UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

FELIPE VEIGA RAMOS

EXTRAÇÃO E ANÁLISE DE PUBLICAÇÕES ASSOCIADAS À CIBERSEGURANÇA NO PASTEBIN

MONOGRAFIA

CAMPO MOURÃO
2018

FELIPE VEIGA RAMOS

EXTRAÇÃO E ANÁLISE DE PUBLICAÇÕES ASSOCIADAS À CIBERSEGURANÇA NO PASTEBIN

Trabalho de Conclusão de Curso de graduação apresentado à disciplina de Trabalho de Conclusão de Curso 2, do Curso de Bacharelado em Ciência da Computação do Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Rodrigo Campiolo

Coorientador: Prof. Dr. Luiz Arthur F. dos Santos

CAMPO MOURÃO
2018

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

Às **13:50** do dia **19 de novembro de 2018** foi realizada na sala **E101** da UTFPR-CM a sessão pública da defesa do Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação do(a) acadêmico(a) **Felipe Veiga Ramos** com o título **Extração e análise de publicações associadas à cibersegurança no pastebin**. Estavam presentes, além do(a) acadêmico(a), os membros da banca examinadora composta por: **Prof. Dr. Rodrigo Campiolo** (orientador), **Prof. Dr. Luiz Arthur Feitosa dos Santos**, **Prof. Dr. Rodrigo Hübner** e **Prof. Dr. Lucio Geronimo Valentin**. Inicialmente, o(a) acadêmico(a) fez a apresentação do seu trabalho, sendo, em seguida, arguido(a) pela banca examinadora. Após as arguições, sem a presença do(a) acadêmico(a), a banca examinadora o(a) considerou _____ na disciplina de Trabalho de Conclusão de Curso **2** e atribuiu, em consenso, a nota _____ (_____). Este resultado foi comunicado ao(à) acadêmico(a) e aos presentes na sessão pública. A banca examinadora também comunicou ao acadêmico(a) que este resultado fica condicionado à entrega da versão final dentro dos padrões e da documentação exigida pela UTFPR ao professor Responsável do TCC no prazo de **onze dias**. Em seguida foi encerrada a sessão e, para constar, foi lavrada a presente Ata que segue assinada pelos membros da banca examinadora, após lida e considerada conforme.

Observações: _____

_____

_____

_____

_____

Campo Mourão, **19 de novembro de 2018**

Prof. Dr. Luiz Arthur Feitosa dos Santos Membro 1	**Prof. Dr. Rodrigo Hübner** Membro 2
Prof. Dr. Lucio Geronimo Valentin Membro 3	**Prof. Dr. Rodrigo Campiolo** Orientador

A ata de defesa assinada encontra-se na coordenação do curso.

Resumo

Ramos, Felipe Veiga. Extração e análise de publicações associadas à cibersegurança no pastebin. 2018. 85. f. Monografia (Curso de Bacharelado em Ciência da Computação), Universidade Tecnológica Federal do Paraná. Campo Mourão, 2018.

O *Pastebin* é uma ferramenta de compartilhamento de texto puro, ou seja, permite a publicação de textos, inclusive de forma anônima. Nesta monografia objetiva-se investigar os textos (*pastes*) postados no *Pastebin* quanto à sua relevância para extrair e identificar informações que possam ser utilizadas para ações proativas ou reativas mais rápidas na proteção de redes de computadores e sistemas. Para identificar tais informações, foram utilizadas expressões regulares, palavras-chave, detecção de idioma e análise manual, que também serviram de entrada para algoritmos de classificação. A coleta foi realizada num intervalo de 21 dias, resultando em uma base com 3650 *pastes*. A partir do pré-processamento e análise da base por meio de processamento de linguagem natural e estatística, foram extraídas características que resultaram em uma base de inteligência para uso na identificação de novos *pastes* de interesse. Verificou-se que existem informações relacionadas à cibersegurança no *Pastebin*, como venda de informações bancárias, vazamento de credenciais (por exemplo *e-mails*), disponibilização de informações pessoais e programas alterados. Essas informações são importantes para ações proativas ou reações mais rápidas contra ciberameaças.

Palavras-chaves: Cibersegurança. Pastebin. processamento de linguagem natural. recuperação de dados. reação a incidentes de cibersegurança

Abstract

Ramos, Felipe Veiga. Extraction and analysis of publication related with cybersecurity in Pastebin. 2018. 85. f. Monograph (Undergradute Program in Computer Science), Federal University of Technology – Paraná. Campo Mourão, PR, Brazil, 2018.

Pure-text sharing tools allows the anonymous sharing of any kind of text. One of the oldest and most used tools is Pastebin. The goal of this monography is to analyse the relevance of texts (known as pastes), that were posted on Pastebin, to Cybersecurity: how to extract and identify information that can be useful to proactive and quickly reactive actions to protect computer networks and systems. In order to identify such information, the methods used were regular expressions, keywords, word count, frequency of bigrams, trigrams and quadgrams and classification's algorithms. The collector ran for 21 days and 3650 pastes were manually inspected. A base of knowledgement was built using the chraracteristics extracted. Because of it,was possible to know that there are sensible information,like financial and personal data and cracked programs hosted on Pastebin. This kind of information allow better answer to cyber threatments.
Keywords: cybersecurity. pastebin. natural language processing. data retrieval. reaction to cybersecurity incidents

Lista de figuras

Lista de tabelas

Siglas

AES: *Advanced Encryption Standard*
API: *Application Programming Interface*

BSD: *Berkeley Software Distribution*
BTC: *Bitcoin*

CPF: Cadastro de Pessoa Física

DDoS: *Distributed Denial of Service*
DoS: *Denial of Service*

GPL: *GNU GENERAL PUBLIC LICENSE*
GT-EWS: *Grupo de Trabalho Early Warning System*

HTML: *HyperText Markup Language*
HTTP: *HyperText Transfer Protocol*
HTTPS: *HyperText Transfer Protocol Secure*

IDF: *inverse document frequency*
IoT: *Internet of Things*
IP: *Internet Protocol*
IRC: *Internet Relay Chat*

JPA: *Java Persistence API*
JSON: *Java Script Object Notation*

KKK: *Ku Klux Klan*
KNN: *K-nearest neighbors*

NLP: *Natural Language Processing*
NLTK: *Natural Language Toolkit*

PLN: Processamento de Linguagem Natural

RNP: Rede Nacional de Pesquisa

SGBD: Sistema Gerenciador de Banco de Dados

SVC: *C-Support Vector Classification*
SVM: *Support vector machines*

TF: *term frequency*
TFN: Taxa de Falsos Negativos
TFP: Taxa de Falsos Positivos

URL: *Uniform Resource Locator*

WWW: *World Wide Web*

Sumário

Capítulo 1

Introdução

A *World Wide Web* (WWW), comumente chamada de *Web*, é um sistema distribuído, mundialmente utilizado para acesso e compartilhamento de recursos e informações. Entretanto, em diversos lugares na *Web*, são disponibilizadas informações sensíveis, possivelmente oriundas de atividades ilícitas, como senhas de e-mails, mídias sociais, dados de cartões de créditos e outras informações pessoais. Estas disponibilizações ilícitas são usualmente denominadas "vazamento de informações" (BOLHUIS et al., 2014).

Os vazamentos de informações podem ser extremamente prejudiciais, tanto para organizações quanto para indivíduos, podendo levar a imensos prejuízos morais e/ou financeiros. Todavia, tais informações, avaliadas por especialistas em cibersegurança, podem ser utilizadas de forma reativa – em resposta a um incidente, para minimizar seus efeitos – ou proativa – para tentar impedi-lo.

Já existem iniciativas que visam monitorar, analisar e, quando necessário, informar a respeito de ciberameaças a partir do processamento de dados não estruturados, especialmente obtidos de mídias sociais. Esses dados são extraídos de fontes como o *Facebook* e *Twitter*, serviços de bate-papo como o IRC, serviços de compartilhamento de texto como *Pastebin*, *Piratepad* e *Github Gist* entre outros (CAMPIOLO, 2016; BENJAMIN, 2016). É o caso do *Hórus CEWS*, sistema de monitoramento e detecção de ameaças criado pelo *Grupo de Trabalho Early Warning System* (GT-EWS)[1] da Rede Nacional de Pesquisa (RNP)[2].

Este trabalho se insere no contexto do GT-EWS, uma vez que analisa informações advindas de uma fonte de dados não estruturados, a saber, o *Pastebin*, um serviço de compartilhamento de texto puro criado em 2002. Estas análises visam compreender as postagens do *Pastebin* no contexto de cibersegurança e gerar alertas de acordo com sua possível relevância, para serem usados de forma proativa ou reativa.

[1] <http://gtews.ime.usp.br/> Acesso em 11/11/2018
[2] <https://www.rnp.br/> Acesso em 11/11/2018

Os serviços de compartilhamento de texto permitem que textos simples, geralmente sem formatação, sejam compartilhados, em geral através de uma URL gerada. Via de regra, este tipo de serviço provê também o anonimato do usuário, o que pode facilitar a divulgação de materiais obtidos de forma ilícita (BRIAN, 2005).

O *Pastebin* foi escolhido como alvo de estudo deste trabalho por ter sido o primeiro *site* para compartilhamento de texto puro que se tem notícia e, mesmo depois de mais de 15 anos de sua criação, ser amplamente utilizado. É mostrado por Castelluccia et al. (2013) que diversas informações, obtidas por *hackers* são postadas no *Pastebin*. Já Heffelfinger (2013) mostra que existem informações, publicadas no *Pastebin*, que podem levar a identificação de alvos de ataques *hackers*.

1.1. Justificativa

Além dos trabalhos já citados, diversos autores abordam a presença de informações sensíveis no *Pastebin*. Tais informações incluem, mas não se limitam a e-mails, senhas, endereços, números de cartões de crédito, *posts* de grupos *hackers* e hacktivistas, endereços de alvos (*Internet Protocol* (IP) e URLs), vendas de informações, entre outros. É possível inclusive usar o *Pastebin* para disseminação de códigos maliciosos (BRIDGE, 2014; SHIFRIN, 2017; HOTFORSECURITY, 2011; AMISSAH, 2014; WARREN; LEITCH, 2013; SHAKARIAN et al., 2015).

Informações oriundas de *keyloggers* foram encontradas no *Pastebin* a partir de maio de 2010. O primeiro caso conhecido na literatura foi o do *Trojan.Keylogger.PBin.A*, que guardava informações a respeito do navegador utilizado, a URL visitada e as teclas pressionadas, entre colchetes (KONTAXIS et al., 2011).

Já existiram, também, ações judiciais envolvendo o *Pastebin*, motivadas pelo compartilhamento de dados obtidos de forma ilícita de empresas e requisitando que o *Pastebin* provesse informações que auxiliassem na confirmação da origem das informações (ATLANTA DIVISION, 2011).

A despeito de diversas ferramentas monitorarem o *Pastebin*, especialmente em busca de vazamentos de informações, não existem análises mais aprofundadas sobre como tornar tal monitoramento mais eficiente, ou estudos que caracterizem as postagens contendo informações relevantes para cibersegurança.

1.2. Objetivos

Objetiva-se nesta monografia realizar a extração e análise de dados do *Pastebin* que sejam relevantes no contexto de cibersegurança, visando a identificação de ameaças (antecipadas ou não) e de informações obtidas ilicitamente por meio de ataques cibernéticos ou *doxing*

(disponibilização de informações pessoais de um alvo específico).

Os objetivos específicos são:

- Caracterizar as postagens públicas do *Pastebin*;
- Identificar características nas postagens que possibilitem a classificação das informações em relevantes ou irrelevantes para o contexto de cibersegurança;
- Criar uma base de inteligência para cibersegurança a partir das características identificadas nas postagens.

1.3. Contribuições

As contribuições deste trabalho se dividem em duas categorias:

- Científica:
  - Identificação de padrões que permitam evidenciar postagens relevantes para a cibersegurança de modo que administradores de redes possam tomar decisões com maior velocidade, visando a proteção dos sistemas ou reações em caso de comprometimento;
  - Disponibilização de uma base de inteligência, criada a partir dos padrões descobertos por este trabalho, para análise de dados oriundos de ferramentas de compartilhamento de texto.
- Tecnológica:
  - Implementação de um sensor para coleta de informações relevantes para a cibersegurança do *Pastebin*.

1.4. Organização do trabalho

Esta monografia está organizada da seguinte maneira: o Capítulo 2 apresenta conceitos importantes para a compreensão deste trabalho. São apresentados, no Capítulo 3, os trabalhos relacionados. A metodologia é detalhada no Capítulo 4. Os resultados são mostrados e discutidos no Capítulo 5. No Capítulo 6 são apresentadas as conclusões e trabalhos futuros. O Apêndice A apresenta uma lista de URLs relevantes de ferramentas citadas neste trabalho. No Apêndice B são listados exemplos de postagens relacionadas com cibersegurança encontrados no *Pastebin*. O Apêndice C explica a nomenclatura adotada para as expressões regulares. No Apêndice D são apresentadas informações detalhadas sobre as análises desenvolvidas. Por fim, o Apêndice E detalha as configurações dos algoritmos usados neste trabalho.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta conceitos importantes para compreensão deste trabalho. A Seção 2.1 apresenta termos comuns para a cibersegurança. A Seção 2.2 destaca alguns ciberataques comuns e dá evidências desses ataques no *Pastebin*. A Seção 2.3 apresenta e discute diversas ferramentas de compartilhamento de texto puro, especialmente o *Pastebin*, foco deste trabalho (Subseção 2.3.1). A Seção 2.4 apresenta os algoritmos de classificação utilizados. Por fim, a Seção 2.5 apresenta conceitos relativos ao processamento de linguagem natural, que foram aplicados durante a análise dos dados.

2.1. Cibersegurança

Esta Seção apresenta definições importantes utilizadas neste trabalho. A Subseção 2.1.1 detalha o significado de cibersegurança,enquanto a Subseção 2.1.2 contém uma lista de termos associados à cibersegurança recorrentes nesta monografia.

2.1.1. Definição

Cibersegurança ou Segurança Cibernética (do inglês *cybersecurity*) é a parte da segurança da informação que refere-se aos métodos para proteção de informações no ciberespaço, em redes ou sistemas computacionais evitando seu roubo, alteração ou comprometimento através de acesso digital. Estratégias de cibersegurança incluem gerenciamento de riscos, gerenciamento de identidades e de incidentes (CAMURCA, 2017; TECHOPEDIA, 2017e; BISHOP, 2004; ISO, 2009). A cibersegurança baseia-se em três pilares: confidencialidade, disponibilidade e integridade.

Confidencialidade: garante que os dados só serão acessados por aqueles que devem ter conhecimento deles. Um dos métodos para garantir a confidencialidade é a criptografia

– que garante que os dados não sejam legíveis exceto para aqueles que possuam as chaves apropriadas. No entanto, se essas chaves são adquiridas de alguma forma, atacantes podem ter acesso ao conteúdo outrora protegido por elas. Comumente são encontrados no *Pastebin* vazamentos, maliciosos ou não, de senhas e chaves privadas (BISHOP, 2004; MIRANTE; CAPPOS, 2013; BRENGEL; ROSSOW, 2018).

Integridade: garante a não alteração indevida de dados. Isso pode significar não permitir que alguém não autorizado altere os dados ou que alguém autorizado altere-os de um modo indevido. Pode ser dividida em detecção (apenas detecta onde ou quando houve falha na integridade) e prevenção (estabelece mecanismos para impossibilitar a quebra de integridade, como, por exemplo, impedir acesso não autorizado) (BISHOP, 2004).

Disponibilidade: garante a disponibilidade do serviço quando desejado. Uma das principais formas de atacar a disponibilidade é sobrecarregar um serviço, o que é extremamente difícil de diferenciar de requisições autênticas (BISHOP, 2004).

2.1.2. Conceitos-chave

A taxonomia na área de Cibersegurança é bem vasta. A seguir, são apresentados alguns conceitos-chave para a compreensão deste trabalho:

- Vulnerabilidades: são falhas, oriundas do projeto, da implementação, manutenção, operação ou configuração de sistemas computacionais. Uma vez exploradas, estas vulnerabilidades violam as bases da segurança da informação, permitindo acesso não autorizado, por exemplo, o que é chamado "ataque". O ato de explorar tais vulnerabilidades também é costumeiramente chamado *exploit* (BISHOP, 2004; CERT.BR, 2012).
- *Hackers*: refere-se a pessoas ou grupos que ultrapassam barreiras de segurança para acessar informações não autorizadas ou restritas, encontrando e explorando falhas e vulnerabilidades (TECHOPEDIA, 2017c). Se dividem, principalmente, em três grupos:
  - *black hat hackers* (*hackers* de chapéu preto – tradução livre): utilizam suas habilidades para causar danos, por meio do vazamento de informações, destruição de dados e outras atividades criminosas;
  - *grey hat hackers* (*hackers* de chapéu cinza – tradução livre): realizam atividades *hacker* ilegais principalmente para provar suas habilidades;
  - *white hat hackers* (*hackers* de chapéu branco – tradução livre): utilizam suas habilidades para proteger redes e sistemas computacionais. Também chamados de *hackers* éticos (*ethical hackers*);
- *Cracker*: frequentemente associados aos *black hat hackers*, *crackers* são indivíduos que invadem dispositivos computacionais ou redes para realizar atividades danosas ou apenas para provar as suas capacidades. Suas atividades incluem invasões para diversão

ou desafio, causar danos a um alvo específico (situação em que o *cracker* costuma se fazer conhecido pela vítima, embora não com sua identidade real), sabotagem, espionagem e roubo de informações para chantagem (TECHOPEDIA, 2017a).

- Hacktivismo (do inglês *hacktivism*): é a utilização de ações *hacker* para deixar mensagens de cunho político ou social, chamando a atenção para uma causa ou situação. Pessoas que o praticam são chamadas hacktivistas (do inglês *hacktivists*) (TECHOPEDIA, 2017f).
- *Key Logger*: são programas que coletam dados sobre as teclas pressionadas no dispositivo da vítima. Os mais avançados são capazes de filtrar a informação (por exemplo, após a entrada de uma certa URL) ou estruturar os dados. Estes dados são enviados, em seguida, para um ambiente chamado *dropzone* (KONTAXIS et al., 2011).
- Códigos de exploração (do inglês *exploit*): termo genericamente utilizado para qualquer forma usada por um *hacker* para acessar informações de forma não autorizada. Também pode se referir ao ato de acessar de forma não autorizada um dispositivo ou rede, através, por exemplo, de vulnerabilidades (BISHOP, 2004; TECHOPEDIA, 2017d).
- *Botnet*: são conjuntos de computadores infectados por códigos maliciosos que podem ser controlados remotamente, usualmente utilizados em ataques de negação de serviço distribuídos - *Distributed Denial of Service* (DDoS) (Subseção 2.2.1). Tais computadores são comumente chamados de "zumbis" (KASPERSKYLAB, 2017; AVAST, 2017).
- *Data breach*, *data leak* ou *data spill*: consiste no vazamento não autorizado, intensionalmente ou não, de dados, em geral de organizações. Estes dados podem ter sido obtidos de forma ilegal, por meio, por exemplo, da exploração de vulnerabilidades. Afeta diretamente a imagem das companhias envolvidas, visto que em geral os dados vazados são confidenciais, como números de cartões de créditos e documentos pessoais dos usuários (TECHOPEDIA, 2017b; TRENDMICRO, 2017).
- *Bitcoin*: O *Bitcoin* (BTC) é uma criptomoeda descentralizada, gerada computacionalmente através de um processo chamado mineração. Esta moeda vem ganhando mais e mais adeptos e seu valor tem aumentado cada vez mais (HUANG et al., 2014; BITCOIN, 2017; EXCHANGEWAR, 2017).

Outros termos eventualmente são utilizados neste trabalho, porém não são tão recorrentes e são explicados à medida em que são utilizados.

2.2. Ciberataques

Esta seção trata sobre alguns tipos de ciberataques, evidenciando a presença de informações relativas ou oriundas dos mesmos no *Pastebin*. A Subseção 2.2.1 aborda ataques DoS. Em seguida, a Subseção 2.2.2 explica o que é a desfiguração de páginas. Por fim, a Subseção 2.2.3 aborda os vazamentos de informações pessoais.

2.2.1. Negação de serviço (DoS)

Segundo CERT.BR (2012), a negação de serviço ou DoS consiste na sobrecarga de um serviço, visando exaurir seus recursos, levando assim a sua indisponibilidade. Quando realizada de forma coordenada por diversos dispositivos, recebe o nome de negação de serviço distribuído ou DDoS.

Ainda segundo CERT.BR (2012), este tipo de ataque não objetiva obter informações e sim causar indisponibilidade ao alvo, de modo que todos que dependam legitimamente dos serviços fornecidos não possam ser atendidos. Isto se dá por meio de técnicas como o envio de excessivas requisições ao alvo, impedindo que todas sejam atendidas (devido, por exemplo, ao número de conexões simultâneas, utilização de todo processamento, memória ou espaço disponível), pela geração de tráfego acima da capacidade da rede ou pela exploração de vulnerabilidades que tornem o serviço indisponível. Este tipo de ataque tem se tornado cada vez mais comum, com a quantidade de dados enviada cada vez maior (CIO, 2016).

No *Pastebin* é possível encontrar códigos fontes para realização de ataques DoS e DDoS, informações a respeito dos alvos e, até mesmo, tutoriais, como os divulgados pelo grupo hacktivista *Anonymous* (G1, 2016).

2.2.2. Desfiguração de página (*defacement*)

A desfiguração de páginas, também chamada de *defacement* ou pichação, consiste na alteração indevida ou sem permissão do conteúdo de páginas *Web*. As principais formas usadas por um atacante, neste caso conhecido como *defacer*, para realizar as desfigurações são: exploração de erros da aplicação *Web*, de vulnerabilidades no servidor, do interpretador/compilador da linguagem de programação ou dos pacotes usados no desenvolvimento da aplicação, invasão do servidor onde a aplicação *Web* se encontra e alteração direta dos arquivos ou furto da senha das interfaces usadas para controle da aplicação *Web* (CERT.BR, 2012).

No *Pastebin* é possível encontrar divulgação de páginas que sofreram *defacement* e, muitas vezes, os motivos desses ataques. Um exemplo foi a divulgação de alvos e informações relativas de cerca de cem páginas israelenses, atacadas por um grupo Palestino, em 2015 (WAQAS, 2015).

2.2.3. *Doxing*

Doxing ou *doxxing* vem de *dropping documents* e se refere a disponibilização de informações pessoais como nomes, endereços, dados governamentais ou fotos da pessoa ou de parentes, em um meio de fácil acesso. Tal ato não possui obrigatoriamente conotação maliciosa ou intenção de provocar danos, mas frequentemente o faz. Eventualmente, o vazamento desses dados é realizado pela própria vítima, por desconhecimento ou incompreensão do uso de ferramentas,

como no caso mostrado por Brengel e Rossow (2018), onde percebeu-se que diversas pessoas postavam chaves privadas de suas carteiras de BTC no *Pastebin*, em mensagens de erro ou confirmação. Mesmo que muitos dos dados sejam públicos, reuni-los e publicá-los é considerado um tipo de *doxing*. Tais informações podem ser reunidas a partir de dados disponíveis na Internet, *sites* governamentais, informações vazadas de companhias ou outros meios ilícitos ou não.

O *doxing* pode ser usado para expor ações consideradas incorretas, humilhação, intimidação, ameaça, punição ou simples perda do anonimato. É uma ferramenta largamente utilizada por grupos *hacktivistas*, expondo informações que levem a identificação de seus alvos e, muitas vezes, conclamando pessoas a persegui-los, virtual ou fisicamente (DOUGLAS, 2016).

Em Moyer (2016) são apresentados diversos casos de *doxing* que foram postados no *Pastebin*. Este tipo de ataque está frequentemente associado ao hacktivismo, como no caso do vazamento de informações pessoais de diversos membros da *Ku Klux Klan* (KKK), conhecido grupo extremista estadunidense (MOYER, 2016). Outro caso de hacktivismo é estudado em Pendergrass e Wright (2014): um indivíduo, parte do grupo hacktivista *Anonymous*, sob a identidade de *KnightSec* promoveu uma série de ações contra diversos grupos e pessoas. A mais conhecida delas em suporte a uma vítima de estupro de Steubenville, Ohio e contra seus estupradores. Tais ações tornaram o caso público e notório, influenciando a rapidez de sua investigação. O *Pastebin* foi amplamente utilizado por tal grupo: foram postadas informações pessoais dos alvos como nome, número telefônico, número do seguro social, dados sobre familiares, entre outras.

2.3. Ferramentas de compartilhamento de texto puro

As ferramentas de compartilhamento de texto possibilitam que textos, originalmente sem formatação, sejam compartilhados na *Web*, em geral através de uma URL gerada. Estes textos costumam ser chamados *pastes*. Este tipo de ferramenta tornou-se bastante utilizada e é um exemplo de fontes de dados não estruturados (SQUIRE; SMITH, 2015; CAMPIOLO, 2016).

A primeira dessas ferramentas a ser desenvolvida e ganhar espaço e, ainda hoje, a ser uma das mais utilizadas, é o *Pastebin* (Seção 2.3.1). No entanto, após sua criação, diversas ferramentas foram desenvolvidas, cada uma com suas particularidades. Uma seleção dessas ferramentas, identificadas nos trabalhos de Campiolo (2016) e Gaikar (2013) são apresentadas a seguir.

- *Piratepad*: Possui um sistema de *chat* integrado em cada *paste*. Alterações podem ser feitas por todos que possuam a URL e serão mantidas síncronas. Apresenta opções para importação e exportação dos conteúdos dos *pastes* em diversos formatos. Possibilita

configuração de diversos parâmetros, como inclusão de marcações (hiperligações, listas ordenadas, listas não ordenadas, etc) no texto, alterar tipo da fonte e diversas opções visuais do texto.

- *Github Gist*: Mantido pelo *Github*, guarda muitas similitudes com este: a capacidade de atribuir estrelas a *gists* (nome dado aos textos compartilhados) ou efetuar *fork* (cópia para edição própria) dos mesmos, realizar comentários, propor edições e *issues*. Permite que os *gists* sejam compartilhados de forma pública, não listadas ou privadas. Pode ser associado à conta do *Github*.
- *Zerobin*: focado na privacidade, criptografando todos os tipos de postagens no navegador com *Advanced Encryption Standard* (AES) de 256 bits, o *Zerobin* permite *pastes* de até 2 MB, um sistema de discussão criptografado que pode ser ativado para cada *paste* e *highlighting syntax* para 54 linguagens de programação. O código fonte do serviço também está disponível no *Github*[1]. Embora ainda esteja disponível, uma versão, compatível e aprimorada desta ferramenta, com maior suporte a *syntaxe hilighting*, possibilidade de postar com ou sem encriptação, uso de *HyperText Transfer Protocol Secure* (HTTPS) ao invés de *HyperText Transfer Protocol* (HTTP) e outras melhorias, chamada *Privatebin*, é a versão oficial e sob manutenção. Seu código fonte também está disponível no *Github*[2].
- *Chopapp*: ferramenta voltada para o compartilhamento de trechos de código para revisão. É possível adicionar anotações a cada linha ou grupos de linhas. Suporta *highlighting syntax* e formatação do código, que podem ser sugeridas pelo *site*. Possui opção para comentários relativos as anotações e ao código. Também suporta que o código seja importado de algum repositório.
- *Snipt*: é, principalmente, um repositório para pequenos trechos de código fonte. Para uso, requer a criação de uma conta, que é gratuita.
- *Codepad*: permite a criação, cópia, exclusão e adição de projetos com trechos de código fonte. Suporta as linguagens *C*, *C++*, *D*, *Haskell*, *Lua*, *ACaml*, *PHP*, *Perl*, *Plain text*, *Python*, *Ruby*, *Scheme* e *Tcl*. Uma vez que um trecho de código fonte seja inserido, é gerada uma URL. Ao acessar esta URL, para além do código fonte escrito, o resultado da execução do código fonte também é exibido. Os *pastes* podem ser privados ou públicos.
- *dPaste*: Extremamente simples, possuindo as opções de manter o *paste*, sempre público, disponível de um dia a um ano, não apresenta nenhum recurso diferente em relação aos anteriormente citados.

As ferramentas supracitadas permitem compreender o funcionamento de ferramentas para compartilhamento de texto puro bem como vislumbrar o tipo e formato dos dados,

[1] <https://github.com/sebsauvage/ZeroBin> Acesso em 11/11/2018
[2] <https://github.com/PrivateBin/PrivateBin> Acesso em 11/11/2018

não estruturados, nelas contido. Uma vez que o *Pastebin* é o foco dessa monografia, ele é detalhado a seguir.

2.3.1. O *Pastebin*

O *Pastebin* (Figura 2.1) é um serviço de compartilhamento de texto puro, criado em 2002. Os textos postados neste *site* frequentemente recebem o mesmo nome, ou seja, são chamados de *pastebin*. A partir disso originou-se o termo comumente usado: *paste* (Figura 2.2). Originalmente, o objetivo era apenas o compartilhamento trechos de código fonte ou arquivos de configuração. Com o passar do tempo, porém, começou a ser usado para postagem de materiais ilícitos e/ou vazamento de dados (BRIAN, 2005).

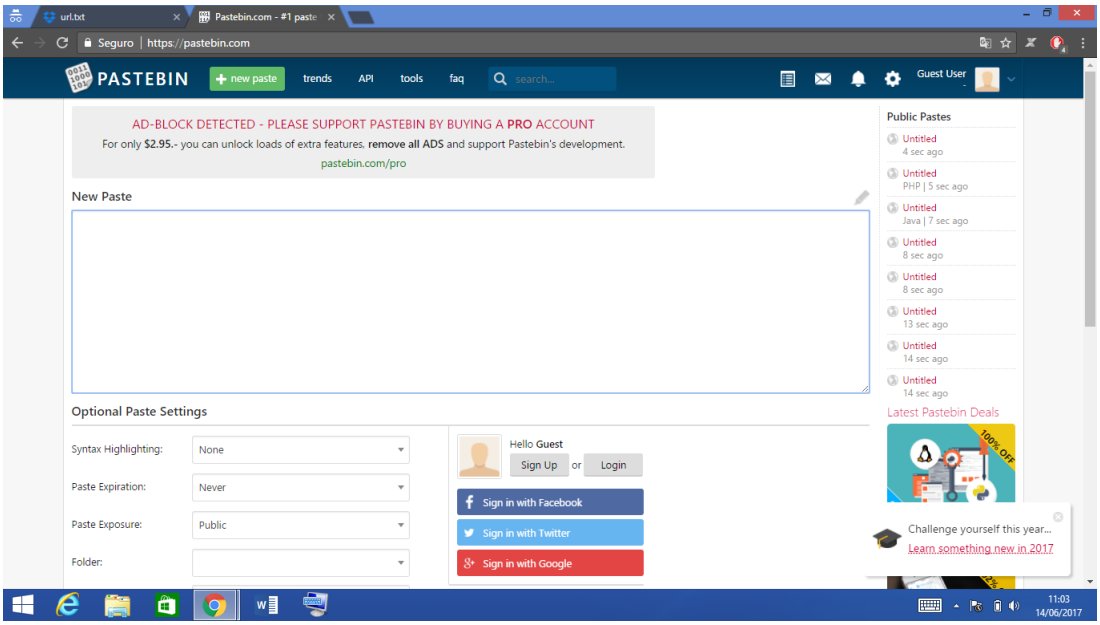

Figura 2.1. Pastebin.com - #1 paste tool since 2002!

O *Pastebin* possibilita que textos (*pastes*) sejam postados e, em consequência, uma URL é gerada. Cada *paste* é associado a uma chave (*key*) de 7 ou 8 caracteres, formadas por letras (maiúsculas ou minúsculas) e números, em qualquer ordem. A URL final é constituída do endereço do *site* (<https://www.pastebin.com/>) e a chave gerada. O *paste* exibido na Figura 2.2, por exemplo, pode ser acessado pelo endereço <https://pastebin.com/LcTY9fRp>. Uma vez que a chave segue regras não descritas no *site* e que as possibilidades de combinações são inúmeras, não é prático tentar encontrar ou extrair informações com base em chaves aleatórias ou mesmo de forma sequencial. A URL é, portanto, anônima.

Os *pastes* podem ser configurados para ficarem disponíveis por dez minutos, uma hora, um dia, um mês ou "nunca expirar" (KONTAXIS et al., 2011). Podem ser configurados também para fornecer *highlighting syntax*. Os *pastes* podem ser públicos – todos podem acessar – não listáveis – o que significa que, embora público, não aparecem na lista de

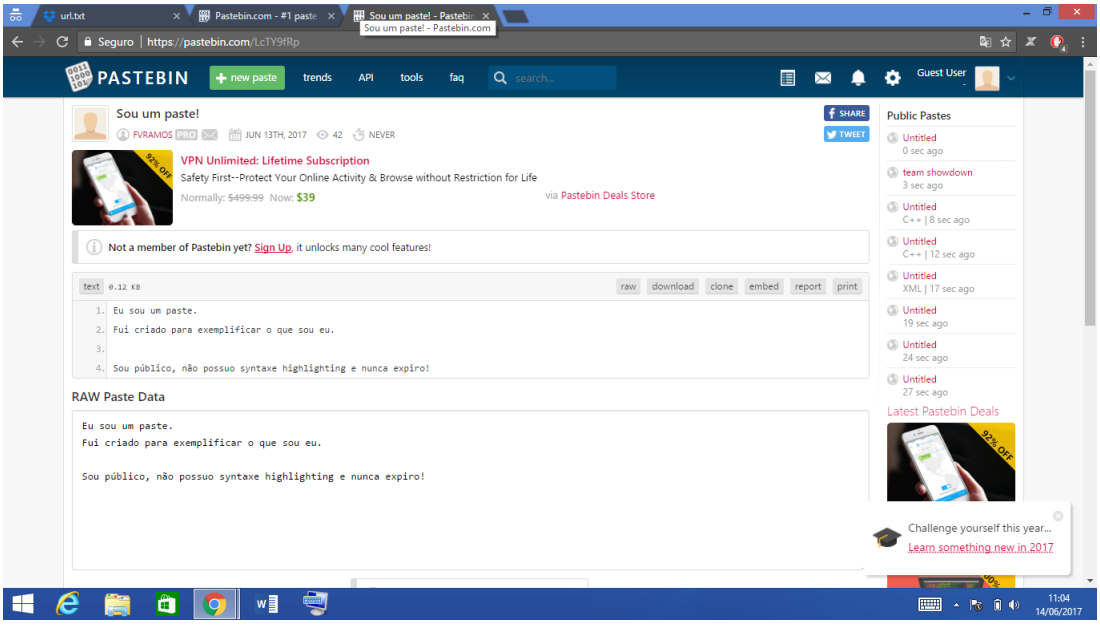

Figura 2.2. Sou um paste! - Pastebin.com

pastes recentes – e privados – apenas acessáveis pela URL. Por padrão, no *Pastebin*, um *paste* é público, nunca expira e não possui *syntaxe highlighting*. Também são armazenadas informações sobre a data de publicação e quantidade de acessos em cada *paste*.

O *Pastebin* possuía uma opção para exibição dos *Trends*, ou seja, os *pastes* mais acessados[3]. No entanto, esta opção foi removida, sem previsão de retorno, em junho de 2018, devido a "abuso" – a administração do serviço não qualificou que tipo de abusos[4]. Os *trends* dividiam-se em "agora" (o padrão: *pastes* mais acessados nas últimas 72 horas), "últimos sete dias", "últimos 30 dias", "últimos 365 dias" e "sempre". As listas eram atualizadas a cada hora e exibiam unicamente os *pastes* com visibilidade pública. É possível, também, visualizar os últimos *pastes* postados, através do *link* <https://www.pastebin.com/archives>.

O *Pastebin* possibilita acesso por uma *Application Programming Interface* (API) HTTPS[5]. Essa API permite criação e deleção de *pastes*, recuperação do conteúdo de um *paste*, lista de *pastes* criados por um usuário e informações sobre um usuário em particular. Uma segunda API, chamada de *scraping API*[6], possibilita aceder a outras informações, porém requer registro e pagamento, como *Pro user*[7]. Esta API permite adicionar um endereço IP à lista de acesso, ou seja, permite que este endereço seja utilizado para capturar as informações. Utilizando-se desta API é possível requisitar até 250 dos *pastes* mais recentes (*most recent pastes*).

Existem algumas restrições quanto aos meios de acesso ao *Pastebin*. Elas se dão, principalmente, pelo tamanho, quantidade e privacidade dos *pastes*. Três são as formas de

[3] <https://pastebin.com/trends> Desativada em Jun/2018
[4] <https://pastebin.com/CfNR87Jr> Acesso em 06/11/2018
[5] <https://pastebin.com/api> Acesso em 11/11/2018
[6] <https://pastebin.com/scraping> Acesso em 11/11/2018
[7] <https://pastebin.com/pro> Acesso em 11/11/2018

acesso ao *Pastebin*: sem autenticação, o que significa que não é necessário criar nenhuma conta (não permite *pastes* privados), conta gratuita, que dá acesso a API do *Pastebin* e, por fim, a conta *Pro*, que permite acesso a API de *scraping*. Os *pastes* postados por usuários não autenticados apresentam, como autor, apenas a palavra *Guest* (Convidado). No caso dos usuários autenticados, podem escolher compartilhar os *pastes* como convidados ou com seu nome de usuário. No caso do compartilhamento com o nome de usuário, é possível ter acesso a todos os *pastes* postados por este usuário, através da página de *pastes* do usuário. No entanto, ao recuperar um *paste* com a API, não é informado o nome do usuário, mesmo que o *paste* o exiba. É possível criar o usuário utilizando *e-mail*, *Facebook* e *Google*.

2.4. Algoritmos de classificação

No presente trabalho, algoritmos de classificação são usados em *pastes* para a separação em "relevantes" e "irrelevantes" no contexto de cibersegurança. Os algoritmos selecionados foram de árvore de decisão, *naive bayes*, *Support vector machines* (SVM) e KNN. Foram selecionados por serem citados na literatura como algoritmos adequados para aplicações envolvendo análise de dados textuais (AGGARWAL; ZHAI, 2012).

2.4.1. Árvore de decisão (*Decision tree*)

Uma árvore de decisão é, basicamente, a decomposição hierárquica dos dados. Para tal, cada atributo ou condição utilizada para a classificação é utilizado para a divisão hierárquica. No caso de dados em texto, essa condição pode ser a presença ou ausência de um termo. A divisão hierárquica é feita recursivamente, até que uma quantidade determinada de registros sejam alcançados nos nós folha. Existindo um grande desbalanceamento entre as classes para serem classificadas, parte das folhas podem ser removidas ("podadas") (AGGARWAL; ZHAI, 2012).

Em relação à análise de dados textuais, os tipos comuns de divisão para os nós, ainda segundo Aggarwal e Zhai (2012) são: divisão por atributos simples, que utiliza palavras-chave especialmente relevante para a classificação ou mesmo frases; divisão baseada em similaridade de diversos atributos, que classifica os dados em relação a sua similaridade com um dado conjunto de palavras-chave; divisão baseada em discriminação de diversos atributos, que ordena os dados com base em um vetor de posições e identifica um ponto deste vetor para dividi-lo.

2.4.2. *Naive Bayes*

O *naive bayes* modela a distribuição dos documentos em cada classe, através de modelos probabilísticos, isto é, não são levadas em conta as posições das palavras no texto

(AGGARWAL; ZHAI, 2012).

São dois os modelos comumente utilizados para análise textual, segundo Aggarwal e Zhai (2012): modelo multivariável de *Bernoulli*, que usa a presença e ausência de palavras num documento de texto como características; modelo *Multinomial*, que utiliza a frequência das palavras para gerar o modelo, de modo que a probabilidade de um dado documento pertencer a uma classe é dada pela probabilidade de existência de cada palavra.

2.4.3. *SVM*

Classificadores SVM objetivam a construção de um hiperplano ótimo, de modo que ele possa separar diferentes classes de dados com a maior margem possível (AGGARWAL; ZHAI, 2012). São especialmente adequados para análises textuais, levando-se em conta que lidam bem com correlação entre termos, onde, muitas vezes, alguns deles não são relevantes para o contexto, caso comum em análises de texto.

2.4.4. KNN

O algoritmo de classificação KNN calcula a distância de cada amostra para suas amostras vizinhas objetivando identificar a que grupo ela pertence. Durante a fase de treinamento do algoritmo, são criados grupos de documentos pertencentes a mesma classe. As distâncias são, então, testadas em relação a estes grupos (AGGARWAL; ZHAI, 2012). A escolha das características e representação dos documentos é extremamente importante para o KNN, já que termos muito frequentes muitas vezes não guardam real relação com as classes a que pertencem.

2.5. Processamento de linguagem natural

O Processamento de Linguagem Natural (PLN) (do inglês *Natural Language Processing* (NLP)) é abordado em Campiolo (2016) e Benjamin (2016) como uma importante ferramenta para a detecção de informações relativas à cibersegurança em fontes de dados não estruturados.

A identificação de entidades, como alvos (através de endereços IP) ou URL, nomes de possíveis alvos e outras informações é realizada através de expressões regulares em Campiolo e Batista (2015).

Expressões regulares são métodos formais para descrição de padrões de texto (JARGAS, 2001). Através de seu uso, é possível identificar, por exemplo, uma URL genericamente, descrevendo seu padrão. Por conta disso, torna-se extremamente valiosa para a identificação de entidades, porém podem existir erros quando um mesmo padrão pode se referir a mais de uma coisa. Por exemplo, quatro números separados por pontos (“127.0.0.1”, por exemplo) pode se referir a um endereço IP ou a um número de versão. Quando um

padrão descrito por uma expressão regular é encontrado num dado texto, isto é chamado "casamento".

Outra forma de encontrar padrões de informações relevantes é através da análise de relações sintagmáticas, paradigmáticas e *collocations*. Relações sintagmáticas identificam palavras que são encontradas frequentemente num mesmo contexto (como nadar e poço). Já as relações paradigmáticas identificam palavras que podem ser substituídas por outras sem prejuízo do sentido, para além de ligar palavras por seu significado (nadar e água). *Collocations* permitem identificar relação entre duas ou três palavras (bigramas ou trigramas) quanto a seu aparecimento em conjunto em textos (MANNING; SCHüTZE, 1999). Estes conceitos foram utilizados para obter informações relativas à cibersegurança por Campiolo e Batista (2015).

Bigramas, trigramas e quadrigramas são, respectivamente, relações entre duas, três e quatro palavras que apareçam em sequência. Isto permite aferir características importantes sobre um determinado grupo de palavras.

2.6. Considerações gerais

Este capítulo abordou conceitos relevantes para a compreensão desta monografia: termos recorrentes na área de Cibersegurança que são utilizados nos próximos capítulos e diversos tipos de ciberataques, destacando sua relação com o *Pastebin*. As diferenças entre algumas ferramentas para compartilhamento de texto puro foram detalhadas e, especialmente, o *Pastebin* (foco deste trabalho) e seu funcionamento foram explicados em detalhes. Por fim, foram abordados conceitos relativos ao PLN como expressões regulares, utilizados neste trabalho para identificação de entidades, *pastes* com informações sensíveis e reconhecimento de padrões em *pastes*. O próximo capítulo aborda trabalhos relacionados ao apresentado nesta monografia.

Capítulo 3

Trabalhos Relacionados

Este capítulo discute trabalhos relacionados a esta monografia. Os trabalhos aqui abordados se dividem em duas partes: na Seção 3.1 são tratados trabalhos originados da proposição do arcabouço para análise de mensagens a respeito de cibersegurança em fontes não estruturadas proposto em Campiolo (2016). Esses trabalhos não abordam o *Pastebin* diretamente, no entanto discutem métodos e técnicas que podem ser aplicadas ao *Pastebin*. Na Seção 3.2 são abordadas ferramentas que extraem informações diretamente do *Pastebin*.

3.1. Arcabouço para extração e análise de mensagens relativas a cibersegurança

Em Campiolo (2016) é proposto um arcabouço para análise e extração de informações relevantes para cibersegurança de fontes não estruturadas. São abordadas questões importantes para cada fase, como a coleta dos dados, pré-processamento, classificação de dados para análise, análise dos dados, extração das informações e divulgação dos alertas. Também apresenta a aplicação do arcabouço para o *Twitter* (Subseção 3.1.1) e o IRC (Subseção 3.1.2), discutindo seus resultados.

O trabalho aqui apresentado está inserido neste contexto (extração e análise de informações a respeito de cibersegurança em fontes de dados não estruturados), uma vez que visa a aplicação do arcabouço proposto por Campiolo (2016) para o *Pastebin*.

3.1.1. *Twitter*

São apresentados em Santos et al. (2013), SANTOS et al. (2012) a extração e análise de mensagens relativas a cibersegurança no *Twitter*.

Em SANTOS et al. (2012) é realizada a análise de um conjunto de mensagens do *Twitter*. Para tal, são seguidas quatro etapas:

- Capturar, armazenar e indexar, no *Lucene*[1] – que indexa textos para futuras buscas e cálculos de similaridade com base em um *score* – mensagens a respeito de cibersegurança postadas no *Twitter*: desenvolveu-se um *software* que utiliza uma API do *Twitter*[2] para monitorar e realizar *download* de mensagens, a cada 1 minuto durante 21 dias. Para filtrar as mensagens relativas a cibersegurança, foram usadas palavras-chave da área. Ao término da execução da coleta constatou-se a presença de diversas mensagens a respeito de cibersegurança no *Twitter*.
- Capturar, armazenar e indexar mensagens a respeito de cibersegurança oriundas de *sites* especializados: utilizando-se de uma API especializada[3] e de um *Web crawler* (WEB. . . , 2017) foram coletados 3.988 *feeds* a respeito de cibersegurança durante um mês.
- Agrupar os *tweets* por similaridade e selecionar os grupos com mais de 10 *tweets*: os *tweets* são agrupados com base no *score* de similaridade do *Lucene*, a partir da comparação de cada *tweet* com todos os *tweets* capturados. São armazenados o *score* de similaridade e número de mensagens por grupo.
- Correlacionar os *tweets* relevantes com as mensagens de *sites* especializados: os *tweets* são correlacionados com as mensagens dos *feeds* e a data de publicação são comparadas.

Este trabalho mostrou a existência de mensagens a respeito de cibersegurança, que se disseminam rapidamente no *Twitter*. Uma expansão deste trabalho foi apresentada em Santos et al. (2013) onde é visada a identificação dessas mensagens apresentando análise semântica mais apurada, bem como filtro de idioma, análise das URLs e criação de uma lista de mensagens que representam *spams*, ou seja, mensagens que, a despeito de primeiramente serem percebidas como relevantes, não o são. Estas mensagens foram usadas para extrair termos e padrões que permitam filtrar as mensagens associadas à segurança, de modo que as não relevantes sejam retiradas da lista.

3.1.2. IRC

A presença de informações relativas a cibersegurança em canais da rede IRC é analisada em Campiolo e Batista (2015). Para tal, canais escolhidos por sua relação com cibersegurança e atividades *hacker* foram monitorados por um software. As interações e mensagens nestes canais foram salvas e analisadas. A partir das mensagens salvas, procedeu-se a normalização

[1] <http://lucene.apache.org/core/> Acesso em 11/11/2018
[2] <http://twitter4j.org/> Acesso em 11/11/2018
[3] Informa: <http://informa.sourceforge.net/> Acesso em 11/11/2018

e estatística descritiva dos dados (como quantidade de mensagens e usuários) e identificação das mensagens potencialmente relevantes.

As análises realizadas no intuito de identificar as mensagens relevantes foram: análise estatística, frequência de palavras, relações entre termos, conteúdo das URLs, identificação de entidades e correlação de padrões com outras fontes.

A análise de frequência de palavras permitiu a criação de bases de termos indicadores tanto de relevância quanto de irrelevância das mensagens. Foram gerados grupos como "gírias", "atividades maliciosas ou ameaças","xingamentos" entre outros.

Foi utilizado PLN para a identificação de relações paradigmáticas, ou seja, palavras que, substituídas por outras de mesma classe, mantém o mesmo sentido da frase, sintagmáticas, ou seja, relação das palavras e *colocations*, expressões compostas por dois ou mais termos para expressar algo. Este tipo de análise permite compreender os padrões apresentados em mensagens relevantes e, assim, expandir as buscas.

Também foram analisadas URLs encontradas. Isto foi feito a partir da análise do domínio de cada URL, remoção das que fizessem parte de domínios irrelevantes e subsequente análise manual. Também foram analisados os títulos dos *sites* acessados por meio da URL, em busca de termos associados à cibersegurança.

A identificação de entidades permite encontrar diversas informações que as análises anteriores podem não permitir. Isto pode ser feito através do uso, por exemplo, de expressões regulares. Foram identificadas entidades nas mensagens postadas na rede IRC que, juntamente com as outras formas de análise utilizadas, permitiam aferir a relevância ou não de mensagens.

O trabalho apresentado por Campiolo e Batista (2015) utiliza diversas formas de classificação, que vão desde busca por palavras-chave até PLN, passando por expressões regulares e análise de URL. Isto torna a análise profunda e permite identificar padrões, criando uma base de inteligência, como objetiva-se na presente monografia.

3.2. Ferramentas para coleta de dados do *Pastebin*

Existem ferramentas que fazem a coleta e análise de *pastes* do *Pastebin*. Entretanto essas ferramentas são limitadas, não se retroalimentam - não conseguem aprender novas características de *pastes* relevantes - ou são extremamente específicas, buscando apenas por alguma palavra ou endereço de e-mail dado.

Nesta seção são abordadas algumas ferramentas que analisam o *Pastebin*. A Subseção 3.2.1 apresenta a *Seek Data Leakage*, uma plataforma para extração de *pastes* de acordo com regras no *Pastebin*. A Subseção 3.2.2 discorre a respeito de uma das mais completas e detalhadas ferramentas que coletam informações a partir do *Pastebin*, denominada *Have I been pwned*, que realiza notificação de comprometimento de *e-mails* e domínios em vazamentos de dados. A Subseção 3.2.3 discorre a respeito do *Dumpmon*, ferramenta que analisa três

ferramentas de compartilhamento de texto puro, utilizando-se de expressões regulares. Por fim, a Subseção 3.2.4, que apresenta dados oriundos do *Pastebin*, mas não é documentado como é realizada a extração.

3.2.1. *Seek Data Leakage*

A *Seek Data Leakage* é uma plataforma *online* para análise de vazamento de dados. São usadas como fontes de dados o *Pastebin* e o *Shodan*[4] (um *site* que permite procura de dispositivos de *Internet of Things* (IoT)). O objetivo da plataforma desenvolvida é possibilitar a identificação e análise de vazamentos de dados nestas duas fontes (SOUSA et al., 2016).

Esta plataforma exige a criação de um usuário e, a partir disso, a adição de regras de busca, que podem se aplicar ao *Shodan* ou ao *Pastebin*. É possível também escolher a periodicidade de busca. Uma vez que algum dado relevante seja encontrado, haverá uma notificação, por e-mail ou no *site*, que poderá classificar a informação como vazamento de dados ou não. As regras criadas são termos de buscas simples. A ferramenta não aprende com os dados já coletados, e não cria, por exemplo, regras novas ou sugere regras relevantes. Tem capacidade de gerar relatórios levando-se em conta as classificações. Esta plataforma também usa uma expressão regular para identificar vazamentos de dados no formato **e-mail:senha** (como no exemplo "testerson@teste.com:t1s2o3n4"). Existem também usuários com poderes administrativos para gerenciar a plataforma, removendo usuários e regras, se necessário. É possível executar uma versão da plataforma. O código está disponível no *Github*[5].

Esta plataforma foi utilizada durante seis meses por 19 pessoas, sendo que 11 delas responderam um questionário desenvolvido pelos idealizadores do projeto que visava avaliar o uso e aceitação da ferramenta. Segundo a avaliação do questionário, a maioria mostrou interesse em utilizar a plataforma. No entanto, o sistema como um todo é pouco flexível, depende da criação de regras rígidas para a busca dos dados e não realiza nenhuma análise dos *pastes*. O foco dela é a criação de regras para monitorar possíveis vazamentos de dados relativos a, por exemplo, uma organização.

Para obter os *pastes* postados no *Pastebin*, esta plataforma realiza requisições constantes à página que lista os últimos *pastes* postados (<https://pastebin.com/archives>), extraindo o endereço do *paste* diretamente da tabela *HyperText Markup Language* (HTML) lá disponível. Esta técnica porém pode levar ao bloqueio do IP por excesso de requisições. Foi utilizada uma lista de *proxies*, para permitir requisições por múltiplos endereços, através da alternância dos endereços de *proxy*. Outra forma é a utilização da API do *Pastebin*. Todavia, para acessar esta API no modo que permite requisições sem bloqueio de IP, é necessário ser usuário *Pro*, ou seja, efetuar pagamento.

[4] <https://www.shodan.io/> Acesso em 11/11/2018
[5] <https://github.com/VSS29/SeekDataLeakage> Acesso em 11/11/2018

3.2.2. *Have I been pwned*

O *Have I been pwned* (Figura 3.1) é um projeto desenvolvido por Troy Hunt, um dos diretores regionais da *Microsoft*, mantenedor de um *site* e diversos cursos a respeito de cibersegurança (HUNT, 2017c, 2017a). O objetivo do *Have I been pwned* é permitir acesso rápido a vazamentos de dados de credenciais de e-mails ou domínios, de maneira simples e gratuita.

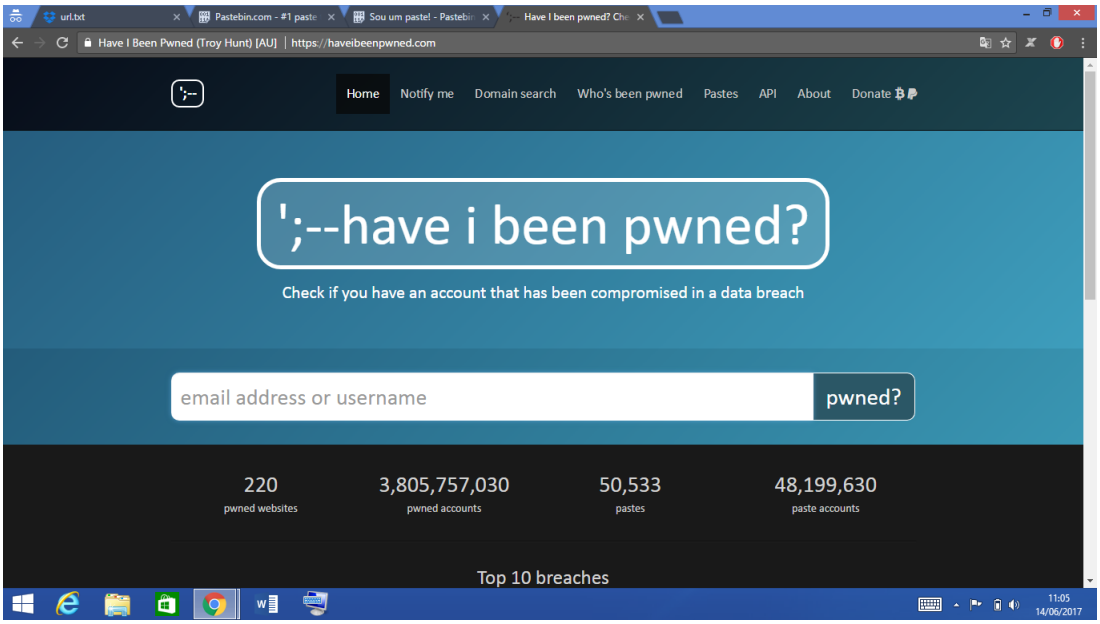

Figura 3.1. Have I been pwned? Check if your email has been compromised in a data breach

O *Have I been pwned* permite que sejam pesquisados e-mails ou domínios, desde que validado o direito de acesso de quem pesquisa por tais informações. Isto se dá através da confirmação do *e-mail* com um *link* enviado quando da inscrição no sistema de notificações que passa, então, a enviar *e-mails* de alerta caso o endereço de *e-mail* ou domínio conste de algum vazamento de dados.

Os dez maiores vazamentos de dados verificados pela plataforma são listados na tabela 3.1. O *Have I been pwned* provê algumas informações sobre a validade e confiabilidade dos dados que são explicadas a seguir. Os rótulos foram mantidos em inglês para compatibilidade com o *Have I been pwned*.

Os vazamentos de dados podem receber os seguintes rótulos:

- *Sensitive*: as informações contidas são potencialmente danosas e só podem ser informadas após a verificação de que o e-mail ou domínio em questão realmente pertença à pessoa que realizou a busca. Isto pode ser feito pelo sistema de notificação do *site*;
- *Retired*: *breaches* que, por alguma razão, foram removidos do sistema do *Have I been pwned*.

Tabela 3.1. Dez maiores vazamentos - *Have I been pwned*

Origem	**Quantidade**	**Informações extras**
Onliner Spambot accounts	711.477.622	lista usada por um *spambot* denominado *Onliner*
contas do <https://exploit.in/>	593.427.119	unverified
Anti Public Combo List	457.962.538	unverified
River City Media Spam List	393.430.309	usado para *Spam marketing*
contas do <https://myspace.com>	359.420.698	
contas do <http://www.netease.com/>	234.842.089	unverified
contas do <https://linkedin.com/>	164.611.595	
contas do <http://www.adobe.com/br/>	152.445.165	
contas do <http://www.exactis.com/>	131.577.763	
contas do <https://www.apollo.io/>	125.929.660	

- *Unverified*: são *breaches* que não puderam ser verificados com certeza absoluta, mas que ainda assim contém dados potencialmente prejudiciais;
- *Fabricated*: *breaches* que, embora possam conter dados relevantes e possivelmente alguns verdadeiros, parecem ter sido fabricados e/ou copiados;
- *Spam list*: *breaches* que contém diversas informações pessoais, embora não necessariamente senhas, principalmente com o fim de envio de *spams*.

O site não guarda nenhuma informação a respeito das senhas encontradas. Nem todos os vazamentos contêm senhas. Muitos se referem a informações pessoais vazadas, como endereço, e-mail, entre outros e nem sempre possuem identificação de usuários (*logins*), não deixando por isso de ser um risco. Uma das fontes de dados usadas pelo *Have I been pwned* é o *Pastebin*. Outra fonte importante são os vazamentos publicados pela conta no *Twitter* "@dumpmon". A partir dessas informações são extraídos as postagens contendo e-mails que são analisados para validação. Os *pastes* são coletados aproximadamente 40 segundos após serem publicados. Os últimos *pastes* que continham vazamentos de dados eram postados para análise e verificação da comunidade, no entanto, segundo Hunt (2017b) estes dados, já pré-processados, começaram a ser utilizados de forma maliciosa.

3.2.3. *Dumpmon*

Desenvolvido por Jordan Wright, esta ferramenta analisa *pastes* de três fontes: *Pastebin*, *pastie*[6] e *Slexy*, utilizando-se de diversas expressões regulares para encontrar dados sensíveis ou que indiquem *data breaches*. Os dados obtidos podem conter informações de bancos de dados, chaves de *Google-api*, arquivos de configuração *Cisco*, dados de *honeypots*, entre outros (WRIGHT, 2013).

[6] Este *site* parece estar indisponível e não provê nenhuma informação a respeito de retorno. <http://pastie.org/> Acesso em 11/11/2018

Uma vez que dados sejam encontrados, são gerados *tweets* de alerta pela conta "@dumpmon" (DUMP..., 2017). O projeto, que é *opensource*, se encontra disponível e aceita colaborações (GITHUB..., 2017).

Este sistema também utiliza a análise da URL onde os *pastes* recentes são postados. Também apresenta a análise dos *pastes* através de expressões regulares: utiliza um conjunto de expressões para exclusão e para o aceite dos *pastes*.

3.2.4. *LeakedIn*

O *leakedin* (LEAKEDIN, 2017) apresenta diversas informações de *breaches* postados em inúmeras fontes. Segundo informações do *site* uma delas é o *Pastebin*. Também é informado que são usadas ferramentas para extrair esses dados, mas nenhuma delas é documentada.

3.2.5. O *Pastehunter*

O *PasteHunter* é uma aplicação desenvolvida pelo auto-entitulado "Kev", dono do *site Tech Anarchy* (KEV, 2017c). Esta aplicação, desenvolvida em *python*[7], recupera informações do *Pastebin*, *Github Gist* e *Dumpz*, utilizando regras definidas com o *Yara*[8], uma ferramenta desenvolvida para identificação de *malwares* com base em informações textuais ou binárias.

Esta aplicação utiliza a API HTTPS do *Pastebin*, portanto é necessário ser *Pro user* para utilizá-la. Usando as regras definidas com o *Yara*, identifica *e-mails*, algumas palavras de interesse (como *tango down*, indicativo de serviços retirados do ar em geral mediante ataques DoS) e *pastes* em *Base64*, comumente utilizado para disseminação de códigos maliciosos (KEV, 2017b). A ferramenta, que objetiva auxiliar pesquisadores de cibersegurança, encontra-se disponível no *Github* (KEV, 2017a).

3.3. Considerações gerais

Este capítulo abordou trabalhos relacionados ao que é proposto nesta monografia. Primeiramente foram abordadas ferramentas que analisavam o *Twitter* e o IRC, baseadas no mesmo arcabouço que pauta esta monografia. Em seguida, foram apresentadas ferramentas que analisam o *Pastebin*, extraindo informações e processando-as. Ficou claro um massivo uso de expressões regulares e palavras-chave para extrair *pastes* com relevância no contexto dessas ferramentas. No entanto, tais ferramentas não levam em conta diversas possibilidades, como o processamento de linguagem natural ou análises mais aprimoradas que expressões regulares. No próximo capítulo é apresentada e discutida a metodologia utilizada.

[7] <https://www.python.org/> Acesso em 11/11/2018
[8] <https://yara.readthedocs.io/en/v3.7.0/\#> Acesso em 11/11/2018

Capítulo 4

Metodologia

Este capítulo apresenta as questões de pesquisa e os métodos utilizados no afã de respondê-las. As questões de pesquisa são apresentadas na Seção 4.1. O método da pesquisa, que aborda a coleta, pré-processamento e análise dos dados e, ao final, a extração de informações de inteligência, é apresentado na Seção 4.2.

4.1. Questões de pesquisa

Neste trabalho objetiva-se a caracterização dos *pastes* postados no *Pastebin* quanto a sua relevância para a cibersegurança. Para tal foram estabelecidas três questões de pesquisa, descritas a seguir.

Q1 Como identificar informações relevantes para a cibersegurança em *pastes* publicados no *Pastebin*?

Q2 Quais são as características das publicações associadas à cibersegurança disponibilizadas no *Pastebin*?

Q3 As informações do *Pastebin* podem ser utilizadas para aviso antecipado de ameaças e/ou reações mais rápidas as mesmas?

Para investigação de **Q1**, foram utilizados conjuntos de expressões regulares, palavras-chave e identificação de idioma do *paste*. As palavras-chave e as expressões regulares foram divididas em *acceptlist* – que indica *pastes* com conteúdo possivelmente relevante para cibersegurança– e *blacklist* – que indica *pastes* com conteúdo provavelmente irrelevante. Também foram avaliados os resultados de algoritmos de classificação.

Para investigação de **Q2**, foi realizada a análise dos *pastes* relevantes. Foram analisadas as palavras-chave mais frequentes, incidência de cada expressão regular, idioma, bigramas, trigramas e quadrigramas mais comuns. Também foram analisadas as portas

(associadas a URLs e endereços IP) e os *e-mails* quanto a frequência e provedores mais comuns.

Para investigação de **Q3**, foram selecionados e analisados *pastes* relevantes, indicando a importância destes para a tomada de medidas proativas ou reativas. Também foram realizadas estatísticas básicas relativas à relevância e visibilidade dos *pastes*.

4.2. Método de pesquisa

A Figura 4.1 apresenta o método de pesquisa utilizado para investigar as questões de pesquisa e compreender as características dos *pastes* associados à cibersegurança no *Pastebin*. O método de pesquisa foi baseado no arcabouço proposto em (CAMPIOLO, 2016).

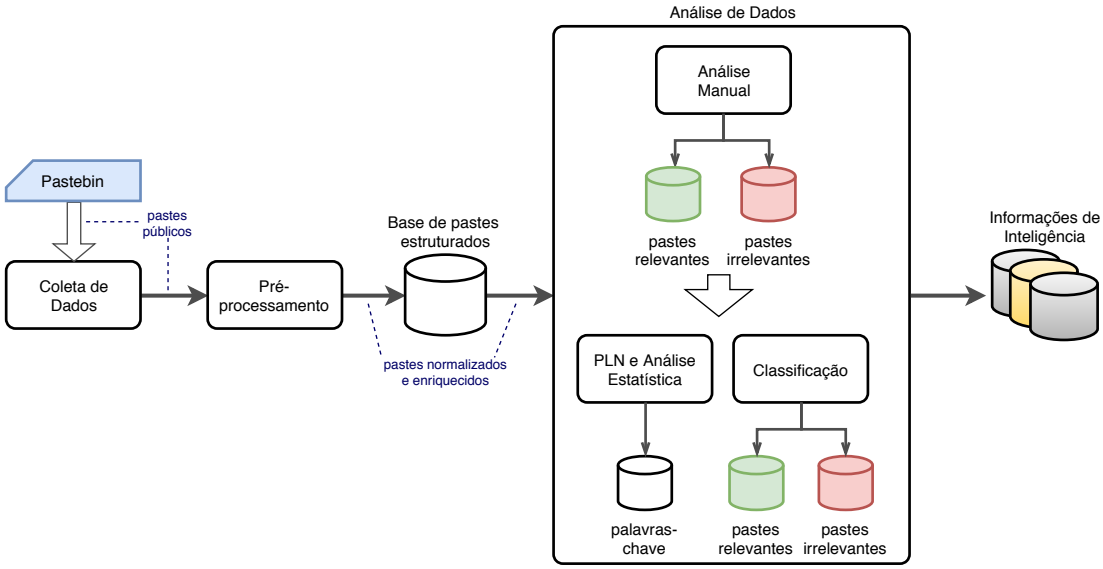

Figura 4.1. Método da pesquisa

Como observado na Figura 4.1, foi realizada a coleta e persistência de *pastes* públicos do *Pastebin* (Subseção 4.2.1). Estes *pastes* são enriquecidos, com o acréscimo de *paste entities* que visam identificar palavras-chave, padrões descritos por expressões regulares e detecção do possível idioma do *paste* (Subseção 4.2.2). Uma amostra aleatória dos *pastes* enriquecidos é selecionada e analisada, por meio de inspeção manual, estatística e de PLN. Os *pastes* desta amostra são classificados quanto a sua relevância ou irrelevância, por meio de algoritmos de classificação (Subseção 4.2.3). A partir desta análise são geradas informações de inteligência para cibersegurança, isto é, palavras-chave, expressões regulares e outras características que poderão ser usadas para a implementação de mecanismos mais eficientes na identificação e no alerta de publicações de *pastes* relacionados à cibersegurança.

4.2.1. Coleta dos dados

Para efetuar a coleta e enriquecimento dos *pastes*, foi desenvolvido um *software* em *Java*[1] que realiza requisições dos 100 *pastes* mais recentes a cada 60 segundos (esta e outras opções são configuráveis).

Para cada *paste* obtido, é verificado se a chave que o identifica já se encontra no banco, situação em que este *paste* é descartado por ser repetido. Caso contrário, ele é pré-processado (Subseção 4.2.2) e salvo, juntamente com as *paste entities* geradas, em formato *Java Script Object Notation* (JSON), em um banco de dados. O formato JSON foi escolhido devido à facilidade de visualização dos dados nesta etapa.

Para persistência dos dados, foi utilizado *Java Persistence API* (JPA)[2] e *Hibernate*[3]. e o Sistema Gerenciador de Banco de Dados (SGBD) *Sqlite3*[4], uma vez que não era necessário a utilização de um SGBD mais complexo, para além do pouco espaço em disco ocupado pelo *Sqlite3*, bem como sua simplicidade de uso. Os dados coletados de cada *paste* são explicados na Tabela 4.1.

Tabela 4.1. Descrição dos dados coletados de cada paste

Campo	**Descrição**	**Obrigatório**
Key	identificador único (7-8 caracteres) do *paste*	sim
Hits	número de acessos ao *paste*	não
Date	data de postagem do *paste*	sim
Title	título do *paste*	não
Format	formato (texto plano ou alguma linguagem de programação)	não
Visibility	privada, não listada ou pública	sim
Expiredate	tempo para que o *paste* seja removido do *Pastebin*	sim
Text	mensagem do paste	sim

Para acessar a API HTTP do *Pastebin*, foi utilizada a biblioteca *JPastebin* (BRIANBB, 2014), uma API sob a licença *GNU GENERAL PUBLIC LICENSE* (GPL) versão 3 (FOUNDATION, 2007), que implementa, para a linguagem de programação *Java*, uma camada de abstração para as APIs HTTPS fornecidas pelo *Pastebin*.

Para obter pleno acesso à API de *scraping* do *Pastebin*, melhor explicada na Subseção 2.3.1, foi necessário pagar pelo status de usuário *pro*, permitindo assim liberar um endereço IP para requisitar os *pastes* mais recentes sem que este fosse bloqueado por excesso de requisições. A coleta foi limitada, principalmente, pelo fato que as APIs do *Pastebin* apenas dão acesso a *pastes* públicos e, além disso, não permitem nenhum tipo de busca prévia nos *pastes*, como por exemplo por palavras-chave. Além disso, a API do *Pastebin* não associa o *paste* ao seu autor, quando o *paste* foi postado por um usuário identificado.

[1] <https://www.java.com> Acesso em 11/11/2018
[2] <http://www.oracle.com/technetwork/java/javaee/tech/persistence-jsp-140049.html> Acesso em 11/11/2018
[3] <http://hibernate.org/> Acesso em 11/11/2018
[4] <https://www.sqlite.org/> Acesso em 11/11/2018

4.2.2. Pré-processamento dos *pastes*

O pré-processamento consistiu na verificação da ocorrência das palavras-chave e aplicação das expressões regulares (tanto para a *acceptlist* quanto para a *blacklist*) e detecção do idioma do *paste*. Este pré-processamento foi realizado pelo mesmo *software* que realizou a coleta. Alguns campos são adicionados ao *paste* ao fim do pré-processamento, que são descritos na Tabela 4.2. As expressões regulares são cobertas em maiores detalhes na Subseção 4.2.2.

Tabela 4.2. Descrição dos dados adicionados ao *paste* no pré-processamento

Campo	**Descrição**	**Obrigatório**
Relevancy	relevância do *paste*	sim
Relevant	se o paste é ou não relevante	sim
Analyzed	determina se o *paste* já foi analisado	sim
Entities	lista de *pasteEntities* associadas (pode ser vazia)	sim
RegexValues	lista de *regexValues* associados (pode ser vazia)	sim

A detecção de idioma é realizada através da API *Langdetect* (NAKATANI, 2010). Essa API, licenciada sob a *Apache License* versão 2.0 (FOUNDATION, 2004), apresenta 99% de precisão para classificação de 56 idiomas utilizando um classificador *Naive Bayes* que utiliza perfis de idiomas.

Para cada tipo de pré-processamento (cada palavra-chave encontrada no *paste*, expressão que se aplique ou idioma detectado) é criada uma entidade chamada *pasteEntity*, que descreve o tipo de informação encontrada e se ela é indicativa de possível relevância ou irrelevância. Esta entidade é detalhada na Tabela 4.3. Os nomes associados a esta entidade são prefixados com o tipo de informação para facilitar a visualização, podendo ser um dentre três tipos: "re" (*regular expression* – expressão regular), "kw" (*keyword* – palavra-chave) ou "lg" (*language* – idioma). Assim, caso fosse utilizado, por exemplo, o termo "URL" como palavra-chave e para nomear uma expressão regular que encontre endereços de URLs não haveria conflito, já que os nomes seriam, respectivamente, "kw_url" e "re_url". Um *paste* pode ser associado a várias *paste entities*, e cada *paste entity* a diversos *pastes*.

Para cada casamento de uma expressão regular é criada uma entidade denominada *regexValue*, que associa o *paste*, *paste Entity* e o valor encontrado para tal expressão, permitindo assim verificar que dado foi obtido. Por exemplo, o casamento de uma expressão regular para capturar URL deverá ter como valor um endereço válido de URL, como <www.exemplo.com>. Detalhes sobre os *regex values* na Tabela 4.4. Cada *paste* pode ser associado a diversos *regex values*. Cada *regex value* é associado a um *paste* e uma *paste entity*. Cada *paste entity* relativa a expressão regular pode ser associada a diversos *expression value*.

Palavras-chave e expressões regulares alteram o valor do campo adicional "relevancy" do *paste*. Integrantes do conjunto da *acceptlist* incrementam o valor, enquanto da *blacklist* decrementam, permitindo assim avaliar essa métrica como indicador de relevância. Palavras-

chave incrementam ou decrementam em 0,1 enquanto expressões regulares possuem valor variável, manualmente atribuído na configuração do coletor. Estes valores foram escolhidos como um primeiro teste, visando analisar a aplicabilidade desta métrica. O valor das palavras-chave é baixo para evitar que a repetição de uma palavra tenha efeitos exacerbados no campo "relevância".

Tabela 4.3. Descrição de *paste entity*

Campo	**Descrição**	**Obrigatório**
Type	palavra-chave, expressão regular ou idioma	sim
Status	*accept* ou *reject*	sim
Name	nome que identifica a entidade (p.ex. re_url, kw_security, lang_pt)	sim
Correct	número de identificações corretas (para expressões regulares)	sim
Incorrect	número de identificações incorretas (para expressões regulares)	sim

Tabela 4.4. Descrição de *regex value*

Campo	**Descrição**	**Obrigatório**
PasteEntity	*pasteEntity* a qual se associa o valor (p.ex. URL)	sim
Value	valor da entidade (p.ex. <www.exemplo.com>)	sim
Correct	se a identificação foi correta ou não (p.ex.o endereço da URL está sintaticamente correto)	sim

Expressões regulares usadas na coleta

São utilizadas expressões regulares para identificar diversas características e padrões dos *pastes*. Além disso, cada expressão regular tem um valor associado, positivo caso faça parte da *acceptlist*, negativo caso da *blacklist*. O objetivo é auxiliar na criação de métricas para a identificação de relevância. Estas expressões regulares foram extraídas de outras ferramentas que monitoram fontes de dados não estruturados (como o *Dumpmon* – Subsection 3.2.3) ou criadas de acordo com tópicos de interesse.

Os tipos de expressões regulares indicativas de *pastes* com conteúdo provavelmente irrelevante são detalhadas na Tabela 4.5. As expressões regulares que indicam *pastes* com conteúdo possivelmente relevante são detalhadas na Tabela 4.6.

As colunas de ambas as tabelas (4.5 e 4.6) são descritas a seguir:

- Tipo da expressão: tipos de informações que se deseja obter (por exemplo números de documentos, cartões de crédito ou formas de endereçamento);
- Quantidade de expressões: número de expressões regulares neste grupo;
- Motivo: a razão para o tipo de informação ser desejado ou indesejado.

Tabela 4.5. *Blacklist*: expressões regulares para exclusões de *pastes*

Tipo da expressão	Quantidade de expressões	Motivo
Trechos de código	8	trechos de código não estão no escopo do trabalho
Logs de erros	2	erros de compilação e/ou lógica
Mensagens diversas	14	diversos padrões de mensagens, não relevantes

Tabela 4.6. *Acceptedlist*: expressões regulares que levam ao aceite de um *paste*

Tipo da expressão	Quantidade de expressões	Motivo
URL	1	possíveis alvos
E-mail	1	possível vazamento de dados, inclusive senhas
IP	4	IP V4 e V6, com ou sem portas, possíveis alvos
Chaves criptográficas e *hashes*	6	possível vazamento de chaves privadas ou senhas encriptadas
Número de documentos	4	números de documentos como CPF
Modos de endereçamento	3	indicativo de doxing
Números de cartão de crédito	6	possível doxing
Número de telefone	2	possível doxing
Termos e nomes de grupos conhecidos	3	termos comumente associados a cibersegurança
Endereço de carteiras *Bitcoin*	1	muito utilizado para venda de informações

4.2.3. Análise dos dados

A análise realizada dividiu-se em três partes: análise manual, visando criar duas bases de *pastes*, uma de *pastes* relevantes e outra de irrelevantes (Subseção 4.2.3); análise estatística e processamento de linguagem natural, visando compreender as características dos *pastes* relevantes (Subseção 4.2.3); e classificação, visando identificar *pastes* relevantes com base nas características extraídas nas análises anteriores (Subseção 4.2.3).

Análise manual

Para realizar a análise manual, foi desenvolvido um *software* em *Java* que seleciona pastes de forma aleatória, dentre todos os coletados. Cada *paste* foi avaliado quanto a sua relevância, sendo as possíveis respostas “sim” e “não”. Esta informação é armazenada no campo “relevant” do *paste*. Foram considerados relevantes todos os *pastes* que contenham informações sensíveis ou que possam de alguma forma ser utilizados em medidas proativas ou reativas dentro do contexto de cibersegurança, incluindo, mas não se limitando a *doxing*, vazamento de dados, venda de dados pessoais e possíveis alvos. A única exceção são códigos fontes que, mesmo que possam ser usados para fins maliciosos, estão fora do escopo deste trabalho, devido às diferenças em sua análise e por suas características semânticas e sintáticas requererem avaliação mais apurada. Quanto a *pastes* envolvendo pirataria, foram considerados apenas os que envolviam *software*, especialmente que necessitavam da utilização de outros programas,

como geradores de chaves de acesso ou programas modificados (*crackeados*), potencialmente danosos no contexto de cibersegurança.

Após a análise do *paste*, para cada *regexValue* associado, se houver, é inquirido se o casamento da expressão regular realmente obteve o tipo de dado esperado. Isto possibilita aferir a taxa de acerto das expressões regulares e identificar as que são passíveis de serem utilizadas em um sistema com menor supervisão, além de permitir encontrar erros (semânticos ou não) nas expressões. Por exemplo, uma expressão regular que identifique endereços IPV4 pode, ocasionalmente, assumir que uma numeração de versão (como 1.2.3.4) é um endereço IP. A informação sobre o acerto do casamento do *regexValue* é armazenada no campo "correct". Além disso, a *paste entity* associada tem um de dois possíveis campos incrementados: caso a identificação tenha sido correta, é incrementado o campo "correct", caso contrário, é incrementado o campo "incorrect".

Após a análise manual, que foi feita por amostragem sobre uma grande base de dados coletados, são gerados dois conjuntos de *pastes*, dos *pastes* "relevantes" e dos "irrelevantes", com suas respectivas *paste entities* e *regex values* associadas. Estes dados são utilizados como entrada para as análises a seguir.

Análise estatística e de PLN

O objetivo principal da análise estatística e de PLN é permitir a caracterização dos *pastes* e, a partir disso, identificar características indicativas de relevância e irrelevância. Tais análises foram realizadas sobre os dados gerados pela análise manual (Subseção 4.2.3). A análise estatística visa identificar os percentuais de acerto das aplicações das expressões regulares e dos idiomas, enquanto o PLN visa identificar as palavras mais frequentes em ambas as bases, bem como os bigramas, trigramas e quadrigramas e discutir sua relevância na identificação especialmente dos *pastes* irrelevantes.

Durante a análise manual (Subseção 4.2.3), foi avaliado o acerto de cada expressão regular. Durante a análise estatística, tal acerto foi utilizada para determinar a taxa de acerto das expressões regulares, o que é um possível indicativo da necessidade de adequação da expressão regular ou mesmo da impraticabilidade de seu uso. Também foi aferido o percentual de *pastes*, relevantes e irrelevantes, em que cada expressão regular aparece. Foi também indicado e discutido o acerto da detecção de idiomas.

Para a extração das palavras mais frequentes, foi desenvolvido um programa em *Python*[5], linguagem escolhida por sua praticidade. Este programa utiliza a *Scikit Learn* (PEDREGOSA et al., 2011), um conjunto de ferramentas para análise e mineração de dados, sob a licença *Berkeley Software Distribution* (BSD)(DISTRIBUTION, 1999). Este programa tokeniza cada *paste*, considerando válido termos puramente alfabéticos, com pelo menos três caracteres. Para tal, foi utilizada a classe *TfidfVectorizer* do *Scikit-learn*, que converte textos

[5] <www.python.org> Acesso em 11/11/2018

não processados em uma matriz de características *term frequency* (TF)–*inverse document frequency* (IDF), ou seja, é a frequência em que o termo aparece, levando-se em conta a quantidade total de documentos e de aparições do termo. Os valores para "max_df" e "min_df" (representando corte de frequência máxima e mínima das palavras no documento) de $0,85$ e $0,05$ respectivamente. Estes valores foram escolhidos após sucessivos testes por conterem o menor ruído.

Os bigramas, trigramas e quadrigramas foram extraídos utilizando *Natural Language Toolkit* (NLTK)(LOPER; BIRD, 2002), um conjunto de ferramentas para PLN em *Python*, sob a licença *Apache* versão 2.0(FOUNDATION, 2004). Foi desenvolvido um programa que calculou, para cada *paste*, os referidos bigramas, trigramas e quadrigramas, agrupando-os em relevantes e irrelevantes e exibindo a quantidade dos mesmos em relação ao conjunto total.

Dados mais apurados, incluindo a configuração completa de cada processo utilizado podem ser encontrados no Apêndice E.

Classificação

A classificação objetivou avaliar a possibilidade de, por meio do uso de algoritmos de classificação, identificar a relevância ou irrelevância de um dado *paste*. Para tal, foram levados em conta dois conjuntos de características: no primeiro caso, foram utilizadas as frequências de palavras, unicamente, através do mesmo processo de tokenização com *scikit-learn* utilizado para a detecção de palavras frequentes (Subseção 4.2.3). No segundo caso, a entrada para os classificadores foi binária, levando em conta a existência ou ausência de cada *paste entity* no *paste*.

Foram utilizadas as implementações dos algoritmos de classificação da *scikit-learn* (PEDREGOSA et al., 2011). Os algoritmos, bem como quaisquer parâmetros modificados do padrão de execução do *scikit-learn* são apresentados a seguir:

- *Naive Bayes*:
  - *multivariate Bernoulli models*: apresentado por Aggarwal e Zhai (2012) como alternativa para análise binária de termos em textos (levando-se em conta apenas a existência ou não de termos). Foi utilizado na classificação das *paste entities*, que são binárias, mas não do texto, que levou em conta a frequência das palavras. Não teve seus parâmetros alterados;
  - *Multinomial*: Segundo Aggarwal e Zhai (2012), ideal para análise de textos quando a frequência de ocorrência das palavras é levada em conta. Neste trabalho, foi utilizado apenas na análise de texto. Não sofreu alteração em seus parâmetros padrão;
- KNN: algoritmo clássico para classificação, utilizado com os valores de "*n_neighbors*" (número de vizinhos) de 3, 5 e 7, que foram escolhidos devido a quantidade de dados

nas amostras e por apresentar melhor resultados que valores mais altos;
- Árvore de decisão: utilizado com balanceamento de classe ("*class_weight*")(o balanceamento de classe tenta gerenciar o desbalanceamento de quantidades de amostras);
- SVM: sempre com "*gamma*" atribuído como "auto", também utilizado com balanceamento de classe "*class_weight*".

Dados mais apurados, incluindo a configuração completa de cada processo utilizado podem ser encontrados no Apêndice E.

4.2.4. Extração das informações de inteligência

A extração de informações de inteligência relativas à cibersegurança objetiva aperfeiçoar a detecção e identificação de novos *pastes*. As características selecionadas para compor esta base de inteligência são obtidas em cada uma das etapas da metodologia, incluindo, por exemplo, a lista de expressões regulares indicando *pastes* irrelevantes obtida a partir da análise manual.

Esta base está disponibilizada no *Github*, no endereço <https://github.com/felipeveigaramos/Pastebin_data>. Também estão disponibilizados, entre outros dados, o coletor de *pastes*, as expressões regulares e outros resultados ainda não explorados.

4.3. Considerações gerais

Este capítulo apresentou as três questões de pesquisa e o método utilizado para respondê-las. Explicou o funcionamento da coleta dos dados, através de um *software* desenvolvido que realiza sucessivas requisições ao *Pastebin* e pré-processa os *pastes* obtidos, utilizando expressões regulares, listas de palavras-chave e detecção de idioma. Em seguida, foi descrito o processo de análise, que é iniciado pela análise manual, a fim de criar as bases de *pastes* relevantes e irrelevantes, passa pela extração dos termos mais comuns e avaliação das expressões regulares aplicadas e termina com a classificação dos dados utilizando os algoritmos de classificação *KNN*, *SVM*, *naive bayes* e árvore de decisão. O próximo capítulo apresenta e discute os resultados obtidos com a aplicação deste método.

Capítulo 5

Resultados e Discussões

Este capítulo apresenta e discute os resultados do método apresentado no capítulo 4. A Seção 5.1 apresenta o resultado da coleta dos dados. A Seção 5.2 discute os resultados da análise manual e caracteriza as bases de *pastes* criadas. A Seção 5.3 apresenta o resultado da análise estatística, incluindo relevância e acerto de expressões regulares e palavras-chave. A Seção 5.4 realiza o PLN, extraindo termos comuns, bigramas, trigramas e quadrigramas. A Seção 5.5 apresenta e discute a classificação dos *pastes* através da utilização de algoritmos de classificação. Por fim, a Seção 5.6 apresenta as respostas para as questões de pesquisa enquanto a Seção 5.7 detalha a disponibilização da base de inteligência gerada.

5.1. Coleta de dados

Para este trabalho, os dados foram coletados pelo *software* desenvolvido para este fim durante 21 dias, iniciando em 07/02/2018 e terminando em 28/02/2018. Devido à natureza da API do *Pastebin*, que restringe o acesso a um único endereço IP pré-cadastrado, a necessidade de ajustes no coletor, indisponibilidade do serviço de *Internet* e outras eventualidades, existiram interrupções no período de coleta. Foi considerada interrupção qualquer período superior a dois minutos sem nenhum *paste* coletado.

Neste período, foram coletados 254.459 *pastes*. A despeito da duração do período de interrupção ser praticamente o dobro do período de coleta, o espaçamento no intervalo de 21 dias, ou seja, com *pastes* coletados em quase todos os dias bem como o número de *pastes* coletados indica suficiente material para iniciar a primeira análise. A média de *pastes* coletados por dia foi de 12117 *pastes*. A Tabela 5.1 detalha a coleta e as médias de *pastes* coletados. Todos os *pastes* coletados foram armazenados, após pré-processamento, em formato JSON, em um banco de dados.

Tabela 5.1. Detalhes da coleta de *pastes*

Característica	**Valor**
Início da coleta	07/02/2018 20:20:12
Fim da coleta	28/02/2018 15:10:13
Total de *pastes* coletados	254.459
Dias executando	21
Média de *pastes* coletados no período por dia	12117
Média de *pastes* coletados no período por hora	505

5.2. Análise manual

Os *pastes* foram analisados com o auxílio do *software* descrito na Subseção 4.2.3 e analisados por três especialistas do domínio, com checagem dupla dos resultados.

Foram analisados 3.650 *pastes*, dos quais 183 foram considerados relevantes e 3.467 irrelevantes. A despeito de 5% para 95% parecer um número pequeno, quando comparado com o número de captura total no período (254.459) é bastante considerável. Foram encontradas 187 *paste entities*, identificando expressões regulares, palavras-chave e idioma. Também foram gerados 95.300 *regex value*, contendo o resultado do casamento das *paste entities* de expressões regulares. Estes dados são detalhados na Tabela 5.2. As Tabelas 5.3 e 5.4 apresentam exemplos de *pastes* encontrados, respectivamente relevantes e irrelevantes.

Tabela 5.2. Detalhes de coleta – *paste entities* e *regex values*

Tipo de dado	**Quantidade**
paste entities totais	187
paste entities de palavras-chave	113
paste entities de idioma	34
paste entities de expressões regulares	40
regex value	95.300
regex value corretos	46477
regex value incorretos	48823

Tabela 5.3. Detalhes de coleta – *pastes* relevantes

Trechos do *paste*	***Paste entities* associadas**
Welcome to Carder007(...) sold cvv credit card(...) Uk (Visa,Master) = 15$	lg_en, re_creditcard_visa, re_url, re_email
Download Free Windows 8 Full Version ISO	re_url, lg_en,
TextAloud 3.0.40 keygen.rar is a stunning interface for programming and the template utility.	kw_protection, re_url, lg_en,kw_password,
Netflix:ku[redacted]@gmail.com Pass: 123456	lg_de, re_email
nome:Stefano(...) eta:22 indirizzo: via: de[redacted]ri citta:(...)	lg_it

Durante a análise manual foram evidenciadas expressões regulares adequadas para a composição de uma *blacklist* mais precisa. Estas expressões regulares não foram encontradas em nenhum *paste* considerado relevante e, em busca por *pastes* que contivessem estas expressões como *paste entities* associadas, especificamente, nenhum *paste* relevante foi

Tabela 5.4. Detalhes de coleta – *pastes* irrelevantes

Trechos do *paste*	***Paste entities* associadas**
sudo apt-get autoremove eclipse*(...)	lg_en
Exception of type System.InvalidOperationException:(...)	lg_en
`?xml version="1.0" encoding="UTF-8"?`	lg_ca,
`#include<iostream> #include<cmath> using namespace std; (...)`	lg_ro
Nocturna_A._Alteriano_Night.Owl.Inc(...)local attack = false(...)	lg_en, kw_attack,re_*1

encontrado. Tais expressões regulares foram: re_#10, re_#13, re_#14, re_#5, re_*1, re_*2, re_canadian_postal_code, re_css, re_html_xml, re_java, re_php, re_python_error, re_skuid, re_sql, re_sqlserver. Para detalhes sobre as nomenclaturas e os resultados destas expressões regulares, consultar respectivamente os apêndices C e D.

5.2.1. Os *pastes* mais acessados

Até meados de 2018, o *Pastebin* disponibilizava listas contendo os *pastes* mais acessados. Estas listas, porém, foram removidas devido a abusos (mais na Subseção 2.3.1). As análises realizadas em meados de 2017 e que são apresentadas a seguir auxiliam a compreender esta situação, bem como permitem compreender o alcance que postagens relacionadas à cibersegurança podem ter no *Pastebin*.

Trends dos últimos 365 dias

O resultado dos *trends* dos últimos 365 dias, disponibilizado em meados de 2017, são resumidos em duas tabelas: uma mostrando os *pastes* relevantes (Tabela 5.5) e irrelevantes (Tabela 5.6). Foram analisados 26 *pastes*, dos quais 10 são relevantes para a cibersegurança.

Tabela 5.5. *Last 365 days*:relevantes

Tipo de informação	**quantidade**	**informações adicionais**
Vazamentos	2	notícias e links contendo vazamentos
Códigos maliciosos	6	diversos códigos, alegadamente maliciosos
Base64	2	códigos com *base64* inclusos

Tabela 5.6. *Last 365 days*: irrelevantes

Tipo de informação	**quantidade**	**informações adicionais**
Jogos	7	links, informações e estratégias
Configurações	2	de máquinas ou sistemas
Textos em geral	2	não relevantes
trechos de código fonte	3	diversas linguagens de programação
Filmes e animes	2	*links*

All trends

O resultado dos *trends* gerais, ou seja, *Trends* avaliados com base em todos os *pastes*, disponibilizado em meados de 2017, são resumidos em duas tabelas: uma mostrando os *pastes* relevantes (Tabela 5.7) e irrelevantes (Tabela 5.8). Foram analisados 34 *pastes*, dos quais 14 são relevantes para a cibersegurança.

Tabela 5.7. *All trends*:relevantes

Tipo de informação	**Quantidade**	**Informações adicionais**
Vazamentos de dados e hacktivismo	8	notícia relacionadas e dados propriamente. 3 a respeito da KKK e outros evidenciando hacktivismo
Possíveis *base64*	2	dentro de atributos de linguagens de programação
Hacking	1	tutorial
Pornografia infantil	1	
Links Deep Web	1	possíveis *links* para análise
Cartões de crédito	1	

Tabela 5.8. *All trends*:irrelevantes

Tipo de informação	**Quantidade**	**Informações adicionais**
Jogos	8	*links*, dicas e informações
Filmes e animes	2	links e informações
Trecho de códigos	1	aparentemente não malicioso
Textos em geral	4	nem sempre em inglês e aparentemente sem dados relevantes
Pornografia	1	lista de links

5.3. Análise estatística

A análise estatística foi realizada sobre os 3.650 *pastes* resultantes da análise manual, bem como suas *paste entities* e *regex values* associadas. Levou em conta o acerto e a relevância das *paste entities* identificadas no pré-processamento. Tabelas contendo o resultado completo da análise de todas as *paste entities* encontradas são exibidas no Apêndice D, enquanto detalhes sobre a nomenclatura das expressões regulares estão no Apêndice C.

Uma das propostas era a avaliação da relevância atribuída a cada *paste*, que foi composta de um valor numérico dado para cada expressão regular de acordo com a sua possível utilidade na identificação de *pastes* relevantes ou irrelevantes – valores positivos para relevantes, negativos para irrelevantes – e pela ocorrência de palavras-chave associadas à *acceptlist*. Este método porém se provou pouco confiável dada a alta taxa de falha das expressões regulares e a necessidade de recalcular as palavras-chave para levar em conta suas bordas. No entanto, os resultados deste trabalho permitiram reavaliar o sistema de pesos para expressões regulares.

5.3.1. Palavras-chave

As palavras-chave utilizadas como parte da *acceptlist* (não existiam palavras-chave na *blacklist*) foram analisadas quanto a sua ocorrência. Inicialmente foram verificadas a existência da cadeia de caracteres no texto, mas, devido a imprecisão que isso trazia, já que considerava quaisquer trechos no meio de palavras, optou-se por buscar pelas palavras-chave com bordas bem definidas, ou seja,o termo buscado, separado de outros termos por espaço ou pontuação. A Tabela 5.9 mostra as dez palavras-chave que apareceram no maior número de *pastes* relevantes, enquanto a Tabela 5.10 as dez mais recorrentes nos irrelevantes. Os valores para ambas são percentuais e relativos ao conjunto de *pastes*, ou seja, 183 relevantes; 3.467 irrelevantes.

Tabela 5.9. Palavras-chaves: 10 mais recorrentes – relevantes

Palavra-chave	**Relevantes (%)**	**Irrelevantes (%)**
crack	19.13	1.33
btc	12.02	0,06
tor	11.48	0,55
password	7.65	1.87
hack	3.83	0,49
hacker	3.28	0,32
patch	3.28	1.21
pass	2.73	1.5
secure	2.73	0,78
cracked	2.19	0,52

Tabela 5.10. Palavras-chaves: 10 mais recorrentes – irrelevantes

Palavra-chave	**Relevantes (%)**	**Irrelevantes (%)**
error	1.09	4.56
down	1.64	3.49
target	1.09	2.97
password	7.65	1.87
attack	0,55	1.5
pass	2.73	1.5
security	1.64	1.41
crack	19.13	1.33
patch	3.28	1.21
patches	1.09	1.1

A despeito das palavras-chave terem sido usadas apenas para compor a *acceptlist*, a incidência de algumas delas no grupo de *pastes* irrelevantes foi alta o bastante para que se considere a utilização das mesmas como parte da *blacklist*. Além disso, a quantidade de termos encontrados, a partir da base utilizada, que foi extraída dos trabalhos de SANTOS et al. (2012), Campiolo e Batista (2015) foi relativamente baixa, indicando que, padrões utilizados em publicações associadas à cibersegurança no *Twitter* e no IRC diferem dos

encontrados no *Pastebin*. Não foi possível comparar estes resultados com os citados na Seção 3.2 uma vez que nenhum disponibiliza quaisquer palavras-chave utilizadas ou extraídas a partir da aplicação de seus métodos.

5.3.2. Expressões regulares

A taxa de acerto das expressões regulares foi avaliada para determinar se os padrões com os quais elas casaram correspondiam ao tipo de informação esperada. Foram identificadas diversas situações em que isto não ocorria. Algumas são exemplificadas a seguir:

- Endereços IP frequentemente associados a número de versão: "Edition 5.3.0.142 incl Crack Rapidshare";
- URL capturando termos separados por ponto sem nenhuma relação com endereços na *web*, em geral como parte de códigos de programação. Exemplo: "c.name = ";
- *E-mails* passando por prefixo de terminais *Unix* ou identificação de trechos de log, por exemplo "vendor.qti.hardware.alarm@1.0-service.rc".

A Tabela 5.11 mostra as dez expressões regulares mais encontradas em *pastes* considerados relevantes, enquanto a Tabela 5.12 mostra as dez com maior incidência nos *pastes* irrelevantes. Os valores de "relevantes" e "irrelevantes" são relativos à quantidade de *pastes*, respectivamente 183 e 3.467, enquanto a "taxa de acerto" refere-se ao percentual de acerto em todos os *pastes* dentre os 3.650 nos quais o padrão representado pela expressão regular foi encontrado. O campo "lista" indica a que grupo a expressão regular pertencia inicialmente, *acceptlist* ou *blacklist*. Detalhes sobre a nomenclatura das expressões regulares são encontrados no Apêndice C.

Tabela 5.11. Expressões regulares: 10 mais recorrentes – relevantes

Expressão regular	Lista	Corretude (%)	Relevantes (%)	Irrelevantes (%)
re_url	*acceptlist*	82.41	69.4	28.61
re_email	*acceptlist*	97.73	26.23	1.79
re_us_zip_code	*acceptlist*	2.75	8.2	0,43
re_us_phone	*acceptlist*	2.78	7.65	0,49
re_creditcard_visa	*acceptlist*	80.11	6.56	0,0
re_ipv4	*acceptlist*	99.74	6.56	1.01
re_us_social_security_	*acceptlist*	12.66	6.01	0,2
re_br_cpf	*acceptlist*	2.22	3.28	0,03
re_br_cep	*acceptlist*	29.41	2.19	0,03
re_hash32	*acceptlist*	2.13	1.09	0,2

A maior dificuldade na avaliação de expressões regulares aconteceu com as que eram compostas por padrões numéricos - como documentos ou números de telefone. O principal motivo foi que, tornada muito rígida, ou seja, com todas as pontuações (traços, parênteses, pontos, entre outras), a chance de detecção cairia, além do que é difícil saber o quanto da

Tabela 5.12. Expressões regulares: 10 mais recorrentes – irrelevantes

Expressão regular	Lista	Corretude (%)	Relevantes (%)	Irrelevantes (%)
re_url	*acceptlist*	82.41	69.4	28.61
re_*1	*blacklist*	98.66	0,0	4.36
re_sql	*blacklist*	81.05	1.09	3.37
re_java	*blacklist*	96.67	0,0	1.93
re_email	*acceptlist*	97.73	26.23	1.79
re_#13	*blacklist*	96.67	0,0	1.5
re_html_xml	*blacklist*	99.02	0,0	1.5
re_css	*blacklist*	94.12	0,0	1.24
re_ipv4	*acceptlist*	99.74	6.56	1.01
re_us_phone	*acceptlist*	2.78	7.65	0,49

divisão em pontuação será seguida (por exemplo, se um Cadastro de Pessoa Física (CPF) é dividido apenas com um hífen – 000000000-00 – ou com pontuação e hífen – 000.000.000-00); tornada muito branda, ou seja, com o mínimo possível de pontuações e marcas, a expressão regular torna-se genérica em demasia e o padrão casará com qualquer tipo de dado (ainda com o exemplo do CPF, onze números contínuos – 00000000000 – podem representar qualquer informação). No entanto, algumas expressões regulares se destacaram devido a seu percentual de acerto. Elas são listadas a seguir:

- Expressões regulares com 100% de taxa de acerto: “re_#14”, “re_#3”, “re_#5”, “re_#6”, “re_*2”, “re_php”, “re_python_error”,
- Expressões regulares com 0% de taxa de acerto: “re_canadian_postal_code”, “re_creditcard_americanexpress”, “re_creditcard_discover”, “re_creditcard_jcb”, “re_creditcard_mastercard”, “re_us_social_insurance_number”,

Certas expressões regulares se destacaram por serem sempre associadas a *pastes* irrelevantes, independente de quais outras *paste entities* sejam associadas. São as seguintes: re_sql, re_php, re_*1, re_*2, re_css, re_python_error, re_java, re_skuid, re_sqlserver, re_#5, re_#10, re_#13, re_#14, re_html_xml, re_canadian_postal_code, re_ipv6_with_port e re_ipv6.

Algumas das expressões utilizadas neste trabalho foram baseadas ou extraídas de algumas das ferramentas descritas na Seção 3.2, principalmente do *Dumpmon* (Subseção 3.2.3). No entanto, o *Dumpmon* não disponibiliza nenhuma informação a respeito do acerto destas expressões. Algumas, como a relacionada a *e-mail* provou-se relevante, enquanto outras, como as relativas a *hashes* não evidenciaram dados relevantes. A *Seek Data Leakage* (Subseção 3.2.1) utiliza uma expressão regular para identificar vazamento de *e-mails* com senhas, mas não provê nenhuma informação sobre seus resultados, impedindo a comparação.

5.3.3. Detecção de idiomas

A detecção de idiomas do *Pastebin* se provou pouco confiável, uma vez que o tipo dos dados disponibilizados é tão diverso – por exemplo, é comum a identificação do idioma Inglês em *pastes* contendo código de programação, seja em inglês ou não. Este método como característica para a definição de relevância foi descartado. No entanto, a título de caracterização, são exibidos, agora, os dez idiomas com maior quantidade de *pastes* encontrados, bem como o respectivo percentual de *pastes* relevantes e irrelevantes, levando-se em conta os 183 *pastes* relevantes e 3.467 irrelevantes. A Tabela 5.13 mostra estes dez idiomas. O idioma é identificado por sua abreviação. A lista completa pode ser encontrada no Apêndice D.

Tabela 5.13. Idiomas: 10 mais recorrentes

Idioma	**Quantidade**	**Relevantes (%)**	**Irrelevantes (%)**
en	2679	76.5	73.23
fr	93	1.64	2.6
sq	92	0,0	2.65
de	73	3.28	1.93
ro	59	1.09	1.64
ca	58	2.73	1.53
es	43	0,0	1.24
pt	43	0,55	1.21
it	37	0,55	1.04
pl	37	0,55	1.04

Nenhuma das ferramentas apresentadas na Seção 3.2 caracteriza o *paste* quanto a seu idioma. Embora este seja um diferencial deste trabalho que auxilia na caracterização e compreensão dos *pastes* postados, não foi significante na determinação da relevância do *paste* no contexto de cibersegurança.

5.3.4. Análise de características específicas

Certas características foram avaliadas visando auxiliar na caracterização dos *pastes*. A Tabela 5.14 mostra os 10 provedores dos *e-mails* mais encontrados no *Pastebin*. Foram avaliadas também as portas de acesso mais utilizadas, em endereços IP e URL. Apenas as URLs irrelevantes retornaram resultados, que são exibidos na Tabela 5.15 e incluem portas para acesso HTTP e servidores HTTP..

As análises específicas evidenciaram a prevalência de *e-mails* associados ao *gmail* nos dados coletados. Permitiram perceber que a avaliação de entidades específicas pode trazer informação relevante, para além de auxiliar na caracterização dos *pastes*.

Tabela 5.14. *E-mails*: 10 provedores mais recorrentes

Provedor	Quantidade
Gmail	620
Yahoo	284
Hotmail	170
Aliceadsl	157
Free	93
Yeah	72
Wanadoo	62
Live	31

Tabela 5.15. 10 portas mais comuns em URLs de *pastes* irrelevantes

Porta	Quantidade
2086	3697
8000	2537
25461	1304
80	657
1985	334
8080	192
1935	191
8020	130
7040	74
7000	70

5.4. Processamento de linguagem natural

Enquanto as análises estatísticas focaram, principalmente, em validar e/ou utilizar recurso de outras fontes – como expressões regulares e palavras-chave – acrescidos de alguns adicionados para este trabalho, o PLN objetivou majoritariamente a extração de novas características com base nos *pastes* analisados. Estas características ajudaram a compor a base de inteligência criada.

5.4.1. Palavras mais comuns

As duas bases de *pastes*, tanto de relevantes quanto de irrelevantes, foram analisadas em busca das palavras mais recorrentes em ambas e que diferissem entre as bases, permitindo uma melhor caracterização dos *pastes*. Foram consideradas palavras com três ou mais caracteres, sem a presença de pontuação ou números. A lista completa foi disponibilizada no repositório com as bases de inteligência, no entanto, seguem algumas amostras:

- Relevantes: As cinco palavras mais comuns foram: “hydra”, “onion”, “center”, “visa” e “country”. Outros termos, embora não estejam entre os cinco mais recorrentes, são bastante significativos: “gmail”, na sexta posição, “crack” na nona, “yahoo”, na décima, “*keygen*”, na 27 e “cvv”, na 38.

- Irrelevantes: As cinco palavras mais comuns entre os *pastes* irrelevantes foram: “null”, “false”, “set”, “error’ e “title”. Ainda com grande recorrência, chama a atenção “type”, na 7a posição, “true”, na 9, “class”, na 10, “return” na 12 e “function” na 13. A maioria destas palavras indicam ser de códigos de programação.

As palavras-chave utilizadas na análise estatística não predominam durante a contagem de palavras, levando à necessidade da criação de listas atualizadas que levem estas palavras em conta.

5.4.2. Bigramas, trigramas e quadrigramas

A análise dos bigramas, trigramas e quadrigramas foi realizada para ambos os conjuntos de *pastes*, relevantes (183) e irrelevantes (3.467), separadamente. Os resultados completos foram disponibilizados no repositório da base de inteligência desenvolvida, descrito na Seção 5.7. Alguns exemplos são evidenciados a seguir.

- Bigramas:
  - Relevantes: “carder007s.com http”, “fullz info”, “from port”, “adobe photoshop” e “account paypal”. A maioria destes guardam relação com a venda de informações financeiras, sejam elas números de cartão de crédito (realizado pelos chamados *carders*) e contas de serviços de pagamento, como o *pay pal*[1]. Também são evidenciados nomes de *software* que possuam versão pagas, em geral associadas a versões *crackeadas* dos mesmos;
  - Irrelevantes: “not set”, “null xture”, “core map”, “client thread/info” e “object reference”. É possível identificar uma forte presença de termos que compõem registros de erros ou trechos de código de programação.
- Trigramas:
  - Relevantes: “block all packets”, “block udp from”, “little war game”, “steam key generator” e “buy bank logs”. A presença de *pastes* contendo conteúdo relacionado a contas bancárias e *softwares* para geração de chaves de acesso fica ainda mais evidenciada;
  - Irrelevantes: “null texture passer”, “core map version”, “object reference not”, “could not resolve” e “the local file”. As referências a registros de erros e exceções ficam ainda mais destacadas.
- Quadrigramas:
  - Relevantes: “buy dumps with pin”, “great ittle war game”, “onion http zefir site”, “all packets from ips”, “all steam key enerator” e “will sell for you”. Este método se evidencia como uma interessante forma para identificar *pastes* fortemente relacionados a vendas de dados;

[1] <https://www.paypal.com/> Acesso em 11/11/2018

– Irrelevantes: “incorrect core map version”, “object reference not set”, “could not resolve crossreference”, “type gl_apicall void gl_apientry” e “morph online resource thread/info”. Os registros de erro são ainda mais evidenciados.

Os bigramas, trigramas e quadrigramas mostrados estão entre os mais recorrentes de seus respectivos grupos, embora não estejam ordenados. O conjunto dos relevantes possui uma quantidade muito menor de termos e possibilidades já que equivale a cerca de 5% da base total de *pastes* manualmente analisados, porém foi possível perceber que este tipo de análise facilita a identificação de *pastes* nos quais haja a venda de informações, usualmente dados de cartões de crédito. Quanto aos irrelevantes, este meio pode ser utilizado para a diminuição dos *pastes* relacionados à programação e registros de erro, diminuindo significativamente a quantidade de *pastes* a analisar.

5.5. Classificadores

Como entrada para a classificação, foram utilizados os dados gerados pela captura e enriquecimento dos *pastes*– etapa de pré-processamento. O foco da análise dos classificadores foi quanto à utilização das *paste entities* e análise direta do próprio texto. A principal dificuldade encontrada foi o número de amostras e o desbalanceamento entre as duas classes (de *pastes* relevantes e irrelevantes). Para melhor compreender o efeito do desbalanceamento, foram realizadas análises com a base completa, ou seja, os 3.650 *pastes* e com amostras de mesmo tamanho de ambas as bases, ou seja, 150 *pastes* relevantes e 150 irrelevantes, escolhidos de forma aleatória, compondo uma base balanceada.

Foram testadas cinco possibilidades:

- Análise textual, através da contagem de frequência de palavras;
- Todas as *paste entities* aplicadas, ou seja, de idioma, palavras-chave e expressões regulares;
- Apenas *paste entities* de palavras-chave e expressões regulares;
- Apenas *paste-entities* de expressões regulares;
- Apenas *paste entities* de palavras-chave.

Para cada um dos itens apresentados foram testados os algoritmos, como descrito na Seção 4.2.3. Para cada algoritmo, são apresentados a precisão (*precision*), a abrangência (*recall*), a Taxa de Falsos Positivos (TFP) e a Taxa de Falsos Negativos (TFN). As matrizes de confusão, bem como detalhes de aplicação estão disponíveis no repositório, descrito na Seção 4.2.4. Todos os valores de métricas foram fornecidos tendo em vista os “relevantes”.

5.5.1. Base de dados completa

Os resultados são apresentados em tabelas individuais para cada tipo de característica avaliada. A Tabela 5.16 apresenta os resultados para a análise textual. A Tabela 5.17 mostra a avaliação de todas as *paste entities*. A Tabela 5.18 mostra a execução dos algoritmos tendo como entrada as *paste entities* de expressões regulares e palavras-chave, apenas. Na Tabela 5.19 é mostrado o resultado com apenas as expressões regulares corretamente aplicadas como entrada, enquanto a Tabela 5.20 apenas com as palavras-chave.

Tabela 5.16. Classificação: análise textual – base completa

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Multinomial NB	1.0	0,0	nan	0,0
KNN n=3	0,74	0,02	0,48	0,26
KNN n = 5	0,74	0,01	0,55	0,26
KNN n=7	0,76	0,01	0,59	0,24
Árvore de Decisão	0,58	0,07	0,24	0,42
SVM *C-Support Vector Classification* (SVC)	0,23	0,21	0,16	0,77

Tabela 5.17. Classificação: todas as *paste entities* – base completa

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,66	0,02	0,45	0,34
KNN n=3	0,79	0,0	0,72	0,21
KNN n = 5	0,79	0,0	0,85	0,21
KNN n=7	0,8	0,0	0,82	0,2
Árvore de Decisão	0,43	0,12	0,2	0,57
SVM SVC	0,24	0,23	0,15	0,76

5.5.2. Base de dados fracionada – 150 *pastes*

Visando compreender melhor o potencial de classificação dos *pastes*, foram realizados os mesmos testes em uma base de dados com a mesma quantidade de *pastes* (150) para relevantes e irrelevantes. Os *pastes* foram selecionados de forma aleatória. A Tabela 5.21 apresenta os resultados para a análise textual. A Tabela 5.22 mostra a avaliação de todas as *paste entities*. A Tabela 5.23 mostra a execução dos algoritmos tendo como entrada as *paste entities* de expressões regulares e palavras-chave, apenas. Na Tabela 5.24 é mostrado o resultado com apenas as expressões regulares corretamente aplicadas como entrada, enquanto a Tabela 5.25 apenas com as palavras-chave.

5.5.3. Considerações sobre a classificação

A aplicação dos algoritmos em condições normais, ou seja, com grande desbalanceamento, indica a necessidade de ajustes nas características de entrada e/ou nas próprias configurações

Tabela 5.18. Classificação: *paste entities*: Palavras-chave expressões regulares – base completa

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,65	0,02	0,44	0,35
KNN n=3	0,69	0,01	0,61	0,31
KNN n = 5	0,79	0,0	0,81	0,21
KNN n=7	0,8	0,0	0,84	0,2
Árvore de Decisão	0,42	0,03	0,47	0,58
SVM SVC	0,25	0,22	0,15	0,75

Tabela 5.19. Classificação: *paste entities*: Expressões regulares – base completa

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,83	0,01	0,42	0,17
KNN n=3	0,8	0,06	0,15	0,2
KNN n = 5	0,73	0,06	0,19	0,27
KNN n=7	0,92	0,0	0,7	0,08
Árvore de Decisão	0,13	0,26	0,15	0,87
SVM SVC	0,14	0,29	0,14	0,86

Tabela 5.20. Classificação: *paste entities*: Palavras-chaves – base completa

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,77	0,02	0,37	0,23
KNN n=3	0,86	0,0	0,76	0,14
KNN n = 5	0,86	0,0	0,89	0,14
KNN n=7	0,86	0,0	0,86	0,14
Árvore de Decisão	0,7	0,03	0,35	0,3
SVM SVC	0,69	0,02	0,51	0,31

Tabela 5.21. Classificação: análise textual – 150 *pastes*

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Multinomial NB	0,23	0,27	0,74	0,77
KNN n=3	0,25	0,23	0,77	0,75
KNN n = 5	0,27	0,22	0,77	0,73
KNN n=7	0,25	0,21	0,78	0,75
Árvore de Decisão	0,26	0,23	0,77	0,74
SVM SVC	0,33	0,2	0,77	0,67

Tabela 5.22. Classificação: todas as *paste entities* – 150 *pastes*

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,35	0,11	0,85	0,65
KNN n=3	0,49	0,12	0,81	0,51
KNN n = 5	0,38	0,1	0,86	0,62
KNN n=7	0,37	0,11	0,85	0,63
Árvore de Decisão	0,25	0,25	0,75	0,75
SVM SVC	0,46	0,14	0,79	0,54

dos algoritmos. Para além do acerto ser baixo, a quantidade de falsos positivos e falsos negativos é alta, de modo que muitos *pastes* relevantes se perdem, e muitos irrelevantes são classificados como relevantes, inundando assim de informações irrelevantes um possível

Tabela 5.23. Classificação: *paste entities*: Palavras-chave expressões regulares – 150 *pastes*

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,31	0,11	0,87	0,69
KNN n=3	0,47	0,05	0,92	0,53
KNN n = 5	0,39	0,06	0,91	0,61
KNN n=7	0,38	0,06	0,91	0,62
Árvore de Decisão	0,36	0,09	0,87	0,64
SVM SVC	0,5	0,11	0,82	0,5

Tabela 5.24. Classificação: *paste entities*: Expressões regulares – 150 *pastes*

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,11	0,29	0,75	0,89
KNN n=3	0,64	0,07	0,84	0,36
KNN n = 5	0,65	0,05	0,87	0,35
KNN n=7	0,65	0,05	0,87	0,35
Árvore de Decisão	0,13	0,29	0,75	0,87
SVM SVC	0,32	0,33	0,68	0,68

Tabela 5.25. Classificação: *paste entities*: Palavras-chaves – 150 *pastes*

Algoritmo	**TFN**	**TFP**	***Precisão***	***Abrangência***
Bernoulli NB	0,62	0,09	0,8	0,38
KNN n=3	0,78	0,02	0,92	0,22
KNN n = 5	0,68	0,03	0,92	0,32
KNN n=7	0,66	0,03	0,91	0,34
Árvore de Decisão	0,65	0,03	0,93	0,35
SVM SVC	0,7	0,02	0,94	0,3

analista.

Quando utilizado em uma situação de balanceamento, o acerto dos algoritmos cresce consideravelmente, indicando que, com os ajustes corretos das características de entrada eles podem ainda ser utilizados para auxiliar na determinação de relevância dos *pastes*.

5.6. Discussão das questões de pesquisa

Q1: Como identificar informações relevantes para a cibersegurança em *pastes* publicados no *Pastebin*?

Para a identificação de informações relacionadas à cibersegurança no *Pastebin*, expressões regulares e os bigramas, trigramas e quadrigramas foram os mais expressivos. A utilização de palavras-chave auxilia a garantir o acerto de expressões regulares, minorando a possibilidade de falhas das mesmas. A criação de uma *blacklist* que diminua a quantidade de informações irrelevantes e uma *acceptlist* que extraia, das previamente filtradas, as que de fato interessem é extremamente importante, dado o fluxo de publicações no *Pastebin*.

A comparação com outros trabalhos no que tange a esta questão é difícil, já que nenhum deles disponibiliza o processo completo utilizado ou seu resultado para apreciação.

Q2: Quais são as características das publicações associadas à cibersegurança disponibilizadas no *Pastebin*?

As características dos *pastes* relacionados à cibersegurança incluem, mas não se limitam as expressões regulares relativas as URLs, *e-mail*, números indicando documentos, telefones dos Estados Unidos e informações bancárias. As palavras-chave e palavras mais comuns incluem termos como *crack*, *btc* (acrônimo para *Bitcoin*), *gmail*, *yahoo* e *keygen*, demonstrando uma alta incidência de termos potencialmente ligados a vazamento de credenciais ou informações financeiras, para além de programas modificados. Dos bigramas, trigramas e quadrigramas foram extraídos termos como "account paypall", "steam key generator", "buy bank logs" e "buy dumps with pin", que reforça a constante aparição de publicações relacionadas à segurança bancária e a disseminação de programas modificados. Estas características são cobertas em maiores detalhes nas Seções 5.3 e 5.4.

Q3: As informações do *Pastebin* podem ser utilizadas para aviso antecipado de ameaças e/ou reações mais rápidas as mesmas?

Sim. Publicações como as exibidas na Tabela 5.3 e detalhadas no Apêndice B podem permitir, entre outras medidas, aviso aos usuários cujos dados tenham sido vazados, intensificação nos procedimentos de segurança de certos grupos e identificação de vulnerabilidades e/ou versões vulneráveis de programas.

A visibilidade e alcance dos *pastes* contendo informações potencialmente maliciosas ficou demonstrada com a apresentação dos *trend pastes* (os *pastes* mais acessados do *Pastebin*). É possível depreender que a quantidade de *pastes* relativos à cibersegurança entre os mais acessados é bastante grande, levando mesmo a retirada do ar desta página.

5.7. Criação e disponibilização da base de inteligência

O Apêndice E apresenta as configurações de todos os programas, algoritmos e classes em geral utilizadas. Além disso, todas as características obtidas, bem como o coletor desenvolvido foram disponibilizados em um repositório no *Github*, disponível em <https://github.com/felipeveigaramos/Pastebin_data>, cumprindo assim com os objetivos propostos neste trabalho.

Os dados disponibilizados são:

- Coletor: *software* utilizado para a coleta e armazenagem dos *pastes*,bem como as expressões regulares e palavras-chave utilizadas;

- *Software* para a auxílio à análise manual dos *pastes*;
- *Software* para a conversão do banco de dados dos *pastes* para os JSON utilizados;
- Programas desenvolvidos para processamento dos dados;
- Características geradas;
- Documentação descrevendo o processo e a utilização destes programas.

A disponibilização da base de dados completa, bem como das características geradas e *softwares* desenvolvidos é um diferencial deste trabalho em relação a todos os outros, citados na Seção 3.2, que, mesmo quando disponibilizam os *softwares* desenvolvidos, não disponibilizam nenhum tipo de dado extraído ou gerado por eles. Além disso, nenhum dos trabalhos citados aplica algoritmos de classificação sobre os dados ou tenta identificar características, como termos frequentes.

5.8. Considerações gerais

Este capítulo apresentou e discutiu os resultados obtidos com a aplicação do método proposto. Foram coletados 254.459 *pastes* em 21 dias, dos quais 3.650 foram manualmente analisados, levando a identificação de 183 relevantes e 3167 irrelevantes. Foram identificados termos comuns, tanto indicativos de relevância como irrelevância, apresentadas as expressões regulares e seus acertos e idiomas mais comuns associados a *pastes*. Também foi explicitada a forma de disponibilização dos dados gerados, através de um repositório no *github*, em <https://github.com/felipeveigaramos/Pastebin_data>. O próximo capítulo conclui este trabalho e apresenta os trabalhos futuros.

Capítulo 6

Conclusões

A disseminação de informações sensíveis no contexto de cibersegurança na *Web* é danosa para pessoas e instituições. Neste trabalho, objetivou-se a identificação e extração de informações relacionadas à cibersegurança em *pastes* do *Pastebin*. Estas informações foram encontradas e foi possível identificar padrões a partir delas.

Para os 3.650 *pastes* resultantes da análise manual, dos quais 183 foram considerados relevantes, foram extraídas as palavras mais comuns, bem como os bigramas, trigramas e quadrigramas. Muitos deles guardaram forte relação com dados bancários e/ou financeiros (como "cvv", "acount paypal" ou "buy bank log"). Foi possível identificar também diversas expressões regulares indicativas de irrelevância para a composição de uma nova *blacklist* mais precisa. Expressões que podem, em combinação com outras, serem usadas para identificar conteúdo relevante também foram encontradas, como as indicativas de número de documentos ou endereços. Foi possível filtrar a lista de expressões regulares com base em seu acerto e relevância.

Algoritmos de classificação foram utilizados (KNN, SVM, *Naive Bayes* e Árvore de Decisão) e indicaram que, dado o grande desbalanceamento entre as classes, as características extraídas não são as mais indicadas para classificar os *pastes*. No entanto, quando balanceado, ou seja, com o mesmo número de *pastes*, foi possível perceber que, com melhorias nas características, a classificação pode ser um meio de minimizar o trabalho de analistas.

Os tipos de *pastes* postados no *Pastebin* possuem formatos e conteúdos variados. Isto torna sua análise e classificação extremamente complexa. O tamanho das publicações, o tipo de informação, o propósito e o formato variam muito, dificultando a criação de automações com base em padrões. Quanto à coleta dos *pastes*, foi possível ter acesso apenas a *pastes* públicos, por limitações da API do *Pastebin*. A despeito disso, as informações encontradas no *Pastebin* podem ser de grande ajuda na tomada de medidas proativas ou reativas em relação a ciberameaças. Tais informações existem. Os processos para encontrá-las só precisam ser

refinados.

6.1. Trabalhos futuros

A partir deste trabalho diversas abordagens e/ou questões foram suscitadas, que pretende-se abordar em trabalhos futuros. Entre eles:

- Retroalimentação: utilização das características encontradas neste trabalho para a elaboração de novas *blacklist* e *acceptlist* e avaliação das mesmas;
- Melhoria nos classificadores e utilização de redes neurais;
- Análise de outros serviços para compartilhamento de texto, incluindo aqueles localizados na *deep web*;

Apêndices

Apêndice

A

Lista de Endereços Web

Chopapp	<http://chopapp.com/>
Codepad	<http://codepad.org/>
dPaste	<http://dpaste.com/>
Dumpz	<dumpz.org>
Facebook	<https://www.facebook.com/>
Github	<http://www.github.com/>
Github Gist	<https://gist.github.com/>
Have I been pwned	<https://haveibeenpwned.com/>
Hórus CEWS	<https://horus.rnp.br/>
IRC	<http://www.irc.org/>
PasteBin	<http://pastebin.com/>
Piratepad	<http://piratepad.net/>
Slexy	<http://slexy.org/>
Snipt	<http://www.snipt.org/>
Twitter	<https://www.twitter.com/>
Zerobin	<http://sebsauvage.net/paste/>

Endereços Web acessados em 12/12/2017.

Apêndice

B

pastes relevantes para a cibersegurança

Neste capítulo serão apresentadas amostras de *pastes*, relevantes e irrelevantes, encontrados durante a análise manual e que compõem a base dos 3.650 *pastes* manualmente analisados. Para cada *paste* será exibido parte do texto, com quaisquer informações sensíveis rajadas e as *paste entities* que foram associadas. A Seção B.1 mostra exemplos de *pastes* relevantes enquanto a Seção B.2, dos irrelevantes. A Seção B.3 apresenta exemplos de *pastes* relacionados à pornografia infantil, obtidos por análise manual. Embora não seja o foco direto deste trabalho, o tema é sensível e relevante no contexto de cibersegurança. Para detalhes sobre a nomenclatura das expressões regulares e a relevância de cada *paste entity*, consultar os APêndices C e D.

B.1. *Pastes* relevantes

Estes *pastes* fazem parte da base de dados dos 183 *pastes* relevantes e exemplificam *pastes* que se deseja obter:

- Texto:

```
arcgis 9 3 crack  - torrentkim12.com:  16-04-03 09:23
 arcgis 9 3 crack [] arcgis 9 3 crack .. README-INSTALL.txt: 2 KB: datainterop:connect.cesta:
 How to Crack ArcGIS 9.3
9- copy *all* files from licenseservercrack to [redacted]-dir in your C .. 16- Now run the lmtoo
```

  Paste entities:lg_en, kw_crack, re_url.

- Texto:

```
You can decrypt the hashed password also if you have the hash . Do not listen
to this individual claiming themself to be a "SA-MP expert".. Samp Account Password Hack Free. Samp Acc
Hack Software at Xentrik.. . if you want to know how to hack Facebook account online for free, . way to
(...)
```

Paste entities:kw_cracking, kw_secure, kw_hacked, kw_crack, re_url, kw_pass, lg_en, kw_hackers, kw_password, kw_hack, kw_decrpypt, kw_hacker, kw_hacking.

- Texto:

```
mi[redacted]nir20@hotmail.com:102blackout9384756S
gi[redacted]tx@gmail.com:ded[redacted]es
zey[redacted]ahtiyan@gmail.com:Port[redacted]01
(...)
```

Paste entities:re_email, lg_en,

- Texto:

```
Win7 KEYS 2018 http://the[redacted]zy.com/
```

Paste entities:lg_de, re_url

- Texto:

```
Free SOCKS4/5 Proxy List - 23/02/18
(...)
82.[redacted].1[redacted].72:51948:NO:Socks5
```

paste entities:re_ipv4, lg_en, re_url;

- Texto:

```
ga[redacted]d@yahoo.com:613[redacted]8 Used: 0/50GB Phone Pics                                    0KB
20150717_[redacted]01.jpg
```

Paste entities: re_email, lg_id;

- Texto:

```
exposed girl pack http://destyy.com/w[redacted]K
```

Paste entities: re_url, lg_en;

- Texto:

```
private australian combolist look after more here :
http://linkshrink.net/7DI[redacted]8
mah[redacted]ad@online.nl:Mas[redacted]ena
(...)
```

Paste entities: re_br_cpf, kw_pirate, kw_op, lg_en, re_url, re_email;

- Texto:

```
sergi[redacted]1@yahoo.com:19[redacted]da|Subscription:Premium For Family|Renew:|Country:US
(...)
```

Paste entities:re_email, lg_en;

- Texto:

```
1. Name : Fai  ab  l ha  d
2. Age 223. Address : 51 tet    od bl  d Ontario
4. Cell number: 519    40
(...)
5. Number linked to address:519     797
6. Skype : live: savag     95_1 (...)
(...)
(...)
15. Facebook:
(...)
Relations / friends : sabrine     Mona
M   m Kadrie A       d
(...)
1. Sister1 : sophia a       id
2. Age:24
(...)
```

paste entities: kw_leak,re_email, re_ipv4, lg_en, kw_password, kw_attacks, kw_vpn, kw_flood, kw_targets, re_url, kw_ddos.

B.2. *Pastes* irrelevantes

Estes *pastes* fazem parte da base de dados dos 3.467 *pastes* irrelevantes e exemplificam os *pastes* que se deseja descartar:

- Texto:

```
sudo apt-get autoremove eclipse*
Reading package lists... Done
(...)
```

Paste entities: lg_en;

- Texto:

```
Exception of type 'System.InvalidOperationException': No free spawnpoint.
   at OpenRA.Mods.Common.Traits.MPStartLocations.ChooseSpawnPoint(World world, List`1 available, List`1
   at OpenRA.Mods.Common.Traits.MPStartLocations.WorldLoaded(World world, WorldRenderer wr)
(...)
```

Paste entities:lg_en;

- Texto:

```
{           dog:[                   {                           name:'Rufus',
breed:'labrador' },
(...)
```

Paste entities:lg_en;

- Texto:

```
#include<iostream>
#include<cmath>
using namespace std;
int main() {
    int n, impare = 0, pare = 0;
(...)
```

Paste entities:lg_ru;

- Texto:

```
$56;EU;Kirin Tor;Rezia;;2;1;110;13:31,15:35,12:39,1:447,14:568,6:680;2;.s1;34;;1136,1,2,2,1,1,2,1,2,1,1
(...)
```

Paste entities: lg_id;

- Texto:

```
--[[Nocturna_A._Alteriano_Night.Owl.Inc]]-- (Edit by kiukiubol & lilh3)
wait(1 / 60)
p = game.Players.LocalPlayer
char = p.Character
torso = char.Torso
neck
(...)
```

Paste entities: re_sql,lg_en, kw_attack, re_*1,

- Texto:

```
shortlog: Commit header (keep it under 50 chars)
Body of the commit message. Feel free to expand on the changes.
Use the Body to explain why and what vs how
(...)
```

Paste entities: lg_en;

- Texto:

```
212th Commander Cody - Last Wednesday at 10:24 PM
Oh, i cant code so I couldnt help
thought it was in game
February 22, 2018
Feds - Last Thursday at 5:21 AM
(...)
```

Paste entities: lg_en;

- Texto:

```
Ron
```

Paste entities: lg_ru;

- Texto:

```
<html>
  <head>
    <title>Readit</title>
    <%= csrf_meta_tags %>
(...)
```

Paste entities:lg_en, re_html_xml.

B.3. *pastes* relacionados à pornografia infantil

Esta seção mostra exemplos de *pastes* relacionados à pornografia infantil, encontrados através de análise manual e não necessariamente parte dos 3.650 *pastes* estudados neste trabalho. Optou-se por apresentar tais exemplos dado que o compartilhamento de pornografia infantil, embora não seja o foco deste trabalho, é um grave problema relacionado à cibersegurança.

- Título: cp sites Texto:

```
1: http://[redacted]teen.w[redacted]k/
2: http://[redacted]ngis.sci[redacted]/
3: http://fap[redacted].w[redacted]k/
(...)
```

- Texto:

```
Random assortment of CP/JB-related links (unvalidated, some of them
might not work or might be fake, and there may be duplicates):

http://jmliqq[redacted].onion/
http://7ha[redacted].onion/
http://7o6[redacted].onion/
(...)
```

- Título:89 CP sites #TangoDown permanently | more to come; Texto:

```
http://www.youtube.com/watch?v=uCPQDfPE5nM

#APH #AnonymousPedoHunters #OpFreeHost #Anonymous
_____________________________________________________________________


*Greetz from the LulzShip

Greetings once again internet users, we are Anonymous.

Our last attack when we took down the 12 child porn sites gained great
support from people everywhere and we thank you for that, without you
the world would not know our cause and this entire operation would be
not nearly as effective.

A post was made to twitter about a formspring user who was posting child
porn to his formspring account. The email for formsprings staff was
posted to twitter and some concerned users emailed it to complain about
```

```
the user. The users account was suspended and we recieved and email
informing us that they were looking into him, he will most likely get
arrested. #Victory

Now we have good news and bad news; the good news being: we have
succesfully attacked and removed 89 child porn sites from the .webs
service. The bad news is that when we run some google queries it shows
that there is about 197,000 websites on the .webs service hosting CP
(this is 100% legitimate). It is depressing to see a number like that
involved with something like this. BUT WE WILL NOT GIVE UP! Our attacks
will not cease until every child porn site on the internet is erased and
every pedophile who has ever done harm is punished.

We invite .webs to aid us in the eradication of child porn websites from
their servers, they already know we mean business and if they refuse to
remove the CP sites we will take them down by force. This is not a
threat, it is a promise.

All these pedos must be getting really butthurt by now. lulz. And they
are about to get a huge dose of extra strength butthurt.

Sites suspended:
_____________________________________________________________________

http://[redacted].webs.com/[redacted]
http://uet[redacted].webs.com/[redacted]
http://[redacted].webs.com/[redacted]
(...)
```

- Título: #OpLiberation - SEMA D0X Texto:

```
#Anonymous #OpLiberation #ShutDownSEMA

We are Anonymous.
We are legion.
We do not forgive.
We do not forget.

Mr. We[redacted]an,
We are Legion. We will never forgive what you have done to the children in
Florida, We will never forget the pain you have inflicted, Expect Us Mr.
We[redacted]an.

Judge Dan V[redacted],
You have allowed criminals who have commited horrific crimes to go free. You
Have failed at your job of keeping them off the streets and you let them go
knowing that the suffering of children by their hand would continue. You are
not true justice, we are. Expect us.


_____________________________________________________________________


Target: Southeastern Military Academy

Reason: http://dar[redacted].tumblr.com/
(...)
Admin emails:
[redacted]we[redacted]an@aol.com    | Registered FB:
facebook.com/al[redacted]an.5
```

```
(...)
DOXES
------------------------


Note: Mo[redacted] is Mo[redacted]
We[redacted]an and lives at the same house as
Al[redacted] We[redacted]an
(...)
```

Apêndice C

Nomenclatura das expressões regulares

Diversas expressões regulares foram utilizadas neste trabalho. Muitas delas foram recolhidas a partir de outras ferramentas ou através de padrões identificados a partir da análise dos dados. Algumas possuem identificações óbvias como, por exemplo, "url", mas, para outras, não existem nomes específicos, por serem partes de algum tipo de registro ou informação desconhecida. A seguir serão descritas as expressões regulares e listadas suas identificações. Todas as expressões regulares são prefixadas com "re_" (*regular expression*). Na Tabela C.1 são detalhadas as expressões da *blacklist*, enquanto na Tabela C.2 as da *acceptlist*.

Tabela C.1. Expressões regulares: detalhe de nomes – *blacklist*

Tipo da expressão	**Lista de nomes**
Trechos de código	re_css, re_html_xml, re_java, re_php, re_python_error, re_skuid, re_sql, re_sqlserver, re_vasya_procedure
Logs de erros	re_*1, re_*2
Mensagens diversas	re_#1, re_#2, re_#3, re_#4, re_#5, re_#6, re_#7, re_#8, re_#9, re_#10, re_#11, re_#12, re_#13, re_#14

Tabela C.2. Expressões regulares: detalhe de nomes – *acceptlist*

Tipo da expressão	**Lista de nomes**
URL	re_url
E-mails	re_email
IP	re_ipv4, re_ipv4_with_port, re_ipv6, re_ipv6_with_port
Chaves criptográficas e *hashes*	re_cisco_hash, re_cisco_pass, re_google_api, re_hash32, re_pgp_private, re_ssh_private
Número de documentos	re_br_cnpj, re_br_cpf, re_us_social_insurance_number, re_us_social_security_number
Modos de endereçamento	re_br_cep, re_canadian_postal_code, re_us_zip_code
Números de cartão de crédito	re_creditcard_americanexpress, re_creditcard_dinersclub, re_creditcard_discover, re_creditcard_jcb, re_creditcard_mastercard, re_creditcard_visa
Número de telefone	re_br_phone, re_us_phone
Termos e nomes de grupos conhecidos	re_fff, re_honeypot, re_lulz
Endereço de carteiras *Bitcoin*	re_bitcoin_address

Apêndice D

Detalhes da análise das *paste entities*

Este capítulo apresenta, em detalhe, a análise das diversas *paste entities* encontradas. São abordadas, na Seção D.1 as expressões regulares, na Seção D.2 as palavras-chave e na Seção D.3 os idiomas. Os resultados obtidos a partir da análise destas *paste entities* são discutidos no capítulo 5.

D.1. *Paste entities*: expressões regulares

Durante a análise, 40 expressões regulares foram encontradas. Elas são cobertas em detalhes na tabela D.1. Detalhes sobre A nomenclatura das mesmas é explicada no Apêndice C. Para cada expressão regular são detalhados os seguintes campos:

- Identificação: identificação da expressão regular (sempre precedida de "re_");
- Lista: a qual lista pertence, *blacklist* – indicativa de irrelevância – ou *acceptlist* – indicativa de relevância;
- Corretude: taxa de acerto da expressão regular com base no contexto: percentual de vezes em que o padrão representado pela expressão regular correspondeu ao tipo de dado esperado para aquele contexto, por exemplo, que os números separados por pontos identificados pela expressão "re_ipv4" realmente eram um endereço IPv4;
- Relevantes: percentual de aparição da expressão regular entre os 183 *pastes* relevantes;
- Irrelevantes: percentual de aparição da expressão regular entre os 3.467 *pastes* irrelevantes.

Tabela D.1. *Paste entities*: expressões regulares

Identificação	*Lista*	**Corretude**	**Relevantes**	**irrelevantes**
re_#10	*blacklist*	91.67	0.0	0.32
re_#13	*blacklist*	96.67	0.0	1.5
re_#14	*blacklist*	100.0	0.0	0.23
re_#3	*blacklist*	100.0	0.0	0.03
re_#5	*blacklist*	100.0	0.0	0.29
re_#6	*blacklist*	100.0	0.0	0.03
re_#8	*blacklist*	33.33	0.0	0.03
re_ast1	*blacklist*	98.66	0.0	4.36
re_ast2	*blacklist*	100.0	0.0	0.37
re_bitcoin_address	*acceptlist*	2.28	0.55	0.14
re_br_cep	*acceptlist*	29.41	2.19	0.03
re_br_cnpj	*acceptlist*	2.82	0.0	0.12
re_br_cpf	*acceptlist*	2.22	3.28	0.03
re_br_phone	*acceptlist*	63.04	0.55	0.0
re_canadian_postal_code	*acceptlist*	0.0	0.0	0.0
re_creditcard_americanexpress	*acceptlist*	0.0	0.0	0.0
re_creditcard_dinersclub	*acceptlist*	20.0	0.0	0.03
re_creditcard_discover	*acceptlist*	0.0	0.0	0.0
re_creditcard_jcb	*acceptlist*	0.0	0.0	0.0
re_creditcard_mastercard	*acceptlist*	0.0	0.0	0.0
re_creditcard_visa	*acceptlist*	80.11	6.56	0.0
re_css	*blacklist*	94.12	0.0	1.24
re_email	*acceptlist*	97.73	26.23	1.79
re_google_api	*acceptlist*	50.0	0.0	0.03
re_hash32	*acceptlist*	2.13	1.09	0.2
re_html_xml	*blacklist*	99.02	0.0	1.5
re_ipv4	*acceptlist*	99.74	6.56	1.01
re_ipv6	*acceptlist*	4.73	0.0	0.09
re_ipv6_with_port	*acceptlist*	1.2	0.0	0.03
re_java	*blacklist*	96.67	0.0	1.93
re_php	*blacklist*	100.0	0.0	0.43
re_python_error	*blacklist*	100.0	0.0	0.26
re_skuid	*blacklist*	97.78	0.0	0.32
re_sql	*blacklist*	81.05	1.09	3.37
re_sqlserver	*blacklist*	85.71	0.0	0.14
re_url	*acceptlist*	82.41	69.4	28.61
re_us_phone	*acceptlist*	2.78	7.65	0.49
re_us_social_insurance_number	*acceptlist*	0.0	0.0	0.0
re_us_social_security_number	*acceptlist*	12.66	6.01	0.2
re_us_zip_code	*acceptlist*	2.75	8.2	0.43

D.2. *Paste entities*: palavras-chave

Foram avaliadas 124 palavras-chave, das quais foram efetivamente encontradas 113, que são descritas na Tabela D.2. Todas as palavras-chave foram utilizadas na *acceptlist*. Para cada palavra-chave é informado:

- Identificação: palavra-chave buscada;

- Relevantes: percentual de aparição da palavra-chave entre os 183 *pastes* relevantes;
- Irrelevantes: percentual de aparição da palavra-chave entre os 3.467 *pastes* irrelevantes.

Tabela D.2. *Paste entities*: palavras-chave

Identificação	**Relevantes**	**Irrelevantes**
abuse	0.0	0.17
anon	0.0	0.17
anons	0.0	0.06
anonymous	0.55	0.32
attack	0.55	1.5
attacked	0.0	0.17
attacker	0.55	0.09
attackers	0.0	0.06
attacking	0.0	0.23
attacks	0.55	0.26
backdoor	0.0	0.03
bitcoin	0.55	0.17
booter	0.0	0.06
bot	0.55	0.58
botnet	0.0	0.06
bots	0.55	0.26
bruteforce	0.55	0.0
btc	12.02	0.06
bug	0.55	0.49
bypass	0.55	0.26
cert	0.0	0.2
crack	19.13	1.33
cracked	2.19	0.52
cracker	0.0	0.03
cracking	0.55	0.09
cve	0.55	0.0
ddos	0.55	0.03
decrypt	0.55	0.09
deface	0.0	0.03
denial	0.0	0.06
dos	0.55	0.49

Continua na próxima página

Tabela D.2. *Paste entities*: palavras-chave

Identificação	**Relevantes**	**Irrelevantes**
down	1.64	3.49
dox	1.09	0.0
dropping	0.0	0.14
encrypt	0.55	0.32
encrypted	0.0	0.43
encryption	0.55	0.17
error	1.09	4.56
evil	0.0	0.69
exploit	0.0	0.06
exploitation	0.0	0.03
exploits	0.0	0.03
expose	0.0	0.09
firewalls	0.0	0.03
flood	0.55	0.2
gpg	0.55	0.03
hack	3.83	0.49
hacked	1.09	0.03
hacker	3.28	0.32
hackers	0.55	0.09
hacking	2.19	0.09
hacks	0.0	0.09
illegal	0.0	0.26
infect	0.0	0.03
inject	0.55	0.32
injection	1.09	0.26
insecure	0.0	0.06
ipsec	0.0	0.03
keylogger	0.0	0.03
kills	0.0	0.23
leak	0.55	0.29
leaks	0.55	0.06
malicious	0.0	0.17
malware	1.09	0.06
offline	2.19	0.58

Continua na próxima página

Tabela D.2. *Paste entities*: palavras-chave

Identificação	**Relevantes**	**Irrelevantes**
op	1.64	0.92
operations	0.0	0.35
overwrite	0.0	0.09
owned	0.55	0.26
pass	2.73	1.5
password	7.65	1.87
pastebin	0.55	0.75
patch	3.28	1.21
patched	0.0	0.03
patches	1.09	1.1
pirate	0.55	0.12
piratebay	0.55	0.0
privacy	0.55	0.32
protection	0.55	0.46
reflected	0.0	0.03
rootkit	0.0	0.03
scan	1.09	0.49
scanner	0.0	0.52
scanning	0.0	0.12
scans	0.0	0.09
sec	0.55	0.69
secure	2.73	0.78
secured	0.55	0.14
security	1.64	1.41
shut	0.0	0.29
shutdown	0.0	0.26
sniff	0.0	0.03
spying	0.0	0.03
ssl	0.0	0.52
stealing	0.0	0.12
stole	0.0	0.09
tango	0.0	0.03
target	1.09	2.97
targeting	0.0	0.03

Continua na próxima página

Tabela D.2. *Paste entities*: palavras-chave

Identificação	Relevantes	Irrelevantes
targets	0.55	0.29
thc	0.0	0.03
threat	0.0	0.23
tor	11.48	0.55
trojan	0.55	0.0
tunnel	0.0	0.17
twitter	1.09	0.37
virus	2.19	0.37
viruses	0.0	0.03
vpn	1.09	0.14
vulnerabilities	0.0	0.03
vulnerability	0.0	0.03
vulnerable	0.55	0.09
xss	0.0	0.03

D.3. *Paste entities*: idiomas

Foram encontrados 33 idiomas diferentes nos *pastes* analisados do *Pastebin*. No entanto, a identificação de idiomas de *pastes* é potencialmente atrapalhada pelo tipo de conteúdo disponibilizado: não são compartilhados apenas textos. Incluem-se entre os conteúdos compartilhados – embora não se limitem a – trechos de código de programação, arquivos codificados e informações puramente numéricas. Portanto, detecção de idioma torna-se uma característica bastante inconstante para utilizá-la na identificação de *pastes* ou na correlação de *pastes* relativos à cibersegurança, embora seja uma forma de caracterizar os *pastes* encontrados. A Tabela

- Identificação: abreviatura do idioma (*EN*: inglês, *pt*: português, etc). *NO* indica os *pastes* cujo idioma não pode ser detectado;
- Quantidade: quantidade de *pastes* associados a este idioma;
- Relevantes: percentual de aparição do idioma entre os 183 *pastes* relevantes;
- Irrelevantes: percentual de aparição do idioma entre os 3.467 *pastes* irrelevantes.

Tabela D.3. *Paste entities*: idiomas

Identificação	Quantidade	Relevantes	irrelevantes
af	7	0.0	0.2
bn	7	0.0	0.2
ca	58	2.73	1.53
cs	3	0.0	0.09
da	27	0.55	0.75
de	73	3.28	1.93
en	2679	76.5	73.23
es	43	0.0	1.24
et	7	0.0	0.2
fi	8	0.0	0.23
fr	93	1.64	2.6
hr	20	0.0	0.58
hu	9	0.0	0.26
id	25	1.09	0.66
it	37	0.55	1.04
ja	1	0.0	0.03
ko	7	3.28	0.03
lt	9	0.0	0.26
nl	14	0.55	0.37
no	20	0.0	0.58
pl	37	0.55	1.04
pt	43	0.55	1.21
ro	59	1.09	1.64
sk	7	0.0	0.2
sl	2	0.0	0.06
so	3	0.0	0.09
sq	92	0.0	2.65
sv	12	0.0	0.35
sw	1	0.0	0.03
te	1	0.0	0.03
tl	29	1.09	0.78
tr	5	1.09	0.09
vi	29	0.55	0.81
zh-tw	13	1.09	0.32

Apêndice E

Configuração dos Algoritmos Utilizados

No decorrer das análises desenvolvidas neste trabalho, diversos programas foram desenvolvidos, para realizar atividades específicas, como contagens de palavras e classificação. A seguir serão listadas as ferramentas utilizadas, suas versões e configurações.

Alguns dados foram os mesmos para todas as análises e, por serem, em geral, dependência, são listados a seguir:

- Python 3.7.0;
- Scipy 1.1.0;
- Numpy 1.15.1

E.1. Contagem de palavras

- API: Scikit-learn (sklearn): 0.20.0;
- Objeto: sklearn.feature_extraction.text.CountVectorizer:
  - decode_error: ignore;
  - strip_accents: unicode; : true;
  - preprocessor: None;
  - tokenizer: None
  - stop_words: None;
  - token_pattern: `(?iu)\b[a-zA-Z][a-zA-Z][a-zA-Z]+\b`;
  - ngram_range: não foi setado;
  - analyzer: word;
  - max_df: 0.85;
  - min_df: 0.05
  - max_features: None;
  - vocabulary: None;

  - binary: False;
  - type: não setado.

E.2. Classificadores

Todos os classificadores utilizados são da API *Scikit Learn*.

- Objeto: sklearn.naive_bayes.BernoulliNB:
  - alpha = 1.0;
  - binarize = 0.0;
  - fit_prior = True;
  - class_prior = None.
- Objeto: sklearn.naive_bayes.MultinomialNB:
  - alpha = 1.0;
  - fit_prior = True;
  - class_prior = None.
- Objeto: sklearn.neighbors.KNeighborsClassifier:
  - n_neighbors = 3,5 e 7;
  - weights = "uniform";
  - algorithm = "auto";
  - leaf_size = 30;
  - p = 2;
  - metric="minkowski";
  - metric_params=None;
  - n_jobs=None.
- Objeto: sklearn.tree.DecisionTreeClassifier:
  - criterion="gini";
  - splitter="best";
  - max_depth=None;
  - min_samples_split=2;
  - min_samples_leaf=1;
  - min_weight_fraction_leaf=0.0;
  - max_features=None;
  - random_state=None;
  - max_leaf_nodes=None;
  - min_impurity_decrease=0.0;
  - min_impurity_split=None;
  - class_weight="balanced";
  - presort=False.

- Objeto: sklearn.svm.SVC:
  - C=1.0;
  - kernel="rbf";
  - degree=3;
  - gamma="auto_deprecated";
  - coef0=0.0;
  - shrinking=True;
  - probability=False;
  - tol=0.001;
  - cache_size=200;
  - class_weight="balanced";
  - verbose=False;
  - max_iter=-1;
  - decision_function_shape="ovr";
  - random_state=None.

Referências

AGGARWAL, Charu C.; ZHAI, ChengXiang. A survey of text classification algorithms. In: *Mining Text Data*. [S.l.: s.n.], 2012.

AMISSAH, John K. Aramco of saudi arabia. Abril 2014.

ATLANTA DIVISION, IN THE UNITED STATES DISTRICT COURT FOR THE NORTHERN DISTRICT OF GEORGIA. *Case 1:11-cv-00458-WSD Document 8-1*. 2011. Http://attrition.org/errata/charlatan/gregory_evans/ligatt23/doe-d8.pdf. Acessada em 15/06/2017.

AVAST. *O que é botnet e como se proteger dela | Avast*. 2017. Https://www.avast.com/pt-br/c-botnet. Acessada em 07/06/2017.

BENJAMIN, Victor. *Securing Cyberspace: Analyzing Cybercriminal Communities through Web and Text Mining Perspectives*. Tese (Doutorado), 2016. Disponível em: <http://hdl.handle.net/10150/613280>.

BISHOP, Matt. *Introduction to Computer Security*. [S.l.]: Addison-Wesley Professional, 2004. ISBN 0321247442.

BITCOIN. *Bitcoin - Open source P2P money*. 2017. Https://bitcoin.org/. Acessada em 07/06/2017.

BOLHUIS, Peter van; SELIJ, Jan-Willem; RASPE, Steven; LADAN, Floris. Information loss to public networks. 2014.

BRENGEL, Michael; ROSSOW, Christian. Identifying key leakage of bitcoin users. In: *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings*. [s.n.], 2018. p. 623–643. Disponível em: <https://doi.org/10.1007/978-3-030-00470-5_29>.

BRIAN, Matt. *Pastebin: How a popular code-sharing site became the ultimate hacker hangout*. 2005. Https://thenextweb.com/socialmedia/2011/06/05/pastebin-how-a-popular-code-sharing-site-became-the-ultimate-hacker-hangout/#.tnw_e69i2AIL. Acessada em 09/05/2017.

BRIANBB. *BrianBB/jPastebin: A complete pastebin.com API wrapper for Java*. 2014. Disponível em: <https://github.com/BrianBB/jPastebin>.

BRIDGE, High-Tech. *300,000 Compromised Accounts Available on Pastebin*. 2014. Https://www.htbridge.com/news/300_000_compromised_accounts_available_on_pastebin.html. Acessada em 10/06/2017.

CAMPIOLO, Rodrigo. *Análise e extração de alertas antecipados sobre ameaças e incidentes de segurança em sistemas computacionais usando fontes de dados não estruturados*. Tese (Doutorado) — Instituto de Matemática e Estatística, Universidade de São Paulo, Setembro 2016.

CAMPIOLO, Rodrigo; BATISTA, Daniel Macêdo. Análise de mensagens associadas à cibersegurança em redes IRC. In: *Anais do XV Simpósio Brasileiro em Segurança da Informação e Sistemas Computacionais (SBSeg 2015)*. Florianópolis: [s.n.], 2015.

CAMURCA, FRANCISCO. *Cibersegurança ou segurança da informação? Explicando a diferença*. 2017. Https://www.welivesecurity.com/br/2017/01/17/ciberseguranca-ou-seguranca-da-informacao/. Acessada em 11/06/2017.

CASTELLUCCIA, Claude; CHAABANE, Abdelberi; DÜRMUTH, Markus; PERITO, Daniele. When privacy meets security: Leveraging personal information for password cracking. *arXiv preprint arXiv:1304.6584*, 2013.

CERT.BR. *Cartilha de Segurança para Internet*. 2. ed. São Paulo (SP): Comitê Gestor da Internet no Brasil, 2012. ISBN 978-85-60062-54-6.

CIO, Redação. *Ataques DDoS estão maiores e mais frequentes, confirma Arbor*. 2016. Https://www.thegoatblog.com.br/goat/index.php/seguranca-informacao/3-ataques-ddos-estao-maiores-e-mais-frequentes-confirma-arbor. Acessada em 05/06/2017.

DISTRIBUTION, Berkeley Software. *The 3-Clause BSD License | Open Source Initiative*. 1999. Disponível em: <https://opensource.org/licenses/BSD-3-Clause>.

DOUGLAS, David M. Doxing: A conceptual analysis. *Ethics and Inf. Technol.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 18, n. 3, p. 199–210, set. 2016. ISSN 1388-1957. Disponível em: <http://dx.doi.org/10.1007/s10676-016-9406-0>.

DUMP Monitor (@dumpmon) | Twitter. 2017. Https://twitter.com/dumpmon. Acessada em 08/06/2017.

EXCHANGEWAR. *9,976.50 | BTC/BRL Bitcoin/Brazilian Real exchange list | Exchange War*. 2017. Https://exchangewar.info/coinprice?btc_brl. Acessada em 11/06/2017.

FOUNDATION, Free Software. *The GNU General Public License v3.0 - GNU Project - Free Software Foundation*. 2007. Disponível em: <https://www.gnu.org/licenses/gpl-3.0.en.html>.

FOUNDATION, The Apache Software. *Apache License, Version 2.0*. 2004. Disponível em: <https://www.apache.org/licenses/LICENSE-2.0>.

G1. *G1 - Site da Anatel caiu após ser alvo de ataque hacker - notícias em Tecnologia e Games*. 2016. Http://g1.globo.com/tecnologia/noticia/2016/04/site-da-anatel-caiu-apos-ser-alvo-de-ataque-hacker.html. Acessada em 05/06/2017.

GAIKAR, Vishal. *Top 5 Best Pastebin Alternative Websites*. [S.l.]: Tricks Machine, 2013. Http://www.tricksmachine.com/2013/06/top-5-best-pastebin-alternative-websites.html. Acessada em 07/06/2017.

GITHUB - jordan-wright/dumpmon: Information Dump Monitor. 2017. Https://github.com/jordan-wright/dumpmon. Acessada em 08/06/2017.

HEFFELFINGER, Christopher. The risks posed by jihadist hackers. *CTC Sentinel*, v. 7, 2013.

HOTFORSECURITY. *Keyloggers Posting on Webpages - HOTforSecurity*. 2011. Https://hotforsecurity.bitdefender.com/blog/keyloggers-posting-on-webpages-831.html. Acessada em 10/06/2017.

HUANG, Danny Yuxing; DHARMDASANI, Hitesh; MEIKLEJOHN, Sarah; DAVE, Vacha; GRIER, Chris; MCCOY, Damon; SAVAGE, Stefan; WEAVER, Nicholas; SNOEREN, Alex C; LEVCHENKO, Kirill. Botcoin: Monetizing stolen cycles. In: . [S.l.: s.n.], 2014. Network and Distributed System Security Symposium.

HUNT, Troy. *Troy Hunt | Profile*. 2017. Https://app.pluralsight.com/profile/author/troy-hunt. Acessada em 08/06/2017.

HUNT, Troy. *Troy Hunt: Pastes on Have I Been Pwned Are No Longer Publicly Listed*. 2017. Https://www.troyhunt.com/pastes-on-have-i-been-pwned-are-no-longer-publicly-listed/. Acessada em 14/10/2018.

HUNT, Troy. *Troy Hunt: Troy Hunt*. 2017. Https://www.troyhunt.com/. Acessada em 08/06/2017.

ISO. *ISO/IEC 27000 Information technology – Security techniques – Information security management systems – Overview and vocabulary*. [S.l.], 2009.

JARGAS, Aurelio. *EXPRESSÕES REGULARES - Livro Online, por Aurelio Jargas*. 2001. Http://aurelio.net/regex/guia/. Acessada em 13/06/2017.

KASPERSKYLAB. *What is a Botnet? - Definition*. 2017. Https://www.kaspersky.com/resource-center/threats/botnet-attacks. Acessada em 07/06/2017.

KEV. *GitHub - kevthehermit/PasteHunter: Scanning pastebin with yara rules*. 2017. Https://github.com/kevthehermit/PasteHunter. Acessada em 03/05/2018.

KEV. *PasteHunter The Results – TechAnarchy*. 2017. Https://techanarchy.net/2017/12/08/pastehunter-the-results.html. Acessada em 03/05/2018.

KEV. *TechAnarchy – A Jekyll theme*. 2017. Https://techanarchy.net/. Acessada em 03/05/2018.

KONTAXIS, G.; POLAKIS, I.; IOANNIDIS, S. Outsourcing malicious infrastructure to the cloud. In: *2011 First SysSec Workshop*. [S.l.: s.n.], 2011. p. 35–42.

LEAKEDIN. 2017. Http://www.leakedin.com/. Acessada em 08/06/2017.

LOPER, Edward; BIRD, Steven. Nltk: The natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Stroudsburg, PA, USA:

Association for Computational Linguistics, 2002. (ETMTNLP '02), p. 63–70. Disponível em: <https://doi.org/10.3115/1118108.1118117>.

MANNING, Christopher D.; SCHüTZE, Hinrich. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1.

MIRANTE, Dennis; CAPPOS, Justin. Understanding password database compromises. *Dept. of Computer Science and Engineering Polytechnic Inst. of NYU, Tech. Rep. TR-CSE-2013-02*, 2013.

MOYER, Julia. Doxing: Dangers and defenses. 2016.

NAKATANI, Shuyo. *Language Detection Library for Java*. 2010. Disponível em: <https://github.com/shuyo/language-detection>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PENDERGRASS, William Stanley; WRIGHT, Michelle. A case study analysis of knightsec and the steubenville rape case. In: *Proceedings of the Conference for Information Systems Applied Research ISSN*. [S.l.: s.n.], 2014. v. 2167, p. 1508.

SANTOS, LUIZ ARTHUR F; CAMPIOLO, Rodrigo; GEROSA, MARCO AURELIO; BATISTA, DANIEL MACEDO. Análise de mensagens de segurança postadas no twitter. *Anais do simpósio brasileiro de sistemas colaborativos (SBSC)*, n. 3, p. 20–28, 2012.

SANTOS, L. A. F.; CAMPIOLO, R.; GEROSA, M. A.; BATISTA, D. M. Detecção de alertas de segurança em redes de computadores usando redes sociais. In: *Anais do XXXI SBRC*. [S.l.: s.n.], 2013. p. 791–804.

SHAKARIAN, Jana; SHAKARIAN, Paulo; RUEF, Andrew. Cyber attacks and public embarrassment: A survey of some notable hacks. *arXiv preprint arXiv:1501.05990*, 2015.

SHIFRIN, Marina. *MasterCard Credit Card Numbers Leaked by WikiLeaks Supporters | MyBankTracker*. 2017. Https://www.mybanktracker.com/news/2010/12/08/mastercard-credit-card-numbers/. Acessada em 10/06/2017.

SOUSA, Vítor Hugo Silva et al. *Plataforma para análise de fugas de informação na World Wide Web*. Tese (Doutorado), 2016.

SQUIRE, Megan; SMITH, Amber K. The diffusion of pastebin tools to enhance communication in floss mailing lists. In: _____. *Open Source Systems: Adoption and Impact: 11th IFIP WG 2.13 International Conference, OSS 2015, Florence, Italy, May 16-17, 2015, Proceedings*. Cham: Springer International Publishing, 2015. p. 45–57. ISBN 978-3-319-17837-0. Disponível em: <http://dx.doi.org/10.1007/978-3-319-17837-0_5>.

TECHOPEDIA. *What is a Cracker? - Definition from Techopedia*. 2017. Https://www.techopedia.com/definition/10257/cracker. Acessada em 02/06/2017.

TECHOPEDIA. *What is a Data Breach? - Definition from Techopedia*. 2017. Https://www.techopedia.com/definition/13601/data-breach. Acessada em 08/06/2017.

TECHOPEDIA. *What is a Hacker? - Definition from Techopedia*. 2017. Https://www.techopedia.com/definition/3805/hacker. Acessada em 02/06/2017.

TECHOPEDIA. *What is an Exploit in Computing? - Definition from Techopedia*. 2017. Https://www.techopedia.com/definition/4275/exploit. Acessada em 02/06/2017.

TECHOPEDIA. *What is Cybersecurity? - Definition from Techopedia*. 2017. Https://www.techopedia.com/definition/24747/cybersecurity. Acessada em 02/06/2017.

TECHOPEDIA. *What is Hacktivism? - Definition from Techopedia*. 2017. Https://www.techopedia.com/definition/2410/hacktivism. Acessada em 02/06/2017.

TRENDMICRO. *Data Breach - Glossário de termos - Trend Micro BR*. 2017. Http://www.trendmicro.com.br/vinfo/br/security/definition/data-breach. Acessada em 08/06/2017.

WAQAS. *OpIsrael: Hackers leak 820 Israeli login data, deface 100+ websites*. 2015. Https://www.hackread.com/opisrael-hackers-leak-israeli-emails-hack-sites/. Acessada em 05/06/2017.

WARREN, Matthew J; LEITCH, Shona. Australia and data retention. *The possibilities of ethical ICT*, Print & Sign University of Southern, p. 511, 2013.

WEB Crawler. 2017. Https://www.linux.ime.usp.br/ cef/mac499-06/monografias/andre/WebCrawler.html. Acessada em 15/06/2017.

WRIGHT, Jordan. *RaiderSec: Introducing dumpmon: A Twitter-bot that Monitors Paste-Sites for Account/Database Dumps and Other Interesting Content*. 2013. Https://raidersec.blogspot.com.br/2013/03/introducing-dumpmon-twitter-bot-that.html. Acessada em 08/06/2017.