北京邮电大学

硕士研究生学位论文开题报告

学　　号：2020111506

姓　　名：熊梦军

学　　院：网络空间安全学院

专业(领域)：计算机科学与技术

研究方向：大数据技术与智能信息处理

导师姓名：秦素娟

攻 读 学 位：工学硕士

2021 年 12 月 10 日

论文题目	面向黑色交易的术语发现和鉴别方法研究		
选题来源	其他	论文类型	应用研究
开题日期	2021-12-02	开题地点	科研楼 610

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，H. Yang [1]等人通过对关键字相关搜索结果来判断是否是术语。只适用于搜索引擎的搜索内容审查。本文集中于文本对话中的关键词提取与搜索引擎的搜索

内容差异较大。

2.基于黑暗术语在不同语料中的词义差异性。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。许多学者[2~6]沿用这一方法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术语。该方法基于词嵌入进行，词义与词嵌入一一对应，无法满足本文对排除多义词干扰的需求。

3.基于情绪分析的方法，J. Taylor[7]从情感极性角度出发，认为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，无法满足本文所期望的自动、大规模的要求。

4.基于术语与指代词上下文相似性。Zhu W[8]通过使用 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。在术语发现的精度上无法满足本文需求。

主要文献：

[1]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.

[2]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

[3]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.

[4]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.

[5]R. Magu and J. Luo, “Determining code words in euphemistic hatespeech using

word embedding networks," inProceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 93–100.

[6]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. [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.

[7]J. Taylor, M. Peignon, and Y .-S. Chen, "Surfacing contextual hate speech words within social media," arXiv preprint arXiv:1711.10093, 2017.

[8]Zhu W, Gong H, Bansal R, et al. Self-supervised euphemism detection and identification for content moderation[J]. arXiv preprint arXiv:2103.16808, 2021.

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）

研究内容：

1. 基于 MLM 模型研究如何兼顾鲁棒性和准确率。为了解决问题 1，需要设计一些合适的后处理步骤，对 MLM 模型的结果进行后处理，排除错误答案，提高准确率同时不影响模型的鲁棒性。为此应该对现有的模型进行错误分析，总结模型缺陷以及错误类型，进行分类总结。在此基础上对术语特点和过去方案进行更加深入的研究，确定过去方案优缺点以及是否能解决或者改善现有模型错误，最后根据研究结果提出合适的后处理办法。

2.基于社交媒体数据特点筛选黑暗语料。黑暗语料在社交媒体数据中只占极小的一部分，相应标签分布呈长尾状态。基于机器学习的分类算法应用于长尾分布数据集时，识别效果并不好。目前通用的长尾分布方法都集中在数据增强、模型损失函数改进等方面，在具体的场景没有针对性，效果往往较为一般。我们需要结合具体场景来进行更为贴合实际的数据分析。首先应该对社交媒体数据类别进行分析，并结合黑暗语料特点进行比对，以此来设计高效且准确率高的应对长尾分布的分类方法。

3. 基于词的上下文研究如何高效区分术语含义。利用研究 1 的结果构建禁用词列表来对内容进行筛查是一个高效的方式，但是由于术语存在的无害含义，会带来大量假阳性例子。需要研究如何确定术语的使用。MLM 模型虽然具有区分不同上下文的能力，但会带来大量的时间消耗，不满足即时性的要求。首先应该研究词在非常规语义下的特点，根据上下文特点设计兼顾及时性和准确率较高的词义判断模型。

拟解决的关键科学问题：

1. 如何提高模型发现术语能力。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）上的差异来检测术语，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表现较好。所使用的自监督方法中的 fine-tune 过程并不稳定导致模型准确度偏低。在不影响鲁棒性的情况下，如何进行改进，提高对术语识别的准确性。

2.如何识别黑暗语料。黑暗语料指的是带有网络犯罪人员交流风格和特色的语料信息。这类文本特色比较鲜明，与正常语料对比差异较大。这类文本主要作用是作为模型的数据或待加工数据。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。虽然不一定所有黑暗语料中都带有术语，但只有获取最新的黑暗语料，才能让模型学习其中存在的最新术语使用。

3.如何即时确定黑暗术语词义。对于内容审查而言，关键部分是确定一句话是否使用了术语来表达违禁含义。过去的研究都着眼于一个词是否具有违禁含义，这样的结果无法直接用于审查系统（大量的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。

研究目标：

目标是设计一个能够发现黑暗术语并检测术语使用的系统，同时系统还具备收集筛选语料进行数据补充的功能，以适应随着时间推移黑暗术语上下文的变化。

现有的识别方法多是集中在单词在不同语料上的词义差距来进行，容易受到词的多义性干扰，同时鲁棒性不强，更替数据集后效果急剧下降。一种新的方法是利用术语与指代词上下文的相似性结合 MLM 掩码模型来预测术语使用，但该方法由于训练过程的特点，准确率较低。过去少有对于黑暗术语词义鉴别的工作， 即使有也是在语料级别上的词义分布，无法在句子层面对语义进行鉴别。传统的词义鉴别工作则严重以来语言学家和词典编纂者创建的带有词义标记的参考语料库，与我们的自动化、大规模的需求场景不符。

近年来随着网络犯罪的兴起，大量社交媒体和网络社区开始花费更多时间和人力到内容审查上。审查自动化是目前许多平台的一项重大需求。自动化审查的最简单的一种方式是基于“禁用词列表”的禁用词系统，通过检索禁用词来完成内容审查。但是这种方式很容易被规避，违法者可以通过发明黑暗术语来代替禁用词，而这些被用来替代的黑暗术语本身往往是一些具有其他无害含义的词，不能被无条件过滤。

由于网络审查的逐渐普遍化，地下论坛的语料变得难以获取。地下论坛的语料数据，里面可能包含着大量的违法者特有的语言风格以及未被识别到的黑暗术语。我们需要考虑如何对这类语料进行补充，使模型表现更加出色同时能适应新的变化。

通过使用黑暗术语，罪犯们能够在主流社交媒体上进行关于黑色交易的交流。即使依赖众多语言学家去研究提取黑暗术语，也可能赶不上术语更新的速度，更遑论对海量数据中的每一句进行词义鉴别。目前并不存在这样的系统能自动化的进行黑暗术语的发现和句子级别的词义鉴别，所以如果能提供一个检测系统，能即时发现新的黑暗术语，同时检测黑暗术语的使用将会对理解各类网络犯罪活动和减轻其威胁提供很大的帮助。

效果：

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。

通过使用设计的检测系统，能发现新的术语的使用情况。同时系统能从大量数据中找到黑暗语料作为模型的数据补充，来保证模型能利用到最新的术语使用信息。系统还具有句子级的鉴定黑暗语义能力，能区分术语是否使用了黑暗语义。通过该系统能即时发现新的黑暗术语的使用，能有效的发现违法犯罪活动的交流，能成为执法者有力的信息辅助。系统的句子级的鉴

别能力也可以帮助审查人员自动化的进行完成内容审查，提高了审查效率，减少了审查人员的负担，为网络空间安全提供更多保障。

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.如何保留 MLM 模型鲁棒性的优点同时提高模型识别能力。利用术语在不同语料中的上下文差异，利用词嵌入技术和余弦相似度量化差异，通过差异对比排除错误的术语。

2.如何识别黑暗语料。如何在大量的媒体数据中筛选出所需的黑暗语料。设计一个粗-细分类器。粗分类器负责筛选与黑暗语料差异较大的类型数据，细分类器负责区分容易与黑暗语料混淆的语料类型（如交易、仇恨言论等）

3.如何确定句子中术语含义。去除掉关键字，直接对上下文进行判断是否术语黑暗风格语料。使用 TF-IDF 主题模型对上下文进行判断，并集成多层感知机分类结果，对上下文风格进行判断。

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. MLM 所使用的术语与指代词上下文相似性和过去利用的术语在黑暗和普通语料库下的上下文差异性，两者并不冲突。后者可以作为 MLM 模型的后处理步骤，通过对比语料库差异，剔除掉一些不合理的预测结果，提高模型准确率。原 MLM 模型得到保留，同时通过调整剔除的相似度阈值，减少对原模型鲁棒性的影响。

2. 两个分类器分别专注不同的功能，前者负责对社交媒体数据进行大致分类，后者则负责在细粒度上对易与黑暗语料混淆的各种语料进行区分。显然对分类器进行分层后每个分类器所需完成的任务都更加简单。

3.语料风格分析。不同于方案二的目的是区分不同类型语料的上下文。本方案的目的是区分术语的无害含义时的上下文和黑暗含义时的上下文。作为刻意选择的黑暗术语，无害和黑暗含义相差较大，相应上下文差异也较大。词频是在差异较大时方便利用的显著特征，TF-IDF 能使用词频特征构建模型来对文本进行分类判断，而感知机能捕捉更多抽象的特征，两者集成将能更好利用上下文特征。方案一的 MLM 模型改为有监督方法后也能用于语义分类，但时间成本高。

技术路线：

1.黑暗术语发现

通过比较不同语料中词义差，对 MLM 的预测结果进行排除。调整后处理的阈值，在准确性和鲁棒性上达到平衡。将发现的新词加入构建的术语库中。

2.动态检测

收集从社交媒体中提取的数据，使用粗分类器进行粗筛，排除与黑暗语料风格差别较大的语料。再通过细分类器对易与黑暗语料风格易混淆的语料进行排除。

3.查询社交媒体数据中数据（以句子为单位）是否含有术语，选择带术语句子对其进行含义鉴别，使用 TF-IDF 主题模型对上下文进行判断，并集成多层感知机分类结果，对上下文风格进行判断，区分术语在上下文中表达的是有害含义还是无害含义。

四、本研究课题可能的创新之处（不少于 500 字）

1. 通过后处理 MLM 模型来提高模型表现，通过处理步骤，剔除掉一些不合理的预测结果，提高模型准确率。 过去的研究依赖于静态词嵌入，希望通过比较在词在不同的语料中上的差异来检测黑暗术语，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，准确度偏低。通过结合两类方法，使用过去方法作为 MLM 模型的后处理办法，尽可能保留 MLM 模型的鲁棒性，同时提高其准确度。

2. 通过粗-细分类器筛选黑暗语料。粗分类器负责筛选与黑暗语料差异较大的类型数据，细分类器负责区分容易与黑暗语料混淆的语料类型（如交易、仇恨言论等）。基于社交媒体数据特点筛选黑暗语料。黑暗语料在社交媒体数据中只占极小的一部分，相应标签分布呈长尾状态。基于机器学习的分类算法应用于长尾分布数据集时，识别效果并不好。粗-细分类器在图像领域有了一些应用，粗分类器负责提取全局特征，细分类器提取局部特征，与本方案想法类似，但基于图像的特征无法直接迁移。

3. 通过集成 tf-idf 与多层感知机确定术语含义。TF-IDF 能使用词频特征构建模型来对文本进行分类判断，而感知机能捕捉更多抽象的特征，两者集成将能更好利用上下文特征。MLM 模型虽然具有区分不同上下文的能力，但会带来大量的时间消耗，不满足即时性的要求。

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 python，需要熟悉的库有 numpy、pytorch，需要学习的模型主要有 MLM 模型、词嵌入模型、粗-细分类器等。

当前已经具备的试验条件，目前可以使用数据量足够的来自 reddit 的黑暗语料数据、机器学习库 sklearn 方便快速使用传统机器学习模型，方便快速配置 NLP 模型的 python 库 transformer。社交媒体数据可以使用来自 kaggle 平台提供的数据集，kaggle 主要为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台，其中还有大量用户自发上传的高质量数据集。可以通过收集整合这些数据得到足够的社交媒体数据。

当前主要缺少用于分析的与黑暗语料相似的仇恨言论等语料、用于后处理的词嵌入方案、对社交媒体的分类方案

2.拟解决途径

仇恨言论等语料可选取 Jigsaw Rate Severity of Toxic Comments 比赛中提供的人工标注的仇恨评论语料。后处理的词嵌入方案可以考虑 peters 等人在 NAACL 中发表的论文《Deep contextualized word representations.》中提出的 ELMO 模型，elmo 计算速度较快同时也能很好的根据上下文生成词义规避多义词问题。对社交媒体的分类方案，可以根据收集的社交媒体数据进行主题建模，人工划分出几个内容主题。具体到细分类器的分类类别的选择上，可以先用粗分类器进行分类，在分类结果上进行分析，选择出具体的易与黑暗语料混淆的语料类型。

学位论文工作计划

时间	研究内容	预期效果
2021 年 11 月-2021 年 12 月	阅读相关方向论文、选题	完成开题报告
2021 年 12 月-2022 年 3 月	通过后处理完成术语识别	可以根据提供的语料数据发现新的术语使用
2022 年 3 月-2022 年 6 月	分析社交媒体数据，完成筛选黑暗语料算法	在海量社交媒体数据中选取所需的黑暗语料
2022 年 6 月-2022 年 9 月	设计分类模型区分行话含义	能根据行话上下文快速区分行话无害含义和有害含义
2022 年 9 月-2022 年 10 月	应用程序框架搭建	完成运行并测试所有流程
2023 年 10 月-2023 年 1 月	论文写作	论文最终完成

	姓 名	职 称	导师类型	单位名称	职务
评定小组成员	金正平	副教授	硕导	北京邮电大学	组长
	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
	秦素娟	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