Inline Text Wrapping Picture

北京邮电大学

硕士研究生学位论文开题报告

学 号: 2019111588

姓 名: 胡杨晔

学 院: 网络空间安全学院

专业(领域): 网络空间安全

研究方向: 应用安全

导师姓名: 张华

攻 读 学 位: 工学硕士

2020年12月11日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 黑暗术语的发现和使用 | | |
| 选题来源 |  | 论文类型 |  |
| 开题日期 | 2020-12-02 | 开题地点 |  |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  研究目的：  本课题旨在研究违法者为了规避审查使用的黑暗术语，通过研究黑暗术语的上下文来确定一个词是否是黑暗术语，为内容审查者提供帮助，更好的维护平台的内容安全，打击犯罪交易。  研究意义：  在过去几十年中，互联网极大地促进了电子商务的发展，同时互联网也成为地下市场的主要平台，网络犯罪分子在这里交换用于犯罪的产品和服务。2013年，全球网络犯罪造成了3750亿美元的损失，几乎等同于全球毒品交易的数量。地下论坛一度是网络罪犯的通信枢纽，帮助他们推广攻击工具包和服务，协调他们的行动，交换信息并寻求合作。例如Silk Road论坛，拥有3万到5万活跃用户，是毒品和其他非法毒品交易的滋生地，每天有两到三百份通讯记录。这些记录提供了对网络犯罪方式、犯罪分子战略、能力、基础设施和商业模式的深刻洞察，甚至可以用来预测他们的下一步行动。  随着监管的扩大，许多地下论坛被封禁，但这不意味着网络罪犯的消失，一些广为人知的社交平台正在被不法分子用于交易通讯。由于地下市场的排他性，网络罪犯发展了一套独特的语言系统，违法者会使用一些黑暗术语来对外隐藏具体交易内容或规避社交平台的内容审查。黑暗术语用来指定产品、服务和其他网络犯罪特定概念。这种术语往往是一些看起来很普通、看很天真、却有秘密含义的词语。罪犯们将其用来隐藏正在讨论的内容。例如，毒贩经常用“ice”代替“Methamphetamine”（冰毒），“pants”代替“Herion”（海洛因）。此类欺骗性内容使得违法者之间的通信不容易被自动化系统和审查人员发现。因此，自动发现和理解这些黑暗术语对于理解各种网络犯罪活动和减轻它们所构成的威胁非常有价值。  国内外研究现状：  针对黑暗术语的发现的研究，目前主流的解决技术主要有五种：  1.基于关键字检测和扩展方法，通过对关键字相关搜索结果来判断是否是黑暗术语。只适用于搜索引擎的搜索内容审查。  2.基于黑暗术语在不同语料中的词义差异性。黑暗术语的正常语料中往往使用普通的含义，在黑暗语料中才会表达违禁词含义。建模潜在黑暗术语在不同语料中的词义表示，比较差异程度确定是否是黑暗术语。该方法容易受到词多义性干扰。  3.基于情绪分析的方法，该方法从情感极性角度出发，认为黑暗术语是指代词的委婉表达，词的情感极性比句子的情感极性低。根据句子极性和词极性差异确定是否是黑暗术语。该方法似乎合理可行，但它需要额外的手动筛选过程来细化候选对象，无法满足所期望的自动、大规模的要求。  4.基于黑暗术语与指代词上下文相似性。通过MLM掩码模型学习指代词的上下文，通过指代词的上下文来预测黑暗术语。该方法规避了复杂的标注工作，但训练目标与任务并不完全一致，准确率也较低。  主要文献：  [1]G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, R. Portnoff, S. Afroz,D. McCoy, K. Levchenko, and V. Paxson, “Identifying products in onlinecybercrime marketplaces: A dataset for fine-grained domain adaptation,”inProceedings of Empirical Methods in Natural Language Processing(EMNLP), 2017, pp. 2598–2607  [2]R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson, “Tools forautomated analysis of cybercriminal markets,” inProceedings ofInternational Conference on World Wide Web (WWW), 2017, pp. 657–666.  [3]K. Yuan, H. Lu, X. Liao, and X. Wang, “Reading thieves’ cant: auto-matically identifying and understanding dark jargons from cybercrimemarketplaces,” inProceedings of 27th USENIX Security Symposium,2018, pp. 1027–1041.  [4]R. Magu and J. Luo, “Determining code words in euphemistic hatespeech using word embedding networks,” inProceedings of the 2ndWorkshop on Abusive Language Online (ALW2), 2018, pp. 93–100.  [5]J. Taylor, M. Peignon, and Y .-S. Chen, “Surfacing contextual hate speech words within social media,”arXiv preprint arXiv:1711.10093, 2017.  [6]Zhu W, Gong H, Bansal R, et al. Self-supervised euphemism detection and identification for content moderation[J]. arXiv preprint arXiv:2103.16808, 2021.  [7]H. Yang, X. Ma, K. Du, Z. Li, H. Duan, X. Su, G. Liu, Z. Geng,and J. Wu, “How to learn klingon without a dictionary: Detection andmeasurement of black keywords used by the underground economy,”inIEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp.751–769.  [8]Zhao K, Zhang Y, Xing C, et al. Chinese underground market jargon analysis based on unsupervised learning[C]//2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016: 97-102. | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  拟解决的关键科学问题：  1.如何建模上下文信息对黑暗术语进行识别。过去的研究依赖于静态词嵌入（如word2vec），希望通过比较在词在不同的语料中（犯罪人员交流的黑暗语料和一般化的交流语料）上的差异来检测黑暗术语，对黑暗术语的标注并不监督，而且模型效果在不同数据集上的鲁棒性较差。一类新的方法使用MLM掩码模型，使用自监督方式规避了标注过程同时在不同的数据上鲁棒性较好。该方法训练方法fine-tune（令MLM模型在黑暗语料上进行类似完型填空的训练）与识别黑暗术语目标并不完全一致，模型准确度偏低。如何对模型进行改进，在不影响鲁棒性的情况下，提高对术语识别的准确性。  2.如何识别黑暗语料。随着现有监管力度的加大，许多黑暗论坛遭到封禁。无论监督模型和自监督模型都需要黑暗语料来作为数据集或待标注数据原料。随着许多论坛的封禁消失，过去爬取特定网页信息的方式已经不太可行。Zhu等人2021年发表在EMNLP的论文，数据却来自2018，缘于18年后许多论坛的消失。但论坛消失不代表犯罪交流的结束，违法者往往会在社交媒体上进行更为隐蔽的交流。对于一个机器学习模型而言，数据是否足够很大程度上会影响模型的鲁棒性，而数据是否够新则决定模型能否捕捉语料中上下文的新的变化。如何在海量社交媒体数据中区分出黑暗语料和非黑暗语料是一个重要问题。  3.如何即时确定黑暗术语使用。对于内容审查而言，关键部分是确定一句话是否使用了黑暗术语来表达违禁含义。过去的研究都着眼于一个词是否被用来表达违禁含义，这样的结果无法直接用于审查系统（大量的黑暗术语不仅仅有违禁含义还有无害含义，给审查带来混淆）。 Zhu等人做了词义鉴别的工作，但鉴别结果是某个词在语料下的语义分布，与我们所需求的判断一句话中黑暗术语是否使用了违禁含义的目标不符。  研究内容：  1.  2.  3.  研究目标：  目标是设计一个能够快速且准确检测重打包应用的系统，并且该系统能够根据当前的相似度数据分布情况，动态调整阈值，能够极大的避免未知数据带来的概念漂移。  现有的应用检测多是使用基于代码特征的静态检测技术对重打包应用进行检测，但是由于混淆、加固技术的普及，重打包应用往往会使用这些技术去逃避基于代码特征静态检测。由于代码特征的局限性，目前也有很多应用检测是根据资源特征进行检测的，但是由于资源特征并不能代表应用的行为特征，并且资源特征通常都是易于修改的，单纯的使用资源特征进行检测会造成很大的误报和漏报。  对于用户来说，重打包应用的威胁是极大的。首先，由于重打包应用通常都是针对受欢迎的良性应用，在图标，界面等用户能够感知的特征上，重打包应用和原始应用并没有任何区别。因此，用户通常无法分辨重打包应用和原始应用有哪些区别，并且由于原始应用的知名度，用户已经默认了重打包应用是安全的。这也就导致了用户完全不能意识到自己触发了恶意行为或观看了重打包应用恶意嵌入的广告。  针对于个人手机在各种应用商场下载的某些应用，这些应用可能没有被应用市场检测，存在各种各样的安全问题，甚至在官方应用商店下载的某些应用也因为审查不力而存在着应用重打包的问题。并没有一个平台能够提供一个检测系统，能够对于手机用户提供一个重打包检测服务。所以如果能提供一个检测系统，它可以快速准确的判断出当前应用是一个重打包应用，存在安全风险，提醒用户需要谨慎安装，这将降低用户被恶意应用攻击的风险。  效果：  对于大多数人来讲，他们可能没有意识到重打包应用的存在。重打包应用有着和正常应用一样的图标、功能和界面，没有明显的能被用户感知的恶意行为。用户即使在遇到相同或相似的应用时，也不会对应用的发布者进行仔细判断，而是随便选择一个应用进行安装。对于用户来说，重打包应用和其他正常应用没有任何区别。如果有一个检测系统可以在应用市场自动发现服务器中存在的重打包应用，并且提供给用户检测重打包应用的服务。对于用户来说，服务器端的自动检测能降低用户被重打包应用攻击的风险，提供给用户的重打包应用检测服务也可以让用户主动进行应用检测，能提高用户的安全意识，进一步降低应用的安全风险。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  研究方法：  方法将对应拟解决的科学问题分为三部分。  1.如何面对未知的数据集时能划分一个合适的阈值，使其能够有效判断应用是否是重打包应用。动态阈值对未知的数据集有着高适应性。设置动态阈值触发机制，只有数据触发了机制，才重新计算新的阈值，动态阈值算法参考了图像分割领域的最大类间方差法，该方法计算迅速，并且生成的阈值对重打包应用有着良好的区分度。  2.如何选取合适的特征，快速过滤出可疑的重打包应用。只使用布局文件，通过布局文件快速生成唯一特征，快速过滤出可疑的重打包应用。  3.如何为混淆或加固应用生成CG。通过向活动的OnCreate和OnDestroy方法注入Android自身的API，动态获取两个方法之间所有线程运行时调用的方法数据。基于这些数据，使用拟合拼接算法生成CG。  可行性分析：  可行性分析方法将对应研究方案分为三部分。  1.动态阈值生成时应该保证快速，并且具有良好的区分度。动态阈值触发机制保证了只有能够对当前数据分布特征产生影响的新数据被输入时，才重新计算阈值，降低了阈值计算的平均时间。相似度分数数据的分布会呈现双峰趋势，并且在数据分布中存在的异常数据数量少。数据分布与图像分割中分割前景与背景的图片像素分布非常相似，因此可以参考图像分割领域的最大类间方差法动态生成阈值。  2.重打包应用和原始应用都具有高度相似的界面以欺骗用户下载使用。因此选取布局文件作为特征快速过滤可疑重打包应用。布局文件与代码紧密相连，删除或者修改布局文件或布局文件中的组件需要修改相对应的代码，这对于应用重打包者的代价是巨大的。在面对应用中增加的冗余布局来说，只需要指定一个约束规则，规定每一个布局最多只允许匹配一个布局的情况下生成最佳布局匹配，就能极大的减小冗余布局文件对相似度判定的影响。  3.重打包应用和原始应用都具有高度相似的界面。在动态检测中，如果事件输入器在对应用进行模拟操作时能够保证在相同或相似的界面下，生成的事件基本相同，那么对于重打包应用和原始应用，他们遍历的界面路径就相同，触发的代码行为也基本一致，这样就可以保证重打包应用和原始应用生成CG图的完整程度基本一致。  技术路线：  1.布局相似度检测  对于应用包含的布局文件进行解析，排除可能插入的冗余布局组件或冗余布局，将剩余的布局文件进行相似度的比较。  2.动态检测  对于过滤出的应用对在沙箱环境中进行模拟输入，通过向方法中注入追踪函数，获取应用运行时的方法调用。  实现方法及步骤：  1.对一定规模的应用进行相似度分析，划分一个合适的初始阈值；  2.将应用按顺序解压，解析布局文件，排除布局文件中可能存在的冗余布局或组件，在特征提取中排除这些文件和组件，然后进行特征提取，并比较相似度，通过动态阈值找出可能存在的重打包应用程序。  3.对所有可疑的重打包应用程序在沙箱环境中模拟执行，通过动态注入追踪函数获取线程间的函数调用信息，使用这些信息拟合生成CG。  4.对生成的CG进行相似度比较，相似度超过动态阈值的应用判断为重打包应用。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  1.采用动态阈值判断应用重打包，提高了对未知数据集的适应性，降低了人工确定阈值时的复杂性。在当前所有的应用重打包检测中，都是采用人工定义阈值去判断重打包应用。人工定义阈值需要使用大量的数据反复进行测试以确定一个合适的阈值。由于阈值是固定的，面对未知数据集时很可能产生概念漂移，并且精度受到人工限制，从而导致误报率和漏报率的上升。  2.只使用布局文件作为静态分析特征，快速过滤可疑的重打包检测。当前采用布局文件作为静态分析特征的重打包检测研究中，都是将布局与代码相结合的方式去除应用中可能存在的冗余布局，往往并没有考虑混淆和加固技术对于提取布局与代码的关联特征产生的影响。或者直接采用动态获取运行时使用的布局文件，以此来获取布局与代码的关联特征，花费的时间成本是昂贵的。  3.使用动态检测生成的数据，拟合生成CG图。CG图作为静态分析中能够反映应用行为的特征，通常结合控制流图等特征一并使用，但是在面对动态加载或者加固技术时，由于重打包应用源代码被隐藏，导致CG图生成不全面，从而导致比较应用相似性时产生误差。而动态检测只能直接获取到基于时间的函数调用序列，由于缺少函数的直接调用关系，这些数据并不能直接用来生成CG。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  1.与本项目相关的研究工作积累基础  这个项目的需求很多，需要多方面的技术，需要掌握的语言有java、python，需要熟悉的框架的有Android框架、Xposed框架，需要学习的技术主要有Android应用逆向分析，树和图的相似度比较，图像分割领域中的动态阈值算法的原理。  当前已经具备的试验条件，目前可以使用大量的开源逆向工具，对应用程序进行逆向解包，提取出应用的布局文件，通过对布局文件的解析得到布局文件中的组件树。动态生成CG可以使用沙箱模拟动态环境，对应用进行仿真运行。动态注入追踪API，然后对应用进行模拟操作，获取应用运行时方法的调用信息，通过调用信息拟合CG。  当前主要缺少对比布局文件的组件树的相似度的方案，CG的相似度快速比较方案。由于重打包应用并没有一个基准的数据集，所以目前还缺少一个已经打上正确标签的重打包应用数据集。  2.拟解决途径  布局文件的组件树可以通过查找文献找到树的相似度比较方案，或者重新从布局文件中提取容易比较相似度的特征。CG相似性比较涉及到图的相似性比较，可以查找最新的关于图比较方向的文献，找到合适的快速比较方法。2019年Li等人在IEEE Transactions on Software Engineering发布的论文《Rebooting Research on Detecting Repackaged Android Apps: Literature Review and Benchmark》中公布了他们使用的基于人工分析和系统分析重打包数据集并在github上开源，目前数据集仍然在不断的更新。可以使用当前公布的数据集进行试验测试。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2020年11月-2020年12月 | 阅读相关方向论文、选题 | 完成开题报告 |
| 2020年12月-2021年3月 | 通过布局文件过滤可疑重打包应用对 | 可以快速过滤重打包应用对；产出当前数据动态生成合适阈值的算法 |
| 2021年3月-2021年6月 | 分析相似度数据分布，完成动态阈值算法的设计 | 动态阈值在未知数据中判断重打包是有效的；设计实现一个完整的系统，基于系统产出一篇论文 |
| 2020年6月-2020年9月 | 使用动态检测结果拟合生成CG | 动态检测生成CG能够精确判断重打包应用；基于动态检测结果生成CG的检测方法 |
| 2020年9月-2021年10月 | 应用程序框架搭建 | 完成运行并测试所有流程 |
| 2021年10月-2022年1月 | 论文写作 | 论文最终完成 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 导师类型 | 单位名称 | 职务 |
| 秦素娟 | 教授 | 博、硕导 | 北京邮电大学 | 组长 |
| 金正平 | 副教授 | 硕导 | 北京邮电大学 | 成员 |
| 李文敏 | 副教授 | 博、硕导 | 北京邮电大学 | 成员 |
| 时忆杰 | 工程师 |  | 北京邮电大学 | 成员 |
| 张华 | 副教授 | 博、硕导 | 北京邮电大学 | 成员 |
|  |  |  |  |  |
| 导师意见：  同意开题。 | | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | | |
| 开题报告小组意见： | | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | | |
| 学院意见（签章）： | | | | | |
| 负责人：  日期： 年 月 日 | | | | | |