方面有助于提升对恶意社交机器人检测的长期能力,营造健康良好的社交网络平台环境;另一方面能帮助研究者了解不同模型的优劣势与适用场景,从而更好地选择合适的模型进行研究,为恶意社交机器人识别领域的后续研究提供参考。

2 研究现状

恶意社交机器人是指通过自动化程序或算法,模拟人类社交行为,以达到欺骗、操纵、传播虚假信息等目的的计算机程序。近年来,随着社交媒体的普及和人工智能技术的发展,恶意社交机器人检测成为了一个备受关注的研究方向,主要识别方法可以分为以下几类[7]:众包社交机器账号识别平台、基于传统机器学习的识别技术、基于深度学习的识别技术和基于图结构的识别技术。

2.1 众包社交机器账号识别平台

Gamallo P等[8]明确提出了一类众包社交机器账号检测平台。研究者强调识别机器账号相较于人类而言,其操作复杂性相对较低,所以创建了在线平台,此平台依托雇用许多专家及人工,对社交网络中的在线账号予以资料分析及判断,平台会将账号资料提供给多个工作人员,并会把多数工作人员的具体意见视为是最终的判定结果。总体来讲,该识别方法浪费大量人力与时间,具有很大局限性;同时社交机器人依靠人工智能技术可以快速、大量产生,成本低廉,基于众包的识别方法明显不能够满足现实的社交机器人检测需求。

2.2 基于机器学习的识别技术

基于机器学习的机器账号识别技术从本质上而言是把恶意社交机器人识别视为一个二分类问题,这项技术凭借对账号特征进行准确的提取,借助分类算法对数据实施建模,由此创建出相应的识别模型,同时借助该模型对需要划分类型的账号加以准确的类型划分。Botometer[9]是一个2014年创建的在线机器账号检测平台,能够对提供的推特账号予以合理的评分,最终的分数越高意味着该账号为机器账号的几率越大。文献[10]借助N-grams来检测机器账号,凭借对推文内容实施严格的语义剖析来判定推文作者是不是机器账号。文献[11]借助朴素贝叶斯算法来检测机器账号,但通过对多个分类器实验加以比较后发现,朴素贝叶斯所获得的结果始终差强人意。文献[12]通过实验证实了随机森林算法在分类器中效果最好。

2.3 基于深度学习的识别技术

与传统的机器学习方法不同,深度学习需要更多

的数据和时间来训练模型，此外它还可以借助无监督亦或是半监督的特征进行学习。同时，借助分层的特征提取算法对手动获取特征加以取代。上述方法能够在很大程度上缩短模型训练时间，同时挖掘出某些隐性的特征。文献[13]把卷积神经网络与LSTM (LSTM, long short-term memory)模型综合起来用来检测机器账号检测；文献[14]借助推特内容还有元数据创建了相应的模型，通过该模型能够在推文级别对社交机器人账号加以准确识别。该模型通过用户元数据对上下文特征加以准确提取，同时把该特征确立为辅助输入并提供给否则对推文进行处理的LSTM网络，仅需一条推文便能够判定是不是机器账号。文献[15]将异构图神经网络应用到了恶意账户的检测之中。

2.4 基于图结构的识别技术

基于图结构的识别技术的核心在于利用社交网络形成的用户关系图，该图可用于理解和分析社交网络平台上用户之间的联系。因此，技术的重点是关注用户之间的关系。文献[16]设计了以随机游走为基础的检测模型SybilWalk，在无向社交图当中实施随机游走。除此之外，Wang [17]、Feng[18]研究均是以社交关系图为基础创建的检测机器账号的相关方法。

3 深度学习理论和实践

深度学习是一种能处理大规模、高维度数据的机器学习方法，相对传统机器学习算法深度学习更加灵活、高效，目前在语音识别、图像识别、自然语言处理等诸多领域得到了大范围的运用[19]。深度学习在社交机器人的识别领域发挥着关键作用，通过对神经网络的各项参数进行不断调整，能够使模型自动学习数据特征，并实现对网络账号进行分类，以识别恶意社交机器人的任务。

其中，全连接层属于神经网络当中出现频率最高的一种层类型，其最鲜明的特征在于各神经元均连接于上一层当中的所有神经元[20]，通过不断调整权重和偏置项，可使模型对各项数据的特征进行自动学习，并进行分类、回归等任务。

本文所构建的深度学习模型由五层组成：

第一层：一个包含10个神经元的全连接层。

第二层：一个包含128个神经元的全连接层，采用ReLU激活函数。ReLU（rectified linear unit）激活函数的主要优势在于计算简便、收敛速度快以及能够缓解梯度消失问题。当输入值大于0时，输出值等于输入值；当输入值小于等于0时，输出为0。这种非线性变换使得神经网络可以表达复杂的非线性函数，从而提高模型的表达能力[1]。

第三层：另一个包含128个神经元的全连接层，同样使用ReLU激活函数。

第四层：一个Dropout层，设置了0.1的丢弃概率。Dropout作为一种正则化技术，通过以一定的概率随机丢弃神经元的输出，降低了神经网络过度拟合的风险，提高了模型的泛化能力，使模型学习数据特征的过程更加稳定。

第五层：一个包含1个神经元的全连接层，使用Sigmoid激活函数。由于Sigmoid函数的输出范围在(0,1)之间，通常用于表示概率值，适用于二分类问题。

模型整体结构如图1所示：

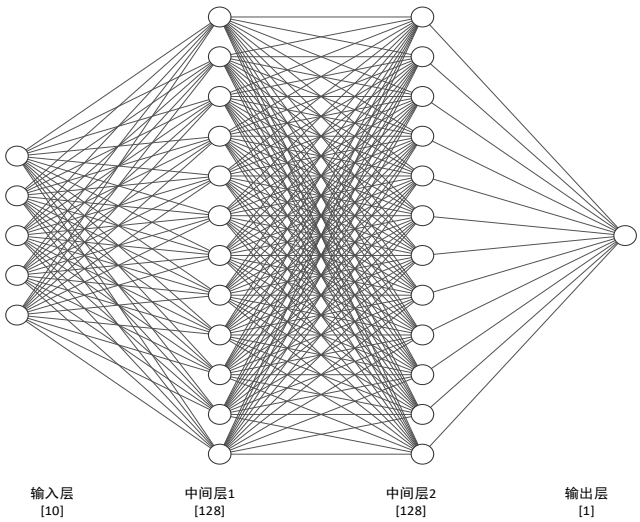

图1 神经网络模型结构图

在神经网络各项参数的更新方面，采用了Adam优化器。该优化器结合了Momentum和RMSProp优化算法的优点，在多数任务中表现出更快的收敛速度和更优的性能。

通过以上结构设计和技术应用，模型有效地提高了对数据的学习和分类能力，同时降低了过拟合的风险，增强了模型的泛化性能。

4 实验设计

4.1 实验模型构建

为科学识别并防御恶意社交机器人，建立基于深

度学习的恶意社交机器人识别模型。实验的流程如图2所示,包含数据获取、数据预处理、传统机器学习算法实践、深度学习模型实践、统计与对比分析在内的5个技术步骤。

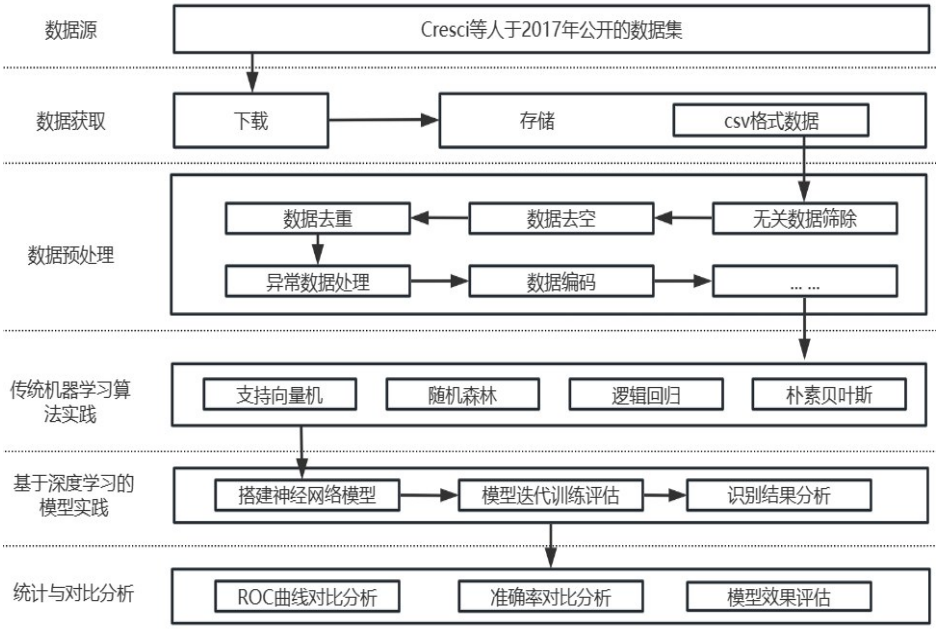

图2 实验流程

采用Cresci等[21]的方法,通过监控推特上的个人账户,寻找恶意社交机器人操作账户之间的关系所收集的公开数据集,并对数据进行异常值和缺失值清洗、数据编码、数据特征抽取等预处理后用于实验模型。实验模型主要分为传统机器学习算法模型和深度学习模型,其中支持向量机、随机森林、逻辑回归和朴素贝叶斯等传统机器学习算法模型主要用于对照实验,通过统计并分析各模型实验结果的精确率、召回率、F1-score等评价指标得分,以评估本文基于深度学习的恶意社交机器人识别模型效果。

4.2 数据来源及预处理

4.2.1 数据获取

数据预处理是数据挖掘中必不可少的重要环节,包括对原始数据集进行清洗、集成、替换等一系列操作。为了确保实验能够挖掘出客观有效的知识,必须提供干净、准确、简洁的数据。伴随人工智能等技术的不断发展,当前阶段的恶意社交机器人已经不满足于短时间内发表或转发大量的社交动态,逐渐演化为模拟人类行为特征的活动,如通过更换吸引力较强的头像、及时评论他人发布的状态的方式获取用户信任。因此,本文以此类新型恶意社交机器人作为检测对象,针对当前国内多数社交平台为保护用户个人隐私,拒绝开放数据接口并设置反爬虫措施的现状,通过对比目前已公开的部分数据集,具体见表1,最终选择文献[21]中的组合数据集作为本文数据集,以期保证获取数据的真实性与有效性,同时避免公开数据集中普遍存在的机器人类别不全、样本量较少等问题。

表1 部分公开的恶意社交机器人数据集对比

文献	数据集描述	相关数据	
		账号总数	推特状态数
文献[13]	包含真实用户账号与已分类的新旧型恶意社交机器人账号	9386	13,253,492
文献[14]	包含真实用户账号和恶意社交机器人账号	3025	10,432
文献[15]	包含真实用户账号和恶意社交机器人账号	167	7,799
文献[16]	包含真实用户账号和恶意社交机器人账号	3535	279,500

上述恶意社交机器人账户拥有从其他真实用户中盗取的照片、简介，并提供了已发布的个人推文及元数据（如发布时间、发布推文的方法、被转发次数等）。此外，该数据集还提供了账户的信息（如朋友和粉丝的数量、Twitter账号、图片是否是默认图片以及该账户创建后的时间等）。每个账户都经过了人工验证，以检查分类是否正确。该数据集包含目前大部分恶意社交机器人类型，可靠性较高。该数据集具体介绍如下表2所示，本文使用真实账号信息与意大利政治风波事件两项数据集作为模型训练数据。

表2 Cresci等人公布的数据集介绍

数据集	内容	用户数量	推特状态数	年份
Genuine accounts	基于人工识别的已验证的真实账户	3474	8,377,522	2011
Social spambots #1	一名意大利政治候选人的社交机器人发送账号	991	1,610,176	2012
Social spambots #2	针对移动设备的付费应用程序的垃圾邮件发送者	3457	428,542	2014
Social spambots #3	在亚马逊网站上销售产品的垃圾邮件发送者	464	1,418,626	2011
Traditional spambots #1	Yang等人（2013）使用的恶意软件垃圾邮件发送者的训练集	1000	1,418,626	2009

4.2.2 数据预处理

数据预处理就是对原始数据集实施清洗、集成等相关操作，此为数据挖掘中的关键环节。为了确保实验能够挖掘出有价值的知识，则必须为其提供干净、精准且简单的数据。但是，在原本就公开的数据集当中，机器人数据集与真实用户数据集间出现了许多有缺失且不正常的数据，由此导致本次模型训练的执行效率明显降低，甚至会造成模型结果出现明显的偏差。所以，针对原始数据集中的缺失值或明显的异常属性值进行去空，如针对属性值为0的账号进行删除，同时为提高分析结果质量与精准度，根据用户注册时间进行二次筛选，截取在数据收集前两周注册的账号信息。避免由于用户注册时间较短，难以辨别其属于恶意社交机器人账号还是正常账号对识别结果的意义不大的问题，最终筛重汇总后，初步拟定数据编码方案，并在预编码过程中对方案进行调整字段编码，最后转存为csv文件导出。

4.3 模型指标特征抽取

Cresci等为每个账户提供了账户信息类型的40个特征。此外，对每条推文，它提供了一个包含25个属性的向量，包括该推文的全文。然而，此类特征仅包括文本特征、名义特征和数字特征的组合，而单分类器只处理数字特征。因此，本文筛选特征向量只包含了代表在Twitter账户中表达行为的数字特征。

从账号信息中提取的特征越多会使模型获得更好的分类性能，然而特征提取是一个耗费较大的过程，且并不是所有的特征都能为模型的训练判断提供有价值的信息。基于之前的工作，选择为每个账号创建一个特征向量，仅包含已被证明可以更好地表示真实账户的相关特征，以期提供更好的机器人账户识别能力。特征向量可分为帐号使用情况和帐号信息两种类型，表3描述了为每个社交帐号获取的10个特征。这些特征是由Cresci等提供的账号原始信息以

表3 特征提取汇总

类型	特征	解释
帐号使用情况	转推比例（retweets）	转推数量与推文总数的比值
	回复比例（replies）	回复数量与推文总数的比值
	每推文收藏数（favoriteC）	用户的收藏总数与推文总数的比值
	每推文主题标签数（hashtag）	推文中包含的主题标签总数与推文总数的比值
	每推文网址数（url）	推文中包含的网址总数与推文总数的比值
	每推文提及数（mentions）	推文中的提及总数与推文总数的比值
	平均推文间隔时间（intertime）	两推文间平均间隔实践
	关注与被关注比例（ffratio）	关注数与被关注数的比值
账号基本信息	收藏数（favorites）	用户的收藏总数
	列表计数（listed）	用户被列入的列表总数

及每个用户相应推文的统计信息进行计算聚合获得的。

5 实验与分析

5.1 基于机器学习算法的社交机器人检测结果分析

5.1.1 支持向量机算法实验及结果分析

支持向量机(SVM, support vector machine)是Vapnilk等[22]于20世纪90年代在统计学理论基础上建立起来的一种机器学习方法,常用于分类问题。其通过寻找最优的超平面(决策边界),将不同类别的数据分开,尽可能使不同类别的数据样本点到超平面的距离最大化,其决策函数表达式见公式(1)。

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}a_i\gamma_i K(x,x_i)+b\right) \tag{1}$$

使用sklearn模块中的SVC类进行SVM建模。在建模过程中需对核函数进行合理的筛选,对使用线性核函数还有sigmoid核函数训练的模型准确率进行详细对比后,选择表现最优的核函数,结果如图3所示。

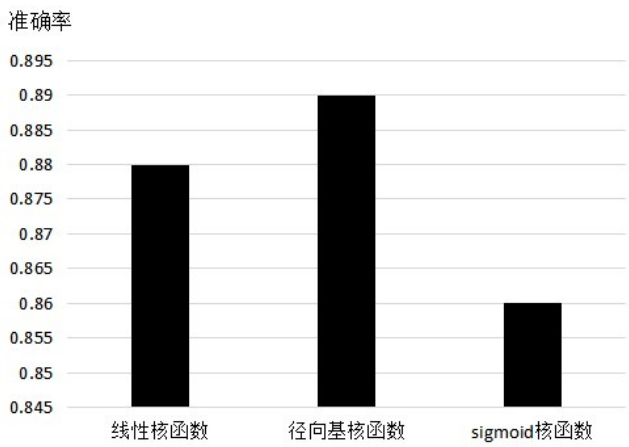

图3 不同核函数准确率计算结果

由图3可见,径向基核函数(rbf)在核函数比较中准确率最高,达到了0.89。因此选择支持向量机模型的径向基核函数创设分类器,相应的参数gamma设置为"auto";惩罚参数C设置为默认值1,然后用训练数据(x_train和y_train)对分类器进行训练。最终使用测试数据(x_test)进行预测,并将预测结果存储在y_pred变量中。结果如下表4所示。

表4 SVM模型预测结果

惩罚参数C	核函数	gamma	训练集平均准确率	测试集预测准确率
1	rbf	auto	0.90	0.88

为了评估分类模型的性能和预测效果,采用三个指标:精确度、召回率、F1-score进行评价。上述指标旨在对模型的性能加以准确衡量,精确度与召回率越大,意味着模型的预测效果越理想。但是,在某种情况下,召回率与精确度之间可能存在彼此矛盾的现象。因此,引入精确度以及召回率对应的调和平均值F1-score来判断分类器的效果。参照训练结果对精确率、召回率以及F1-score加以核算,随机森林模型对应的评价效果详情在表5中加以展示。

表5 SVM模型评价指标

精确率	召回率	F1-score
0.6395	0.9957	0.7802

分析表5不难发现,支持向量机分类模型对应的精确率等于0.6395,召回率等于0.9957,F1-score等于0.7802,只有召回率的结果相对理想,因此分类效果较差。

5.1.2 随机森林算法实验及结果分析

随机森林属于一种以决策树为基础的Bagging类型集成算法。这种算法从本质上来讲是多个决策树的一种集合,其核心理念是借助对多个决策树实施投票对最终结果加以确立。

对已提取的十种社交账号特征建立随机森林分类模型,同样通过sklearn模块实现模型。本文首先对数据进行初步的建模,在各项参数采用默认值的情况下,测试集上的预测准确率为98.6%。由于初步拟合的随机模型参数均为默认值,因此为了提升预测效果,提高分类器性能,对模型进行进一步改进,需要合理地调整随机森林模型当中的各项参数。

随机森林模型涉及的参数相对较多,在这之中有两项参数尤为重要,一是n_estimators(决策树的数目,默认值等于100),二是max_depth(树的最大深度值,默认为100)。及时调整参数,可以有效防止模型过拟合,提升了模型的泛化能力使模型更加优化,运行速度更快,所以需要优化该参数,通过十折交叉验证来实施参数调整,将n_estimators的参数范围设置为[10,200],测试步长等于10,输出结果如图4所示。

由图4所示,当决策树个数为161时,交叉验证的平均得分最高,因此决策树的最优个数为161。得到最优参数后再次进行建模训练,最终得到测试集平均准确率为0.9907。

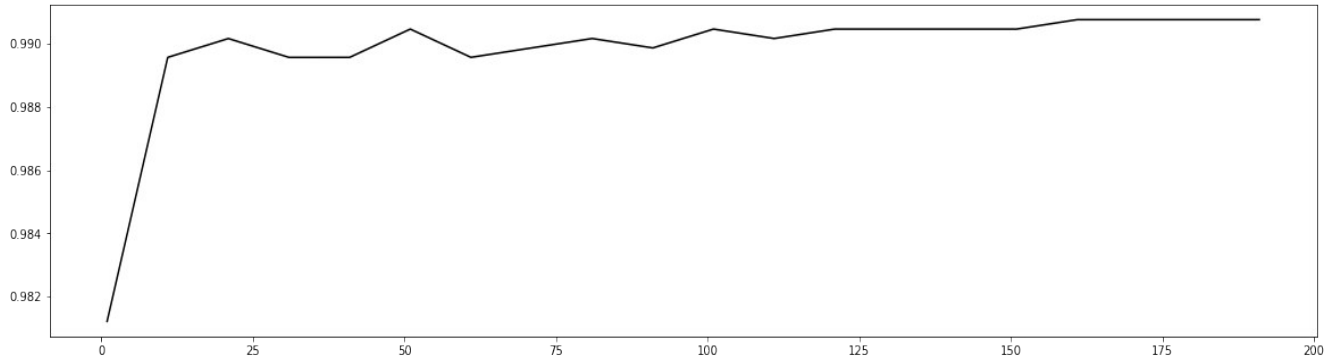

图4 最优决策树个数调参图

随机森林模型的训练结果评价指标如表6所示：

表6 随机森林模型指标评价

准确率	精确率	召回率	F1-score
0.9907	0.9745	0.9704	0.9728

从表6中可以看出，随机森林分类模型的召回率为0.9704，F1-score为0.9728，精确率为0.9745，该模型分类效果较好。

5.1.3 逻辑回归算法实验及结果分析

逻辑回归属于一种广义层面上的分类回归分析模型，凭借对给定训练集的n组数据进行训练，并在训练结束后对给定测试集的一组或多组数据进行分类。首先通过自变量利用回归分析的思想得到因变量的预测值y，然后通过逻辑函数把线性回归的结果从逻辑函数映射到(0,1)之间的概率值，最后根据设定的阈值ρ进行判别，逻辑函数的图像如图5所示，可以看出y的取值为(0,1)。

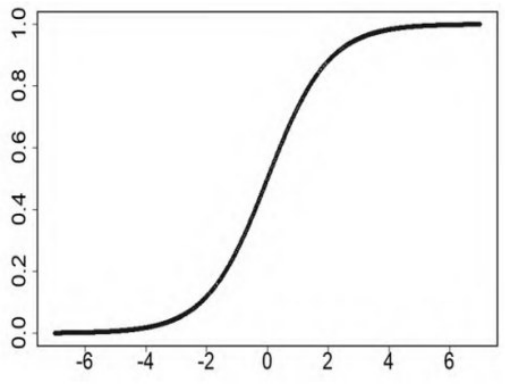

图5 逻辑函数图像

预测值y和逻辑函数的计算方法见公式(2)与公式(3)，其中，x_i是自变量，y是预测值，θ_i是待求系数，i=1,2,⋯,n。

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{2}$$

$$\rho(y) = \frac{1}{1 + e^{-y}} \tag{3}$$

为实现恶意社交机器人识别的任务，调用python中sklearn库构建基于逻辑回归的恶意社交机器人识别模型，具体训练结果与模型评价见表7：

表7 逻辑回归模型训练效果

准确率	精确率	召回率	F1-score
0.89	0.6836	0.2407	0.3561

5.1.4 朴素贝叶斯算法实验及结果分析

朴素贝叶斯属于一种以贝叶斯定理为基础的分类算法，其核心思想是通过先验概率和条件概率计算后验概率，以此实现分类。具体来说，朴素贝叶斯是一种以贝叶斯定理为基础的分类算法，其核心思想是通过先验概率和条件概率计算后验概率，以此实现分类。后验概率的计算公式如下式(4)：

$$p(Y = c_k | X = x) = \frac{p(Y = c_k) P(X = x | Y = c_k)}{P(X = x)} \tag{4}$$

在scikit-learn库中，使用sklearn.naive_bayes模块中的五种不同的朴素贝叶斯分类算法，根据特征数据的先验分布不同而区分，包括伯努利朴素贝叶斯（BernoulliNB）、类朴素贝叶斯（CategoricalNB）、高斯朴素贝叶斯（GaussianNB）、多项式朴素贝叶斯（MultinomialNB）和压缩感知朴素贝叶斯（ComplementNB）。本文调用python中sklearn库中函数GaussianNB()与ComplementNB()，选择构建基于贝叶斯的恶意社交机器人识别模型，具体训练结果与模型评价见表8。

表8 朴素贝叶斯模型训练效果

模型类型	准确率	精确率	召回率	F1-score
GaussianNB	0.93	0.7749	0.9877	0.8684
ComplementNB	0.85	0.5855	0.9959	0.7374

5.2 基于神经网络模型的社交机器人检测结果分析

实现过程使用model.compile方法来编译模型。本模型对优化器optimizer进行了设置。优化器的主要功能是对权重进行及时更新,同时有效地减小损失函数。Adam优化器兼具Momentum及RMSProp优化算法的所具备的优点,对各参数的学习率实施合理的自适应调节。相较于其他优化算法,Adam在多数任务中展现出更快的收敛速度和更优的性能,因此本文使用了Adam优化器;损失函数loss,用于衡量模型的预测与真实值之间的差异。由于模型的第五层使用了Sigmoid激活函数,因此选择BinaryCrossentropy损失函数,并设置from_logits=True,即损失函数于内部将模型的输出转换为概率;metrics:评估模型性能的指标,使用准确率(accuracy)作为评价指标。最后调用model.fit方法训练模型。具体训练结果与模型评价见表9:

表9 逻辑回归模型训练效果

准确率	精确率	召回率	F1-score
0.97	0.9688	0.9840	0.9764

5.3 对比分析

通过上述评价指标和ROC曲线对构造的恶意社交机器人识别模型进行评价,将五种模型的评价指标汇总到一起进行比较,结果如表10所示:

表10 模型各评价指标对比

模型	测试集准确率	精确率	召回率	F1-score
支持向量机模型	0.88	0.6395	0.9957	0.7802
随机森林模型	0.99	0.9745	0.9704	0.9728
逻辑回归模型	0.89	0.6836	0.2407	0.3561
高斯朴素贝叶斯模型	0.93	0.7749	0.9877	0.8684
补充朴素贝叶斯模型	0.85	0.5855	0.9959	0.7374
深度学习模型	0.97	0.9688	0.9840	0.9764

从表10中可以看出,深度学习模型与随机森林模型表现出色。随机森林模型的准确率最高,达到0.99,精确率为0.9745,召回率为0.9704,F1-score为0.9728。深度学习模型的准确率为0.97,精确率为0.9688,召回率为0.9840,F1-score为0.9764,在F1-score和召回率上略优于随机森林模型。这表明深度学习模型在正确识别正样本的同时,保持了较低的误报率,体现了良好的泛化能力。相比之下,其他模型的性能指标相对较低,支持向量机和逻辑回归模型的精确率和F1-score较低,可能存在较高的误报或漏报率。综上所述,深度学习模型在精确率和召回率之间取得了较好的平衡,整体性能优异,适用于对准确性和全面性要求较高的应用场景。

绘制出五个模型的ROC曲线图,可以直观地评价模型对社交机器人的识别效果,五个模型的ROC曲线如图6所示:

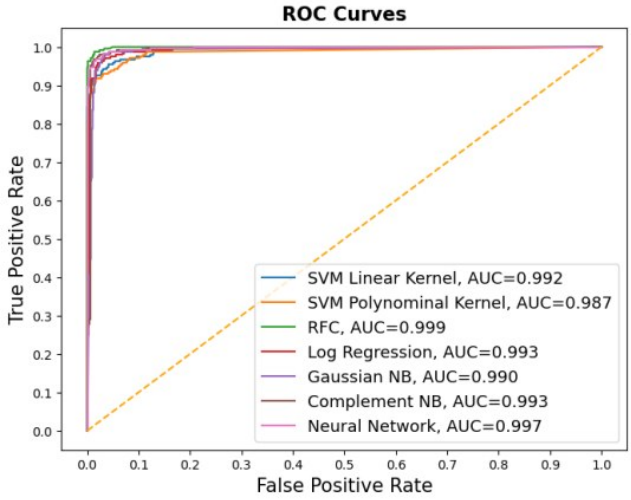

图6 模型ROC曲线对比图

从ROC曲线图可以直观地看出,随机森林模型对应ROC曲线与横坐标之间的面积最大,即AUC值最大,为0.999,模型的区分能力最强,性能最优。本文搭建的神经网络模型AUC值为0.997排第二。

综上可知,在支持向量机、随机森林、逻辑回归、朴素贝叶斯、深度学习五个模型中,随机森林模型的预测效果是最好的,其准确率、F1-score、AUC值均高于其他模型,能够很好地实现对于恶意社交机器人的识别,并且本文搭建的恶意社交机器人识别模型综合能力处于上游,具有较强的可用性与有效性。

6 总结与展望

本文针对在线社交网络中广泛出现的恶意机器人账号的特征,提出了一种基于深度学习的识别模型。在阐述了传统机器学习以及深度学习相关分类算法的定义及原理的基础上,介绍了模型评估方法以及评价指标的相关概念。设计社交机器人检测框架,在实验过程中对Twitter上恶意社交机器人数据进行了预处理,包括异常值处理、数据去重去空等步骤。对原始数据进行数字化特征提取后对账号特征向量构建分类模型。在此基础上建立基于传统机器学习

方法的支持向量机分类模型、随机森林分类模型、逻辑回归分类模型与两种朴素贝叶斯模型进行检测。

最后探讨并实践了使用TensorFlow框架针对二分类问题构建一个简单神经网络模型的过程，涉及到了全连接层、激活函数（ReLU）、Dropout技术、二元交叉熵损失函数以及Adam优化器等多种深度学习技术及其相关原理。通过将深度学习与传统机器学习的不同算法进行对比，借助模型评价指标检验模型效果，结果表明该深度学习模型准确率达到了0.97，效果最佳，证实了该模型的可行性与有效性。

后续研究可以尝试结合多种技术和方法，提高神经网络模型的性能，以更好地解决实际问题。同时细化模型指标，提取具有关系属性的恶意社交机器人数字特征，从而针对性优化模型训练过程。

参考文献(References)：

[1] 张玉清，吕少卿，范丹. 在线社交网络中异常账号检测方法研究[J]. 计算机学报，2015，38(10)：2011-2027.

[2] 杨舟. 社交网络机器人检测综述[J]. 网络安全技术与应用，2022(03)：135-136.

[3] Newberry C. 36X(Twitter)stats that matter to marketers in 2024[EB/OL]. (2024-04-03)[2024-10-28]. https://blog.hootsuite.com/twitter-statistics/.

[4] Ferrara E, Varol O, Davis C, et al. The rise of social bots [J]. Communications of the ACM, 2016, 59 (7) : 96-104.

[5] Wang G, Mohanlal M, Wilson C, et al. Social turing tests: crowdsourcing sybil detection [C]// NDSS Symposium, 2013.

[6] Davis C A, Varol O, Ferrara E, et al. BotOrNot: a system to evaluate social bots[C]// Proceedings of the 25th International Conference Companion on World Wide Web, 2016: 273-274.

[7] Pizarro J. Using n-grams to detect bots on Twitter [C]// Conference and Labs of the Evaluation Forum, 2019.

[8] Gamallo P, Almatarneh S. Naive-bayesian clas-sification for bot detection in Twitter[C]// Conference and Labs of the Evaluation Forum, 2019.

[9] Fazil M, Abulaish M. Identifying active, reactive, and inactive targets of socialbots in Twitter[C]//Proceedings of the International Conference on Web Intelligence, 2017: 573-580.

[10] Ping H, Qin S. A social bots detection model based on deep learning algorithm[C]// IEEE 18th International Conference on Communication Technology (ICCT), 2018: 1435-1439.

[11] Kudugunta S, Ferrara E. Deep neural networks for bot detection[J].Information Sciences, 2018, 467: 312-322.

[12] Liu Z, Chen C, Yang X, et al.Heterogeneous graph neural networks for malicious account detection[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018: 2077-2085.

[13] Jia J, Wang B, Gong N Z. Random walk based fake account detection in online social networks[C]// 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2017: 273-284.

[14] Wang B, Gong N Z, Fu H. Gang: detecting fraudu lent users in online social networks via guilt-by-association on directed graphs [C]// IEEE International Conference on Data Mining (ICDM), 2017: 465-474.

[15] Wang B, Jia J, Zhang L, et al. Structure-based sybil detection in social networks via local rule-based propagation[J]. IEEE Transactions on Network Science and Engineering, 2018, 6(3): 523-537.

[16] Gao P, Wang B, Gong N Z, et al. Sybilfuse: combining local attributes with global structure to perform robust sybil detection [C]// IEEE Conference on Communications and Network Security (CNS), 2018: 1-9.

[17] Wang B, Zhang L, Gong N Z. SybilSCAR: sybil detection in online social networks via local rule based propagation [C]//IEEE INFOCOM 2017-IEEE Conference on Computer Communications, 2017: 1-9.

[18] Feng S, Tan Z, Wan H, et al. Twibot-22: towards graph-based twitter bot detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 35254-35269.

[19] 赵蓓，张洪忠. 有关北京冬奥会的社交机器人叙事与立场偏向——基于Twitter数据的结构主题模型分析[J]. 新闻界，2022(05)：62-70.

[20] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological review, 1958, 65(6): 386.

[21] Cresci S, Pietro R D, Petrocchi M, et al. The paradigm-shift of social spambots: evidence, theories, and tools for the arms race[C]//Proceedings of the 26th International Conference on World Wide Web Companion, 2017: 963-972.

[22] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer Science & Business Media, 2000.

[23] 陈虹，张文青. Twitter社交机器人在涉华议题中的社会传染机制——以2022年北京冬奥会为例[J]. 新闻界，2023(02)：87-96.

[24] 刘蓉，陈波，于泠，等. 恶意社交机器人检测技术研究[J]. 通信学报，2017，38(S2)：197-210.

[25] Rao S, Verma A K, Bhatia T. A review on social spam detection: challenges, open issues, and future directions[J]. Expert Systems with Applications, 2021, 186: 115742.

编辑：王谦