Identificação de *Posts* Maliciosos na *Dark Web* Utilizando Aprendizado de Máquina Supervisionado

Sebastião Alves de Jesus Filho

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2023

Sebastião Alves de Jesus Filho

Identificação de *Posts* Maliciosos na *Dark Web* Utilizando Aprendizado de Máquina Supervisionado

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Rodrigo Sanches Miani

Uberlândia
2023

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

J58i
2023

Jesus Filho, Sebastião Alves de, 1982-
Identificação de posts maliciosos na dark web utilizando Aprendizado de Máquina Supervisionado [recurso eletrônico] / Sebastião Alves de Jesus Filho. - 2023.

Orientador: Rodrigo Sanches Miani.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2023.8127
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Miani, Rodrigo Sanches, 1983-, (Orient.). II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU: 681.3

André Carlos Francisco
Bibliotecário - CRB-6/3408

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação de Mestrado, 6/2024, PPGCO				
Data:	29 de janeiro de 2024	Hora de início:	14:05	Hora de encerramento:	16:40
Matrícula do Discente:	12112CCP028				
Nome do Discente:	Sebastião Alves de Jesus Filho				
Título do Trabalho:	Identificação de ameaças de segurança usando mineração de dados de redes sociais e darkweb (Iniciativa privada: DataRisk)				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Sistemas de Computação				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Silvio Ereno Quincozes - UNIPAMPA, Bruno Bogaz Zarpelão - CCE/UEL e Rodrigo Sanches Miani- FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Silvio Ereno Quincozes - Alegrete/Rio Grande do Sul, Bruno Bogaz Zarpelão- Londrina/PR . O Orientador e o aluno participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof . Dr. Rodrigo Sanches Miani, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação

interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Rodrigo Sanches Miani**, **Professor(a) do Magistério Superior**, em 31/01/2024, às 18:46, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Bogaz Zarpelão**, **Usuário Externo**, em 01/02/2024, às 09:20, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Silvio Ereno Quincozes**, **Usuário Externo**, em 01/02/2024, às 12:20, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5095787** e o código CRC **48FC5B03**.

Referência: Processo nº 23117.002448/2024-31 SEI nº 5095787

À Giovana e ao Pedro, vocês são a minha inspiração.

Agradecimentos

Muito tenho a agradecer a todos aqueles que, de uma forma ou de outra, contribuíram para que eu chegasse até aqui. Tenho certeza de que sozinho jamais teria sido capaz.

Em primeiro lugar, agradeço a Deus, que me deu a vida, me capacitou e sempre guiou os meus caminhos.

Agradeço aos meus pais, Tião Toco e Geralda, meus primeiros educadores, que sempre confiaram na minha capacidade e me incentivaram a continuar lutando diante de cada desafio encontrado.

Agradeço à minha esposa, Núbia, que sempre foi minha parceira, capaz de se sacrificar para que eu tivesse as condições necessárias para seguir em frente e jamais desistir.

Agradeço aos meus filhos, Giovana e Pedro, que, à sua maneira, transmitem-me muito carinho e inspiração.

Agradeço aos meus irmãos, Kezia, Nadson, Eucy e Fabiana, que também sempre me incentivaram e acreditaram em mim.

Gostaria de agradecer a todos os meus professores desde as séries iniciais até a pós-graduação; todos vocês foram muito importantes na minha formação. Obrigado por compartilharem o conhecimento de vocês comigo.

Agradeço aos meus colegas de trabalho da FEQUI-UFU pelo incentivo e apoio, que foram exemplos para mim na formação acadêmica.

Agradeço ao meu orientador, professor Rodrigo Sanches Miani, que me apontou o caminho a percorrer durante esses anos. Muito obrigado pelo apoio e dedicação.

Agradeço a todos os colegas do Grupo de Pesquisa: Previsão de Incidentes. Saibam que vocês foram fundamentais para que eu conseguisse desenvolver este trabalho.

"Nas grandes batalhas da vida, o primeiro passo para a vitória é o desejo de vencer."
(Mahatma Gandhi)

Resumo

Diante do crescimento constante e da sofisticação dos ataques cibernéticos, a segurança cibernética não pode mais depender exclusivamente de técnicas e ferramentas tradicionais de defesa. A detecção proativa de ameaças cibernéticas torna-se uma necessidade nos dias atuais para que as equipes de segurança possam identificar potenciais ameaças e adotar medidas de mitigação eficazes. A área de *Cyber Threat Intelligence* (CTI), ou Inteligência de Ameaças Cibernéticas, desempenha um papel fundamental ao fornecer aos analistas de segurança conhecimento fundamentado em evidências sobre ameaças cibernéticas. A extração de informações de CTI pode ocorrer por meio de diversas técnicas e envolver diferentes fontes de dados; no entanto, o uso de aprendizado de máquina tem se mostrado uma abordagem promissora nessa área. Quanto à fonte de dados, as redes sociais e fóruns de discussão online têm sido comumente explorados. Nesta dissertação, aplicam-se técnicas de mineração de texto, Processamento de Linguagem Natural (PLN) e aprendizado de máquina em dados coletados de fóruns da *Dark Web* com o objetivo de identificar *posts* maliciosos. A base de dados para treinamento foi rotulada levando em consideração a ocorrência de *Indicadores de Comprometimento* (IoCs), palavras-chave contextuais, além de análise manual. Diferentes algoritmos de classificação foram testados utilizando diversas formas de representações de texto para encontrar o melhor modelo. Os resultados revelaram que o modelo com o algoritmo *Light Gradient Boosting Machine* (LightGBM) e *Term Frequency* (TF) - *Inverse Document Frequency* (IDF) - *(TF-IDF - Unigram)* como representação de texto alcançou as melhores métricas de acurácia, precisão, revocação e medida-F. Adicionalmente, novos *posts* não rotulados foram submetidos ao classificador, apresentando resultados promissores ao serem analisados com auxílio de um algoritmo de modelagem de tópicos - *Latent Dirichlet Allocation* (LDA).

Palavras-chave: Ataques Cibernéticos, Segurança Cibernética, Inteligência de Ameaças Cibernéticas, Processamento de Linguagem Natural, Aprendizado de Máquina, Modelagem de Tópicos.

Abstract

In the face of the constant growth and sophistication of cyber attacks, cybersecurity can no longer rely solely on traditional defense techniques and tools. Proactive detection of cyber threats has become a necessity in today's world, enabling security teams to identify potential threats and adopt effective mitigation measures. The field of Cyber Threat Intelligence (CTI) plays a fundamental role by providing security analysts with evidence-based knowledge about cyber threats. Information extraction from CTI can occur through various techniques and involve different data sources; however, machine learning has proven to be a promising approach in this area. Regarding data sources, social networks and online discussion forums have been commonly explored. In this dissertation, text mining, Natural Language Processing (NLP), and machine learning techniques are applied to data collected from Dark Web forums with the aim of identifying malicious posts. The training dataset was labeled considering the occurrence of Indicators of Compromise (IoCs), contextual keywords, and manual analysis. Different classification algorithms were tested using various text representations to find the best model. The results revealed that the model using the LightGBM algorithm and TF-IDF (Term Frequency-Inverse Document Frequency) with Unigram representation achieved the best metrics of accuracy, precision, recall, and F1-score. Additionally, new unlabeled posts were submitted to the classifier, showing promising results when analyzed using Topic Modeling with Latent Dirichlet Allocation (LDA).

Keywords: Cyber Attacks, Cybersecurity, Cyber Threat Intelligence, Natural Language Processing, Machine Learning, Topic Modeling.

Lista de ilustrações

Lista de tabelas

Lista de siglas

API *Application Programming Interface*

CNN *Convolutional Neural Networks*

CVE *Common Vulnerabilities and Exposures*

CPE *Common Platform Enumeration*

CVSS *Common Vulnerability Scoring System*

CTI *Cyber Threat Intelligence*

CBOW *Continuous Bag of Words*

GBM *Gradient Boosting Machine*

GOSS *Gradient-based One-Side Sampling*

IP *Internet Protocol*

IoT *Internet of Things*

IoCs *Indicadores de Comprometimento*

IDF *Inverse Document Frequency*

LDA *Latent Dirichlet Allocation*

LightGBM *Light Gradient Boosting Machine*

NIST *National Institute of Standards and Technology*

NVD *National Vulnerability Database*

PLN *Processamento de Linguagem Natural*

SVM *Support Vector Machine*

Tor *The Onion Router*

TF *Term Frequency*

TI *Tecnologia da Informação*

TTPs *Tactics, Techniques and Procedures*

XGBoost *eXtreme Gradient Boosting*

Sumário

Capítulo 1

Introdução

Os avanços tecnológicos impulsionam a crescente conectividade global. Atividades como trabalho, lazer, educação, compras e transações financeiras têm motivado a necessidade de dispositivos computacionais interconectados em rede. No entanto, a crescente visibilidade online também traz consigo um aumento nos riscos de ataques cibernéticos, que frequentemente resultam em consequências adversas significativas para indivíduos e entidades.

A partir do primeiro trimestre de 2020, houve um aumento significativo de ações ilícitas na Internet como roubo de dados financeiros, extorsão e espionagem cibernética (BROOKS, 2021). Recentemente, empresas como Lojas Renner, JBS e *Colonial Pipeline* foram alvo de ataques de *ransomware*. Ao mesmo tempo, ocorreram diversos vazamentos de senhas, como o caso da empresa *RockYou* (GUGELMIN, 2021), e o vazamento de dados pessoais de cidadãos brasileiros (Olhar Digital, 2021), o que ilustra o impacto financeiro e social dos ciberataques. Um dos motivos desse aumento no número de ataques cibernéticos nos últimos anos se dá em consequência da pandemia de COVID-19, uma vez que muitas organizações adotaram o trabalho remoto sem tomar as devidas medidas de segurança contra esses ataques (DUTTA; KANT, 2020). Segundo Mador (2021), as comunidades de *hackers* confirmam um aumento de postagens discutindo a exploração da pandemia como uma nova oportunidade para ataques, principalmente direcionados a ferramentas de trabalho remoto.

Nesse contexto, a *Dark Web*, reconhecida por preservar o anonimato, torna-se um ambiente propício para o compartilhamento de informações entre cibercriminosos. A *Dark Web* é uma parte da Internet não acessível aos mecanismos de pesquisa e requer o uso de um navegador especial. Essa rede oculta e anônima abre caminho para atividades ilegais, colaborando para que cibercriminosos executem ciberataques de forma planejada e coordenada (SALEEM; ISLAM; KABIR, 2022).

Ainda de acordo com Saleem, Islam e Kabir (2022), especialistas em segurança cibernética concordam que as atividades criminosas online estão aumentando exponencialmente e estão se tornando cada vez mais desenfreadas e intensificadas. Essas atividades ciberné-

ticas ilegais incluem violação de dados, ataques de *ransomware* e diversos outros crimes. Assim, a preservação da privacidade e do sigilo dos dados se torna um novo dilema da era atual.

Basheer e Alkhatib (2021) também ressaltam que, na era da tecnologia, diante do constante avanço de técnicas e ferramentas de *hacking*, tornou-se uma necessidade urgente que as organizações adotem medidas preventivas eficazes para se proteger contra ataques cibernéticos e criminosos virtuais. A detecção proativa de ameaças à segurança cibernética representa uma iniciativa crucial e desafiadora para antecipar e identificar potenciais ataques antes que se concretizem.

Devido ao rápido desenvolvimento técnico, grande parte das atividades de *hacking* deixaram de ser apenas atos isolados de vandalismo e se transformaram em ações bem organizadas e muitas vezes financiadas visando grande retorno financeiro. Uma espécie de crime organizado que além do ganho financeiro pode envolver até motivações políticas (TOUNSI, 2019).

Diante desse cenário, ferramentas tradicionais como os filtros de pacotes e sistemas de detecção de intrusão parecem não ser mais suficientes para evitar o comprometimento da informação. O avanço da capacidade computacional dos sistemas digitais, aliado às *Tactics, Techniques and Procedures* (TTPs) (Táticas, Técnicas e Procedimentos) aprimorados pelos cibercriminosos, não corresponde aos mecanismos de segurança convencionais para detecção de intrusões e prevenção de ameaças no atual panorama de segurança cibernética (DUTTA; KANT, 2020). Com isso, cada vez mais pesquisadores e profissionais de segurança cibernética têm voltado suas atenções para uma nova geração sob demanda de ferramentas de segurança cibernética, conhecida como CTI ou Inteligência de Ameaças Cibernéticas (BASHEER; ALKHATIB, 2021). A CTI consiste em uma coleção de dados cujo objetivo é fornecer conhecimento baseado em evidências sobre ameaças cibernéticas. Tendo o conhecimento adquirido, as organizações podem tomar decisões quando a segurança, o que envolve a detecção, prevenção e recuperação de ataques cibernéticos (TOUNSI, 2019), (SARI, 2018).

De acordo com Sapienza et al. (2017), agentes maliciosos seguem uma série de etapas para conduzir ataques cibernéticos. Esses passos incluem a identificação de vulnerabilidades, a aquisição de ferramentas e habilidades, a seleção do alvo, a criação ou obtenção de infraestrutura, e o planejamento e a execução do ataque. Durante essas fases, os agentes maliciosos podem deixar rastros associados a atividades específicas, como tentativas de acesso a URLs incomuns ou a manipulação de listas de e-mails corporativos. Esses vestígios são conhecidos como IoCs (JO; LEE; SHIN, 2022), funcionando como uma espécie de impressão digital que pode ser observada por especialistas em segurança da informação. Além disso, exemplos adicionais de IoCs compreendem endereços *Internet Protocol* (IP), nomes de domínio e *hashes* de arquivos.

O termo IoCs está diretamente relacionado à CTI, de forma que o compartilhamento

de CTI pode se dar através dos IoCs. No entanto, eles não devem ser o único foco das estratégias de segurança cibernética, pois os IoCs oferecem apenas uma indicação de uma ameaça cibernética e podem se tornar imprecisos ao longo do tempo (JO; LEE; SHIN, 2022), (PREUVENEERS; JOOSEN, 2021). Tendo em vista o crescente número de ameaças cibernéticas e a sofisticação dos métodos de ataque, o compartilhamento de informações de inteligência que visa mitigar os riscos de comprometimento da informação é fundamental para a comunidade de segurança cibernética; entretanto, é necessário aprimorar as fontes de CTI. Basheer e Alkhatib (2021) destacam que a análise de conteúdo em plataformas *Dark Web* pode contribuir para a detecção e prevenção de crimes cibernéticos. Esses fatores indicam a importância do desenvolvimento de ferramentas que sejam capazes de extrair CTI de fontes como a *Dark Web* envolvendo, dentre outras coisas, a presença de IoCs.

Diante da necessidade de constante evolução da CTI, a incorporação de tecnologias avançadas se faz necessária; nesse contexto, o aprendizado de máquina surge como uma ferramenta fundamental. De acordo com Preuveneers e Joosen (2021), cada vez mais as plataformas de CTI têm adotado o aprendizado de máquina para processar grandes quantidades de dados estruturados e não estruturados de segurança cibernética. Basheer e Alkhatib (2021) destacam a importância do aprendizado de máquina e do processamento de linguagem para o futuro da CTI. Essas tecnologias são consideradas fundamentais para permitir uma resposta proativa e rápida diante das ameaças cibernéticas em constante evolução.

Entre os anos de 2017 e 2023, diversos trabalhos têm se concentrado na identificação de ameaças de segurança a partir da análise de conteúdo presente em fóruns da *Dark Web* (SAPIENZA et al., 2017), (DONG et al., 2018), (SARKAR et al., 2019) e (ARNOLD et al., 2019). No âmbito da análise e extração de IoCs, Niakanlahiji et al. (2019) investigam a existência de IoCs no *Twitter*, enquanto Zhang et al. (2019) extraíram IoCs realizando buscas na *Surface Web*, usando como entrada indicadores como domínios e endereços *Internet Protocol* (IP) que tenham sido considerados suspeitos de acordo com informações de ameaças de código aberto. Por sua vez, (CABALLERO et al., 2023) direcionam seus esforços para extrair IoCs de seis diferentes fontes: *Blogs RSS, Twitter* e *Telegram*, bem como *Malpedia, APTnotes* e *ChainSmith*, que são repositórios de projetos relacionados à segurança cibernética. No entanto, nenhum desses trabalhos que investigaram *IoCs* foi conduzido na *Dark Web*. Já os trabalhos (QUEIROZ; MCKEEVER; KEEGAN, 2019) e (KOLOVEAS et al., 2021), abordaram o uso de aprendizado de máquina em dados coletados da *Dark Web* para obtenção de CTI. Entretanto, esses trabalhos ainda deixaram lacunas quanto à identificação de conteúdo relevante em *posts* coletados da *Dark Web*, bem como na criação e compartilhamento de conjuntos de dados rotulados sobre o assunto.

A ausência de trabalhos que abordem conteúdo no idioma português do Brasil é notável. Além disso, os processos de rotulagem adotados nesses trabalhos não foram bem

detalhados. Alguns mencionaram a rotulagem manual, enquanto outros recorreram ao uso de palavras-chave. Contudo, não há relatos de um processo que tenha considerado simultaneamente a presença de IoCs, palavras-chave e análise manual.

Adicionalmente, a indisponibilidade de dados rotulados representa um desafio. Embora o trabalho de (QUEIROZ; MCKEEVER; KEEGAN, 2019) tenha fornecido uma URL para download da base, a mesma não se encontrava ativa durante o desenvolvimento deste trabalho. Vale destacar também que os trabalhos mencionados não incluíram testes envolvendo os algoritmos LightGBM e XGBoost, além disso, não há relatos de avaliação dos modelos usando novos dados não rotulados.

Diante dos desafios discutidos anteriormente, esta dissertação visa contribuir com o desenvolvimento de métodos para a identificação de *posts* maliciosos na *Dark Web*. Dentre os pontos abordados neste trabalho, destaca-se a criação e compartilhamento de um conjunto de dados rotulados, coletados de fóruns na *Dark Web* com conteúdo no idioma português do Brasil. O processo de rotulagem envolve a ocorrência de IoCs, palavras-chave contextuais e análise manual. Além disso, propõe-se o desenvolvimento de um modelo para a identificação de *posts* maliciosos, incluindo testes de diferentes algoritmos, tais como LightGBM e XGBoost, seguido pela avaliação do modelo em novos dados não rotulados.

1.1 Objetivos e Desafios da Pesquisa

Este trabalho tem como objetivo principal o desenvolvimento de um modelo que seja capaz de identificar se um determinado *post* coletado da *Dark Web* tem conteúdo malicioso, ou seja, se seu conteúdo é relevante para a comunidade de segurança cibernética na detecção de ameaças, vulnerabilidades (Subseção 2.1.1), vazamentos de dados e ataques cibernéticos. Para atingir esse propósito, este trabalho analisa *posts* provenientes de fóruns da *Dark Web* utilizando técnicas de mineração de texto, *Processamento de Linguagem Natural* (PLN) e aprendizado de máquina supervisionado. Esse processo inclui a rotulagem de dados para o treinamento do modelo, o qual envolve a identificação de IoCs, palavras-chave contextuais e análise manual.

Os objetivos específicos são:

- ❏ Desenvolver uma ferramenta computacional para extração de IoCs;
- ❏ Construir conjuntos de dados rotulados a partir de dados coletados na *Dark Web*;
- ❏ Desenvolver um modelo de detecção de *posts* relevantes para segurança cibernética em fóruns da *Dark Web* utilizando aprendizado de máquina supervisionado;
- ❏ Avaliar o desempenho do modelo de detecção de *posts* relevantes em fóruns da *Dark Web*;

- ❑ Avaliar o comportamento do modelo desenvolvido na classificação de novos dados não rotulados.

1.2 Hipótese

A detecção de mensagens de conteúdo relevante que auxiliem a comunidade de segurança cibernética na obtenção e compartilhamento de CTI pode ser realizada através de um modelo computacional que utiliza aprendizado de máquina supervisionado, tendo a *Dark Web* como fonte de dados.

1.3 Organização do Trabalho

Esta dissertação está organizada da seguinte forma. O Capítulo 2 é dedicado à fundamentação teórica, fornecendo a base essencial para o desenvolvimento do trabalho. No Capítulo 3, são apresentados os trabalhos relacionados que serviram como referencial bibliográfico para esta dissertação. As etapas do desenvolvimento deste trabalho são detalhadas no Capítulo 4. Os resultados obtidos são apresentados e discutidos no Capítulo 5, e, por fim, o Capítulo 6 apresenta as conclusões, destacando as principais contribuições deste trabalho e apontando possíveis direções para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentados os conceitos fundamentais para a compreensão deste trabalho. Para uma melhor organização, o capítulo foi dividido em três seções. Na Seção 2.1, definem-se os pilares da segurança de computadores, juntamente com os conceitos de ameaças e vulnerabilidades, bem como a exploração do papel da CTI (Inteligência de Ameaças Cibernéticas). Na Seção 2.2, diferenciam-se os conceitos de *Deep Web* e *Dark Web*, além de ser discutida a estrutura básica de um fórum da *Dark Web*. Por fim, na Seção 2.3, abordam-se os conceitos de mineração de texto e PLN, incluindo a coleta de documentos, pré-processamento de texto, representação de texto, mineração de dados e aprendizado de máquina, bem como a avaliação e interpretação de resultados.

2.1 Segurança Cibernética

O termo segurança cibernética, também conhecido como cibersegurança, está relacionado ao tema de segurança da informação, embora esta última seja um conceito mais amplo que abrange questões não totalmente relacionadas à computação (RABII et al., 2020). De acordo com Taherdoost (2022), enquanto a segurança da informação busca proteger a informação de qualquer forma de ameaça, seja digital ou física, a segurança cibernética concentra-se especificamente nas informações no ciberespaço. Em outras palavras, a segurança cibernética está focada na proteção de sistemas de computadores e redes contra ameaças e ataques cibernéticos. O escopo deste trabalho está diretamente relacionado à segurança cibernética.

Em algumas fontes da literatura, percebe-se que a terminologia segurança cibernética é tida como sinônimo de segurança de computadores. No *Manual de Segurança de Computadores* do *National Institute of Standards and Technology* (NIST), a segurança de computadores é definida como: *"a proteção oferecida para um sistema de informação automatizado a fim de alcançar os objetivos de preservar a integridade, a disponibilidade e a confidencialidade dos recursos do sistema de informação"* (GUTTMAN; ROBACK, 1995). Essa definição introduz três conceitos que são considerados os pilares da segurança

de computadores:

- **Confidencialidade:** Este princípio garante que informações privadas ou confidenciais não sejam acessadas por pessoas não autorizadas.
- **Integridade:** Para o NIST, o termo integridade em segurança de computadores abrange dois conceitos relacionados:
  - **Integridade de Dados:** Garante que informações e programas sejam alterados somente de maneira especificada e autorizada.
  - **Integridade do Sistema:** Garante que um sistema execute suas funcionalidades de forma íntegra, livre de qualquer manipulação.
- **Disponibilidade:** Este princípio garante que os sistemas funcionem prontamente e não fiquem indisponíveis para usuários autorizados.

Stallings e Brown (2014) ressaltam que *Confidencialidade*, *Integridade* e *Disponibilidade* formam o que é chamado de tríade CID, representada na Figura 1. Esses três conceitos abrangem os objetivos de segurança fundamentais para dados e informações, bem como para serviços de computação.

Figura 1 – A tríade de requisitos de segurança - Adaptado de Stallings e Brown (2014).

Embora a tríade CID defina bem os objetivos de segurança, alguns especialistas da área acreditam ser necessária a inclusão de mais alguns conceitos. Os mais comumente mencionados são dois: *autenticidade*, que é a propriedade de ser genuína e verificável, visando garantir que as partes envolvidas em uma comunicação sejam quem dizem ser e que todos os dados que chegam ao sistema provenham de uma fonte confiável; e *responsabilização*, que exige que ações de uma entidade possam ser atribuídas exclusivamente a

ela, ou seja, por meio de registros de atividades mantidos nos sistemas, deve ser possível identificar o responsável pela ação (STALLINGS; BROWN, 2014).

2.1.1 Ameaças e Vulnerabilidades na Segurança Cibernética

Na terminologia de segurança de computadores, conforme citado por Stallings e Brown (2014), uma ameaça é algo que tem potencial para violar a segurança de um sistema e pode ocorrer quando há circunstância, capacidade, ação ou evento capaz de causar danos. É um perigo possível que pode explorar um ponto fraco de um sistema de segurança. Já as vulnerabilidades são justamente os pontos fracos de um sistema, uma falha, um defeito ou qualquer fraqueza que permitem sua exploração por agentes mal-intencionados a fim de violar suas políticas de segurança. A exploração de uma vulnerabilidade de um determinado sistema que resulta no surgimento de uma ameaça de segurança se faz com o uso de algumas ferramentas, como técnicas de exploração e códigos maliciosos, a isso se dá o nome de *exploits*.

É importante destacar que, embora a exploração de vulnerabilidades possa resultar no surgimento de uma ameaça à segurança, nem todas as ameaças originam-se de vulnerabilidades de software. Em cenários como o roubo de senhas por meio de técnicas de engenharia social, o ponto fraco explorado não é o sistema em si, mas sim o usuário do sistema. Além disso, existem casos em que softwares maliciosos são instalados de maneira legítima e executam ações prejudiciais ao sistema, mesmo que não tenham sido baseados em vulnerabilidades específicas.

Conforme apontado por Ghaffarian e Shahriari (2017), um dos principais desafios da segurança de computadores são as vulnerabilidades de segurança de software, devido ao seu potencial impacto significativo. Ao longo das últimas décadas, diversas abordagens foram propostas, incluindo técnicas de aprendizado de máquina e mineração de dados, com o objetivo de mitigar esses riscos.

Falhas no processo de desenvolvimento de software podem ocorrer de forma não intencional ou serem inseridas de maneira maliciosa, frequentemente resultando na atribuição de responsabilidade aos desenvolvedores. Assal e Chiasson (2019) observam que, embora em muitos casos haja cuidado para evitar falhas e a maioria dos desenvolvedores seja bem-intencionada, ainda persistem vulnerabilidades de segurança. Mesmo com esforços contínuos para aprimorar a segurança do software e adotar as melhores práticas disponíveis, essas vulnerabilidades persistem e podem impactar milhões de usuários.

Na literatura, quando se fala em vulnerabilidades, duas siglas são comumente usadas pelos autores, a saber:

- *Common Vulnerabilities and Exposures* (CVE): As vulnerabilidades divulgadas publicamente são catalogadas no CVE, um dicionário que contém informações básicas de todas as vulnerabilidades de softwares já conhecidas. O CVE foi fundado em

1999 pelo MITRE, uma organização americana de pesquisa que presta serviços ao governo. (The MITRE Corporation, 2022)

- *National Vulnerability Database* (NVD): Um repositório do governo dos Estados Unidos para gerenciamento de vulnerabilidades mantido pelo NIST. O NVD traz informações específicas de cada vulnerabilidade como *checklists* de segurança, descrição de falhas, erros de configuração e também as chamadas *Common Vulnerability Scoring System* (CVSS) que são métricas de impacto.

Frequentemente CVE e NVD são confundidas quanto as suas funções, talvez pelo motivo da NVD fazer uso do CVE e trazer informações contidas no site dessa base, mas na verdade elas são complementares (NIST, 2022).

Este trabalho tem por objetivo colaborar na identificação de *posts* maliciosos que possam implicar em incidentes de segurança. Acredita-se que assuntos envolvendo vulnerabilidades e ameaças, bem como técnicas de exploração, códigos maliciosos (*exploits*) e vazamento de dados podem ser discutidos na *Dark Web* antes de serem aplicados na prática. Os trabalhos (SAPIENZA et al., 2017), (SHU et al., 2018), (SARKAR et al., 2019) fornecem indicações sobre isso. Monitorar esses fóruns pode ser uma forma de se obter informações essenciais para organizações e até antecipar eventuais ataques.

2.1.2 *Cyber Threat Intelligence (CTI)* ou Inteligência de Ameaças Cibernéticas

Conforme descrito por Saxena e Gayathri (2022), CTI é uma estrutura de defesa baseada em evidências que reage proativamente ao monitorar e trocar informações relacionadas à segurança de ameaças cibernéticas avançadas entre diferentes setores. Por meio da coleta, análise e interpretação de informações relevantes relacionadas às ameaças cibernéticas, a CTI capacita as organizações a se manterem cientes das ameaças existentes e a tomar medidas proativas para reduzir os riscos. No mesmo contexto, de acordo com Jo, Lee e Shin (2022), a CTI é definida como uma coleção de dados que descreve as características das ameaças cibernéticas. Isso permite os profissionais de segurança a compreenderem essas ameaças e desenvolverem estratégias eficazes de defesa.

Sun et al. (2023) destacam que, de maneira geral, o fluxo de processamento de entrada de CTI consiste nos dados brutos relacionados à cibersegurança. Por outro lado, a saída desse processo é o conhecimento que orienta decisões futuras para a defesa proativa da cibersegurança, abrangendo estratégias destinadas a limitar a extensão e prevenir ataques cibernéticos. Exemplos das informações resultantes desse processo incluem relatórios detalhados descrevendo ameaças específicas, dados sobre vulnerabilidades recentemente descobertas e listas de IoCs, como endereços IP, *hashes* de arquivos, URLs, entre outros.

A implementação prática da CTI pode envolver o monitoramento de diversas fontes de informações como, sites de notícias relacionadas a segurança cibernética, redes sociais, fóruns da *Dark Web*, blogs dentre outras.

Uma maneira de obter informações cruciais sobre ameaças cibernéticas, aplicando os conceitos de CTI, é por meio dos IoCs (ZHANG et al., 2019). Os IoCs são utilizados para identificar atividades maliciosas em um sistema ou rede, funcionando como uma espécie de impressão digital. Eles consistem em sinais específicos, tais como endereços IP, nomes de domínio, URLs e *hashes* de arquivos, que auxiliam os analistas de segurança a determinar se um sistema foi comprometido. Entre os tipos de dados de CTI, os IoCs têm recebido uma atenção significativa em pesquisas relacionadas ao tema (JO; LEE; SHIN, 2022).

Niakanlahiji et al. (2019) destacam que os IoCs podem ser classificados de diferentes maneiras, sendo a mais comum aquela que se baseia na granularidade dos dados representados pelos IoCs. Nesta classificação, os IoCs são divididos em três grupos: IoCs atômicos, computados e comportamentais. Exemplos de IoCs atômicos são: endereços IP, nomes de domínio e chaves de registro. Já os IoCs computados são aqueles calculados a partir de dados observados durante um ataque, como por exemplo valores de *hash* de instâncias de *malware*. Por fim, os IoCs comportamentais são uma combinação dos dois outros tipos.

Para Asiri et al. (2023), além de estar ciente dos agentes de ameaças e tipos de ataques, as equipes de segurança precisam também conhecer os dados associados a esses ataques cibernéticos, chamados de IoCs. Esse conhecimento pode melhorar o tempo de resposta a um incidente de segurança. Por outro lado, segundo Jo, Lee e Shin (2022) os IoCs são apenas um dos tipos de dados CTI e não devem ser o único foco das estratégias de segurança cibernética, já que as ameaças estão em constante evolução e se tornando cada vez mais sofisticadas.

2.2 *Deep Web* e *Dark Web*

Os termos *Deep Web* e *Dark Web* são muitas vezes tratados como sinônimos, no entanto existe diferença entre os dois. De acordo com Akhgar et al. (2021), a primeira camada da *Web*, chamada de *Surface Web*, é a parte da Internet prontamente disponível para o público em geral, acessível através de navegadores comuns, como *Mozilla, Opera, Edge*, e pesquisável por meio de mecanismos de busca convencionais, tais como *Google, Yahoo e Bing*, uma vez que seu conteúdo está indexado. Akhgar et al. (2021) destacam ainda que essa camada superficial faz parte da *Web* desde a invenção do primeiro navegador. Por outro lado, a *Deep Web*, ao contrário da *Surface Web*, não é indexada. Nela, hospeda-se conteúdo protegido por senha, como contas bancárias *online* e serviços de e-mail privados, entre outros, que também são acessíveis ao público, mas exigem métodos diferentes de acesso, geralmente envolvendo o uso de credenciais de acesso, como nome de usuário e senha, incluindo o uso de criptografia. A *Dark Web*, por sua vez, representa a camada

mais profunda da *Deep Web*. Além de não ser indexada, seu conteúdo é intencionalmente ocultado, tornando-o inacessível através de navegadores comuns utilizados nas primeiras camadas da *Web*.

Para Bradbury (2014), o conceito de *Deep Web* e *Dark Web* refere-se a duas entidades distintas. A *Deep Web* consiste em páginas da Web que são acessíveis na Internet pública, porém não podem ser encontradas por meio de mecanismos de pesquisa convencionais, como o *Google*. Por outro lado, a *Dark Web* representa uma parte da Internet que geralmente é acessível publicamente, mas requer conhecimento específico para encontrá-la, pois está situada em uma camada alternativa da Internet onde o anonimato é valorizado. Ainda segundo Bradbury (2014), a *Dark Web* tem sido associada a atividades criminosas, como venda de armas e drogas, redes de *hackers* e lavagem de dinheiro. No entanto, também é utilizada por pessoas que buscam contornar medidas de censura impostas por regimes não democráticos.

Basheer e Alkhatib (2021) destacam que a *Deep Web* é a parte da *Web* que os motores de busca não podem acessar por diferentes razões relacionadas às funções operacionais dos sites. Estima-se que essa parte representa mais de 90% de toda a Internet. A *Dark Web* faz parte da *Deep Web*, nela técnicas de criptografia especial são usadas para ocultar as identidades e endereços IP dos usuários. Saleem, Islam e Kabir (2022) reforçam que a rede oculta e o anonimato da *Dark Web* abrem caminho para atividades ilegais e colaboram com os cibercriminosos na execução de ciberataques.

De acordo com Akhgar et al. (2021), a analogia do *iceberg* é frequentemente utilizada para representar as diferentes camadas da Internet: a parte visível do *iceberg* representa a *Surface Web*, que corresponde a apenas cerca de 4% de todo o conteúdo disponível na rede mundial de computadores. A grande maioria, ou seja, a *Deep Web*, encontra-se abaixo da superfície e não é acessível por meio de ferramentas de busca convencionais. Já a *Dark Web*, que representa uma pequena parte do fundo do *iceberg*, é composta por sites que requerem o uso de navegadores especializados e é conhecida por ser um ambiente frequentado por criminosos para realizar atividades ilegais.

2.2.1 Fórum da *Dark Web*

Um fórum da *Dark Web* é uma plataforma online onde os usuários podem discutir sobre vários tópicos. Geralmente um fórum é dividido em categorias, sendo cada uma dedicada a um tema específico, como *hacking*, drogas, dinheiro e mercados. No entanto, de acordo com Al-Ramahi, Alsmadi e Davenport (2020), os usuários nem sempre seguem essas categorias predefinidas e podem postar conteúdo relacionado a um tópico específico em uma categoria diferente. Akhgar et al. (2021), enfatizam que para acessar um fórum na *Dark Web*, os usuários precisam usar navegadores especializados, como o *The Onion Router* (Tor), que roteia o tráfego da Internet através de vários servidores em todo o mundo, tornando difícil rastrear a identidade do usuário.

Para inserir conteúdo em um fórum na *Dark Web*, o usuário segue alguns passos básicos: escolhe uma categoria, define um título e, em um campo separado, pode detalhar o conteúdo de sua postagem. Tanto no título quanto no conteúdo, é comum compartilhar informações ou fazer perguntas. A postagem inicial pode receber respostas e comentários de outros usuários, o que é conhecido como interações e depende da relevância do conteúdo. Embora qualquer resposta, comentário ou mensagem inicial possa ser considerado um *post*, neste trabalho um *post* é definido como um conjunto de mensagens relacionadas a um tópico específico. O número de interações e se elas ocorreram não afetam a definição de um *post*.

Neste trabalho, considera-se um *post* como sendo *malicioso* ou *relevante* se o seu conteúdo estiver relacionado a alguma ameaça, vulnerabilidade, *exploit*, vazamento de dados ou algo que indicar algum risco à segurança cibernética. Caso contrário, ele é classificado como *não relevante*. A Tabela 1 mostra exemplos de quatro *posts*, sendo dois *relevantes* e dois *não relevantes*.

Tabela 1 – Exemplos de *posts relevantes* e *não relevantes* (Fonte: O autor (2023)).

ID	**Título**	**Conteúdo**	**Relevante**
144	"Free Hacking T**** ..."	"Estava navegando na Deep Web, quando me deparei com um site que diz distribuir ferramentas de hacking ... Link do site citado : http://********.onion/"	Sim
815	"dados pessoais, alguém compra?"	"tenho dados pessoais em grande quantidade, com nome completo, cpf, data de nascimento, cell endereço entre outros ..."	Sim
2550	"Decretos de Governadores"	"Qual a opinião de vocês sobre esses decretos de governadores em todo Brasil? ..."	Não
11108	"fake news"	"Mano com essa m***a toda rolando, você acredita que pode rolar manipulação da verdade? ... "	Não

Conforme mostrado na Tabela 1, os *posts* com IDs 144 e 815 foram rotulados como relevantes, sendo que o primeiro aborda a distribuição de ferramentas de *hacking* e o segundo trata de um vazamento de dados. Por outro lado, os *posts* com IDs 2550 e 11108 não foram considerados relevantes, pois abordam temas não relacionados a ameaças de segurança cibernética, sendo, respectivamente, sobre política e notícias falsas.

2.3 Mineração de Texto e Processamento de Linguagem Natural

Para Dang e Ahmad (2014), mineração de texto é definida como uma forma de descobrir informações valiosas de textos não estruturados. Um texto não estruturado, como o próprio nome sugere, pode ser visto como uma grande quantidade de informações armazenadas em diferentes lugares, de maneira não estruturada. Um texto assim ainda não se encontra pronto para ser processado. Para se obter informações preciosas nesse caso, é necessário o uso de alguma técnica que seja útil. É neste contexto que se aplica a mineração de texto.

Justamente por lidar com informações não estruturadas, a mineração de texto passa por etapas bastante onerosas, desde o pré-processamento até a extração de características. Nesse contexto, utiliza-se o (PLN), uma área da computação cujo objetivo é extrair uma representação de significado mais completa a partir de texto livre. Isso geralmente envolve o uso de conceitos linguísticos, além do tratamento de anáfora e ambiguidades relacionadas à estrutura gramatical e ao conjunto lexical (KAO; POTEET, 2007).

Chowdhary e Chowdhary (2020) destacam que PLN é uma forma de utilizar os computadores para compreender e manipular texto ou fala em linguagem natural, com o objetivo de realizar tarefas úteis. Através da compreensão de como os seres humanos entendem e utilizam a linguagem, é possível desenvolver sistemas computacionais que compreendam e manipulem linguagens naturais para realizar tarefas desejadas.

No contexto da mineração de texto, as fontes de dados são chamadas de coleções de documentos. Um documento pode ser representado de diversas formas, como um arquivo em diferentes formatos *(doc, pdf, txt)*, um *e-mail*, uma página *Web* ou um *post* da *Dark Web*, por exemplo.

A mineração de textos envolve várias etapas essenciais. Na literatura, destacam-se as seguintes etapas fundamentais nos processos de mineração de texto: coleta de documentos, pré-processamento de texto, representação de texto, aplicação de técnicas de mineração de dados e aprendizado de máquina, bem como a avaliação e interpretação dos resultados. Em um cenário onde os dados são coletados de fontes como redes sociais e fóruns da *Dark Web*, como é o caso deste trabalho, essas etapas podem ser representadas e visualizadas conforme ilustrado na Figura 2.

2.3.1 Coleta de Documentos

A primeira etapa, *coleta de documentos*, é frequentemente realizada por meio de uma *Application Programming Interface* (API) especializada, projetada para extrair informações de fontes específicas. No contexto da análise de fóruns na *Dark Web*, onde a disponibilidade pública de dados é limitada, a coleta de informações é ainda mais desafiadora.

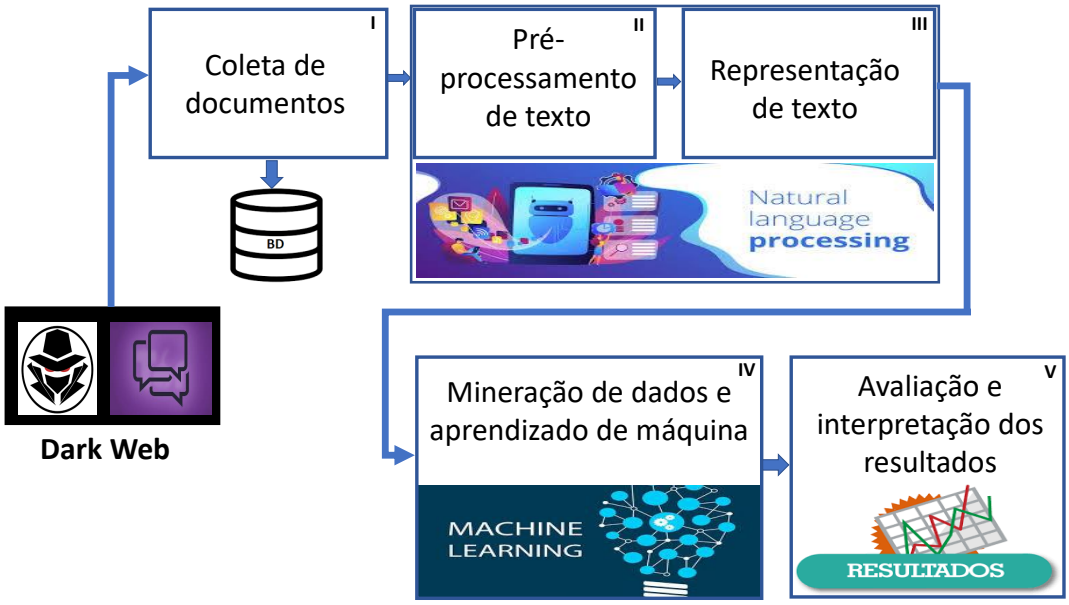


Figura 2 – Etapas básicas do processo de mineração de texto (Fonte: O autor (2023)).

É comum recorrer a ferramentas de extração de dados, como os *crawlers*, que são programas desenvolvidos para realizar uma varredura sistemática e minuciosa em busca de informações relevantes.

Vale ressaltar que, ao lidar com fóruns na *Dark Web*, é essencial considerar questões relacionadas à privacidade e à segurança, uma vez que esses ambientes são conhecidos por seu anonimato e natureza sigilosa. A coleta de documentos nesses cenários exige uma abordagem cuidadosa e planejada, incluindo medidas para proteger a identidade dos pesquisadores e a integridade dos dados coletados.

2.3.2 Pré-processamento de Texto

A segunda etapa é o *pré-processamento de texto*. Nessa fase, são aplicadas técnicas de PLN para realizar a limpeza e preparação do texto. De acordo com Anandarajan, Hill e Nolan (2019), o pré-processamento de texto recebe um documento em formato bruto e retorna *tokens (palavras únicas ou grupos de palavras contabilizados por sua frequência, que servem como características na análise)* após terem sido devidamente limpos. Durante esse procedimento, algumas tarefas são executadas, tais como:

- **Padronização e limpeza**: Envolve a conversão de caracteres maiúsculos em minúsculos, remoção de números, sinais de pontuação e caracteres especiais.

- **Remoção de *stop words***: Consiste na eliminação de palavras comuns que não fornecem informações relevantes sobre o conteúdo ou o assunto do texto, tais como artigos, preposições e verbos auxiliares. Também é possível criar dicionários personalizados para remover palavras comuns em um determinado idioma.

- **Derivação ou Lematização**: Utiliza conceitos linguísticos, como classes gramaticais *(parte do discurso)*, para identificar a classe gramatical de cada palavra no texto. Aqui, as palavras são reduzidas à sua raiz. A derivação envolve a remoção de sufixos, enquanto a lematização incorpora informações sobre a parte do discurso das palavras. Ambos os métodos agrupam palavras que compartilham a mesma raiz em um único token, reduzindo assim o número de tokens exclusivos no conjunto de análise (ANANDARAJAN; HILL; NOLAN, 2019).

- **Unitização e tokenização**: Inclui a escolha da unidade de texto a ser analisada e a separação do texto com base na unidade de análise. Essa unidade pode ser uma única palavra *(unigrama)*, duas palavras lado a lado *(bigrama)* ou uma sequência de palavras consecutivas com comprimento n *(n-grams, onde n é um número inteiro)*.

A Figura 3 ilustra as etapas do pré-processamento de texto. É importante observar que a ordem dessas tarefas pode variar, embora faça mais sentido realizar a *Padronização e Limpeza* e a *Remoção de Stop Words* no início do processo de pré-processamento de texto, uma vez que isso tende a diminuir a carga de trabalho nas próximas etapas e contribui para a eficiência do processo. Por outro lado, a *Derivação ou Lematização* são opções que podem ser aplicadas, mas não são estritamente obrigatórias; a decisão de utilizá-las ou não dependerá das características específicas de cada caso.

Figura 3 – Etapas do pré-processamento de texto - Adaptado de Anandarajan, Hill e Nolan (2019).

2.3.3 Representação de Texto

A terceira etapa, *representação de texto*, também faz uso de técnicas de PLN para extrair informações significativas do texto. Um conceito amplamente utilizado nesta etapa

é o *saco de palavras* ou *bag of words* em inglês, que representa um documento como um conjunto de palavras contidas nele, juntamente com a contagem de quantas vezes cada palavra aparece no documento (ANANDARAJAN; HILL; NOLAN, 2019). Essa abordagem, também chamada de modelo de espaço vetorial, é amplamente empregada na extração de informações relevantes de documentos e na classificação de documentos para treinar classificadores baseados na frequência das palavras (MCTEAR; CALLEJAS; GRIOL, 2016).

A frequência de uma palavra, ou TF como é chamado na literatura, é um conceito frequentemente aplicado na área de mineração de texto e análise de documentos. Ela é usada para medir a importância relativa de um termo em um documento, contando quantas vezes esse termo ocorre em relação ao tamanho total do documento. Como destacado por Qaiser e Ali (2018), a TF é especialmente útil porque os documentos variam em tamanho, e um termo pode ocorrer mais frequentemente em documentos maiores do que em documentos menores. Para lidar com essa variação, a TF é calculada dividindo o número de ocorrências do termo (*t1*) pelo total de termos no documento (T), conforme ilustrado na Equação 1.

$$TF(t1) = \frac{quantidade(t1)}{quantidade(T)} \tag{1}$$

Onde:

- *TF(t1)*: é a Frequência de Termo para o termo *t1*.
- *quantidade(t1)*: representa o número de ocorrências do termo *t1* em um documento específico.
- *quantidade(T)*: indica o total de termos presentes no mesmo documento onde o termo *t1* está sendo avaliado.

Ainda segundo Qaiser e Ali (2018), ao calcular a frequência de um termo, percebe-se que o algoritmo trata todos os termos como sendo igualmente importantes, o que pode levar o sistema a dar um peso maior para uma palavra comum que não seja muito relevante no contexto. Para tratar essa questão, existe a chamada IDF, que atribui menor peso às palavras mais frequentes e maior peso às menos frequentes. A IDF é calculada conforme a Equação 2.

$$IDF(t1) = \log_e \left(\frac{quantidade(D)}{quantidade(Dt1)} \right) \tag{2}$$

Onde:

- *IDF(t1)*: representa a Frequência Inversa de Documentos para o termo *t1*.
- *quantidade(D)*: é o número total de documentos em seu conjunto de dados ou corpus.

❏ *quantidade(Dt1)*: é o número de documentos em que o termo *t1* ocorre.

Ao combinar a frequência de termo *(TF)* com o elemento que dá peso as palavras raras, frequência inversa de documentos *(IDF)*, obtêm-se o *(TF-IDF)*, que nada mais é do que a multiplicação de *(TF)* x *(IDF)*, conforme mostrado na Equação 3.

$$TFIDF(t1) = TF(t1) \times IDF(t1) \tag{3}$$

Existem outras formas de representação de texto em PLN, uma delas é o *word2vec*, que foi desenvolvido por *Tomas Mikolov* e sua equipe na *Google Research* em 2013 (MIKOLOV et al., 2013). Desde então, o *word2vec* tem sido amplamente utilizado em sistemas de tradução automática, recomendação, agrupamento de documentos e muito mais. Ele adota duas arquiteturas de modelo principais: o modelo contínuo de saco de palavras ou *Continuous Bag of Words* (CBOW) e o modelo contínuo de *skip-gram*, para aprender representações vetoriais de palavras. No modelo CBOW, a palavra atual é prevista com base em seu contexto, enquanto o modelo *skip-gram* prevê as palavras circundantes com base na palavra atual.

2.3.4 Mineração de Dados e Aprendizado de Máquina

Na quarta etapa, emprega-se o aprendizado de máquina, que vai além da abordagem tradicional de programação de computadores. De acordo com Alpaydin (2016), na programação convencional de computadores, utiliza-se um algoritmo, que consiste em uma sequência de instruções, para processar uma entrada e gerar uma saída. Isso é evidente em problemas como a ordenação de números, em que a entrada é um conjunto de números e a saída é uma lista ordenada. No entanto, em alguns cenários, como a previsão do comportamento de um cliente ou a distinção entre e-mails de spam e e-mails legítimos, não existe um algoritmo pré-definido pronto para uso.

No contexto da classificação de e-mails, a entrada é um documento de e-mail, geralmente uma mensagem de texto simples, e a saída desejada é binária: SIM (indicando que é um e-mail de *spam*) ou NÃO (indicando que é um e-mail legítimo). A complexidade aqui reside em converter essa entrada em uma saída precisa, uma vez que a definição de spam pode variar ao longo do tempo e de pessoa para pessoa.

No entanto, como apontado por Alpaydin (2016), a falta de conhecimento específico pode ser compensada com a quantidade de dados disponíveis. É possível reunir facilmente milhares de mensagens, algumas das quais já estão identificadas como *spam* e outras como legítimas. Com essa amostragem, o computador (ou a máquina) pode aprender a distinguir o que constitui um e-mail de *spam* e, assim, extrair automaticamente um algoritmo para essa tarefa.

Em problemas de classificação, o computador recebe exemplos de entradas e saídas fornecidos por um *supervisor*, geralmente, em forma de uma base de dados rotulada. Sua

tarefa é aprender uma regra geral que mapeia as entradas para as saídas, e esse tipo de aprendizado é denominado *aprendizado supervisionado*. Alguns dos algoritmos de aprendizado de máquina supervisionado frequentemente utilizados em problemas que envolvem classificação incluem: *Support Vector Machine, Random Forest, Logistic Regression, LightGBM* e *XGBoost*.

No aprendizado de máquina, quando não se dispõe de uma base de dados rotulada, é empregado o chamado *aprendizado não supervisionado*. Nesse cenário, o algoritmo não recebe rótulos predefinidos, sendo responsável por identificar padrões nos dados de forma autônoma (HAHNE et al., 2008). Esse tipo de aprendizado é comumente aplicado em tarefas de agrupamento e associação. Um exemplo relevante no campo de aprendizado de máquina e processamento de linguagem natural é a LDA, uma técnica estatística amplamente empregada para análise de texto e extração de tópicos e que foi utilizada ao longo deste trabalho.

2.3.4.1 *Support Vector Machine (SVM)* ou Máquina de Vetores de Suporte

O SVM é um algoritmo amplamente utilizado em problemas de classificação. Ele busca encontrar o hiperplano ideal de separação entre classes, maximizando a margem entre os pontos mais próximos das classes (MEYER; WIEN, 2015). Embora o SVM também possa ser aplicado a problemas de regressão, ele se destaca em lidar com tarefas de classificação, especialmente quando se trata de conjuntos de dados complexos, alta dimensionalidade e separação não linear. Vale destacar que o treinamento do SVM pode ser computacionalmente mais intensivo em grandes conjuntos de dados.

A Figura 4 exemplifica a separação de classes do SVM, onde os pontos situados nos limites são os chamados vetores de suporte e o meio da margem é onde está o hiperplano de separação ideal.

2.3.4.2 *Random Forest* ou Floresta Aleatória

Random Forest (Floresta Aleatória) é um *ensemble*, ou seja, uma técnica que combina vários modelos para melhorar o desempenho geral do sistema, e é baseado em árvores. Como o nome sugere, cada árvore nessa floresta depende de uma coleção de variáveis aleatórias (CUTLER; CUTLER; STEVENS, 2012). As Florestas Aleatórias têm características que as tornam atraentes tanto para problemas de classificação quanto para regressão. Diversas vantagens das Florestas Aleatórias foram destacadas por Cutler, Cutler e Stevens (2012), tais como:

- ❏ Lidam naturalmente com problemas de regressão e classificação, incluindo problemas multiclasse;
- ❏ São relativamente rápidas no treinamento e previsão;

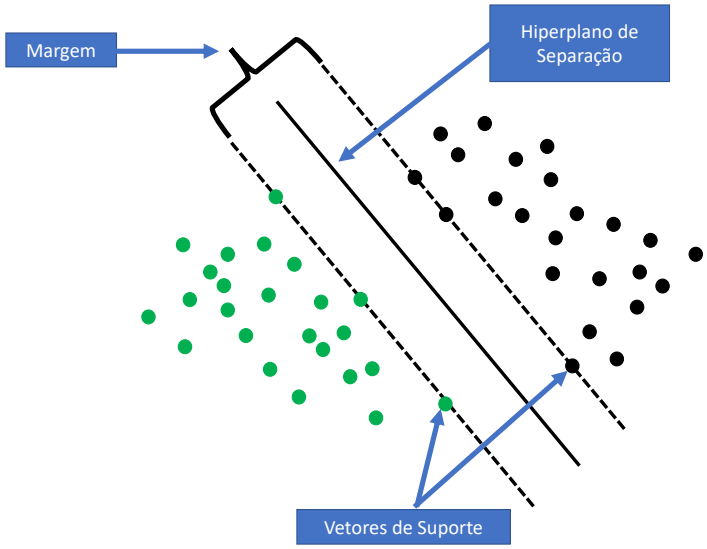


Figura 4 – Classificação linear usando o algoritmo SVM - Adaptado de Meyer e Wien (2015).

- ❏ Dependem apenas de um ou dois parâmetros de ajuste;
- ❏ Podem ser aplicadas diretamente a problemas de alta dimensionalidade;
- ❏ Fornecem medidas de importância das variáveis;
- ❏ Permitem ponderação diferencial de classes;
- ❏ São úteis na detecção de valores atípicos.

2.3.4.3 *Logistic Regression* ou Regressão Logística

A Regressão Logística é um algoritmo amplamente empregado em tarefas de classificação, sendo especialmente útil em problemas de classificação binária, mas também adaptável para cenários multiclasse. Essencialmente, a Regressão Logística constrói um modelo matemático que avalia a probabilidade de um exemplo ser atribuído à classe positiva (1) ou à classe negativa (0). Esse cálculo é realizado por meio da aplicação de uma função, chamada função logística (HASTIE et al., 2009).

De acordo com Fagerland e Hosmer (2012), o modelo de regressão logística (binária) estabelece uma relação entre uma variável de resultado que possui dois possíveis valores (binária) e uma ou mais variáveis preditoras. Em outras palavras, ele busca entender como as variáveis preditoras influenciam a probabilidade de o resultado pertencer a uma das duas categorias possíveis.

2.3.4.4 *Light Gradient Boosting Machine (LightGBM)*

O LightGBM é um algoritmo de aprendizado de máquina supervisionado que se destaca pela sua eficiência e precisão. O termo *Gradient Boosting Machine* (GBM) (Máquina de Impulsionamento por Gradiente) refere-se a um método de aprendizado de máquina baseado em árvores de decisão, com o objetivo de aprimorar a precisão dos modelos preditivos. Ele opera pela construção sequencial de árvores de decisão, onde cada árvore é projetada para corrigir os erros das árvores anteriores. No entanto, o LightGBM adota uma abordagem única conhecida como *Gradient-based One-Side Sampling* (GOSS), que o diferencia de outros algoritmos GBM (KE et al., 2017).

O GOSS concentra-se nas amostras com erros de grande magnitude, reduzindo significativamente a quantidade de dados de treinamento necessária. Isso se traduz em maior eficiência, redução do tempo de treinamento e menor necessidade de ajustes de hiperparâmetros. Além disso, o *LightGBM* utiliza uma técnica chamada *Histogram-based Learning* (Aprendizado Baseado em Histograma), que agrupa os valores dos recursos em histogramas para acelerar a construção da árvore de decisão. Essa combinação de estratégias faz do *LightGBM* uma referência em termos de velocidade de treinamento e eficiência.

2.3.4.5 *eXtreme Gradient Boosting (XGBoost)*

O XGBoost, é um algoritmo de aprendizado de máquina supervisionado amplamente reconhecido por sua eficácia em problemas de classificação e regressão. Sua base é o método de impulsionamento por gradiente, que combina várias árvores de decisão fracas em um modelo preditivo robusto. O XGBoost opera em um processo iterativo, onde cada nova árvore é construída com o objetivo de corrigir os erros das árvores anteriores. Durante cada iteração, o algoritmo atribui pesos às observações mal classificadas, focando mais nelas para melhorar a precisão do modelo. Isso permite o algoritmo lidar bem com base de dados desbalanceadas, ou seja, quando há um número significativamente maior em uma classe em comparação com outra (CHEN et al., 2015).

Outra característica de destaque do XGBoost é sua capacidade de regularização incorporada, que ajuda a prevenir o superajuste *(overfitting)* e aumentar a generalização do modelo. Além disso, o XGBoost utiliza técnicas de pré-processamento eficientes, como particionamento de dados, para acelerar o treinamento. Sua flexibilidade permite a otimização de hiperparâmetros, incluindo a profundidade máxima das árvores, a taxa de aprendizado e o número de árvores, para alcançar o melhor desempenho (CHEN; GUESTRIN, 2016).

2.3.4.6 *Latent Dirichlet Allocation (LDA)*

O uso da técnica de modelagem de tópicos é amplamente reconhecido como uma das abordagens mais poderosas em mineração de texto para descobrir informações laten-

tes, identificar padrões e estabelecer conexões entre dados e documentos. Dentro desse contexto, LDA se destaca como uma das técnicas mais populares e eficazes. Conforme ressaltado por Jelodar et al. (2019), a LDA tem sido amplamente adotada na área de mineração de dados e texto devido à sua capacidade de revelar tópicos subjacentes em coleções de documentos, fornecendo uma compreensão mais profunda e estruturada do conteúdo textual.

A LDA, conforme discutida por Jelodar et al. (2019), é um modelo probabilístico generativo utilizado para analisar uma coleção de textos, conhecida como corpus de documentos. Sua abordagem fundamental envolve a representação de documentos como combinações aleatórias de tópicos latentes, onde cada tópico é caracterizado por uma distribuição de palavras. Em outras palavras, a LDA modela tópicos com base em probabilidades de palavras, sendo que as palavras com maiores probabilidades em cada tópico fornecem *insights* sobre o seu conteúdo. Como um método de aprendizado não supervisionado, a LDA é amplamente adotada na modelagem de tópicos e análise de tópicos em coleções de documentos, destacando-se como uma técnica essencial nesse contexto.

Ao tomar como exemplo o conteúdo exibido na Tabela 1, que apresenta quatro *posts* coletados da *Dark Web*, a aplicação da técnica LDA para dividir todo o conteúdo em três tópicos resulta no que é apresentado na Tabela 2.

Tabela 2 – Exemplo da geração de tópicos usando LDA (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	estado, povo, decretos, pessoas, dia, lockdown, correto, virus, necessario, brasil
2	news, fake, censura, beneficio, proprio, acham, memes, todos, deveriam, pior
3	site, dados, confiavel, full, pessoais, maquina, onion, hacking, virtual, compra

Documento(s) do Tópico 1:	2550
Documento(s) do Tópico 2:	11108
Documento(s) do Tópico 3:	144, 815
Base de Dados:	**Tabela 1**

Conforme apresentado na Tabela 2, o tópico 1 inclui palavras relacionadas à política, o que é coerente com o *post* de ID 2549, exibido na Tabela 1. Da mesma forma, o tópico 2 contém palavras associadas a notícias falsas, refletindo o conteúdo do *post* de ID 11107 na mesma tabela. Já o tópico 3 é composto por dois documentos, representando os *posts* de IDs 143 e 814, e apresenta palavras relacionadas a ameaças de segurança cibernética e vazamento de dados, alinhando-se ao conteúdo desses *posts*.

2.3.5 Avaliação e Interpretação dos Resultados

Durante a avaliação e interpretação dos resultados, algumas métricas clássicas são calculadas para avaliar o desempenho do modelo. Retomando o exemplo de classificação de e-mails mencionado anteriormente na Subseção 2.3.4, quando um classificador identifica um e-mail verdadeiramente como sendo um *spam*, isso contribui para a taxa de verdadeiros positivos *(TP)* - do inglês *True Positive*. Por outro lado, quando um e-mail verdadeiramente legítimo é identificado como *legítimo*, isso contribui para a taxa de verdadeiros negativos *(TN)* - do inglês *True Negative*.

Seguindo a mesma lógica, quando um e-mail verdadeiramente *spam* é identificado falsamente como sendo *legítimo*, isso contribui para a taxa de falsos negativos *(FN)* - do inglês *False Negative*. Por fim, quando um e-mail verdadeiramente legítimo é identificado falsamente como sendo um *spam*, isso contribui para a taxa de falsos positivos *(FP)* - do inglês *False Positive*. Com essas taxas, é possível construir a chamada Matriz de Confusão, citado por Susmaga (2004) e ilustrado na Figura 5.

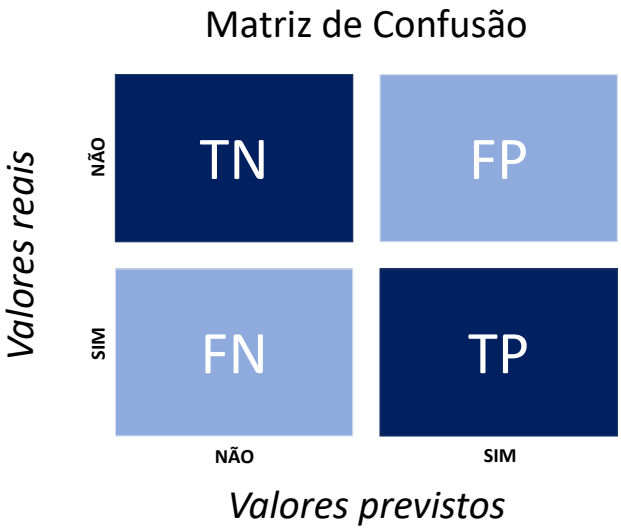


Figura 5 – Matriz de confusão para problemas binários (Fonte: O autor (2023)).

Através da matriz de confusão, importantes métricas de avaliação do modelo podem ser extraídas, tais como: acurácia, precisão, revocação e a medida F, onde:

$$\textbf{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

A acurácia, representada na Equação 4, é a razão entre as amostras preditas corretamente e o total de amostras.

$$\textbf{Precisão} = \frac{TP}{TP + FP} \tag{5}$$

Já a precisão, mostrada na Equação 5, é a razão entre o número de amostras corretamente preditas como positivas e o número total de amostras preditas como positivas.

$$\textbf{Revocação} = \frac{TP}{TP + FN} \tag{6}$$

A revocação, representada na Equação 6 é a razão entre o número de amostras corretamente preditas como positivas e o número total de amostras verdadeiramente positivas na base de dados. A revocação também é conhecida como sensibilidade ou taxa de verdadeiros positivos.

$$\textbf{\textit{Medida F}} = 2 \cdot \frac{Precisão \cdot Revocação}{Precisão + Revocação} \tag{7}$$

Por fim, a medida F, representada na Equação 7, é a média ponderada da precisão e da revocação, é uma métrica que combina precisão e revocação em uma única medida.

Com essas métricas, é possível avaliar o desempenho do modelo de classificação, o que proporciona uma visão abrangente e precisa de sua capacidade de prever e distinguir entre as diferentes classes ou categorias do problema em questão.

Uma abordagem adicional para avaliar e interpretar os resultados reside na realização de agrupamentos por tópicos, principalmente utilizando a técnica de LDA. Ao examinar as palavras-chave vinculadas a cada tópico, é possível obter *insights* valiosos sobre o conteúdo abordado. A visualização das palavras mais significativas em cada tópico proporciona uma compreensão nítida das características e temas subjacentes a cada um, facilitando assim a interpretação do significado e da relevância de cada tópico identificado.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão apresentados os principais trabalhos relacionados a esta dissertação com o objetivo de fornecer uma visão geral do estado da arte em relação à extração de informações para CTI, utilizando fontes de dados não estruturadas, como redes sociais e fóruns da *Dark Web*. Essas propostas partem do pressuposto de que nessas fontes é possível identificar sinais de ameaças existentes e, até mesmo, prever novos ataques cibernéticos antes que sejam executados.

A Seção 3.1 apresenta um resumo de trabalhos relacionados que exploraram redes sociais e a *Dark Web* para identificar ameaças cibernéticas, empregando diversas abordagens. A Seção 3.2 destaca os principais trabalhos que se concentraram na extração de informações para CTI através de IoCs. A Seção 3.3 aborda os trabalhos que utilizaram aprendizado de máquina para extrair CTI da *Dark Web*. Já a Seção 3.4 traz as considerações finais sobre os trabalhos relacionados. O presente trabalho centrou-se na identificação de *posts* maliciosos, utilizando técnicas de mineração de texto, PLN e aprendizado de máquina supervisionado em dados coletados de fóruns da *Dark Web*. Este processo compartilha algumas etapas comuns com os trabalhos relacionados, já que todos têm como objetivo central a obtenção de CTI em fontes de dados não estruturados, como redes sociais, fóruns da *Surface Web* e *Dark Web*.

3.1 Identificação de Ameaças Cibernéticas em Redes Sociais e Fóruns da *Dark Web*

Sapienza et al. (2017) propuseram uma estrutura que aproveita sensores de mídia social, em especial *Twitter* e fóruns da *Dark Web*, para gerar alertas antecipados de ameaças cibernéticas. Os autores usaram como estratégia o monitoramento de contas no *Twitter* de especialistas, pesquisadores e *hackers* éticos escolhidos manualmente, a fim de encontrar postagens de exploração de vulnerabilidades. Usando técnicas de mineração de texto, selecionam termos importantes e removem os irrelevantes, tendo como base

dicionários pré-definidos. Em seguida, verifica se os termos descobertos também já foram mencionados em fóruns de *hackers* da *Dark Web* previamente selecionados. Durante o período observado, 84% dos avisos gerados foram considerados relevantes por especialistas. Os autores relataram que importantes eventos de segurança como o ataque *Mirai* que ocorreu em outubro de 2016 através de uma exploração de vulnerabilidade de dispositivos *Internet of Things* (IoT) e ataques de violação de dados como o *AdultFriendFinder* e *BrazzersForum* foram identificados nos testes realizados.

Por outro lado, Alves et al. (2021) apresentaram um monitor de ameaças que utiliza o *Twitter* para gerar um resumo continuamente atualizado do cenário de ameaças relacionado a uma infraestrutura monitorada. Inicialmente, é criada manualmente uma lista de palavras-chave que descrevem a infraestrutura de *Tecnologia da Informação* (TI) a ser monitorada. Além disso, é previamente definida uma lista de contas do *Twitter* relacionadas à segurança cibernética. O sistema utiliza uma API para capturar os *posts* das contas monitoradas e aplica um filtro com base na lista de palavras-chave pré-definida. Apenas *tweets* que incluem pelo menos uma das palavras da lista passam pelo filtro. Em seguida, os dados são preparados para processamento por meio de técnicas de mineração de texto. Nos passos seguintes, algoritmos de aprendizado de máquina supervisionado são aplicados para classificar os *tweets* de acordo com sua relevância de segurança. Antes da geração de alertas de segurança, os *tweets* passam por uma etapa de agrupamento, na qual *tweets* semelhantes e *retweets* são agrupados. Os autores destacam que essa etapa é fundamental para evitar alertas redundantes.

Os testes realizados em uma infraestrutura de TI de exemplo demonstraram que o sistema teve um bom desempenho na identificação de ameaças de segurança, com uma taxa de verdadeiro positivo acima de 90%, enquanto a taxa de falso positivo ficou abaixo de 10%. Os alertas gerados pelo sistema também foram considerados relevantes, levando em conta a pontuação CVSS, a disponibilidade de *patches* ou *exploits*, e as datas de divulgação de ameaças do NVD.

A abordagem utilizada pelos autores neste trabalho mostrou-se eficiente, e os resultados foram animadores. Um desafio identificado no processo foi a rotulagem da base de dados para treinamento, a qual foi realizada manualmente por um analista responsável por marcar todos os *tweets* julgados como irrelevantes.

Subroto e Apriyana (2019) apresentaram um modelo que utiliza análise de *big data* de mídias sociais, especialmente o *Twitter*, e aprendizado de máquina estatístico para prever riscos cibernéticos. Os dados são coletados do *Twitter* por meio de uma API, utilizando algumas palavras-chave, incluindo a palavra *CVE*. Apenas os *tweets* que contêm IDs CVE publicados no site CVE são considerados como ameaças. Após passarem por um processo de limpeza, uma análise de *big data* é realizada para detectar, validar e eliminar informações irrelevantes.

Após a coleta e limpeza, ocorre a extração de recursos e a análise de dados, com o

objetivo de extrair e analisar palavras individuais, bem como determinar o número de ocorrências no conjunto de documentos. Durante a fase de treinamento e teste, utilizando aprendizado de máquina estatístico, o rótulo *(1)* é atribuído aos *tweets* cujo ID CVE está registrado no site *(cvedetails.com)*, indicando ser um risco cibernético. O rótulo *(0)* é utilizado para IDs CVE não registrados no site, considerados como não representando um risco cibernético.

Os modelos de previsão alcançaram altas taxas de acerto, atingindo uma precisão de 96,73% no melhor dos casos. O modelo proposto pelos autores mostrou-se eficiente na identificação de riscos cibernéticos usando o *Twitter* como fonte de dados. No entanto, uma questão em aberto é se há algum prejuízo ao considerar apenas *tweets* que mencionam um *CVE*.

Por sua vez, Sceller et al. (2017) propõem uma estrutura automática e autodidata capaz de detectar, geo-localizar e categorizar eventos de segurança cibernética quase em tempo real no fluxo do *Twitter*. A coleta de dados ocorre por meio da API do *Twitter* com base em uma lista de palavras-chave, sendo posteriormente processados e armazenados em um banco de dados. Em seguida, o algoritmo realiza uma busca para identificar um conjunto de *tweets* que discutem o mesmo tópico durante o mesmo período, sendo considerado um evento. Os eventos identificados são geolocalizados, classificados e exibidos em uma interface do sistema.

Para expandir a lista de palavras-chave inicial, o sistema incorpora um componente que rastreia automaticamente novas palavras-chave relevantes nos *tweets* processados. Em uma tentativa de evitar falsos positivos, o sistema utiliza uma lista negra para descartar *tweets* irrelevantes. No processo de classificação, é adotada uma taxonomia de cinco níveis, cada um composto por um conjunto de palavras-chave, de modo que cada *tweet* é mapeado na categoria da palavra-chave utilizada para recuperá-lo. O processo de geolocalização é realizado por meio da API de geocodificação do *Google Maps*. Quanto à detecção de eventos, o algoritmo agrupa os documentos processados com base em sua similaridade.

Quanto aos resultados, os autores relatam que o sistema apresentou um bom desempenho, sendo capaz de detectar eventos importantes de segurança cibernética. A estrutura proposta neste trabalho traz contribuições significativas para a detecção de eventos de segurança cibernética. No entanto, alguns pontos de melhoria são destacados, sendo o principal deles o alto índice de falsos positivos. De acordo com os autores, em uma amostra de 100 eventos detectados pelo sistema, apenas 23 foram considerados altamente relevantes.

Rodriguez e Okamura (2020) apresentam um sistema em tempo real que utiliza análise de dados recuperados do Twitter para agregar grandes quantidades de *tweets* e gerar informações de conscientização cibernética. O sistema examina o contexto contido nos *tweets* por meio da análise de sentimentos para obter informações sobre o nível de risco da ameaça.

Na etapa de coleta de dados, os autores selecionaram contas do *Twitter* relacionadas à segurança cibernética e aplicaram um filtro de palavras-chave. Para evitar a recuperação de dados fora do contexto, o sistema analisa o corpo do texto em relação à lista de palavras-chave de segurança. Novas palavras-chave são descobertas aplicando as técnicas de TF e TF-IDF.

O sistema conta com um classificador de sentimentos que utiliza aprendizado de máquina supervisionado, onde cada *tweet* é classificado como tendo um sentimento negativo ou positivo. Um analista pode então observar a interface gráfica detalhada do sistema e compreender a situação no momento. Um grande número de *tweets* negativos que contenham uma palavra-chave pode indicar a ocorrência de um ataque.

Em um estudo de caso, os autores relatam que o sistema foi capaz de fornecer informações relevantes que permitiriam a um analista compreender o nível de risco atual em sua organização e, assim, tomar decisões informadas. Eles também destacam que o sistema foi capaz de detectar algumas ameaças antes de serem publicadas em determinados sites de segurança ou tornarem-se conhecidas na comunidade em geral.

De maneira geral, a proposta apresentada nesse trabalho mostrou-se útil na identificação de ameaças de segurança cibernética. A classificação de *tweets* usando análise de sentimentos mostrou bons resultados, embora os autores não tenham apresentado números relacionados a falsos positivos.

Khandpur et al. (2017) apresentam uma abordagem um pouco diferente das demais. Nesse trabalho, a detecção de ataques cibernéticos é realizada de maneira fracamente supervisionada, dispensando a necessidade de uma fase de treinamento ou amostras rotuladas. Os autores consideram o uso da mídia social como um sensor de *crowdsourcing* para obter informações sobre ataques cibernéticos em andamento.

A estrutura para detecção de eventos de segurança cibernética utiliza uma expansão de consulta dinâmica que requer apenas um conjunto reduzido de gatilhos iniciais. Esses gatilhos podem ser palavras-chave, padrões específicos ou características relevantes que, quando identificados em postagens, indicam a possibilidade de eventos de segurança cibernética. A estratégia é fundamentada em padrões de árvore de dependência e emprega um método de expansão de gatilho de evento dinâmico baseado em *kernels* de convolução e análise de dependência. Além disso, incorpora uma estratégia de *embedding* de palavras para capturar semelhanças entre gatilhos de eventos e relatórios de eventos candidatos. As sementes usadas pelo sistema consistem em uma lista de palavras-chave organizadas em três categorias: violação de dados, ataques de negação de serviço e sequestro de conta.

Os autores relatam que os experimentos foram conduzidos utilizando um volume bruto de mais de cinco bilhões de *tweets*. Durante o período de teste, o sistema detectou um grande número de incidentes de segurança de diferentes categorias. Para avaliar o método, foi utilizado um banco de dados organizado pelos autores, composto por fontes confiáveis de divulgação de incidentes de segurança cibernética.

Os resultados apresentados mostram que o sistema teve um bom desempenho e foi capaz de identificar importantes incidentes de segurança cibernética que ocorreram no período observado. No entanto, os autores não abordaram o desempenho do sistema quanto ao processamento. Devido ao processamento de um grande volume de dados, a detecção de eventos em tempo real pode ser comprometida.

Por fim, Arnold et al. (2019) propuseram uma ferramenta de CTI que envolve o uso de análise de grafos e redes complexas para identificar ameaças cibernéticas em fontes de dados da *Dark Web*. Os autores relataram ter utilizado dados coletados de oito fóruns nos idiomas inglês, russo e finlandês para formar uma rede social de múltiplos nós.

Através de consultas SQL, foram identificados 132 nomes de organizações mencionadas nesses fóruns, incluindo empresas conhecidas como *Amazon, PayPal* e *Microsoft*. Utilizando bibliotecas do *Python* e *Gephi* (uma ferramenta de análise de grafos e redes complexas), os autores relataram que identificaram um grande número de ameaças, sendo o maior número relacionado a fraudes, seguido por violação de contas e ferramentas de *hacking* disponíveis para ataques a empresas e seus clientes.

3.2 Extração de Informações para CTI Através de IoCs

Niakanlahiji et al. (2019) apresentaram um *framework* escalável para extração automática IoCs do *Twitter*, utilizando uma combinação de teoria dos grafos, aprendizado de máquina e técnica de mineração de texto. O sistema conta com um modelo de reputação para descobrir perfis confiáveis que publicam informações de CTI e apenas rastreia o fluxo de *tweets* desses perfis. Os autores relataram que, ao longo de quatro semanas, o sistema identificou mais de 1.200 IoCs, incluindo URLs maliciosas.

Já Zhang et al. (2019) apresentaram um sistema capaz de extrair automaticamente IoCs da *Surface Web*, verificando indicadores suspeitos com a ajuda de informações de ameaças de código aberto. O sistema recebe como entrada indicadores considerados suspeitos, como domínios e endereços IP, e verifica se são de ameaças reais, coletando e analisando ativamente suas informações relevantes sobre ameaças de código aberto na *Surface Web*. Com base nos resultados da verificação, o sistema gera uma lista de IoCs. Em seguida, extrai automaticamente novos indicadores das páginas da web relacionadas aos IoCs como novas entradas, repetindo o processo de verificação para gerar mais IoCs.

No trabalho Al-Ramahi, Alsmadi e Davenport (2020), foi apresentada uma abordagem sistemática para extrair automaticamente Tópicos de Interesse (ToIs) de sites de *hackers*, visando utilizá-los como entradas para controles de segurança acionáveis ou coletores de IoCs. Em um primeiro momento, os autores analisaram postagens de *hackers* em um conjunto de dados público. Como segundo experimento, desenvolveram um rastreador para extrair ToIs em um fórum da *Dark Web*. Os resultados foram positivos, porém

os autores relataram vários desafios relacionados ao rastreamento e extração de ToIs relevantes.

Por fim, Caballero et al. (2023) apresentaram uma plataforma para extrair IoCs de seis diferentes fontes: *Blogs RSS, Twitter e Telegram*, bem como *Malpedia, APTnotes* e *ChainSmith*, que são repositórios de projetos relacionados à segurança cibernética. Além de terem desenvolvido a ferramenta de extração de IoCs, os autores relataram que fizeram uma análise para avaliar a precisão de outras 7 ferramentas de extração de IoCs. Os resultados mostraram que a ferramenta desenvolvida obteve maior precisão em 11 dos 13 tipos de IoCs extraídos.

3.3 Extração de Informações para CTI na *Dark Web* Utilizando Aprendizado de Máquina

A estrutura proposta por Sapienza et al. (2017) (Seção 3.1) serviu de base para a abordagem adotada por Dong et al. (2018). No entanto, diferentemente da proposta anterior, a coleta de dados concentra-se exclusivamente em fóruns da *Dark Web*, incorporando o uso de aprendizado de máquina supervisionado na etapa de classificação. O principal objetivo é a identificação de novas ameaças cibernéticas.

A metodologia consiste em monitorar alguns dos maiores mercados da *Dark Web* para coletar itens relacionados à segurança cibernética, equivalentes aos chamados *posts*. Os itens são classificados em quatro categorias *(1–data, 2-carding, 3-hack, 4-others)*, sendo a categoria *3-hack* o foco principal do sistema. Essa categoria engloba vulnerabilidades, ferramentas de *hacking*, *malwares*, tutoriais de exploração, entre outros.

Utilizando técnicas de mineração de texto, os caracteres especiais e palavras irrelevantes são removidos, resultando apenas nos termos considerados importantes. Os autores destacaram que manualmente rotularam 8.000 amostras para treinar o modelo de classificação e que usaram TF-IDF para a representação de texto.

A abordagem adotada parte do princípio de que novos termos, ainda não conhecidos pelo sistema, podem indicar o surgimento de novas vulnerabilidades ou *malwares*. O processo subsequente envolve a verificação para determinar se o termo recém-descoberto representa uma nova ameaça ou se refere a uma já existente. Por fim, o sistema gera avisos para os termos recém-descobertos. Os resultados dos testes na etapa de classificação tiveram alta taxa de acurácia, atingindo 94%. Contudo, é importante destacar que os autores observaram uma considerável incidência de falsos positivos nos alertas gerados pelo sistema.

Já Sarkar et al. (2019) utilizaram dados de fóruns na *Dark Web*, analisando a estrutura de respostas dos usuários para prever ataques cibernéticos corporativos. Essa estrutura captura como as interações estão conectadas entre si, formando uma espécie de rede ou grafo. O sistema proposto tenta prever se haverá um ataque cibernético em um determi-

nado dia para uma organização, para isso aplica modelos de aprendizado supervisionado em um conjunto de recursos extraído dos fóruns. Os dados da *Dark web*, de acordo com os autores, foram adquiridos por meio de uma interface de programação comercial. Primeiramente, foram selecionados um conjunto de fóruns que consideraram mais relevantes. Em seguida, realizaram uma busca por menções de vulnerabilidades nesse conjunto de fóruns, computando, assim, o número total de CVEs mencionados nessas postagens. Os CVEs foram posteriormente agrupados utilizando o esquema de nomenclatura estruturado *Common Platform Enumeration* (CPE) do banco de dados da NVD mantido pelo NIST.

O sistema emprega grafos direcionados para extrair um conjunto de usuários especializados, chamados de *especialistas*, cujos *posts* contendo menções a vulnerabilidades capturam a atenção de outros usuários em um período específico. Em seguida, são geradas séries temporais para capturar as interações desses usuários especialistas nos fóruns da *Darkweb*. Por fim, é aplicado um modelo de aprendizado para tentar prever ataques cibernéticos antes que ocorram. Os resultados apresentados mostraram que o sistema teve uma boa precisão e foi capaz de prever importantes incidentes de segurança que ocorreram no período de testes.

Queiroz, Mckeever e Keegan (2019) propuseram uma abordagem para aprimorar modelos de classificação utilizando modelos de linguagem para a representação de características. Empregaram técnicas de incorporação de palavras e incorporação de sentenças a fim de identificar propriedades contextuais semânticas de palavras e frases, permitindo assim a detecção de ameaças cibernéticas relacionadas a vulnerabilidades em fóruns e redes sociais na *Surface Web* e na *Dark Web*.

Os autores destacaram que o objetivo do trabalho foi investigar o desempenho de modelos de incorporação como o *Word2Vec* na detecção de ameaças de *hackers* em fóruns online. Para isso, foi feita uma comparação com um trabalho anterior realizado em modelos de linguagem clássica. Foram utilizados dois algoritmos de aprendizado de máquina supervisionado para realização dos testes, o SVM e o *Convolutional Neural Networks* (CNN).

Foram utilizados dados provenientes de cinco fontes, incluindo quatro fóruns da *Dark Web* e o *Twitter*, totalizando 9.470 amostras. Especialistas realizaram a rotulagem manual desses dados, identificando aproximadamente 11,8% das mensagens como maliciosas e o restante como não maliciosas. Devido a baixa taxa de revocação nos testes iniciais, técnicas de balanceamento de classes foram aplicadas, o que elevou de 11,8% para cerca de 37,2% as mensagens maliciosas. O modelo que utilizou *Word2Vec* para representação de texto obteve o melhor desempenho, alcançando acurácia de 96% e uma taxa de revocação de 93%, superando os resultados obtidos anteriormente.

Os resultados foram promissores; no entanto, o processo de rotulagem pode ser uma fonte de ruídos, uma vez que inicialmente os dados apresentavam três classes: *Sim, Não*

e *Indeciso*, mas ao final, tudo que havia sido marcado como *Indeciso* acabou sendo considerado como *Sim*.

Já Koloveas et al. (2021) apresentaram uma estrutura integrada para mineração e extração de CTI de diversas fontes, como *Surface Web*, redes sociais e *Dark Web*. Além da coleta, o trabalho envolve análise, gerenciamento e compartilhamento de CTI. O sistema usou *Word2Vec* para representação de texto e dois diferentes algoritmos de aprendizado de máquina supervisionado.

Para o treinamento do modelo, foi construído um conjunto de dados rotulado, o qual teve como foco a segurança em IoT. Adotou-se o conceito de páginas, onde aquelas que mencionavam os termos *Segurança* e *IoT* simultaneamente foram consideradas relevantes, enquanto aquelas que mencionavam apenas um dos termos foram consideradas irrelevantes.

Na implementação prática, os autores relataram a utilização de um conjunto de onze palavras-chave para a coleta de dados no *Twitter*, resultando em um total de 1.677 *tweets*, onde 5.54% foram marcados como relevantes e o restante como irrelevantes. O melhor resultado alcançou 95% de acurácia, 61% de precisão, 73% de revocação e 64% de medida-F.

Embora o trabalho tenha mencionado o uso de várias fontes de dados, a implementação prática relatada utilizou apenas o *Twitter* como fonte de dados. Dependendo da fonte, o uso de páginas inteiras na coleta de dados pode acarretar um grande volume de informações, o que pode dificultar o armazenamento e análise do conteúdo.

3.4 Discussão

A abordagem adotada para esta dissertação apresenta algumas semelhanças com os trabalhos relacionados, desde aqueles que buscaram identificar ameaças cibernéticas por diferentes abordagens onde a fonte de dados principal foi o *Twitter* (Seção 3.1), passando pelos que focaram na extração de CTI através de IoCs (Seção 3.2), e sobretudo aqueles que fizeram uso de aprendizado de máquina supervisionado e utilizaram a *Dark Web* como fonte de dados (Seção 3.3). No entanto, a proposta implementada neste trabalho abrange essas subáreas, trazendo algumas diferenças relevantes. Destacam-se a escolha específica de fóruns no idioma português do Brasil, o processo de rotulagem dos dados, envolvendo a identificação de IoCs, palavras-chave contextuais e análise manual. Além disso, a base rotulada é disponibilizada, e são conduzidos testes com diferentes algoritmos de aprendizado de máquina, utilizando diversas formas de representação de texto. Este trabalho se destaca, sobretudo, pelos testes e pela análise do comportamento do sistema na classificação de novos *posts* não rotulados apresentados a ele.

As propostas apresentadas por (SAPIENZA et al., 2017) e (ARNOLD et al., 2019), embora tenham envolvido a *Dark Web* como fonte de dados, não incorporaram o uso de

aprendizado de máquina na identificação de ameaças. Em contrapartida, os trabalhos conduzidos por (ALVES et al., 2021), (SUBROTO; APRIYANA, 2019), (SCELLER et al., 2017), (RODRIGUEZ; OKAMURA, 2020) e (KHANDPUR et al., 2017) distinguem-se pela escolha do *Twitter* como fonte de dados para a identificação de ameaças cibernéticas.

No âmbito da extração de informações para CTI através de IoCs, os trabalhos (NIAKANLAHIJI et al., 2019), (ZHANG et al., 2019), (AL-RAMAHI; ALSMADI; DAVENPORT, 2020) e (CABALLERO et al., 2023) não abordaram especificamente a extração de *IoCs* em fóruns da *Dark Web*, como realizado nesta dissertação durante a fase de rotulagem dos dados.

Os trabalhos destacados na Seção 3.3 guardam mais semelhanças com esta dissertação, ao envolverem aprendizado de máquina e terem a *Dark Web* como fonte principal de dados. No entanto, há diferenças relevantes, conforme já destacado no início desta Seção (3.4).

Capítulo 4

Materiais e Métodos

Neste capítulo, será detalhado o método adotado para a identificação de *posts* maliciosos na *Dark Web* por meio da aplicação de técnicas de aprendizado de máquina supervisionado. O trabalho foi dividido em três fases principais de desenvolvimento. A primeira compreendeu a construção de conjuntos de dados rotulados, a segunda envolveu o desenvolvimento de um modelo de classificação de *posts*, e a terceira fase incluiu o teste do modelo em um novo conjunto de dados não rotulado. A Seção 4.1 detalha a construção dos conjuntos de dados rotulados, enquanto a Seção 4.2 trata do desenvolvimento e teste do modelo de classificação de *posts*.

Este capítulo irá fornecer uma visão abrangente das etapas e metodologias empregadas ao longo desta dissertação, permitindo uma compreensão mais sólida do processo de identificação de conteúdo malicioso na *Dark Web* por meio de técnicas de aprendizado de máquina. Todas as implementações e recursos relacionados a este trabalho estão disponíveis no *GitHub*, no seguinte endereço: <https://github.com/sebastiaoafilho/Malicious_Posts_Identification>

4.1 Construção de Conjuntos de Dados Rotulados

A primeira fase deste trabalho foi dedicada à construção de conjuntos de dados rotulados, essenciais para o treinamento de algoritmos de aprendizado de máquina supervisionado. Isso se deve à escassez de dados rotulados e às particularidades do projeto, que ficaram evidentes ao perceber que trabalhos como (DONG et al., 2018), (KOLOVEAS et al., 2021) e (QUEIROZ; MCKEEVER; KEEGAN, 2019) também tiveram que construir suas bases de dados rotuladas. Este último até citou uma URL para acesso à base de dados rotulada, porém a mesma não estava disponível durante o desenvolvimento deste trabalho. Além disso, mesmo que esses dados pudessem ser acessados, eles não atenderiam completamente à demanda desta dissertação, que optou por usar mensagens de fóruns no idioma português do Brasil.

A Figura 6 ilustra cada etapa desta fase do desenvolvimento do trabalho, que incia

com a coleta dos *posts*, passando pelo pré-processamento inicial, extração de IoCs, segundo pré-processamento, modelagem de tópicos até chegar a rotulagem dos dados.

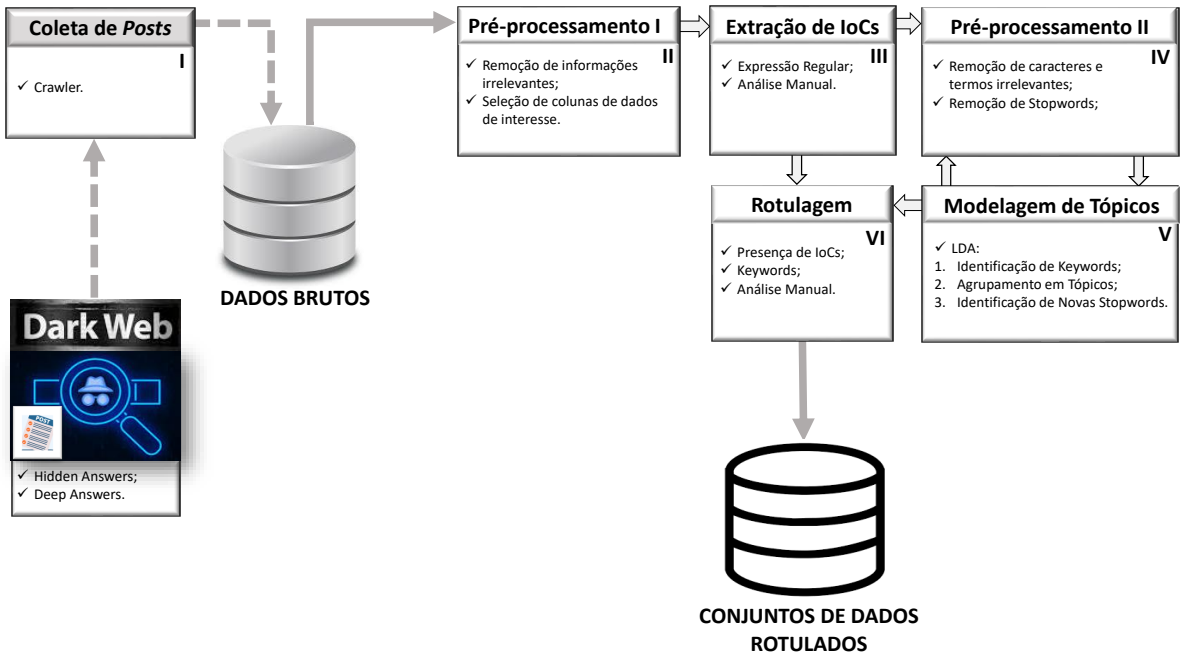


Figura 6 – Etapas da fase de construção dos conjuntos de dados rotulados (Fonte: O autor (2023)).

4.1.1 Etapa I - Coleta de *Posts*

A primeira etapa do processo consiste na coleta de *posts* de dois fóruns da *Dark Web* em língua portuguesa do Brasil: o *Hidden Answers* e o *Deep Answer*. Esses fóruns foram escolhidos por serem considerados abertos, ou seja, acessíveis a qualquer usuário que tenha o *link* ou a URL correspondente. Ao contrário de fóruns restritos ou privados, que exigem convites ou senhas para permitir o acesso, os fóruns abertos são mais acessíveis e costumam ter um fluxo maior de informações e atividades.

Para executar o processo de coleta de dados, foi utilizado um *crawler* – um sistema automatizado que varre os fóruns em busca de postagens. Após a coleta, os dados foram armazenados em uma base de dados no formato JSON, cujos atributos são apresentados na Tabela 3, totalizando 26.575 *posts* para este estudo.

4.1.2 Etapa II - Pré-processamento I

Na segunda etapa do processo, foi desenvolvido o primeiro módulo de pré-processamento dos dados coletados, utilizando a biblioteca *pandas* do *Python*. A finalidade deste estágio inicial de pré-processamento foi realizar uma limpeza e organização dos dados. Isso incluiu a seleção de atributos pertinentes para a análise, a padronização dos nomes desses

Tabela 3 – Atributos presentes nos arquivos JSON coletados dos fóruns *Hidden Answers* e *Deep Answers* (Fonte: O autor (2023)).

Fórum *Hidden Answers*	
1	category
2	title
3	content
4	answers
5	created_at
6	author
7	tags
8	coments
9	best_answer
10	up_votes
11	down_votes

Fórum *Deep Answers*	
1	category
2	title
3	question
4	answers
5	dataCreated
6	author
7	tags
8	type
9	votes
10	points

atributos, a consolidação de todos os *posts* em um único conjunto de dados, a concatenação de alguns atributos do tipo texto em um novo denominado *full_text* e a inclusão do atributo ID para enumerar os *posts* sequencialmente.

Essa etapa foi realizada para garantir que os dados estivessem em condições ideais para a etapa subsequente: a Extração de IoCs (Subseção 4.1.3). Além de eliminar inconsistências, esse processo resultou em dados mais organizados e coesos. Por exemplo, o atributo 3 do fórum *Hidden Answers*, conforme apresentado na Tabela 3, é equivalente ao atributo 3 do fórum *Deep Answers*. Contudo, enquanto em um fórum esse atributo recebe o nome de *content*, no outro ele é chamado de *question*. Algo semelhante ocorreu com o atributo 5, responsável por armazenar a data de criação do *post*. Nesse caso, além dos nomes diferentes, o formato também não era o mesmo. A Tabela 4 apresenta como ficou a definição dos atributos do conjunto de dados nesta etapa de pré-processamento.

Tabela 4 – Atributos definidos na etapa de pré-processamento (Fonte: O autor (2023)).

Atributo			**Descrição**
1	ID		Código sequencial dos *posts*
2	category		Categoria em que o *post* foi incluído
3	full_text	title	Contém o título do *post*
		content	Contém o texto principal do *post*
		answers	Contém respostas e comentários dos usuários
4	created_at		Contém a data em que foi criado o *post*

4.1.3 Etapa III - Extração de IoCs

A terceira etapa deste trabalho engloba o desenvolvimento do módulo de extração de IoCs, utilizando as bibliotecas *pandas*[1] e *re*[2] do *Python*. O propósito central deste módulo é identificar e marcar todos os *posts* que contêm IoCs, sendo essa marcação um dos parâmetros usados para a subsequente rotulagem dos dados. Expressões regulares adaptadas para cada tipo de IoC procurado foram empregadas para identificar padrões nos dados. Por exemplo, uma expressão regular específica foi utilizada para buscar IoCs do tipo e-mail:

```
(r"[a-zA-Z0-9.]+@[a-zA-Z0-9]+.[a-zA-Z]+(.[a-zA-Z]+)*")
```

O *script* completo utilizado para este propósito está disponível no repositório do *GitHub* mencionado no início deste capítulo.

Uma parte dos IoCs foi extraída com o auxílio do *ioc-finder*[3], versão 7.2.4, uma ferramenta de código aberto desenvolvida por Forrest Hightower (HIGHTOWER, 2017) e disponível no *GitHub*. A integração dessa ferramenta no projeto permitiu ampliar a busca e extração de tipos de indicadores relevantes, otimizando a análise de potenciais ameaças. Vale ressaltar que, neste trabalho, o foco está na extração de IoCs do tipo atômico, tais como endereços IP, e-mails, URLs e nomes de domínio.

A Tabela 5 mostra os tipos de IoCs que foram buscados em cada *post*, bem como a ferramenta usada na busca. O sistema criou um atributo para cada tipo de IoC definido na Tabela 5, com o objetivo de registrar sua presença ou ausência em cada *post*. Adicionalmente, um outro atributo denominado *IOC* foi criado para indicar se pelo menos um IoC foi encontrado no *post*. Inicialmente, todos os valores desse atributo foram definidos como *NÃO*.

Tabela 5 – Tipos de IoCs procurados e a ferramenta de extração utilizada (Fonte: O autor (2023)).

Tipo de IoC	**Ferramenta de Busca**
URL	Própria
E-mail	Própria
Domínio	Própria
Hash | MD5, SHA1, SHA256, SHA512 e SSDEEP	IoC-Finder
IPv4	Própria
IPv6	IoC-Finder
ASN	IoC-Finder
CVE	Própria
MAC	IoC-Finder
Registry Key Path	IoC-Finder

[1] Site oficial: <https://pandas.pydata.org/>
[2] Documentação: <https://docs.python.org/3/library/re.html>
[3] ioc-finder: <https://github.com/fhightower/ioc-finder>

Em seguida, o sistema percorreu cada *post* e para aqueles que continham pelo menos um IoC, a coluna correspondente foi marcada com o valor *1*, enquanto a coluna *IOC* foi atualizada para *SIM*. Conforme já citado anteriormente, essa marcação será usada como um dos parâmetros para rotulagem da base de dados para treinamento dos modelos de aprendizado de máquina supervisionado.

A Tabela 6 mostra a marcação em dois *posts*. No *post* de ID 829, foi encontrado pelo menos um IoC do tipo *IP* e pelo menos um do tipo *URL*, logo, as colunas correspondentes foram marcadas com o valor *1*. Consequentemente, a coluna *IOC* foi atualizada para *SIM*. No *post* de ID 830, não foram encontrados IoCs, e portanto nenhuma coluna foi marcada. A coluna *IOC* permaneceu como *NÃO*. Todos os IoCs encontrados foram extraídos e salvos em arquivos de texto, sendo um arquivo para cada tipo de IoC.

Tabela 6 – Marcações feitas para indicar a presença/ausência e o tipo de IoC encontrado nos *posts* (Fonte: O autor (2023)).

ID	category	full_text	created_at	IOC	IP	URL	EML	HSH	CVE	DOM	ASN	IP6	MAC	RKP
829	"Knowledge and info"	"IP DESTE SITE!!! eu tentei pegar o ip deste site https://www.****.com ... Primeiro site: 186.***.***.28 "...	27/08/2020	SIM	1	1								
830	"Other"	"quem é s t a c k z? eu vi q é um youtuber ai sla mas pq quando falam aqui geralmente falam mal ... Ele faz vídeos, como pode perceber. Ontem mesmo vi um, ahaush."...	09/08/2020	NÃO										

No final do processo, uma verificação manual foi realizada para eliminar IoCs que possuíam formatos semelhantes, mas não correspondiam a IoCs legítimos. Como exemplo, a sequência *4.2.0.2* foi identificada durante a busca, apresentando características compatíveis com o formato de um endereço IPv4. Entretanto, neste caso específico, se tratava da versão de um software.

4.1.4 Etapa IV - Pré-processamento II

A preparação dos dados para a extração de IoCs, conforme descrito na Subseção 4.1.2, difere da preparação dos dados destinados à alimentação dos modelos de aprendizado de máquina. Alguns caracteres, como ponto e arroba, utilizados em determinados tipos de IoCs, como endereços IPv4 e e-mails, são indesejáveis quando se trata de mineração de texto. Portanto, tornou-se necessário implementar duas etapas distintas de pré-processamento neste estudo.

Nesta etapa específica, foram conduzidas a remoção de caracteres especiais, números, termos irrelevantes, como *QuestionID* e *AnswerID*, bem como sequências repetitivas, como *kkkkkk* e *aaaaaaa*. Também foram eliminadas *stopwords* (através do pacote *stopwords* da biblioteca *nltk*), tags HTML, URLs e espaços em branco adicionais. Além disso, caracteres acentuados foram substituídos por suas formas não acentuadas, e todo o texto

foi convertido para letras minúsculas. Esta etapa também foi desenvolvida em *Python* e incluiu o uso das bibliotecas *pandas, re, nltk, BeautifulSoup* e *unidecode*.

4.1.5 Etapa V - Modelagem de Tópicos

Esta etapa foi desenvolvida com foco em dois objetivos principais: a identificação e marcação de todos os *posts* que contenham alguma das *palavras-chave* existentes em uma lista pré-definida de *palavras-chave* relacionadas à segurança cibernética, e a identificação de palavras irrelevantes no contexto, que foram tratadas como novas *stopwords*. Este processo foi conduzido utilizando modelagem de tópicos LDA. Além da identificação das palavras-chave, a abordagem permitiu a organização dos dados em tópicos, o que possibilitou uma análise mais profunda do contexto tanto daqueles que continham palavras-chave quanto daqueles que não continham.

Da mesma forma que a detecção de IoCs mencionada na Subseção 4.1.3, a presença de palavras-chave também será usada como um parâmetro no processo de rotulagem, conforme descrito na próxima Subseção (4.1.6). Para indicar a presença de palavras-chave nos *posts*, foi introduzido um atributo denominado *KEYWORD*. Inicialmente, esse atributo foi preenchido com o valor *NÃO*. A Tabela 7 ilustra a marcação em dois *posts*. No *post* de ID 990, foi identificada pelo menos uma palavra-chave da lista pré-definida, consequentemente, o atributo *KEYWORD* correspondente foi atualizado para *SIM*. Por outro lado, no *post* de ID 989, não foram encontradas palavras-chave, resultando na manutenção do atributo *KEYWORD* como *NÃO*.

Tabela 7 – Marcações feitas para indicar a presença/ausência de palavras-chave nos *posts* (Fonte: O autor (2023)).

ID	category	full_text	created_at	IOC	IP	URL	EML	HSH	CVE	DOM	ASN	IP6	MAC	RKP	KEYWORD
989	"Other"	"duvidas produzir video vhs pessoal algumas"...	26/03/2017	NÃO											NÃO
990	"Knowledge and info"	"agora cpfs burguesia link download cpfs"...	25/01/2021	SIM		1	1								SIM

A Tabela 8 mostra as palavras-chave que foram consideradas na busca. Parte dessas palavras foram definidas usando como referência o trabalho de (DELIU; LEICHTER; FRANKE, 2018) e outras foram inseridas levando em consideração o próprio contexto da base de dados.

Já a Tabela 9, mostra um exemplo de palavras que após a análise usando o modelagem de tópicos LDA, foram consideradas como novas *stopwords* e portanto removidas do texto. É importante destacar que o LDA pode ser executado várias vezes, com diferentes números de tópicos pré-definidos, visando identificar o maior número possível de novas *stopwords*.

4.1.6 Etapa VI - Rotulagem

Ter uma base de dados rotulada é um requisito fundamental para o desenvolvimento de modelos de aprendizado de máquina supervisionado. Em alguns casos, é possível

Tabela 8 – Lista de palavras-chave consideradas relevantes no contexto de segurança cibernética (Fonte: O autor (2023)).

Palavras-chave consideradas
cpf, cpfs, cve, password, passwords, senha, senhas, hack, hacker, hackers, hacking, virus, malware, spyware, phishing, fishing, spam, trojan, criptografia, rootkit, backdoor, worm, botnet, vazamento, vazamentos, dados, spoofing, wordlist, ransomware, injection, sqlinjection, ddos, exploit, keylogger, vulnerabilidade, vulnerabilidades, hash, hashes

Tabela 9 – Novas *stopwords* encontradas e removidas do texto (Fonte: O autor (2023)).

Novas *stopwords* encontradas
pra, etc, none, vai, ter, nan, user, author, title, none, name, score, content, down, votes, created, comments, comment, answercontent, vote, type, points, aqui, pode, sobre, fazer, alguem, tudo, regular, coisa, bem, vou, sei, boca, algum, alguns, alguma, algo, nada, bom, entao, acho, quer, the, and, you, cara, coisas, sim, ainda, ver, usar, assim, index

utilizar bases de dados públicas para treinar modelos, o que pode simplificar o processo. No entanto, neste trabalho, devido à natureza da fonte de dados - a *Dark Web* - que ainda é pouco explorada para a identificação de incidentes de segurança, não é comum encontrar dados rotulados disponíveis para uso. Diante dessa dificuldade e considerando a especificidade do estudo, optou-se por construir uma base de dados rotulada própria. Isso representa uma contribuição significativa para a comunidade de segurança da informação, uma vez que essa base poderá ser disponibilizada mediante solicitação.

A etapa de rotulagem dos dados foi conduzida inicialmente considerando a ocorrência simultânea de *IoCs* e *palavras-chave* nos *posts*. Em outras palavras, todos os *posts* que continham pelo menos um *IoC* e pelo menos uma *palavra-chave* foram categorizados como *Relevantes*. Por outro lado, aqueles que não continham nenhuma dessas ocorrências foram classificados como *Não Relevantes*. Já os *posts* que continham apenas uma dessas ocorrências foram sinalizados para posterior análise. Esse primeiro conjunto de dados rotulado foi chamado de *CONJUNTO DE DADOS I*.

Sabe-se, porém, que o processo de rotulagem de dados é um trabalho minucioso que exige uma avaliação cuidadosa. Portanto, esse primeiro conjunto de dados foi empregado apenas para verificar o desempenho inicial dos algoritmos de aprendizado de máquina. A rotulagem final envolveu análise manual, na qual algumas marcações que inicialmente consideraram apenas a presença de IoCs e palavras-chave foram alteradas, levando em consideração o conteúdo dos *posts* e outras características, como a categoria à qual pertenciam. A versão final da base rotulada compreendeu todos os 26.575 *posts* iniciais, de forma que, após a análise, aqueles que continham apenas IoCs ou apenas palavras-chave e que haviam sido retirados da base, foram marcados como *Relevantes* ou *Não Relevan-*

tes e inseridos de volta na base. Esta versão final da base rotulada foi nomeada como *CONJUNTO DE DADOS II*.

4.2 Desenvolvimento do Modelo de Classificação de *Posts* e Testes

Com os dados rotulados, o trabalho avançou para a fase de desenvolvimento do modelo de classificação de *posts* e, em seguida, para a fase de testes de identificação de *posts* relevantes ou potencialmente maliciosos em novos dados coletados da *Dark Web*.

A Figura 7, exemplifica as etapas do desenvolvimento do modelo de classificação de *posts*, iniciando na representação vetorial do texto que foi cuidadosamente processado na primeira fase, passando pelos algoritmos de aprendizado de máquina, classificação, modelagem de tópicos e resultados.

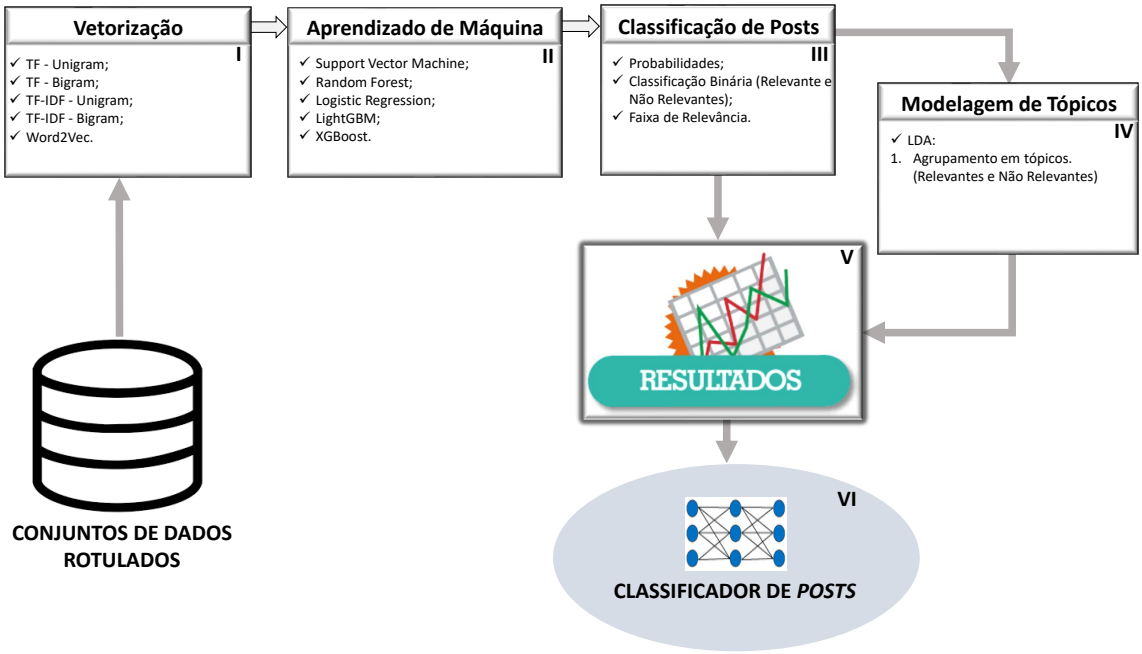


Figura 7 – Etapas da fase de desenvolvimento do modelo de classificação de *posts* (Fonte: O autor (2023)).

4.2.1 Etapa I - Vetorização

A etapa de vetorização consiste na representação do texto em forma de vetores numéricos, o que inclui dois passos essenciais: a extração de atributos e a geração de *tokens*. Neste processo, o texto foi representado de cinco maneiras diferentes: *TF Unigram, TF Bigram, TF-IDF Unigram, TF-IDF Bigram e Word2Vec*.

Essas diferentes representações garantem que os dados estejam prontos para serem processados pelos algoritmos de aprendizado de máquina, permitindo a extração de infor-

mações e a realização de análises mais aprofundadas. Alguns trabalhos correlatos, como (DONG et al., 2018), (QUEIROZ; MCKEEVER; KEEGAN, 2019), (SAMTANI; ZHU; CHEN, 2020) e (KOLOVEAS et al., 2021), citaram o uso de algumas dessas representações, principalmente envolvendo *Word2Vec* e TF-IDF. No entanto, durante a investigação, não foi encontrado algum trabalho que tenha testado todas essas representações.

A decisão de explorar diversas representações de texto foi motivada pela capacidade única de cada técnica de vetorização em capturar informações específicas do texto, conforme detalhado na Seção 2.3.3. Enquanto as abordagens TF e TF-IDF enfatizam a importância da frequência e frequência inversa de termos, respectivamente, o método *Word2Vec* destaca-se ao considerar o contexto semântico. O objetivo subjacente a essa variedade de abordagens foi identificar qual delas se ajusta de maneira mais eficaz ao contexto em questão.

4.2.2 Etapa II - Aprendizado de Máquina

Nesta etapa, foram aplicados cinco algoritmos diferentes de aprendizado de máquina supervisionado com o objetivo de selecionar aquele que apresentasse as melhores métricas de desempenho. A maioria dos parâmetros dos algoritmos foi mantida nos valores padrões, com algumas poucas alterações específicas para cada um. Os algoritmos testados incluíram: *Support Vector Machine, Random Forest, Logistic Regression*, LightGBM e XGBoost.

A escolha desses algoritmos baseou-se, principalmente, em seu desempenho comprovado na literatura relacionada. O *SVM* foi adotado em trabalhos como os de (DELIU; LEICHTER; FRANKE, 2018), (DONG et al., 2018), (QUEIROZ; MCKEEVER; KEEGAN, 2019) e (KOLOVEAS et al., 2021). Já o *Random Forest* foi utilizado por (KOLOVEAS et al., 2021), e a *Logistic Regression* foi testada por (SARKAR et al., 2019) e (KOLOVEAS et al., 2021). Embora LightGBM e XGBoost não tenham sido especificamente abordados nos estudos relacionados, a decisão de incluí-los se justifica pelos reconhecidos benefícios em termos de eficiência, tempo de treinamento e habilidade para lidar com classes desbalanceadas. Essas características foram identificadas durante a revisão da literatura realizada para a fundamentação teórica (Capítulo 2) desta dissertação. Além disso, esses algoritmos são frequentemente utilizados em competições do *Kaggle*, uma plataforma online de ciência de dados que hospeda desafios, conjuntos de dados e recursos para entusiastas e profissionais de aprendizado de máquina e ciência de dados (BOJER; MELDGAARD, 2021).

Inicialmente, a divisão entre treino e teste foi definida como 80% para treino e 20% para teste; posteriormente, também foi testado com uma divisão de 90% para treino e 10% para teste. Essas configurações de divisão foram adotadas para permitir uma análise comparativa bem como a identificação da melhor estratégia para a aplicação em questão.

4.2.3 Etapa III - Classificação de *Posts*

Utilizando o *CONJUNTO DE DADOS I* — a primeira base de dados rotulada conforme descrito na Subseção 4.1.6 — foram testados os cinco algoritmos mencionados na Subseção 4.2.2, utilizando as cinco diferentes representações de dados citadas na Subseção 4.2.1. Isso resultou em um total de 25 combinações de algoritmo e representação de dados. Foram selecionados os melhores modelos, ou seja, aqueles que alcançaram valores acima de 60% em todas as principais métricas, incluindo acurácia, precisão, revocação e medida F. Esses modelos foram escolhidos para o treinamento com o segundo conjunto de dados, o *CONJUNTO DE DADOS II*, que é a base de dados rotulada contendo todos os *posts* previamente coletados.

Após o treinamento, foi selecionado o melhor modelo para realizar a classificação dos *posts*. Inicialmente, a classificação ocorreu na própria base rotulada como parte do processo de validação. Em seguida, para avaliar o comportamento do modelo na classificação de novos *posts*, utilizou-se uma nova base contendo 7.498 *posts* coletados dos mesmos fóruns, que nunca haviam sido previamente analisados pelo modelo. Essa nova base foi nomeada como *CONJUNTO DE DADOS III*. A Figura 8 ilustra o fluxo completo do processo de classificação, incluindo a seleção do melhor modelo de classificação.

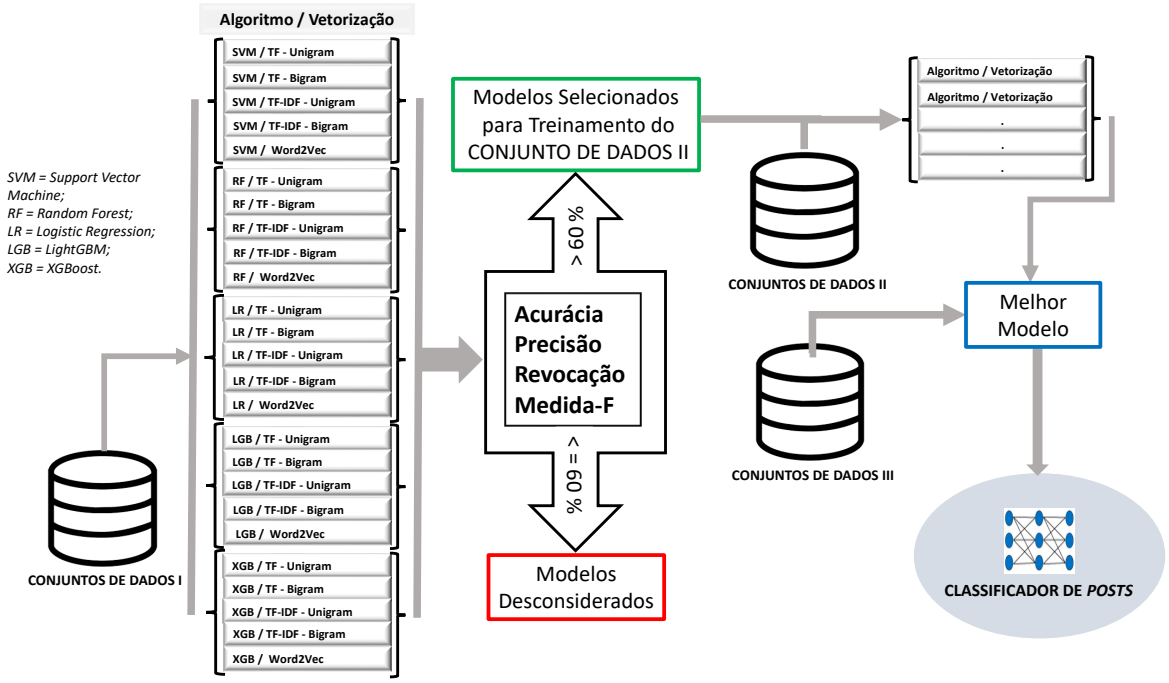


Figura 8 – Fluxo do processo de classificação de *posts* (Fonte: O autor (2023)).

A etapa de classificação foi conduzida aplicando a função *predict* do algoritmo de classificação. Essa função atribui uma probabilidade entre 0 e 1 para indicar a relevância do *post*. Essa probabilidade reflete a confiança do modelo na classificação da instância como relevante ou não relevante. Para interpretação dos resultados, foi estabelecido um limiar de 0,5. Probabilidades inferiores a 0,5 são consideradas como *Não Relevante*, enquanto

aquelas iguais ou superiores a 0,5 são consideradas como *Relevante*. Uma abordagem adicional também foi testada, na qual três faixas de relevância foram definidas: menor que 0,3 como *Baixa Relevância*, entre 0,3 e 0,7 como *Relevância Média*, e acima de 0,7 como *Alta Relevância*.

4.2.4 Etapa IV - Modelagem de Tópicos

Tendo os *posts* classificados como *Relevantes* e *Não Relevantes*, foi aplicada a modelagem de tópicos usando LDA a fim de identificar tópicos no conjunto de dados. Cada tópico é composto por um conjunto de documentos, com palavras atribuídas com base em uma distribuição de probabilidade. A modelagem foi realizada nos dois conjuntos de dados, *II e III*. Inicialmente, todos os *posts* nos conjuntos de dados foram agrupados em 20 e depois em 10 tópicos para obter uma visão geral do agrupamento. Em seguida, usando apenas os *posts* de cada grupo - *Não Relevantes e Relevantes* - foi realizado o agrupamento em 10 tópicos. A Tabela 10 apresenta detalhes sobre essa divisão em tópicos.

Tabela 10 – Agrupamento em tópicos realizado nos conjuntos de dados II e III usando LDA (Fonte: O autor (2023)).

Conjunto de Dados	**Número de Tópicos**	**Abrangência**
CONJUNTO DE DADOS II	20	Todos os *posts*
	10	Todos os *posts*
	10	*Posts* Não Relevantes
	10	*Posts* Relevantes
CONJUNTO DE DADOS III	20	Todos os *posts*
	10	Todos os *posts*
	10	*Posts* Não Relevantes
	10	*Posts* Relevantes

O agrupamento em tópicos realizado tanto na base de dados rotulada *CONJUNTO DE DADOS II* quanto na base de dados não rotulada *CONJUNTO DE DADOS III*, que foi classificada por meio do melhor modelo de classificação treinado, teve como objetivo avaliar a semelhança entre os tópicos de cada conjunto e, consequentemente, obter uma visão geral do desempenho do modelo, mesmo sem o rótulo das amostras.

4.2.5 Etapa V - Resultados

Nesta etapa, os resultados obtidos nas etapas anteriores são apresentados e discutidos. Em relação aos algoritmos de aprendizado de máquina supervisionado, foram apuradas as principais métricas de desempenho, como acurácia, precisão, revocação e medida F, além da análise das matrizes de confusão geradas para cada modelo treinado.

No que diz respeito à modelagem de tópicos, foram analisadas as palavras mais significativas de cada tópico a fim de compreender características relevantes de cada um. Além

disso, realizou-se uma comparação entre os tópicos gerados a partir dos conjuntos de dados rotulados e não rotulados. Isso permitiu uma visão mais detalhada do comportamento do modelo na classificação de dados ainda não vistos.

4.2.6 Identificação de *Posts* Relevantes em Novos Dados Coletados da *Dark Web*

Para testar o classificador de *posts* desenvolvido foi usado o conjunto de dados não rotulado que contém *posts* novos não conhecidos pelo modelo, conforme já descrito na Subseção 4.2.3. A Figura 9, exemplifica as etapas do processo de teste do classificador.

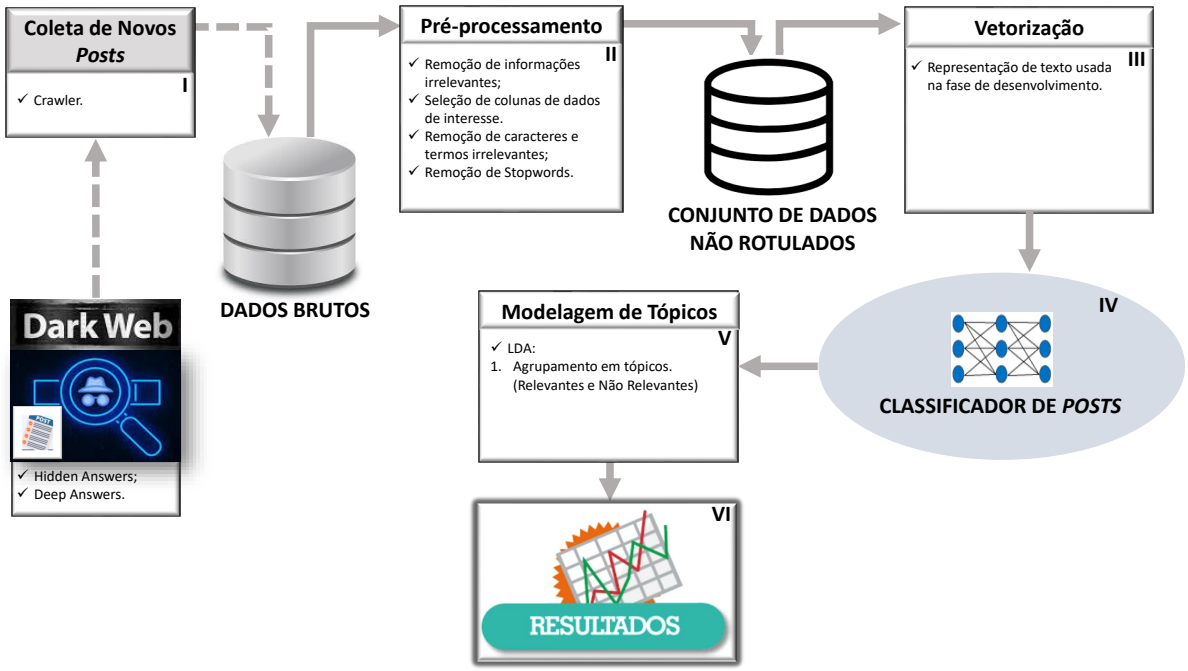


Figura 9 – Etapas da fase de testes do modelo de identificaçaõ de *posts* relevantes em novos dados coletados da *Dark Web* (Fonte: O autor (2023)).

O processo se inicia com a coleta de novos *posts*, que, após passarem por um pré-processamento, dão origem a um novo conjunto de dados não rotulado, denominado *CONJUNTO DE DADOS III*. Em seguida, o texto é vetorizado usando o mesmo padrão de vetorização definido na fase de desenvolvimento e submetido ao classificador. Após a classificação, é realizada uma modelagem de tópicos, gerando tópicos para ambas as classes, que serão posteriormente comparados com os tópicos da base de dados rotulada para análise dos resultados.

Capítulo 5

Experimentos e Análise dos Resultados

Neste capítulo, serão detalhados os experimentos realizados conforme planejados no Capítulo 4 e os resultados obtidos. A Seção 5.1 detalha os conjuntos de dados, a Seção 5.2 detalha o processo para escolha dos melhores modelos de classificação, a Seção 5.3 mostra os resultados dos testes dos melhores classificadores, a Seção 5.4 mostra o desempenho do modelo na classificação de novos *posts*, a Seção 5.5 mostra a análise de tópicos LDA, a Seção 5.6 exibe a análise das palavras mais frequentes nas classes de *post Relevantes* e *Não Relevantes* dos conjuntos de dados e por fim a Seção 5.7 traz uma comparação desta dissertação com alguns trabalhos relacionados.

5.1 Conjuntos de Dados

Conforme já mencionado no Capítulo 4, os dados analisados neste estudo foram coletados de dois fóruns da *Dark Web*: o *Hidden Answers* e o *Deep Answer*. O processo de coleta foi realizado utilizando um *crawler* desenvolvido em linguagem *Go*. Para treinamento dos modelos de aprendizado de máquina supervisionado, foram usados 26.575 *posts*, os quais foram rotulados dando origem aos *CONJUNTOS DE DADOS I* e *II*, detalhados na Subseção 5.1.1.

A Tabela 11 apresenta os detalhes da base de dados coletada, incluindo a quantidade de *posts*, o período de postagem e o idioma das mensagens dos dois fóruns que compõem os conjuntos de dados. Observa-se um intervalo nos períodos de postagem do fórum *Hidden Answers*, isso de deu devido a um período de inatividade do fórum.

Tabela 11 – Detalhes dos *posts* coletados para a base de dados de treinamento dos modelos de aprendizado de máquina supervisionado (Fonte: O autor (2023)).

Fórum	**Período de Postagem**	***Posts***	**Idioma**
Hidden Answers	Entre 26/11/2016 e 12/04/2021	19.652	Português do Brasil
Hidden Answers	Entre 31/07/2021 e 15/07/2022	6.681	Português do Brasil
Deep Answers	Entre 24/08/2021 e 14/09/2022	242	Português do Brasil
	Total de *posts*:	**26.575**	

Para testar o modelo de classificação desenvolvido, foram utilizados 7.498 novos *posts* extraídos dos mesmos fóruns, conforme detalhado na Tabela 12. O objetivo foi apresentar ao modelo mensagens desconhecidas e observar seu comportamento. Conforme mencionado na Subseção 4.2.3, esse conjunto de dados, composto por novos *posts* não rotulados, foi denominado *CONJUNTO DE DADOS III*.

Tabela 12 – Detalhes dos *posts* coletados para a base de dados de testes do modelo de classificação de *posts* (Fonte: O autor (2023)).

Fórum	**Período de Postagem**	***Posts***	**Idioma**
Hidden Answers	Entre 10/09/2022 e 10/07/2023	7.343	Português do Brasil
Deep Answers	Entre 16/09/2022 e 01/01/2023	155	Português do Brasil
	Total de *posts*:	**7.498**	

5.1.1 Rotulagem dos Conjuntos de Dados

Conforme descrito na Subseção 4.1.6, o processo de rotulagem dos *posts* foi realizado por meio de duas abordagens diferentes. Na primeira abordagem, considerou-se a ocorrência simultânea de IoCs e *palavras-chave*. Já na segunda abordagem, além de considerar a ocorrência de IoCs e *palavras-chave*, realizou-se uma análise manual, levando em consideração o conteúdo dos *posts* e outras características, como a categoria.

Na primeira abordagem, dos 26.575 *posts* detalhados na Tabela 11, verificou-se que 16.010 não continham nenhum IoC nem nenhuma palavra-chave, portanto, foram marcados como *Não Relevantes*. Em 6.926 *posts*, foi encontrado pelo menos um IoC, enquanto em 5.304 constatou-se a presença de palavras-chave. A interseção da ocorrência de *IoC* com a presença de *palavras-chave* totalizou 1.665 *posts*, os quais foram marcados como *Relevantes*. A Figura 10 ilustra essa distribuição com os percentuais aproximados de cada rótulo.

Com essa abordagem, construiu-se o que foi denominado de *CONJUNTO DE DADOS I*, ou seja, uma primeira versão de uma base de dados rotulada para o treinamento dos modelos de aprendizado de máquina supervisionado. Essa base contém um total de 17.675 *posts*, dos quais 1.665 estão marcados como *Relevantes*, representando aproximadamente 9%, enquanto os outros 16.010, cerca de 91%, estão marcados como *Não Relevantes*. A Figura 11 ilustra a configuração deste conjunto de dados, apresentando dois gráficos em forma de pizza. À esquerda, destacam-se a fatia amarela, que representa os *posts* que contêm apenas IoC, e a fatia rosa, que representa os *posts* que contêm apenas palavras-chave, sendo destacadas do círculo. Já no gráfico à direita, é mostrado o círculo menor, onde aparecem apenas as fatias vermelha e verde, que representam os *posts Relevantes* e *Não Relevantes*, respectivamente.

Na segunda abordagem foram considerados todos os 26.575 *posts*, de forma que aqueles que continham apenas IoCs ou apenas palavras-chave e que haviam sido retirados da

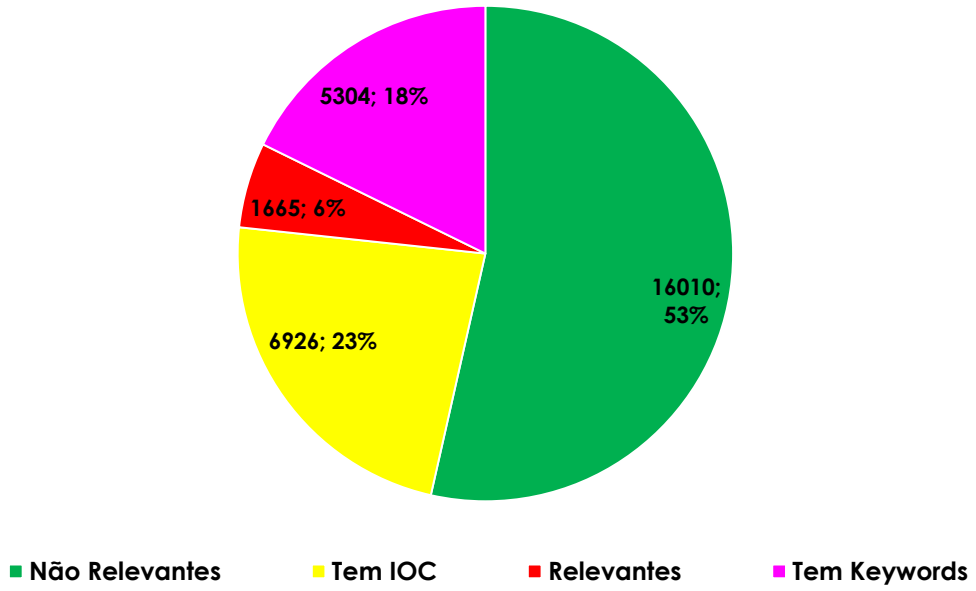


Figura 10 – Primeira rotulagem considerando apenas a presença/ausência de IoCs e palavras-chave nos *posts* (Fonte: O autor (2023)).

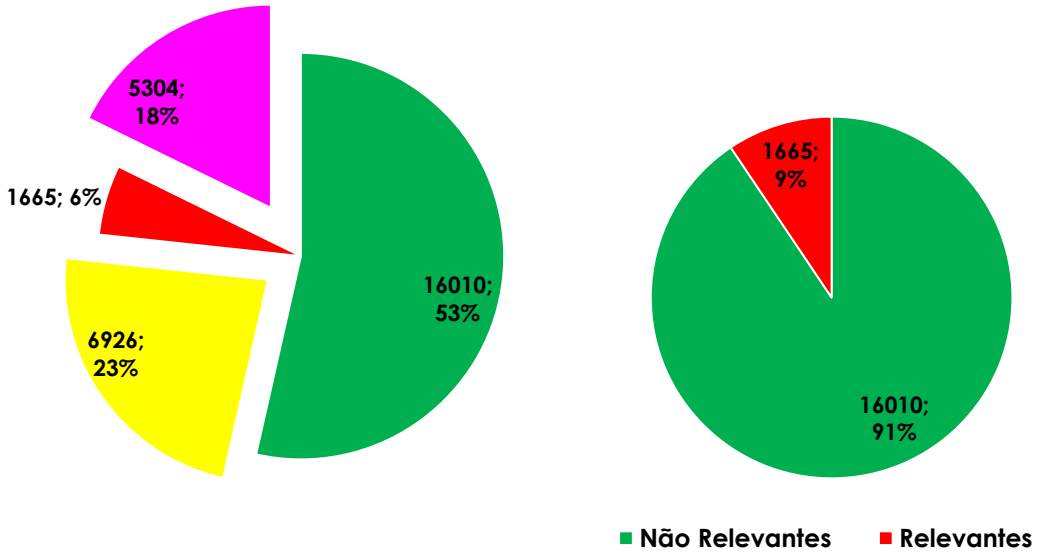


Figura 11 – *CONJUNTO DE DADOS I* rotulado para teste inicial dos modelos de aprendizado de máquina supervisionado (Fonte: O autor (2023)).

base na primeria etapa de rotulagem, foram marcados como *Relevantes* ou *Não Relevantes* e inseridos de volta. Essa versão final, foi chamada de *CONJUNTO DE DADOS II*. A Figura 12 mostra como a rotulagem final ficou em comparação com a marcação inicial. À esquerda, encontra-se a marcação inicial, na qual ainda estão presentes as cores amarela e rosa, que representam os *posts* que contêm apenas IoCs ou apenas palavras-chave, res-

pectivamente. À direita, encontram-se apenas as cores vermelha e verde, que indicam os *posts Relevantes* e *Não Relevantes*, nessa ordem. Um total de 3.341 *posts*, o que representa aproximadamente 13% da base, foi marcado como *Relevantes*, enquanto 23.234, cerca de 87%, foram marcados como *Não Relevantes*.

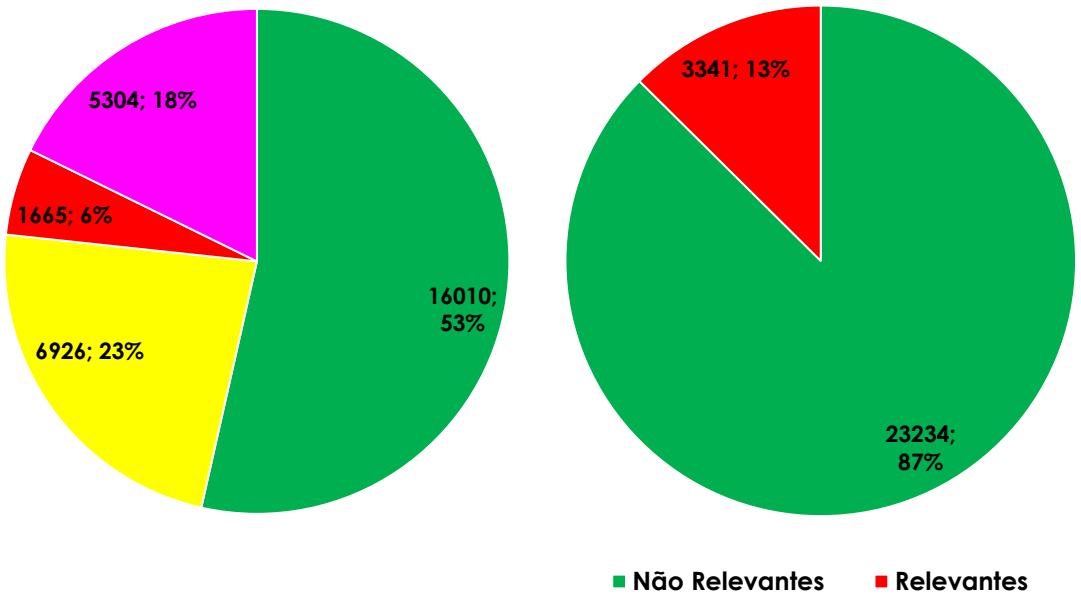


Figura 12 – *CONJUNTO DE DADOS II* rotulado para treinamento dos modelos de aprendizado de máquina supervisionado (Fonte: O autor (2023)).

A Tabela 13 apresenta um resumo dos conjuntos de dados utilizados no desenvolvimento desta pesquisa. É importante ressaltar que os *posts* dos *CONJUNTOS DE DADOS I* e *II* fazem parta da mesma base de dados detalhada na Tabela 11, diferindo apenas no processo de rotulagem. Por outro lado, os *posts* do *CONJUNTO DE DADOS III* são distintos, coletados com o propósito de testar o modelo de classificação, conforme previamente mencionado nesta seção.

Tabela 13 – Detalhes dos conjuntos de dados usados no desenvolvimento da pesquisa (Fonte: O autor (2023)).

Conjunto de Dados	**Total de Posts**	**Rótulo**		
CONJUNTO DE DADOS I	17.675	Relevante	1.665	9,42%
		Não Relevante	16.010	90,58%

Conjunto de Dados	**Total de Posts**	**Rótulo**		
CONJUNTO DE DADOS II	26.575	Relevante	3.341	12,57%
		Não Relevante	23.234	87,43%

Conjunto de Dados	**Total de Posts**	**Rótulo**
CONJUNTO DE DADOS III	7.498	Sem Rótulo

5.2 Seleção dos Melhores Modelos de Classificação

Para a seleção dos melhores classificadores, foram avaliados cinco algoritmos de classificação em combinação com cinco representações de dados distintas. A divisão dos dados consistiu em 80% para treinamento e 20% para teste.

O conjunto de dados usado nesta fase de testes foi o *CONJUNTO DE DADOS I*, o qual foi rotulado levando em consideração a ocorrência simultânea de IoCs e *palavras-chave* nos *posts*. Essa abordagem de rotulagem serviu justamente como base para a avaliação das combinações de algoritmos de aprendizado e representações de texto. A seguir, são listados os cinco algoritmos de aprendizado de máquina supervisionado utilizados nos testes, juntamente com as cinco formas de representação de texto aplicadas. Cada algoritmo foi avaliado com as cinco representações de texto.

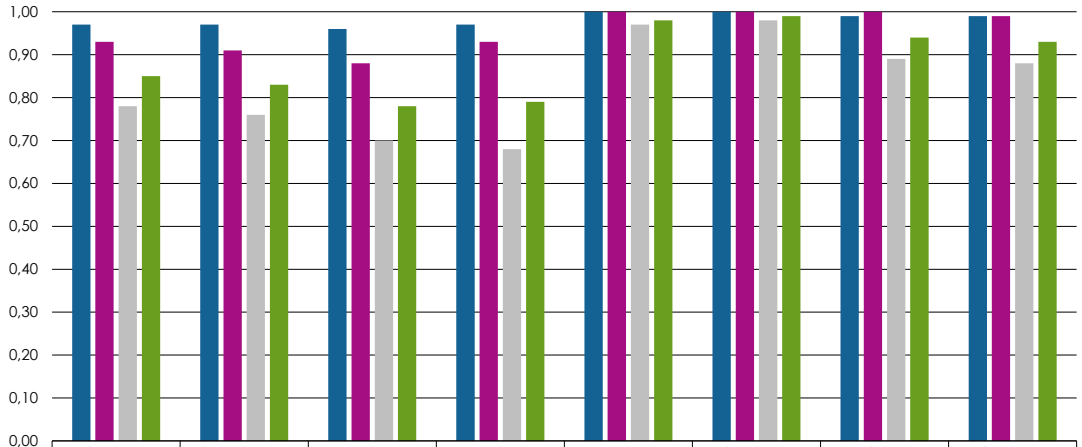
- ❏ Algoritmos: SVM, *Random Forest*, *Logistic Regression*, LightGBM e XGBoost.
- ❏ Representações de Texto: TF - *Unigram*, TF - *Bigram*, TF-IDF - *Unigram*, TF-IDF - *Bigram* e *Word2Vec*.

Dentre as 25 combinações testadas (algoritmo de aprendizado / representação de texto), foram escolhidas aquelas que alcançaram métricas de acurácia, precisão, revocação e medida F superiores a 60%. É importante ressaltar que, devido ao desbalanceamento das classes, a acurácia isolada pode ser enganosa, pois favorece a classe majoritária. As combinações de algoritmos *SVM*, *Logistic Regression*, LightGBM e XGBoost com as representações de *TF-Unigram* e *TF-IDF-Unigram* foram as que obtiveram valores de métricas acima do mínimo definido. Essas combinações, que apresentaram os melhores resultados, são destacadas na Tabela 14, enquanto a Figura 13 mostra o gráfico com os valores das métricas de desempenho de cada classificador para a classe 1 *(Posts Relevantes)*.

Tabela 14 – Algoritmos de aprendizado de máquina supervisionado e representações de texto que alcançaram métricas acima de 60% (Fonte: O autor (2023)).

Representação Vetorial do Texto	**Algoritmos de Aprendizado de Máquina Supervisionado**				
	Support Vector Machine	Random Forest	Logistic Regression	LightGBM	XGBoost
TF – Uingram	✓		✓	✓	✓
TF – Bigram					
TF-IDF - Unigram	✓		✓	✓	✓
TF-IDF - Bigram					
Word2Vec					

Observa-se que os resultados são altamente satisfatórios, com a maioria das métricas acima de 90%, algumas atingindo até 100%. Apenas alguns valores se situam na faixa dos 70%. Esses resultados são coerentes com as expectativas, pois acredita-se que o processo de rotulagem adotado para o *CONJUNTO DE DADOS I* efetivamente segmentou os



	SVM TF - Unigram	SVM TF-IDF - Unigram	LR TF - Unigram	LR TF-IDF - Unigram	LGB TF - Unigram	LGB TF-IDF - Unigram	XGB TF - Unigram	XGB TF-IDF - Unigram
Acurácia	0,97	0,97	0,96	0,97	1,00	1,00	0,99	0,99
Precisão	0,93	0,91	0,88	0,93	1,00	1,00	1,00	0,99
Revocação	0,78	0,76	0,70	0,68	0,97	0,98	0,89	0,88
Medida F	0,85	0,83	0,78	0,79	0,98	0,99	0,94	0,93

Figura 13 – Métricas de desempenho dos melhores classificadores testados no *CONJUNTO DE DADOS I* (Fonte: O autor (2023)).

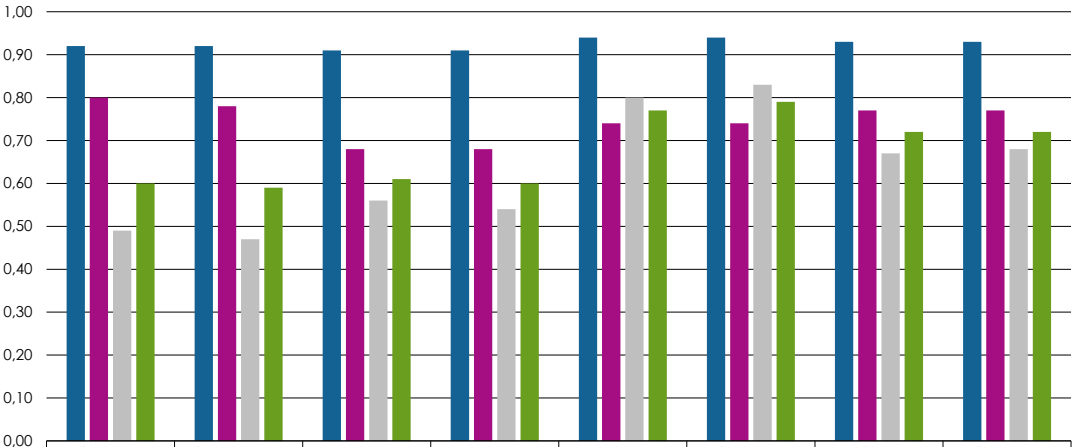
posts em duas classes distintas. De maneira geral o algoritmo LightGBM usando *TF-IDF - Unigram* foi o que apresentou melhor resultado.

5.3 Testes dos Melhores Modelos de Classificação

Os melhores modelos de classificação, selecionados conforme descrito na Seção 5.2, foram treinados usando o conjunto de dados *CONJUNTO DE DADOS II*. O objetivo foi encontrar o melhor modelo de classificação de *posts* dentre as oito combinações (algoritmo de aprendizado / representação de texto) que apresentaram as melhores métricas de desempenho no experimento anterior.

Na primeira etapa deste treinamento, realizou-se uma divisão de dados, alocando 80% para treinamento e 20% para teste. Posteriormente, testou-se o modelo de melhor desempenho com uma divisão de 90% para treino e 10% para teste, contudo, observou-se que essa alteração não resultou em mudanças significativas nos resultados. O resumo dos resultados para a classe 1 *(Posts Relevantes)*, usando a divisão de dados 80% para treino e 20% para teste é apresentado no gráfico da Figura 14. Observa-se que os resultados não atingem o mesmo nível alcançado no primeiro conjunto de dados, o que era esperado devido à abordagem adotada na rotulagem dos dados. No entanto, os valores ainda são satisfatórios.

O algoritmo LightGBM, utilizando a abordagem *TF-IDF - Unigram*, mais uma vez apresentou os melhores resultados, consolidando-se como um forte candidato a modelo de classificação a ser adotado. Os resultados detalhados de cada algoritmo com a res-



	SVM TF - Unigram	SVM TF-IDF - Unigram	LR TF - Unigram	LR TF-IDF - Unigram	LGB TF - Unigram	LGB TF-IDF - Unigram	XGB TF - Unigram	XGB TF-IDF - Unigram
Acurácia	0,92	0,92	0,91	0,91	0,94	0,94	0,93	0,93
Precisão	0,80	0,78	0,68	0,68	0,74	0,74	0,77	0,77
Revocação	0,49	0,47	0,56	0,54	0,80	0,83	0,67	0,68
Medida F	0,60	0,59	0,61	0,60	0,77	0,79	0,72	0,72

Figura 14 – Métricas de desempenho dos melhores classificadores testados no *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

pectiva representação de texto são mostrados nas subseções a seguir, sendo: Subseção 5.3.1 desempenho do algoritmo SVM, Subseção 5.3.2 desempenho do algoritmo *Logistic Regression*, Subseção 5.3.3 desempenho do algoritmo LightGBM e 5.3.4 desempenho do algorimo XGBoost.

5.3.1 Desempenho do Algoritmo SVM

A Tabela 15 e a Figura 15 mostram respectivamente as métricas de desempenho e a matriz de confusão do algoritmo de classificação SVM usando *TF - Unigram*.

Tabela 15 – Métricas de desempenho do algoritmo de classificação SVM usando *TF - Unigram* (Fonte: O autor (2023)).

		Acurácia		*0,92*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.639*	*0,93*	*0,98*	*0,96*
Relevantes	*676*	*0,80*	*0,49*	*0,60*
Total de posts	***5.315***			

Já na Tabela 16 e na Figura 16 são mostrados respectivamente os valores das métricas de desempenho e a matriz de confusão do algoritmo SVM usando *TF-IDF - Unigram*.

O modelo apresentou um desempenho semelhante para ambas as formas de representação vetorial de texto. Na classe dos *posts Não Relevantes*, o índice de acerto foi muito bom, com um baixo número de falsos positivos. No entanto, na classe dos *posts Relevan-*

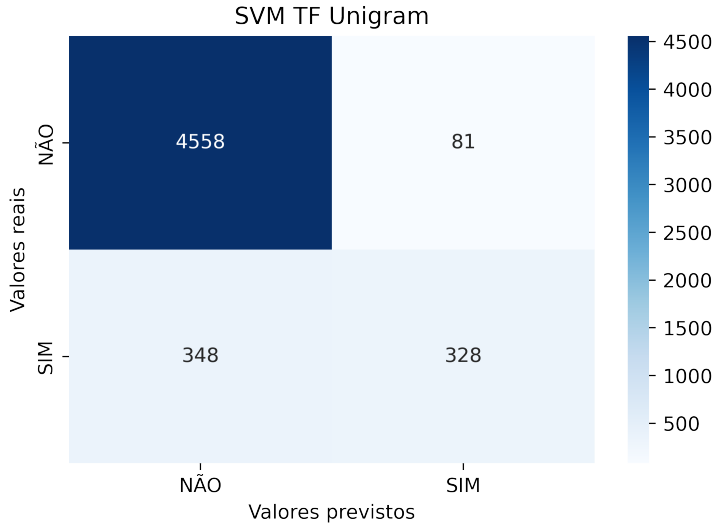


Figura 15 – Matriz de confusão do algoritmo SVM usando *TF - Unigram* (Fonte: O autor (2023)).

Tabela 16 – Métricas de desempenho do algoritmo de classificação SVM usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

			Acurácia	*0,92*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.657*	*0,93*	*0,98*	*0,95*
Relevantes	*658*	*0,78*	*0,47*	*0,59*
Total de posts	***5.315***			

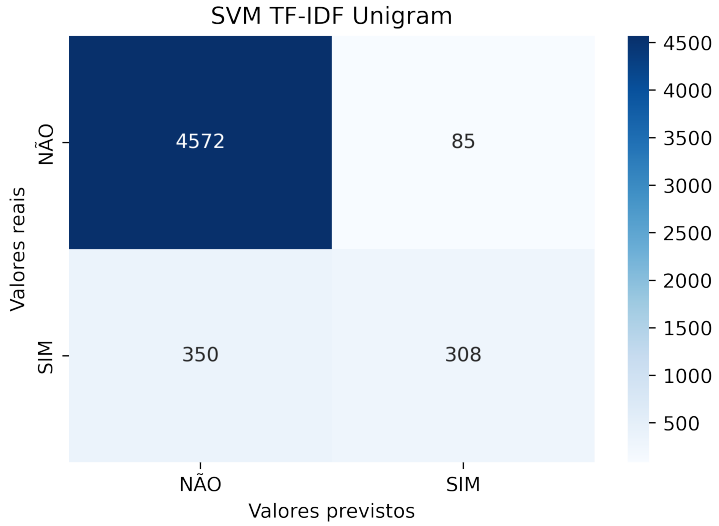


Figura 16 – Matriz de confusão do algoritmo SVM usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

tes, o índice de acerto foi menor, com uma taxa de revocação inferior a 50%. Em outras palavras, o modelo identificou como relevantes menos de 50% dos *posts* verdadeiramente

relevantes apresentados a ele.

5.3.2 Desempenho do Algoritmo Regressão Logística

Os resultados do desempenho do algoritmo de Regressão Logística são apresentados a seguir para duas abordagens distintas: *TF - Unigram* e *TF-IDF Unigram*. Na Tabela 17, são mostrados os valores das métricas de desempenho correspondentes à abordagem *TF - Unigram*, enquanto a Tabela 18 apresenta as métricas para a abordagem *TF-IDF Unigram*. Além disso, as Figuras 17 e 18 mostram as matrizes de confusão associadas a cada abordagem, respectivamente.

Tabela 17 – Métricas de desempenho do algoritmo de classificação Regressão Logística usando *TF - Unigram* (Fonte: O autor (2023)).

		Acurácia		*0,91*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.657*	*0,94*	*0,96*	*0,95*
Relevantes	*658*	*0,68*	*0,56*	*0,61*
Total de posts	***5.315***			

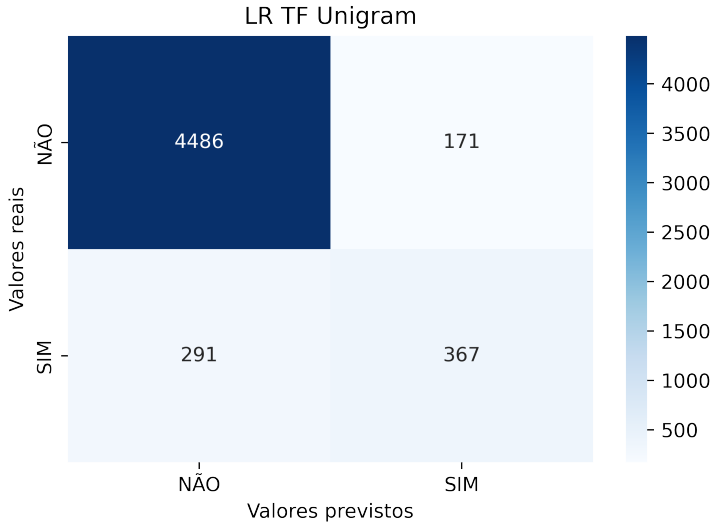


Figura 17 – Matriz de confusão do algoritmo Regressão Logística usando *TF - Unigram* (Fonte: O autor (2023)).

Em comparação com o SVM, o modelo de Regressão Logística apresentou um desempenho ligeiramente superior, com uma taxa de acerto acima de 50% para os *posts Relevantes* em ambas as abordagens.

Tabela 18 – Métricas de desempenho do algoritmo de classificação Regressão Logística usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

			Acurácia	*0,91*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.672*	*0,94*	*0,97*	*0,95*
Relevantes	*643*	*0,68*	*0,54*	*0,60*
Total de posts	***5.315***			

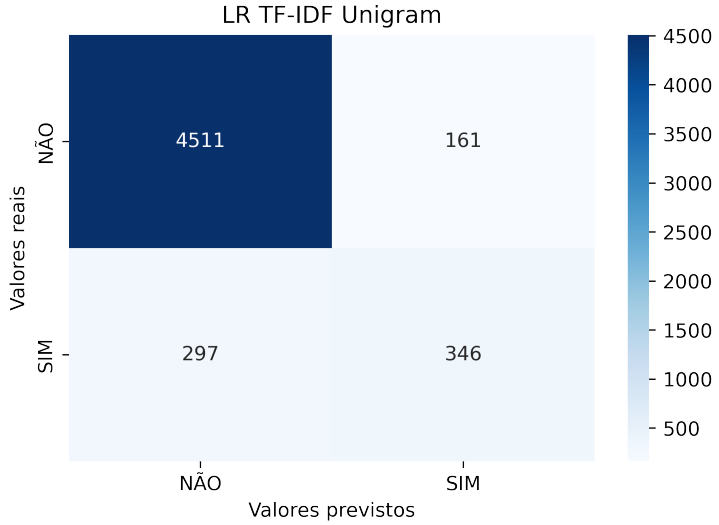


Figura 18 – Matriz de confusão do algoritmo Regressão Logística usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

5.3.3 Desempenho do Algoritmo LightGBM

Os resultados de desempenho para o algoritmo LightGBM são apresentados a seguir. A Tabela 19 contém as métricas de desempenho relacionadas à abordagem *TF - Unigram*, enquanto a Tabela 20 exibe as métricas para a abordagem *TF-IDF Unigram*. Adicionalmente, as Figuras 19 e 20 ilustram as matrizes de confusão correspondentes a cada abordagem.

Tabela 19 – Métricas de desempenho do algoritmo de classificação LightGBM usando *TF - Unigram* (Fonte: O autor (2023)).

			Acurácia	*0,94*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.639*	*0,97*	*0,96*	*0,96*
Relevantes	*676*	*0,74*	*0,80*	*0,77*
Total de posts	***5.315***			

Os resultados demonstraram que o algoritmo LightGBM apresentou um bom desempenho, caracterizado por um baixo número de falsos positivos e uma alta taxa de acerto.

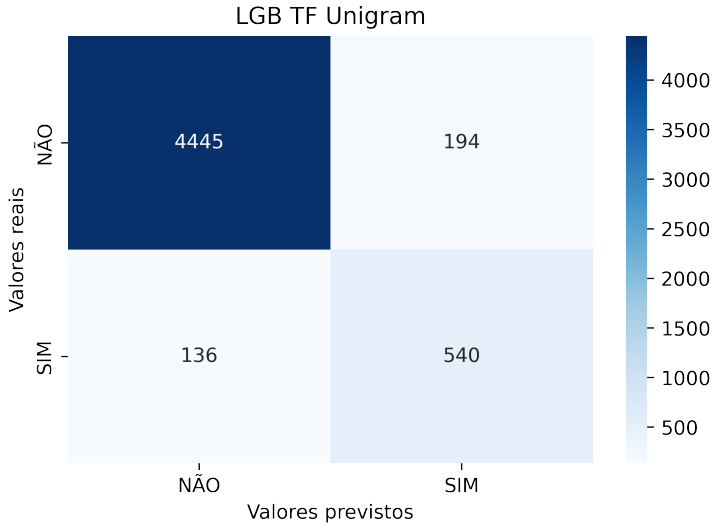


Figura 19 – Matriz de confusão do algoritmo LightGBM usando *TF - Unigram* (Fonte: O autor (2023)).

Tabela 20 – Métricas de desempenho do algoritmo de classificação LightGBM usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

		Acurácia		*0,94*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.657*	*0,98*	*0,96*	*0,97*
Relevantes	*658*	*0,74*	*0,83*	*0,79*
Total de posts	***5.315***			

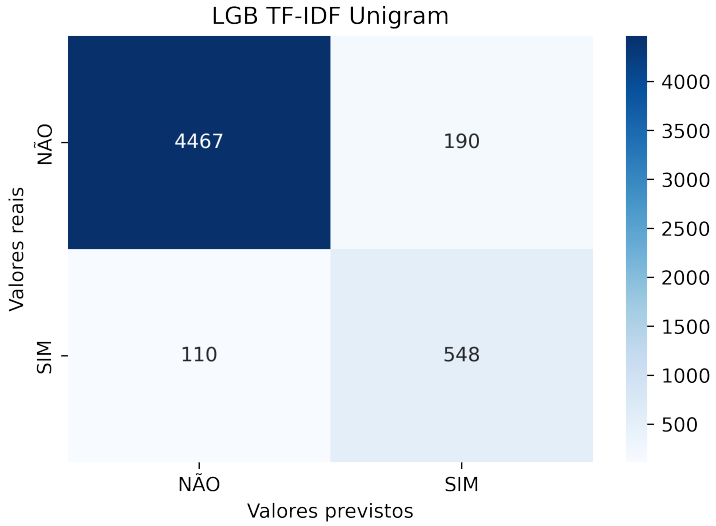


Figura 20 – Matriz de confusão do algoritmo LightGBM usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

Esse desempenho foi particularmente notável ao utilizar a abordagem *TF-IDF - Unigram*, atingindo uma taxa de acerto de 83% na classe dos *posts Relevantes*.

5.3.4 Desempenho do Algoritmo XGBoost

Os resultados de desempenho do algoritmo XGBoost estão apresentados abaixo. A Tabela 21 contém as métricas de desempenho associadas à abordagem *TF - Unigram*, enquanto a Tabela 22 exibe as métricas para a abordagem *TF-IDF Unigram*. Adicionalmente, as Figuras 21 e 22 ilustram as matrizes de confusão correspondentes a cada abordagem.

Tabela 21 – Métricas de desempenho do algoritmo de classificação XGBoost usando *TF - Unigram* (Fonte: O autor (2023)).

			Acurácia	*0,93*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.639*	*0,95*	*0,97*	*0,96*
Relevantes	*676*	*0,77*	*0,67*	*0,72*
Total de posts	***5.315***			

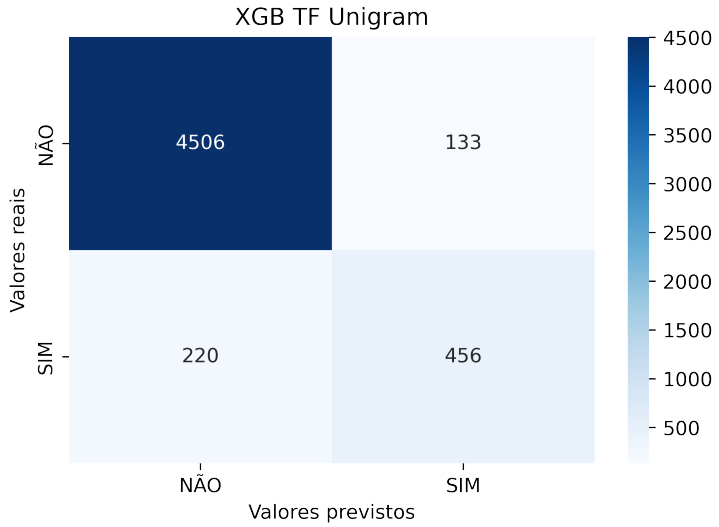


Figura 21 – Matriz de confusão do algoritmo XGBoost usando *TF - Unigram* (Fonte: O autor (2023)).

Tabela 22 – Métricas de desempenho do algoritmo de classificação XGBoost usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

			Acurácia	*0,93*
Amostra		***Precisão***	***Revocação***	***Medida F***
Não Relevantes	*4.657*	*0,96*	*0,97*	*0,96*
Relevantes	*658*	*0,77*	*0,68*	*0,72*
Total de posts	***5.315***			

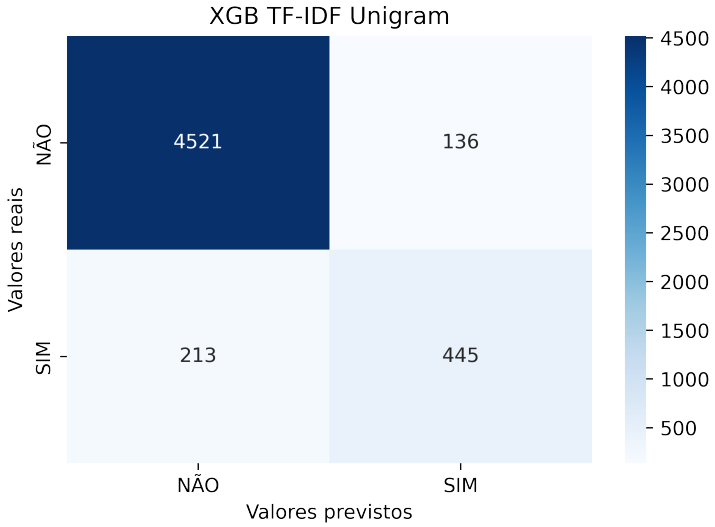


Figura 22 – Matriz de confusão do algoritmo XGBoost usando *TF-IDF - Unigram* (Fonte: O autor (2023)).

Como pode ser observado, os resultados obtidos pelo algoritmo XGBoost foram consideráveis, superando o desempenho dos algoritmos SVM e *Regressão Logística*. No entanto, eles não alcançaram os excelentes resultados alcançados pelo LightGBM.

5.3.5 Conclusão do Teste dos Modelos de Classificação

Como mencionado no início desta seção, os testes nesta etapa tinham como objetivo identificar o melhor modelo entre os selecionados na primeira fase de testes, conforme detalhado na Seção 5.2. Os resultados confirmaram as tendências observadas nos primeiros testes. A combinação do algoritmo LightGBM com a abordagem de representação de texto *TF-IDF - Unigram* superou os demais modelos em todas as métricas avaliadas. Notavelmente, a revocação superou 83%, com o modelo corretamente classificando 548 das 658 amostras apresentadas.

Além disso, é importante observar que, em todos os algoritmos testados, não foram identificadas diferenças significativas entre as abordagens *TF - Unigram* e *TF-IDF - Unigram*. Outra vantagem notável do modelo LightGBM foi o tempo de treinamento, que, embora não tenha sido cronometrado, foi significativamente menor do que o dos demais modelos.

Por fim, é importante destacar que, devido à falta de diferenças significativas, os resultados relativos às diferentes proporções de divisão de dados entre treino e teste não foram apresentados. Essa consistência no desempenho sugere que o modelo pode ser robusto e capaz de generalizar bem, independentemente da proporção de divisão de dados adotada.

5.4 Teste do Modelo de Identificação de *posts* Relevantes em Novos Dados não Rotulados

O modelo de identificação de *posts* relevantes, implementado por meio do algoritmo LightGBM e utilizando a representação de texto TF-IDF - Unigram, foi testado em novos dados não rotulados, o *CONJUNTO DE DADOS III*. Conforme descrito na Subseção 4.2.3, para cada *post* analisado, o algoritmo atribui uma probabilidade entre 0 e 1 quanto à sua relevância. Durante a fase de testes, foram aplicadas duas formas de classificação: a primeira, classificando-os como *Relevante* ou *Não Relevante*, e a segunda, considerando faixas de relevância *(Baixa, Média* ou *Alta)*.

Dos 7.498 *posts* presentes no conjunto de dados, um total de 1.158 (cerca de 15%) foram identificados pelo modelo como *Relevantes*, ou seja, mensagens potencialmente maliciosas, enquanto 6.340 (aproximadamente 85%) foram marcados como *Não Relevantes*. O gráfico da Figura 23 ilustra o resultado dessa classificação.

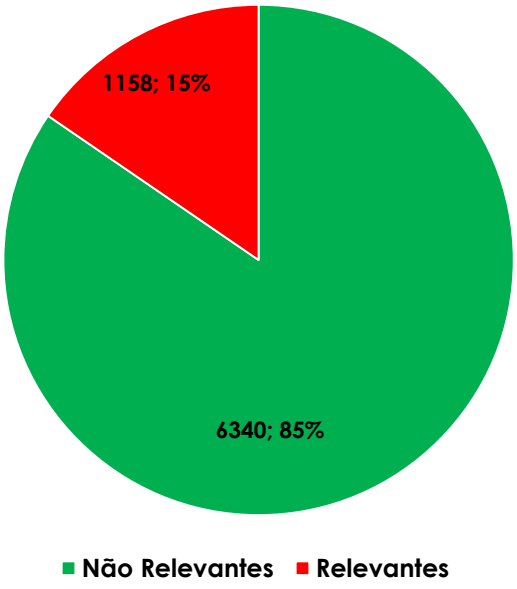


Figura 23 – *Posts* do *CONJUNTO DE DADOS III* classificados pelo modelo como *Relevantes* ou *Não Relevantes* (Fonte: O autor (2023)).

Uma análise comparativa entre o *CONJUNTO DE DADOS II* previamente rotulado, e o resultado da classificação feita pelo modelo no *CONJUNTO DE DADOS III* revela uma proximidade dos valores. No *CONJUNTO DE DADOS II* 13% dos *posts* estão marcados como *Relevantes*, enquanto o modelo de classificação identificou 15% dos *posts* no *CONJUNTO DE DADOS III* como *Relevantes*.

Na classificação por faixa de relevância, dos 7.498 *posts* no conjunto de dados, a distribuição foi a seguinte: 595 *posts* (cerca de 8%) foram identificados como *Relevância Alta*, 849 (cerca de 11%) como *Relevância Média* e 6.054 (aproximadamente 81%) como

Baixa Relevância. O gráfico na Figura 24 ilustra o resultado dessa classificação, a qual leva em consideração a probabilidade de maliciosidade dos *posts*.

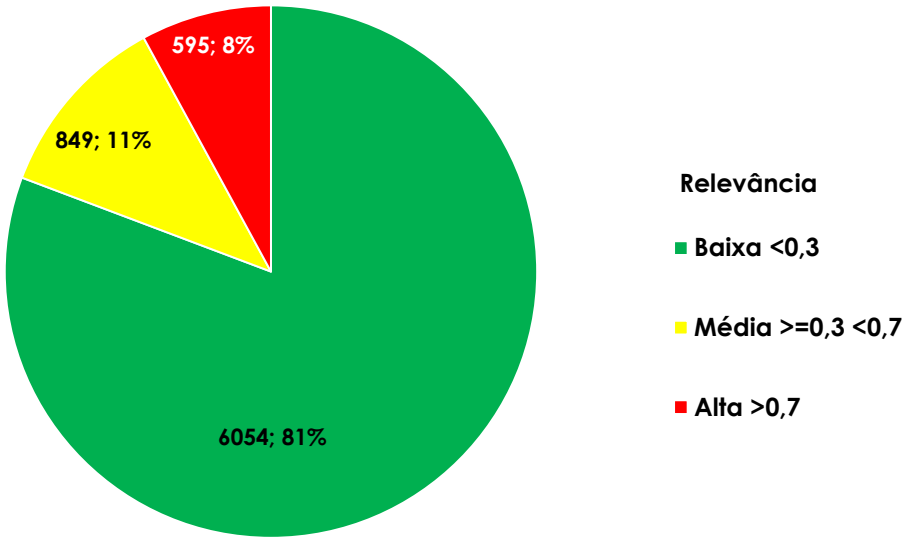


Figura 24 – *Posts* do *CONJUNTO DE DADOS III* classificados por faixa de relevância (Fonte: O autor (2023)).

Esse resultado indica que a maioria significativa dos *posts* (81%) não contém conteúdo relevante no que se refere à segurança da informação. Isso mantém uma coerência com a base de dados rotulada, onde a maioria dos *posts* também é irrelevante nesse contexto.

5.5 Análise da Modelagem de Tópicos LDA

Conforme descrito na Subseção 4.2.4, uma visão geral do desempenho do modelo de identificação de *Posts Relevantes* foi obtido gerando tópicos por meio da técnica LDA em ambos os conjuntos de dados: o *CONJUNTO DE DADOS II* (conjunto de dados rotulado) e o *CONJUNTO DE DADOS III*, que foi classificado pelo modelo. As Tabelas 23 e 24 apresentam os 20 tópicos gerados a partir de todos os *posts* desses conjuntos de dados.

Os 20 tópicos dos conjuntos de dados oferecem uma visão abrangente do conteúdo presente neles. As principais palavras de cada tópico indicam uma ampla variedade de assuntos discutidos nos fóruns da *Dark Web*. É possível identificar semelhanças entre os conjuntos de dados, como por exemplo o tópico 3 da Tabela 23 e o tópico 2 da Tabela 24, ambos contendo palavras relacionadas à política. Além disso, o tópico 8 da Tabela 23 guarda semelhanças com o tópico 19 da Tabela 24, uma vez que ambas apresentam palavras relacionadas a vazamento de dados.

Tabela 23 – 20 Tópicos de todos os *posts* do *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	filme, filmes, assistir, serie, comando, assisti, cachorro, votos, preto, gato
2	celular, rede, internet, redes, numero, app, possivel, codigo, whatsapp, facebook
3	brasil, pais, governo, guerra, estado, povo, paises, mundo, eua, poder
4	vida, tempo, melhor, sempre, ficar, pessoa, quero, pessoas, boa, anos
5	dinheiro, comprar, cartao, ganhar, grana, vender, bitcoin, conta, reais, compra
6	dia, dormir, dias, ficar, corpo, vez, tempo, horas, pornografia, noite
7	arte, persuasao, proxy, tag, buscadores, servidores, lista, reddit, tecnicas, postagem
8	dados, nome, conta, pessoa, email, informacoes, telegram, conseguir, saber, cpf
9	livro, ler, historia, parte, russo, grande, lendo, ordem, primeiro, maconaria
10	tor, linux, sistema, arquivo, uso, windows, arquivos, baixar, maquina, vpn
11	site, link, links, web, deep, sites, surface, forum, achar, tipo
12	pessoas, deus, mundo, existe, acredito, vida, apenas, forma, porque, verdade
13	curso, aprender, cursos, programacao, hacking, estudar, livros, ingles, python, conhecimento
14	mulher, mano, homem, mina, mulheres, p****, p**, sexo, fica, anos
15	casa, drogas, maconha, arma, cidade, onde, agua, comprar, armas, carro
16	brodie, deg, sonho, sonhos, cabelo, pac, faw, top, sus, downvote
17	nunca, sempre, dia, mim, tempo, falar, anos, tava, uns, hoje
18	forum, pergunta, perguntas, pontos, respostas, conta, agora, usuarios, resposta, tempo
19	pessoas, pessoa, porque, ninguem, merda, tipo, gente, faz, caso, qualquer
20	hoje, anos, dia, vacina, estados, muitos, causa, virus, saude, outros

Após a geração dos 20 tópicos para cada conjunto de dados, foi realizado um segundo agrupamento com 10 tópicos para cada conjunto, como apresentado nas Tabelas 25 e 26.

Percebe-se que, mesmo com um número menor de tópicos, ainda é possível ter uma visão geral dos assuntos discutidos, e as semelhanças entre os conjuntos de dados também são visíveis, o que era esperado, uma vez que ambos os conjuntos são provenientes dos mesmos fóruns, porém de períodos diferentes (Seção 5.1).

Até este ponto, a modelagem de tópicos usando o LDA revelou tópicos gerais discutidos nos fóruns e destacou semelhanças entre os conjuntos de dados analisados. Agora, para avaliar a capacidade do modelo na identificação de *Posts Relevantes*, tópicos foram gerados em ambos os conjuntos de dados, dividindo-os de acordo com seus rótulos. Nessa análise, considera-se que *posts* com probabilidade inferior a 0,5 são classificados como Não Relevantes, enquanto aqueles com probabilidade igual ou superior a 0,5 são considerados Relevantes.

As Tabelas 27 e 28 apresentam os 10 tópicos gerados nos conjuntos de dados conside-

Tabela 24 – 20 Tópicos de todos os *posts* do *CONJUNTO DE DADOS III* (Fonte: O autor (2023)).

Tópico	Principais Palavras
1	informacoes, minimo, caracteres, pontos, acrescente, livro, ler, desse, pergunta, forum
2	brasil, governo, pais, povo, porque, bolsonaro, brasileiro, lula, desde, presidente
3	obrigado, mano, resposta, talvez, sempre, tempo, pensamentos, forma, boa, horas
4	jogo, vlw, conta, forum, jogar, man, agora, fiz, apenas, nunca
5	dia, tempo, ficar, passar, merda, mano, anos, noite, queria, melhor
6	estado, forma, pessoas, talvez, problemas, disse, amor, caso, melhor, apenas
7	dinheiro, ganhar, bitcoin, lula, governo, melhor, certo, hoje, economia, mercado
8	tor, rede, linux, sistema, windows, seguranca, uso, aprender, maquina, acesso
9	deus, jesus, biblia, logica, romano, faz, igreja, verdade, pro, carregar
10	pena, chan, todo, igual, achei, porque, f***, tao, melhor, epoca
11	dinheiro, faz, quero, trabalho, boa, pessoas, ano, pagar, grana, vida
12	email, nome, filme, dados, anime, mail, pessoa, endereco, gosto, assistir
13	vida, verdade, pessoas, todos, tipo, sabe, acredito, saber, nunca, apenas
14	corpo, gosto, casa, comer, dia, disso, musica, iria, acido, dar
15	vida, pessoa, pessoas, tempo, mim, sempre, faz, vezes, gente, melhor
16	pessoas, mundo, guerra, poder, contra, anos, vao, grande, porque, parte
17	web, forum, deep, surface, dark, conteudo, lembro, video, gore, link
18	porque, forma, vida, pessoas, mundo, geracao, homem, arma, fato, todos
19	site, telegram, link, sites, dados, conta, comprar, links, onde, saber
20	tbm, imperio, preso, derrubar, melhor, site, outro, cada, open, todo

Tabela 25 – 10 Tópicos de todos os *posts* do *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

Tópico	Principais Palavras
1	jogo, gosto, musica, filme, jogos, arte, filmes, assistir, musicas, serie
2	celular, senha, numero, conta, rede, pessoa, acesso, social, possivel, dados
3	brasil, pais, governo, estado, guerra, povo, paises, poder, eua, mundo
4	vida, pessoa, tempo, pessoas, sempre, melhor, ficar, dia, mim, apenas
5	dinheiro, ganhar, vale, boa, grana, pena, trabalho, bitcoin, estudar, reais
6	dia, mano, nunca, uns, agora, anos, mulher, tava, vez, ficar
7	forum, pergunta, deep, pessoas, perguntas, web, site, porque, tipo, pontos
8	comprar, dados, cartao, nome, telegram, onde, site, conta, facil, sabe
9	pessoas, deus, mundo, existe, porque, acredito, apenas, forma, verdade, pois
10	site, link, links, tor, curso, sites, linux, cursos, aprender, google

rando apenas os *posts Não Relevantes*. Por outro lado, as Tabelas 29 e 30 exibem os 10 tópicos gerados considerando apenas os *posts Relevantes*.

A análise dos tópicos gerados a partir dos *posts Não Relevantes* novamente revela que

Tabela 26 – 10 Tópicos de todos os *posts* do *CONJUNTO DE DADOS III* (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	minimo, informacoes, pontos, caracteres, acrescente, forum, vida, dia, ficar, faz
2	brasil, pais, estado, governo, porque, mundo, guerra, povo, paises, contra
3	pessoas, vida, forma, deus, apenas, todos, verdade, fato, mundo, existe
4	site, conta, dados, link, forum, telegram, sites, nome, links, web
5	tempo, dia, vida, melhor, casa, ficar, corpo, dar, menos, faz
6	pessoas, pessoa, vida, tipo, porque, sempre, mim, gente, talvez, mulher
7	dinheiro, comprar, melhor, grana, faz, valor, ganhar, vender, pagar, certo
8	rede, tor, linux, sistema, windows, uso, curso, aprender, seguranca, maquina
9	faz, deus, video, entender, problema, twitter, romano, logica, pois, verdade
10	gosto, f***, tbm, filme, falar, musica, anime, gente, todo, melhor

Tabela 27 – 10 Tópicos dos *posts Não Relevantes* do *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	pessoas, vida, pessoa, porque, mundo, apenas, forma, outros, todos, tipo
2	deus, acredito, mundo, pessoas, existe, verdade, tempo, apenas, pois, nunca
3	forum, pergunta, perguntas, conta, respostas, tempo, pontos, agora, usuarios, porque
4	mano, anos, nunca, mulher, mae, tava, casa, falar, uns, amigo
5	cartao, agua, numero, faca, metodo, ingles, possivel, use, corpo, precisa
6	dinheiro, comprar, ganhar, grana, vender, boa, facil, conta, compra, mercado
7	musica, jogo, estudar, aprender, gosto, jogos, filme, melhor, escola, ler
8	brasil, pais, governo, guerra, estado, povo, paises, poder, contra, eua
9	site, link, links, sites, web, deep, surface, tor, curso, google
10	vida, tempo, dia, sempre, ficar, melhor, vezes, nunca, mim, sinto

Tabela 28 – 10 Tópicos dos *posts Não Relevantes* do *CONJUNTO DE DADOS III* (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	vida, melhor, pessoas, sinto, pessoa, deve, faz, tempo, gosto, f***
2	site, web, forum, deep, mano, surface, link, rede, tor, links
3	dinheiro, brasil, mundo, lula, hoje, governo, sempre, sendo, fez, errado
4	deus, jesus, verdade, biblia, porque, pessoas, logica, faz, escravidao, romano
5	pessoas, tempo, forma, talvez, vida, pessoa, dia, corpo, melhor, mim
6	obrigado, mano, vida, sempre, tipo, mal, porque, boa, sabe, vezes
7	estado, pais, governo, povo, pois, gente, porque, paises, dizer, brasil
8	gosto, anime, falar, porque, livro, apenas, conta, internet, pois, sempre
9	linux, windows, faz, obrigado, porque, todo, brodie, gostei, pois, fico
10	dinheiro, melhor, faz, comprar, site, sabe, dar, saber, conta, boa

Tabela 29 – 10 Tópicos dos *posts Relevantes* do *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	arquivo, arquivos, pontos, malware, criptografia, programa, chave, hash, senha, ransomware
2	pessoas, vida, mundo, porque, todo, grupo, todos, brasil, apenas, hacker
3	site, hacker, forum, tempo, hacking, pessoas, tipo, web, boa, sempre
4	rede, site, ataque, pagina, servidor, phishing, criar, vitima, ferramenta, social
5	curso, linux, programacao, hacking, aprender, python, linguagem, hacker, cursos, estudar
6	link, site, links, dados, sites, forum, cursos, senha, google, senhas
7	pessoa, numero, cpf, nome, dados, telegram, celular, social, engenharia, telefone
8	conta, senha, facebook, email, trojan, virus, escola, hackear, recuperar, entrar
9	tor, vpn, linux, celular, sistema, acesso, rede, maquina, internet, windows
10	dados, cartao, conta, dinheiro, comprar, informacoes, pessoa, banco, nome, onde

Tabela 30 – 10 Tópicos dos *posts Relevantes* do *CONJUNTO DE DADOS III* (Fonte: O autor (2023)).

Tópico	**Principais Palavras**
1	forum, hacking, gosto, boa, tecnologia, massa, gente, sentido, comunidade, muita
2	windows, site, virus, arquivos, tipo, agora, derrubar, fato, queria, exemplo
3	dados, agora, porque, conta, hackear, forma, melhor, grupo, sistema, possivel
4	faz, pergunta, dar, melhor, conta, passar, disse, nota, todo, valeu
5	senha, rede, site, tor, vpn, caso, conta, acesso, senhas, sistema
6	dados, conta, informacoes, dinheiro, posso, boa, pessoas, saber, pessoa, comprar
7	linux, kali, maquina, uso, usando, conta, gente, twitter, realmente, forma
8	hacking, dados, curso, obrigado, aprender, tipo, links, programacao, cursos, quero
9	dados, cpf, numero, pessoa, telegram, cartao, site, nome, telefone, puxar
10	email, dados, nome, tipo, acesso, tor, site, qualquer, informacoes, onde

os assuntos tratados estão relacionados a diversos temas semelhantes entre os conjuntos de dados. Por exemplo, a política é um tema presente no tópico 8 da Tabela 27 e no tópico 3 da Tabela 28. Além disso, há tópicos relacionados a compras, como o tópico 6 da Tabela 27 e o tópico 10 da Tabela 28. Entretanto, nenhum dos tópicos faz referência a temas relacionados à ameaças cibernéticas, o que é esperado, já que esses temas devem aparecer na classe dos *posts Relevantes*.

Já praticamente todos os tópicos gerados a partir dos *posts Relevantes* contêm palavras

diretamente relacionadas a ameaças cibernéticas, com alguns destaques sugerindo vazamento de dados. Por exemplo, o tópico 7 da Tabela 29 e o tópico 9 da Tabela 30 revelam palavras relacionadas ao tema. Novamente é importante ressaltar que há muitas semelhanças entre os tópicos de cada conjunto de dados. Isso reforça que o modelo foi capaz de adquirir um bom nível de aprendizado durante a fase de treinamento, permitindo-lhe classificar com precisão os novos *posts* que lhe são apresentados.

A Tabela 31 exibe quatro *posts* que foram identificados como altamente relevantes pelo modelo de classificação no CONJUNTO DE DADOS III. Conforme pode ser observado, os *posts* de ID 899 e 1048 tratam de vazamento de dados de pessoas e empresas, enquanto os *posts* de ID 1010 e 6632 abordam vulnerabilidades de software. Esses resultados indicam que o modelo desenvolvido identificou *posts* contendo informações potencialmente úteis para CTI.

5.6 Análise das Palavras mais Frequentes de Cada Classe

A modelagem de tópicos revelou os grupos de palavras dominantes em cada tópico gerado. No entanto, para uma compreensão mais aprofundada, realizou-se uma análise adicional com o propósito de visualizar a frequência absoluta das palavras mais comuns. Essa análise contribuiu para verificar as semelhanças entre os dados rotulados e os dados classificados pelo modelo. Essa análise quantitativa foi apresentada em forma de gráficos de barras, que exibem as 100 palavras mais frequentes para cada rótulo.

As Figuras 25 e 26 apresentam os gráficos das 100 palavras mais frequentes nos *posts Não Relevantes* dos *CONJUNTO DE DADOS II* e *III*, enquanto as Figuras 27 e 28 exibem os gráficos das 100 palavras mais frequentes nos *posts Relevantes* de ambos os conjuntos de dados.

A análise das palavras mais frequentes em cada conjunto de dados reforçou o que já havia sido evidenciado no teste de identificação de *posts Relevantes* (Seção 5.4) e na análise da modelagem de tópicos LDA (Seção 5.5). No caso dos novos *posts* do *CONJUNTO DE DADOS III*, a classificação realizada pelo modelo demonstrou uma correspondência consistente com a base de dados rotulada, o *CONJUNTO DE DADOS II*.

Várias palavras são comuns em ambos os conjuntos de dados. Por exemplo, na classe dos *posts Não Relevantes* (Figuras 25 e 26), palavras como *pessoas, vida, tempo* e *porque* figuram no topo da lista em ambos os conjuntos, afastando uma relação direta com ameaças cibernéticas. Já na classe dos *posts Relevantes* (Figuras 27 e 28), palavras como *dados, site* e *conta* estão no topo da lista, sugerindo possível vazamento de dados. Outras palavras, como *senha* e *CPF*, reforçam essa hipótese. Além disso, várias outras palavras mostradas nos gráficos, como *hacker, hacking* e *vírus*, estão diretamente relacionadas a ameaças cibernéticas.

Tabela 31 – Exemplo de *posts* identificados como altamente relevantes no CONJUNTO DE DADOS III

ID	full_text	created_at	probabilidade
899	"record sabem site deep web vazaram dados record pastor enganava fieis obter dinheiro vazamento conhecimento .... vazado sendo vendidos soubesse endereco onion ... hackers ... comecam divulgacao dados sensiveis roubados durante ataque vazamento parece maior avaliacao inicial documentos fotocopia passaporte .... planilhas detalhadas despesas receitas alem correspondencias internas departamento juridico empresa pastor enganava fieis obter dinheiro "...	17/10/2022	0,77194924
1010	"vender vulnerabilidade tres meses achei vulnerabilidade risco medio tiktok ganhei recentemente writeup apple disse vender vulnerabilidades deep web ... falha empresa ... garantir confidencialidade pontas exploracao vulnerabilidade quanto compromete confidencialidade integridade disponibilidade diretamente .... vulnerabilidade poderia invadir contas tiktok poderia invadir servidor database tipo claramente ganharia agora apenas derrubar tiktok"...	23/10/2022	0,94460218
1048	"indica plataforma boa vender dados forma segura possuo dados pessoais privados pessoas empresas vender gostaria encontrar plataforma segura vender mesmos atraves network telegram site google ... interesse dados completos endereco telefone cpf ... sabe acredito pagam pro telegram ... vender dados dados tipo cpf nome data nascimento onde mora tals site"...	24/10/2022	0,861122741
6632	"vulnerabilidade isc bind explorando servidor achei isc bind versao queria tentar explorar vulnerabilidade nele achei exploits exploit queria saber trabalhou servidor trabalhou falhas qualquer ajuda dica vindanone"	24/05/2023	0,820266759

5.7 Comparação com Trabalhos Relacionados

Todos trabalhos relacionados citados nesta dissertação (Capítulo 3), tem algo em comum com esta dissertação no sentido de buscar informações relevantes para CTI em fontes de dados não estruturadas como a *Dark Web* e redes sociais. Porém, vários deles seguem diferentes abordagens que dificultam realizar uma comparação direta com este trabalho.

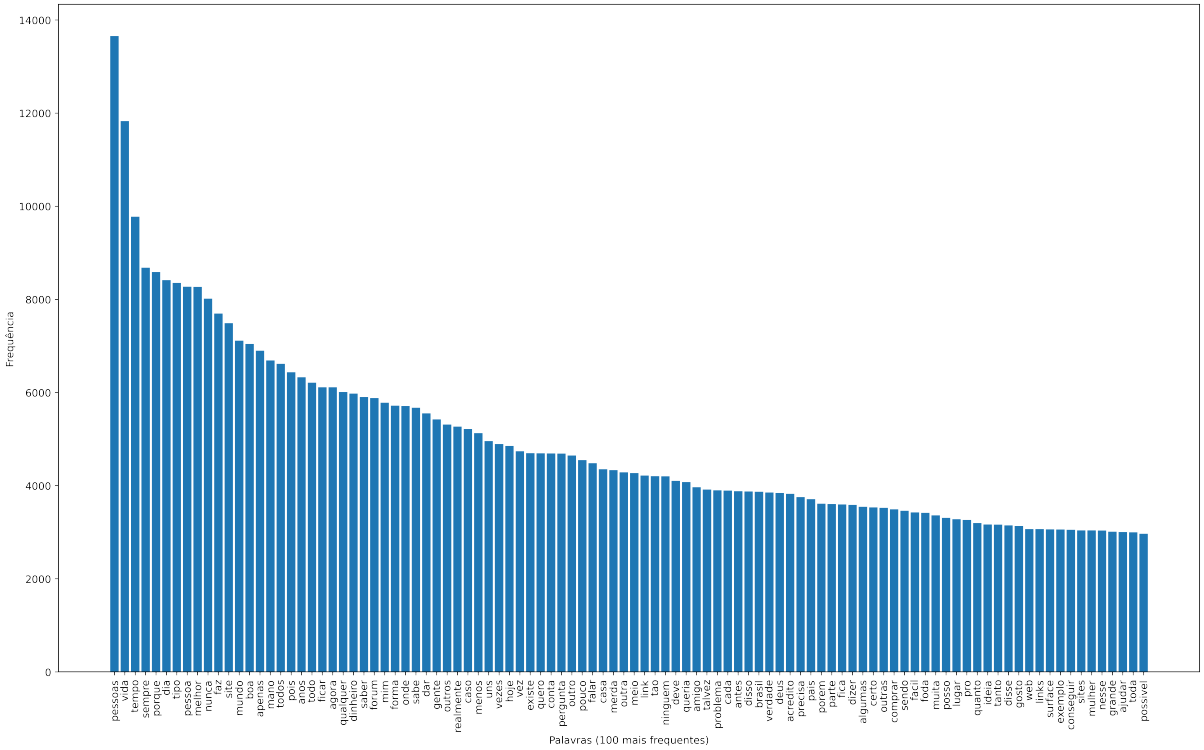


Figura 25 – 100 palavras mais frequentes nos *posts Não Relevantes* do *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

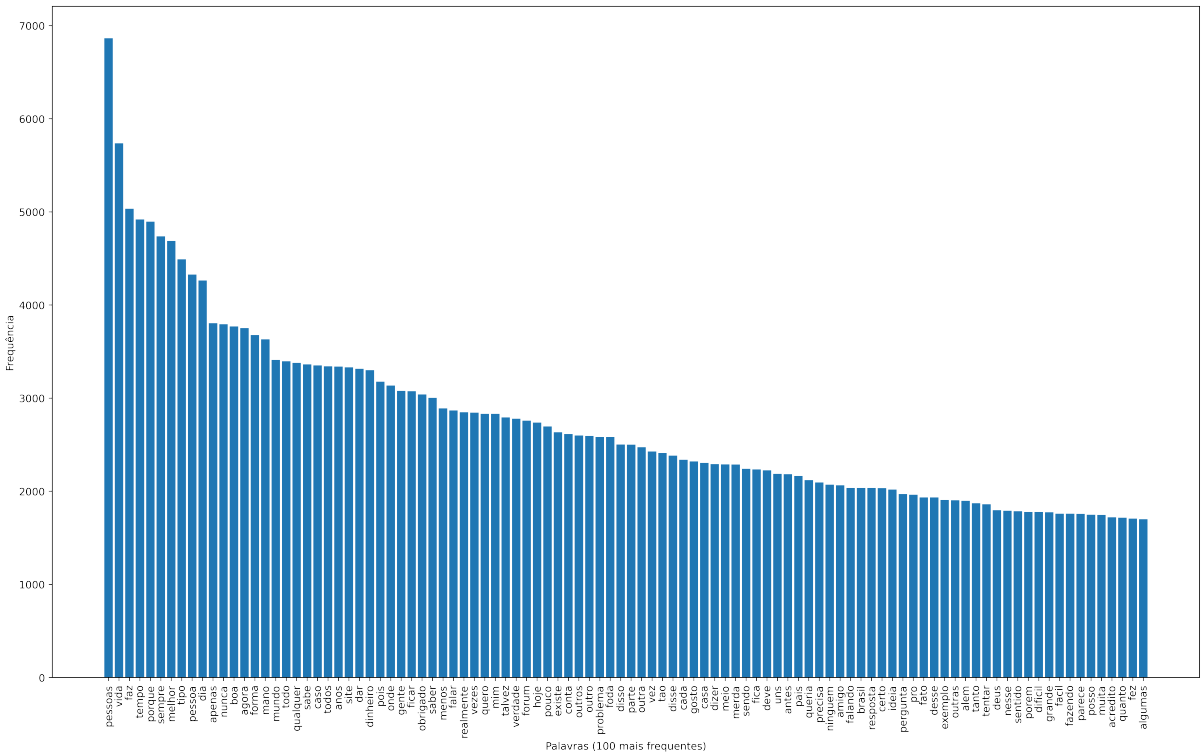


Figura 26 – 100 palavras mais frequentes nos *posts Não Relevantes* do *CONJUNTO DE DADOS III* (Fonte: O autor (2023)).

Três trabalhos apresentaram uma metologia mais aproximada envolvendo informações sobre conjunto de dados, técnicas de vetorização, algoritmos de classificação e resulta-

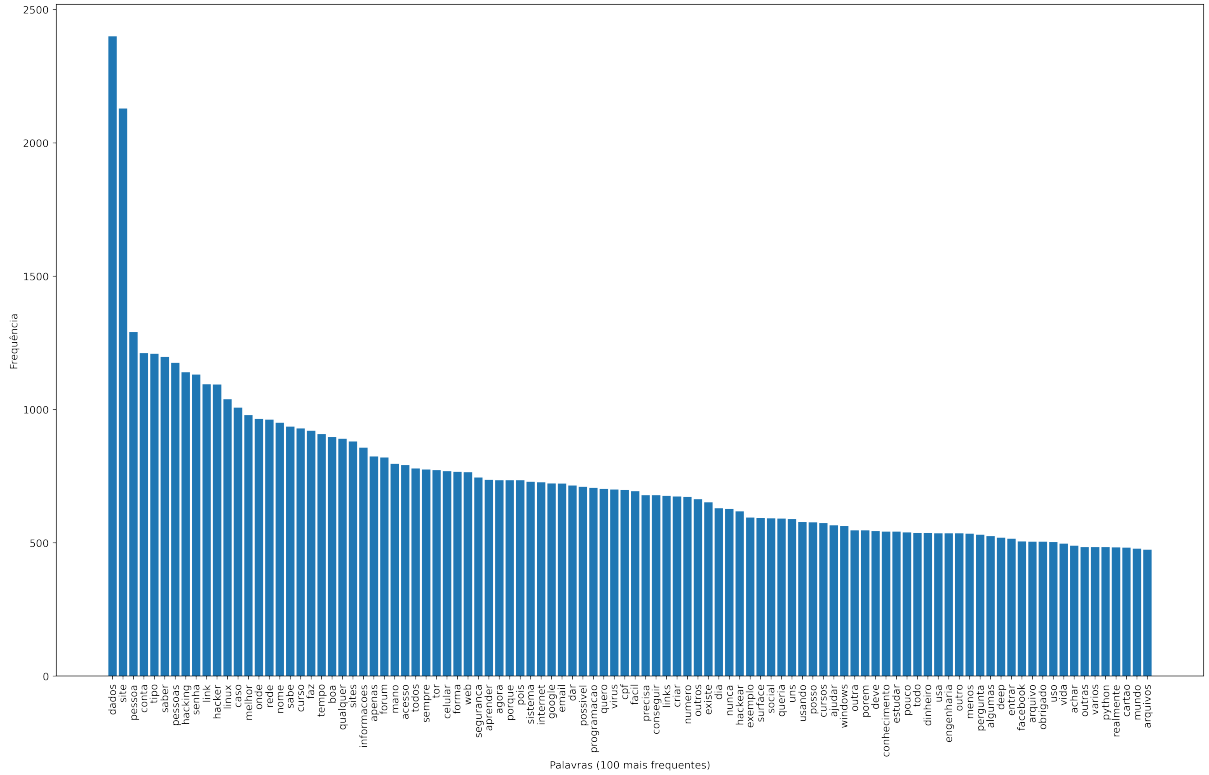


Figura 27 – 100 palavras mais frequentes nos *posts Relevantes* do *CONJUNTO DE DADOS II* (Fonte: O autor (2023)).

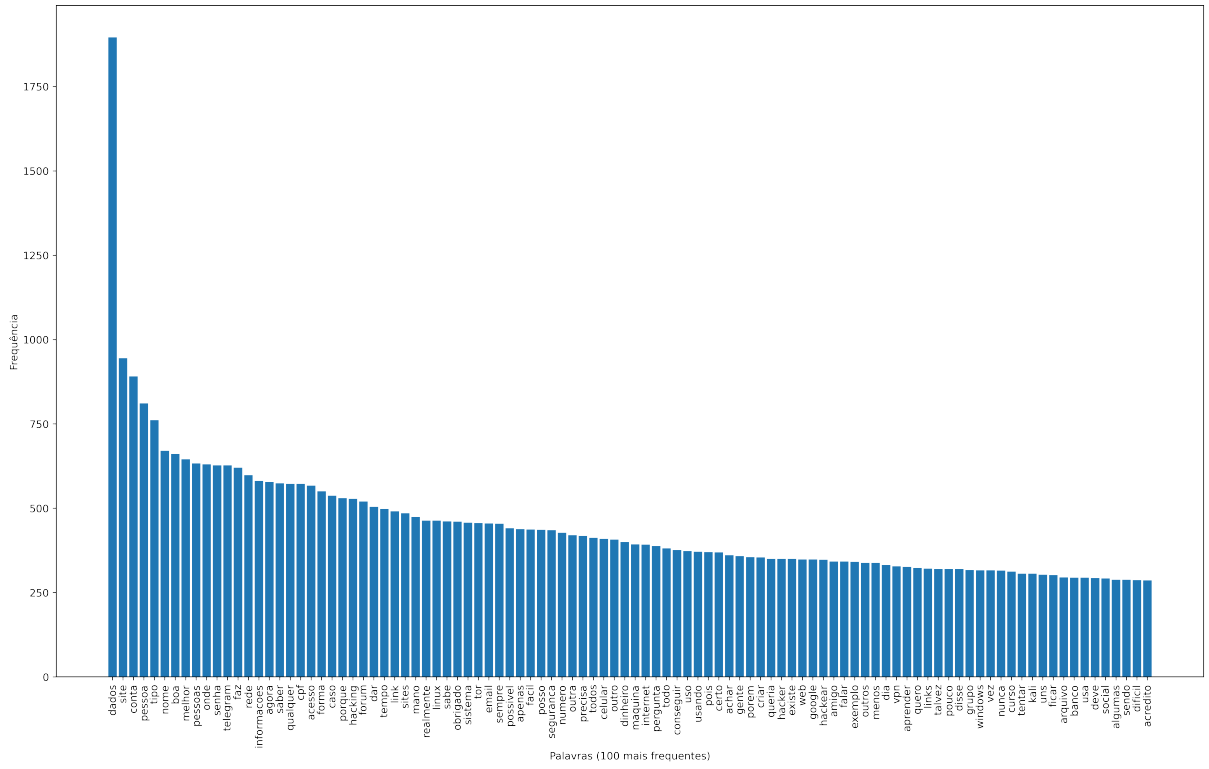


Figura 28 – 100 palavras mais frequentes nos *posts Relevantes* do *CONJUNTO DE DADOS III* (Fonte: O autor (2023)).

dos alcançados, permitindo assim uma comparação mais direta com esta dissertação. A Tabela 32 exibe uma comparação desta dissertação com esses três trabalhos.

Tabela 32 – Comparação com alguns trabalhos relacionados

Trabalho	**Conjunto de Dados**					**Vetorização**	**Algoritmos**	**Objetivo**	**Avaliação**	**Desempenho**	**Observações**
	Fonte	***Rotulagem***	***Amostras***	**Língua**	***Público***						
(DONG et al., 2018)	*Dark Web*	Manual	8.000	EN	Não	TF-IDF	MLP	Identificar novas ameaças	Usou a plataforma de inteligência de ameaças (AlienVault OTX)	94% de acurácia	Altas taxas de falso positivo
(QUEIROZ; MCKEEVER; KEEGAN, 2019)	*Dark Web* / *Surface Web*	Manual	9.470	EN	URL não funciona	Word2Vec/ Glove Sent2vec / InferSent / SentEncoder	SVM / CNN	Aprimorar métodos de classificação usando modelos de incorporação	Usou como base um trabalho desenvolvido anteriormente	96% de acurácia 93% de revocação	Altas taxas de falsos positivos causando baixas taxas de revocação
(KOLOVEAS et al., 2021)	*Dark Web* / *Surface Web*	Ocorrência dos termos segurança e IoT	1.677	EN	Não	Word2Vec	SVM / RF	Identificar informações de CTI	não houve relato	95% de acurácia 61% de precisão 73% de revocação 64% de medida F	Embora o artigo mencione uso de dados da Dark Web e teste de vários algoritmos, no teste relatado foi apresentado apenas dados do Twitter e o uso dois algoritmos CNN e RF
Esta Dissertação	*Dark Web*	Ocorrência de IoCs, palavras-chave contextuais e análise manual	26.575	PT-BR	Sim	TF (Unigram e Bigram) / TF-IDF (Unigram e Bigram) / Word2Vec	SVM / RF / LR / LightGBM / XGBoost	Identificação de posts relevantes para CTI	O modelo foi avaliado usando novos dados não rotulados usando modelagem de tópicos LDA	94% de acuráia 74% de precisão 83% de revocação 79% de medida F	Taxa de acerto para posts da classe relevante acima de 83%

Embora os trabalhos mencionados na tabela compartilhem semelhanças com esta dissertação, seus objetivos apresentam algumas diferenças. O trabalho de Dong et al. (2018), por exemplo, buscou identificar novas ameaças cibernéticas na *Dark Web*. Já Koloveas et al. (2021) focaram na obtenção de informações relevantes para CTI, utilizando dispositivos IoT como base. Por outro lado, Queiroz, Mckeever e Keegan (2019) se aproximou mais do objetivo desta dissertação, que é identificar informações gerais relevantes para CTI nos *posts* analisados.

Conforme demonstrado na Tabela 32, em relação ao conjunto de dados, esta dissertação se destaca pelo maior número de amostras, pelo critério adotado para a rotulagem dos dados e pela disponibilização do conjunto de dados. Quanto à vetorização, esta dissertação testou técnicas de frequência de palavras e uma técnica incorporação de palavras, enquanto os demais trabalhos utilizaram uma ou outra. Vale ressaltar que Queiroz, Mckeever e Keegan (2019), além da incorporação de palavras, também empregou a incorporação de sentenças. Os autores desse trabalho destacam que o objetivo era comparar essas técnicas em relação às técnicas de frequência de palavras adotadas em um trabalho anterior.

No que diz respeito aos algoritmos de aprendizado de máquina, esta dissertação testou um número maior de opções. Embora Koloveas et al. (2021) tenha mencionado no texto o uso de vários outros algoritmos, apenas dois deles foram apresentados nos resultados. Quanto às métricas de desempenho, os valores obtidos nesta dissertação ficaram dentro da média dos demais trabalhos, embora alguns desses trabalhos tenham divulgado apenas algumas métricas.

Para avaliar o modelo de identificação de conteúdo relevante para CTI, Dong et al. (2018), cujo objetivo era identificar novas ameaças, utilizou como base uma plataforma aberta de ameaças. Por sua vez, Queiroz, Mckeever e Keegan (2019) comparou os resultados alcançados usando técnicas de vetorização diferentes das testadas em um trabalho anterior. Já Koloveas et al. (2021) não relatou alguma forma de avaliação do modelo desenvolvido. Nesta dissertação, a avaliação do modelo foi realizada em uma base de dados não rotulada, envolvendo o uso de modelagem de tópicos LDA, além de uma análise de palavras mais frequentes para comparar a semelhança entre o conjunto de dados rotulado e o conjunto de dados classificado pelo modelo.

Capítulo 6

Conclusão

Diante das diversas abordagens disponíveis na literatura em busca de novas técnicas e fontes de dados para a obtenção de CTI, este trabalho se propôs a explorar fóruns da *Dark Web*. Assim, através do uso de técnicas de mineração de texto e aprendizado de máquina supervisionado desenvolver um modelo computacional capaz de identificar *posts* relevantes que auxiliem a comunidade de segurança cibernética na detecção de ameaças, vulnerabilidades, vazamento de dados e ataques cibernéticos.

Inicialmente, foram analisados 26.575 *posts* coletados de dois fóruns da *Dark Web*: o *Hidden Answers* e o *Deep Answers*. Esses *posts* foram rotulados, resultando em dois conjuntos de dados: o *CONJUNTO DE DADOS I*, onde o critério adotado foi a ocorrência simultânea de IoCs e palavras-chave contextuais, e o *CONJUNTO DE DADOS II*, que, além da ocorrência de IoCs e palavras-chave, envolveu uma análise manual considerando fatores como a categoria do *post*. Foram testadas cinco representações de texto diferentes: *TF - Unigram, TF - Bigram, TF-IDF - Unigram, TF-IDF - Bigram* e *Word2Vec*, com cinco algoritmos de aprendizado de máquina distintos: SVM, *Random Forest, Logistic Regression*, LightGBM e XGBoost. Usando o *CONJUNTO DE DADOS I*, foram selecionadas as melhores combinações de algoritmo de aprendizado e representação de texto, as quais foram empregadas para treinar o modelo de classificação com o *CONJUNTO DE DADOS II*. Após os testes, o melhor desempenho foi alcançado com o algoritmo LightGBM e *TF-IDF - Unigram*, atingindo 94% de acurácia, 74% de precisão, 79% de medida F e uma taxa de revocação de 83%, representando a classificação correta de 548 das 658 amostras apresentadas.

Posteriormente, foram analisados 7.498 novos *posts* coletados dos mesmos fóruns, mas em um período de postagem diferente daqueles dos *posts* anteriormente analisados. Esses novos *posts* deram origem ao *CONJUNTO DE DADOS III*, o qual, propositalmente, não foi rotulado. A ideia nesta etapa foi apresentar ao modelo de classificação *posts* nunca antes analisados por ele, a fim de verificar o seu comportamento.

A classificação dos novos *posts*, realizada pelo algoritmo, revelou proximidade com a base de dados rotulada. Na base rotulada, cerca de 13% dos *posts* estão marcados como

relevantes, enquanto nos novos *posts* foi identificado um percentual de aproximadamente 15%. Além da classificação binária (relevante e não relevante), a categorização por faixa de relevância (baixa, média e alta), calculada com base na probabilidade atribuída a cada *post* pelo algoritmo de classificação, também confirmou essa proximidade, sendo que a maioria dos *posts* foi enquadrada na categoria de baixa relevância.

Uma análise utilizando a modelagem de tópicos LDA revelou uma semelhança entre os tópicos gerados nos conjuntos de dados II e III. Para a classe dos *posts* não relevantes, as principais palavras nos tópicos de ambos os conjuntos de dados não fazem referência ao contexto de segurança cibernética. No entanto, para a classe dos *posts* relevantes, as principais palavras estão diretamente relacionadas a ameaças cibernéticas e vazamentos de dados. Isso reforça a capacidade do modelo de adquirir um bom nível de aprendizado, a ponto de classificar com precisão os novos *posts* apresentados a ele.

Por fim, realizou-se uma análise das palavras mais frequentes em ambos os conjuntos de dados, contabilizando as 100 palavras mais recorrentes. Essa análise reforçou o que já havia sido evidenciado nas análises anteriores quanto à precisão do modelo de classificação de *posts*. Observou-se a presença de muitas palavras comuns nas classes dos *posts* relevantes e não relevantes, sendo que as palavras da classe dos *posts* relevantes estão diretamente relacionadas a ameaças cibernéticas ou sugerem vazamento de dados.

6.1 Principais Contribuições

As principais contribuições deste trabalho são listadas a seguir:

1. Desenvolvimento de uma ferramenta para extração de IoCs em fontes de dados não estruturados, como a *Dark Web*;
2. Construção de um conjunto de dados rotulado para o treinamento de modelos de aprendizado de máquina supervisionado no contexto de segurança cibernética;
3. Desenvolvimento de um modelo de classificação de *posts* relevantes para a comunidade de segurança cibernética na obtenção de CTI;
4. Avaliação do desempenho do modelo de classificação de *posts* relevantes em conjunto de dados não rotulado usando modelagem de tópicos e análise de frequência de palavras.

6.2 Trabalhos Futuros

Como trabalhos futuros os itens a seguir podem ser objetos de estudo:

- ❑ Investigar o conteúdo dos *posts* relevantes para identificar os diferentes tipos de ameaças presentes na *Dark Web*.

- ❏ Incluir outras fontes de dados além da *Dark Web* como canais do *Telegram*.
- ❏ Avaliar o desempenho do modelo em outras fontes de dados.
- ❏ Testar outras formas formas de representação de texto como o *BERT*, a fim de verificar se há algum ganho significativo em relação a melhor representação identificada neste trabalho que foi *TF-IDF - Unigram*.
- ❏ Incluir o modelo gerado em um sistema computacional que coleta, armazena, analisa e gera alertas de atividades maliciosas em diferentes fontes de dados.

6.3 Contribuições em Produção Bibliográfica

O artigo intitulado *"Extração e Análise de Indicadores de Comprometimento (IoCs) em Fóruns da Dark Web"* (FILHO; GABRIEL; MIANI, 2023) apresentou uma análise da incidência de IoCs em fóruns da *Dark Web*, abordando a extração e análise de 10 diferentes tipos de IoCs. Este trabalho foi aceito e apresentado durante o *XXIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais* (*SBSeg 2023*), um evento promovido pela Sociedade Brasileira de Computação que ocorreu em setembro de 2023 na Universidade Federal de Juiz de Fora (UFJF) em Juiz de Fora, MG.

Referências

AKHGAR, B. et al. **Dark Web Investigation**. Springer, 2021. Disponível em: <https://doi.org/10.1007/978-3-030-55343-2>.

AL-RAMAHI, M.; ALSMADI, I.; DAVENPORT, J. Exploring hackers assets: topics of interest as indicators of compromise. In: **Proceedings of the 7th Symposium on Hot Topics in the Science of Security**. [s.n.], 2020. p. 1–4. Disponível em: <https://doi.org/10.1145/3384217.3385619>.

ALPAYDIN, E. **Machine learning: the new AI**. [S.l.]: MIT press, 2016.

ALVES, F. et al. Processing tweets for cybersecurity threat awareness. **Information Systems**, Elsevier, v. 95, p. 101586, 2021. Disponível em: <https://doi.org/10.1016/j.is.2020.101586>.

ANANDARAJAN, M.; HILL, C.; NOLAN, T. Text preprocessing. In: _____. **Practical Text Analytics: Maximizing the Value of Text Data**. Cham: Springer International Publishing, 2019. p. 45–59. ISBN 978-3-319-95663-3. Disponível em: <https://doi.org/10.1007/978-3-319-95663-3_4>.

ARNOLD, N. et al. Dark-net ecosystem cyber-threat intelligence (cti) tool. In: IEEE. **2019 IEEE International Conference on Intelligence and Security Informatics (ISI)**. 2019. p. 92–97. Disponível em: <https://doi.org/10.1109/ISI.2019.8823501>.

ASIRI, M. et al. Understanding indicators of compromise against cyber-attacks in industrial control systems: a security perspective. **ACM transactions on cyber-physical systems**, ACM New York, NY, 2023. Disponível em: <https://doi.org/10.1145/3587255>.

ASSAL, H.; CHIASSON, S. 'think secure from the beginning' a survey with software developers. In: **Proceedings of the 2019 CHI conference on human factors in computing systems**. [s.n.], 2019. p. 1–13. Disponível em: <https://doi.org/10.1145/3290605.3300519>.

BASHEER, R.; ALKHATIB, B. Threats from the dark: a review over dark web investigation research for cyber threat intelligence. **Journal of Computer Networks and Communications**, Hindawi Limited, v. 2021, p. 1–21, 2021. Disponível em: <https://doi.org/10.1155/2021/1302999>.

BOJER, C. S.; MELDGAARD, J. P. Kaggle forecasting competitions: An overlooked learning opportunity. **International Journal of Forecasting**, Elsevier, v. 37, n. 2, p. 587–603, 2021. Disponível em: <https://doi.org/10.1016/j.ijforecast.2020.07.007>.

BRADBURY, D. Unveiling the dark web. **Network security**, Elsevier, v. 2014, n. 4, p. 14–17, 2014. Disponível em: <https://doi.org/10.1016/S1353-4858(14)70042-X>.

BROOKS, C. **Alarming cybersecurity stats: What you need to know for 2021**. 2021. <https://www.forbes.com/sites/chuckbrooks/2021/03/02/alarming-cybersecurity-stats-------what-you-need-to-know-for-2021/?sh=6ae7f87158d3>. Acesso em: 20/01/2022.

CABALLERO, J. et al. The rise of goodfatr: A novel accuracy comparison methodology for indicator extraction tools. **Future Generation Computer Systems**, Elsevier, v. 144, p. 74–89, 2023. Disponível em: <https://doi.org/10.1016/j.future.2023.02.012>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [s.n.], 2016. p. 785–794. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

CHEN, T. et al. Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1–4, 2015.

CHOWDHARY, K.; CHOWDHARY, K. Natural language processing. **Fundamentals of artificial intelligence**, Springer, p. 603–649, 2020. Disponível em: <https://doi.org/10.1007/978-81-322-3972-7_19>.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. **Ensemble machine learning: Methods and applications**, Springer, p. 157–175, 2012. Disponível em: <https://doi.org/10.1007/978-1-4419-9326-7_5>.

DANG, S.; AHMAD, P. H. Text mining: Techniques and its application. **International Journal of Engineering & Technology Innovations**, v. 1, n. 4, p. 22–25, 2014.

DELIU, I.; LEICHTER, C.; FRANKE, K. Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation. In: IEEE. **2018 IEEE International Conference on Big Data (Big Data)**. 2018. p. 5008–5013. Disponível em: <https://doi.org/10.1109/BigData.2018.8622469>.

DONG, F. et al. New cyber threat discovery from darknet marketplaces. In: IEEE. **2018 IEEE Conference on Big Data and Analytics (ICBDA)**. 2018. p. 62–67. Disponível em: <https://doi.org/10.1109/ICBDAA.2018.8629658>.

DUTTA, A.; KANT, S. An overview of cyber threat intelligence platform and role of artificial intelligence and machine learning. In: SPRINGER. **Information Systems Security: 16th International Conference, ICISS 2020, Jammu, India, December 16–20, 2020, Proceedings 16**. 2020. p. 81–86. Disponível em: <https://doi.org/10.1007/978-3-030-65610-2_5>.

FAGERLAND, M. W.; HOSMER, D. W. A generalized hosmer–lemeshow goodness-of-fit test for multinomial logistic regression models. **The Stata Journal**, SAGE Publications Sage CA: Los Angeles, CA, v. 12, n. 3, p. 447–453, 2012. Disponível em: <https://doi.org/10.1177/1536867X1201200307>.

FILHO, S. Alves de J.; GABRIEL, P.; MIANI, R. Extração e análise de indicadores de comprometimento (iocs) em fóruns da dark web. In: . [S.l.: s.n.], 2023.

GHAFFARIAN, S. M.; SHAHRIARI, H. R. Software vulnerability analysis and discovery using machine-learning and data-mining techniques: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 50, n. 4, p. 1–36, 2017. Disponível em: <https://doi.org/10.1145/3092566>.

GUGELMIN, F. **RockYou2021**. 2021. <https://canaltech.com.br/seguranca/rockyou2021-84-bilhoes-de-senhas-sao-reveladas-no-maior-vazamento-da-historia-186656/>. Acesso em: 20/01/2022.

GUTTMAN, B.; ROBACK, E. A. **An introduction to computer security: the NIST handbook**. Diane Publishing, 1995. v. 800. Disponível em: <http://dx.doi.org/10.6028/NIST.SP.800-12>.

HAHNE, F. et al. Unsupervised machine learning. **Bioconductor case studies**, Springer, p. 137–157, 2008. Disponível em: <https://doi.org/10.1007/978-0-387-77240-0_10>.

HASTIE, T. et al. **The elements of statistical learning: data mining, inference, and prediction**. Springer, 2009. v. 2. Disponível em: <https://doi.org/10.1007/978-0-387-21606-5>.

HIGHTOWER, F. **IOC Finder**. 2017. Disponível em: <https://github.com/fhightower/ioc-finder>.

JELODAR, H. et al. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. **Multimedia Tools and Applications**, Springer, v. 78, p. 15169–15211, 2019. Disponível em: <https://doi.org/10.1007/s11042-018-6894-4>.

JO, H.; LEE, Y.; SHIN, S. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. **Computers & Security**, Elsevier, v. 120, p. 102763, 2022. Disponível em: <https://doi.org/10.1016/j.cose.2022.102763>.

KAO, A.; POTEET, S. R. **Natural language processing and text mining**. [S.l.]: Springer Science & Business Media, 2007.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017.

KHANDPUR, R. P. et al. Crowdsourcing cybersecurity: Cyber attack detection using social media. In: **Proceedings of the 2017 ACM on Conference on Information and Knowledge Management**. [s.n.], 2017. p. 1049–1057. Disponível em: <https://doi.org/10.1145/3132847.3132866>.

KOLOVEAS, P. et al. intime: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. **Electronics**, MDPI, v. 10, n. 7, p. 818, 2021. Disponível em: <https://doi.org/10.3390/electronics10070818>.

MADOR, Z. Keep the dark web close and your cyber security tighter. **Computer Fraud & Security**, MA Business London, v. 2021, n. 1, p. 6–8, 2021. Disponível em: <https://doi.org/10.1016/S1361-3723(21)00006-3>.

MCTEAR, M. F.; CALLEJAS, Z.; GRIOL, D. **The conversational interface**. Springer, 2016. v. 6. Disponível em: <https://doi.org/10.1007/978-3-319-32967-3>.

MEYER, D.; WIEN, F. Support vector machines. **The Interface to libsvm in package e1071**, v. 28, n. 20, p. 597, 2015.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Disponível em: <https://doi.org/10.48550/arXiv.1301.3781>.

NIAKANLAHIJI, A. et al. Iocminer: Automatic extraction of indicators of compromise from twitter. In: IEEE. **2019 IEEE International Conference on Big Data (Big Data)**. 2019. p. 4747–4754. Disponível em: <https://doi.org/10.1109/BigData47090.2019.9006562>.

NIST. **NVD - General Information**. 2022. Disponível em: <https://nvd.nist.gov/general>.

Olhar Digital. **ANPD abre investigação de vazamento de dados de quase todos os brasileiros**. 2021. <https://olhardigital.com.br/2021/02/04/noticias/anpd-abre-investigacao-de-vazamento-de-dados-de-quase-todos-os-brasileiros/>. Acesso em: 20/01/2022.

PREUVENEERS, D.; JOOSEN, W. Sharing machine learning models as indicators of compromise for cyber threat intelligence. **Journal of Cybersecurity and Privacy**, MDPI, v. 1, n. 1, p. 140–163, 2021. Disponível em: <https://doi.org/10.3390/jcp1010008>.

QAISER, S.; ALI, R. Text mining: use of tf-idf to examine the relevance of words to documents. **International Journal of Computer Applications**, v. 181, n. 1, p. 25–29, 2018. Disponível em: <https://doi.org/10.5120/ijca2018917395>.

QUEIROZ, A. L.; MCKEEVER, S.; KEEGAN, B. Detecting hacker threats: Performance of word and sentence embedding models in identifying hacker communications. In: **AICS**. [S.l.: s.n.], 2019. p. 116–127.

RABII, A. et al. Information and cyber security maturity models: a systematic literature review. **Information & Computer Security**, Emerald Publishing Limited, v. 28, n. 4, p. 627–644, 2020. Disponível em: <https://doi.org/10.1108/ICS-03-2019-0039>.

RODRIGUEZ, A.; OKAMURA, K. Social media data mining for proactive cyber defense. **Journal of Information Processing**, Information Processing Society of Japan, v. 28, p. 230–238, 2020. Disponível em: <https://doi.org/10.2197/ipsjjip.28.230>.

SALEEM, J.; ISLAM, R.; KABIR, M. A. The anonymity of the dark web: A survey. **IEEE Access**, IEEE, v. 10, p. 33628–33660, 2022. Disponível em: <https://doi.org/10.1109/ACCESS.2022.3161547>.

SAMTANI, S.; ZHU, H.; CHEN, H. Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (d-gef). **ACM Transactions on Privacy and Security (TOPS)**, ACM New York, NY, USA, v. 23, n. 4, p. 1–33, 2020. Disponível em: <https://doi.org/10.1145/3409289>.

SAPIENZA, A. et al. Early warnings of cyber threats in online discussions. In: IEEE. **2017 IEEE International Conference on Data Mining Workshops (ICDMW)**. 2017. p. 667–674. Disponível em: <https://doi.org/10.1109/ICDMW.2017.94>.

SARI, A. Context-aware intelligent systems for fog computing environments for cyber-threat intelligence. **Fog Computing: Concepts, Frameworks and Technologies**, Springer, p. 205–225, 2018. Disponível em: <https://doi.org/10.1007/978-3-319-94890-4_10>.

SARKAR, S. et al. Predicting enterprise cyber incidents using social network analysis on dark web hacker forums. **The Cyber Defense Review**, JSTOR, p. 87–102, 2019. Disponível em: <https://www.jstor.org/stable/26846122>.

SAXENA, R.; GAYATHRI, E. Cyber threat intelligence challenges: Leveraging blockchain intelligence with possible solution. **Materials Today: Proceedings**, Elsevier, v. 51, p. 682–689, 2022. Disponível em: <https://doi.org/10.1016/j.matpr.2021.06.204>.

SCELLER, Q. L. et al. Sonar: Automatic detection of cyber security events over the twitter stream. In: **Proceedings of the 12th International Conference on Availability, Reliability and Security**. [s.n.], 2017. p. 1–11. Disponível em: <https://doi.org/10.1145/3098954.3098992>.

SHU, K. et al. Understanding cyber attack behaviors with sentiment information on social media. In: SPRINGER. **Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11**. 2018. p. 377–388. Disponível em: <https://doi.org/10.1007/978-3-319-93372-6_41>.

STALLINGS, W.; BROWN, L. Segurança de computadores. **Princípios e Práticas. Trad.: Arlete Simille Marques. 2ª Ed. Rio de Janeiro: Elsevier Editora**, Elsevier, 2014.

SUBROTO, A.; APRIYANA, A. Cyber risk prediction through social media big data analytics and statistical machine learning. **Journal of Big Data**, Springer, v. 6, n. 1, p. 1–19, 2019. Disponível em: <https://doi.org/10.1186/s40537-019-0216-1>.

SUN, N. et al. Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. **IEEE Communications Surveys & Tutorials**, IEEE, 2023. Disponível em: <https://doi.org/10.1109/COMST.2023.3273282>.

SUSMAGA, R. Confusion matrix visualization. In: SPRINGER. **Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM '04 Conference held in Zakopane, Poland, May 17–20, 2004**. 2004. p. 107–116. Disponível em: <https://doi.org/10.1007/978-3-540-39985-8_12>.

TAHERDOOST, H. Cybersecurity vs. information security. **Procedia Computer Science**, Elsevier, v. 215, p. 483–487, 2022. Disponível em: <https://doi.org/10.1016/j.procs.2022.12.050>.

The MITRE Corporation. **CVE - Frequently Asked Questions**. 2022. Disponível em: <https://cve.mitre.org/about/faqs.html>.

TOUNSI, W. What is cyber threat intelligence and how is it evolving? **Cyber-Vigilance and Digital Trust: Cyber Security in the Era of Cloud Computing and IoT**, Wiley Online Library, p. 1–49, 2019. Disponível em: <https://doi.org/10.1002/9781119618393.ch1>.

ZHANG, P. et al. imcircle: Automatic mining of indicators of compromise from the web. In: IEEE. **2019 IEEE Symposium on Computers and Communications (ISCC)**. 2019. p. 1–6. Disponível em: <https://doi.org/10.1109/ISCC47284.2019.8969570>.

Apêndices

APÊNDICE **A**

Lista das Palavras mais Frequentes dos CONJUNTOS DE DADOS II e III

Esta seção traz as listas ordenadas das palavras mais frequentes de ambas as classes de *posts* dos CONJUNTOS DE DADOS II e III. A Tabela 33 exibe as 100 palavras mais frequentes da classe dos *posts* não relevantes, enquanto a Tabela 34 exibe as 100 palavras mais frequentes da classe dos *posts* relevantes.

Tabela 33 – Lista das 100 palavras mais frequentes da classe dos posts não relevantes dos CONJUNTO DE DADOS II e III

Palavras mais Frequentes	
***Posts* Não Relevantes**	
Conjunto de Dados II	**Conjunto de Dados III**
1 - pessoas, 2 - vida, 3 - tempo, 4 - sempre, 5 - porque, 6 - dia, 7 - tipo, 8 - pessoa, 9 - melhor, 10 - nunca, 11 - faz, 12 - site, 13 - mundo, 14 - boa, 15 - apenas, 16 - mano, 17 - todos, 18 - pois, 19 - anos, 20 - todo, 21 - ficar, 22 - agora, 23 - qualquer, 24 - dinheiro, 25 - saber, 26 - forum, 27 - mim, 28 - forma, 29 - onde, 30 - sabe, 31 - dar, 32 - gente, 33 - outros, 34 - realmente, 35 - caso, 36 - menos, 37 - uns, 38 - vezes, 39 - hoje, 40 - vez, 41 - existe, 42 - quero, 43 - conta, 44 - pergunta, 45 - outro, 46 - pouco, 47 - falar, 48 - casa, 49 - merda, 50 - outra, 51 - meio, 52 - link, 53 - tao, 54 - ninguem, 55 - deve, 56 - queria, 57 - amigo, 58 - talvez, 59 - problema, 60 - cada, 61 - antes, 62 - disso, 63 - brasil, 64 - verdade, 65 - deus, 66 - acredito, 67 - precisa, 68 - pais, 69 - porem, 70 - parte, 71 - fica, 72 - dizer, 73 - algumas, 74 - certo, 75 - outras, 76 - comprar, 77 - sendo, 78 - facil, 79 - f***, 80 - muita, 81 - posso, 82 - lugar, 83 - pro, 84 - quanto, 85 - ideia, 86 - tanto, 87 - disse, 88 - gosto, 89 - web, 90 - links, 91 - surface, 92 - exemplo, 93 - conseguir, 94 - sites, 95 - mulher, 96 - nesse, 97 - grande, 98 - ajudar, 99 - toda, 100 - possivel	1 - pessoas, 2 - vida, 3 - faz, 4 - tempo, 5 - porque, 6 - sempre, 7 - melhor, 8 - tipo, 9 - pessoa, 10 - dia, 11 - apenas, 12 - nunca, 13 - boa, 14 - agora, 15 - forma, 16 - mano, 17 - mundo, 18 - todo, 19 - qualquer, 20 - sabe, 21 - caso, 22 - todos, 23 - anos, 24 - site, 25 - dar, 26 - dinheiro, 27 - pois, 28 - onde, 29 - gente, 30 - ficar, 31 - obrigado, 32 - saber, 33 - menos, 34 - falar, 35 - realmente, 36 - vezes, 37 - quero, 38 - mim, 39 - talvez, 40 - verdade, 41 - forum, 42 - hoje, 43 - pouco, 44 - existe, 45 - conta, 46 - outros, 47 - outro, 48 - problema, 49 - f***, 50 - disso, 51 - parte, 52 - outra, 53 - vez, 54 - tao, 55 - disse, 56 - cada, 57 - gosto, 58 - casa, 59 - dizer, 60 - meio, 61 - merda, 62 - sendo, 63 - fica, 64 - deve, 65 - uns, 66 - antes, 67 - pais, 68 - queria, 69 - precisa, 70 - ninguem, 71 - amigo, 72 - falando, 73 - brasil, 74 - resposta, 75 - certo, 76 - ideia, 77 - pergunta, 78 - pro, 79 - fato, 80 - desse, 81 - exemplo, 82 - outras, 83 - alem, 84 - tanto, 85 - tentar, 86 - deus, 87 - nesse, 88 - sentido, 89 - porem, 90 - dificil, 91 - grande, 92 - facil, 93 - fazendo, 94 - parece, 95 - posso, 96 - muita, 97 - acredito, 98 - quanto, 99 - fez, 100 - algumas

Tabela 34 – Lista das 100 palavras mais frequentes da classe dos posts relevantes dos CONJUNTOS DE DADOS II e III

Palavras mais Frequentes	
***Posts* Relevantes**	
Conjunto de Dados II	**Conjunto de Dados III**
1 - dados, 2 - site, 3 - pessoa, 4 - conta, 5 - tipo, 6 - saber, 7 - pessoas, 8 - hacking, 9 - senha, 10 - link, 11 - hacker, 12 - linux, 13 - caso, 14 - melhor, 15 - onde, 16 - rede, 17 - nome, 18 - sabe, 19 - curso, 20 - faz, 21 - tempo, 22 - boa, 23 - qualquer, 24 - sites, 25 - informacoes, 26 - apenas, 27 - forum, 28 - mano, 29 - acesso, 30 - todos, 31 - sempre, 32 - tor, 33 - celular, 34 - forma, 35 - web, 36 - seguranca, 37 - aprender, 38 - agora, 39 - porque, 40 - pois, 41 - sistema, 42 - internet, 43 - google, 44 - email, 45 - dar, 46 - possivel, 47 - programacao, 48 - quero, 49 - virus, 50 - cpf, 51 - facil, 52 - precisa, 53 - conseguir, 54 - links, 55 - criar, 56 - numero, 57 - outros, 58 - existe, 59 - dia, 60 - nunca, 61 - hackear, 62 - exemplo, 63 - surface, 64 - social, 65 - queria, 66 - uns, 67 - usando, 68 - posso, 69 - cursos, 70 - ajudar, 71 - windows, 72 - outra, 73 - porem, 74 - deve, 75 - conhecimento, 76 - estudar, 77 - pouco, 78 - todo, 79 - dinheiro, 80 - usa, 81 - engenharia, 82 - outro, 83 - menos, 84 - pergunta, 85 - algumas, 86 - deep, 87 - entrar, 88 - facebook, 89 - arquivo, 90 - obrigado, 91 - uso, 92 - vida, 93 - achar, 94 - outras, 95 - varios, 96 - python, 97 - realmente, 98 - cartao, 99 - mundo, 100 - arquivos	1 - dados, 2 - site, 3 - conta, 4 - pessoa, 5 - tipo, 6 - nome, 7 - boa, 8 - melhor, 9 - pessoas, 10 - onde, 11 - senha, 12 - telegram, 13 - faz, 14 - rede, 15 - informacoes, 16 - agora, 17 - saber, 18 - qualquer, 19 - cpf, 20 - acesso, 21 - forma, 22 - caso, 23 - porque, 24 - hacking, 25 - forum, 26 - dar, 27 - tempo, 28 - link, 29 - sites, 30 - mano, 31 - realmente, 32 - linux, 33 - sabe, 34 - obrigado, 35 - sistema, 36 - tor, 37 - email, 38 - sempre, 39 - possivel, 40 - apenas, 41 - facil, 42 - posso, 43 - seguranca, 44 - numero, 45 - outra, 46 - precisa, 47 - todos, 48 - celular, 49 - outro, 50 - dinheiro, 51 - maquina, 52 - internet, 53 - pergunta, 54 - todo, 55 - conseguir, 56 - uso, 57 - usando, 58 - pois, 59 - certo, 60 - achar, 61 - gente, 62 - porem, 63 - criar, 64 - queria, 65 - hacker, 66 - existe, 67 - web, 68 - google, 69 - hackear, 70 - amigo, 71 - falar, 72 - exemplo, 73 - outros, 74 - menos, 75 - dia, 76 - vpn, 77 - aprender, 78 - quero, 79 - links, 80 - talvez, 81 - pouco, 82 - disse, 83 - grupo, 84 - windows, 85 - vez, 86 - nunca, 87 - curso, 88 - tentar, 89 - kali, 90 - uns, 91 - ficar, 92 - arquivo, 93 - banco, 94 - usa, 95 - deve, 96 - social, 97 - algumas, 98 - sendo, 99 - dificil, 100 - acredito