Uma análise dos dados provenientes de vazamentos disponíveis para venda em marketplaces da dark web

Kelvin Lopes[1], **Luciano Ignaczak**[1]

[1] Universidade do Vale do Rio do Sinos (UNISINOS) – São Leopoldo, RS – Brasil

kelvingl@edu.unisinos.br, lignaczak@unisinos.br

Abstract. *The use of personal data on websites has been increasing gradually in recent years, and cybersecurity incidents resulting in data breaches have become frequent. As a consequence, cybercriminals make data available in dark web marketplaces aiming for profits. This study analyzed announcements publicized in four marketplaces to present the current scenario of leaked data available for sale. Aiming to perform the analysis, we applied information extraction and implemented Named Entity Recognition (NER) to identify organizations in announcements. We evaluated two approaches: the XLM-RoBERTa and the NLTK. Based on the analysis, we identified that "Interactive Media & Services" is the industry sector with the highest number of data for sale.*

Resumo. *O uso de dados pessoais em sites tem aumentado gradualmente nos últimos anos, ampliando o número de incidentes de segurança cibernética que resultam em violações de dados. Uma consequência das violações é a disponibilização de dados vazados em marketplaces da dark web. Este estudo analisou anúncios divulgados em quatro marketplaces para apresentar o cenário atual de dados disponíveis para venda. A análise fez uso de extração de informações a partir da implementação de Named Entity Recognition (NER) para identificar as organizações em anúncios. Duas abordagens foram avaliadas: o XLM-RoBERTa e o NLTK. A análise identificou que "Interactive Media & Services" é o setor da indústria com o maior número de dados para venda.*

1. Introdução

Com o decorrer do tempo tornou-se natural para maior parte dos internautas informar seus dados aos sites que eles costumam usar [Fang et al. 2019]. A medida que as empresas da internet solicitam informações dos seus clientes para provê-los uma melhor experiência, são gerados grandes volumes de dados. Quando os devidos cuidados não são tomados, criminosos cibernéticos podem aproveitar-se de vulnerabilidades apresentadas pelos sites e roubar essas informações. O termo violação de dados, originado do inglês *data breach*, é utilizado globalmente para denominar essa ação. Nos últimos anos, em virtude da coleta massiva de dados por empresas, as violações de dados vem impactando dos usuários na internet. Por isso, as empresas que fazem esse tipo de coleta devem redobrar o cuidado com os dados de seus clientes. O custo médio global estimado para cada incidente resultante da violação de dados em empresas é de cerca de 3,86 milhões de dólares [IBM 2019].

Após a obtenção da base de dados roubada, os criminosos cibernéticos buscam a obtenção de lucros através da comercialização da mesma. Visto que esse tipo de ação é ilegal, os criminosos precisam garantir anonimato e ter a segurança para anunciar seu

produto (base de dados) em um meio que não seja monitorado por indexadores de busca. A maneira que os criminosos encontram pra fazer esse tipo de venda é chamada *dark web*, que, segundo [Fu et al. 2010], provê mecanismos de segurança de rede e anonimato. Buscando estas mesmas garantias, os criminosos realizam suas transações usando criptomoedas. Diferentemente das transações financeiras comuns, que podem ser rastreadas com facilidade, as criptomoedas dificultam significativamente o rastreio.

O crescimento desse mercado ilegal evidenciou a necessidade de investimento no monitoramento desses meios de comercialização na *dark web*, para auxiliar as organizações na identificação e mitigação de vazamentos de seus dados. Na *dark web*, há várias soluções corporativas para monitorar os meios de comercialização, porém ainda em menor quantidade quando comparado a web convencional. Devido a existência de vários fóruns e *marketplaces* na *dark web*, a técnica mais apropriada para o monitoramento é a extração de palavras ou expressões que identifiquem a organização a qual o vazamento pertence, o que consiste basicamente em aplicação de mineração de texto. Este tipo de análise é realizada através de tarefas de mineração de texto, as quais compõe um conjunto de ferramentas e técnicas que tem como objetivo extrair informações relevantes de um dado texto, e disponibilizá-las de uma forma compreensível ao usuário [Feldman et al. 2007]

O objetivo desse artigo é analisar os dados provenientes de vazamentos que estão disponíveis para venda na *dark web*, identificando quais tipos de dados são mais vendidos, qual setor da indústria é mais afetado, e se algum tipo de dado gera mais interesse aos possíveis compradores. Para atingir esse objetivo será utilizada mineração de texto. Para obtenção dos dados que serão analisados através de mineração de texto, foi criado um *web crawler* para coleta de dados das páginas dos *marketplaces* na *dark web*. Após a obtenção dos dados, serão aplicados algoritmos de aprendizado de máquina e técnicas de mineração de texto, extraindo informações detalhadas da página.

Este artigo visa contribuir com a comunidade de segurança da informação provendo um panorama dos dados provenientes de vazamentos comercializados em quatro *marketplaces* escolhidos da *dark web*, os quais são apresentados de forma sistematizada pelos autores. Outra contribuição é a avaliação de dois modelos de reconhecimento de entidades, baseados em aprendizado de máquina, para identificação de entidades nos anúncios de *marketplaces*.

As demais seções deste artigo estão estruturadas da seguinte forma. A Seção 2 apresenta os trabalhos relacionados ao estudo, que abordam assuntos no entorno ao tema, em formatos diferentes. A Seção 3 apresenta a metodologia aplicada no estudo, e como foram obtidos os resultados. Na Seção 4, são apresentados os resultados e abertas discussões. Por fim, na Seção 5 são apresentadas as considerações finais sobre o trabalho, além de sugestões de trabalhos futuros.

2. Trabalhos relacionados

Esta seção apresenta trabalhos considerados relevantes para o contexto desse artigo. A pesquisa por artigos relacionados foi realizada utilizando o mecanismo de indexação Google Schoolar. Nele, foi utilizada a *string* de busca " *'dark web' AND 'threat intelligence' AND 'text mining'* " e aplicado um filtro para exibir apenas artigos dos últimos 5 anos. Com base no resultado obtido, os autores consideraram apenas artigos científicos publica-

dos nas seguintes bases: ACM Digital Library, IEEE Xplore, Springer, Elsevier e Taylor & Francis. O resultado da busca considerando apenas essas bases contabilizou 18 estudos. Os autores realizaram o *download* dos estudos e, baseados na leitura dos *abstracts*, selecionaram os oito trabalhos que possuem maior relação com o presente estudo. Todos os trabalhos selecionados estão associados ao uso de mineração de texto para predição e identificação de ameaças cibernéticas em fóruns e *marketplaces*.

Dentre os trabalhos relacionados destacam-se os estudos que propuseram um *framework* de identificação proativa de ameaças cibernéticas. [Sapienza et al. 2018] introduz o *framework* DISCOVER, capaz de monitorar blogs relacionados com a área de segurança e perfis selecionados do Twitter de especialistas na área. O *framework* utiliza mineração de texto nos dados coletados para remover *stopwords*, ameaças já conhecidas e vocabulário técnico. Com os dados já filtrados, o *framework* implementa um mecanismo de geração de alertas para possíveis novas ameaças, de acordo com a frequência de ocorrência dos termos. Em sua avaliação, o *framework* obteve uma precisão de 0,84 para contas do Twitter e 0,59 para os blogs. Já o estudo de [Samtani et al. 2020] propõe o *Diachronic Graph Embedding Framework* (D-GEF), que utiliza técnicas de processamento de linguagem natual como *Graph of Words* (GoW), *graph embedding* não-supervisionado e *Diachronic Word Embeddings*. Os dados analisados pelo D-GEF foram obtidos através do desenvolvimento de um *crawler* para *darkweb*. A avaliação apresentou que o D-GEF obteve performance superior aos outros estudos nas modalidades de identificação de palavras predominantes e, também, na modalidade de *graph embedding*.

Com objetivo semelhante ao estudo de [Sapienza et al. 2018], [Dong et al. 2018] visa classificar itens encontrados nos *marketplaces* da *darkweb* e extrair nomes de possíveis novas ameaças. Para isso, os autores desenvolveram um *crawler* que monitora oito *marketplaces* da *darkweb*, classificando os produtos anunciados por categoria: *data*, *carding*, *hack* e *others*. O estudo utilizou uma rede neural *Multi-layer Perceptron* para realizar a classificação. O experimento obteve precisão de 0,94.

Outra abordagem encontrada nos estudos foi a tentativa de reconhecimento do perfil e comportamento de autores e vendedores nos fóruns e *marketplaces*. Com o objetivo de identificar perfis e autores chaves, [Huang and Chen 2016] utilizaram como fonte de dados o fórum Baidu. O estudo utilizou modelagem de tópicos, o modelo não supervisionado *Self-Organizing Map* (SOM) e, por fim, um conjunto de três técnicas para classificação, *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e *deep learning*. Os resultados mostraram que a técnica de *deep learning*, empregado sem o KNN e o SVM, tem acurácia de 0,89; enquanto que, quando utilizadas em conjunto, mostram acurácia de 0,99. Em outro estudo, [Huang and Ban 2019] avaliam o reconhecimento de perfis e o comportamento de vendedores em fóruns e *marketplaces* utilizando a mesma base de dados do estudo de [Huang and Chen 2016], porém empregando o modelo de modelagem de tópicos *Growing Hierarchical Self-Organizing Map* (GHSOM). Os resultados do estudo mostram que a predição dos tópicos dos fóruns obteve precisão de 0,83. Em outro estudo que visa identificar o papel de cada perfil e sua influência, [Park et al. 2018] coletaram todos os *posts* de três fóruns e realizaram análise textual, agrupando perfis relacionados ao mesmo usuário e classificando suas publicações com base em *keywords*. Diferente dos estudos anteriores, o estudo de [Park et al. 2018] apresentou resultados não sumarizados, com precisão para cada perfil analisado, o que dificulta o apontamento de

uma precisão geral do estudo.

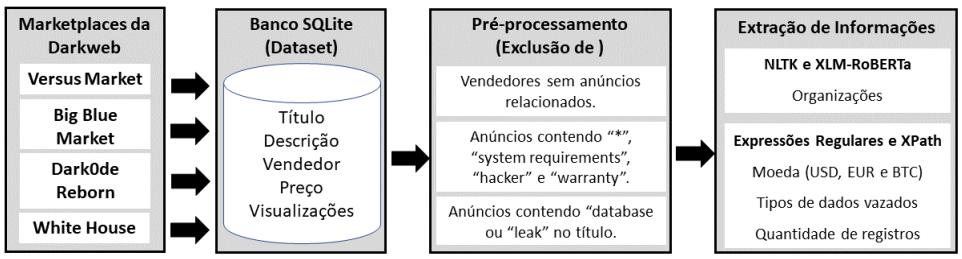
Além do estudo de [Huang and Ban 2019], [Al-Ramahi et al. 2020] também utilizaram modelagem de tópicos para extrair os pontos de interesse de um *dataset* público do fórum CrackingFire, usando *Latent Dirichlet Allocation* (LDA). O trabalho também realizou um segundo experimento, no qual os autores desenvolveram um *crawler* para *darkweb* e extraiu informações do fórum chamado Torum, que já possui categorizações de *posts*, para comparação com os resultados do *dataset* do CrackingFire. O estudo não divulgou a acurácia da extração. Outro estudo que também utilizou modelagem de tópicos, [Mendsaikhan et al. 2019] visa ajudar pessoas que estão analisando as ameaças, através da extração de informações de *data sources* como Reddit e *National Vulnerability Database* (NVD). Para filtragem, foi usado o modelo *doc2vec*, treinado com sete conjuntos de parâmetros diferentes, e a melhor acurácia obtida foi 0,83.

Da mesma forma que os trabalhos de [Samtani et al. 2020], [Dong et al. 2018], e [Al-Ramahi et al. 2020], este estudo implementou um *crawler* para gerar um *dataset* com base em informações de *marketplaces* da *darkweb*. Da mesma forma que o trabalho de [Park et al. 2018], este estudo usou extração de informações para mapear os dados dos *marketplaces*, enquanto que a maioria dos trabalhos citados nesta seção usaram modelagem de tópicos. Diferentemente dos estudos apresentados, este trabalho propõe uma análise dos tipos de dados disponíveis para vendas em *marketplaces* da *darkweb*. Apesar de outros estudos explorarem a *darkweb*, o presente trabalho se destaca pelo foco nas características associadas aos anúncios oferecendo dados de usuários e nos vendedores que atuam neste tipo de venda. Por fim, este trabalho avaliou a performance de redes neurais baseadas em atenção para o reconhecimento das organizações associadas aos anúncios, através do uso do modelo BERT, proposto por [Devlin et al. 2018],

3. Método de Trabalho

Este trabalho implementou a abordagem tradicional para mineração de texto para extrair e analisar dados dos *marketplaces*. Inicialmente foi feita a coleta dos dados e utilizadas técnicas de pré-processamento. Posteriormente foram usadas diversas técnicas para extração das informações com o objetivo de estruturar os dados para análise. As etapas do método são apresentadas na Figura 1 e detalhadas em subseções específicas. O método de trabalho foi estruturado para obter as seguintes informações: (i) os tipos de dados mais encontrados em vazamentos; (ii) os setores da indústria mais afetados; (iii) a variação de preço dos anúncios; (iv) os dados que geram maior interesse pelos possíveis compradores.

Figura 1. Etapas do processo de mineração de texto.



3.1. Coleta de dados

A etapa da coleta de dados visa gerar um *dataset* com as informações das páginas nos *marketplaces* que estão associadas com o trabalho. Para isso foi desenvolvido um *crawler* na linguagem de programação Python[1], extendendo um *framework* chamado Scrapy[2] e especializando implementações para cada *marketplace*. O código-fonte está disponível no GitHub do autor[3]. O *crawler* foi executado no dia 3 de fevereiro de 2021, e os *marketplaces* escolhidos foram Versus Market, Big Blue Market, Dark0de Reborn e White House Market. A escolha desses *marketplaces* foi resultado do acompanhamento na mesma data do *ranking* dos *marketplaces* mais acessados da *darkweb*, segundo o *site* onion.live.

Devido a complexidade de login automatizado pela presença de mecanismos de desafios para humanos (CAPTCHAS) nos *marketplaces* escolhidos, foi necessário o login manual e a configuração dos *cookies* de sessão no crawler para cada *marketplace*. Com a sessão configurada, o *crawler* realiza uma busca no mecanismo do próprio *marketplace* utilizando a *string* "*database*". Após isso, o *crawler* varre todas as páginas de resultado da pesquisa, acessando cada um dos anúncios e extraindo o título, descrição, vendedor, preço e a quantidade de visualizações para cada anúncio. Por fim, foi necessária a criação de um banco de dados local com SQLite a fim de armazenar os dados extraídos, e todos os resultados foram gravados em uma tabela.

3.2. Pré-processamento

A primeira atividade de pré-processamento realizada foi a limpeza dos dados, que envolveu a remoção dos anúncios obtidos dos *marketplaces* que não tem necessariamente vínculo com as análises realizadas neste estudo. Esta atividade removeu todos anúncios que não oferecem bases de dados vazadas. Para isso, os autores criaram um *script* em *Python* para realizar uma filtragem simples de anúncios no banco de dados. A atividade foi dividida em três estágios.

No primeiro estágio foram removidos alguns vendedores, pois estes só haviam publicado anúncios que não estão relacionados com o tópico deste estudo. No segundo estágio, ainda em razão do alto número de anúncios que não representavam dados vazados, foram removidos todos os anúncios que continham alguma das seguintes *strings* no título ou descrição: "$\star$" , "*system requirements*", "*hacker*" e "*warranty*". Por fim, no terceiro estágio foram removidos todos os anúncios que não possuíam no título as *strings* "*database*" ou "*leak*". Os autores avaliaram o uso somente do terceiro estágio, porém não foi possível excluir todos os anúncios almejados, em virtude de alguns dos mesmos fazerem uso das *strings* mencionadas. Um exemplo de anúncio que não seria removido caso o segundo estágio não fosse executado é "Text Utils - hack and get any web database!"

A segunda atividade de pré-processamento realizada viabiliza a utilização dos algoritmos de NER do Natural Language Toolkit (NLTK)[4], conforme seu livro de documentação[5]. Para isso, foram aplicadas duas técnicas de pré-processamento do *toolkit*. A primeira técnica utilizada foi a tokenização, que recebe uma lista de sentenças

[1] www.python.org
[2] scrapy.org
[3] github.com/kelvingl/onion-juicer
[4] www.nltk.org
[5] www.nltk.org/book/ch07.html

e as separa em *tokens*, representando cada uma das palavras. A segunda técnica de pré-processamento aplicada foi a PoS *Part-of-Speech tagging*, a qual recebe a lista de *tokens* gerados pela técnica aplicada anteriormente e os rotula com a respectiva classe gramatical.

3.3. Extração de informações

Os autores utilizaram duas abordagens independentes para manipular e extrair informações dos anúncios. A primeira abordagem consiste na implementação de modelos para extração de informações para reconhecer as organizações responsáveis pelos dados vazados nos anúncios. Este trabalho considera como organização sites ou empresas. Além disso, o *script* implementado nessa abordagem também busca informações sobre o ramo de atuação das organizações reconhecidas. A implementação foi executada com aprendizado de máquina, utilizando como entrada o título e a descrição do anúncio. Para outras informações, como preço, tipos de dados e número de registros no vazamento, foi utilizada a segunda abordagem com expressões regulares e seletores XPath, que tem como objetivo a descoberta de informações contidas em páginas HTML.

3.3.1. Reconhecimento de organizações

A primeira abordagem implementa o processo de Named Entity Recognition and Classification (NERC), proposto por [Nadeau and Sekine 2007]. Para decidir sobre o melhor modelo para a extração da entidade no cenário avaliado neste trabalho, os autores realizaram um experimento para avaliar dois modelos de aprendizado de máquina.

O primeiro modelo a ser utilizado é o modelo padrão do *framework* NLTK para *Python*, desenvolvido na Universidade Stanford ([Bird et al. 2009]). A escolha do NLTK foi baseada por sua utilização no trabalho relacionado ([Samtani et al. 2020]). O NLTK já é um *framework* estabelecido para trabalhar com linguagem natural. O classificador é baseado no conceito estatístico de Máxima Entropia e treinado com o *dataset* "*ACE - Multilingual Training Corpus*". A implementação do *framework* NLTK foi baseada na utilização de três funções. A primeira função aplicada é um *tokenizer* e a segunda é para *PoS tagging*. Por fim, a terceira função utilizada tem como entrada o resultado do *PoS tagging* e constrói uma estrutura de árvore com as classes identificadas. Com base nos resultados foram selecionadas apenas as sequências que apresentavam a classe "organization", e "person".

O segundo modelo utilizado é o XLM-RoBERTa, especializado em NER. O uso do modelo XLM-RoBERTa busca avaliar um mecanismo novo de *self-attention*, o qual não foi implementado em artigos incluídos na Seção 2. Este trabalho implementou o modelo XLM-RoBERTa, que é público e foi está disponível no site huggingface.co[6]. Esse modelo funciona com a biblioteca de aprendizado de máquina Pytorch[7], disponível para *Python*. O título foi utilizado como entrada da função *tokenizer*, e o resultado da função foi usado como entrada do modelo. A implementação classificou cada *token* com a classe do dado inferido. Com os *tokens* classificados, os autores selecionaram apenas resultados que apresentaram a classe "organization".

[6]huggingface.co/xlm-roberta-large-finetuned-conll03-english
[7]pytorch.org

O experimento para o processo de NERC avaliou o desempenho individual de cada modelo, e se a união deles resulta em uma maior precisão na identificação das entidades. Os autores avaliaram se o experimento obtém o mesmo resultado que o trabalho de [Jiang et al. 2016], que aumenta a acurácia quando os modelos são unidos. Nesse experimento, além da acurácia, foram utilizadas as seguintes métricas de performance ”precisão”, ”recall” e ”F1” para avaliação. A definição foi baseada nas métricas utilizadas em outro trabalho baseado em NERC [Santos et al. 2019].

Para avaliação dos resultados das implementações do NLTK e do XLM-RoBERTa foi necessário o uso de uma biblioteca para consulta em site externo, que forneça informação sobre a existência da organização. A biblioteca escolhida foi a da Wikipédia para Python, que permite a automatização de buscas e acesso a páginas. Os autores verificaram que algumas entidades reconhecidas terminavam com a *string* ”com” e ”org”, e não possuíam o ponto, tornando-a um domínio inválido. Um exemplo é a entidade reconhecida como ”Mopcom”. Para tratar destes caso foi implementado um *script* em Python para adicionar um ponto antes da *string* com o TLD, tornando o nome da entidade um domínio válido.

Cada organização reconhecida em ambos os métodos de reconhecimento de entidades (NLTK e XLM-RoBERTa), foi usada como entrada em um *script* desenvolvido pelos autores. Nesse *script*, as entradas eram passadas para a função de busca da biblioteca da Wikipédia. Com o resultado dessa função de busca, foram filtradas apenas as organizações que possuíam uma página cujo nome combinasse exatamente com o da entidade. Depois de filtrados, foi realizada a extração das informações da barra lateral da página, que possui o tipo de organização, bem como o tipo de serviço prestado.

3.3.2. Seletores XPath e expressões regulares

Esta abordagem tem como objetivo obter as informações que não são referentes à entidade envolvida no vazamento. As informações como título, descrição, vendedor, quantidade de visualizações, foram capturadas através de *scripts* com capacidade de extração de dados diretamente da estrutura HTML. Para isto, os *scripts* foram implementados para identificar as *tags* com os dados necessários para este trabalho. Para extração do preço, o *script*, além de identificar a *tag* HTML, necessitou utilizar uma expressão regular para considerar o tipo da moeda, como USD, EUR e BTC.

Após a extração desses dados, foi necessária a implementação de três *scripts* em *Python* para extrair as informações que estão implícitas na descrição e/ou título dos anúncios, além da conversão dos valores obtidos dos mesmos. O primeiro *script* converte todos os valores monetários obtidos para dólar americano, utilizando a cotação do dia da extração, obtida pelo Google. Apenas um dos *marketplaces* tinha anúncios em que a moeda não era o dólar, e a alteração foi feita em 35 registros.

O segundo *script* identifica os tipos de dados vazados através de expressões regulares aplicadas na descrição do anúncio. Para isto foram analisadas descrições de anúncios aleatórios, e com base nos tipos de dados que viu, criou expressões regulares para identificar os padrões observados. Os autores repetiram o processo até que atingisse a marca de 30 tipos de dados com suas respectivas expressões regulares.

O terceiro *script* visa extrair a quantidade de registros no vazamento, com base no título e na descrição dos anúncios. Assim como no segundo *script*, os autores analisaram os padrões encontrados nos anúncios, como "4 Million", ou "24M", e criou expressões regulares para vários casos, enquanto ainda houvessem casos não identificados. O *script* não conseguiu identificar a quantidade de registros de alguns anúncios, os quais se tratavam de várias bases agregadas, ou em outros casos, que não possuíam a informação sobre a quantidade de registros na descrição e tampouco no título. Estes anúncios não foram considerados nos resultados que envolvem a quantidade de registros.

3.4. Organização dos dados

Para a análise dos vazamentos de sites e empresas foi necessário obter a informação sobre o setor da industria a qual ele pertencia. Para isso foi utilizado o padrão de classificação *The Global Industry Classification Standard* (GICS), organizado pelas empresas MSCI e S&P Global, conforme [Hrazdil and Zhang 2012]. Os autores optaram pela utilização desse modelo, e não de outro, como o da Organização Internacional do Trabalho, porque o GICS tem maior granularidade nos ramos de atividades econômicas associadas com negócios digitais. Conforme já mencionado, os autores implementaram um *script* para encontrar automaticamente, com base na Wikipédia, o ramo de atuação associado a cada uma das organizações envolvidas no vazamento. Para as organizações que o *script* retornou o ramo de atuação, os autores inferiram um setor da indústria equivalente do GICS.

No entanto, como houve um número pequeno de combinações exatas de páginas, foi necessária a classificação manual das entidades que não foram localizadas. Para isso, os autores pesquisaram cada uma das entidades não classificadas e definiram seu setor da industria conforme o GICS. Os autores analisaram os sites das empresas quando eles estavam no ar e buscou informações sobre o ramo da empresa na seção "sobre" do site. A partir disso, os autores localizaram no GICS a classificação equivalente. Quando o site das empresas não existiam mais, os autores fizeram uso de pesquisa histórica, tanto no Google[8], quanto no *Wayback Machine*[9] e no Reddit[10]. Em casos onde o site da empresa não apontava exatamente para um setor da industria no GICS, os autores inferiram com base nas descrições do site e nas descrições contidas no padrão de classificação.

Mesmo com a generalização, não foi possível a classificação de todas as entidades. Alguns sites não foram encontrados nos métodos mencionados, ou não se encaixavam em nenhuma das categorias propostas no GICS, como ONGs. Por esse motivo, 23 anúncios não foram considerados nos resultados.

4. Resultados e discussão

Nesta seção são apresentados os resultados do estudo. O *dataset* gerado possui um total de 1.424 anúncios resultantes dos quatro *marketplaces* mencionados na Seção 3. Os autores necessitaram realizar uma limpeza no *dataset* e removeu 29 anúncios que não estavam ofertando dados resultantes de vazamentos. Considerando os 1.395 anúncios restantes, tampouco puderam ser considerados aqueles associados a pacotes com vazamentos de dados de vários sites, ou anúncios nos quais não foram possíveis obter o setor da indústria.

[8] google.com
[9] archive.org/web/
[10] reddit.com

Com estes dois critérios, outros 78 anúncios foram excluídos. O *dataset* final analisado possui 1.317 anúncios. Além disso, no restante desta discussão os autores interpretaram a quantidade de visualizações dos anúncios como interesse no anúncio visualizado.

Diante do desafio de reconhecer as entidades contidas no título dos anúncios, os autores aplicaram, de forma experimental, dois modelos de NERC baseados em aprendizado de máquina e avaliou suas respectivas performances com uma automação que realiza uma busca no Wikipédia. A avaliação dos modelos é apresentada na Tabela 1.

Tabela 1. Avaliação dos modelos para NERC.

Método	**Anúncios (n=1.317)**				
	Acurácia	**Precisão**	**Recall**	**F1**	**Entidades identificadas pela busca na Wikipédia**
NLTK	0,39	0,87	0,41	0,56	0,20
XLM-RoBERTa	0,60	0,92	0,63	0,75	0,17
NLTK + XLM-RoBERTa	0,87	0,98	0,88	0,93	0,25

Na Tabela 1 é possível destacar a métrica de acurácia dos modelos. O NLTK apresentou baixa acurácia (0,39), enquanto que o XLM-RoBERTa foi melhor (0,60). O valor da acurácia foi impactado porque em muitos anúncios não foram encontradas as entidades. No NLTK, não foram identificados domínios como entidades, o que resultou em 724 falsos negativos, dos 1317 no total. O XLM-RoBERTa reconheceu domínios, o que fez com que sua acurácia fosse melhor, mas ainda ocorreram 408 falsos negativos. O uso conjunto dos modelos causou um impacto positivo nas métricas, resultando em uma acurácia de 0,87. A aplicação conjunta dos modelos resultou em apenas 139 falsos negativos.

Os autores consideram que a acurácia de 0,60 para o uso isolado do XLM-RoBERTa é baixa e acredita que um possível motivo é os dados com quais o modelo foi treinado. O modelo utilizado foi treinado com o *dataset* de noticias da Reuters, do ano de 2003. Uma possível melhoria seria treinar o XLM-RoBERTa com dados mais recentes, e também expor o modelo a outros tipos de fontes, e não somente notícias. Outra dificuldade percebida envolveu a pesquisa na Wikipédia, a qual resultou na descoberta de poucas organizações. Os autores acreditam que isto foi causado porque parte dos anúncios possuíam somente a informação do domínio do site vazado, e não é comum o registro de páginas na Wikipédia cujo nome seja um domínio.

Outra análise realizada envolve os vendedores e informações referentes aos anúncios. Na Tabela 2 é apresentada a distribuição dos anúncios de cada vendedor, por *marketplace* e setor da indústria. Além destas informações, também é apresentada a soma de visualizações relacionadas aos anúncios de cada vendedor. Vale destacar que alguns dos *marketplaces* não publicam a quantidade de visualizações de seus anúncios, o que gerou os valores em branco na tabela.

Na Tabela 2 é possível verificar que o vendedor mais ativo na *dark web* relacionado com a venda de dados vazados é o ”drunkdragon”, que possui 746 anúncios em dois *marketplaces*. Este vendedor, assim como outros, anuncia vazamentos de diversos setores. Através da análise não foi possível constatar a existência de vendedor especializado em algum setor da indústria. Com base na tabela também é possível afirmar que poucos

Tabela 2. Lista de vendedores associados aos anúncios de vazamento de dados.

Vendedor	Anúncios (n=1.317)			
	Anúncios	**Marketplaces**	**Setores**	**Visualizações**
drunkdragon	746	2	26	77.894
empireshop	188	2	22	31.163
emperorman	151	1	17	-
goldapple	107	4	11	7.889
topvendor	82	1	22	11.289
eternos	22	1	5	907
mountaindew	17	1	10	3.091
ccity	2	1	1	3
loyaldemon	1	1	1	-
wick7	1	1	1	226

vendedores publicam anúncios de vazamentos em mais de um *marketplace*, com o mesmo nome de usuário. Os autores acreditam que a dificuldade de publicar em vários *marketplaces* está associada a volatilidade deles. Por exemplo, durante o desenvolvimento deste estudo, os quatro *marketplaces* analisados foram desativados por autoridades policiais, enquanto que novos *marketplaces* surgiram.

Através da análise da média de visualizações para cada anúncio é possível concluir que nenhum vendedor se destaca em relação aos demais. Alguns possuem mais visualizações pois o anúncio foi publicado há mais tempo, como é o caso do ”wick7”. Já outros possuem poucas visualizações em virtude de anúncios mais recentes, como o caso do ”ccity”. Com exceção destes casos, todos os vendedores tiveram médias parecidas de visualizações por anúncio.

A partir da análise dos dados foi possível apresentar o percentual de vazamentos para cada um dos setores da industria, conforme a Tabela 3. A tabela apresenta apenas os dez setores com o maior número de anúncios. Destaca-se nesse cenário o setor da indústria descrito como ”*Interactive Media & Services*”. Este setor possui um grande número de vazamentos, pois ele contempla fóruns e redes sociais. Considerando os 1.317 anúncios presentes no *dataset*, fóruns estão associados a 307, enquanto que redes sociais estão relacionadas com 126. É possível afirmar que fóruns e redes sociais juntos representam 32,87% dos anúncios de vazamento, o que justifica o posicionamento desse setor com maior número de vazamentos.

Um dos motivos considerados pelos autores para este grande número de vazamentos na categoria ”*Interactive Media & Services*”, é o uso dos sistemas mais comuns de fóruns: o vBulletin e o phpBB. Foi realizada uma pesquisa por vulnerabilidades desses dois sistemas na *Common Vulnerabilities and Exposures* (CVE), e foi possível encontrar seis com score igual ou maior que 7,5, que não necessitam de autenticação e viabilizam vazamentos. A partir de uma consulta no site *expoit-db.com*, é possível verificar que quatro vulnerabilidades já possuem *exploits* desenvolvidos e publicados, o que pode facilitar a exploração delas.

O estudo também analisou a média de registros por anúncio, apresentada na Tabela 3. Com base nos dados foi possível identificar que o vazamento envolvendo o maior

Tabela 3. Setores da indústria com mais vazamentos.

Setor da indústria	Anúncios (n=1.317)			
	Qtde. anúncios	%	Qtde. visualizações	Média de registros por anúncio*
Interactive Media & Services	564	42,82	56.519	15,41
Entertainment	258	19,59	25.131	8,89
IT Services	93	7,06	10.304	15,42
Hotels, Restaurants & Leisure	55	4,18	5.693	0,88
Software	45	3,42	4.414	41,32
Capital Markets	37	2,81	3.703	1,31
Professional Services	36	2,73	3.592	11,99
Diversified Consumer Services	31	2,35	2.938	8,73
Diversified Financial Services	26	1,97	2.692	2,97
Food & Staples Retailing	24	1,82	2.541	8,24

* Em milhões

número de registros no *dataset* está relacionado com a rede social *MySpace*, totalizando 360 milhões. A pesquisa também permite destacar que o setor da indústria com a maior média de registros por anúncio foi o de "*Software*" com um número superior a 41 milhões de registros médios por anúncio. Este setor da indústria contempla anúncios relacionados com dados da *Adobe*, do *iMesh* e do *mSpy*.

Outro campo analisado foi o valor dos anúncios. Na Tabela 4 são apresentados os valores sumarizados por setor da indústria, em dólar americano. É preciso destacar que entre a mediana (50%) e o quartil superior (75%) não há nenhuma variação relevante. As variações de valores começam por volta do percentil 85 em alguns setores, como é apresentado na tabela. Avaliando o percentil 85, é possível destacar o setor da indústria "Diversified Financial Services", que sofreu a maior variação de valor, apesar de seu valor máximo não superior ao dos demais setores. Também é apresentado o custo por milhão de registros, que é afetado diretamente pela quantidade de anúncios envolvendo o setor da indústria, apresentado na Tabela 3.

A Tabela 4 apresenta que o setor "*Hotels, Restaurants & Leisure*" possui o custo por registro mais caro dentre todos os outros setores. Além disso, seus anúncios possuem uma média de 258 mil registros, que é menor do que a média dos outros setores. A tabela também permite destacar o "*Interactive Media & Services*", pois ele apresenta muitos registros com baixo valor, tornando-se um bom custo-benefício para criminosos. Outra característica encontrada é que a maioria dos anúncios (863) custa entre 10 e 11 dólares.

Outra análise realizada contempla os tipos de dados ofertados em cada anúncio. Um resumo contendo o número de anúncios associados com os 10 tipos de dados mais frequentes é apresentado na Tabela 5. Esta contém o percentual de anúncios em relação ao total dos que tiveram algum dos 30 tipos de dados identificados. Outro dado presente na tabela é a soma das visualizações de anúncios que envolvem cada dado, que crescem de acordo com o número de anúncios.

Na análise da Tabela 5, é importante salientar que a categoria "Texto codificado" foi identificada a partir de vários nomes de algoritmos de *hashing* e criptografia. Ainda

Tabela 4. Valores dos vazamentos por indústria (Valores em USD).

Setor da indústria	Anúncios (n=1.317)			
	Mínimo	Percentil 85	Máximo	Preço médio*
Interactive Media & Services	1,00	10,01	200,00	1,00
Entertainment	1,00	10,01	999,00	2,34
IT Services	6,21	12,08	999,00	2,24
Hotels, Restaurants & Leisure	5,00	10,01	500,00	36,06
Software	5,00	33,60	135,77	0,56
Capital Markets	6,21	10,01	145,48	10,69
Professional Services	6,21	10,01	119,95	1,00
Diversified Consumer Services	10,00	17,67	274,00	3,37
Diversified Financial Services	6,21	69,00	165,00	11,11
Food & Staples Retailing	8,28	41,70	176,92	3,53

* Por milhão de registros

Tabela 5. Tipos de dados mais vazados.

Tipo de dado	Anúncios (n=1.117)		
	Qtde. anúncios	%	Soma de visualizações
Senha	868	77,71	85.373
E-mail	858	76,81	84.849
Texto claro	559	50,04	56.920
Nome de usuário	559	50,04	53.373
Texto codificado	344	30,80	32.741
Endereço IP	309	27,66	29.622
Nome	209	18,71	20.104
Data de nascimento	182	16,29	18.311
Número de telefone	142	12,71	15.099
Endereço	139	12,44	13.751

assim, foram encontrados mais registros ”plaintext”, ou seja, sem aplicação de controles de segurança. Há casos de vazamentos com um percentual dos dados codificados e o restante em texto claro, e nesses casos, o vazamento foi classificado em ambos os tipos.

Os autores acreditam que uma grande quantidade dos sites envolvidos nos anúncios exigem poucos dados para criação de um cadastro, como fóruns, e tornam a diversidade dos tipos de dados vazados menor. Por esse motivo, alguns tipos de dados que são extremamente cobiçados não apareceram na tabela. Para exemplificar, dados de pagamentos são apresentado em apenas 20 anúncios, e ficam fora da tabela por não possuírem uma quantidade de anúncios relevante diante a outros tipos de dados.

O tipo de dado identificado com maior frequência foi a ”Senha”. Na visão dos autores, uma das razões para isso ocorrer é a relação do dado ”Senha” com outros campos para autenticação, como e-mail ou nome de usuário, que são exibidos separadamente na tabela. No entanto, um ponto importante a ser notado sobre esse tipo de dado é o interesse despertado por pessoas mal-intencionadas, devido a possibilidade de acesso a outros sites que não o vazado. Conforme [Das et al. 2014], há muito reuso de senhas na

web, além de pequenas variações quando a alteração de senha é exigida. Para resolução desse reuso, uma possível solução é a aplicação de políticas de senha mais fortes nos sites, que exigissem rotatividade e validassem nível de semelhança com senhas antigas.

Outro ponto importante é a descoberta de 116 anúncios que possuem dados necessários para cadastro de pessoas em sites (nome, data de nascimento, número de telefone e endereço). Estes dados permitem que criminosos criem perfis falsos em muitos sites, por exemplo, redes sociais, que possibilitam a eles personificar a vítima. Por fim, na visão dos autores não há nenhuma relação entre os tipos de dados vazados e o interesse de compradores (número de visualizações), pois nenhum dos tipos de dados se destacou. Além disso, também não foi identificado nenhuma relação de interesse com o vendedor ou setor da indústria. Por esses motivos, os autores consideram que apenas o tempo de publicação do anúncio interfere em suas visualizações.

5. Conclusão

Os resultados do estudo apresentam um panorama geral dos dados disponíveis para venda em *marketplaces* da *dark web*. Com base no experimento realizado é possível afirmar que os modelos de aprendizado de máquina avaliados funcionam para reconhecimento de entidades em anúncios disponíveis em *marketplaces*. Neste estudo, os modelos apresentaram uma acurácia baixa (0,39 e 0,60), porém o uso em conjunto dos modelos melhorou significativamente a acurácia (0,87).

A análise dos dados apresenta que mais da metade dos anúncios estão associados com dois setores da indústria. O primeiro setor, Interactive Media & Services, contempla, entre outros, redes sociais e fóruns; já o setor Entertainment contempla plataformas de jogos e *streaming* online. Outra análise foi direcionada à quantidade de visualizações dos anúncios, o qual foi interpretado como interesse pelos mesmos. A partir dela não foi possível atrelar um maior interesse com um tipo específico de dado ou setor da indústria. A relação que pôde ser feita foi com a data de publicação de cada anúncio, pois as visualizações aumentam com o tempo desde anunciado.

A partir desse trabalho, é possível sugerir melhoria como a utilização de aprendizado de máquina para extração de outras informações, por exemplo, os tipos de dados vazados. Outra sugestão é a avaliação de outros modelos de aprendizado de máquina para NERC, como modelos mais recentes, ou baseados em redes neurais recorrentes ou convolucionais.

Referências

Al-Ramahi, M., Alsmadi, I., and Davenport, J. (2020). Exploring hackers assets: Topics of interest as indicators of compromise. In *ACM Int. Conf. Proceeding Ser.*, pages 38–41, New York, NY, USA. Association for Computing Machinery.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Das, A., Bonneau, J., Caesar, M., Borisov, N., and Wang, X. (2014). The tangled web of password reuse. In *NDSS*, volume 14, pages 23–26.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, F., Yuan, S., Ou, H., and Liu, L. (2018). New cyber threat discovery from darknet marketplaces. In *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 62–67.

Fang, Y., Guo, Y., Huang, C., and Liu, L. (2019). Analyzing and identifying data breaches in underground forums. *IEEE Access*.

Feldman, R. et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Fu, T., Abbasi, A., and Chen, H. (2010). A focused crawler for dark web forums. *Journal of the American Society for Information Science and Technology*.

Hrazdil, K. and Zhang, R. (2012). The importance of industry classification in estimating concentration ratios. *Economics Letters*, 114(2):224–227.

Huang, S.-Y. and Ban, T. (2019). A topic-based unsupervised learning approach for online underground market exploration. In *2019 18th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. IEEE Int. Conf. Big Data Sci. Eng.*, pages 208–215. IEEE.

Huang, S.-Y. and Chen, H. (2016). Exploring the online underground marketplaces through topic-based social network and clustering. In *2016 IEEE Conf. Intell. Secur. Informatics*, pages 145–150. IEEE.

IBM (2019). 2018 cost of data breach study: Impact of business continuity management. Disponível em: https://www.ibm.com/. Acesso em: ago. 2019.

Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27.

Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., and Shimada, H. (2019). Identification of cybersecurity specific content using the doc2vec language model. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*.

Park, A. J., Frank, R., Mikhaylov, A., and Thomson, M. (2018). Hackers hedging bets: A cross-community analysis of three online hacking forums. In *2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, pages 798–805. IEEE.

Samtani, S., Zhu, H., and Chen, H. (2020). Proactively identifying emerging hacker threats from the dark web. *ACM Trans. Priv. Secur.*, 23(4):1–33.

Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., and Vieira, R. (2019). Assessing the impact of contextual embeddings for portuguese named entity recognition. In *2019 8th Brazilian Conf. on Intelligent Systems (BRACIS)*, pages 437–442. IEEE.

Sapienza, A., Ernala, S. K., Bessi, A., Lerman, K., and Ferrara, E. (2018). Discover: Mining online chatter for emerging cyber threats. In *Web Conf. 2018 - Companion World Wide Web Conf. WWW 2018*, pages 983–990. Association for Computing Machinery.