# 基于目标识别与主题引导对话的黑灰产威胁情报挖掘

罗双春[1], 黄 诚[1], 孙恩博[2]

[1]四川大学 网络空间安全学院 成都 中国 610207
[2]中国电子科技集团公司第三十研究所 成都 中国 610093

**摘要** 随着网络技术的不断发展, 在巨额利益的驱动下, 黑灰产活动日益泛滥, 黑灰产从业者利用互联网社交媒体和地下论坛进行业务推广, 如何挖掘更多的黑灰产威胁情报信息成为监管者打破网络空间治理攻防僵局、推动网络空间有效治理的关键一环。然而, 现有研究通过开发爬虫工具或利用开源数据进行被动式的数据分析, 难以获取全面、准确、实时的威胁情报信息。为此, 本文提出一种基于目标识别与主题引导对话的主动式黑灰产威胁情报挖掘方法, 能够从社交媒体群聊中自动识别黑灰产人员, 并采用主动引导对话的方式与其一对一交流, 挖掘威胁情报信息。首先, 根据黑灰产人员在群聊中的发言文本进行分类, 实现人员目标识别, 同时, 为使模型能有效理解黑灰产行话, 微调黑灰产领域词向量进行文本语义表征; 其次, 构建对话系统与黑灰产人员主动对话, 对话过程中通过识别其话语的意图, 引入基于规则匹配、场景记忆和深度学习三种策略自动化构建问答内容, 引导黑灰产人员暴露情报信息。实验结果表明, 本文提出的方法人员目标识别准确率达到 98.78%, 对话意图识别的准确率达到 90.80%, 并在真实场景下验证了方法的有效性。

**关键词** 地下产业; 威胁情报; 人员识别; 主题引导对话
**中图法分类号** TP DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.05.03

# Threat Intelligence Mining Based on Target Recognition and Topic-Guided Dialogue in Underground Markets

LUO Shuangchun[1], HUANG Cheng[1], SUN Enbo[2]

[1]School of Cyber Science and Engineering, Sichuan University, Chengdu 610207, China
[2]The 30th Research Institute of China Electronics Technology Group Corporation, Chengdu 610093, China

**Abstract** With network technology's continuous development, driven by substantial profits, underground market activities are becoming increasingly widespread, and the practitioners in this clandestine realm employ various Internet platforms, including social media and underground forums, to promote their illicit ventures. Consequently, unearthing a plethora of underground-market threat intelligence has emerged as a crucial imperative for regulatory bodies to break the deadlock between cyberattacks and cybersecurity within cyberspace governance, ultimately promoting more effective cyber governance practices. However, the existing research uses crawler tools or open source data through passive collection, making obtaining comprehensive, accurate, real-time threat intelligence information difficult. To this end, we propose an active underground-market threat intelligence mining method based on target identification and topic-guided dialog, which can automatically identify underground market personnel from social media group chats and communicate with them one-on-one in an active-guided dialog to mine threat intelligence information. Firstly, classify the text according to the speeches of the underground market personnel in the group chat to realize the personnel target identification, and at the same time, in order to make the model understand the underground market jargon effectively, fine-tune the underground market domain word vector for text semantic characterization; Secondly, construct a dialogue system to have an active dialogue with the underground market personnel, and in the course of the dialogue, by identifying the intent of their words, introduce three strategies based on rule matching, scene memory, and deep learning automatically construct the Q&A content to guide the underground market producers to expose much more intelligence information that concerned by the regulators. The experimental results unequivocally demonstrate the effectiveness of this method. The accuracy rate of personnel target recognition attains an impressive 98.78%, while the accuracy rate of dialogue intent recognition stands at 90.80%. The real-world deployment of the underground-market target recognition module and the intelligent dialogue module further substantiates the method’s efficacy when applied in practical scenarios.

**通讯作者:** 黄诚, 博士, 副教授, Email: codesec@scu.edu.cn。
本课题得到国家重点研发项目(No. 2021YFB3100500), 四川省科技厅重点研发项目(No. 2023YFG0162)资助。
收稿日期: 2023-06-05; 修改日期: 2023-10-09; 定稿日期: 2025-02-27

**Key words** underground market; threat intelligence; personnel identification; topic-guided dialogue

## 1 引言

近年来, 随着网络技术的不断发展, 网络黑灰产行为日益增多。不法分子通常会利用社交媒体来宣传他们的产品或服务以此来吸引购买者从而获得产品变现。例如, 黑灰产从业者在社交媒体发布消息, 宣传被泄露的用户名账号数据, 寻找买家进行交易从而获取利益[1]。手机黑卡、网络账号成为网络黑灰产产业的源头和基础[2]。通过采集分析这种地下交易活动产生的通信数据, 可以为打击网络违法犯罪活动提供有用的威胁情报。

现有研究主要通过开发爬虫工具或利用开源数据获取社交媒体和地下论坛上的文本数据[3], 分析犯罪人物[4]、犯罪供应链条[5]、行业术语[6]来挖掘威胁情报信息。该类方法采用机器学习和自然语言处理技术, 通过文本分类、实体识别、语义分析、相似度计算等来分析论坛文本数据, 检测网络犯罪主题, 识别潜在受害者, 确定可疑群体, 挖掘犯罪供应链[7]。现有的分析方法以被动式为主, 然而黑灰产人员在社交媒体和地下论坛发表的帖子很少暴露高价值的情报信息, 导致被动式分析在情报获取的及时性和有效性上存在挑战。例如, 在社交媒体贩卖手机 SIM 卡的人员只会以其拥有大量资源来进行宣传, 而不会透露更多的隐私信息(如卡的来源、联系方式等)。为了获取深层次的情报信息, 可以潜入黑灰产社交群组, 与其进行一对一的交流。但黑灰产团伙众多, 群聊数据量巨大, 采用人工难以实现这一目标。

为此, 本文提出了一种主动式的黑灰产威胁情报挖掘方法, 采用细粒度的人员目标识别方法, 从社交媒体群聊中自动识别黑灰产人员, 然后进行主动引导对话, 从而实现威胁情报挖掘。该方法包括人员目标识别与主动引导对话两个阶段。在人员目标识别阶段, 首先通过黑灰产领域词向量来表征文本数据, 并提取文本主题词进行语义增强, 采用文本分类模型自动识别黑灰产人员; 在主动引导对话对话阶段, 首先通过故事编排设计对话场景, 自定义引导话术模板, 通过识别黑灰产人员对话意图, 采用基于规则匹配、场景记忆和深度学习三种策略选择引导话术模板进行提问, 诱导黑灰产人员暴露深层次情报信息。综上所述, 本文的主要贡献如下:

(1) 提出一种主动式的黑灰产威胁情报挖掘方法, 能够从社交媒体群聊中自动识别黑灰产人员, 并进行主动引导对话, 从而收集威胁情报信息。

(2) 提出一种基于主题词增强的黑灰产人员目标识别方法, 针对短文对话缺少关键语义信息的问题, 微调黑灰产领域词向量, 并提取群聊文本主题词增强语义, 训练神经网络模型构建分类器, 实现黑灰产人员自动识别。

(3) 构建黑灰产情报自动提取系统, 采用意图与实体联合识别模型识别黑灰产人员的意图, 引入基于规则匹配、场景记忆和深度学习三种策略自动化构建问答内容, 从而诱导黑灰产人员暴露情报信息。

## 2 相关工作

本节从黑灰产治理相关技术和智能对话技术两个方面介绍相关工作。

### 2.1 黑灰产治理相关技术

随着黑灰产人员利用社交媒体进行交易越来越频繁, 研究人员利用社交媒体通信数据进行研究也越来越普遍, 主要集中在犯罪主题挖掘、关键人物识别、供应链识别、行业术语检测、虚假账号识别等方面。随着机器学习和自然语言处理技术的不断发展, 相关技术也被运用到黑灰产治理研究中。例如, Iqbal 等人[8]基于 WordNet 从可疑聊天日志中识别和提取实体信息, 通过分类模型将对话总结为不同的犯罪主题。Pastrana 等人[9]利用逻辑回归模型与 K-Means 聚类和社交网络分析相结合的方法, 识别地下论坛中的关键人物。Bhalerao 等人[5]采用自然语言处理技术对地下论坛帖子进行自动分类识别, 然后构建一个交互图, 并使用图遍历算法来发现相关产品购买和后续销售帖子的链接。Lv 等人[10]结合金融和社会特征, 构建了联邦学习模型用于检测金融诈骗账户, 检测效果好于逻辑回归与决策树模型; Xu 等人[11]采用图神经网络和图自监督学习用于微信虚假账号检测, 不仅实验指标优于现有方法, 而且在真实环境中也取得较好的泛化效果。此外, 喻海松[12]分析了手机黑卡、网络账号、网络流量、资金通道等黑灰产业链, 并从法律层面提出了标本兼治的方法。

总体来说, 现有工作在特定场景下能够较好地挖掘黑灰产威胁情报, 然而, 随着黑灰产业逐步呈现智能化、链条化、多样化, 使用手段越来越隐蔽, 单一场景挖掘到的情报可靠性、时效性、全面性难以满足黑灰产治理的需求, 因此迫切需要对现有技术和方法进行扩展和优化。

### 2.2 智能对话相关技术

为了获取更加全面可靠的威胁情报信息, 利用

智能对话与黑灰产人员进行主动对话是一个可行的解决方案。现有的对话系统通常可以分为任务型和闲聊型[13]。任务型对话系统是通过对话完成既定任务，其技术重点关注任务的完成，而闲聊型对话系统重点关注对话回答的流畅性、一致性、多样性。智能对话技术也被应用到不同领域，在金融服务方面，Yu 等人[14]利用 BERT 预训练模型开发了一个聊天机器人，用于处理金融投资服务中客户提出的问题，该机器人可以处理客户提出的多种问题，并在超出能力时将不相关或不确定的问题提交人工处理；在软件开发方面，Yuan 等人[15]开发了 API 聊天机器人，开发者可以通过该机器人从 API 文档获取重要信息，Abdellatif 等人[16]使用聊天机器人来回答开发人员软件开发和维护中最常见的一些问题；在网络安全方面，Mcdermott 等人[17]提出了一种用于检测网络异常流量的会话代理，Franco 等人[18]提出一个安全聊天机器人，用于网络安全规划与管理。然而，现有智能对话系统不能直接用来与黑灰产人员进行对话。此外，Wang 等人[19]提出通过聊天机器人与从事电子商务的人员进行自主聊天来自动收集相关情报信息，然而，该机器人基于有限状态机实现，难以处理复杂的对话逻辑。为了能够与黑灰产人员进行自然的对话，获取全面、可靠、及时的威胁情报信息，必须能够对不同类型的黑灰产人员进行识别，针对性地构建对话模型，理解其回答话语，采用有效的策略引导其暴露更多威胁情报信息。

## 3 方法设计

为实现对黑灰产威胁情报智能、高效、准确地挖掘，本文提出了一种基于目标识别与主题引导对话的主动式威胁情报挖掘方法，主要由人员目标识别和主题引导对话两个部分组成，方法框架如图 1 所示。

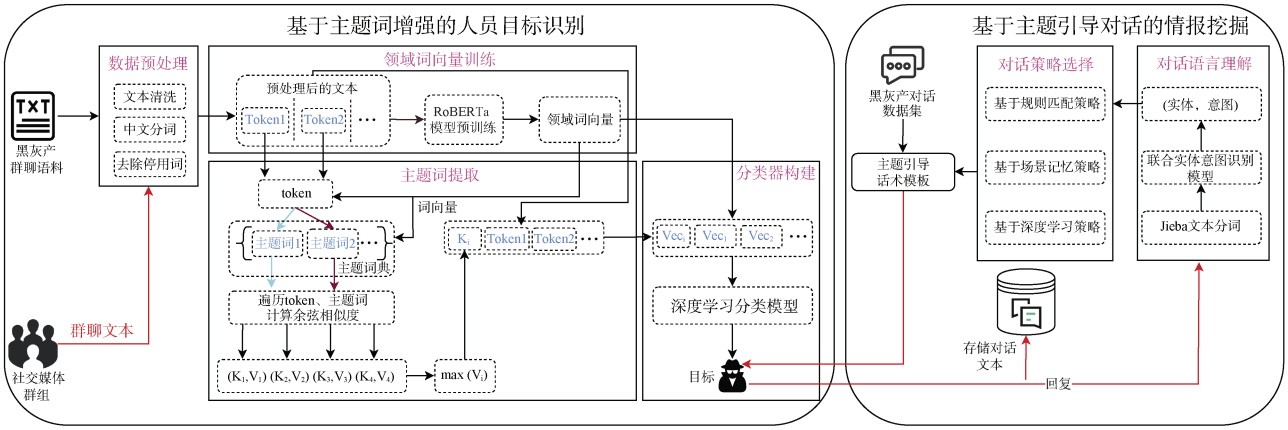


**图 1 方法框架图**

**Figure 1 Framework of the method**

具体来说，首先从黑灰产社交媒体群聊中采集群聊数据，进行数据预处理，然后利用 RoBERTa[20]模型微调基于中文词的黑灰产领域词向量，用来表征人员对话语义信息。接下来，利用领域词向量和自定义的主题词词典从对话文本中提取主题词，与原始对话文本拼接后用领域词向量表示为词向量后输入神经网络模型进行分类，输出人员类型。识别人员类型后开启主动提问，通过识别人员目标回答的话语中包含的意图和实体，引入基于规则匹配、场景记忆、深度学习三种策略选择引导话术模板进行下一轮提问，从而挖掘更多网络黑灰产威胁情报信息。

### 3.1 数据采集与处理

利用开源工具 Telethon①和 Mirai②采集社交媒体黑灰产群聊中的聊天记录内容，获得原始数据集。

数据采集。首先，通过关键字搜索黑灰产相关的群组，使用社交媒体账号加入群聊中；然后，利用开源工具 Telethon 和 Mirai-API-HTTP③创建客户端程序，持续采集群聊数据；最后，将采集的数据存入数据库中。收集的数据包括群组名称、群 ID、群头像、创建日期、群成员 ID、群成员昵称、群成员头像、发言文本、发言时间。

数据处理。由于采集到的群聊文本不是纯文本数据，还包含表情图标、特殊符号、繁体中文、网址

① https://github.com/LonamiWebs/Telethon
② https://github.com/mamoe/mirai
③ https://github.com/project-mirai/mirai-api-http

链接等冗余信息, 需要对数据预处理。具体来说, 首先删除文本中的表情图标、特殊符号、标点符号; 其次构建字典实现繁体中文和简体中文的转换。最后使用 Jieba[①]分词工具对文本进行分词, 并删除停用词。

### 3.2 基于主题词增强的人员目标识别

黑灰产人员通常在社交媒体群组中群发消息来宣传其售卖的产品或提供的服务, 通过识别他们群发的消息内容, 可以判断其人员类型。

#### 3.2.1 领域词向量微调

黑灰产人员在社交媒体群聊中会使用暗语来交流, 以此躲避监管。这类暗语通常是短语形式, 由 2 个或 2 个以上的字组成。如, “水房”表示“专门的洗钱集团”, “鹅场”表示“微信支付”。神经网络模型处理文本数据时需要将其转换为数值表示, 以便能够直接进行计算。词向量模型的作用就是将文本数据转换为向量表示, 现有的中文词向量模型多以单个字为基本单位, 将语句拆分为一个一个的字进行向量表示[21]。由于黑灰产人员采用的暗语多以词为单位, 为了能够更好表征黑灰产人员的对话文本语义信息, 本文在 RoBERTa 模型上微调一个基于中文词的黑灰产领域词向量。在将黑灰产语料输入模型之前, 需要将文本切分为中文词。分词过程按以下步骤进行:

(1) 根据公开资料[②]收集整理一个包含黑灰产暗语的字典 $D_1$;

(2) 将中文字和常用中文词和字典 $D_1$ 加入词表 $V_1$;

(3) 输入一个文本语句, 利用 Jieba 分词工具和字典 $D_1$ 进行分词, 得到文本序列 $L=[w_1,w_2,\cdots,w_i]$;

(4) 遍历 $L$, 如果 $w_i$ 在 $V_1$ 中, 则保留, 否则, 将其按字分开;

(5) 将每个 $w_i$ 的分词结果有序拼接起来, 作为最后的分词结果。

获得分词后, 将文本序列输入 RoBERTa 模型, 并采用随机掩码策略进行微调。初始化阶段, 将每个词切分为单个字, 然后用字向量的平均值作为词向量的初始化。

#### 3.2.2 主题词提取

黑灰产人员群发的消息文本中短文本占比大, 而短文本语义稀疏, 因此采用微调的黑灰产领域词向量从消息文本中提取主题词加并入原始文本来增强语义信息, 提升人员目标识别的效果。首先, 自定义一个主题词字典 $D_2$, 主题词是能够区分不同人员类型的词语或短语, 对于不同人员目标类型具有代表性且词语的相似性很低。如: 短语“注册卡”可以成为“卡商”人员类型的主题词, 主题词字典后续通过专家经验人工进行迭代优化。其次, 对消息文本进行预处理和分词, 加入词表。然后, 用领域词向量对词语进行向量化表示, 遍历词表分别计算与主题词的余弦相似度[22]。两个词语的词向量用 $E_m=[e_{m,1},e_{m,2},\cdots,e_{m,k}]$ 和 $E_n=[e_{n,1},e_{n,2},\cdots,e_{n,k}]$ 表示, 则余弦相似度可用公式(1)计算, 其中 $k$ 表示词向量的维度。

$$\mathrm{Cos}(E_m,E_n)=\frac{E_m\cdot E_n}{|E_m|\times|E_n|}=\frac{\sum_{i=1}^{i=k}(e_{m,i}\times e_{n,i})}{\sqrt{\sum_{i=1}^{i=k}(e_{m,i})^2}\times\sqrt{\sum_{i=1}^{i=k}(e_{n,i})^2}} \tag{1}$$

最后, 输出与词表中词语相似度最大的主题词作为原始文本的主题词。提取主题词的具体步骤如算法 1 所示。

**算法 1**.主题词提取算法

输入: 主题词字典 $D_2$, 原始文本 $T$
输出: 主题词 $K$
1 对原始文本 $T$ 进行分词, 获得词表 $V_2$
2 加载主题词字典 $D_2$
3 foreach $v$ in $V_2$ *do*
4 foreach $d$ in $D_2$ *do*
5 $Sim_k=\mathrm{Cos}(E_v, E_d)$
6 $Sim$.append($Sim_k$)
7 end
8 end
9 $K=\max(Sim)$

#### 3.2.3 人员目标识别

将黑灰产人员目标识别看作文本分类任务, 定义黑灰产人员群聊文本集合 $S=\{s_1,s_2,\cdots,s_n\}$, $s_i$ 为黑灰产人员在社交媒体群聊中发表的会话文本, $n$ 为数据集大小, 定义人员类别集合 $Y=\{y_1,y_2,\cdots,y_C\}$, $y_i$ 为黑灰产人员的类型标签, $C$ 为标签个数。模型输入为 $s_{n+1}$, 输出为 $y_i\in Y$。黑灰产人员目标识别方法主要包括向量表示层、语义学习层和类别计算层, 方法框架如图 2 所示。

① https://github.com/fxsjy/jieba

② https://wiki.y1ng.org/0x7_%E9%BB%91%E7%81%B0%E4%BA%A7%E7%A0%94%E7%A9%B6/7x7_%E9%BB%91%E8%AF%8D_%E9%BB%91%E8%AF%9D%E5%A4%A7%E5%85%A8/

**表 1 引导话术模板示例**

**Table 1 Example of guide speech template**

| 情报类型 | 情报挖掘需求 | 话术模板 | 情报目的 |
|---|---|---|---|
| 浅层情报 | 打招呼 | 你好, 有卡卖吗? | 提出购买需求, 引起黑灰产人员注意 |
| | 询问卡价格 | 这个卡每张多少钱呢? | 挖掘黑卡价格 |
| | 询问卡类型 | 你们卖的都有什么类型的卡啊? | 挖掘售卖的黑卡类型 |
| | 询问发货时间 | 现在买的话什么时候可以发货? | 挖掘发货时间 |
| | 询问卡号 | 问一下你的卡都有哪些号码段啊? | 挖掘黑卡号段 |
| | 询问卡运营商 | 这个卡是什么运营商的卡啊? | 挖掘黑卡的运营商 |
| | 询问卡归属地 | 这个卡归属地是哪里啊? | 挖掘黑卡归属地 |
| | 询问卡库存量 | 这个卡可以开多少张啊? | 挖掘黑卡库存量 |
| | 询问卡发货地点 | 这个卡从哪里发货的啊? | 挖掘黑卡的来源地 |
| | 询问卡应用场景 | 这个卡支持哪些平台啊? | 挖掘黑卡的应用场景 |
| 深层情报 | 询问支付平台 | 通过什么方式支付啊? | 挖掘黑卡支付平台 |
| | 询问销售方式 | 你们是代理还是公司直销? | 挖掘黑灰产人员售卖方式 |
| | 询问开卡平台 | 可不可以给个开卡平台啊? | 挖掘黑卡销售渠道 |
| | 询问售后服务 | 这种卡后期有什么售后服务吗? | 挖掘黑卡售后服务信息 |
| | 询问联系方式 | 方便给个联系方式? | 挖掘黑灰产人员联系方式 |
| | 询问用户群体 | 这个卡什么人用得比较多呢? | 挖掘黑卡用户群体 |
| 隐秘情报 | 询问卡的销量 | 这种卡一天能卖多少张呢? | 挖掘黑卡销量情况 |
| | 询问接码平台 | 你有没有合作的接码平台呢? | 挖掘与黑灰产人员合作的接码平台 |
| | 询问监管措施 | 这种卡有什么使用限制呢? | 挖掘黑灰产人员掌握的监管措施 |

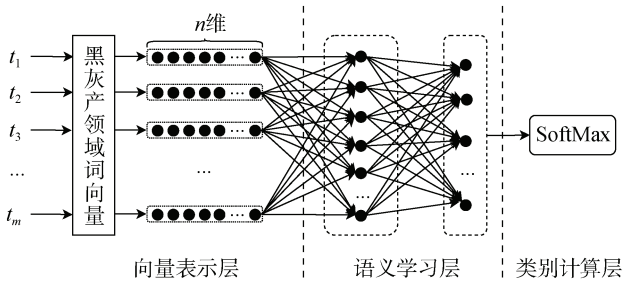


**图 2 人员目标识别框架图**

**Figure 2 Framework of target identification**

向量表示层。向量表示层是将文本语句转化为神经网络模型可以直接计算的数值化向量。采用微调后的黑灰产产领域词向量来对文本语句进行向量表示。首先, 对待分类的文本语句 $s$ 进行 Jieba 分词和主题词提取, 获得文本序列 $T=[t_1,t_2,\cdots,t_m]$,通过领域词向量表征后获得到语句 $s$ 的向量表示矩阵 $R_{m\times k}$。其中, $m$ 表示对话文本分词和加上主题词后的长度; $k$ 表示词向量的维度。

语义学习层。语义学习层是在获得文本语句句子级词向量矩阵的基础上, 通过输入神经网络进行计算, 学习更深层次的语义信息和上下文信息并输出。

类别计算层。类别计算层将语义学习层输出的特征向量通过 Softmax 分类器计算语句 $s$ 在分类标签中的概率向量 $P=[p_1,p_2,\cdots,p_C]$, 选择概率最大值所代表的类别标签作为最终输出的类别。Softmax 分类器将输入向量的第 $i$ 项转化为概率 $P_i$, 计算方法如公式(2)所示。

$$P_i=\frac{e^i}{\sum_{c=1}^{C}e^c},c=1,2,\cdots,C \tag{2}$$

其中, $C$ 表示标签类别数。采用交叉熵 Loss 作为损失函数, 训练过程中, 将交叉熵最小化, 计算过程如公式(3)所示。

$$Loss=-\sum_{i=0}^{C-1}y_i\log(p_i) \tag{3}$$

其中, $y_i$ 是样本标签的 one-hot 向量表示, 当样本属于第 $i$ 类别时 $y_i=1$, 否则 $y_i=0$。

### 3.3 基于主题引导对话的情报挖掘

完成黑灰产人员目标识别后, 主动发送消息与其进行对话来挖掘威胁情报, 话术模板如表 1 所示。对话过程包括构建引导话术模板、对话语言理解和对话策略选择。将对话过程建模为实体、意图、动作、策略和回复 5 元组。其中, 实体是黑灰产人员对话响应内容中包含的目标情报信息, 意图是对黑灰产人员响应内容进行分析, 确定其对话的真实意图, 动作是情报挖掘需要询问的问题集合, 策略是根据当前对话状态选择下一步系统执行动作的一个映射

函数, 回复则是根据系统执行动作选择引导话术模板对黑灰产人员进行提问。

### 3.3.1 引导话术模板构建

根据群聊文本数据集和情报挖掘需求, 人工构建引导话术模板。引导话术模板由各种问句组成, 是回复黑灰产人员的消息集合。当对话系统识别黑灰产人员意图并成功预测下一个动作时, 根据系统动作, 随机选择话术模板进行回答。防止系统回复消息过快, 随机等待时间阈值进行消息回复。

### 3.3.2 对话语言理解

在与黑灰产人员对话过程中, 通过实体抽取和意图识别来理解黑灰产人员的对话内容, 并通过意图和实体来判断下一轮对话选择的引导话术模板。例如, 黑灰产人员回答“可以微信支付”。这句话包含的实体为“微信支付”, 包含的意图是“回答支付平台”。意图和实体示例如表 2 所示。

**表 2 意图与实体示例**

**Table 2 Examples of intentions and entities**

| 黑灰产人员回答文本 | 实体 | 意图 |
|---|---|---|
| 你想要什么样的卡? | | 回答购买需求 |
| 联通卡 12 元一张 | 价格: 12 元 | 回答卡价格 |
| 全是流量卡 | 类型: 流量卡 | 回答卡类型 |
| 在 48 小时之前发货 | 时间: 48 小时 | 回答发货时间 |
| 有电信 141 号段 | 号段: 141 号段 | 回答卡号 |
| 我们的卡有用友的 | 运营商: 用友 | 回答运营商 |
| 我们卡是天津市的 | 归属地: 天津市 | 回答归属地 |
| 库存还有 25 张 | 库存量: 25 张 | 回答库存量 |
| 黑龙江发货 | 发货地点: 黑龙江 | 回答发货地点 |
| 支持注册令币平台 | 应用场景: 令币 | 回答应用场景 |
| 可以微信支付 | 支付平台: 微信支付 | 回答支付平台 |
| 我们就是个人销售 | 销售方式: 个人 | 回答销售方式 |
| 可以关注信阳卡邦 | 开卡平台: 信阳卡邦 | 回答开卡平台 |
| 服务电话: 4000121 | 服务电话: 4000121 | 回答售后服务 |
| 添加微信号: xx1212 | 联系方式: xx1212 | 回答联系方式 |
| 很多做自媒体的在用 | 用户: 自媒体 | 回答用户群体 |
| 每天出货 1000 张以上 | 销量: 1000 张 | 回答卡的销量 |
| 云短信接码平台 | 接码平台: 云短信 | 回答接码平台 |
| 不要实名, 不打电话 | | 回答监管措施 |

实体抽取本质上是序列标注任务, 而意图识别本质上是文本分类任务, 基于句子层面所理解的文本语义, 确定黑灰产人员意图所属的类别。采用一种实体与意图联合识别的深度学习方法[23], 来识别对话过程中的实体和意图, 具体过程如算法 2 所示。

**算法 2**.实体与意图联合识别算法

输入: 对话文本集合 $X=\{X_1,X_2,\cdots,X_n\}$;意图集合 $I=\{I_1,I_2,\cdots,I_n\}$; 实体集合 $E=\{E_1,E_2,\cdots,E_n\}$

输出: 意图 $I$; 实体 $E$

1 foreach $x$ *in* $X$

2 对 $x$ 进行分词, 获得文本序列 $T=[token_1, token_2,\cdots,token_n]$

3 在文本序列 $T$ 后面添加分类 *token[cls]*, 随机选择 $token_i$

替换为 *[mask]*, 得到新的序列 $W=[token_1, mask,\cdots,token_n,cls]$

4 foreach $w$ *in* $W$

5 用 one-hot 方式编码 $w$ 获得稀疏特征输入一个全连接层

6 同时用 BERT 编码 $w$ 获得稠密特征输入一个全连接层

7 拼接全连层输出的稀疏特征与稠密特征

8 将拼接输出向量输入一个 2 层的 Transformer 结构对整个序列进行编码

9 将 Tranformer 层输出序列输入 CRF 层, 输出预测实体 $E$, 采用对数似然函数作为实体提取的损失函数 $L_E$

10 将 *token*[*cls*]输出向量与意图标签向量进行相似度计算, 输出意图 $I$, 采用交叉熵函数作为意图识别的损失函数 $L_I$

11 用 *token[mask]*与真实 $token_i$ 向量进行点积计算相似度, 采用交叉熵函数作为 MASK 任务的损失函数 $L_M$

12 计算总损失 $L_{total}=L_E+L_I+L_M$

首先, 将对话文本分词为文本序列, 一个字为一个 *token*; 采用 one-hot 编码方式获取 *token* 的稀疏特征, 采用 BERT 编码方式获取 *token* 的稠密特征, 将稀疏特征和稠密特征进行拼接; 通过 CRF[24]层来预测实体序列, 使用句子向量用于在所有可能的意图标签上排序; 在训练中随机屏蔽输入词符, 通过采用向量点积计算相似性, 进行损失反馈。意图分类损失计算如公式(4)所示。

$$L_I = -<S_I^+ - \log(e^{S_I^+} + \sum_{\omega_I^-}^{\infty} e^{S_I^-})> \tag{4}$$

通过 Transformer 层输出的[*cls*]词符 $a_{cls}$ 和意图标签 $y_{intent}$ 被嵌入一个相同的语义向量空间中, $h_{cls}=E(a_{cls}), h_{intent}=E(y_{intent})$。使用点积损失最大化与目标标签 $y_{intent}^+$ 的相似性 $S_I^+=h_{cls}^T h_{intent}^+$, 并最小化与负样本 $y_{intent}^-$ 的相似性 $S_I^-=h_{cls}^T h_{intent}^-$。其中, 求和是在一组负样本 $\omega_I^-$ 上进行的, 平均值 $<\bullet>$ 在所有

样本中进行。

在训练过程中, 采用随机掩码策略, 随机选择15%的 *token* 进行替换, 其中, 将选择的 *token* 中的

70%替换为*[mask]*, 10%随机替换, 20%保留为原始 token。MASK 损失计算如公式(5)所示。

$$L_M = -< S_M^+ - \log(e^{S_M^+} + \sum_{\omega_M^-}^{\infty} e^{S_M^-}) > \tag{5}$$

每个被选定*[mask]*的字符 $y_{token}$ 与 Transformer 层输出的字符 $a_{mask}$，被嵌入到一个相同语义向量空间中 $h_{mask} = E(a_{mask})$，$h_{token} = E(y_{token})$。使用点积损失最大化与目标标签 $y_{token}^+$ 的相似性 $S_M^+ = h_{mask}^T h_{token}^+$ 并最小化与负样本 $y_{token}^-$ 的相似性 $S_M^- = h_{mask}^T h_{token}^-$。其中, 求和是在一组负样本 $\omega_M^-$ 上进行的, 平均值 $<\bullet>$ 在所有样本中进行。

实体抽取损失计算如公式(6)所示。

$$L_E = L_{CRF}(a, y_{entity}) \tag{6}$$

式中 $L_{CRF}(\bullet)$ 表示 CRF 的对数似然可能性, $a$ 表示输入序列, $y_{entity}$ 表示实体标签序列。

模型总损失计算如公式(7)所示。

$$L_{total} = L_I + L_M + L_E \tag{7}$$

#### 3.3.3 对话策略选择

对话策略是将实体和意图映射到系统动作的一个函数。在生成对应函数之前, 采用故事编排的方式定义对话流程, 按照情报挖掘需求以及黑灰产人员目标类型, 为每种对话编排不同故事。每个故事由意图和动作组成, 是对话流程的抽象表示, 涵盖不同的对话顺序, 故事编排按照循序渐进, 由浅入深的原则, 通过浅层情报挖掘, 营造对黑灰产业非常感兴趣的假象, 降低黑灰产人员警惕性, 而后基于深层情报信息对话, 获取更多情报, 对话结束前, 基于隐秘信息进行询问, 获取隐秘情报信息。基于故事编排, 采用基于规则匹配、基于场景记忆、基于深度学习三种对话策略。

(1) 基于规则匹配策略。根据先验知识和业务逻辑, 自定义规则, 当输入意图是 $I_1$ 时, 直接返回系统动作为 $A_1$。例如, 当黑灰产人员的意图是“回答卡价格”时, 系统执行的动作为“询问卡类型”, 这种规则是可以确定, 但通常规则只能解决某几轮, 无法覆盖复杂多轮对话。

(2) 基于场景记忆。通过编排的故事执行顺序, 检查当前对话是否与故事匹配, 如果匹配, 它将从训练数据的匹配故事中预测下一个动作。如某轮对话输入的是意图 $I_2$, 从历史故事中, 发现意图 $I_2$ 接下来的系统动作都是 $A_2$, 则输出系统动作 $A_2$。

(3) 基于深度学习。采用一种基于 Transformer[25] 架构的深度学习对话策略[26]用于下一动作预测。模型输入用户意图和实体, 然后通过自注意力层编码, 接入池化层, 并通过 Dropout 层防止过拟合, 最后通过 Softmax 层获取每个行为的概率, 最终返回概率最大的动作作为下一个对话动作。

## 4 实验与分析

### 4.1 实验设置与数据

实验环境。本文实验在安装 Ubuntu 22.04.1 LTS 操作系统的服务器进行, 配置 2.30GHz 的 40 核 Intel Xeon CPU 和 256GB 内存, 显卡为 NVIDIA RTX3090, 显卡驱动为 CUDA 11.6。神经网络模型基于 Pytorch 1.12 构建, 编程语言采用 Python 3.8。

分类数据集。从黑灰产聊天群中爬取 60 万条聊天记录数据, 根据黑灰产人员发言内容, 去除重复内容, 并根据人员类型进行手工标注, 用于人员目标识别模型的训练, 标注的人员类型包括卡商、号商、料商、接码、赌博、色情 6 类, 如表 3 所示。

**表 3 人员目标识别数据集分布**

**Table 3 Distribution of target recognition dataset**

| 人员标签 | 训练集(条) | 测试集(条) | 验证集(条) |
|---|---|---|---|
| 号商 | 2746 | 342 | 343 |
| 卡商 | 968 | 120 | 120 |
| 料商 | 2098 | 261 | 262 |
| 接码 | 1752 | 218 | 219 |
| 赌博 | 3892 | 486 | 486 |
| 色情 | 2041 | 254 | 255 |

分类数据集主要是短文本, 文本长度小于 30 个字符的占比 75.85%, 特别是小于 15 个字符的文本占所有文本的 50%, 具体如图 3 所示。

对话数据集。通过关键字搜索获取与黑灰产相关的群组, 然后使用社交媒体账号加入群聊中, 判断群成员所属黑灰产类别后, 假扮感兴趣的购买者与其进行对话, 并收集聊天日志存入数据库中。收集完数据后, 将聊天日志按照一问一答的方式进行梳理, 同时, 对群聊文本原始数据集进行整理, 对黑灰产人员回答文本进行意图和实体标注, 最后形成多轮对话数据集。

主题词字典。黑灰产人员有着不同的业务分工, 采集整理不同类型常用词可以更加快速、准确对黑

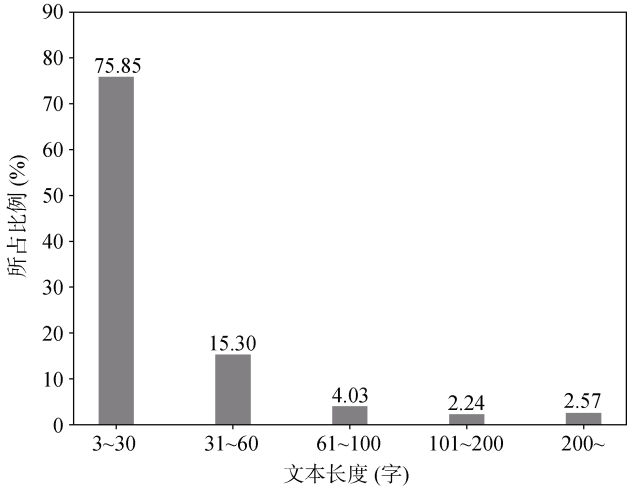


**图 3　文本字符占比**

**Figure 3　Text character ratio**

灰产人员进行分类。在不同的业务领域交流中, 存在不同的常用词语, 例如, 在从事各类手机卡售卖的黑灰产人员交流中会大量出现“注册卡”“流量卡”, 为此将其收录到主题词字典中。通过计算分类数据集中不同类别黑灰产人员群聊文本中出现次数最多的 30 个词语, 且人工判定删除与黑灰产人员类别不相关的词语, 总共收录 6 类 100 个主题词, 如表 4 所示。

**表 4　主题词示例**

**Table 4　Examples of subject words**

| 人员标签 | 主题词数量(个) | 主题词示例 |
|---|---|---|
| 号商 | 20 | 号商、账号、白号 |
| 卡商 | 15 | 卡商、黑卡、注册卡 |
| 料商 | 16 | 内料、卡料、洗料 |
| 接码 | 12 | 接码、短信、验证码 |
| 赌博 | 16 | 菠菜、埋雷、猪蹄 |
| 色情 | 21 | 黄色、情色、偷拍 |

## 4.2　实验设计与结果分析

### 4.2.1　领域词向量有效性实验

评价指标。计算 Precision-at-K(*P@K*)值作为领域词向量有效性的评价指标。选择 20 个候选词, 然后用不同词向量模型计算与候选词相似度最大的 $K$ 个词, 人工筛查这 $K$ 个词语与候选词相关的个数, 当判定两个词相关时取值为 1, 否则为 0。计算方法如公式(8):

$$I_{w_0}(w)=\begin{cases}1, w_0\text{与}w\text{相关}\\ 0, \text{其他}\end{cases} \tag{8}$$

其中, $w$ 是输入词, $w_0$ 是候选词。

$P@K$ 值计算方法如公式(9)所示:

$$P@K(w_q)=\frac{\sum_{i=1}^{i=K} I_{w_q}(w_{(i)}^{q})}{K} \tag{9}$$

其中, $w_q$ 是遍历所有候选词, $w^q_{(i)}$是通过相似性计算获得的 $K$ 个词中的某一个词。

最后采用 20 个候选词 $P@K$ 值的平均值作为评价性能指标, 计算方法如公式(10)所示:

$$\overline{P@K}=\frac{\sum_{i=1}^{i=20} P@K(w_{qi})}{20} \tag{10}$$

实验步骤。首先, 利用黑灰产暗语字典对群聊数据进行分词操作, 采用 MLM(Masked Language Model)方式进行预训练, MASK 比例采用 15%, 在这 15%选出的部分中, 将其中的 80%替换成*[mask]*, 10%替换成一个随机的 *token*, 剩下的 10%保留原来的 *token*, 并用字向量的平均作为词向量的初始化值。在 RoBERTa 模型上继续进行微调, 文本最大序列长度设置为 256, 学习率设置为 5e–5, batch_size 设置为 32, 词表最大长度设置为 30000, epoch 设置为 50, steps_per_epoch 设置为 2000, 其他参数默认。

领域词向量对理解黑灰产人员对话内容至关重要。实验中, 选择 20 个黑灰产领域的常用的行业术语作为候选词, 如“猪蹄”“注册卡”“跑分”等, 然后利用模型生成的词向量来评估领域词向量的有效性。选取 2 个不同的词向量模型与黑灰产领域词向量进行对比实验。

Word2Vec[27]: Word2Vec 是 Google 公司在 2013 年提出的用于计算词向量的模型, 实验中采用 Skip-gram 方式在群聊数据集语料上进行训练, 获得词向量。

WoBERT[28]: WoBERT 是以中文词为基础预训练的词向量, 经过大量中文语料训练, 涵盖常用中文词语的词向量, 但没有使用黑灰产领域语料训练。

结果分析。领域词向量微调训练过程中准确率和损失变化如图 4 所示。

从图 4 中可以看到, 领域词向量训练模型在经过 50 轮微调训练后 MLM 任务准确率达到 0.9766, 损失为 0.1277。并且在经过 10 轮微调训练后, 模型的准确率和损失都达到了相对稳定的状态, 因为领域词向量模型是在 RoBERTa 模型上进行的再训练, 而 RoBERTa 模型已经经过大量语料训练, 再使用黑灰产领域语料训练时, 能够很快收敛。词向量有效性结果如表 5 所示。由实验结果可知, 领域词向量的表现最好, WoBERT 性能次之, Word2Vec 的性能最低。这是因为基于深度学习技术的词向量模型, 如 WoBERT,

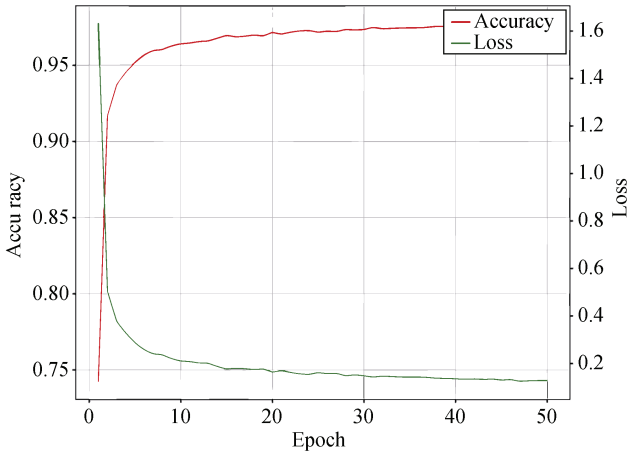


**图 4　领域词向量微调过程准确率与损失曲线**

**Figure 4　Accuracy and loss curve of domain word vector fine-tuning process**

**表 5　不同词向量有效性结果**

**Table 5　The validity results of different word vectors**

| 模型 | P@10 | P@20 | P@30 | P@50 |
|---|---|---|---|---|
| Word2Vec | 0.31 | 0.27 | 0.25 | 0.20 |
| WoBERT | 0.68 | 0.65 | 0.62 | 0.56 |
| 领域词向量 | 0.80 | 0.78 | 0.75 | 0.65 |

可以根据训练语料中词语在不同语境下捕捉动态语义信息, 并在实际下游任务动态调整词语的词向量, 获得更好的语义表达, 而 Word2Vec 只生成静态单词向量, 表示整个语料库中单词的上下文的平均值, 因此无法捕捉上下文语义信息。而由于 WoBERT 使用的是常规中文语料训练, 缺少黑灰产领域语境中行业术语的语义, 只能在大规模和广泛的语料库中理解单词的原义, 而无法理解黑灰产领域行话。

### 4.2.2　黑灰产人员目标识别实验

评价指标。在黑灰产人员目标识别实验中, 采用准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、*F*1 值(*F*1-*Score*)作为评价指标。

(1) 准确率: 代表所有类别都预测准确的数量占比, 计算方法如公式(11)所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

(2) 精确率: 代表正确预测为正例占全部预测为正例的比例, 计算方法如公式(12)所示。

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

(3) 召回率: 代表正确预测为正例的占所有实际为正例的比例, 计算方法如公式(13)所示。

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

(4) *F*1 值: 综合精确率和召回率, 计算两个指标的调和平均数, 代表模型的整体效率, 计算方法如公式(14)所示。

$$F1 = \frac{2 \times Precison \times Recall}{Precison + Recall} \tag{14}$$

实验步骤。利用黑灰产领域词向量表征黑灰产人员对话文本语义, 通过构建神经网络模型进行文本分类, 输出黑灰产人员目标类型。采用不同神经网络模型进行对比实验, 并利用领域词向量和主题词增强进行多种消融实验。实验选择的神经网络模型如下。

(1) TextCNN[29]: 利用卷积神经网络 CNN 处理文本分类任务, 通过词嵌入层将文本向量化, 然后通过卷积层提取文本特征, 通过池化层, 将向量转化为定长表示, 最后通过全连接层输出每个类别的概率。

(2) TextRNN[30]: 利用循环神经网络 RNN 处理文本分类任务, 通过词嵌入层将文本向量化, 利用 BiLSTM 捕获变长且双向的上下文信息, 最后通过 Softmax 层输出标签类别。

(3) TextRCNN[31]: 结合卷积神经网络(CNN)和循环神经网络(RNN)的一种文本分类方法, 使用循环结构捕获上下文信息, 并且使用卷积神经网络构造文本表示, 最后通过 Softmax 层输出类别。

(4) DPCNN[32]: 一种基于 Word-Level 级别的神经网络模型, 由于 TextCNN 不能通过卷积获得文本的长距离关系依赖, 而 DPCNN 通过不断加深网络, 可以抽取长距离的文本依赖关系。

(5) RoBERTa[20]: 基于 BERT[33]模型架构采取改进方法, 采用动态掩码语言模型(Dynamic Masking Language Model)和去掉下一句话预测(Next Sentence Prediction, NSP)模式, 增加训练 Batch Size, 采用基于字节的单词表示方式(Byte-Pair Encodeing, BPE), 在更大规模语料上进行预训练, 可以进行各种各样的下游任务, 并取得较好结果, 本实验中用该模型直接进行文本分类任务。重复多次实验后, 选取参数的最优值进行实验, 实验参数设置如表 6 所示。

**表 6　实验参数设置**

**Table 6　Experimental parameter settings**

| 参数 | 说明 | 取值 |
|---|---|---|
| learning_rate | 模型学习率 | 5e-6 |
| batch_size | 一次训练选取的样本数 | 0.65 |
| epoch | 训练轮次 | 100 |
| maxlen | 文本处理最大长度 | 256 |
| optimizer | 模型优化器 | Adam |
| loss | 损失函数 | cross_entropy |

结果分析。不同分类模型对人员目标识别的效果如表 7 所示。从实验结果可以看出，在 RoBERTa 词向量下，选取的 5 个模型中，黑灰产人员目标识别准确率最高为 95.24%，最低为 93.68%，这说明从黑灰产群聊中识别出目标人员类型是可行的；当使用黑灰产领域词向量后，每个模型分类准确率都有提升，且在 TextRNN 模型上人员目标识别准确率提升到了 98.44%，所有模型中准确率最高提升了 4.62%；进一步在采用主题词增强后，效果有进一步的提升，人员目标识别准确率最高提升到了 98.78%，最高提升 4.82‰ 通过实验数据可以看出，经过领域词向量和主题词增强后的分类模型在理解黑灰产人员的行话和短文本方面表现出显著效果，实验结果表明，提出的方法能够有效地从黑灰产社交媒体群聊中识别出黑灰产人员类型。

**表 7　不同分类模型目标识别效率**

**Table 7　The target recognition efficiency of different classification models**

| 模型 | 准确率(%) | 精确率(%) | 召回率(%) | *F1* 值(%) |
|---|---|---|---|---|
| RoBERTa | 93.68 | 94.52 | 97.17 | 92.97 |
| RoBERTa_CNN | 94.50 | 94.70 | 92.29 | 93.37 |
| RoBERTa_DPCNN | 95.11 | 94.70 | 93.56 | 94.05 |
| RoBERTa_RCNN | 94.16 | 94.23 | 92.26 | 93.14 |
| RoBERTa_RNN | **95.24** | 94.62 | 94.34 | 94.41 |
| RoBERTa* | 98.30(+**4.62**) | 97.50 (+2.98) | 98.50 (+1.33) | 97.98(+5.01) |
| RoBERTa*_CNN | 98.23(+3.73) | 97.44(+2.74) | 98.10(+5.81) | 97.76(+4.39) |
| RoBERTa*_DPCNN | 97.69(+2.58) | 97.27(+2.57) | 98.14(+4.58) | 97.67(+3.62) |
| RoBERTa*_RCNN | 97.62(+3.46) | 97.06(+2.83) | 97.91(+5.65) | 97.46(+4.32) |
| RoBERTa*_RNN | **98.44**(+3.20) | 97.96(+3.34) | 98.55(+4.21) | 98.25(+3.84) |
| RoBERTa*_T | 98.50 (+**4.82**) | 97.64(+3.12) | 98.54(+1.37) | 97.99(+5.02) |
| RoBERTa*_CNN_T | 98.55(+4.05) | 97.44(+2.74) | 98.65(+6.36) | 97.91(+4.54) |
| RoBERTa*_DPCNN_T | 97.99(+2.88) | 98.14(+3.44) | 98.70 (+5.14) | 98.37(+4.32) |
| RoBERTa*_RCNN_T | 98.23(+4.07) | 97.53(+3.30) | 98.35(+6.09) | 98.23(+5.09) |
| RoBERTa*_RNN_T | **98.78** (+3.54) | 98.37(+3.75) | 98.74(+4.40) | 98.55(+4.14) |

(RoBERTa*: 代表本文构建的黑灰产领域词向量模型; RoBERTa_代表利用 RoBERTa 模型进行向量表征后输入不同神经网络模型进行分类; _T 表示在模型分类前进行主题词增强, (+*)代表在相同分类模型下，领域词向量模型比 RoBERTa 词向量模型分类效果的提升量)

### 4.2.3　智能对话有效性实验

评价指标。在智能对话有效实验中，采用精确率(*Precision*)、召回率(*Recall*)、*F*1 值(*F*1-*Score*)作为评价指标，计算方法如公式(12)～式(14)所示。

实验步骤。实验中，采用 Rasa[34]开源软件编写智能对话系统。Rasa 是一个开源的对话系统框架，用于构建聊天机器人。采用对话数据集，训练意图和实体联合识模型训练，并用测试数据验证模型对黑灰产人员意图识别的有效性。

结果分析。通过对话数据集训练后的模型意图识别准确率达到 90.80%，实体识别准确率达到 92.70%。模型在测试样本数据上取得良好效果，从表 8～表 9 可以看出，定义的各类意图预测精度平均达到了 90%以上，模型能够对黑灰产人员的对话意图进行有效的识别，并且模型对文本内容中包含的实体识别准确率平均也达到 90%以上。

意图识别混淆矩阵如图 5 所示，实体识别混淆矩阵如图 6 所示。实验结果表明，智能对话模型能够

**表 8　意图识别结果**

**Table 8　Result of intention recognition**

| 意图 | 精确率(%) | 召回率(%) | *F*1 值(%) |
|---|---|---|---|
| answer_operator | 98.60 | 98.60 | 98.60 |
| answer_user | 97.20 | 100.00 | 98.60 |
| answer_storage | 97.70 | 98.40 | 98.00 |
| answer_price | 97.30 | 98.20 | 97.80 |
| answer_location | 98.80 | 96.30 | 97.50 |
| answer_time | 95.00 | 99.00 | 96.90 |
| answer_receivePlat | 96.90 | 93.90 | 95.40 |
| answer_type | 90.90 | 96.80 | 93.70 |
| answer_cardNumber | 97.20 | 89.70 | 93.30 |
| answer_platform | 87.50 | 100.00 | 93.30 |
| answer_use | 94.20 | 92.50 | 93.30 |
| answer_service | 97.60 | 87.20 | 92.10 |
| answer_measures | 83.00 | 97.80 | 89.80 |
| answer_source | 97.70 | 80.80 | 88.40 |
| answer_saleMode | 89.30 | 83.30 | 86.20 |
| answer_otherCard | 95.50 | 77.80 | 85.70 |
| ask_aim | 85.70 | 81.80 | 83.70 |
| answer_payWay | 71.40 | 88.20 | 78.90 |
| answer_contactWay | 86.70 | 72.20 | 78.80 |
| answer_sales | 61.90 | 100.00 | 76.50 |

表 9 实体识别结果

Table 9 Result of entity identification

| 实体 | 精确率(%) | 召回率(%) | *F*1 值(%) |
|---|---|---|---|
| quantity | 98.90 | 98.20 | 98.60% |
| platform | 94.90 | 99.50 | 97.10 |
| location | 99.20 | 94.70 | 96.90 |
| price | 99.20 | 93.20 | 96.10 |
| company | 98.60 | 92.80 | 95.60 |
| phonenumber | 98.70 | 92.60 | 95.50 |
| time | 97.60 | 93.10 | 95.30 |
| receivePlat | 97.20 | 90.80 | 93.90 |
| model | 96.60 | 87.50 | 91.80 |
| payWay | 97.30 | 85.70 | 91.10 |
| card | 99.10 | 82.70 | 90.20 |
| use | 97.70 | 78.30 | 86.90 |
| user | 97.60 | 61.50 | 75.50 |

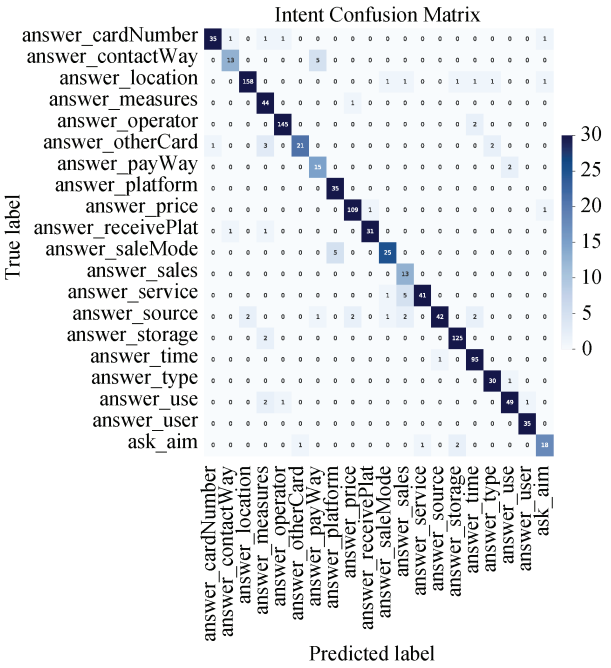


图 5 意图识别混淆矩阵

Figure 5 Intention recognition confusion matrix

有效理解黑灰产人员对话内容并提取威胁情报实体信息。

### 4.2.4 真实场景评估

为评估本文提出方法在真实场景下的效果，本文使用云服务器部署实验代码。服务器硬件配置为 1 核 CPU、4GB 内存、20GB 硬盘、10Mbps 网络带宽; 操作系统为 Ubuntu18.04。同时部署黑灰产目标识别模块和智能对话模块，并通过 Telethon 创建聊天客户端加入黑灰产群组。一共接入 174 个黑灰产相关的群组并对群聊内容进行持续监控分析，最终从 1036 个活跃群成员中识别出号商类黑灰产人员 48 个、卡商类黑灰产人员 65 个、料商类黑灰产人员 36 个、接码类黑灰产人员 35 个、赌博类黑灰产人员 32 个、色情类黑灰产人员 22 个。对系统识别出的每类黑灰产人员进行随机抽样 20 个 ID，根据其在群聊中发布的群聊内容进行人工判断，最终判定号商类 17 个账号、卡商类 18 个账号、料商类 16 个账号、接码类 16 个账号、赌博类 17 个账号、色情类 16 个账号实际在群组中从事相关黑灰产活动，如表 10 所示。

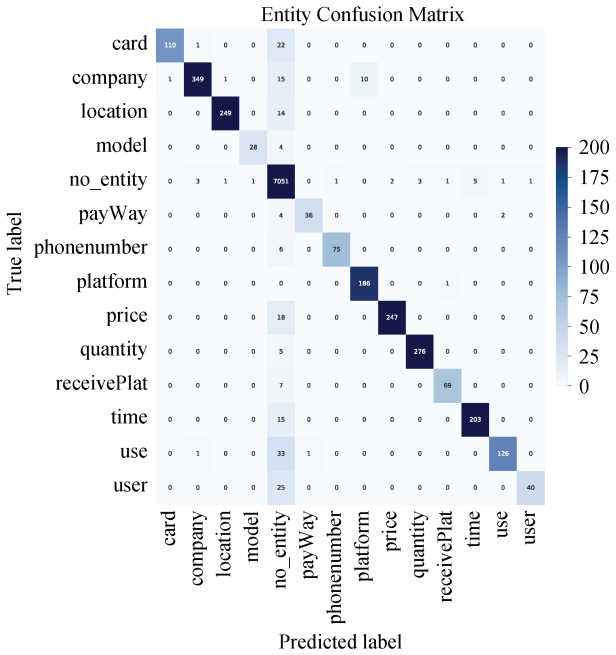


图 6 实体识别混淆矩阵

Figure 6 Entity recognition confusion matrix

表 10 黑灰产人员识别及抽样分析结果

Table 10 Identification and sampling analysis results

| 类别 | 系统识别数量 | 抽样 20 人工判定数量 |
|---|---|---|
| 号商 | 48 | 17 |
| 卡商 | 65 | 18 |
| 料商 | 36 | 16 |
| 接码 | 35 | 16 |
| 赌博 | 32 | 17 |
| 色情 | 22 | 16 |

选取 20 个卡商类别黑灰产人员账号进行对话测试。同时向 20 个卡商账号发送消息“你好，有卡卖吗？”以此来引起黑灰产人员注意。测试中有 18 个账号进行了回复，其中，对话机器人与 15 个账号进行了连续 10 轮以上完整对话，通过收集对话内容，获取了黑灰产人员的微信账号、产品代理方式，而这些信息都不是其主动群发披露的，难以通过开源采集获取。而有 2 个账号没有回复对话机器人发送的第一条消息，有可能是账号处于离线状态。

同时, 使用 APIfox[①]软件对机器人服务端进行了压力测试, 实验中将服务端设置为单线程模式, 同时可支持 100 个处理请求, 平均时延 4.3s。

#### 4.2.5 讨论

本文的研究表明, 通过构建基于深度学习的智能对话系统在黑灰产威胁情报挖掘方面具有较好的效果, 方法可以自动识别群聊中的黑灰产人员, 为挖掘深层次黑灰产威胁情报提供了支撑。然而, 本文的方法也存在局限性, 一是方法的准确性依赖于数据集, 由于缺少开源的黑灰产领域分类和对话数据集, 本文通过人工参与自建数据集用于模型训练, 而数据集的质量和大小影响方法最终效果; 二是目前的研究主要关注中文领域黑灰产群聊, 方法仅支持对中文语言, 但后续可通过扩展不同语言类型的数据集, 并进行模型训练, 满足多种语言; 三是方法无法保证挖掘情报的准确性, 智能对话是按照提问模板进行提问, 诱导黑灰产人员回答问题, 方法无法保证其按照问话回答, 也无法保证回答的情报的准确性, 后续可扩展情报挖掘需求, 增加更多的提问模板, 以适应更加复杂的对话场景。

## 5 结论

本文提出一种基于目标识别与主题引导对话的方法来采集黑灰产相关的威胁情报信息。该方法能够从社交媒体群聊中自动识别黑灰产人员目标, 并主动开启对话, 通过主题引导驱动, 在对话过程中进行意图和实体联合识别, 并采用多种对话策略实现了自主对话, 以引导黑灰产人员暴露更多信息。实验表明, 黑灰产人员目标识准确率达到 98.78%, 意图识别准确率达到 90.80%, 实体识别准确率达到 92.70%。下一步的研究方向是扩展黑灰产人员分类数据集, 优化对话模型, 采集更多真实对话数据进行训练, 使智能对话过程更类似人类互动, 降低暴露的风险, 以便获取更多的黑灰产威胁情报。

## 参考文献

[1] Thomas K, Li F, Zand A, et al. Data Breaches, Phishing, or Malware?: Understanding the Risks of Stolen Credentials[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 1421-1434.

[2] Wei C L, Guo C, Li X Z, et al. Research on Account Regulation under the Background of Network Black and Gray Production[J]. *Netinfo Security*, 2021, 21(S1): 17-20.
(位春亮, 郭聪, 李晓壮, 等. 网络黑灰产背景下的账号规制研究[J]. *信息网络安全*, 2021, 21(S1): 17-20.)

[3] Pastrana S, Thomas D R, Hutchings A, et al. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale[C]. *The 2018 World Wide Web Conference on World Wide Web - WWW '18*, 2018: 1845-1854.

[4] Zheng P P, Yuan S H, Wu X T, et al. Hidden Buyer Identification in Darknet Markets *via* Dirichlet Hawkes Process[C]. *2021 IEEE International Conference on Big Data*, 2021: 581-589.

[5] Bhalerao R, Aliapoulios M, Shumailov I, et al. Towards Automatic Discovery of Cybercrime Supply Chains[EB/OL]. 2018: 1812.00381. https://arxiv.org/abs/1812.00381v2.

[6] Zhao K Z, Zhang Y, Xing C X, et al. Chinese Underground Market Jargon Analysis Based on Unsupervised Learning[C]. *2016 IEEE Conference on Intelligence and Security Informatics*, 2016: 97-102.

[7] Basheer R, Alkhatib B. Threats from the Dark: A Review over Dark Web Investigation Research for Cyber Threat Intelligence[J]. *Journal of Computer Networks and Communications*, 2021, 2021(1): 1302999.

[8] Iqbal F, Fung B C M, Debbabi M, et al. Wordnet-Based Criminal Networks Mining for Cybercrime Investigation[J]. *IEEE Access*, 2019, 7: 22740-22755.

[9] Pastrana S, Hutchings A, Caines A, et al. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum[C]. *Research in Attacks, Intrusions, and Defenses*, 2018: 207-227.

[10] Lv B, Cheng P, Zhang C, et al. Research on Modeling of E-banking Fraud Account Identification Based on Federated Learning[C]. *2021 IEEE International Conference on Dependable, Autonomic and Secure Computing*, 2021: 611-618.

[11] Xu Z Q, Li L Y, Li H, et al. Self-Supervised Graph Representation Learning for Black Market Account Detection[C]. *The Sixteenth ACM International Conference on Web Search and Data Mining*, 2023: 330-338.

[12] Yu H S. Modality and Regulation of the Underground Industry Chain of Cybercrime[J]. *Journal of National Prosecutors College*, 2021, 29(1): 41-54.
(喻海松. 网络犯罪黑灰产业链的样态与规制[J]. *国家检察官学院学报*, 2021, 29(1): 41-54.)

[13] Ni J J, Young T, Pandelea V, et al. Recent Advances in Deep Learning Based Dialogue Systems: A Systematic Survey[J]. *Artificial Intelligence Review*, 2023, 56(4): 3055-3155.

[14] Yu S, Chen Y X, Zaidi H. AVA: A Financial Service Chatbot Based on Deep Bidirectional Transformers[J]. *Frontiers in Applied Mathematics and Statistics*, 2021, 7: 604842.

[15] Tian Y, Thung F, Sharma A, et al. APIBot: Question Answering Bot for API Documentation[C]. *2017 32nd IEEE/ACM International Conference on Automated Software Engineering*, 2017: 153-158.

[16] Abdellatif A, Badran K, Shihab E. MSRBot: Using Bots to Answer Questions from Software Repositories[J]. *Empirical Software Engineering*, 2020, 25(3): 1834-1863.

[17] McDermott C D, Jeannelle B, Isaacs J P. Towards a Conversational Agent for Threat Detection in the Internet of Things[C]. *2019*

① https://apifox.com/

*International Conference on Cyber Situational Awareness, Data Analytics and Assessment*, 2019: 1-8.

[18] Franco M F, Rodrigues B, Scheid E J, et al. SecBot: A Business-Driven Conversational Agent for Cybersecurity Planning and Management[C]. *2020 16th International Conference on Network and Service Management*, 2020: 1-7.

[19] Wang P, Liao X J, Qin Y, et al. Into the Deep Web: Understanding E-Commerce Fraud from Autonomous Chat with Cybercriminals[C]. *Proceedings 2020 Network and Distributed System Security Symposium*, 2020.

[20] Liu Y H, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[EB/OL]. 2019: 1907.11692. https://arxiv.org/abs/1907.11692v1.

[21] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]. *The 27th International Conference on Neural Information Processing Systems - Volume 2*, 2013: 3111-3119.

[22] Zheng L M, Jia K, Bi T S, et al. Cosine Similarity Based Line Protection for Large-Scale Wind Farms[J]. *IEEE Transactions on Industrial Electronics*, 2021, 68(7): 5990-5999.

[23] Bunk T, Varshneya D, Vlasov V, et al. DIET:Lightweight Language Understanding for Dialogue Systems[EB/OL]. arXiv Preprint arXiv: 2004.09936, 2020.

[24] Lafferty J D, McCallum A, Pereira F C N, et al. Conditional Random Fields[C]. *The Eighteenth International Conference on Machine Learning*, 2001: 282-289.

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. *Advances in Neural Information Processing Systems*, 2017, 30.

[26] Vlasov V, Mosig J E M, Nichol A. Dialogue Transformers[EB/OL]. 2019: 1910.00486. https://arxiv.org/abs/1910.00486v3.

[27] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[EB/OL]. 2013: 1301.3781. https://arxiv.org/abs/1301.3781v3.

[28] Su J L. 2020. WoBERT: Word-based Chinese BERT model ZhuiyiAI[EB/OL]. https://github.com/ZhuiyiTechnology/WoBERT.

[29] Chen Y. Convolutional Neural Network for Sentence Classification[J]. *Proc. EMNLP*, 2014, 1746-1751.

[30] Liu P F, Qiu X P, Huang X J, et al. Recurrent Neural Network for Text Classification with Multi-Task Learning[EB/OL]. 2016: 1605.05101. https://arxiv.org/abs/1605.05101v1.

[31] Lai S W, Xu L H, Liu K, et al. Recurrent Convolutional Neural Networks for Text Classification[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, 29(1).

[32] Johnson R, Zhang T. Deep Pyramid Convolutional Neural Networks for Text Categorization[C]. *The 55th Annual Meeting of the Association for Computational Linguistics*, 2017: 562-570.

[33] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. 2018: 1810.04805. https://arxiv.org/abs/1810.04805v2.

[34] Bocklisch T, Faulkner J, Pawlowski N, et al. Rasa: Open Source Language Understanding and Dialogue Management[EB/OL]. arXiv Preprint arXiv: 1712.05181, 2017.

**罗双春** 于 2012 年在国防科技大学网络工程专业获得学士学位。现在四川大学网络空间安全专业攻读硕士学位。研究领域为网络空间安全。研究兴趣包括: Web 安全、威胁检测、机器学习和自然语言处理等。Email: luoshuangchun@scu.edu.cn

**孙恩博** 于 2016 年在电子科技大学计算机应用技术专业获得硕士学位。现就职于中国电子科技集团公司第三十研究所, 工程师。研究兴趣包括: 网络空间安全、黑产治理、匿名网络协议分析等。Email: suneb429@163.com

**黄诚** 于 2017 年在四川大学信息系统 z 安全专业获得博士学位。现任四川大学副教授。研究领域为网络空间安全。研究兴趣包括: 攻击检测、威胁溯源、数据挖掘、社交网络、机器学习和自然语言处理等。Email: codesec@scu.edu.cn