Universidade Federal do Rio de Janeiro -- UFRJ

Instituto de Computação -- IC

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais – NCE

Programa de Pós-Graduação em Informática -- PPGI

Felipe Adrian Moreno Vera

Detecção Automatizada de Exploração de Vulnerabilidades em Fóruns de Hacking Clandestinos

Rio de Janeiro

2024

Universidade Federal do Rio de Janeiro -- UFRJ

Instituto de Computação -- IC

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais – NCE

Programa de Pós-Graduação em Informática -- PPGI

Felipe Adrian Moreno Vera

Detecção Automatizada de Exploração de Vulnerabilidades em Fóruns de Hacking Clandestinos

Dissertação de Mestrado submetida ao Programa de Pós-graduação em Informática do Instituto de Computação (IC) e do Insti-tuto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE), da Universidade Fed-eral do Rio de Janeiro (UFRJ). Como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Supervisor: Ph.D. Daniel Sadoc Menasché

Rio de Janeiro

2024

Este trabalho é dedicado ao apoio e orientações recebidas por familiares, companheiros alunos e professores colaboradores. Também aos meus familiares que sempre estão me apoiando nas minhas decisões e me motivando a continuar.

AGRADECIMENTOS

Os agradecimentos principais são direcionados aos alunos e professores que estiveram envolvidos e contribuíram com este trabalho. Além disso, agradecemos à Universidade de Cambridge, por compartilhar conosco os dados do CrimeBB, usados nesta proposta como estudo de caso. Agradecemos também às entidades de desenvolvimento científico, incluindo **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior** (CAPES – 315110/2020-1), **Conselho Nacional de Desenvolvimento Científico e Tecnológico** (CNPq – E-26/211.144/2019) e **Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro** (FAPERJ – E-26/201.376/2021) que ajudaram neste período de Mestrado em Informática na **Universidade Federal do Rio de Janeiro** (UFRJ), no **Programa Pós-Graduação em Informática** (PPGI).

RESUMO

Este trabalho propõe uma abordagem baseada em aprendizado de máquina para identificar e classificar a exploração de vulnerabilidades, o escopo e o impacto de softwares maliciosos por meio do monitoramento de fóruns de hackers clandestinos. O volume crescente de postagens discutindo a exploração de vulnerabilidades exige uma abordagem automatizada para processar tópicos e postagens que possam acionar alarmes com base em seu conteúdo. Para ilustrar o sistema proposto, utilizamos o conjunto de dados CrimeBB, composto por dados extraídos de vários fóruns clandestinos, e desenvolvemos um modelo de aprendizado de máquina supervisionado capaz de filtrar tópicos que citam CVEs e rotulá-los como prova de conceito (PoC), armamento (Weaponization), exploração (Exploitation), entre outros. Aplicamos técnicas de aprendizado de máquina e processamento de linguagem natural para pré-processar os textos e, em seguida, treinamos diversos modelos lineares e não lineares. Para avaliar a eficácia e a precisão dos modelos e comparar os resultados, utilizamos as métricas de exatidão (accuracy), precisão (precision) e recuperação (recall). Além disso, para maior compreensão e explicação sobre o motivo de o modelo diferenciar entre as classes, utilizamos métodos de explicação de modelos para determinar a relevância das palavras nas predições. No geral, este trabalho destaca as diferenças entre as naturezas das postagens rotuladas e sua importância na análise de vulnerabilidades.

Palavras-chave: Cíber segurança; fóruns online; mineração de dados; processamento de linguagem natural; segurança da informação; modelos de linguagem; interpretabilidade.

ABSTRACT

This work proposes a machine learning-based approach to identify and classify vulnerability exploitation, the scope, and the impact of malicious software through the monitoring of underground hacker forums. The growing volume of posts discussing vulnerability exploitation demands an automated approach to process topics and posts that may trigger alarms based on their content. To illustrate the proposed system, we used the CrimeBB dataset, which contains data extracted from various underground forums, and developed a supervised machine learning model capable of filtering topics that cite CVEs and labeling them as proof of concept (PoC), weaponization, exploitation, among others. We applied machine learning and natural language processing techniques to preprocess the texts and then trained several linear and non-linear models. To evaluate the models' effectiveness and accuracy and compare results, we used metrics such as accuracy, precision, and recall. Additionally, to gain a better understanding and explanation of why the model can differentiate between classes, we employed model explanation methods to determine the relevance of words in predictions. Overall, this work highlights the differences in the nature of the labeled posts and their importance in vulnerability analysis.

Keywords: Cybersecurity; online forums; data mining; natural language processing; information security; large language models; Generative Pre-trained Transformer; interpretability.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
EPSS	Exploit Prediction Scoring System
GPT	Generative Pre-trained Transformer
NLP	Natural Language Processing
NVD	National Vulnerability Database

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, descrevemos as principais motivações para o desenvolvimento deste trabalho. Também definimos o problema que pretendemos abordar e os objetivos que buscamos alcançar.

1.1 CONTEXTO E MOTIVAÇÃO

O impacto global do cibercrime sobre corporações e organizações governamentais é significativo, com investimentos em contramedidas que alcançam cerca de 2 trilhões de dólares ao longo dos anos (ANDERSON et al., 2019; MORROW, 2019). Ao explorar vulnerabilidades em softwares e hardwares, diversos ecossistemas da internet podem ser facilmente enganados ou atacados. Isso representa uma ameaça para usuários finais, empresas e para a estabilidade geral da internet. Portanto, a detecção precoce da instrumentalização (weaponization) e das tentativas de exploração é essencial para a defesa contra esses ataques.

A exploração de vulnerabilidades envolve o uso de um exploit instrumentalizado para atacar um alvo, permitindo que um atacante explore uma vulnerabilidade para obter acesso não autorizado a um sistema ou roubar informações sensíveis. Embora bases de dados públicas, como [1]ExploitDB, forneçam continuamente informações atualizadas sobre como explorar vulnerabilidades, fóruns clandestinos de hackers ainda oferecem informações exclusivas e mais recentes sobre a disponibilidade e o desenvolvimento de exploits, bem como sobre o possível uso desses exploits no mundo real. Esses fóruns são documentados em estudos recentes (BASHEER; ALKHATIB, 2021; CAMPOBASSO; ALLODI, 2022; PASTRANA; THOMAS et al., 2018).

Em particular, o monitoramento de fóruns clandestinos de hackers é crucial para detectar e neutralizar a exploração de vulnerabilidades no mundo real e identificar os principais usuários envolvidos nessas atividades. Alguns fóruns também fornecem informações sobre o desenvolvimento de novos exploits, preços mais recentes e instruções sobre como tornar ataques completamente indetectáveis (FUD – *Fully Undetectable*) (ABLON; LIBICKI; GOLAY, 2014). Nesse contexto, compreender o que os usuários discutem nesses fóruns ajuda a rastrear preços de exploits, seu uso, demanda e os principais alvos.

O desenvolvimento de um ciberataque envolve diversas etapas, incluindo a instrumentalização (*weaponization*) e a exploração. Embora a maior parte da literatura tenha se concentrado na instrumentalização (HANKS et al., 2022; ALLODI, 2017), isto

[1] https://www.exploit-db.com/

é, na construção de exploits para vulnerabilidades, a exploração no mundo real recebeu menos atenção. A exploração refere-se ao uso efetivo de um exploit instrumentalizado para atacar um alvo ou obter acesso não autorizado a um sistema ou a informações sensíveis (BASHEER; ALKHATIB, 2021). Essa menor atenção ocorre, em parte, devido ao envolvimento de dados sensíveis e a acordos rigorosos de confidencialidade na análise da exploração no mundo real.

Neste estudo, utilizamos o conjunto de dados CrimeBB, uma compilação de dados coletados de diversos fóruns clandestinos, fornecida pelo Cambridge Cybercrime Centre (PASTRANA; THOMAS et al., 2018). Em nosso trabalho (MORENO-VERA et al., 2023), propusemos uma abordagem de aprendizado de máquina, utilizando o CrimeBB, para classificar postagens de fóruns que discutem Vulnerabilidades e Exposições Comuns (*Common Vulnerabilities and Exposures* – CVE) em categorias como *Proof-of-Concept* (PoC) e Exploração, alcançando valores de F1 superiores a 90%. Informações de *ground truth* não estavam disponíveis no CrimeBB, e um grande esforço em trabalhos anteriores envolveu a rotulação manual de postagens e tópicos. Também aplicamos uma abordagem de aprendizado não supervisionado para inferir tópicos com base no conteúdo das threads (MORENO-VERA, 2023), nas quais diversas palavras como "server" ou "tool" podem pertencer a discussões de *Proof-of-Concept* ou *Weaponization*. Além disso, uma nova extensão do CrimeBB, chamada PostCog, fornece rótulos de *ground truth* (com ruído) para um subconjunto de postagens do CrimeBB. Para análises adicionais, também integramos a extensão PostCog ao conjunto de dados, permitindo uma rotulação aprimorada para cada tipo de postagem, intenção e tipo de crime. Notavelmente, termos como "fud" (completamente indetectável) e "pm" (*private message*) emergiram como indicadores significativos de exploração. Esses termos já haviam sido identificados como relevantes em (MORENO-VERA et al., 2023), e nossos novos resultados, utilizando a extensão PostCog, reforçam algumas de nossas descobertas anteriores.

1.2 OBJETIVOS

Nesta seção, apresentamos de forma concisa os objetivos deste trabalho. Nosso objetivo principal é investigar métodos para prever discussões relacionadas ao cibercrime. Para isso, utilizamos o conjunto de dados CrimeBB e propomos uma metodologia para realizar essa tarefa. A seguir, apresentamos os objetivos definidos neste trabalho.

1.2.1 **Objetivo geral**

O objetivo geral deste trabalho é avaliar diferentes modelos baseados em texto para prever discussões relacionadas ao cibercrime em postagens e threads utilizando o conjunto de dados *CrimeBB* e a extensão PostCog. Para atingir esse objetivo, exploraremos,

analisaremos e apresentaremos os resultados obtidos em nosso estudo.

1.2.2 **Objetivos específicos**

Em particular, listamos os seguintes objetivos específicos:

- Propomos uma metodologia para explorar e analisar o conjunto de dados *CrimeBB*, que contém 117.365.492 postagens provenientes de 37 fóruns clandestinos. Nosso objetivo é fornecer insights sobre o comportamento e os padrões presentes nos dados. O foco principal do nosso estudo é o fórum *HackForums*.
- Propomos e apresentamos um modelo baseado em texto capaz de classificar de forma eficiente discussões relacionadas ao cibercrime em conteúdos de threads e postagens, considerando o comportamento e a distribuição dos dados analisados.
- Aplicamos métodos de explicabilidade para compreender e analisar as previsões do modelo, com o objetivo de identificar palavras relevantes que forneçam insights sobre as discussões presentes em fóruns.
- Utilizamos *Generative Pre-trained Transformers* (GPT) para analisar e obter novas informações a partir das postagens e seus conteúdos. Buscamos avaliar o desempenho de modelos GPT na classificação de informações não estruturadas, como o conteúdo de fóruns.

1.3 METODOLOGIA

A metodologia proposta, ilustrada na Figura 1, consiste em três tarefas principais: (i) preparação dos dados, rotulação, extração de características e treinamento de um modelo preditivo; (ii) predição de novas amostras; e (iii) explicação dos modelos. Na fase de preparação dos dados (etapas 1–3), coletamos dados tanto do NVD quanto do CrimeBB, filtrando as postagens do CrimeBB para manter apenas aquelas que fazem referência a identificadores CVE e extraindo todos os CVEs citados. Em seguida, recuperamos as threads que contêm essas postagens e geramos automaticamente um conjunto de características para cada thread. Posteriormente, utilizando rótulos fornecidos pelo PostCog, anotações de especialistas e GPT, treinamos um modelo de aprendizado de máquina para classificar o conteúdo das discussões nas threads (etapa 4). Na fase de inferência (etapa 5), aplicamos o modelo treinado para classificar o conteúdo de novas postagens em threads, prevendo os rótulos mais relevantes com base no conteúdo textual. Por fim, aplicamos técnicas de explicação (etapa 6) para analisar as previsões e obter insights a partir das inferências.

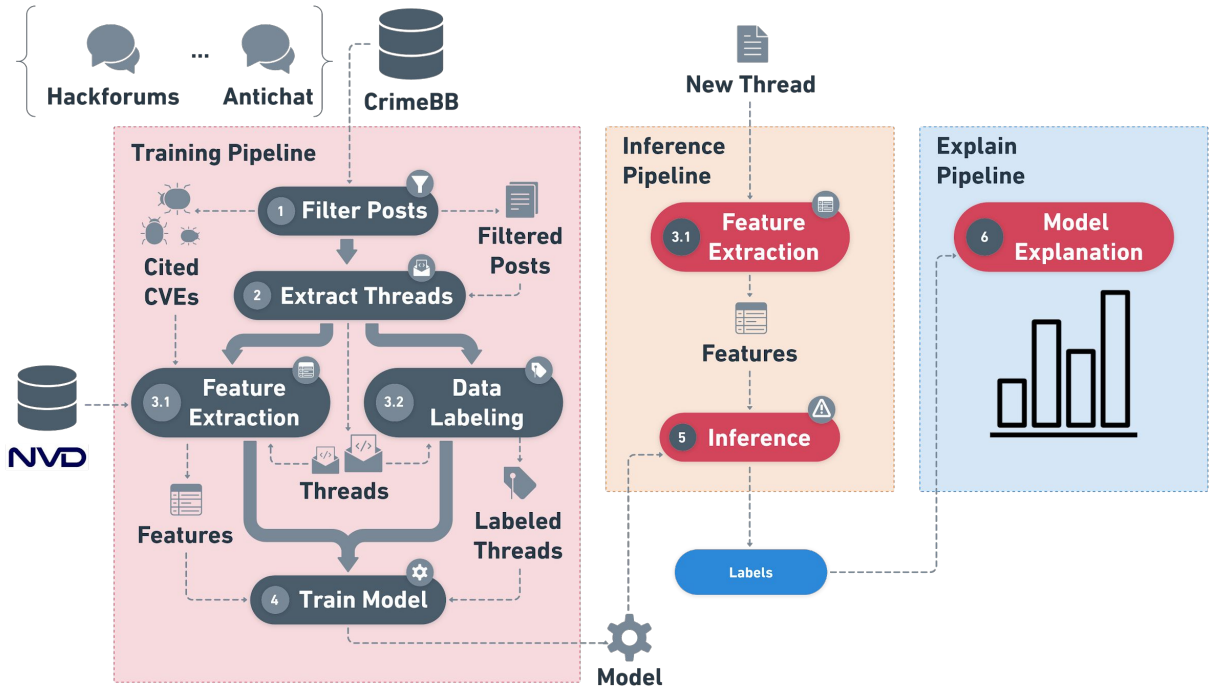


Figura 1 – Framework proposto composto por duas etapas principais: (1) pré-processamento de threads e postagens, incluindo extração de características e rotulação; (2) modelo classificador de três classes para prever o conteúdo de novas threads; (3) geração de explicações do modelo para compreender as previsões.

1.4 CONTRIBUIÇÕES

Já publicamos insights e resultados obtidos a partir de experimentos realizados: (i) estudamos o conteúdo de threads no HackForums, analisando o CVSS e o EPSS de códigos CVE presentes nas postagens; esse trabalho foi publicado no IEEE CSR: *IEEE International Conference on Cyber Security and Resilience (CSR '23)*. (ii) utilizamos modelagem de tópicos para analisar e inferir o que está sendo discutido dentro das threads e postagens sem anotações; esse trabalho foi publicado no IEEE ICTC: *IEEE International Conference on Information and Communication Technology Convergence (ICTC '23)*. (iii) analisamos a extensão PostCog e as saídas do GPT, comparamos com modelos em ensemble e aplicamos explicações de modelos; esse trabalho foi apresentado no *International Symposium on Cyber Security, Cryptology and Machine Learning (CSCML '24)*.

- Felipe Moreno-Vera, Daniel S. Menasché e Cabral Lima. "Beneath the Cream: Unveiling Relevant Information Points from CrimeBB Underground Forums with Its Ground Truth Labels". In *The International Symposium on Cyber Security, Cryptology and Machine Learning* (CSCML '24), dezembro de 2024, Be'er Sheva, Israel (MORENO-VERA; MENASCHE; LIMA, 2024).

- Felipe Moreno-Vera, Mateus Nogueira, Cainã Figueiredo, Daniel S. Menasché, Miguel Bicudo, Ashton Woiwood, Enrico Lovat, Anton Kocheturov e Leandro Pfleger de Aguiar. "Cream Skimming the Underground: Identifying Relevant Information Points from Online Forums". In *IEEE International Conference on Cyber Security and Resilience* (CSR '23), agosto de 2023, Veneza, Itália (MORENO-VERA et al., 2023).

- Felipe Moreno-Vera. "Inferring Discussion Topics about Exploitation of Vulnerabilities from Underground Hacking Forums". In *International Conference on Information and Communication Technology Convergence* (ICTC '23), outubro de 2023, Seul, Coreia do Sul (MORENO-VERA, 2023).

1.5 ESTRUTURA DO DOCUMENTO

O restante deste trabalho está organizado da seguinte forma. O Capítulo 2 apresenta conceitos e definições importantes para uma melhor compreensão deste documento. O Capítulo 3 discute trabalhos relacionados. O Capítulo 4 descreve o conjunto de dados utilizado, a análise exploratória realizada e a metodologia adotada. O Capítulo 5 apresenta os principais experimentos e resultados, e o Capítulo 6 apresenta as conclusões.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, descrevemos os conceitos fundamentais de aprendizado de máquina, aprendizado profundo e suas aplicações em processamento de linguagem natural. Por meio de uma descrição abrangente de métodos, técnicas e modelos importantes, buscamos apresentar e definir os conceitos básicos necessários para a compreensão dos capítulos seguintes.

2.1 APRENDIZADO DE MÁQUINA E APRENDIZADO PROFUNDO

O aprendizado de máquina e o aprendizado profundo são dois ramos poderosos da inteligência artificial que transformaram a forma como abordamos problemas complexos em diversas áreas. Em essência, ambas as abordagens buscam permitir que computadores aprendam a partir de dados, possibilitando que realizem previsões ou tomem decisões sem a necessidade de programação explícita para cada tarefa específica (BENGIO, 2007). A combinação de aprendizado de máquina e aprendizado profundo tem impulsionado avanços significativos em uma ampla variedade de aplicações, incluindo carros autônomos, assistentes virtuais, diagnósticos médicos, análise de percepção urbana e sistemas de recomendação.

O aprendizado de máquina engloba uma variedade de algoritmos e técnicas, desde métodos clássicos como regressão linear e árvores de decisão até abordagens mais avançadas, como máquinas de vetor de suporte e florestas aleatórias. Ele permite que computadores aprendam padrões e relações a partir dos dados, possibilitando a realização de previsões e decisões informadas. Além disso, o aprendizado profundo tornou-se uma abordagem poderosa para lidar com dados complexos e de alta dimensionalidade, especialmente em áreas como o processamento de linguagem natural (PLN), onde dados textuais apresentam desafios específicos. Em sua essência, o aprendizado profundo baseia-se em redes neurais inspiradas na estrutura dos neurônios humanos. Essas redes são compostas por camadas de neurônios interconectados que processam e transformam dados, aprendendo representações da linguagem diretamente a partir de textos brutos (GOODFELLOW; BENGIO; COURVILLE, 2016). Em PLN, modelos como redes neurais recorrentes (RNNs) lidam com dados sequenciais, capturando dependências temporais, enquanto os transformers revolucionaram o processamento de linguagem ao utilizar mecanismos de autoatenção para compreender o contexto em sentenças ou documentos inteiros. O aprendizado profundo em PLN pode ser aplicado por meio de diferentes tipos de aprendizado: supervisionado (utilizando textos

rotulados para classificação ou tradução), não supervisionado (descoberta de padrões sem rótulos, como modelagem de tópicos) e semi-supervisionado ou auto-supervisionado (como nos modelos Generative Pre-trained Transformer). Esses tipos de aprendizado abrangem diferentes tarefas e formulações de problemas, permitindo que modelos de aprendizado profundo obtenham excelentes resultados em diversas aplicações de PLN, desde análise de sentimentos até geração de linguagem.

2.1.1 Tipos de aprendizado

Um tipo de aprendizado refere-se à abordagem geral ou paradigma que define como um algoritmo de aprendizado de máquina aprende a partir de dados. Os principais tipos de aprendizado incluem:

- **Aprendizado Supervisionado**: No aprendizado supervisionado, o modelo é treinado com um conjunto de dados rotulado, no qual cada exemplo está associado a um rótulo ou resultado esperado. O sistema aprende a associar dados de entrada a rótulos de saída minimizando uma função de perda, o que melhora gradualmente sua capacidade de predição. Exemplos de tarefas incluem classificação de texto, tradução automática, etiquetagem de partes do discurso (Part-of-Speech Tagging – POS) e reconhecimento de entidades nomeadas (Named Entity Recognition – NER).
- **Aprendizado Não Supervisionado**: No aprendizado não supervisionado, o modelo é treinado com dados não rotulados, nos quais não existem rótulos ou resultados esperados. Nesse caso, o modelo busca descobrir padrões, estruturas ou relações presentes nos dados, como agrupamentos ou características latentes. Exemplos de tarefas incluem agrupamento (clustering), modelagem de tópicos, representações vetoriais de palavras (word embeddings) e redução de dimensionalidade.
- **Aprendizado Semi-supervisionado / Auto-supervisionado**: O aprendizado semi-supervisionado combina aspectos do aprendizado supervisionado e não supervisionado. Nesse caso, o sistema é treinado com um conjunto de dados que contém uma pequena quantidade de exemplos rotulados e uma grande quantidade de dados não rotulados. O modelo utiliza os dados rotulados para orientar o aprendizado e os dados não rotulados para melhorar seu desempenho e capacidade de generalização. Exemplos incluem geração de texto, embeddings de sentenças e sumarização automática.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O Processamento de Linguagem Natural (PLN) é uma área da inteligência artificial focada em permitir que máquinas compreendam, interpretem e gerem linguagem humana.

A necessidade do PLN surge do fato de que a linguagem humana é inerentemente complexa, rica em nuances e frequentemente ambígua, tornando seu processamento um desafio para sistemas computacionais. A Figura 3 apresenta um resumo dos resultados de representação textual utilizando diferentes técnicas.

2.2.1 Técnicas de processamento de texto

No Processamento de Linguagem Natural, o pré-processamento dos dados é essencial para melhorar a precisão e a eficiência da análise textual. A seguir são apresentadas algumas etapas importantes de limpeza e preparação do texto.

Remoção de stopwords

Stopwords são palavras frequentemente utilizadas em uma língua (como “o”, “é”, “e”, “em”, etc.) que geralmente não carregam significado relevante. A remoção dessas palavras ajuda a reduzir ruído nos dados textuais, permitindo que os modelos se concentrem em termos mais informativos. Ao filtrar palavras muito frequentes e pouco informativas, a dimensionalidade dos dados diminui, melhorando o desempenho do modelo e a eficiência computacional. Por exemplo, na frase “O gato está sobre o tapete”, ao remover stopwords obtemos “gato tapete”, o que é mais significativo para o modelo.

Remoção de emojis

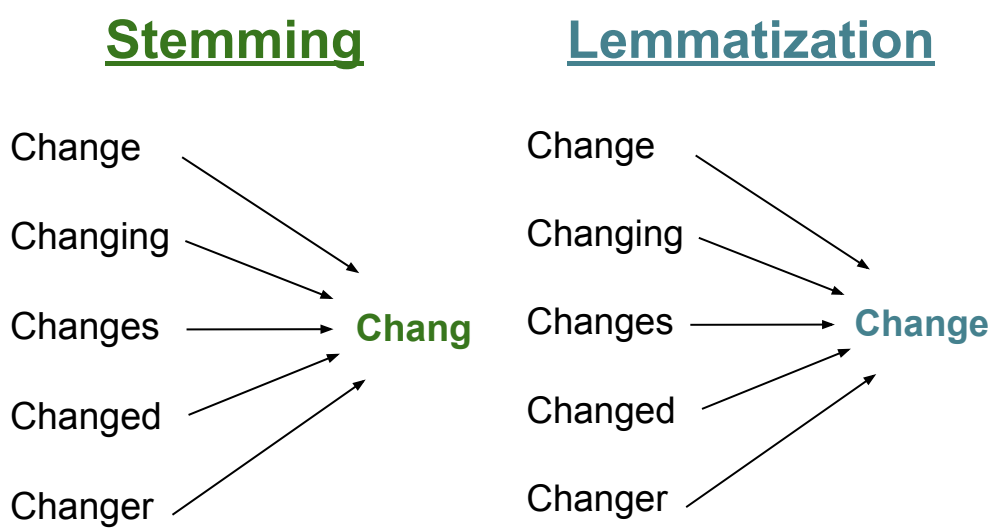
Emojis são ícones utilizados em comunicações digitais para expressar emoções, ações ou objetos. Dependendo do contexto, eles podem introduzir ruído ou fornecer informações úteis sobre sentimento. Em tarefas gerais de PLN focadas em palavras, a remoção de emojis pode simplificar os dados. Entretanto, em análise de sentimentos, eles podem fornecer informações relevantes se processados adequadamente.

Remoção de pontuação

Sinais de pontuação (como “.”, “,”, “!”, “?”) são usados para estruturar frases e esclarecer significados, mas muitas vezes não são necessários em modelos de PLN. Em tarefas focadas apenas no conteúdo lexical, como classificação de textos ou modelagem de tópicos, a pontuação pode ser redundante. Sua remoção reduz a complexidade textual e facilita o processo de tokenização.

Remoção de caracteres especiais

Caracteres especiais (como “#”, “@”, “$” e “&”) são símbolos que geralmente não trazem informação relevante para tarefas comuns de PLN. Eles são particularmente



(a) Stemming | (b) Lemmatization

Figura 2 – Técnicas de redução de palavras em PLN: (a) Stemming e (b) Lemmatization.

frequentes em dados provenientes de redes sociais e podem introduzir ruído nos dados. Removê-los melhora a tokenização e reduz o tamanho do vocabulário.

Redução de palavras

Stemming e Lemmatization são técnicas utilizadas para reduzir palavras às suas formas básicas.

- **Stemming**: método baseado em regras que remove terminações das palavras para obter sua raiz. Esse processo pode gerar palavras que não existem formalmente.
- **Lemmatization**: método mais sofisticado que reduz palavras à sua forma de dicionário (lema), considerando o contexto e a classe gramatical da palavra.

Em geral, a lematização é mais precisa por considerar o contexto, enquanto o stemming é mais rápido, porém menos preciso.

2.2.2 Abordagens iniciais de representação textual

O processamento de texto começou com tarefas simples como tokenização, que consiste em dividir frases em unidades menores chamadas tokens. Métodos iniciais baseados em regras conseguiam identificar padrões simples, mas tinham dificuldade em lidar com a complexidade da linguagem humana.

Uma das técnicas mais influentes foi o modelo *Bag-of-Words* (BoW) (HARRIS, 1954), que transforma textos em frequências de palavras por meio de uma matriz Documento-Termo (DTM), onde linhas representam documentos e colunas representam termos presentes no corpus.

Corpus

Document 1	I like cats
Document 2	cats are the best, they are awesome
Document 3	also dogs are nice

Document-Term Matrix

Words	I	like	cats	are	the	best	they	awesome	also	dogs	nice
Document 1	1	1	1	0	0	0	0	0	0	0	0
Document 2	0	0	1	2	1	1	1	1	0	0	0
Document 3	0	0	0	1	0	0	0	0	1	1	1

Doc2Vec

Vectors	x1	x2	x3	x4	x5	x6	x7	...
Document 1	0.35	0.86	1.82	3.48	1.05	10.15	8.63	...
Document 2	0.84	0.45	3.45	4.49	2.64	2.87	13.97	...
Document 3	0.39	1.0	0.98	7.92	5.14	6.19	20.98	...

Bag-Of-Words (1-2-gram)

Words	I	like	cats	I like	like cats	are	the	best	they	awesome	cats are	the best	they are	...
Document 1	1	1	1	1	1	0	0	0	0	0	0	0	0	...
Document 2	0	0	1	0	0	2	1	1	1	1	1	1	1	...
Document 3	0	0	0	0	0	1	0	0	0	0	0	0	0	...

TF-IDF (1-2-gram)

Words	I	like	cats	I like	like cats	are	the	best	they	awesome	cats are	the best	they are	...
Document 1	0	0	0.47	0.62	0.62	0	0	0	0	0	0	0	0	...
Document 2	0	0	0.33	0	0	0.56	0.43	0	0	0	0.43	0.43	0.13	...
Document 3	0	0	0	0	0	0.35	0	0	0	0	0	0	0	...

Figura 3 – Exemplo de métodos Bag-of-Words, TF-IDF e Doc2Vec. Os dois primeiros utilizam a matriz Documento-Termo (DTM), enquanto o Doc2Vec gera representações vetoriais de sentenças ou parágrafos.

Bag-of-Words (BoW)

O modelo Bag-of-Words representa um documento como uma coleção não ordenada de palavras, ignorando gramática e ordem das palavras.

$$v_d = [freq(w_1, d), freq(w_2, d), \ldots, freq(w_k, d)] \tag{2.1}$$

Onde $freq(w_i, d)$ representa o número de ocorrências da palavra w_i no documento d.

TF-IDF

TF-IDF é uma medida estatística que avalia a importância de uma palavra em um documento em relação ao corpus.

$$TF(w, d) = \frac{freq(w, d)}{\sum_{i \in d} freq(w_i, d)} \tag{2.2}$$

$$IDF(w, D) = log(\frac{|D|}{1 + |d \in D : w \in d|}) \tag{2.3}$$

O valor final TF-IDF é obtido multiplicando TF e IDF.

2.2.3 Aprendizado de máquina e representações vetoriais

A verdadeira revolução no processamento de texto ocorreu com o avanço do aprendizado profundo. Modelos como Word2Vec (MIKOLOV et al., 2013) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) passaram a representar palavras como vetores densos em espaços multidimensionais, capturando relações semânticas entre palavras.

Posteriormente, Doc2Vec (LE; MIKOLOV, 2014) estendeu esse conceito para representar documentos completos como vetores.

Word2Vec

Word2Vec utiliza redes neurais para aprender representações vetoriais distribuídas de palavras. Dois modelos principais são utilizados:

- Continuous Bag of Words (CBOW)
- Skip-gram

$$J(\theta) = -\sum_{t=1}^{T} \sum_{-m \leq j \leq m, j \neq 0} log(p(w_{t+j}|w_t)) \tag{2.4}$$

Doc2Vec

Doc2Vec estende Word2Vec para representar documentos completos como vetores.

$$J(\theta) = -\sum_{t=1}^{T} \sum_{-m \leq j \leq m, j \neq 0} log(p(w_{t+j}|w_t, d)) \tag{2.5}$$

2.2.4 Aprendizado profundo e Transformers

Apesar dos avanços, ainda havia espaço para melhorias. A introdução da arquitetura Transformer revolucionou o PLN ao utilizar mecanismos de autoatenção para capturar relações entre palavras independentemente da distância entre elas. Isso levou ao desenvolvimento de modelos como BERT (DEVLIN et al., 2018) e GPT (RADFORD; NARASIMHAN, 2018).

GPT (Generative Pre-trained Transformer)

GPT é um modelo de linguagem causal baseado em abordagem autoregressiva, que prevê o próximo token dado o contexto anterior.

$$J(\theta) = -\sum_{t=1}^{T} log(p(w_{t+1}|w_1, w_2, \ldots, w_t)) \tag{2.6}$$

Onde T representa o número total de tokens na sequência e θ representa os parâmetros da rede neural.

2.3 CONSIDERAÇÕES FINAIS

Este capítulo apresentou e descreveu as principais metodologias, pipelines e modelos aplicados no processamento de texto utilizando processamento de linguagem natural. Foram abordados desde algoritmos clássicos de aprendizado de máquina até modelos baseados em aprendizado profundo, destacando os principais tipos de aprendizado utilizados para extrair características relevantes de informações textuais. Além disso, discutimos técnicas amplamente utilizadas como Bag-of-Words, TF-IDF e Doc2Vec, fundamentais para a representação e análise de textos.

3 TRABALHOS RELACIONADOS

Neste capítulo, apresentamos uma visão geral da literatura relacionada e do referencial teórico relevante para os principais temas deste trabalho.

3.1 FÓRUNS BLACKHAT

Fóruns blackhat são compostos por postagens não estruturadas. Todas as postagens incluem seu conteúdo, autor e assunto. Esses fóruns oferecem um meio para que pesquisadores, assim como usuários mal-intencionados, troquem conhecimento sobre invasão e exploração de sistemas. Eles fornecem informações que vão desde habilidades básicas de hacking até ferramentas funcionais que qualquer pessoa pode obter com facilidade, às vezes gratuitamente. Os chamados *script-kiddies*, isto é, usuários domésticos com conhecimento computacional limitado, podem, por exemplo, utilizar essas ferramentas para iniciar ciberataques. Um de nossos objetivos é distinguir atividades de pesquisa que representam ameaça potencial da exploração no mundo real, na qual criminosos discutem ameaças.

Apesar de os fóruns blackhat fornecerem uma fonte rica de informações sobre vulnerabilidades, curar e utilizar dados provenientes desses fóruns é uma tarefa desafiadora. Entre os desafios, destacamos os seguintes.

1. **Tempestividade.** Devido à sua natureza pública, democrática e distribuída, os fóruns blackhat são candidatos naturais a serem os primeiros locais onde vulnerabilidades são estudadas e discutidas. Hackers normalmente encontram nesses fóruns um ambiente fértil para discutir exploits e estratégias de *weaponization*. No entanto, a filtragem e a detecção precoce de atividades relacionadas a uma determinada vulnerabilidade não são tarefas triviais. Embora seja possível encontrar retrospectivamente, de forma eficaz, informações nesses fóruns sobre vulnerabilidades já bem conhecidas, por exemplo, após terem sido catalogadas pelo NIST, detectar atividade relevante nos fóruns blackhat antes que os sistemas sejam comprometidos continua sendo um grande desafio.

2. **Confiabilidade do relato e impacto.** A detecção precoce de atividades relacionadas a vulnerabilidades é desafiadora, em parte devido ao compromisso entre tempestividade e confiabilidade do relato. Mensagens iniciais sobre uma determinada vulnerabilidade geralmente tratam de tentativas de desenvolvimento de armas *proof-of-concept* (PoC), que podem acabar se mostrando ineficazes, tornando difícil determinar seu impacto sobre o risco associado à vulnerabilidade.

3. **Correspondência entre postagens e vulnerabilidades.** As postagens em fóruns blackhat são não estruturadas. Não é trivial relacioná-las a vulnerabilidades reportadas recentemente, sobre as quais pode haver pouca informação disponível. Embora algumas postagens citem explicitamente a vulnerabilidade a que se referem, outras podem mencionar apenas os produtos ou fabricantes afetados.

Em resumo, para que fóruns blackhat forneçam dados úteis e tempestivos sobre vulnerabilidades, é fundamental superar os desafios associados tanto à avaliação da qualidade das informações presentes nas postagens quanto ao processo de correspondência entre essas postagens e vulnerabilidades. Com esse objetivo, neste trabalho empregamos uma abordagem automatizada de mineração de texto para explorar o amplo espectro de dados disponíveis em fóruns blackhat e extrair eventos relevantes nos estágios iniciais do ciclo de vida das vulnerabilidades.

3.1.1 Por que utilizar fóruns blackhat para a análise do ciclo de vida de vulnerabilidades?

As informações coletadas em fóruns blackhat e em *feeds* de *Threat Intelligence* (TI) podem servir como *ground truth* para modelos que predizem explorabilidade, como EPSS e *Expected Exploitability* (SUCIU et al., 2022; JACOBS; ROMANOSKY et al., 2021). Essas plataformas também podem melhorar a *correspondência entre vulnerabilidades e exploits*, que frequentemente não é trivial, ao identificar CVEs que ainda não foram associados a exploits correspondentes. Além disso, fóruns blackhat e *feeds* de TI podem ser utilizados para *encontrar novos exploits* antes que eles sejam reportados em bases públicas como o ExploitDB. A presença de uma citação de CVE em um fórum blackhat também pode motivar um *aumento no escore temporal do CVSS*, pois pode indicar maior risco associado à vulnerabilidade. Por fim, fóruns blackhat podem ser usados para identificar automaticamente *atividades de desenvolvimento de recursos*, isto é, atividades nas quais adversários estabelecem infraestrutura, contas ou capacidades para apoiar seus objetivos. Ao analisar dados desses fóruns, é possível obter insights sobre o desenvolvimento desses recursos, identificando novos métodos de ataque e vulnerabilidades.[1]

3.2 PLN E INTELIGÊNCIA DE AMEAÇAS (TI)

O uso de Processamento de Linguagem Natural (PLN) para a análise de fóruns de hackers foi considerado em diversos trabalhos anteriores (RAHMAN et al., 2021; PASTRANA; HUTCHINGS et al., 2019; PASTRANA; THOMAS et al., 2018; MORENO-VERA et al., 2023). Neste trabalho, complementamos esse corpo da literatura ao focar

[1] Ver Mitre Att&ck Develop Capabilities https://attack.mitre.org/matrices/enterprise/

em discussões sobre vulnerabilidades de software em fóruns do CrimeBB, que já foi anteriormente utilizado na análise de *eWhoring* (PASTRANA; HUTCHINGS et al., 2019) e de outros tipos de cibercrime (PASTRANA; THOMAS et al., 2018; DEGUARA et al., 2022).

Nossa pesquisa tem como objetivo analisar discussões sobre vulnerabilidades de software nos fóruns do CrimeBB. O sistema Threat Miner (DEGUARA et al., 2022) foi desenvolvido para identificar ameaças com base em fóruns de hackers, classificando notificações ou relatos como "bons" quando representam uma ameaça cibernética que pode ser vinculada a um CVE conhecido. Outras aplicações de PLN incluem classificação de sentimento com base nas palavras utilizadas em threads e postagens, como a identificação de linguagem agressiva (CAINES et al., 2018a), intenção e tipo de conteúdo (CAINES et al., 2018b), e a busca por conteúdos associados a ideologias específicas, como misoginia, antissemitismo e racismo (CHUA; WILSON, 2023).

3.3 EXPLORAÇÃO DE VULNERABILIDADES NO MUNDO REAL

A análise da exploração de vulnerabilidades no mundo real apresenta desafios significativos. Enquanto atacantes exploram ativamente vulnerabilidades para comprometer sistemas, frequentemente compartilhando técnicas e informações em fóruns online, pesquisadores precisam compreender esses exploits para construir estratégias de defesa eficazes (LIANG et al., 2018). Métodos manuais, como engenharia reversa e *fuzzing*, são particularmente úteis para descobrir novas vulnerabilidades e fraquezas, contribuindo para o fortalecimento das defesas contra ameaças, sejam elas identificadas em ambientes controlados ou expostas em fóruns blackhat (SUTTON; GREENE; AMINI, 2007).

A colaboração e o compartilhamento de informações desempenham papel fundamental no combate às vulnerabilidades. Plataformas e bases de dados, como o sistema Common Vulnerabilities and Exposures (CVE) e a National Vulnerability Database (NVD), fornecem informações padronizadas sobre vulnerabilidades, incluindo gravidade e correções disponíveis (CHEN et al., 2016). Além disso, práticas de divulgação coordenada, como *responsible disclosure* e programas de *bug bounty*, facilitam o reporte e a correção de vulnerabilidades (EDKRANTZ; TRUVÉ; SAID, 2015). Outras aplicações baseiam-se na identificação de atividades em Shodan (BADA; PETE, 2020) ou destacam atividades de hackers e vazamentos de dados (FANG et al., 2019). Além disso, é possível estudar as interações sociais entre usuários e seus comportamentos criminais e não criminais (MAN; SIU; HUTCHINGS, 2023; PASTRANA et al., 2018), trocas, transações, compras e vendas criminosas e não criminosas (BUTLER, 2020; ALLODI, 2017; SIU; COLLIER; HUTCHINGS, 2021), bem como analisar a maturidade de softwares compartilhados associ-

ados a CVE, CVSS e EPSS (SUN et al., 2021; MORENO-VERA, 2023; MORENO-VERA et al., 2023).

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo, revisamos de forma abrangente o corpo de trabalhos relacionados ao nosso estudo, com foco em três áreas principais: fóruns blackhat, processamento de linguagem natural (PLN) e inteligência de ameaças, e a exploração de vulnerabilidades no mundo real. O CrimeBB, conjunto de dados fornecido pelo Cambridge Cybercrime Centre (CCC), tem se destacado como a fonte mais utilizada para esse tipo de análise, oferecendo uma grande quantidade de dados sobre discussões cibercriminosas. Por meio desta revisão, identificamos diversas aplicações derivadas da análise textual, incluindo análise de sentimentos, detecção de linguagem agressiva, rastreamento de comércio e transações, identificação de atividades criminosas, detecção de racismo e análise detalhada de CVEs, entre outras.

Essa revisão orientou a definição do nosso próprio caminho de pesquisa, centrado na classificação e compreensão das discussões em threads e postagens de fóruns clandestinos. Diferentemente de estudos anteriores, que frequentemente enfatizam temas amplos de cibercrime, nossa abordagem concentra-se especificamente nas vulnerabilidades discutidas e nos riscos associados aos CVEs referenciados em comentários e interações entre usuários. Ao focar nas citações de CVEs, nossa pesquisa busca fornecer insights mais profundos sobre ameaças emergentes, enfatizando o risco potencial que cada vulnerabilidade representa sob a perspectiva de cibercriminosos. Esse foco inédito em uma análise mais granular de vulnerabilidades permite uma abordagem mais direcionada para compreender discussões em fóruns clandestinos, com implicações relevantes para o aprimoramento da inteligência de ameaças e das práticas de cibersegurança.

4 ANÁLISE EXPLORATÓRIA DO CONJUNTO DE DADOS *CRIMEBB*

Neste capítulo, apresentamos nossa metodologia e detalhamos as etapas envolvidas. Também apresentamos os resultados de nossa análise exploratória, incluindo estatísticas e insights obtidos a partir do conjunto de dados CrimeBB e da extensão PostCog.

4.1 DESCRIÇÃO DO CONJUNTO DE DADOS

Nesta seção, fornecemos uma visão abrangente dos conjuntos de dados utilizados neste estudo, detalhando suas fontes, características e relevância para os objetivos da pesquisa. Discutimos a estrutura de cada conjunto de dados, incluindo o tipo e o volume de dados, bem como as etapas de pré-processamento aplicadas para garantir consistência e adequação à nossa análise.

4.1.1 National Vulnerability Database (NVD)

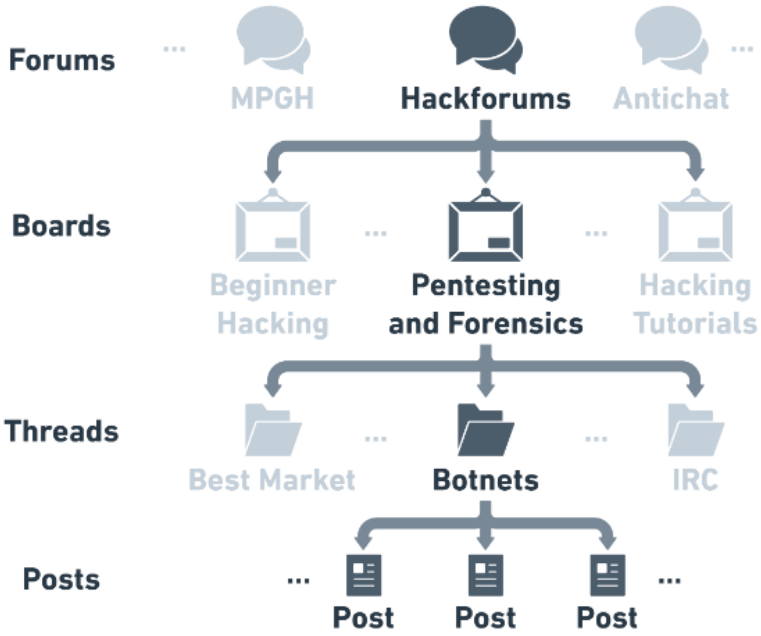
A National Vulnerability Database (NVD)[1] é um repositório abrangente de informações sobre vulnerabilidades de software e problemas de segurança. Ela é mantida pelo National Institute of Standards and Technology (NIST), nos Estados Unidos. O conjunto de dados da NVD fornece informações detalhadas sobre vulnerabilidades conhecidas em diversos produtos de software, incluindo sistemas operacionais, aplicações, bibliotecas e hardware.

4.1.2 CrimeBB e PostCog

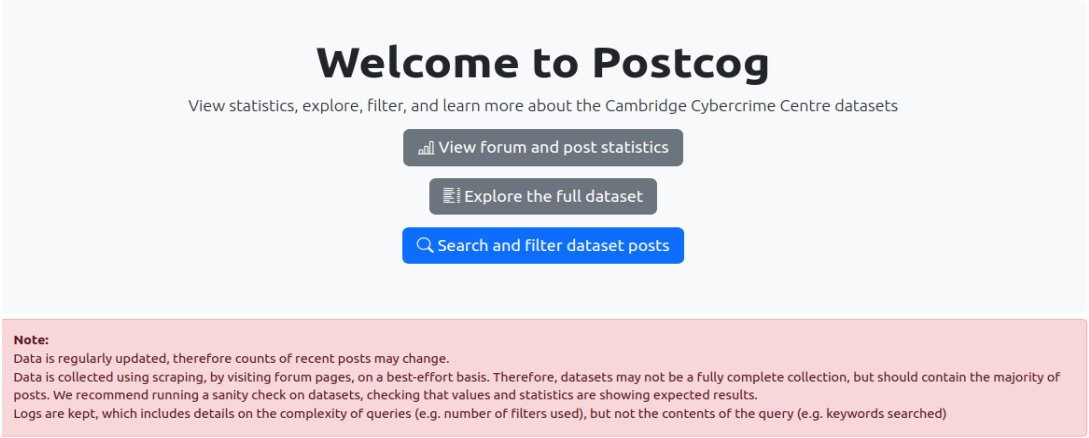
O Cambridge Cybercrime Centre (CCC) disponibiliza 38 fóruns clandestinos coletados no conjunto de dados CrimeBB (PASTRANA; THOMAS et al., 2018). O framework PostCog (PETE et al., 2022) estende o CrimeBB ao fornecer uma aplicação web para navegar pelos dados do CrimeBB, buscar palavras-chave e rotular postagens seguindo uma abordagem de *crowdsourcing*.

CrimeBB. A maioria das postagens em fóruns online consiste principalmente em texto simples, embora também possam incorporar elementos adicionais, como imagens, vídeos ou anexos. Por meio do CrimeBot, o CCC identifica e anota diversos elementos presentes nas postagens, incluindo imagens, vídeos (via *iframes*), trechos de código-fonte,

[1] https://nvd.nist.gov/



(a)



(b)

Figura 4 – (a) Conjunto de dados CrimeBB, mostrando a composição hierárquica de websites, boards, threads e posts. (b) Framework PostCog, por meio do qual podemos navegar pelos dados do CrimeBB.

links para sites externos, links internos para threads e anexos (PASTRANA; THOMAS et al., 2018).

O conjunto de dados CrimeBB é composto por dados hierárquicos baseados em websites, contendo cerca de 45 GB de informação textual. A estrutura dos fóruns clandestinos é composta por fóruns ou "websites", seguidos por boards relacionados a "tópicos", threads correspondentes a "perguntas" e posts correspondentes a "respostas" (ver Figura 4 (a)). Em 28 de agosto de 2024, o CrimeBB possuía 6.739.073 usuários interagindo em 37 websites, 10.600.580 threads de discussão e 117.365.492 postagens (ver Tabela 1).

PostCog. O PostCog (PETE et al., 2022) é uma aplicação web que fornece dados curados provenientes dos fóruns clandestinos do CrimeBB. A Figura 4 (b) mostra a página inicial do framework PostCog; temos acesso a esse conjunto de dados em função de um acordo e de um contrato de confidencialidade (NDA) firmado. O PostCog encontra-se em sua segunda fase de desenvolvimento. O protótipo inicial foi construído com NodeJS, ExpressJS e PostgreSQL. Posteriormente, com base em insights obtidos em testes com usuários, a pilha tecnológica foi expandida para integrar ReactJS e ElasticSearch. Recentemente, o PostCog passou a permitir que usuários pesquisem palavras por fórum, subfórum, data e etiquetas de PLN. Os resultados podem ser exportados em arquivos CSV para análise em aplicações externas.

4.2 ANÁLISE DE FÓRUNS CLANDESTINOS

Para produzir o conjunto de dados, consideramos as etapas descritas na Figura 1. Primeiramente, utilizamos o framework PostCog para pesquisar e baixar posts, threads e boards relevantes do CrimeBB. Também obtivemos a data de criação do post, o nome de usuário e três rótulos atribuídos pelo PostCog com base no conteúdo do post, como intenção, tipo de post e tipo de crime discutido. Dividimos esse processo em quatro subetapas: (i) processamento de threads; (ii) rotulação das classes-alvo; (iii) processamento de texto; e (iv) tokenização e extração de características.

4.2.1 Processamento de threads

Empregamos um processo de filtragem para identificar todos os posts que fazem referência a pelo menos um código Common Vulnerabilities and Exposures (CVE). Concatenamos e unimos todos os posts com sua thread pai correspondente. Em seguida, utilizando a expressão regular *case-insensitive* `cve-[0-9]{4}-[0-9]{4,}` (ligeiramente mais específica do que `cve(-id)?(?i)` utilizada em (ALLODI, 2017) e refinada em (MORENO-VERA et al., 2023)), a Figura 5 apresenta o processo de processamento de threads. Buscamos e filtramos posts que se referem a vulnerabilidades por meio de seus identificadores CVE, ignorando threads-posts que não citam nenhuma referência CVE. Em todos os fóruns do CrimeBB, encontramos cerca de 15.645 posts citando 2.218 CVEs únicos (11.420 citações de CVEs no total) distribuídos em 6.496 threads de discussão. Em seguida, concatenamos cada um desses posts contidos em uma thread juntamente com o título da thread.

4.2.2 Rotulação do alvo

Consideramos três tipos de rotulação-alvo: (i) rotulação do PostCog; (ii) rotulação manual por especialistas; e (iii) rotulação por ChatGPT.

Tabela 1 – Estatísticas dos fóruns do CrimeBB extraídas do PostCog. Os fóruns são divididos em boards, e os boards são divididos em threads. Cada thread contém um número de posts. Os dados foram baixados em 28 de agosto de 2024.

Fórum	**#Boards**	**#Threads**	**#Posts**	**Primeiro post**	**Post mais recente**
Hack Forums	212	4,301,893	42,686,891	2007-01-27	2024-05-24
Zismo	39	546,832	12,194,525	2010-05-26	2024-05-04
MPGH	770	918,439	12,193,797	2005-12-26	2024-05-26
Blackhatworld	112	1,017,226	12,132,290	2005-10-31	2024-05-20
Nulled	169	687,522	9,546,230	2013-04-02	2024-05-16
lolzteam	292	577,642	6,196,005	2013-03-10	2019-09-01
Cracked	163	419,517	3,911,032	2018-04-03	2024-04-17
OGUsers	58	244,766	3,608,306	1990-01-01	2019-04-09
UnKnoWnCheaTs	248	182,667	2,837,509	2002-11-02	2024-05-24
Antichat	80	254,810	2,642,161	2002-05-29	2024-03-15
V3rmillion	40	456,262	2,459,519	2016-02-02	2019-11-11
Raidforums	88	114,450	1,231,126	2015-03-20	2022-02-20
Elhacker	53	212,081	987,039	2002-08-21	2024-05-28
Probiv	168	123,023	909,007	2014-11-05	2024-04-25
Breached	72	34,412	737,922	2022-03-16	2023-03-19
Hackers Armies	53	42,548	468,880	2009-06-01	2024-04-01
Forum Team	201	44,404	433,901	2017-10-31	2024-03-26
BreachForums	76	28,800	331,357	2023-05-12	2024-05-14
Indetectables	72	32,274	328,539	2006-02-20	2024-05-19
XSS Forum	49	48,718	310,796	2004-11-13	2023-04-27
Dread	446	75,122	294,596	2018-02-15	2020-01-09
Runion	19	16,867	240,632	2012-01-11	2020-01-05
Offensive Community	71	119,251	161,492	2012-06-30	2018-12-11
Underc0de	73	27,054	95,723	2010-02-10	2024-05-26
The Hub	62	11,286	88,753	2014-01-09	2019-08-09
Ifud	65	11,827	72,851	2012-05-10	2022-12-19
PirateBay Forum	33	11,526	60,678	2013-10-23	2020-12-03
OnniForums	27	3,542	45,094	2023-02-08	2024-05-24
Torum	11	4,346	28,485	2017-05-25	2019-08-07
Safe Sky Hacks	50	12,963	27,018	2013-03-28	2019-01-23
Kernelmode	11	3,606	26,815	2010-03-11	2019-11-29
Freehacks	228	5,106	26,471	2013-07-27	2023-04-23
Deutschland im Deep Web	43	4,075	20,185	2018-11-22	2020-06-04
GreySec	28	2,232	11,925	2015-06-10	2022-01-04
Garage for Hackers	47	2,329	8,710	2010-07-06	2018-10-13
Stresser Forums	17	708	7,069	2017-04-09	2018-04-09
Envoy Forum	93	454	2,163	2019-07-06	2019-08-09
Total	4,339	10,600,580	117,365,492		

Rótulos do PostCog

Obtivemos três conjuntos de rótulos a partir do PostCog; esses rótulos estão disponíveis apenas para threads-posts do HackForums:

1. **Tipo de post**, que se refere ao conteúdo do post e é obtido de (CAINES et al., 2018b);

2. **Intenção**, que se refere à intenção do usuário (por exemplo, negativa, positiva ou

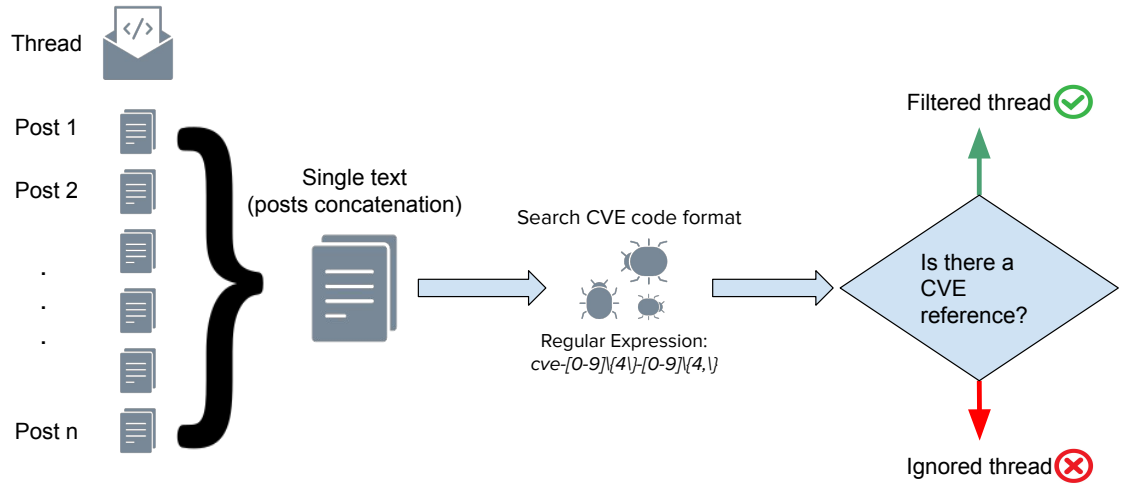


Figura 5 – Concatenação de posts por thread: se ao menos um post cita um código CVE, consideramos todos os demais posts da mesma thread como uma única amostra textual. Caso contrário, a thread completa é ignorada e excluída do conjunto de dados. Essa é a razão pela qual não utilizamos todas as threads rotuladas.

neutra) e é obtida de (CAINES et al., 2018b);

3. **Tipo de crime**, que se refere à atividade criminosa identificada na thread e é obtido de (SIU; COLLIER; HUTCHINGS, 2021).

Esses rótulos estão atualmente disponíveis apenas para o site HackForums; portanto, a maior parte da nossa análise concentra-se nele.

Anotações de especialistas no HackForum

Aproveitamos a rotulação manual realizada por especialistas e reportada em (MORENO-VERA et al., 2023).[2] Solicitamos a especialistas que rotulassem manualmente um subconjunto de 1.037 threads do conjunto de dados CrimeBB, utilizando o website HackForum. Os especialistas utilizaram o seguinte guia de codificação para rotular manualmente as threads:[3]

1. **Exploitation:** (i) menciona um grupo hacker conhecido; (ii) faz referência a criptomoedas e palavras-chave como bitcoin, exploitation e attack (no contexto de ataques no mundo real); (iii) discute abordagens para tornar exploits completamente indetectáveis; (iv) envolve mercados de exploits.

[2] Em (MORENO-VERA et al., 2023) já dispúnhamos de todos os rótulos de especialistas considerados neste trabalho, mas utilizamos apenas três deles (exploração, PoC e weaponization).

[3] Considerar gírias e abreviações típicas dessas comunidades é deixado como trabalho futuro.

2. **Proof-of-Concept (PoC):** (i) contém palavras-chave como PoC, tutorial ou guide (no contexto da produção de ferramentas em laboratório ou ambiente controlado); (ii) fornece um tutorial para construir um PoC; ou (iii) discute vulnerabilidades sem utilizar exploits no mundo real.

3. **Weaponization:** (i) contém palavras-chave como vulnerability e exploit (no contexto de weaponization); (ii) discute a disponibilidade de exploits totalmente funcionais ou altamente maduros, fornecendo referências ou código-fonte.

4. **Warning:** contém conselhos ou avisos sobre a detecção (por exemplo, por alguma empresa) da exploração de determinada vulnerabilidade.

5. **Help:** contém palavras-chave como help ou coding, referindo-se a usuários pedindo ajuda para explorar alguma vulnerabilidade ou executar algum código de propósito geral.

6. **Scam:** contém informações sobre transações, ou sobre a venda de um exploit ou vulnerabilidade que não funciona ou que já foi publicado.

7. **Others:** outras discussões não relacionadas a nenhuma das categorias acima.

Na Tabela 2, resumimos o número de amostras de cada classe para cada conjunto de rótulos analisado neste trabalho. Observe o desbalanceamento das amostras em cada conjunto de rótulos: “not criminal”, “infoRequest” e “neutral” são as classes mais representativas para tipo de crime, tipo de post e intenção, respectivamente.

Rotulação com ChatGPT

Solicitamos ao ChatGPT[4] que rotulasse nossas threads utilizando *prompts* dados o contexto e o conteúdo de cada thread. Pedimos que, para cada conjunto de rótulos, ele resumisse: (1) tipo de post, (2) tipo de crime, (3) intenção e (4) anotações de especialistas no HackForums.

Um de nossos objetivos foi reduzir o número de classes e obter um conjunto representativo de até quatro rótulos para cada uma das quatro dimensões acima. Utilizamos o ChatGPT 3.5-turbo e realizamos *few-shot learning* (AHMED; DEVANBU, 2022), fornecendo cinco exemplos de entrada e saída para cada classe de cada conjunto de rótulos (tipo de crime, tipo de post, intenção e anotações de especialistas). A seguir apresentamos um exemplo do template de prompt utilizado para a rotulação do tipo de crime:

```
I have a list of possible labels {not criminal, bots/malware, ...} related to cyber criminal activites.

I want to perform two tasks: the first one is to group the list of possible labels into a smaller list of lab
```

[4] OpenAI ChatGPT: https://www.openai.com/chatgpt

Rótulos de tipo de crime	
Rótulos	Amostras
Not criminal	2,307
Bots/Malware	604
Sql Injection	208
Credentials	41
VPN/proxy	34
DDoS/booting	12
Spam/marketing	7
CurrencyXchange	4
Identity fraud	2
eWhoring	1

(a)

Rótulos de intenção	
Rótulos	Amostras
Neutral	2,184
Other	494
Positive	197
Gratitude	170
Aggression	53
Negative	37
PrivateMessage	30
Moderate	28
Vouch	27

(b)

Rótulos de tipo de post	
Rótulos	Amostras
InfoRequest	912
Comment	909
Other	494
OfferX	490
Exchange	137
RequestX	137
Tutorial	76
Social	65

(c)

Rótulos das anotações de especialistas	
Rótulos	Amostras
Weaponization	397
PoC	242
Others	190
Exploitation	102
Warning	55
Help	41
Scam	10

(d)

Tabela 2 – Apresentamos as 3.220 threads e os conjuntos de rótulos obtidos a partir do framework PostCog: (a) os rótulos de tipo de crime correspondem aos tipos de cibercrime discutidos em um post; (b) os rótulos de intenção referem-se à intenção expressa em um post; e (c) os rótulos de tipo de post correspondem ao conteúdo do post. Além disso, em (d), apresentamos as 1.037 threads do HackForum anotadas por especialistas.

```
The second task is to set a new label using the new smaller list of labels given an input raw {text} and thei

Based on the following samples { (text 1, label 1, new label 1), \dots, (text 5, label 5, new label 5)}, I wo

Please return in a list of tuples: [ ( "input", {text},   "new label", {new label}), ...]
```

Consequentemente, executamos esse prompt para obter um novo agrupamento de rótulos de cada conjunto de rótulos e os novos rótulos para cada amostra. Na Tabela 3, apresentamos o novo conjunto de rótulos, classes e o número de amostras atribuídas pelo prompt do ChatGPT. Em todos os casos, reduzimos significativamente o número de classes, preservando, no entanto, um elevado desbalanceamento nos rótulos de tipo de crime e intenção. A partir disso, o ChatGPT aprende a classificar as threads utilizando os seguintes critérios:

1. **Tipo de crime:** recategorizado em “not criminal”, definido como atividades que não são consideradas criminosas; “cybercrime activities”, definido como atividades ilegais relacionadas ao cibercrime; e “cybercrime support services”, definido como serviços que facilitam atividades de cibercrime.

Rótulos de tipo de crime	
Rótulos	Amostras
Not criminal	2,307
Cybercrime activities	875
Cybercrime support services	38

(a)

Rótulos de intenção	
Rótulos	Amostras
Sentiment	2,418
Other	494
Expression	280
Intensity	28

(b)

Rótulos de tipo de post	
Rótulos	Amostras
Communication/Interaction	1050
Requests	1049
Other	494
Offer/exchanges	627

(c)

Rótulos de especialistas	
Rótulos	Amostras
Malicious activity	509
Informal	297
Others	190
Support and assistance	41

(d)

Tabela 3 – Apresentamos os novos rótulos atribuídos pelo ChatGPT: (a) os rótulos de tipo de crime foram agrupados em três novos rótulos; (b) os rótulos de intenção foram agrupados em quatro novos rótulos; (c) os rótulos de tipo de post foram agrupados em quatro novos rótulos; e (d) as anotações de especialistas no HackForums foram agrupadas em quatro novos rótulos.

2. **Tipo de post:** recategorizado em "communication/interaction", definido como formas de interação ou tipo de conteúdo; "requests", definido como busca por informação ou serviços; "offer/exchanges", definido como oferta de serviços ou trocas; e "others".

3. **Intenção:** recategorizada em "sentiment", definida como emoções ou atitudes; "expression", definida como formas de comunicação ou expressão; "intensity", definida como níveis de intensidade ou veemência; e "others".

4. **Anotações de especialistas no HackForum:** recategorizadas em "malicious activity", definida como conteúdo relacionado a weaponization, exploitation ou scam; "informational", definida como conteúdo que inclui PoC, alertas, tutoriais ou avisos; "support and assistance", definida como conteúdo pedindo ajuda; e "others".

4.2.3 Extração de características textuais

Dividimos nosso pré-processamento em três etapas: (i) pré-processamento de PLN; (ii) avaliação de linguagem; e (iii) extração de características.

Pré-processamento de PLN

Implementamos um pré-processador textual que auxilia na identificação e preservação de palavras relevantes a serem consideradas. Desenvolvemos uma biblioteca para pré-processar textos e avaliar a linguagem, de modo a facilitar a preparação do conjunto de dados. Devemos ter cuidado ao selecionar quais palavras devem ser filtradas. Para isso, filtramos os seguintes elementos:

- **Stopwords:** são palavras de qualquer idioma que não agregam muito significado a uma frase. Alguns exemplos são pronomes, advérbios, artigos etc.
- **Pontuações:** são símbolos adicionados ao texto para indicar divisões entre suas diferentes partes, como pontos, vírgulas, ponto e vírgula, pontos de interrogação, apóstrofos e parênteses.
- **Caracteres Especiais:** são símbolos utilizados na escrita, digitação etc., que representam algo diferente de uma letra (fora das 26 letras usadas no inglês dos EUA) ou número, como §, à, é, î, œ, ü, ñ etc.
- **Emojis:** são uma forma de linguagem pictórica utilizada para expressar uma ideia, também chamada de "imagens digitais". Esses símbolos denotam uma emoção ou uma ação. Pode ser uma imagem ou um símbolo textual composto, como ":)", ":C", ":-)"etc.

Após filtrar esses elementos da entrada bruta, prosseguimos para identificar a língua à qual o texto de entrada pertence. Subsequentemente, optamos por converter todas as palavras utilizando lematização em vez de stemming. Essa escolha se justifica porque, em classificação de texto, a lematização facilita a criação de representações de características de alta qualidade, contextualmente precisas e semanticamente significativas. Isso, por sua vez, contribui para melhorar a acurácia da classificação, reduzir ruído e aprimorar a generalização em diferentes tipos de dados textuais.

Avaliação de linguagem

Definimos a **Indicator Language Function** (ILF), denotada por $\mathbb{I}_{ilf}$. Na Equação 4.1, definimos nossa ILF, tomando dois parâmetros: a palavra w e a língua L a ser avaliada:

$$\mathbb{I}_{ILF}(w, L) = \begin{cases} 1, & \text{se } w \in L \\ 0, & \text{caso contrário} \end{cases} \tag{4.1}$$

Além disso, definimos a **Language Ratio Function** (LRF) para um conjunto de palavras e línguas. Essa função nos permite identificar e calcular qual porcentagem de palavras em uma frase, parágrafo ou texto, em geral, pertence a uma língua específica L. Na Equação 4.2, definimos nossa LRF, tomando como parâmetros o texto t, com n palavras, e a língua L a ser avaliada:

$$\text{Ratio}_{LRF}(t, L) = \frac{1}{\text{n}} \times \sum_{\substack{i=1 \\ \text{w}_i \in t}}^{n} \mathbb{I}_{ILF}(w_i, L) \tag{4.2}$$

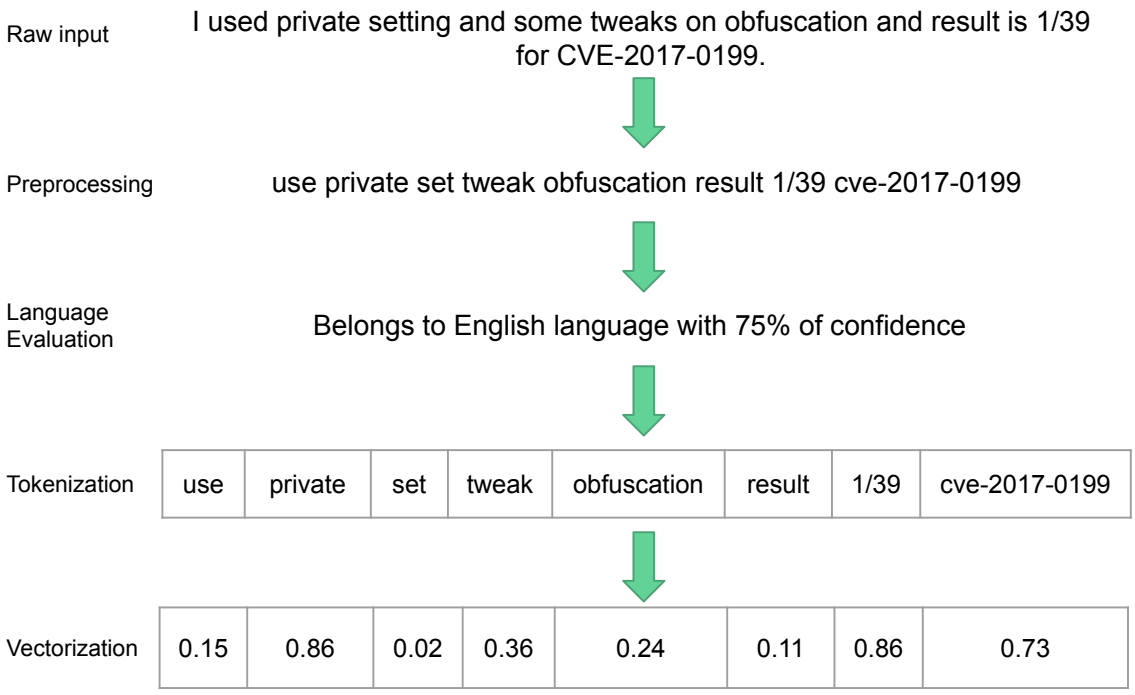


Figura 6 – Pipeline de pré-processamento textual, mostrando todas as etapas desde a entrada bruta até a vetorização. Observe que apenas mantemos e lematizamos palavras para sua forma raiz básica, mas não todas as palavras. Este post foi extraído do website HackForums.

Por fim, a **Determine Language Function** (DLF) é definida utilizando o fator de reconhecimento de linguagem (LRF) calculado anteriormente. Essa função determina a língua mais provável à qual o texto de entrada pertence. Na Equação 4.3, definimos nossa DLF, tomando um texto t como parâmetro. Esse texto é avaliado em um conjunto de línguas $\mathbb{L}$ (inglês e russo por padrão):

$$\text{Language}_{DLF}(t) = \max_{\forall L \in \mathbb{L}} \text{Ratio}_{LRF}(t, L) \tag{4.3}$$

Essa etapa nos ajuda a identificar a língua principal em uma thread (focamos apenas na língua inglesa; caso contrário, filtramos a thread). Após o pré-processamento das concatenações thread-post para identificar línguas, identificamos 3.220 thread-posts em inglês, 1.977 thread-posts em russo, 1.223 thread-posts em espanhol, 21 thread-posts em português e 4 thread-posts em alemão. Após essa etapa, prosseguimos com o embedding textual, convertendo texto em vetores.

Tokenização e Extração de Características

Nesta etapa, realizamos a extração de características a partir de informações textuais. Após o pré-processamento de todas as informações textuais e a tokenização das

palavras, utilizamos métodos de codificação baseados em texto, como Bag-Of-Words (BoW) (HARRIS, 1954), Term Frequency - Inverse Document Frequency (TF-IDF)(SALTON; BUCKLEY, 1988) e Doc2Vec (MIKOLOV et al., 2013). A Figura 6 mostra todo o pipeline de processamento textual seguido para a extração de características.

4.3 ACHADOS EMPÍRICOS

Nesta seção, reportamos achados empíricos dos fóruns CrimeBB, incluindo atividade dos usuários, riscos, preços de exploits, atrasos, e escores CVSS e EPSS provenientes da NVD.

4.3.1 **Atividade dos usuários**

A Figura 7 (a) mostra como a atividade variou entre os usuários. Em particular, Trillium é o usuário mais ativo, normalmente anunciando novidades sobre a ferramenta Trillium Security MultiSploit Tool e artefatos FUD. Trillium citou mais de 175 CVEs. Os demais usuários que aparecem em threads citando CVEs abrangem um conjunto menos diverso de CVEs. Em particular, é desafiador relacionar usuários reais a contas virtuais, uma vez que usuários reais podem criar múltiplas contas virtuais, por exemplo, para aumentar privacidade ou poder de voto. De fato, os fóruns do CrimeBB não utilizam reputação para fins de controle de admissão. Isso contrasta fortemente com o mercado russo, que exige admissão explícita por administradores do mercado, condicionada ao fato de o usuário ser ativo em comunidades relacionadas, incentivando usuários a manter suas contas ativas por períodos mais longos (ALLODI, 2017). Vale notar que, neste trabalho, focamos exclusivamente em threads referentes a vulnerabilidades. Para uma análise mais ampla sobre a reputação de usuários nos fóruns do CrimeBB, remetemos o leitor a (PARACHA; ARSHAD; KHAN, 2023).

4.3.2 **Riscos**

A Figura 7 (b) mostra os CVEs mais citados no Hackforums. Ela indica que o número de citações cai fortemente após as 10 vulnerabilidades mais citadas. Entre as principais vulnerabilidades, temos CVE-2017-0199, CVE-2017-8570, CVE-2017-8759 e CVE-2017-11882. Para essas quatro vulnerabilidades, encontramos exploits no GitHub, conforme citado pela NVD. Essas quatro vulnerabilidades, juntamente com CVE-2010-3333 e CVE-2012-0158, estão relacionadas a produtos Microsoft. Algumas vulnerabilidades citadas no CrimeBB também são mencionadas em outras fontes relevantes. Esse é o caso, por exemplo, da CVE-2012-0158, conhecida como Microsoft MSCOMCTL.OCX Remote Code Execution Vulnerability, e da CVE-2015-1641, que constam no catálogo Known

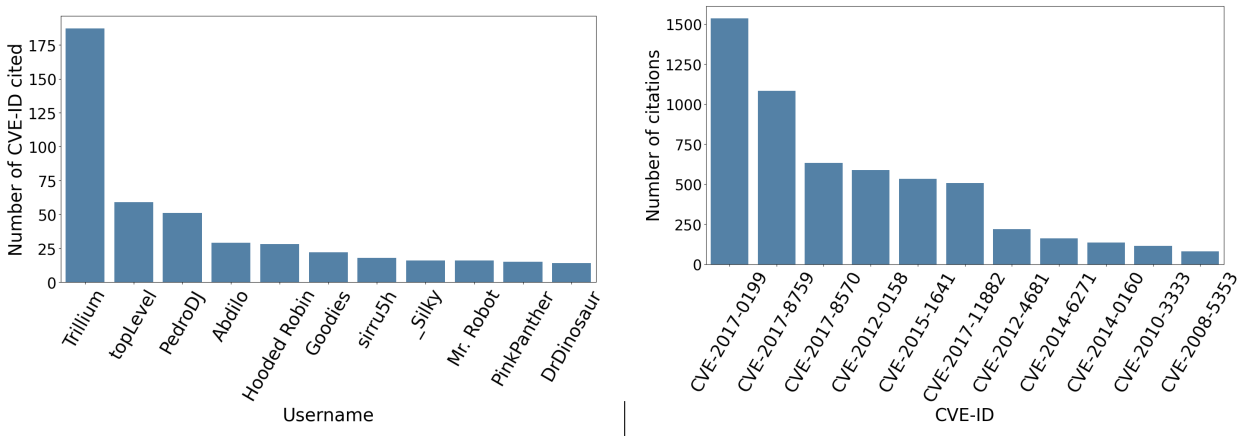


Figura 7 – Estatísticas: (a) atividade dos usuários: a maioria dos usuários que cita CVEs cita menos de 50 CVEs únicos. Uma exceção notável é Trillium. (b) Popularidade dos CVEs no Hackforums.

Exploited Vulnerabilities (KEV) da CISA.[5] Para outras vulnerabilidades, como CVE-2008-5353, CVE-2012-4681, CVE-2010-0094 e CVE-2011-3544, há exploits disponíveis no ExploitDB ou no toolkit Metasploit[6]. Essas vulnerabilidades estão relacionadas ao Java Runtime Environment. Outras vulnerabilidades notáveis incluem CVE-2014-0160 e CVE-2014-6271, correspondentes ao Heartbleed e ao Shellshock.

4.3.3 Preços

A seguir, examinamos os preços dos artefatos referenciados por usuários nos fóruns do CrimeBB. A Figura 8 (a) apresenta a CDF dos preços em dólares e, para fins comparativos, também apresentamos a CDF dos preços de exploits reportados no mercado russo analisado em (ALLODI, 2017). Os valores mínimo, mediano e máximo nos fóruns CrimeBB foram 1 USD, 100 USD e 4.400 USD, respectivamente, enquanto os valores correspondentes no mercado russo foram 100, 2.000 e 8.000 USD, respectivamente. Além disso, observamos que mais de 80% das referências a ferramentas de hacking correspondiam a preços inferiores a 1.000 USD. Os preços mais altos no mercado russo em comparação aos fóruns do CrimeBB podem ser atribuídos ao fato de que o mercado russo requer admissão explícita por administradores. Como resultado, os usuários nesse mercado tendem a discutir artefatos mais maduros, que consequentemente são mais caros.

Por outro lado, nos fóruns do CrimeBB, observou-se que usuários frequentemente propõem o reempacotamento de exploits existentes, por exemplo em novas versões FUD, ou oferecem assinaturas de websites que tendem a ser mais baratas do que exploits (VALEROS; GARCIA, 2020). Apesar das diferenças de preços, algumas semelhanças também foram observadas. Os preços máximos não excederam 8.000 USD em ambas as plataformas, a

[5] Ver também https://www.cisa.gov/binding-operational-directive-22-01
[6] https://www.exploit-db.com/exploits/16293

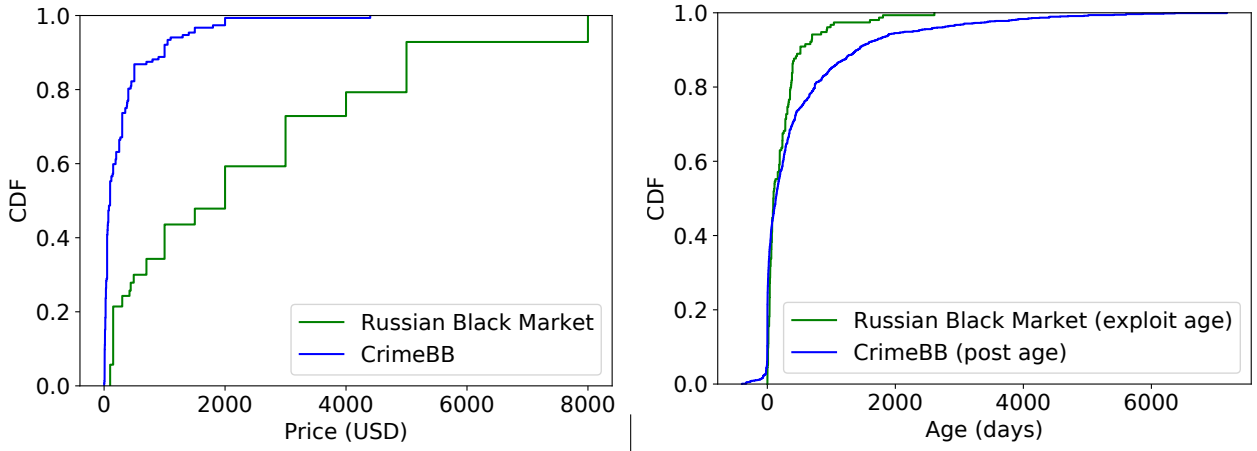


Figura 8 – Mercado Russo vs CrimeBB: (a) CDF dos preços de ferramentas de hacking: os preços no CrimeBB são relativamente baixos em comparação ao mercado russo – alguns preços correspondem a assinaturas, e outros a reempacotamento e FUD. Estatísticas de preços: CrimeBB (Mín: 1, Mediana: 100, Máx: 4400), mercado russo (Mín: 100, Mediana: 2000, Máx: 8000). (b) CDF da diferença, em dias, entre a citação no CrimeBB e a data de publicação na NVD. Valores negativos correspondem a citações de CVEs ocorridas antes da publicação da vulnerabilidade correspondente na NVD. Estatísticas de idade: CrimeBB (Mín: -396, Mediana: 132, Máx: 7181), mercado russo (Mín: 1, Mediana: 95,5, Máx: 2610)

maioria dos preços ficou abaixo de 2.000 USD e aproximadamente 20% dos preços estavam próximos de 100 USD. Esses números sugerem que participar desses fóruns pode ser financeiramente vantajoso, com recompensas alinhadas à maioria dos programas de *bug bounty*, que oferecem até 3.000 USD por uma falha crítica. No entanto, esses valores ainda estão longe das recompensas milionárias já relatadas na literatura.[7]

4.3.4 Atrasos

Em seguida, consideramos os atrasos entre a publicação das vulnerabilidades e o aparecimento de posts nos fóruns. Para os fóruns do CrimeBB, calculamos a idade do post como a diferença entre o dia do post e o dia em que a vulnerabilidade correspondente foi publicada na NVD, `PostAge = PostPubDate - CVEPubDate`. De forma semelhante, a idade do exploit reportada por (ALLODI, 2017) é a diferença entre o dia em que um exploit foi publicado no mercado russo e o dia em que a vulnerabilidade correspondente foi publicada na NVD, `ExplAge = ExplPubDate - CVEPubDate`. A Figura 8 (b) mostra a CDF de `PostAge` e `ExplAge`, para os fóruns do CrimeBB e o mercado russo, respectivamente.

Observe que mais de 50% dos exploits discutidos no CrimeBB dizem respeito a vulnerabilidades que haviam sido divulgadas nos 69 dias anteriores à sua citação no CrimeBB. Considerando que 50% dos sistemas de controle industrial (ICS) não são

[7] https://portswigger.net/daily-swig/million-dollar-bug-bounties-the-rise-of-record-breaking-payouts

corrigidos 60 dias após a divulgação da vulnerabilidade(WANG et al., 2017), o uso de fóruns blackhat é fundamental para estimar os riscos associados às vulnerabilidades. Entre as semelhanças entre os fóruns do CrimeBB e o mercado negro russo, observamos que aproximadamente 60% da atividade ocorre muito próximo à data de publicação do CVE. As discussões tendem a desaparecer para praticamente todas as vulnerabilidades seis anos após seu lançamento.

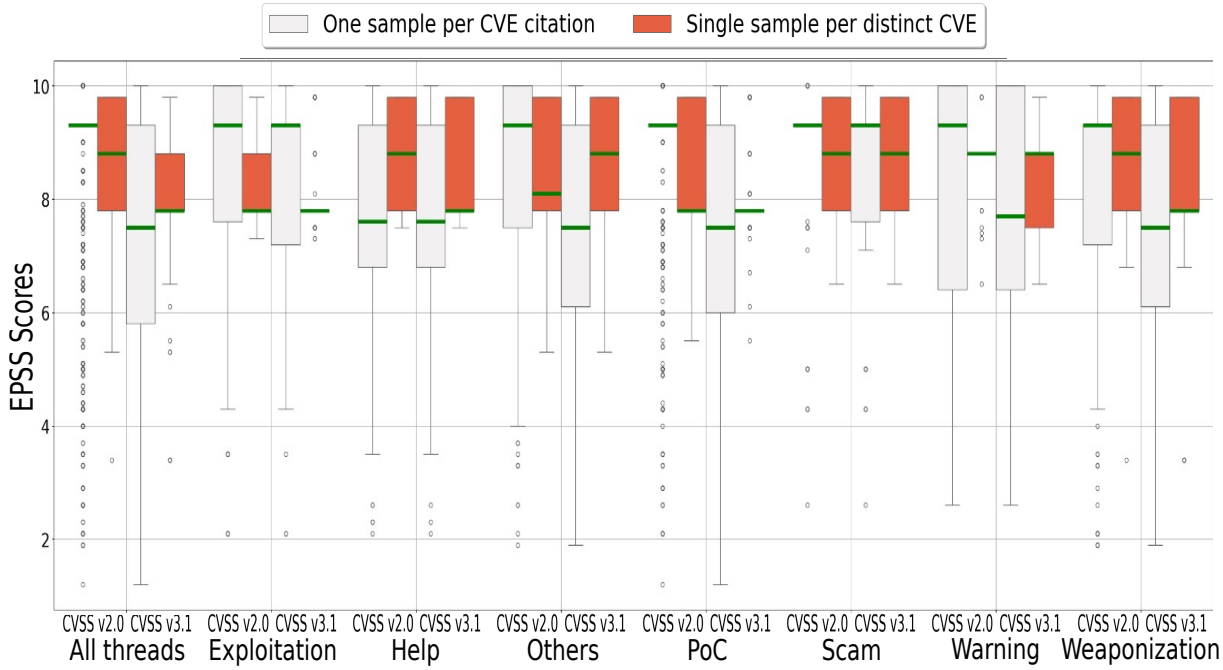
Observamos que, no mercado russo, temos apenas valores positivos de idade, enquanto no CrimeBB há uma pequena fração de valores negativos. Isso se explica pela natureza distinta dos dois fóruns, como discutido na seção anterior: enquanto o CrimeBB também contém mensagens perguntando sobre vulnerabilidades e discutindo estratégias para produzir armas *proof-of-concept*, o mercado russo contém majoritariamente discussões sobre exploits maduros a serem vendidos por valores mais altos, normalmente sendo disponibilizados apenas após a publicação do CVE na NVD. Exploits totalmente funcionais ou de alta maturidade raramente são produzidos antes da data de publicação da vulnerabilidade, isto é, `ExplPubDate` é maior que `CVEPubDate`. Discussões sobre vulnerabilidades, porém, podem começar antes de sua publicação, uma vez que identificadores CVE são anunciados ao público antes de serem publicados na NVD.

Dados da NVD

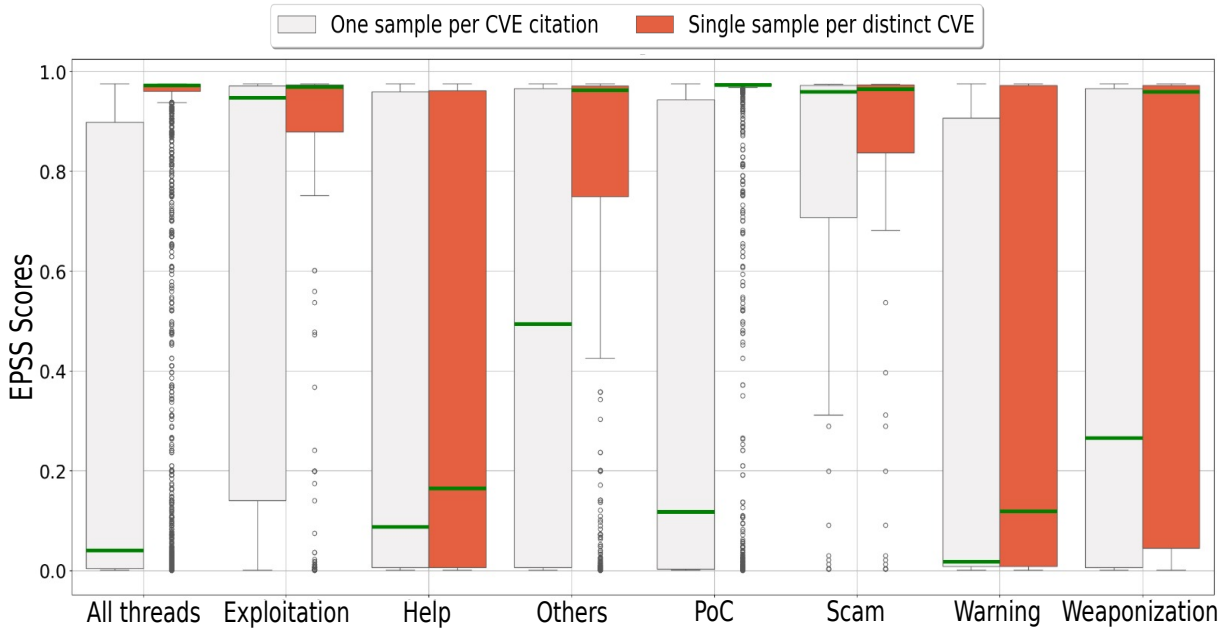
Utilizamos dados da NVD para determinar propriedades das vulnerabilidades consideradas, como o nível de severidade (CVSS[8]). O CVSS varia entre 0 e 10, e valores acima de 8 correspondem a alto risco. Em particular, a NVD fornece uma breve descrição de cada vulnerabilidade, juntamente com sua data de publicação, produtos afetados e recursos externos. Observe que o CVSS considera um subescore temporal para capturar o nível de maturidade dos exploits. Esse nível varia entre PoC, totalmente funcional e alta maturidade. Em particular, o CVSS não considera exploração no mundo real. O Expected Exploitability, em contraste, foca exclusivamente em exploits totalmente funcionais e de alta maturidade, referidos, neste trabalho, como weaponization, e o EPSS[9] tem como alvo a exploração. Neste trabalho, generalizamos o escopo de CVSS e EPSS considerando os três níveis de maturidade em uma estrutura unificada.

Considerando todas as vulnerabilidades, a Figura 9 mostra a distribuição dos escores CVSS e EPSS condicionados à classe da thread e ao conjunto total de threads. Os boxplots cinza consideram uma amostra por citação de CVE (com repetições), e os boxplots vermelhos consideram uma única amostra por CVE distinto (sem repetições). A Figura 9 (a) apresenta a distribuição dos escores CVSS, para as versões 2.0 e 3.1 do CVSS (esta última disponível para um subconjunto de CVEs). Em todas as categorias, os

[8] https://nvd.nist.gov/vuln-metrics/cvss
[9] https://www.first.org/epss/



(a) CVSS



(b) EPSS

Figura 9 – Distribuição dos escores CVSS e EPSS em diferentes classes anotadas por especialistas. Observe que 91% dos posts se referem a CVEs cujo escore CVSS é superior à média de CVSS entre todos os CVEs da NVD (não mostrada na figura).

valores medianos de CVSS excedem 7,0. Considerando o CVSS 2.0, a figura mostra que exploitation geralmente corresponde a valores de CVSS mais altos em comparação com PoC e weaponization.

Além disso, ainda considerando o CVSS 2.0, a figura revela que os CVEs com maiores valores de CVSS e severidade são os mais frequentemente citados (cerca de 60,8%

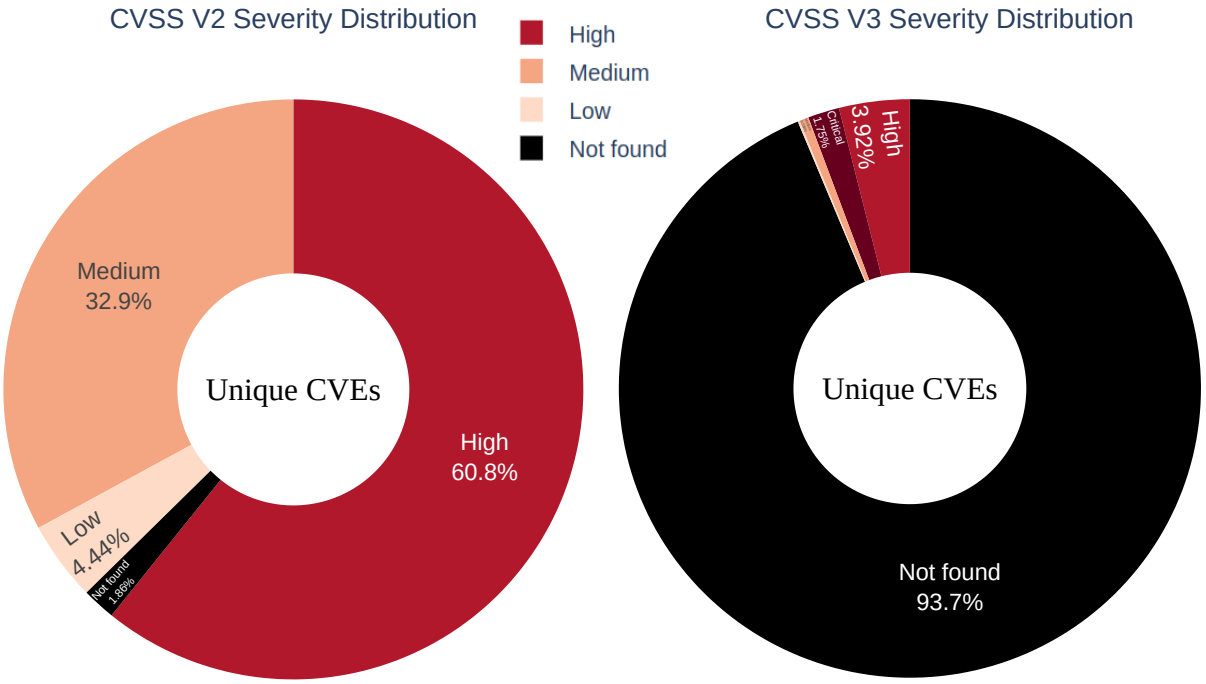


Figura 10 – Nível de severidade do Common Vulnerability Scoring System (CVSS); comparamos as versões 2 e 3.1 dos escores CVSS. Observamos que cerca de 93,7% dos escores CVSS v3.1 não estão disponíveis.

dos CVEs são de alto risco). Isso é evidenciado pelo fato de que as medianas dos boxplots vermelhos estão todas acima de 9,0, indicando que os escores de risco são amplificados por citações repetidas (ver Figura 10). Uma tendência semelhante é observada na Figura 9 (b) para o EPSS. O EPSS foi lançado em 2021, e utilizamos sua versão mais recente em 28 de agosto de 2024, que representa a probabilidade de exploração no mundo real nos próximos 30 dias. Apesar da defasagem entre as citações em posts e a disponibilização dos escores EPSS, observamos que os escores EPSS conseguem capturar riscos elevados para as vulnerabilidades presentes em nosso conjunto de dados.

4.3.5 Nuvens de palavras do conteúdo

A Figura 11 mostra as nuvens de palavras obtidas a partir dos posts do CrimeBB agrupados por classes e considerando todos os posts. A presença da palavra "multisploit" na categoria PoC indica que a ferramenta Trillium Security MultiSploit Tool é uma ferramenta popular utilizada por hackers. Além disso, observamos como a linguagem utilizada por hackers nas threads é fundamental para distinguir a exploração no mundo real do restante dos posts. A presença da palavra "thanks" na nuvem de exploitation refere-se ao comportamento de usuários que confirmam que uma determinada exploração funcionou.

Note que o termo "clean" também é prevalente em posts sobre exploração no mundo real. De fato, ele aparece no contexto de exploits FUD. Usuários geralmente indicam que

(a) Todos (b) Exploitation

(c) Help (d) Others

(e) PoC (f) Scam

(g) Warning (h) Weaponization

Figura 11 – Nuvens de palavras mostrando as palavras-chave mais frequentes que aparecem nos posts agrupados pelas anotações de especialistas: (a) todos os posts; (b) PoC; (c) weaponization; e (d) exploitation.

conseguiram executar o exploit e que ele não foi detectado por nenhum antivírus. Como exemplo, em um único post encontramos 35 ocorrências da palavra "clean" ao se referir a um exploit "silent" para a CVE-2011-3544. Os autores do exploit compartilham seu hash MD5 e reportam uma taxa de detecção de 0/35. Para fornecer evidência da não detecção, os autores apresentam uma saída de exemplo: `AVG Free: Clean; ArcaVir: Clean; Avast 5: Clean`. A lista continua com um total de 35 antivírus.

4.4 CONSIDERAÇÕES FINAIS

Neste capítulo, descrevemos os frameworks CrimeBB e PostCog utilizados para baixar e obter nossos dados. Também descrevemos o processo de preparação do conjunto de dados, iniciando com pré-processamento textual, corpus textual e técnicas de rotulação. Solicitamos a especialistas do domínio que anotassem cerca de 1.037 posts do fórum clandestino Hackforums. Além disso, utilizamos o modelo ChatGPT para classificar threads com base em seu conteúdo e re-rotular tanto as anotações de especialistas quanto os rótulos do PostCog. Também descrevemos nossos achados utilizando o conjunto de dados da NVD para obter CVSS e EPSS. Por fim, analisamos e comparamos preços mencionados nas threads de discussão, atrasos em respostas ou negociações, riscos e atividade dos usuários.

5 DISTINGUINDO AMEAÇAS POTENCIAIS DE AMEAÇAS IMINENTES

Neste capítulo, apresentamos a discussão e os resultados obtidos em nossos experimentos.

5.1 PREPARAÇÃO DO AMBIENTE

Em nosso estudo, comparamos três técnicas de codificação e verificamos que o BoW é mais interpretável, enquanto TF-IDF e doc2vec produzem maior acurácia.

5.1.1 **Extração de características**

Para BoW e TF-IDF, consideramos os quatro parâmetros a seguir: 1) as 30.000 palavras mais frequentes, 2) que apareçam pelo menos 5 vezes, e 3) em pelo menos 90% dos posts do corpus sejam consideradas para análise. Além disso, consideramos 4) *n*-gramas, variando de uma a três palavras. No geral, esses parâmetros definem os critérios para selecionar quais palavras ou grupos de palavras são incluídos na análise e com que frequência devem aparecer no corpus. Para o Doc2Vec, codificamos os posts em vetores de 5000 dimensões. Utilizamos técnicas padrão de pré-processamento em PLN, por exemplo, filtrando *stopwords* em inglês e pontuação dos posts.

5.1.2 **Configuração dos modelos e métricas**

Realizamos tarefas de classificação utilizando modelos lineares, como Ridge Classifier (TIKHONOV, 1943), LASSO (TIBSHIRANI, 1996), regressão logística (HOSMER; LEMESHOW; STURDIVANT, 2005), SVC linear (CORTES; VAPNIK, 1995) e SVM (BOSER; GUYON; VAPNIK, 1992). Também adicionamos modelos em ensemble, como Decision Tree (SPEYBROECK, 2012) e Random Forests (BREIMAN, 2001). Utilizamos *Accuracy*, *Precision*, *Recall* e *F1 score* para comparar o desempenho dos modelos, mas nos baseamos principalmente nos valores reportados pelo *F1 score*. Essas métricas são calculadas da seguinte forma:

Tabela 4 – Número de posts (threads) citando CVEs nos 10 principais boards do Hackforums, ordenados pelo número de posts rotulados

Board	Número de posts (threads) citando CVEs							
	Posts rotulados como							
	PoC		Weapon		Exploit		Todos os posts	
Pentesting and Forensics	271	(55)	210	(57)	11	(3)	557	(166)
Premium Tools and Programs	198	(1)	28	(3)	142	(4)	433	(20)
Website and Forum Hacking	93	(34)	139	(43)	16	(12)	333	(132)
Hacking Tools and Programs	10	(7)	57	(28)	174	(7)	260	(59)
Premium Sellers Section	–	–	81	(28)	89	(26)	210	(66)
Beginner Hacking	86	(43)	58	(47)	6	(6)	219	(143)
Botnets, IRC, and Zombies	24	(4)	85	(34)	22	(5)	160	(62)
Hacking Tutorials	58	(21)	8	(4)	3	(3)	74	(33)
Secondary Sellers Market	8	(4)	33	(21)	–	–	91	(40)
News and Happenings	9	(9)	11	(5)	1	(1)	75	(54)
Total, todos os boards	757	(244)	710	(397)	464	(102)	3,037	(1,162)

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{5.1}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{5.2}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{5.3}$$

$$F1_{score} = 2\frac{Precision * Recall}{Precision + Recall} \tag{5.4}$$

Onde T_P significa *True Positive*, T_N significa *True Negative*, F_P significa *False Positive* e F_N significa *False Negative*. Além disso, começamos com um conjunto de dados desbalanceado em todas as categorias (ver Tabela 2 e Tabela 3). Para lidar com esse desbalanceamento, utilizamos a heurística de *Random Over-Sampling* (LEMAîTRE; NOGUEIRA; ARIDAS, 2017) para produzir um conjunto de dados balanceado. Dividimos o conjunto de dados em 75% para treino+validação e 25% para teste. Para cada experimento, realizamos então uma busca em grade com validação cruzada StratifiedKFold com cinco partições para encontrar os melhores hiperparâmetros, tais como parâmetro de regularização, taxa de aprendizado, profundidade da árvore, número de características consideradas em cada divisão da árvore, número mínimo de amostras exigidas para dividir um nó interno e grau máximo do nó. Todos os experimentos foram treinados em uma GPU NVIDIA GeForce GTX 1070 com 8 GB de VRAM e processador Intel Core i5 com 6 GB de RAM.

Tabela 5 – Desempenho de Decision Tree (DT) e Random Forest (RF).

	Codificação textual	Classes-alvo	Accuracy	Precision	Recall	F1
DT	BoW	PoC, Weaponization, Exploitation	0.71	0.71	0.72	0.70
DT	TF-IDF	PoC, Weaponization, Exploitation	0.73	0.73	0.74	0.72
DT	doc2vec	PoC, Weaponization, Exploitation	0.74	0.74	0.74	0.73
DT	BoW	Exploitation vs Non-exploitation	0.85	0.86	0.85	0.85
DT	TF-IDF	Exploitation vs Non-exploitation	0.91	0.91	0.91	0.91
DT	doc2vec	Exploitation vs Non-exploitation	**0.92**	**0.93**	**0.92**	**0.92**
DT	BoW	PoC vs Non-PoC	0.75	0.75	0.75	0.75
DT	TF-IDF	PoC vs Non-PoC	0.77	0.78	0.77	0.77
DT	doc2vec	PoC vs Non-PoC	0.70	0.71	0.70	0.70
DT	BoW	Weaponization vs Non-weapon.	0.68	0.68	0.68	0.68
DT	TF-IDF	Weaponization vs Non-weapon.	0.63	0.64	0.63	0.62
DT	doc2vec	Weaponization vs Non-weapon.	0.59	0.59	0.59	0.59
RF	BoW	PoC, Weaponization, Exploitation	0.85	0.84	0.85	0.84
RF	TF-IDF	PoC, Weaponization, Exploitation	0.86	0.87	0.86	0.86
RF	doc2vec	PoC, Weaponization, Exploitation	0.86	0.90	0.86	0.86
RF	BoW	Exploitation vs Non-exploitation	0.98	0.98	0.98	0.98
RF	TF-IDF	Exploitation vs Non-exploitation	0.98	0.98	0.98	0.98
RF	doc2vec	Exploitation vs Non-exploitation	**0.99**	**0.99**	**0.99**	**0.99**
RF	BoW	PoC vs Non-PoC	0.84	0.84	0.84	0.84
RF	TF-IDF	PoC vs Non-PoC	0.87	0.87	0.87	0.87
RF	doc2vec	PoC vs Non-PoC	0.88	0.90	0.88	0.87
RF	BoW	Weaponization vs Non-weapon.	0.67	0.67	0.67	0.67
RF	TF-IDF	Weaponization vs Non-weapon.	0.69	0.70	0.69	0.69
RF	doc2vec	Weaponization vs Non-weapon.	0.58	0.58	0.58	0.58

5.2 IDENTIFICANDO PONTOS DE INFORMAÇÃO EM FÓRUNS ONLINE

Neste experimento, focamos nas 1.037 anotações de especialistas do HackForums. Para os experimentos, utilizamos apenas os rótulos Proof-Of-Concept (PoC), Weaponization e Exploitation.

5.2.1 **Conjunto de dados**

A Tabela 4 mostra os boards que contêm a maior quantidade de posts citando CVEs. Nos dois principais boards, encontramos usuários vendendo e comprando exploits, o que indica que as discussões sobre vulnerabilidades geralmente tratam de exploits já disponíveis no mercado. Além disso, a Tabela 4 também mostra a distribuição dos posts entre diferentes classes nos diferentes boards do Hackforums. Observe que, no board de *pentesting*, por exemplo, encontramos atividade significativa relacionada a weaponization e exploitation. Em contraste, poucos posts citam explicitamente identificadores CVE no board de tutoriais de hacking.

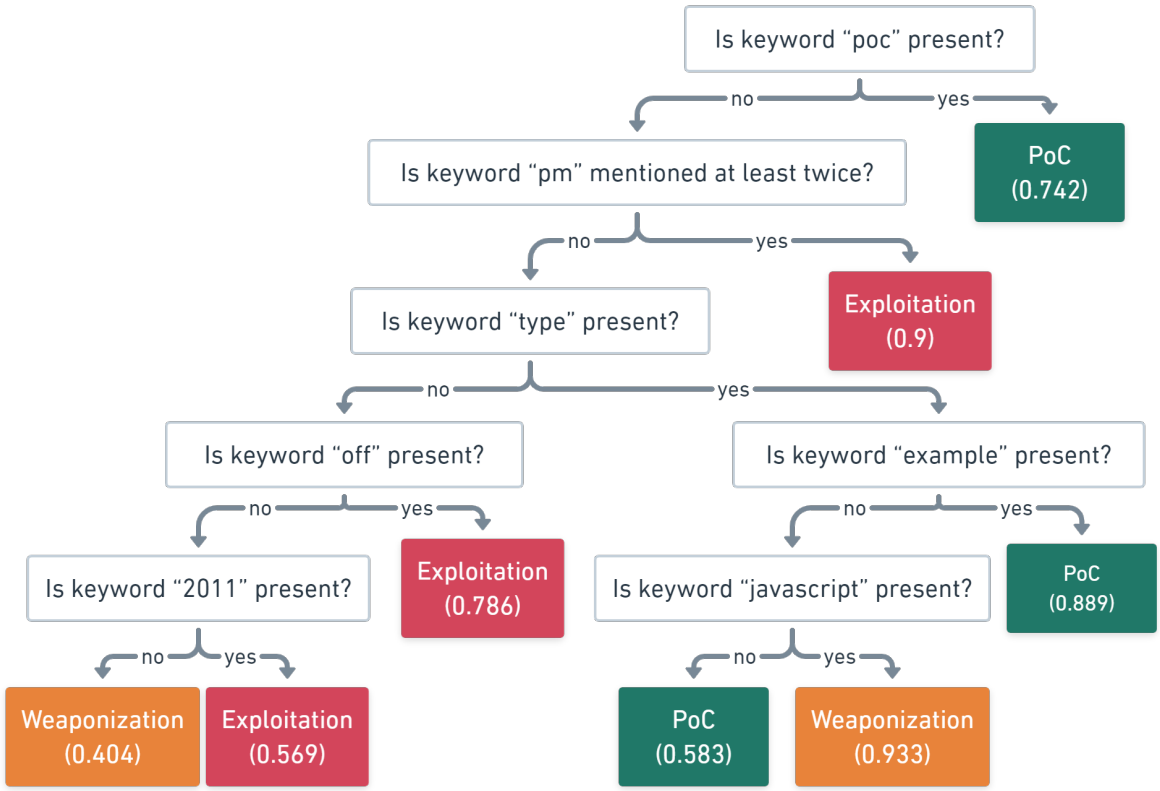


Figura 12 – Árvore de decisão para classificar PoC, weaponization e exploitation.

5.2.2 Interpretações dos modelos

Reportamos o desempenho dos classificadores considerados. Para isso, utilizamos quatro métricas: accuracy, precision, recall e F1. A Tabela 5 mostra os resultados obtidos, considerando o melhor conjunto de hiperparâmetros para cada configuração, conforme descrito anteriormente. Em particular, nossas configurações variam em função da estratégia de codificação textual e das classes-alvo. Observamos aumento de acurácia ao migrar da codificação mais simples e interpretável (BoW) para a mais complexa, porém menos interpretável (doc2vec). De fato, o doc2vec supera BoW e TF-IDF, exceto nos casos "PoC versus Non-PoC" e "Weaponization versus Non-Weaponization". Ainda assim, o BoW é fundamental para produzir a árvore interpretável apresentada na Figura 12.

Com relação às classes-alvo, consideramos PoC vs Weaponization vs Exploitation e três classificadores adicionais do tipo um-contra-todos, nos quais cada classificador binário separa membros de uma classe dos membros das demais classes. Os melhores resultados foram obtidos ao filtrar exploitation in the wild do restante das threads, o que é, possivelmente, o primeiro passo para identificar informação relevante em fóruns clandestinos, já que exploitation representa o risco mais iminente. Quanto ao modelo classificador, árvores de decisão são mais simples do que random forests, produzindo previsões menos precisas, mas sendo interpretáveis, como ilustrado a seguir.

A Figura 12 ilustra a árvore de decisão usada para classificar entre PoC, weaponization e exploitation (primeira linha da Tabela 5). Cada nó interno na árvore contém uma regra que divide o conjunto de dados, e cada folha indica a classe predominante naquela divisão e sua frequência. Apesar de nem todas as regras de divisão que aparecem na Figura 12 serem interpretáveis, já podemos extrair insights interessantes. Na raiz da árvore, encontramos a regra com maior poder de divisão, de acordo com o critério do índice de Gini. De fato, a raiz, junto com a folha imediatamente abaixo dela, indica que, se a thread contém a palavra-chave "poc", com 74,2% de chance ela é, de fato, um *proof-of-concept*. A regra seguinte indica que posts nos quais os usuários demonstram preocupação com privacidade, isto é, contendo a palavra-chave "pm", que significa "private message" no jargão dos fóruns blackhat, correspondem à exploitation in the wild. Por fim, também observamos que JavaScript é uma linguagem comumente utilizada para produzir exploits, por exemplo, que injetam código por meio de campos de entrada não verificados.

5.2.3 Conclusões do experimento

Neste experimento, utilizamos CrimeBB e métodos de aprendizado de máquina para aprender conteúdo textual e distinguir entre: (1) ameaça potencial (*proof of concept*), (2) ameaça iminente (*weaponization*) e (3) criminosos discutindo uma ameaça (*exploitation in the wild*). Verificamos que os CVEs mais citados estão tipicamente relacionados a maiores riscos e que é viável filtrar automaticamente threads de exploitation com acurácia superior a 99%.

5.3 INFERINDO TÓPICOS A PARTIR DE FÓRUNS DE HACKING

Neste experimento, utilizamos modelagem de tópicos para analisar a exploração de vulnerabilidades em fóruns clandestinos de hacking. Aplicamos Latent Dirichlet Allocation (LDA) (BLEI; NG; JORDAN, 2001) para inferir tópicos, ou seja, com base no conteúdo das threads de discussão, identificamos o tópico sendo discutido.

5.3.1 Conjunto de dados

Empregamos os rótulos como tópicos: Proof of Concept (PoC), weaponization, exploitation e "other" (incluindo others, scam, warning e help), rotulados com os IDs correspondentes 1, 2, 3 e 4, respectivamente. O objetivo é identificar temas-chave relacionados a técnicas de exploit, vulnerabilidades e potenciais alvos, fornecendo insights valiosos sobre o cenário da exploração de vulnerabilidades.

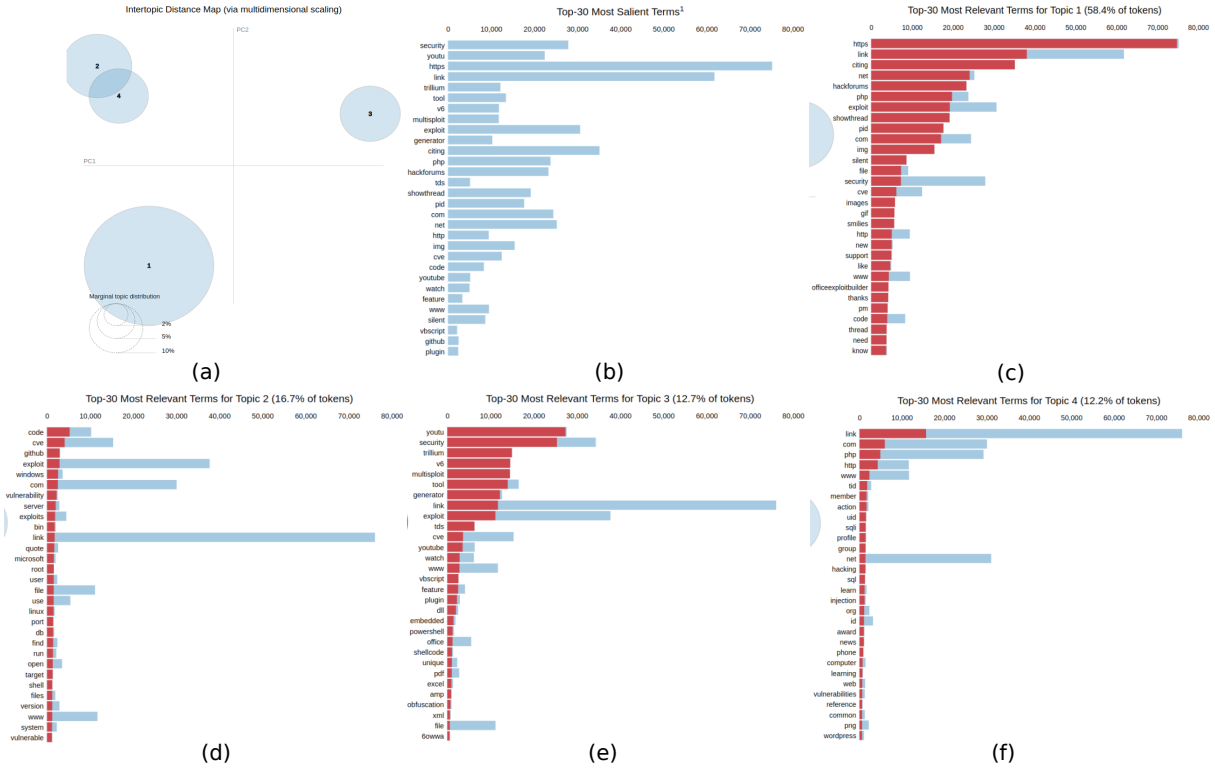


Figura 13 – Projeção dos grupos de tópicos por componentes principais. (a) O raio de cada grupo determina a distribuição marginal do tópico, (b) os 30 termos mais salientes, (c) os 30 termos mais relevantes para o tópico *PoC*, (d) os 30 termos mais relevantes para o tópico *Weaponization*, (e) os 30 termos mais relevantes para o tópico *Exploitation* e (f) os 30 termos mais relevantes para o tópico *Others*.

5.3.2 Modelagem de tópicos

Utilizamos a biblioteca Gensim (REHUREK; SOJKA, 2011) para realizar modelagem de tópicos com o algoritmo LDA. Na Figura 13 (a), mostramos a projeção dos grupos de tópicos por componentes principais; observamos que os tópicos *Weaponization* e *exploitation* apresentam interseção. Esse é um resultado interessante devido à proximidade entre uma vulnerabilidade em weaponization e exploitation. Na Figura 13 (b), mostramos as 30 palavras mais relevantes em todos os tópicos; observamos que palavras como "https", "link", "citing" e "exploit" são as mais relevantes.

Na Figura 13 (c), usando apenas 58,4% dos tokens, as palavras mais relevantes para o tópico *PoC* são "https", "link", "php" etc. Observamos que, em geral, essas palavras são relevantes para todos os tópicos, exceto "code" e "security". Na Figura 13 (d), mostramos as palavras relevantes para o tópico *weaponization*, usando apenas 16,7% dos tokens, sendo elas "code", "cve", "github" etc. Na Figura 13 (e), mostramos as palavras relevantes para o tópico *exploitation*, usando apenas 12,7% dos tokens, sendo elas "security", "trillium" (o usuário que cita a maior quantidade de códigos CVE (MORENO-VERA et al., 2023)), "multisploit", "tool", "exploit" etc.

Na Figura 13 (f), mostramos as palavras relevantes para o tópico *others*, usando apenas 12,2% dos tokens, sendo elas “link”, “com”, “php”, “member”, “profile”, “learn” etc. Observamos que, para cada tópico, há palavras relevantes que nos permitem compreender a principal discussão. Na interseção entre os tópicos *Weaponization* e *exploitation*, temos palavras relevantes como “code”, “cve”, “exploit”, “link” e “vulnerability”. Além disso, no tópico *PoC*, as palavras mais relevantes são “https”, “link”, “citing”, “net” etc. Isso ocorre devido à natureza da discussão, que envolve compartilhamento de links, tutoriais, código etc. A partir disso, observamos que a modelagem de tópicos pode inferir o tema da discussão em uma thread. Além disso, sabemos que em cada thread podem ser discutidos vários temas, mas apenas um tópico predominante.

5.3.3 Conclusões do experimento

Neste experimento, aplicamos uma técnica de aprendizado não supervisionado, modelagem de tópicos, para analisar a exploração no mundo real. Utilizamos anotações de especialistas e as recategorizamos em PoC, exploitation, weaponization e others. Conseguimos identificar grupos separados de palavras correspondentes a cada tópico, além de identificar palavras relevantes como “code”, “cve”, “exploit”, “link” e “vulnerability”, comumente utilizadas em discussões sobre exploitation e weaponization. Verificamos que é possível identificar ameaças emergentes, ajudando profissionais de segurança e pesquisadores a se manterem informados e a priorizarem suas estratégias de defesa de forma apropriada.

5.4 REVELANDO PONTOS DE INFORMAÇÃO EM FÓRUNS

Neste experimento, utilizamos os rótulos do PostCog, anotações de especialistas e rótulos de classificação gerados por GPT.

5.4.1 Conjunto de dados

Comparamos e avaliamos o desempenho dos modelos descritos utilizando a seguinte rotulação:

- **Rótulos do PostCog**: esses rótulos foram fornecidos por especialistas de domínio do PostCog e serviram como linha de base para nossa análise (ver Tabela 2 (a), (b) e (c)).
- **Rótulos de especialistas**: esses rótulos foram fornecidos por especialistas de domínio utilizando um guia de codificação reportado em (MORENO-VERA et al., 2023) (ver Tabela 2 (d)). Comparamos nossos novos resultados com esses anteriores, focando em três classes: “poc”, “weaponization” e “exploitation”.

- **Rótulos do ChatGPT**: utilizamos o ChatGPT para gerar novos rótulos, com o objetivo de explorar o potencial de anotações guiadas por IA e contrastar os resultados de classificação usando esses novos rótulos (ver Tabela 3) com os rótulos do PostCog.

O HackForum contém 4 milhões de threads, das quais apenas aproximadamente 3.200 estão rotuladas com seu tipo de crime no PostCog. No caso dos rótulos de tipo de post e intenção, há ainda menos rótulos no PostCog (2700), enquanto o restante foi catalogado como "other" (correspondendo à ausência de rótulo). Portanto, uma de nossas tentativas ao utilizar o ChatGPT foi rotular mais threads, para preencher dados ausentes.

5.4.2 Rotulador GPT e classificadores

Reportamos o desempenho dos classificadores considerados. Apresentamos as quatro métricas descritas: accuracy, precision, recall e F1 score. Em geral, o RandomForest destacou-se como o modelo de melhor desempenho, alcançando a maior acurácia e robustez em nossos experimentos. Em todos os experimentos com RandomForest, obtivemos acurácia superior a 74%-99%. Outros métodos produziram acurácias inferiores a 60%. Além disso, observamos que os resultados com a rotulação do ChatGPT nem sempre melhoram os resultados da linha de base, especialmente ao tentar prever tipo de post e tipo de crime. A Tabela 6 reporta o desempenho de random forests para inferir rótulos fornecidos pelo PostCog, ChatGPT e por nossas próprias anotações de especialistas. Temos um RandomForest separado para inferir tipo de post, tipo de crime, intenção e anotações de especialistas. Observamos que, para tipo de crime e intenção, modelos treinados com rótulos do ChatGPT e do PostCog produzem desempenho semelhante. No entanto, para tipo de post e anotações de especialistas, os rótulos do ChatGPT foram mais difíceis de prever. Além disso, utilizamos resultados de nossos trabalhos anteriores (CAINES et al., 2018b; SIU; COLLIER; HUTCHINGS, 2021; MORENO-VERA et al., 2023) e os comparamos com nossos novos resultados, indicando que os novos resultados, independentemente do uso de rótulos do ChatGPT, correspondem a melhor desempenho. A inspeção manual dos rótulos de especialistas e dos rótulos do ChatGPT sugere que o ChatGPT nem sempre é melhor do que as anotações de especialistas humanos.

5.4.3 Explicações dos modelos

Para explicar os resultados produzidos pelos RandomForests, utilizamos o SHAP (SHapley Additive exPlanations), que permite a interpretação de classificadores textuais, oferecendo insights claros e consistentes sobre as previsões do modelo (LUNDBERG; LEE, 2017; RIBEIRO; SINGH; GUESTRIN, 2016).

Tabela 6 – Resumo do desempenho de Random Forest (RF) para todos os experimentos.

	Classes-alvo	Accuracy	Precision	Recall	F1
Crime type	Rótulos do PostCog	**0.97**	0.97	0.99	0.98
	Rótulos do ChatGPT	0.95	0.98	0.94	0.96
	Trabalho anterior (SIU; COLLIER; HUTCHINGS, 2021)	0.89	0.9	0.89	0.89
Intention	Rótulos do PostCog	0.98	0.95	0.97	0.95
	Rótulos do ChatGPT	**0.99**	0.97	0.99	0.98
	Trabalho anterior (CAINES et al., 2018b)	–	0.78	0.49	0.61
Post type	Rótulos do PostCog	**0.81**	0.79	0.89	0.82
	Rótulos do ChatGPT	0.74	0.75	0.76	0.75
	Trabalho anterior (CAINES et al., 2018b)	–	0.91	0.78	0.84
Expert annotations	Rótulos de especialistas	**0.96**	0.97	0.98	0.97
	Rótulos do ChatGPT	0.91	0.92	0.93	0.92
	Trabalho anterior (MORENO-VERA et al., 2023)	0.86	0.87	0.86	0.86

Tipo de Crime

A Figura 14 reporta os resultados de classificação e as explicações geradas por SHAP para os rótulos de tipo de crime do PostCog, que alcançaram acurácia de 97%, e para os rótulos de tipo de crime do ChatGPT, que alcançaram acurácia de 95%. Na Figura 14 (a), focamos nos rótulos do PostCog. Observe que temos oito classes, mas classes com baixo número de amostras têm relevância mínima nas explicações geradas pelo SHAP. As palavras "injection", "fud" (fully undetectable),[1] "hacked" e "open" apresentam alta relevância para as classes "bots/malware", "not criminal" e "wrong access/sql injection".

A palavra "pm" é relevante ao avaliar a classe "trading credential". "pm" significa *private message* e indica que uma mensagem privada será usada para compartilhar detalhes sobre o post, por questões de privacidade. Palavras como "btc" (para Bitcoin) e "selling" também são relevantes para "spam-related/marketing". Na Figura 14(b), as explicações com rótulos do ChatGPT concentram-se em duas classes: "not criminal" e "cybercrime activities". As palavras "server" e "prices" mostram relevância para a classe "cybercrime support services". Também verificamos que "php", "malicious", "scan" e "free" são palavras relacionadas a "cybercrime support services".

Intenção

Na Figura 15, apresentamos os resultados de classificação e as explicações geradas por SHAP para os rótulos de intenção do PostCog, correspondendo a uma acurácia de 98%, e para os rótulos de intenção do ChatGPT, com acurácia de 99%. Na Figura 15 (a), mostramos a relevância de diferentes palavras para os rótulos de intenção do PostCog. A palavra "thanks" tem alta relevância para as classes "neutral" e "gratitude", e contribuição mínima para "positive". A palavra "vouch" é informativa para a classe de mesmo nome.

A Figura 15 (b) mostra que a palavra "thanks" é relevante para determinar as

[1] Um exploit *fully undetectable* é um exploit que não é detectado por nenhuma ferramenta antivírus conhecida.

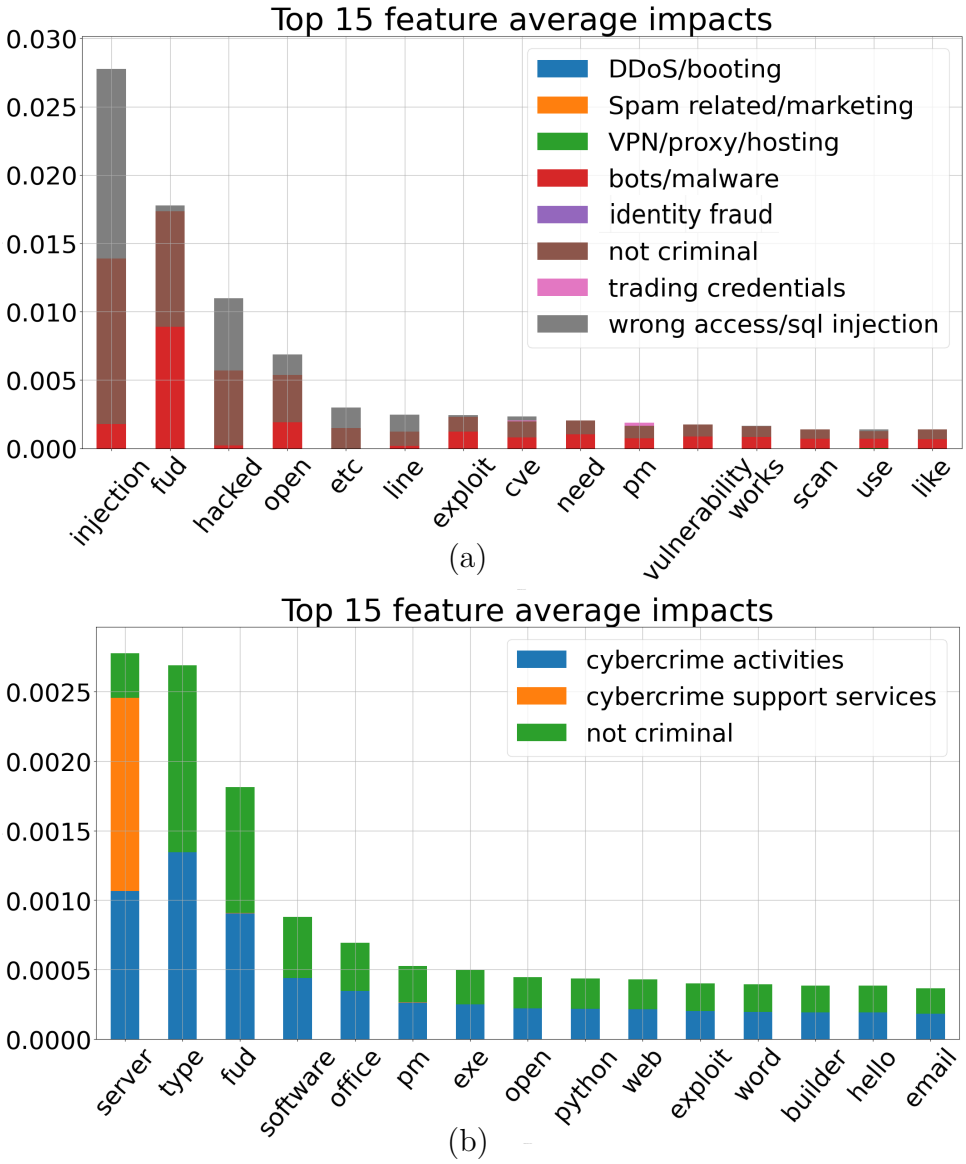
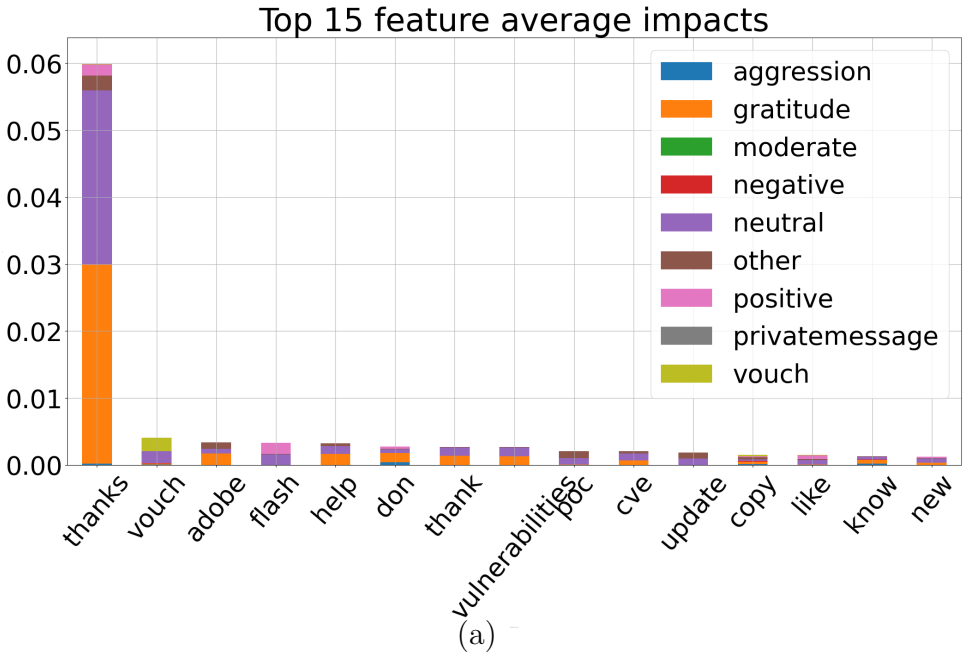


Figura 14 – As 15 características mais relevantes com base nos valores de explicação SHAP foram determinadas a partir de modelos treinados com (a) os **rótulos de tipo de crime do PostCog** e (b) os **rótulos de tipo de crime do ChatGPT**.

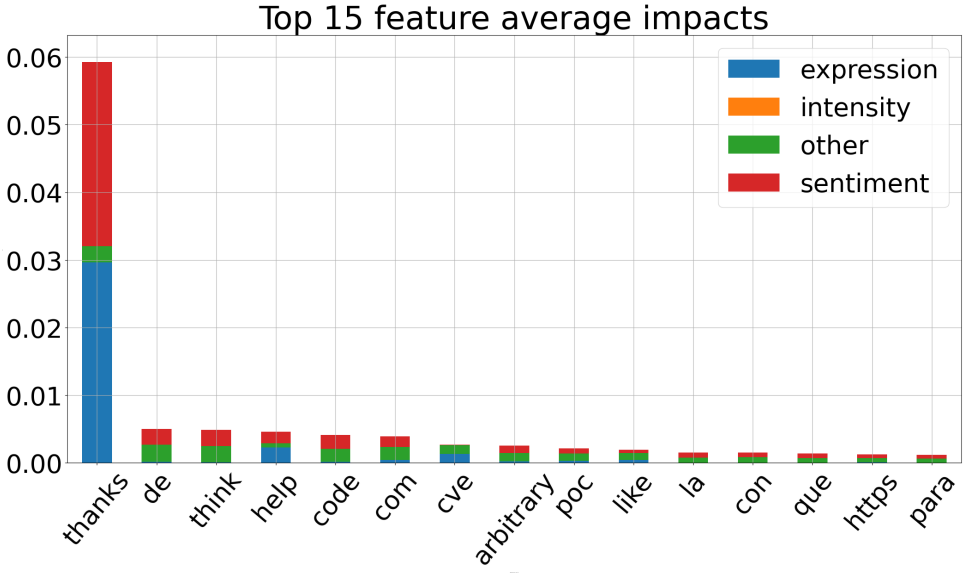
classes de intenção definidas pelo ChatGPT. De fato, “thanks” ajuda a diferenciar entre as classes “sentiment” e “expression”. Também identificamos palavras cujo significado não representa algo relevante, como “com”, “de”, “la”, “para” etc. Isso indica que, embora a precisão da random forest tenha sido alta, nem sempre é simples encontrar explicações baseadas em SHAP para os resultados de classificação.

Tipo de Post

A Figura 16 apresenta os resultados de classificação e as explicações baseadas em SHAP para os rótulos de tipo de post do PostCog, com acurácia de 81%, e para os rótulos de tipo de post do ChatGPT, com acurácia de 74%. Na Figura 16(a), apresentamos as explicações para o modelo treinado com os rótulos de tipo de post do PostCog, observando



(a)



(b)

Figura 15 – As 15 características mais relevantes com base nos valores de explicação SHAP foram determinadas a partir de modelos treinados com (a) os **rótulos de intenção do PostCog** e (b) os **rótulos de intenção do ChatGPT**.

que a palavra-chave “fud” é relevante para distinguir entre infoRequest, comment e offerX.

Outras palavras-chave, como “poc”, “https” e “found”, são mais relevantes para identificar classes como comment e infoRequest. A Figura 16(b) ilustra que as explicações baseadas em SHAP usando os rótulos do ChatGPT exibem melhor equilíbrio de importância entre classes para cada palavra. Por exemplo, palavras como “found” e “public” estão mais associadas a requests, communication e offers. Além disso, termos como “www”, “poc” e “exploit” podem ajudar a distinguir se a classe correspondente é communication ou offers.

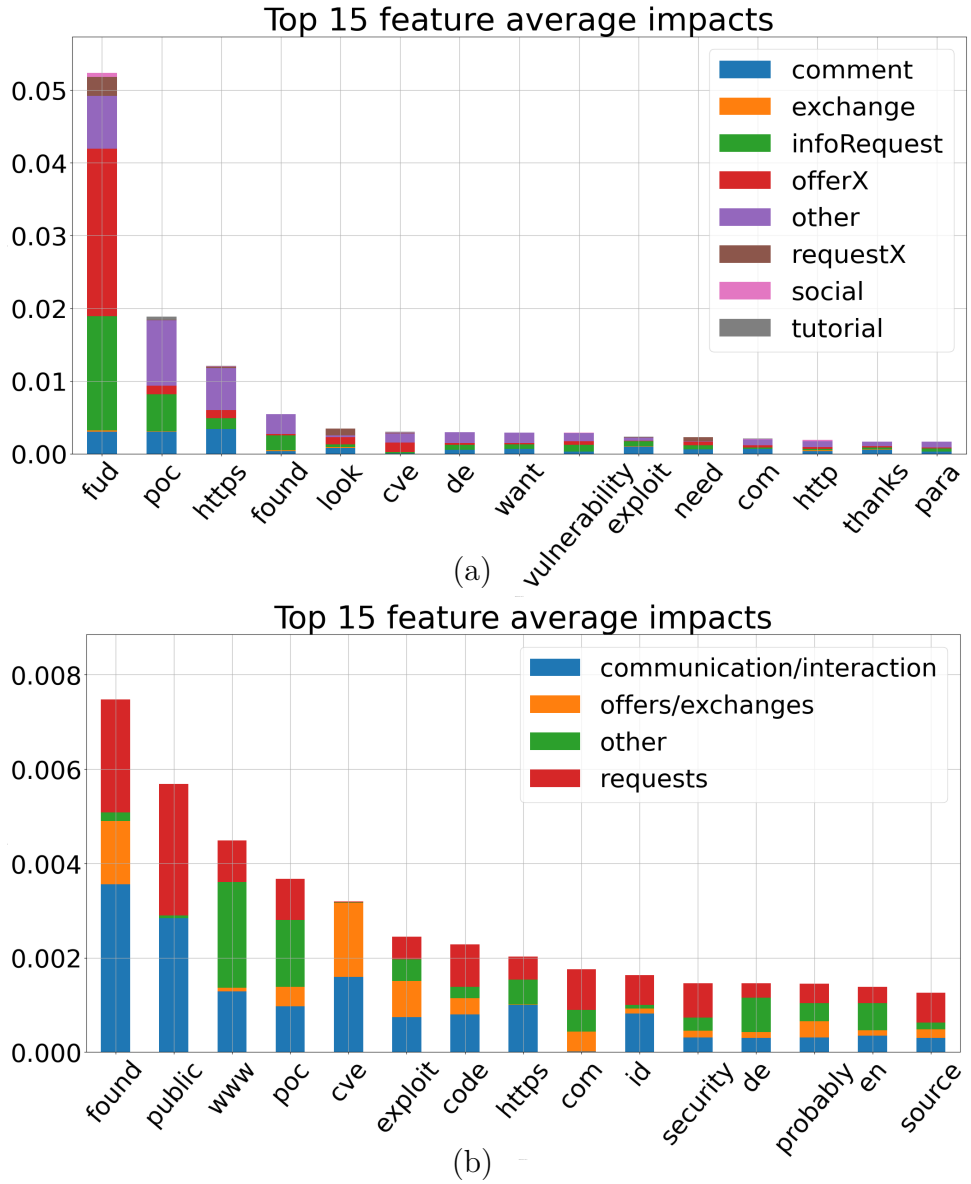
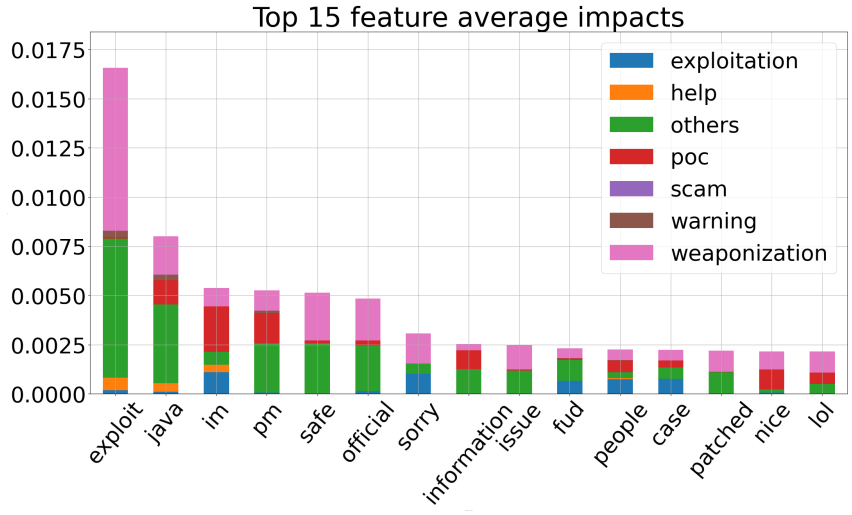


Figura 16 – As 15 características mais relevantes com base nos valores de explicação SHAP foram determinadas a partir de modelos treinados com (a) os **rótulos de tipo de post do PostCog** e (b) os **rótulos de tipo de post do ChatGPT**.

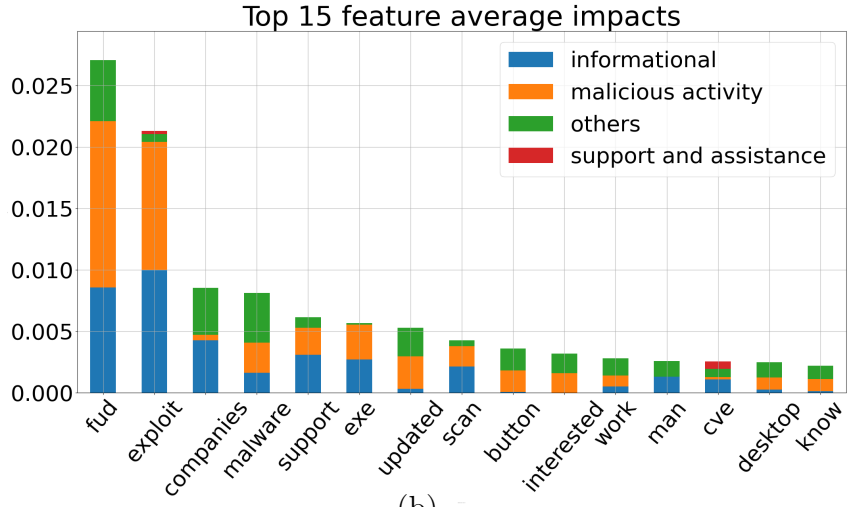
Anotações de Especialistas

Na Figura 17, mostramos a explicação baseada em SHAP dos modelos random forest treinados com rótulos de anotações de especialistas (acurácia de 96%) e com os rótulos correspondentes gerados pelo ChatGPT (acurácia de 91%). Na Figura 17(a), observamos que a palavra "exploit" está fortemente relacionada a "weaponization", enquanto a palavra "sorry" está relacionada à classe "exploitation". Também encontramos palavras importantes como "pm" (*private message*), "fud", "issue" e "information" relevantes para as classes "poc" e "weaponization".

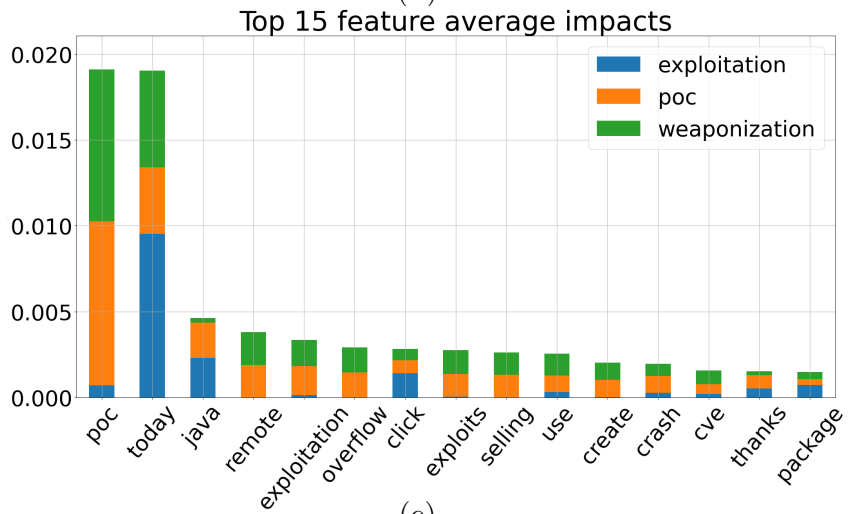
Na Figura 17 (b), a relevância das palavras "fud", "exploit", "companies", "malware", "scan", "interested" e "man" é informativa para "informational" e "malicious activities". Es-



(a)



(b)



(c)

Figura 17 – As 15 características mais relevantes segundo os valores de explicação SHAP calculados a partir de modelos treinados usando os **rótulos de anotações de especialistas** (a), os **rótulos de anotações de especialistas do ChatGPT** (b) e os **resultados de trabalhos anteriores** (c).

sas explicações mostram a importância da seleção de características; algumas palavras são valiosas por causa de seu contexto, por exemplo, palavras como “nice”, “lol”, entre outras. Na Figura 17 (c), reportamos resultados do nosso modelo anterior, com 86% de acurácia, conforme reportado em (MORENO-VERA et al., 2023); usando SHAP, identificamos que as palavras “remote”, “pm”, “today”, “java”, “selling”, “crash” e “use” são relevantes para fins de classificação, em concordância com os achados reportados em (MORENO-VERA et al., 2023) por meio de métodos alternativos.

5.4.4 Conclusões do experimento

Neste experimento, classificamos posts de fóruns clandestinos por atividade cibercriminosa, intenções textuais e tipo de conteúdo utilizando TF-IDF, um classificador RandomForest, e também utilizamos o método SHAP para análise de explicação dos modelos. Além disso, observamos que o ChatGPT introduziu rótulos ruidosos em várias instâncias. Sua tendência a gerar informações plausíveis, mas nem sempre precisas ou relevantes, pode levar a características enganosas ou irrelevantes nos dados de treinamento. Isso pode degradar o desempenho do classificador ao distorcer a importância das características ou introduzir viés. Assim, embora o ChatGPT possa ser uma ferramenta valiosa, é crucial utilizá-lo com cuidado e complementá-lo com etapas rigorosas de validação e pré-processamento de dados para mitigar possíveis limitações.

6 CONCLUSÕES

"Dados são poder, desde que você saiba como utilizá-los."

No campo da inteligência de ameaças, tem havido uma tendência crescente de utilizar abordagens orientadas por dados para apoiar decisões de segurança. Seja para classificação de ameaças, gerenciamento de correções ou detecção de ameaças, os dados vêm desempenhando um papel cada vez mais transformador nas práticas modernas de cibersegurança. Tradicionalmente, operadores de segurança têm recorrido a *feeds* de inteligência de ameaças (TI), incluindo listas negras de IPs, assinaturas de arquivos e descrições textuais, para orientar sua tomada de decisão. Em síntese, nossos experimentos demonstram o potencial das técnicas de aprendizado de máquina e de processamento de linguagem natural para analisar e categorizar conteúdos em fóruns clandestinos, auxiliando na identificação e priorização de ameaças emergentes em cibersegurança.

Utilizamos o CrimeBB, o framework PostCog e técnicas de aprendizado de máquina para analisar conteúdo textual e classificá-lo com base no nível de exploração. Nossos achados empíricos, obtidos a partir da correlação entre escores CVSS e EPSS e os CVEs identificados no CrimeBB, revelaram que os CVEs mais frequentemente citados estão tipicamente associados a maiores riscos. Isso sugere que é viável filtrar automaticamente, com alta acurácia, threads relacionadas à exploração por meio do nosso classificador de conteúdo. Além disso, realizamos uma análise não supervisionada por meio de modelagem de tópicos, que demonstrou que as discussões podem ser agrupadas com base em termos relevantes. Ao aplicar modelos GPT para analisar threads de discussão, conseguimos gerar novos rótulos com base no conteúdo, fornecendo insights valiosos sobre aquilo que os modelos GPT priorizam em comparação com as anotações de especialistas.

De modo geral, nossos experimentos evidenciam a eficácia da combinação de aprendizado de máquina, aprendizado não supervisionado e técnicas de inteligência artificial explicável na detecção de ameaças e na inteligência cibernética, ao mesmo tempo em que destacam a importância de uma validação rigorosa no uso de métodos assistidos por IA. Essas abordagens podem aprimorar significativamente as capacidades de análise de ameaças, permitindo uma postura de cibersegurança mais proativa e bem fundamentada.

REFERÊNCIAS

ABLON, L.; LIBICKI, M. C.; GOLAY, A. A. *Markets for cybercrime tools and stolen data: Hackers' bazaar*. [S.l.]: Rand Corporation, 2014. Citado na página 13.

AHMED, T.; DEVANBU, P. Few-shot training llms for project-specific code-summarization. In: *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. [S.l.: s.n.], 2022. p. 1–5. Citado na página 34.

ALLODI, L. Economic factors of vulnerability trade and exploitation. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017. Disponível em: https://api.semanticscholar.org/CorpusID:20070349. Citado 6 vezes nas páginas 13, 27, 31, 39, 40 e 41.

ANDERSON, R. et al. Measuring the changing cost of cybercrime. *The 2019 Workshop on the Economics of Information Security*, 2019. Citado na página 13.

BADA, M.; PETE, I. An exploration of the cybercrime ecosystem around shodan. *2020 7th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, p. 1–8, 2020. Disponível em: https://api.semanticscholar.org/CorpusID:231204238. Citado na página 27.

BASHEER, R.; ALKHATIB, B. Threats from the dark: a review over dark web investigation research for cyber threat intelligence. *Journal of Computer Networks and Communications*, Hindawi Limited, v. 2021, p. 1–21, 2021. Citado 2 vezes nas páginas 13 e 14.

BENGIO, Y. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, v. 2, p. 1–127, 2007. Disponível em: https://api.semanticscholar.org/CorpusID:207178999. Citado na página 18.

BLEI, D. M.; NG, A.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, v. 3, p. 993–1022, 2001. Citado na página 51.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. *Proceedings of the fifth annual workshop on Computational learning theory*. [S.l.], 1992. p. 144–152. Citado na página 47.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. Citado na página 47.

BUTLER, S. Cyber 9/11 will not take place: A user perspective of bitcoin and cryptocurrencies from underground and dark net forums. In: *Workshop on Socio-Technical Aspects in Security and Trust*. [s.n.], 2020. Disponível em: https://api.semanticscholar.org/CorpusID:235600176. Citado na página 27.

CAINES, A. et al. Aggressive language in an online hacking forum. In: *Workshop on Abusive Language Online*. [s.n.], 2018. Disponível em: https://api.semanticscholar.org/CorpusID:53651435. Citado na página 27.

CAINES, A. et al. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, v. 7, 2018. Citado 5 vezes nas páginas 27, 32, 33, 54 e 55.

CAMPOBASSO, M.; ALLODI, L. Threat/crawl: a trainable, highly-reusable, and extensible automated method and tool to crawl criminal underground forums. In: *APWG eCrime 2022*. [S.l.: s.n.], 2022. ArXiv:2212.03641. Citado na página 13.

CHEN, D. D. et al. Towards automated dynamic analysis for linux-based embedded firmware. In: *Network and Distributed System Security Symposium*. [S.l.: s.n.], 2016. Citado na página 27.

CHUA, Y. T.; WILSON, L. Beyond black and white: the intersection of ideologies in online extremist communities. *European Journal on Criminal Policy and Research*, v. 29, p. 337–354, 2023. Disponível em: https://api.semanticscholar.org/CorpusID:261128636. Citado na página 27.

CORTES, C.; VAPNIK, V. N. Support-vector networks. *Machine Learning*, v. 20, p. 273–297, 1995. Disponível em: https://api.semanticscholar.org/CorpusID:52874011. Citado na página 47.

DEGUARA, N. et al. Threat miner: A text analysis engine for threat identification using dark web data. In: *Big Data*. [S.l.: s.n.], 2022. p. 3043–3052. Citado na página 27.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Disponível em: https://arxiv.org/abs/1810.04805. Citado na página 23.

EDKRANTZ, M.; TRUVÉ, S.; SAID, A. Predicting vulnerability exploits in the wild. *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*, p. 513–514, 2015. Citado na página 27.

FANG, Y. et al. Analyzing and identifying data breaches in underground forums. *IEEE Access*, v. 7, p. 48770–48777, 2019. Disponível em: https://api.semanticscholar.org/CorpusID:131776921. Citado na página 27.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. Citado na página 18.

HANKS, C. et al. Recognizing and extracting cybersecurity entities from text. In: *ICML*. [S.l.: s.n.], 2022. Citado na página 13.

HARRIS, Z. S. Distributional structure. In: . [s.n.], 1954. Disponível em: https://api.semanticscholar.org/CorpusID:86680084. Citado 2 vezes nas páginas 21 e 39.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. Applied logistic regression: Hosmer/applied logistic regression. In: . [S.l.: s.n.], 2005. Citado na página 47.

JACOBS, J.; ROMANOSKY, S. et al. Exploit prediction scoring system (EPSS). *Digital Threats: Research and Practice*, ACM New York, NY, USA, v. 2, n. 3, p. 1–17, 2021. Citado na página 26.

LE, Q. V.; MIKOLOV, T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. [s.n.], 2014. Disponível em: https://api.semanticscholar.org/CorpusID:2407601. Citado na página 23.

LEMAîTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, v. 18, n. 17, p. 1–5, 2017. Citado na página 48.

LIANG, H. et al. Fuzzing: State of the art. *IEEE Transactions on Reliability*, v. 67, p. 1199–1218, 2018. Citado na página 27.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Neural Information Processing Systems*. [S.l.: s.n.], 2017. Citado na página 54.

MAN, J.; SIU, G. A.; HUTCHINGS, A. Autism disclosures and cybercrime discourse on a large underground forum. *2023 APWG Symposium on Electronic Crime Research (eCrime)*, p. 1–14, 2023. Disponível em: https://api.semanticscholar.org/CorpusID:268499519. Citado na página 27.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: *International Conference on Learning Representations*. [s.n.], 2013. Disponível em: https://api.semanticscholar.org/CorpusID:5959482. Citado 2 vezes nas páginas 23 e 39.

MORENO-VERA, F. Inferring discussion topics about exploitation of vulnerabilities from underground hacking forums. In: *ICTC*. [S.l.: s.n.], 2023. p. 816–821. Citado 3 vezes nas páginas 14, 17 e 28.

MORENO-VERA, F.; MENASCHE, D.; LIMA, C. Beneath the cream: Unveiling relevant information points from crimebb underground forums with its ground truth labels. In: *International Symposium on Cyber Security, Cryptology and Machine Learning*. [S.l.]: Springer, 2024. p. 280–290. Citado na página 16.

MORENO-VERA, F. et al. Cream skimming the underground: Identifying relevant information points from online forums. In: *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*. [S.l.: s.n.], 2023. p. 66–71. Citado 11 vezes nas páginas 14, 17, 26, 28, 31, 33, 52, 53, 54, 55 e 60.

MORROW, T. C. S. The future of cybercrime & security. 2019. Citado na página 13.

PARACHA, A.; ARSHAD, J.; KHAN, M. M. SUS You're SUS!-Identifying Influencer Hackers on Dark Web Social Networks. *Computers and Electrical Engineering*, Elsevier, 2023. Citado na página 39.

PASTRANA, S. et al. Characterizing eve: Analysing cybercrime actors in a large underground forum. In: *International Symposium on Recent Advances in Intrusion Detection*. [s.n.], 2018. Disponível em: https://api.semanticscholar.org/CorpusID:51686519. Citado na página 27.

PASTRANA, S.; HUTCHINGS, A. et al. Measuring ewhoring. In: *Proceedings of the Internet Measurement Conference*. [S.l.: s.n.], 2019. p. 463–477. Citado 2 vezes nas páginas 26 e 27.

PASTRANA, S.; THOMAS, D. R. et al. CrimeBB: Enabling cybercrime research on underground forums at scale. In: *Proceedings of the 2018 World Wide Web Conference*. [S.l.: s.n.], 2018. p. 1845–1854. Citado 6 vezes nas páginas 13, 14, 26, 27, 29 e 30.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Conference on Empirical Methods in Natural Language Processing*. [s.n.], 2014. Disponível em: https://api.semanticscholar.org/CorpusID:1957433. Citado na página 23.

PETE, I. et al. Postcog: A tool for interdisciplinary research into underground forums at scale. *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, p. 93–104, 2022. Citado 2 vezes nas páginas 29 e 31.

RADFORD, A.; NARASIMHAN, K. Improving language understanding by generative pre-training. In: . [s.n.], 2018. Disponível em: https://api.semanticscholar.org/CorpusID:49313245. Citado na página 23.

RAHMAN, M. R. et al. What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. *ACM Computing Surveys*, ACM New York, NY, 2021. Citado na página 26.

REHUREK, R.; SOJKA, P. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, v. 3, n. 2, 2011. Citado na página 52.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *SIGKDD*, 2016. Citado na página 54.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, v. 24, p. 513–523, 1988. Citado na página 39.

SIU, G. A.; COLLIER, B.; HUTCHINGS, A. Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. *EuroS&PW*, p. 191–201, 2021. Citado 4 vezes nas páginas 27, 33, 54 e 55.

SPEYBROECK, N. Classification and regression trees. *International Journal of Public Health*, v. 57, p. 243–246, 2012. Citado na página 47.

SUCIU, O. et al. Expected exploitability: Predicting the development of functional vulnerability exploits. In: *31st USENIX Security Symposium*. [S.l.: s.n.], 2022. p. 377–394. Citado na página 26.

SUN, J. et al. Generating informative cve description from exploitdb posts by extractive summarization. *ArXiv*, abs/2101.01431, 2021. Disponível em: https://api.semanticscholar.org/CorpusID:230523926. Citado na página 28.

SUTTON, M. S.; GREENE, A. R.; AMINI, P. F. Fuzzing: Brute force vulnerability discovery. In: . [S.l.: s.n.], 2007. Citado na página 27.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series b-methodological*, v. 58, p. 267–288, 1996. Citado na página 47.

TIKHONOV, A. N. On the stability of inverse problems. In: *Dokl. Akad. Nauk SSSR*. [S.l.: s.n.], 1943. v. 39, p. 195–198. Citado na página 47.

VALEROS, V.; GARCIA, S. Growth and commoditization of remote access trojans. In: IEEE. *2020 IEEE EuroS&PW*. [S.l.], 2020. p. 454–462. Citado na página 40.

WANG, B. et al. Characterizing and modeling patching practices of industrial control systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, ACM New York, NY, USA, v. 1, n. 1, p. 1–23, 2017. Citado na página 42.