Identificação de Ameaças em Fóruns da Dark Web e Surface Web: Um Estudo sobre a Evolução Temporal das Discussões e Generalização de Modelos

Miguel Henrique de Brito Pereira

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

Uberlândia
2026

Miguel Henrique de Brito Pereira

Identificação de Ameaças em Fóruns da Dark Web e Surface Web: Um Estudo sobre a Evolução Temporal das Discussões e Generalização de Modelos

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
2026

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

P436i
2025

Pereira, Miguel Henrique de Brito, 1993-
Identificação de ameaças em fóruns da Dark Web e Surface Web [recurso eletrônico] : um estudo sobre a evolução temporal das discussões e generalização de modelos / Miguel Henrique de Brito Pereira. - 2025.

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.2026.5021
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

Nelson Marcos Ferreira
Bibliotecário-Documentalista - CRB-6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em 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, 03/2026, PPGCO				
Data:	28 de Janeiro de 2026	Hora de início:	09:05	Hora de encerramento:	11:00
Matrícula do Discente:	12222CCP016				
Nome do Discente:	Miguel Henrique de Brito Pereira				
Título do Trabalho:	Identificação de ameaças em fóruns da Dark Web e Surface Web: um estudo sobre a evolução temporal das discussões e generalização de modelos				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Sistemas de Computação				
Projeto de Pesquisa de vinculação:	Identificação de ameaças de segurança usando mineração de dados de redes sociais e Dark Web - Financiado pela DataRisk.				

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, Diego Luis Kreutz - Unipampa e Rodrigo Sanches Miani - FACOM/UFU, orientador do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Silvio Ereno Quincozes - Alegrete/RS, Diego Luis Kreutz - Alegrete/RS. O outro membros da banca e o aluno(a) participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Rodrigo Sanches Miani, apresentou a Comissão Examinadora e o(á) candidato(a), agradeceu a presença do público, e concedeu ao(á) Discente a palavra para a exposição do seu trabalho.

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

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 28/01/2026, às 16:52, 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 **Diego Luis Kreutz**, **Usuário Externo**, em 28/01/2026, às 17:35, 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 29/01/2026, às 15:20, 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 **6975758** e o código CRC **3765F062**.

Referência: Processo nº 23117.000800/2026-66 SEI nº 6975758

À Natália, Angela, Thaís e Clara, vocês são a minha inspiração.

"Não há triunfo sem perda, não há vitória sem sofrimento."
(J.R.R Tolkien)

Resumo

Diante da crescente estruturação do cibercrime e do aumento de ameaças articuladas em ambientes anônimos, a Inteligência de Ameaças Cibernéticas (*Cyber Threat Intelligence* (CTI)) torna-se essencial para uma defesa proativa. Esta dissertação propõe dois estudos de caso com o intuito de avaliar aplicações práticas de CTI por meio da integração de técnicas de mineração de dados em fóruns da *Surface Web* e *Dark Web*. O primeiro estudo consiste em investigar a evolução temporal das discussões entre 2015 e 2024 utilizando modelagem de tópicos *Latent Dirichlet Allocation* (LDA), identificando padrões sazonais e uma transição temática de debates técnicos para práticas criminosas, como a comercialização de dados pessoais em língua portuguesa. O segundo estudo consiste em avaliar a eficácia da transferência de aprendizado para superar a escassez de dados rotulados no domínio da segurança. Para isso, foi utilizado um modelo baseado no algoritmo *LightGBM* com representação TF-IDF, desenvolvido previamente por um integrante do mesmo projeto de pesquisa, sendo este aplicado em domínios distintos e ambientes multilíngues, português e inglês. Os resultados demonstram que o modelo utilizado possui capacidade de generalização ao isolar vocabulários de risco em novas fontes, como mercados da *Dark Web* e fóruns de discussão genéricos, embora apresente alta sensibilidade a termos técnicos. Este estudo contribui para o desenvolvimento de modelos de CTI com alta eficiência e adaptabilidade em fontes de dados heterogêneas, auxiliando na antecipação de incidentes mesmo diante da escassez de dados rotulados.

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

Abstract

In light of the increasing structuring of cybercrime and the rise of coordinated threats in anonymous environments, Cyber Threat Intelligence (CTI) becomes essential for proactive defense. This dissertation proposes two case studies aimed at evaluating practical applications of CTI through the integration of data mining techniques in *Surface Web* and *Dark Web* forums. The first study investigates the temporal evolution of discussions between 2015 and 2024 using topic modeling with LDA, identifying seasonal patterns and a thematic transition from technical debates to criminal practices, such as the commercialization of personal data in the Portuguese language. The second study evaluates the effectiveness of transfer learning to overcome the scarcity of labeled data in the security domain. To this end, a model based on the *LightGBM* algorithm with TF-IDF representation was employed, previously developed by a member of the same research project, and applied across distinct domains and multilingual environments, Portuguese and English. The results demonstrate that the model exhibits generalization capability by isolating risk-related vocabularies in new sources, such as *Dark Web* marketplaces and generic discussion forums, although it shows high sensitivity to technical terms. This study contributes to the development of CTI models with high efficiency and adaptability across heterogeneous data sources, supporting the anticipation of incidents even in the presence of limited labeled data.

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*

BERT *Bidirectional Encoder Representations from Transformers*

CVE *Common Vulnerabilities and Exposures*

CPF *Cadastro de Pessoa Física*

CTI *Cyber Threat Intelligence*

DistilBERT *Distilled version of BERT*

HTML *HyperText Markup Language*

JSON *JavaScript Object Notation*

LSTM *Long Short Term Memory*

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*

PQC *Parameterized Quantum Circuits*

RoBERTa *Robustly Optimized BERT Pretraining Approach*

SCAP *Security Content Automation Protocol*

Tor *The Onion Router*

TF-IDF *Term Frequency–Inverse Document Frequency*

ULMFiT *Universal Language Model Fine-tuning*

XML *Extensible Markup Language*

XLM-R *Cross-lingual Language Model — RoBERTa*

Sumário

Capítulo 1

Introdução

A Internet impulsionou uma transformação digital sem precedentes e um aumento exponencial na geração de dados, redefinindo a sociedade e a forma como interagimos. Nesse vasto ecossistema de conectividade global, compreender a magnitude e a importância desse volume de dados e suas implicações, como as de segurança, é crucial.

Em contraste com os benefícios da conectividade, essa popularidade da internet trouxe consigo um aumento significativo e organizado das ameaças cibernéticas. Cibercriminosos têm se estruturado, transformando ataques isolados em crime organizado (RAHMAN; HEZAVEH; WILLIAMS, 2023). Fóruns online, especialmente na *Dark Web*, tornaram-se o ambiente principal para o compartilhamento e comercialização de vulnerabilidades, ataques e dados vazados (AVANZI et al., 2023). O monitoramento dessas discussões pode revelar padrões de comportamento, permitindo uma resposta proativa (KOLOVEAS et al., 2021), como ilustrado pelo caso de *ransomware* na Universidade de Utah em 2020 (CIMPANU, 2020) e pela exposição de vulnerabilidades do *Windows* em 2015 em fóruns da *Dark Web* (NUNES et al., 2016).

Conforme evidenciado por Nunes et al. (2016), grande parte desses ataques é discutida em fóruns da *Dark Web* antes de ocorrer. Um exemplo significativo, mencionado no mesmo artigo, refere-se a uma vulnerabilidade no sistema operacional *Microsoft Windows* em fevereiro de 2015. Em um curto período de pouco mais de um mês, surgiram relatos indicando que um fórum na *Dark Web* estava comercializando informações relacionadas a essa vulnerabilidade específica. A Figura 1 mostra a ocorrência dos fatos citados.

Para enfrentar esses crescentes desafios, o conceito de Inteligência de Ameaças Cibernéticas (*Cyber Threat Intelligence*, CTI) surge como essencial. CTI é o conhecimento fundamentado em evidências (mecanismos, indicadores e informações) sobre ameaças existentes ou emergentes, que permite tomar decisões de segurança proativas e mais rápidas (SHACKLEFORD, 2015). Neste contexto, o monitoramento de fóruns e comunidades da *Dark Web* emerge como uma fonte de CTI de alto valor.

Neste contexto, o monitoramento de fóruns e comunidades da *Dark Web* emerge como uma fonte de CTI de alto valor. Contudo, o estudo conduzido por (RAHMAN; HEZAVEH;

Timeline	Event
Feb. 2015	Microsoft identifies Windows vulnerability MS15-010/CVE 2015-0057 for remote code execution. There was no publicly known exploit at the time the vulnerability was released.
April 2015	An exploit for MS15-010/CVE 2015-0057 was found on a darknet market on sale for 48 BTC (around $10,000-15,000).
July 2015	FireEye identified that the Dyre Banking Trojan, designed to steal credit card number, actually exploited this vulnerability[1].

Figura 1 – Exemplo de vulnerabilidade exposta na *Dark Web*. Retirado de (NUNES et al., 2016)

WILLIAMS, 2023) aponta lacunas na aplicação de técnicas avançadas de aprendizado de máquina e na generalização desse conhecimento. Sistemas de CTI eficazes devem ser capazes de extrair informações de segurança de uma fonte e aplicá-las de forma generalizada para proteger contra novas ameaças.

A generalização de aprendizado refere-se, portanto, à capacidade de um modelo aprender padrões em um domínio de origem e aplicá-los com precisão a novos dados não vistos de um domínio alvo. Este conceito, também conhecido como Transferência de Aprendizado (*Transfer Learning*), tem sido amplamente abordado na literatura, incluindo trabalhos importantes como (ZHUANG et al., 2020), (WEISS; KHOSHGOFTAAR; WANG, 2016), (BARBON; AKABANE, 2022) e (ZHUANG et al., 2014).

Por exemplo, Zhuang et al. (2020) oferece uma análise abrangente das técnicas de generalização, cobrindo manipulação de dados e modelos. Os autores demonstram a aplicabilidade dessas técnicas em experimentos de classificação de texto em conjunto de dados distintos, sendo eles *Amazon Reviews, Reuters-21578* e *Office-31*. O trabalho de Weiss, Khoshgoftaar e Wang (2016) complementa essa visão com uma pesquisa abrangente, buscando formalizar o conceito e analisar as soluções de aprendizagem por transferência existentes. Embora não realize testes de aplicabilidade, o estudo fornece um panorama detalhado sobre as diferentes abordagens e *softwares* disponíveis no campo.

Outros trabalhos como (BARBON; AKABANE, 2022) e (ZHUANG et al., 2014), apresentam abordagens aplicáveis de duas formas distintas. No estudo conduzido em Barbon e Akabane (2022), os autores exploraram o uso de dois modelos, *Bidirectional Encoder Representations from Transformers* (BERT) e *Distilled version of BERT* (DistilBERT), em duas bases de dados, uma em português brasileiro e outra em inglês. Já em Zhuang et al. (2014), os pesquisadores desenvolveram um novo modelo e conduziram testes para avaliar sua aplicabilidade, embora o artigo tenha sido publicado quase uma década atrás, em 2014.

Além das questões teóricas de transferência de aprendizado, a aplicação prática de CTI enfrenta desafios de disponibilidade de dados e idioma. Primeiramente, a maioria

dos fóruns de ameaças cibernéticas e *Dark Web* é dominada pelo idioma inglês, o que restringe a capacidade de desenvolver soluções de CTI robustas para a língua portuguesa sem um esforço significativo de tradução ou coleta em fontes específicas, que são escassas. Em segundo lugar, há uma notável ausência de grandes conjuntos de dados rotulados publicamente no domínio da segurança cibernética. A rotulagem de postagens de fóruns para identificar ameaças requer conhecimento especializado em segurança, sendo um processo oneroso, demorado e sujeito a erros, o que reforça a necessidade de abordagens que sejam eficientes mesmo com a escassez de dados rotulados, como as técnicas não supervisionadas ou a generalização de aprendizado.

O presente trabalho propõe dois estudos de caso para avaliar aplicações práticas de CTI sob dados coletados em fóruns da *Dark Web* e *Surface Web*, contrastando com as abordagens exploradas na literatura. O primeiro consiste na investigação da evolução temporal dos tópicos abordados em fóruns da *Dark Web* e *Surface Web* (2015-2024), utilizando Modelagem de Tópicos (LDA). O segundo, é a aplicação da generalização de aprendizado, ou transferência de aprendizado, como estratégia de adaptação de modelos de CTI. Foi empregado um modelo de classificação previamente treinado (FILHO, 2024) para ser adaptado e avaliado em sua capacidade de classificar textos em diferentes fóruns, fontes e idiomas. A expectativa é que este trabalho contribua para o desenvolvimento de modelos de CTI com alta eficiência e adaptabilidade em fontes de dados futuras e heterogêneas, mesmo diante da escassez de dados rotulados e das barreiras linguísticas.

1.1 Objetivos e Questões

Com base na lacuna de conhecimento apresentada e na relevância da CTI preditiva, este trabalho se propõe a alcançar os seguintes objetivos principais e complementares:

- ❏ Investigar a evolução temporal de tendências de ameaças cibernéticas, analisando postagens em fóruns da *Dark Web* e da *Surface Web* entre 2015 e 2024, visando aprimorar estratégias de CTI preditiva.

- ❏ Avaliar o desempenho da transferência de aprendizado de modelos para identificação de ameaças em diferentes cenários, permitindo que o conhecimento adquirido em uma fonte seja transferido para outras.

Os objetivos específicos são:

- ❏ Construir um conjunto de dados rotulado e abrangente para análise temporal e de generalização, por meio da integração e complementação do conjunto de dados existente em (FILHO, 2024) com novas coletas de postagens de fóruns da *Dark Web* e *Surface Web* contendo indícios de ameaças cibernéticas.

- ❏ Aplicar técnicas de aprendizagem por transferência em um modelo previamente treinado, adaptando-o para a classificação de postagens em novos fóruns.

Alinhado aos objetivos delineados, formulou-se um conjunto de *Questões de Pesquisa* (QP) com o intuito de nortear a investigação. Cada questão decompõe o objetivo geral em problemas específicos e operacionalizáveis, permitindo uma análise pautada em métricas mensuráveis e hipóteses testáveis:

QP1 : De que maneira os tópicos de ameaças cibernéticas evoluíram temporalmente nos fóruns da *Dark Web* e *Surface Web*?

QP2 : Quais padrões sazonais ou picos de atividade podem ser correlacionados a tópicos de interesse específicos?

QP3 : Quais são as principais distinções estruturais e temáticas entre os conteúdos de fóruns da *Dark Web* e da *Surface Web* no contexto de ameaças cibernéticas?

QP4 : Qual é a eficácia da transferência de aprendizado na generalização de um modelo de classificação de *posts* maliciosos para domínios distintos?

1.2 Organização do Trabalho

O restante desta dissertação está estruturado nos seguintes capítulos. O Capítulo 2 estabelece a base teórica, abordando conceitos fundamentais sobre as camadas da *internet*, o modelo LDA, o conceito de Transferência de Aprendizado, dentre outros. Em seguida, o Capítulo 3 discute os principais trabalhos relacionados, com foco em CTI, mineração de dados e em como a Transferência de Aprendizado potencializa a Generalização em diferentes domínios. A metodologia completa é detalhada no Capítulo 4, apresentando as etapas de desenvolvimento do estudo. Por fim, o Capítulo 5 expõe e analisa os resultados obtidos em relação às questões de pesquisa, e o Capítulo 6 resume as conclusões e sugere trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, apresentam-se os conceitos fundamentais necessários para a compreensão desta dissertação, estruturada em nove seções que abrangem desde os conceitos básicos de cibersegurança e os pilares da tríade CID na Seção 2.1 até as distinções técnicas entre *Surface Web*, *Deep Web* e *Dark Web*, com foco no funcionamento da rede Tor, na Seção 2.2. São detalhados os processos de mineração de texto e *Processamento de Linguagem Natural* (PLN) na Seção 2.3, incluindo etapas de coleta e pré-processamento, seguidos pelos métodos de representação de dados, como *Bag of Words* e *Term Frequency–Inverse Document Frequency* (TF-IDF), na Seção 2.4. Adicionalmente, os paradigmas de classificação por aprendizado de máquina na Seção 2.5 e a modelagem de tópicos via LDA na Seção 2.6, bem como as métricas para avaliação de resultados e a matriz de confusão na Seção 2.7. Por fim, abordam-se as técnicas de transferência de aprendizado na Seção 2.8 e o conceito e ciclo de vida da Inteligência de Ameaças Cibernéticas (CTI) para defesa proativa na Seção 2.9.

2.1 Conceitos básicos de cibersegurança

Cibersegurança compreende um conjunto de estratégias e procedimentos empregados para proteger sistemas, redes e dados contra ameaças e ataques maliciosos. Segundo o *National Institute of Standards and Technology* (NIST) o conceito é definido como *"a proteção oferecida para um sistema de informação automatizado a fim de se alcançar os objetivos de preservar a integridade, a disponibilidade e a confidencialidade dos recursos do sistema de informação"* (GUTTMAN; ROBACK, 1995). Os três conceitos considerados fundamentais em segurança de computadores, conforme citados no NIST, incluem:

- **Confidencialidade**: Refere-se à proteção dos dados, garantindo que apenas as pessoas autorizadas tenham acesso a eles.
- **Integridade**: Dividida em duas vertentes, *integridade de dados* diz respeito à preservação dos dados contra modificações não autorizadas, garantindo não apenas a

preservação dos dados, mas também o não repúdio e a autenticidade das informações. Já a *integridade do sistema*, outra vertente dessa propriedade, é a qualidade que um sistema possui quando opera de maneira íntegra, sem manipulações não autorizadas, seja de forma intencional ou acidental.

- **Disponibilidade**: Assegura que os sistemas estejam disponíveis e que não ocorra interrupção de serviço para usuários autorizados.

Ao lidar com esses pilares, a segurança cibernética busca criar um ambiente digital confiável e resistente às ameaças. A tríade CID (Confidencialidade, Integridade e Disponibilidade), incorpora os objetivos de segurança fundamentais para dados, informações e serviços. Qualquer violação desses princípios é considerada um ataque ao sistema (STALLINGS; BROWN, 2014). Outros conceitos relevantes mencionados por outros autores incluem o **Não repúdio**, cujo propósito é garantir que ações de uma entidade sejam rastreadas e atribuídas apenas a essa entidade. Além disso, destaca-se a **Autenticidade**, que tem o objetivo de assegurar que a identidade do usuário ou sistema seja realmente quem ou o que afirmam ser.

2.1.1 Ameaças, vulnerabilidades e *exploits*

Segundo Stallings e Brown (2014), uma ameaça é definida como o potencial para uma violação na segurança do sistema, manifestando-se quando há circunstâncias, capacidades, ações ou eventos capazes de comprometer a segurança e causar danos. Em outras palavras, uma ameaça representa um perigo possível com a capacidade de explorar vulnerabilidades existentes. Além da definição de ameaça, a vulnerabilidade refere-se a uma falha, defeito ou fragilidade presente no projeto, implementação, operação e gerenciamento de um sistema. Essa vulnerabilidade apresenta o potencial de ser explorada, comprometendo a integridade do sistema e violando sua política de segurança. Explorar vulnerabilidades é uma prática muito conhecida entre os criminosos cibernéticos. Um ataque de exploração de vulnerabilidades ocorre quando um atacante, utilizando-se de uma vulnerabilidade do sistema, tenta executar ações maliciosas, como invadir um sistema, acessar informações confidenciais, disparar ataques contra outros computadores ou tornar um serviço inacessível. Essa tática é conhecida como *exploit*.

Com o propósito de catalogar e monitorar o histórico de cada vulnerabilidade, o MITRE, uma organização americana de pesquisa que presta serviços ao governo dos Estados Unidos, estabeleceu em 1999 o *Common Vulnerabilities and Exposures* (CVE), um banco de dados que registra vulnerabilidades e exposições relacionadas à segurança da informação (The MITRE Corporation, 2018). Da mesma forma, o *National Vulnerability Database* (NVD) é um repositório do governo dos Estados Unidos de dados para o gerenciamento de vulnerabilidades, baseado em padrões representados usando o *Security Content Automation Protocol* (SCAP), mantido pelo NIST. Esses sistemas possibilitam

a automação no gerenciamento de vulnerabilidades, assim como a medição da segurança e a avaliação de conformidade.

2.2 *Surface Web, Deep Web* e *Dark Web*

O termo *Surface Web* refere-se à parte da Internet prontamente acessível ao público em geral e pesquisável por meio de mecanismos de busca padrão, utilizando navegadores convencionais, também chamados de *browsers* (KAVALLIEROS et al., 2021). Esses programas permitem que os usuários interajam com documentos *HyperText Markup Language* (HTML) hospedados em servidores. Outro componente essencial da *Surface Web* são os indexadores, frequentemente chamados em trabalhos de *crawler* ou *spider*. Esses *softwares*, a partir de um ou mais *URLs*, fazem o *download* das páginas *web* associadas a esses endereços, extraem os *hiperlinks* contidos nelas e continuam o processo de forma recursiva, baixando novas páginas identificadas por esses *hiperlinks* (NAJORK, 2009).

Sites de busca como o *Google* utilizam técnicas avançadas de indexação. Após a descoberta de uma página, o *Google* analisa seu conteúdo, catalogando arquivos de imagens e vídeos incorporados, e tenta identificar o tema abordado. Essas informações são registradas no índice do *Google*, um grande banco de dados armazenado em diversos servidores (GOOGLE, 2021). O método da empresa é mais complexo do que apenas buscar novos hiperlinks, envolvendo também a catalogação detalhada de cada URL descoberta e acessada.

Segundo Kavallieros et al. (2021), a *Deep Web* corresponde a outra parte da web. De forma simplificada, ela é o oposto da *Surface Web*, pois seus mecanismos de busca não conseguem indexar seu conteúdo. Essa é a principal diferença entre ambas em termos de acessibilidade aos dados. Enquanto os *sites* na *Surface Web* são indexados e podem ser encontrados por mecanismos de busca, a *Deep Web* permanece fora desse processo.

Por fim, a *Dark Web* é uma parte da *Deep Web*, mas possui uma diferença fundamental: não pode ser acessada por meio de navegadores convencionais. Para isso, é necessário o uso de navegadores alternativos, como o Tor, que opera de maneira distinta dos navegadores tradicionais (KAVALLIEROS et al., 2021). O projeto Tor (*The Onion Routing*) foi criado para fornecer um método eficiente e seguro para que os usuários protejam suas identidades *online*. Ele foi desenvolvido pelo Laboratório de Pesquisa Naval dos Estados Unidos (NRL - *Naval Research Laboratory*) em meados da década de 1990, com o objetivo de proteger comunicações *online* na comunidade de inteligência dos Estados Unidos da América. Em 2004, o Tor foi lançado como um *software open-source*, disponível gratuitamente para o público e, atualmente, é mantido por voluntários e financiado por diversas fontes, incluindo o governo dos EUA, grupos de defesa dos direitos digitais e doadores individuais (SWAN, 2016). Hoje, o software está disponível para todos os sistemas operacionais. A utilização e navegação na *Dark Web* é ilustrada na Figura 2, que resume

os passos principais: (1) o usuário deve baixar o navegador Tor; (2) em seguida, deve encontrar *URLs* de interesse, as quais tipicamente possuem o domínio especial .onion (indicando que o endereço é um serviço oculto do Tor); e (3) finalmente, navegar na *Dark Web* acessando esses endereços. A descoberta de URLs .onion frequentemente depende de diretórios especializados, comunidades e fóruns dedicados, ou *links* compartilhados por outros usuários.

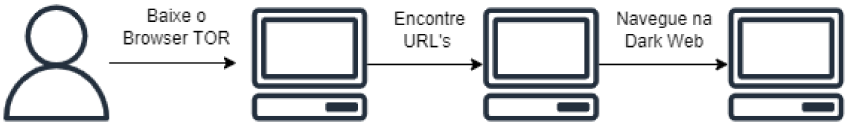

Figura 2 – Como navegar na *Dark Web* com Tor (Fonte: O autor (2025)).

A arquitetura da rede Tor é formada por sistemas descentralizados, um coletivo de servidores hospedados por voluntários em todo o mundo, chamados de nós. Um nó é um ponto de conexão em uma rede, com um endereço atribuído, e pode ser um roteador, terminal de computador, dispositivo periférico ou dispositivo móvel. Os nós são responsáveis pelo acesso e transferência de dados dos usuários, conectando-os aos seus destinos. Agindo como intermediários entre o usuário do Tor e o destino final, os nós também protegem o usuário contra o rastreamento de suas informações (SWAN, 2016). O funcionamento é mostrado na Figura 3.

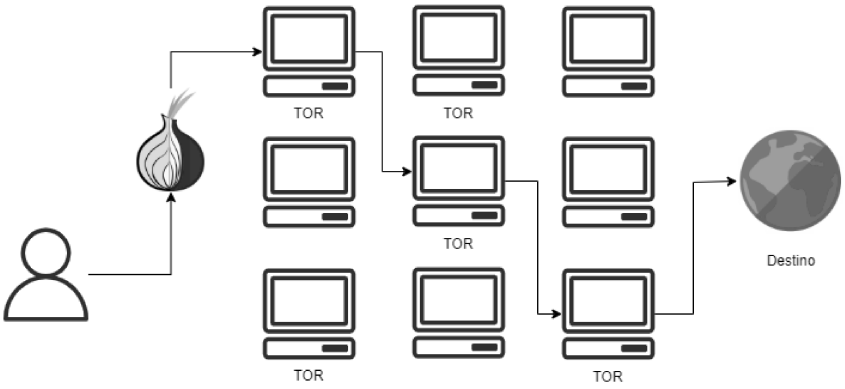

Figura 3 – Arquitetura projeto TOR (Fonte: O autor (2025)).

2.2.1 Fóruns de discussão na *Dark Web*

Fóruns de discussão na *Dark Web* são comunidades anônimas onde os usuários discutem temas que vão desde tecnologia e cibersegurança até assuntos de natureza ilegal. Esses fóruns operam em redes como o Tor, permitindo que os participantes mantenham o anonimato. Comunidades semelhantes também existem na *Surface Web* e na *Deep Web*, mas, diferentemente da *Dark Web*, nelas não há mecanismos que dificultem a identificação e localização dos usuários.

Conforme destacado por (SARKAR et al., 2018), a estrutura dos fóruns na *Dark Web* é hierárquica: cada fórum consiste em várias *threads* independentes. Cada *thread* corresponde a uma discussão sobre um tema específico. Dentro de uma *thread*, diversas postagens são feitas por diferentes usuários ao longo do tempo. Um mesmo usuário pode aparecer várias vezes na sequência de postagens, dependendo da frequência e do momento em que contribuiu para aquela discussão.

Para publicar em um fórum de discussão, basta criar uma conta, caso seja exigido, e seguir as regras da comunidade. O processo envolve escolher uma categoria apropriada, definir um título descritivo e elaborar o conteúdo da postagem. Alguns fóruns permitem publicações anônimas, enquanto outros exigem registro ou convite. Além disso, fóruns de acesso restrito podem requerer pagamento em criptomoedas. A aparência dos fóruns pode ser observada na Figura 4.

Figura 4 – Fórum Respostas Ocultas (Fonte: O autor (2025)).

2.3 Mineração de texto e processamento de linguagem natural

A mineração de texto, também denominada mineração de dados textuais, refere-se ao processo de converter texto não estruturado em um formato estruturado, com o intuito de identificar padrões, conceitos, tópicos, palavras-chave e outros atributos nos dados (STEDMAN, 2023). O texto representa uma das formas de informação mais frequentemente armazenadas em bancos de dados, esse conteúdo pode ser organizado como:

- **Dados estruturados**: Esses dados são padronizados de maneira tabular, com várias linhas e colunas, o que simplifica tanto o armazenamento quanto o proces-

samento por meio de algoritmos de análise e aprendizado de máquina. Os dados estruturados incluem entradas como nomes, endereços e números de telefone.

- **Dados não estruturados**: Esses dados não seguem um formato predefinido, podendo conter textos oriundos de diversas fontes, como redes sociais, avaliações de produtos, fóruns, arquivos de vídeo e áudio.

- **Dados semiestruturados**: Esses dados, o nome sugere, são uma combinação entre os formatos de dados estruturados e não estruturados. Embora possuam alguma organização, não tem estrutura suficiente para atender aos requisitos de um banco de dados relacional. Exemplos de dados semiestruturados englobam arquivos nos formatos *Extensible Markup Language* (XML), *JavaScript Object Notation* (JSON) e HTML(IBM, 2023).

Segundo Dialani (2020), cerca de 80% dos dados no mundo estão em formato não estruturado, e sendo que cerca de 90% desse montante foi gerado nos últimos dois anos. Dentro desse vasto conjunto de dados, apenas 0,5% são analisados e utilizados atualmente. Ferramentas de mineração de texto e técnicas de processamento de linguagem natural (PLN) auxiliam na conversão desses documentos não estruturados para um formato estruturado, possibilitando a análise.

A área de Processamento de Linguagem Natural (PLN) se dedica à pesquisa que tem como objetivo investigar e propor métodos e sistemas de processamento computacional da linguagem humana. A palavra **Natural** na sigla se refere a línguas faladas pelos humanos, distinguindo-as das demais linguagens como matemáticas, visuais, gestuais, de programação e outras. No campo do PLN, busca-se desenvolver soluções para problemas computacionais, abrangendo tarefas, sistemas, aplicações ou programas que requerem o tratamento computacional de uma língua como português ou inglês, seja escrita ou falada (CASELI; NUNES, 2023). O PLN se divide em duas grandes áreas, a saber:

- **Interpretação ou Compreensão de Linguagem Natural – NLU (*Natural Language Understanding*)**: Todas as atividades relacionadas ao processamento voltado para a análise e interpretação da linguagem estão contempladas nesta abordagem. Por análise, compreende-se a segmentação e classificação dos componentes linguísticos, abrangendo palavras e suas classes morfológicas e gramaticais, assim como seus traços semânticos ou ontológicos. Por outro lado, a interpretação envolve a tentativa de apreender significados construídos pelo ser humano(CASELI; NUNES, 2023).

- **Geração de Linguagem Natural – NLG (*Natural Language Generation*)**: O objetivo da NLG é a geração de linguagem natural. Um exemplo atual é o *ChatGPT*, que é capaz de gerar linguagem de forma tão ou mais fluente quanto muitos humanos(CASELI; NUNES, 2023).

Na literatura, o conceito de coleção de documentos é dado as fontes de dados. Um documento, elemento básico de uma coleção textual que se deseja analisar, pode ser representado por uma página da *web*, um *e-mail* ou um *tweet*, por exemplo (KONCHADY, 2006). Na prática, a maioria das soluções de mineração de texto tem como objetivo identificar padrões em coleções de documentos muito grandes, abrangendo desde muitos milhares até dezenas de milhões (FELDMAN; SANGER et al., 2007).

A mineração de textos pode ser realizada em várias etapas. A Figura 5 mostra uma sequência potencial dessas fases aplicadas à classificação de textos oriundos de redes sociais, fóruns da *Dark Web* e *Deep Web*. Ressalta-se que a adoção de cada procedimento está condicionada à aplicação particular, sendo, em alguns casos, indispensável a inclusão de passos adicionais.

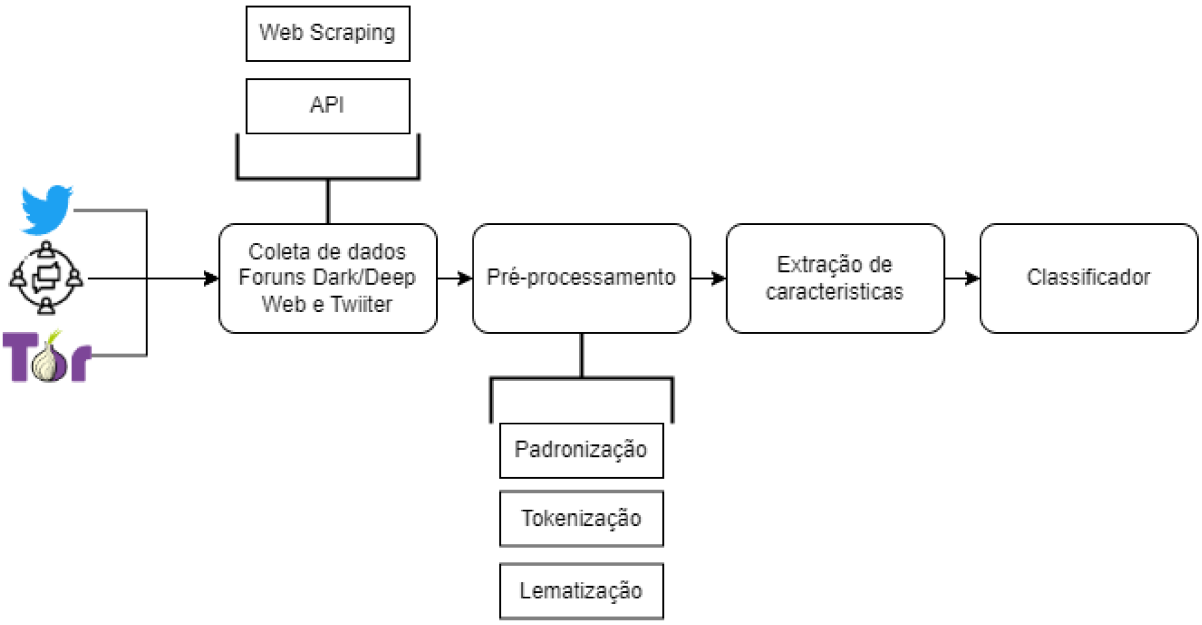

Figura 5 – Etapas processo de mineração de texto (Fonte: O autor (2025)).

2.3.1 Coleta de dados

O processo inicial da mineração de texto envolve a coleta de dados. *Crawler* ou *web crawler* são algoritmos de coleta de dados na *web*, frequentemente referidos também como *spider* ou *scraper*. Eles são utilizados para extrair dados que podem então ser processados e analisados em tarefas como a mineração de dados. Esses algoritmos têm a função de vasculhar *sites* ou bancos de dados para diversos propósitos, como catalogar páginas, indexar conteúdo ou até mesmo extrair partes específicas (CRAWLY, 2021).

Os dados coletados pelos *crawlers* podem abranger desde conteúdos disponíveis na *Surface Web*, com estrutura padronizada e de fácil acesso, até informações mais restritas presentes na *Dark Web*, onde a navegação exige anonimato e os *sites* frequentemente apresentam formatos não convencionais. Nessas camadas, a atuação dos algoritmos de

coleta demanda estratégias mais refinadas, como a extração seletiva de informações e a identificação de padrões em ambientes dinâmicos e descentralizados.

Além da coleta direta via navegação automatizada, algumas plataformas oferecem APIs que permitem o acesso estruturado e autorizado aos seus dados. Um exemplo bastante utilizado na mineração de texto é a *Application Programming Interface* (API) do *X/Twitter*, que possibilita a coleta de *tweets*, metadados, interações e outros elementos textuais. Após a extração, os dados costumam ser organizados em formatos estruturados, como arquivos JSON, XML ou planilhas, o que facilita sua análise posterior e integração com outras fontes.

2.3.2 Pré-processamento

A etapa seguinte ocorre o pré-processamento, sendo responsável por transformar os dados linguísticos brutos em formatos estruturados e mais adequados para análises posteriores. Segundo Nayak et al. (2016), essa fase inicial tem como objetivo remover ruídos, padronizar termos e reduzir a dimensionalidade do conjunto de dados, possibilitando maior eficiência nos algoritmos de mineração de texto. O processo de pré-processamento é composto por diversas etapas complementares, destacando-se:

- Padronização e Limpeza: Inclui a conversão de letras maiúsculas para minúsculas, remoção de pontuação, números, caracteres especiais e espaços em branco. Essa normalização é para garantir que diferentes formas gráficas de um mesmo termo sejam tratadas como equivalentes.

- Tokenização: Consiste na segmentação do texto em unidades menores chamadas *tokens*, geralmente palavras e símbolos. O critério principal para essa divisão geralmente ocorre na presença de espaços em branco ou sinais de pontuação. Essa etapa é fundamental, pois permite filtrar termos irrelevantes durante as fases posteriores(TABASSUM; PATIL, 2020). Por exemplo, considere a sentença:

  – "NLP is the future of Speech Recognition Systems!"

  Após a tokenização, a frase será dividida nos seguintes tokens:

  – "NLP", "is", "the", "future", "of", "Speech", "Recognition", "Systems", "!"

- Remoção de Palavras-Vazias (*Stopwords*): A remoção de *stopwords* refere-se ao processo de eliminação dessas palavras da lista de *tokens*. Segundo (JO, 2024), uma *stopword* é uma palavra cuja função é puramente gramatical e que não possui valor semântico relevante para a análise do conteúdo. Entre os exemplos mais comuns desse grupo, incluem-se preposições, como "em", "sobre", "para"; conjunções, como "e", "ou", "mas" e "porém"; e artigos definidos e indefinidos, como "o", "a", "um" e "uma".

- *Stemming* e Lematização: São técnicas que visam reduzir as palavras às suas formas básicas. O *stemming* refere-se ao processo de mapeamento de cada *token*, gerado na etapa anterior, para sua forma de raiz, por meio da aplicação de regras para remoção de sufixos (JO, 2024). A lematização é o processo de agrupar diferentes formas flexionadas de uma palavra e reduzi-las à sua forma canônica ou lema, com preservação do significado. Diferentemente do *stemming*, que apenas remove sufixos, a lematização leva em consideração o contexto gramatical e semântico da palavra (KHYANI et al., 2021). Por exemplo, observe como as palavras abaixo são reduzidas ao seu lema em português:
  - correr, correu, correndo, correria → correr
  - bons, boa, boas → bom
  - homens → homem

2.4 Representação de Texto

A representação de texto é a etapa responsável por transformar dados textuais brutos em formas vetoriais que possam ser processadas por algoritmos de aprendizado de máquina. Trata-se de uma etapa no processamento de linguagem natural, pois estabelece a ponte entre o texto não estruturado e os modelos computacionais.

Em conjunto com a representação de texto, a extração de características envolve diversas técnicas de vetorização, como *Bag of Words* (BoW), TF-IDF e *N-Grams*. Essas abordagens têm como objetivo mapear os dados textuais em um espaço vetorial, preservando informações relevantes de frequência, contexto e semântica (GARG; SHARMA, 2022).

2.4.1 *Bag of Words*

A técnica *Bag of Words* desempenha um papel fundamental na extração de características do texto que são então alimentadas a um classificador. O modelo representa o texto como uma coleção de palavras sem atribuir importância ao contexto ou à ordem em que essas palavras aparecem.

Dessa forma, parte-se da intuição de que textos ou documentos que contêm as mesmas palavras são semelhantes em contexto. Esse modelo é amplamente utilizado em tarefas como classificação de documentos, correspondência de palavras-chave, entre outras.

As palavras selecionadas geralmente representam o conteúdo central de uma sentença. Embora a gramática e a ordem de aparecimento das palavras sejam ignoradas, a frequência das palavras é contabilizada e pode ser utilizada posteriormente para identificar os pontos de foco dos documentos(KOWSARI et al., 2019a; TABASSUM; PATIL, 2020). Um exemplo da aplicação da *Bag of Words*, como resultado visual, temos a Figura 6:

- Documento: “Eu gosto de café. Ela também gosta de café com leite.”
- *Bag of Words*: “Eu”, “gosto”, “de”, “café”, “Ela”, “também”, “com”, “leite”
- Frequência das palavras: 1, 2, 2, 2, 1, 1, 1, 1

Figura 6

2.4.2 TF-IDF

O *Term Frequency–Inverse Document Frequency* (TF-IDF) é uma técnica utilizada na recuperação da informação e na mineração de texto, com o objetivo de avaliar a importância de uma palavra em um documento, dentro de um conjunto de documentos. Essa técnica atribui um peso a cada termo de um documento, por meio da multiplicação entre a frequência do termo no documento (TF) e o inverso da frequência do termo em todos os documentos de um corpus ou coleção (IDF) (CASELI; NUNES, 2023).

A frequência de um termo (TF) mede sua importância em um documento, indicando quantas vezes determinado termo aparece nesse documento. É conhecido que o comprimento dos documentos pode variar, com isso tem a possibilidade de que certos termos ocorram com maior frequência em documentos mais extensos, em comparação aos mais curtos. Para corrigir essa distorção, a frequência de um termo em um documento é calculada dividindo-se o número de ocorrências do termo t em um documento d pelo total de termos presentes nesse mesmo documento d (QAISER; ALI, 2018). O resultado representa a frequência do termo.

A sua contraparte, a frequência inversa no documento (IDF), determina a importância de uma palavra em um conjunto de documentos. Basicamente, ela reduz a influência de palavras menos relevantes, calcula-se dividindo-se o número total de documentos da coleção pelo número de documentos que contêm a palavra em questão e tomando o logaritmo desse resultado (CASELI; NUNES, 2023; TABASSUM; PATIL, 2020).

$$\mathrm{TF}(t,d) = \frac{\mathrm{numero\_ocorrências}(t,d)}{\mathrm{total\_termos}(d)} \qquad \mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{total\_documentos}(t)}\right) \tag{1}$$

Dessa forma, quanto maior a ocorrência de uma palavra nos documentos, maior será o valor de TF. Por outro lado, quanto menor a ocorrência dessa palavra nos documentos,

maior será o valor de IDF, especialmente quando essa palavra for buscada em um documento específico (QAISER; ALI, 2018). Quanto menor for o número de documentos que contêm determinado termo, maior será o valor de TF-IDF atribuído a esse termo. Em síntese, termos que aparecem com frequência em diversos documentos recebem um peso menor do que aqueles mais específicos de um determinado documento. O TF-IDF nada mais é do que o produto entre o valor de TF e o valor de IDF.

$$TF - IDF(t, d) = tf(t, d) * idf(t) \tag{2}$$

2.4.2.1 *N-Gramas*

O n-grama é um conceito utilizado em processamento de linguagem natural e estatística de texto. Ele representa uma sequência de n elementos consecutivos de um dado texto ou fala. Esses elementos podem ser palavras, caracteres ou até mesmo símbolos, dependendo do contexto.

O modelo *Bag of Words* consiste em uma representação de texto utilizando apenas suas palavras (1-grama), desconsiderando a ordem sintática dessas palavras. Trata-se de um modelo no qual o texto é representado por um vetor, geralmente de tamanho razoável. Por sua vez, o n-grama pode ser entendido como uma extensão do modelo *Bag of Words*, permitindo a representação de texto por meio de sequências de n-gramas. É comum o uso de 2-gramas e 3-gramas. Dessa forma, as características extraídas do texto permitem capturar mais informações em comparação ao uso de 1-gramas (KOWSARI et al., 2019b).

Exemplo de 1-grama, são apenas palavras individuais:

- "O", "gato", "dorme"

Exemplo de 2-gramas, pares de palavras consecutivas:

- "O gato", "gato dorme"

Exemplo de 3-gramas, sequências de três palavras:

- "O gato dorme"

2.5 Classificação de Texto e Aprendizado de Máquina

O processo de atribuir documentos a classes, ou seja, associar um ou mais rótulos de classe a cada documento, é comumente denominado classificação de textos. Uma classe pode ser definida como um grupo rotulado, um conjunto no qual podemos categorizar documentos cujo conteúdo pode ser descrito pelo seu respectivo rótulo. Um exemplo de classes seria a distinção entre crítica boa e crítica ruim (BAEZA-YATES; RIBEIRO-NETO, 2013).

A etapa mais importante da classificação de texto é a escolha do melhor classificador. Sem uma compreensão completa de cada algoritmo, não é possível determinar com eficácia o modelo mais eficiente para uma aplicação de classificação de texto.

O paradigma de classificação adotado nos modelos de classificação de texto geralmente é supervisionado, uma vez que se utilizam conjuntos de treinamento rotulados. Nesses modelos, é empregado um classificador que passa por um processo de ajuste antes de ser aplicado a coleções de documentos. Esse ajuste é realizado com base em um conjunto de treinamento, com o objetivo de aprimorar a acurácia do classificador, e com isso a sua capacidade de classificação. Para avaliar o desempenho do classificador, utiliza-se um conjunto de documentos pré-selecionados, cujas classes corretas são conhecidas — o conjunto de teste. Este conjunto de teste não é utilizado no treinamento, permitindo uma avaliação imparcial. Após o treinamento e a validação, o classificador pode ser utilizado para classificar documentos novos e não vistos (BAEZA-YATES; RIBEIRO-NETO, 2013). As etapas mencionadas anteriormente são ilustradas na Figura 7. Dentre os algoritmos de aprendizado supervisionado mais conhecidos, destacam-se: *K-Nearest Neighbors, Naive Bayes, Support Vector Machines e Decision Trees.*

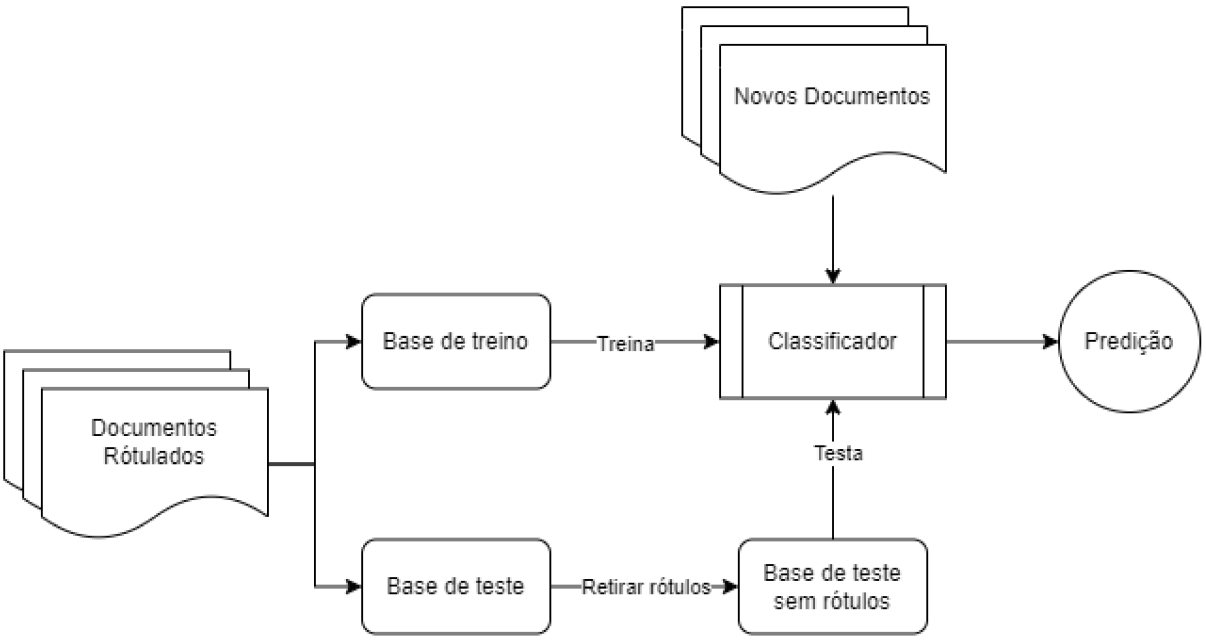

Figura 7 – Processo de classificação e de avaliação de um classificador supervisionado (Fonte: O autor (2025)).

2.6 Modelagem de tópicos - LDA

O *Latent Dirichlet Allocation* (LDA) constitui um modelo probabilístico não supervisionado empregado na modelagem de tópicos(BLEI; NG; JORDAN, 2003). Ele opera sob a premissa de que os documentos são gerados a partir de uma combinação de temas subjacentes, sendo esses temas caracterizados pela frequência com que certas palavras

coocorrem. O LDA infere a proporção de cada tema em cada documento, revelando a organização temática intrínseca do *corpus*. Empregada no processamento de linguagem natural, essa abordagem permite detectar padrões e relações entre textos, facilitando a identificação de conteúdos semelhantes.

A primeira etapa do algoritmo consiste no pré-processamento do texto para a extração de tópicos, como a remoção de palavras de parada ou termos irrelevantes. Essa prática visa eliminar elementos que não contribuem para o significado desejado da análise, preservando a integridade da informação. Nesse sentido, palavras como “as”, “os”, “uns”, “de”, “para”, “com” e “por” são consideradas palavras de parada. Outra etapa essencial é a tokenização, que consiste em dividir o texto em palavras individuais, como vimos anteriormente, removendo pontuações e espaços extras.

Por fim, o algoritmo gera listas de palavras-chave com as respectivas probabilidades de acordo com a frequência de palavras e coocorrências. Rastreando a frequência de coocorrências, o LDA pressupõe que as palavras que ocorrem juntas provavelmente fazem parte de tópicos semelhantes. Por exemplo, suponha que tenhamos um LDA para uma coleção de artigos de notícias, teríamos esse resultado parcial na Figura 8(MUREL, 2024).

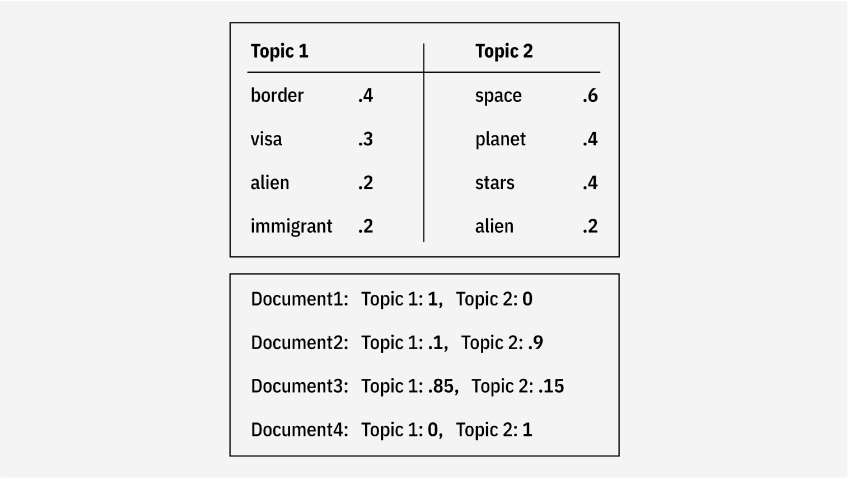

Figura 8 – Tópicos que provavelmente podem ser descritos como imigração (Tópico 1) e astronomia (Tópico 2). Fonte: Retirado de (MUREL, 2024)

As pontuações associadas a cada palavra na Figura 8 representam a probabilidade de essa palavra-chave aparecer em um determinado tópico. Da mesma forma, as probabilidades atribuídas a cada documento indicam a chance de ele pertencer a uma combinação de tópicos, considerando a distribuição e a coocorrência de palavras em seu conteúdo.

2.7 Avaliação e Interpretação de Resultados

A correta aplicação de métodos e métricas de avaliação é essencial para garantir estimativas confiáveis de desempenho e evitar vieses durante o desenvolvimento de modelos de aprendizado de máquina. De acordo com Powers (2020), medidas como acurácia, preci-

são, revocação e medida F1 são fundamentais para analisar a capacidade de generalização de um classificador e compreender suas limitações em cenários reais. Estudos recentes reforçam essa importância ao demonstrar que a utilização dessas métricas permite não apenas a comparação justa entre diferentes algoritmos, mas também o aprimoramento de soluções práticas.

Além disso, a aplicação correta dos métodos e métricas de avaliação possibilita identificar pontos fortes e limitações do modelo, orientando ajustes no processo de treinamento e na seleção de algoritmos. Dessa forma, não apenas se obtém uma estimativa confiável de desempenho, mas também se cria uma base sólida para aprimorar continuamente a qualidade das soluções desenvolvidas.

2.7.1 *Hold-out*

O método *Hold-out* é uma técnica de avaliação de modelos de aprendizado de máquina, onde o conjunto de dados é dividido em dois subconjuntos: um para treinamento e outro para teste. Essa abordagem permite estimar o desempenho do modelo em dados não vistos, minimizando os riscos de *overfitting*, fenômeno caracterizado pelo ajuste excessivo do modelo aos dados de treino, comprometendo sua capacidade de generalização. Convencionalmente, utiliza-se uma proporção comum, como 70% para treinamento e 30% para teste (70/30) ou 80/20, mas essas divisões podem variar conforme o tamanho da base de dados.

Apesar de sua simplicidade, a principal limitação do método está na sua dependência de uma única partição dos dados, o que pode gerar estimativas de desempenho enviesadas, especialmente em bases pequenas (KOHAVI, 2001). Para mitigar esse problema, outras técnicas, como a validação cruzada (*k-fold cross-validation*), são frequentemente utilizadas, oferecendo maior robustez estatística (WONG, 2015). Ainda assim, o *hold-out* é vantajoso em situações com grandes volumes de dados, em que múltiplas divisões não são necessárias para obter estimativas confiáveis.

2.7.2 Matriz de confusão

A matriz de confusão é uma estrutura matricial que resume o desempenho de um modelo de classificação, indicando sua capacidade de prever corretamente as classes atribuídas a determinados exemplos. Um dos eixos da matriz representa os rótulos previstos pelo modelo, enquanto o outro corresponde aos rótulos reais (BURKOV, 2019), conforme ilustrado na Figura 1.

A classificação binária é um tipo fundamental de problema de aprendizado de máquina, onde o objetivo é separar dados em duas classes mutuamente exclusivas. Para esse tipo de classificação, convenciona-se que P representa a classe positiva e N representa a classe negativa. Uma amostra desconhecida é classificada como P ou N por um mo-

delo previamente treinado. Esse modelo pode produzir dois tipos de saída: discreta ou contínua.

A saída discreta refere-se à atribuição direta de um rótulo de classe à amostra de teste, como por exemplo, classe positiva (P) ou negativa (N). Esse tipo de saída representa uma decisão categórica do modelo, sem indicar o grau de confiança na previsão.

Por outro lado, a saída contínua corresponde à estimativa da probabilidade de pertencimento da amostra a uma determinada classe. Nesse caso, o modelo retorna um valor numérico, geralmente entre 0 e 1, que expressa o nível de confiança associado à classificação. Por exemplo, uma probabilidade de 0,85 indica que o modelo estima 85% de chance de a amostra pertencer à classe positiva. Essa probabilidade pode ser convertida em uma saída discreta por meio da aplicação de um limiar de decisão, como 0,5.

		Valor Real	
		Sim	**Não**
Valor Predito	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 9 – Representação de uma matriz de confusão 2 × 2. A matriz considera duas classes reais: P (positiva) e N (negativa), com previsões classificadas como verdadeiras ou falsas. Fonte: Adaptado de (THARWAT, 2018)

Na matriz, a diagonal verde representa as previsões corretas, enquanto a diagonal rosa indica as previsões incorretas. Quando uma amostra positiva é corretamente classificada como positiva, ela é considerada um verdadeiro positivo (VP). Caso seja classificada como negativa, trata-se de um falso negativo (FN). Por outro lado, se a amostra negativa for corretamente classificada como negativa, é considerada um verdadeiro negativo (VN); caso seja incorretamente classificada como positiva, configura-se um falso positivo (FP), também conhecido como alarme falso (THARWAT, 2018).

2.7.3 Acurácia

A acurácia é uma métrica que expressa a proporção de previsões corretas feitas por um modelo de classificação em relação ao total de previsões realizadas. É utilizada na avaliação de modelos, principalmente quando as classes do conjunto de dados estão balanceadas. No contexto de tarefas de classificação, a acurácia pode ser definida como a fração das previsões corretas, podendo ser definida como:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

Onde VP é o número de amostras positivas corretamente classificadas; VN, o número de amostras negativas corretamente classificadas; FP, o número de amostras negativas incorretamente classificadas como positivas; e FN, o número de amostras positivas incorretamente classificadas como negativas (CASELI; NUNES, 2023).

2.7.4 Precisão

A precisão é a métrica que indica a proporção de casos previstos como positivos que realmente são positivos. Em outras palavras, mede o quanto das previsões positivas feitas por um modelo estão corretas. Essa métrica é relevante em áreas como *machine learning, data mining* e *information retrieval*, sendo para avaliar a qualidade dos resultados obtidos(POWERS, 2020).

$$Precisão = \frac{VP}{VP + FP} \tag{4}$$

Sendo a métrica, o número de amostras positivas classificadas corretamente dividido pelo número de amostras que o modelo rotulou como positivas.

2.7.5 Revocação

A revocação de um classificador corresponde à proporção de amostras positivas corretamente identificadas em relação ao número total de amostras positivas. Essa métrica é especialmente relevante quando falsos negativos (FN) têm impacto crítico. Ela expressa a eficácia do classificador na identificação de rótulos positivos(SOKOLOVA; LAPALME, 2009).

$$Revocação = \frac{VP}{VP + FN} \tag{5}$$

A revocação é calculada como o número de amostras positivas classificadas corretamente dividido pelo número total de amostras positivas no conjunto de dados.

2.7.6 Medida F1 (*F1-Score*)

A medida F1 (*F1-Score* é a média harmônica entre a precisão e a revocação. Seu valor varia de zero a um, sendo que valores altos indicam um bom desempenho de classificação. O *F1-Score* próximo de 1 demonstra que o modelo mantém um bom equilíbrio entre detectar casos positivos e evitar falsos alarmes (THARWAT, 2020).

$$F1 - Score = 2\frac{precisão \times revocação}{precisão + revocação} \tag{6}$$

2.8 Transferência de Aprendizado de Máquina

A generalização de aprendizado ou aprendizagem por transferência é uma técnica no campo do aprendizado de máquina na qual o conhecimento previamente adquirido por um modelo treinado é aplicado a um problema distinto, porém no mesmo domínio relacionado. Em outras palavras, é a capacidade de utilizar as habilidades e entendimentos adquiridos por um modelo em uma tarefa específica para aprimorar o desempenho em uma tarefa diferente. Por exemplo, ao treinar um classificador simples para identificar a presença de uma mochila em imagens, é possível empregar o conhecimento assimilado durante o treinamento para aprimorar a capacidade do modelo em reconhecer outros objetos, como óculos de sol (DONGES, 2022). Um exemplo visual do cenário anterior é apresentado na Figura 10.

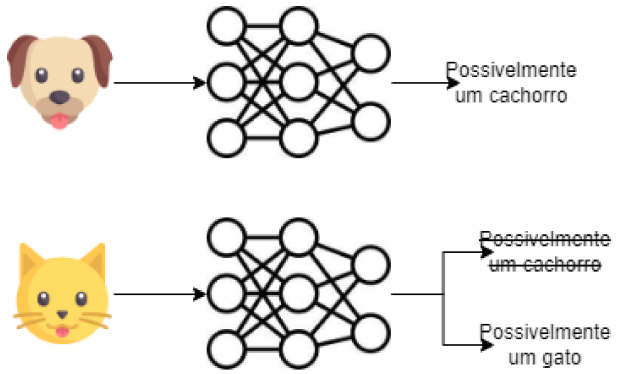

Figura 10 – Modelo previamente treinado para a predição de imagens de cachorros, agora para reconhecer gatos.

A necessidade de aprendizagem por transferência ocorre quando há uma limitação no fornecimento de dados de treinamento para o domínio alvo. Essa limitação pode ser devida à escassez de dados disponíveis, tornando a coleta e rotulação bastante custosas, ou à inacessibilidade dos dados. Com repositórios de *big data* tornando-se mais predominantes, a utilização de conjuntos de dados relacionados, mas não exatamente iguais ao domínio alvo de interesse, torna as soluções de aprendizagem por transferência atrativas para essa abordagem (WEISS; KHOSHGOFTAAR; WANG, 2016).

Segundo (WEISS; KHOSHGOFTAAR; WANG, 2016), problemas de aprendizagem por transferência podem ser divididos em três categorias:

- **Transferência Indutiva**: A aprendizagem por transferência indutiva é exemplificada pelo cenário em que dados rotulados do domínio alvo estão prontamente disponíveis.

- **Transferência Transdutiva**: A aprendizagem por transferência transdutiva caracteriza situações em que dados rotulados estão presentes no domínio de origem, mas ausentes no domínio alvo.

- ❑ **Transferência Não Supervisionada**: A aprendizagem por transferência não supervisionada diz respeito a cenários em que nem o domínio de origem nem o domínio alvo possuem dados rotulados.

Na presente dissertação, a metodologia empregada é a transferência transdutiva, pois os dados do domínio alvo encontram-se disponíveis (modelo previamente treinado nos dados da *Dark Web*) mas ausentes nos domínios alvo (outros tipos de fóruns).

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

À medida que se enfrenta a crescentes desafios digitais, torna-se necessário buscar métodos mais eficazes para prevenir ataques cibernéticos, vazamentos de dados e vulnerabilidades. Um conceito contemporâneo nesse cenário é a inteligência de ameaças cibernéticas, conhecida como *CyberThreat Intelligence* (CTI). De acordo com (SHACKLEFORD, 2015), a inteligência de ameaças cibernéticas é definida como "conhecimento fundamentado em evidências, mecanismos, indicadores e exploração de informações relacionadas a uma ameaça ou perigo existente ou emergente para sistemas. Esses elementos são essenciais para tomar decisões de segurança mais rápidas diante dessa ameaça ou perigo." Em linhas gerais, a CTI envolve a coleta de dados brutos provenientes de fontes de segurança cibernética, resultando em conhecimento que pode nos ajudar na tomada de decisões de defesa proativa na segurança cibernética, abrangendo também a prevenção de potenciais ataques.

Compreender as novas técnicas e métodos empregados por esses grupos criminosos, bem como as ferramentas utilizadas para prevenir tais ataques, é essencial. Por exemplo, em 2021, a JBS pagou o equivalente a US$ 11 milhões em resgate de dados após o ataque de *ransomware* à empresa. O objetivo, segundo a companhia, foi reduzir potenciais problemas relacionados à invasão (GLOBO, 2021).

Outra vertente desses ataques diz respeito aos vazamentos de dados. Em 2022, a Uber foi alvo de um ataque cibernético no qual o invasor obteve acesso ao sistema ao adquirir informações de um funcionário por meio de um fórum na *Dark Web* e conseguiu as credenciais através de *phishing* (KOST, 2022). Semelhantemente aos fóruns na *Dark Web*, o *Twitter* é uma plataforma de mídia social muito utilizada por usuários e especialistas em segurança, servindo como um meio para antecipar a divulgação de vulnerabilidades e ataques.

Nesse contexto, a inteligência de ameaças cibernéticas (CTI) desempenha um papel fundamental na defesa proativa contra esses ataques. O trabalho (RAHMAN; HEZAVEH; WILLIAMS, 2023) identificou uma arquitetura composta por etapas comuns, presente na maioria das publicações analisadas. Nesse sentido, oferece uma revisão dos estudos

existentes sobre extração automática de CTI, utilizando novas técnicas de processamento de linguagem natural e aprendizado de máquina. O processo de extração e análise de CTI (*pipeline*) proposto, mostrado na Figura 11, pode ser adaptado conforme os objetivos de extração de CTI, proporcionando aos pesquisadores de segurança cibernética opções para projetar seus próprios pipelines de extração de CTI.

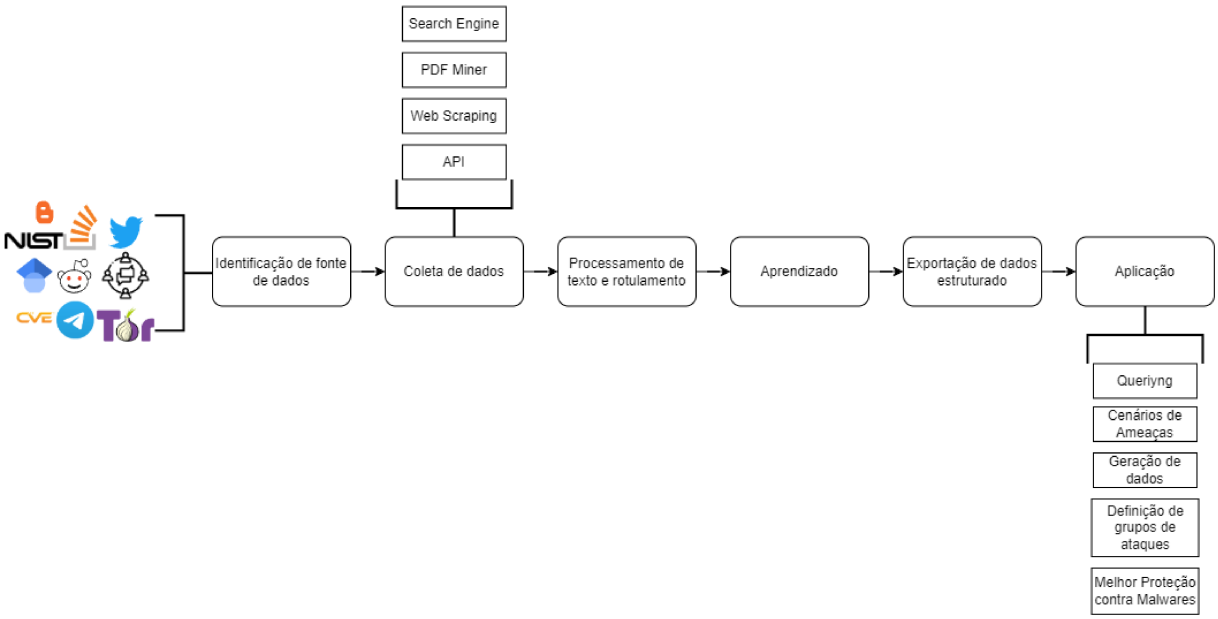

Figura 11 – *Pipeline* proposto para extração de CTI. Adaptado de (RAHMAN; HEZAVEH; WILLIAMS, 2023)

A metodologia proposta fundamenta-se no ciclo de vida de CTI, estruturando-se em seis etapas: (i) identificação da fonte de dados, (ii) coleta, (iii) pré-processamento e rotulagem, (iv) aprendizado, (v) exportação para formatos estruturados e (vi) aplicação. Esta estrutura fundamenta-se na literatura da área, com exceção da exportação e aplicação, todos os passos foram recorrentes nos trabalhos analisados pela revisão de (RAHMAN; HEZAVEH; WILLIAMS, 2023).

Capítulo 3

Trabalhos Relacionados

Neste capítulo, apresentam-se os principais trabalhos relacionados a esta dissertação, com o objetivo de oferecer uma visão abrangente do estado da arte em extração de informações para CTI, a partir de fontes de dados não estruturadas, como redes sociais e fóruns da *Dark Web*. Essas abordagens partem do pressuposto de que tais ambientes podem revelar sinais de ameaças existentes e, em alguns casos, permitir a previsão de novos ataques cibernéticos antes de sua efetivação.

A Seção 3.1 foca nas diversas publicações na área de fontes de dados para CTI. Destacam-se, para os fins deste estudo, as pesquisas que investigaram de forma aprofundada o uso de redes sociais e fóruns especializados da *Dark Web* como insumos para modelos de IA. Tais estudos demonstram que é possível identificar indícios de ataques cibernéticos futuros a partir da análise de conteúdo textual nessas plataformas.

Na Seção 3.2, a vertente de estudos correlatos é direcionada à generalização do aprendizado de máquina. Serão apresentados trabalhos que exploram a transferência de aprendizado em textos, discutindo a aplicação de um modelo de IA, treinado e validado em uma base de dados, para realizar previsões sobre outra fonte distinta. Esta seção visa analisar os resultados e as limitações na generalização de modelos de detecção de ameaças.

Por fim, a Seção 3.3 é dedicada à Discussão, onde será realizada uma análise dos trabalhos revisados. Esta seção irá sintetizar os principais achados, identificar as lacunas de pesquisa e posicionar a contribuição desta dissertação no contexto dos trabalhos correlatos apresentados.

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

O trabalho conduzido por Rahman, Hezaveh e Williams (2023) apresenta uma abordagem abrangente, examinando os métodos de coleta empregados na atualidade, bem como as fontes utilizadas na análise. Ao final do estudo, contribui significativamente ao propor

uma arquitetura completa para um sistema de inteligência de ameaças cibernéticas. Já o estudo conduzido por (SARKAR et al., 2018) propõe uma abordagem inovadora para analisar os fóruns na *Dark Web*, visando antecipar problemas a partir da identificação de vulnerabilidades discutidas antes da ocorrência de ataques. Ao coletar dados provenientes de 53 fóruns ao longo de um período de 12 meses. O trabalho procurou prever incidentes de segurança em dois casos do mundo real. Uma das principais contribuições foi a técnica de mineração de rede que identifica "especialistas", cujas postagens contendo menções populares de vulnerabilidades atraem a atenção de outros usuários em períodos específicos.

O trabalho de Kühn, Wittorf e Reuter (2024) analisou a *Dark Web* como uma fonte de CTI, combinando a exploração manual de 144 fóruns, mercados negros e lojas de fornecedores com uma varredura semi-automatizada de mais de 1,1 milhão de endereços do domínio *.onion*. O estudo destaca a natureza dupla da *Dark Web* para o CTI que apesar de rica e subutilizada, sua automação é dificultada por barreiras técnicas. Constata-se que, fontes da *Surface Web*, como o *X/Twitter* ou banco de dados CVE, são mais acessíveis e abundantes, enquanto a *Dark Web* fornece dados de nicho e alerta precoce, porém mais difíceis de extrair.

Sapienza et al. (2017) propõem um *framework* de baixo custo computacional, visando monitorar mídias sociais, como *X/Twitter* e fóruns na *Dark Web*, para gerar alertas como avisos antecipados de ameaças cibernéticas. O sistema monitora os *feeds* de perfis relevantes ou com grau de ameaça elevado, procurando por postagens relacionadas a vulnerabilidades e tópicos relevantes de segurança cibernética, aplicando técnicas de mineração e refinamento de texto para, assim, gerar alertas antecipados para sistemas de segurança. Ao longo do artigo, os autores abordam como é o funcionamento, a arquitetura e a avaliação do sistema que eles propuseram. No final, são apresentados os alertas gerados a partir do sistema.

Na linha de *crawlers* e indexadores, destacam-se trabalhos que investigam formas de analisar e coletar informações da *Dark Web* por meio de fóruns e páginas, com o propósito de qualificar esses dados. O estudo de Fu, Abbasi e Chen (2010) apresenta uma metodologia voltada para a coleta de informações em uma variedade de fóruns. Os autores demonstram o algoritmo utilizado no indexador e no *crawler*, além de sua classificação, com foco na identificação de grupos extremistas. Ao final do trabalho, é apresentado um gráfico que exibe a distribuição desses grupos dentro dos domínios analisados, classificando-os com base nas palavras utilizadas nas perguntas e respostas dos fóruns.

Além do trabalho citado, Liakos et al. (2015) apresenta uma abordagem mais abrangente ao ampliar sua área de obtenção de dados na *Deep Web* e focar seu algoritmo de *crawler* em várias páginas com assuntos diversos, tentando extrair e verificar se o assunto abordado e conteúdo do respectivo endereço correspondem à premissa. Os autores se de-

dicaram em classificar os dados coletados de vários sites, verificando as informações com as referentes áreas relacionadas, assim verificando se o conteúdo está correlacionado com a página em questão. Ao final, são apresentadas imagens, gráficos e uma visão abrangente de cada categoria, exemplificando dados relevantes para o esporte e para a política.

Por fim, o estudo conduzido por Sun et al. (2023) analisou diversos trabalhos com o objetivo de investigar as metodologias atualmente empregadas na mineração de CTI e os algoritmos associados. Os autores abordam os passos comumente presentes no processo de mineração de CTI, incluindo a análise do cenário cibernético, extração de dados, destilação de informações relacionadas ao CTI, aquisição de conhecimento CTI, avaliação de desempenho e tomada de decisão.

3.2 Transferência de Aprendizado em Textos

Zhuang et al. (2020) buscam estabelecer conexões e sistematizar estudos existentes sobre aprendizagem por transferência, resumindo e interpretando de maneira abrangente os mecanismos e estratégias envolvidos na transferência de conhecimento. O objetivo principal é proporcionar aos pesquisadores uma compreensão mais aprofundada do estado atual das pesquisas e das ideias que estão sendo delineadas. Este artigo analisa mais de quarenta abordagens de transferência de aprendizagem, oferecendo uma visão abrangente das aplicações desse campo. Para demonstrar o desempenho de modelos de aprendizagem por transferência, o trabalho realiza experimentos com mais de vinte modelos representativos, contribuindo para uma análise mais aprofundada.

Além do trabalho citado, Weiss, Khoshgoftaar e Wang (2016) delineiam a aprendizagem por transferência, proporcionando percepções sobre as soluções existentes e examinando aplicações práticas dessa abordagem. Além disso, o documento oferece informações sobre *downloads* de *software* relacionados a diversas soluções de aprendizagem por transferência, seguidas por uma discussão sobre possíveis direções para futuros trabalhos de pesquisa. É relevante destacar que as soluções investigadas são independentes em relação ao tamanho dos dados, possibilitando sua aplicação em ambientes de *big data*.

Barbon e Akabane (2022) propõe um estudo de caso ao examinar um extenso conjunto de dados de diferentes contextos, incluindo conjuntos de dados de diferentes idiomas, como inglês e português brasileiro. O objetivo é analisar o desempenho de dois modelos (BERT e DistilBERT), os quais foram treinados do zero por meio da classificação de texto de notícias *online*. Os pesquisadores ajustaram tanto o BERT quanto o DistilBERT, utilizando a camada agregada como incorporação de texto. Em seguida, compararam os dois modelos usando diversos conjuntos de dados selecionados. Como resultado, destacaram o DistilBERT, que demonstrou ser quase 40% menor e cerca de 45% mais rápido que o BERT. Além disso, os pesquisadores salientaram que o DistilBERT preserva cerca de 96% das habilidades de compreensão linguística, tanto para o inglês quanto para o português

brasileiro, em conjuntos de dados balanceados.

Com o avanço dos modelos de linguagem pré-treinados, Howard e Ruder (2018) propuseram o *Universal Language Model Fine-tuning* (ULMFiT), considerado um dos primeiros métodos de aprendizagem por transferência para tarefas de processamento de linguagem natural, ao aplicar (*fine-tuning*) completo de modelos de linguagem de forma universal. O modelo ULMFiT usa uma rede *Long Short Term Memory* (LSTM) que foi pré-treinada em um grande conjunto de textos para aprender as estruturas e padrões gerais da linguagem. Depois disso, ele é ajustado para tarefas específicas, como classificação de sentimentos ou tópicos, usando técnicas que ajudam a aproveitar melhor o que já foi aprendido.

Devlin et al. (2019) apresentaram o BERT, um modelo que revolucionou o campo do processamento de linguagem natural ao introduzir uma forma eficaz de modelar o contexto das palavras de maneira bidirecional, considerando simultaneamente os *tokens* à esquerda e à direita por meio da arquitetura *Transformer*. O BERT foi pré-treinado em duas tarefas principais: a previsão de palavras mascaradas no meio das frases e a verificação de relação entre pares de sentenças. Graças a esse pré-treinamento geral, o modelo alcançou bons resultados em diversas tarefas de linguagem natural, exigindo apenas pequenas adaptações para cada aplicação específica.

Mais recentemente, os avanços em tecnologias quânticas abriram novas possibilidades para a aprendizagem por transferência em Processamento de Linguagem Natural. O trabalho de Chen e Lou (2025) propõe um modelo híbrido que combina componentes da computação clássica e quântica para tarefas de classificação de texto. A arquitetura integra mecanismos de atenção inspirados em redes neurais com *complex kernels* (CKSA) e circuitos quânticos parametrizados (*Parameterized Quantum Circuits* (PQC)), capazes de processar informações em espaços de alta dimensão. Os experimentos mostraram que o modelo alcança resultados promissores, especialmente em cenários com escassez de dados. Esses achados reforçam a ideia de que a combinação entre PLN e computação quântica representa uma direção para o avanço da área.

Nesse contexto, Narde et al. (2024) investiga o uso de técnicas baseadas e em transferência de aprendizagem e modelos *Transformer* clássicos para a classificação da veracidade de postagens no *Twitter* em português. Ao contrário de abordagens emergentes, o foco aqui recai sobre modelos consolidados, como BERT, *Robustly Optimized BERT Pretraining Approach* (RoBERTa), *Cross-lingual Language Model — RoBERTa* (XLM-R) e *ELECTRA*, com diferentes graus de especialização na língua portuguesa. A pesquisa demonstra que, mesmo com recursos computacionais acessíveis e uma base de dados limitada, é possível alcançar bons resultados, especialmente com o modelo BERT pré-treinado em português que superou os demais modelos com uma acurácia de 95.1%.

Nesse sentido, ampliando a discussão sobre transferência de aprendizagem em processamento de linguagem natural, o trabalho proposto por Alyafeai, AlShaibani e Ahmad (2020) apresenta uma análise das principais abordagens e modelos utilizados na área. O

trabalho revisa a evolução dos modelos de linguagem, destacando a transição dos modelos baseados em RNNs e CNNs para arquiteturas baseadas em atenção, especialmente os modelos *Transformer*, como *BERT* e seus derivados. A nomenclatura proposta pelos autores categoriza as estratégias de transferência em dois grandes grupos: transferência transdutiva, que envolve adaptação entre domínios ou idiomas, e transferência indutiva, que trata da transferência entre tarefas distintas.

Raffel et al. (2023) propuseram uma abordagem unificada para tarefas de processamento de linguagem natural por meio do modelo *Text-to-Text Transfer Transformer* (T5), que converte tarefas como tradução, sumarização e classificação em um único formato de entrada e saída textual. O estudo apresenta uma investigação empírica sobre técnicas de aprendizado por transferência em PLN, avaliando arquiteturas, objetivos de pré-treinamento, métodos de *fine-tuning* e escalabilidade do modelo. Para isso, os autores introduziram o corpus *Colossal Clean Crawled Corpus* (C4), contendo centenas de *gigabytes* de texto limpo em inglês, que foi utilizado no pré-treinamento dos modelos. A pesquisa destaca a importância da padronização e da escalabilidade no contexto do aprendizado por transferência aplicado à linguagem natural.

3.3 Discussão

A análise dos trabalhos revisados mostra que tanto as pesquisas voltadas à identificação de ameaças cibernéticas em redes sociais e fóruns da *Dark Web* (Seção 3.1), quanto aquelas relacionadas à aprendizagem por transferência aplicada a textos (Seção 3.2), oferecem contribuições significativas para o avanço da área de CTI. Contudo, observa-se que a integração entre essas duas vertentes ainda está em estágio inicial, o que revela uma lacuna e, ao mesmo tempo, uma oportunidade de contribuição que este trabalho busca explorar.

Os estudos que abordaram a *Dark Web* e as redes sociais destacaram-se pelas informações presentes nesses ambientes para a detecção de incidentes cibernéticos futuros, bem como pela constatação de que tais fontes apresentam obstáculos técnicos relevantes para sua exploração e automatização (RAHMAN; HEZAVEH; WILLIAMS, 2023; KüHN; WITTORF; REUTER, 2024). Em grande parte, as pesquisas existentes recorreram a abordagens baseadas em mineração manual ou no desenvolvimento de *crawlers* especializados, com foco na extração de dados em larga escala (FU; ABBASI; CHEN, 2010; LIAKOS et al., 2015). Diferentemente, a proposta desta dissertação busca avançar nesse cenário ao integrar a coleta de informações oriundas desses espaços com modelos de aprendizado supervisionado.

No que se refere à aprendizagem por transferência, a literatura revisada evidencia que modelos de linguagem pré-treinados, quando devidamente ajustados a domínios e idiomas específicos, alcançam resultados satisfatórios, mesmo em cenários caracterizados

pela escassez de dados rotulados (ZHUANG et al., 2020; DEVLIN et al., 2019; BARBON; AKABANE, 2022). Essa constatação é relevante para o presente trabalho, dado que a análise de fóruns e redes sociais em português brasileiro carece de grandes bases de dados classificadas. A utilização de arquiteturas como BERT, DistilBERT e suas variantes adaptadas à língua portuguesa tem se mostrado eficaz não apenas pela capacidade de compreensão semântica em contextos específicos, mas também pela sua viabilidade (NARDE et al., 2024).

Nesse contexto, a principal contribuição deste trabalho consiste justamente na integração dessas duas linhas de investigação: de um lado, a exploração de fóruns da *Dark Web* e *Surface Web* como fontes de dados relevantes para a geração de CTI; de outro, a aplicação de técnicas de aprendizagem por transferência para viabilizar a classificação automática e a análise desses conteúdos em língua portuguesa e inglesa. Essa integração, além de distinguir a presente proposta em relação aos trabalhos revisados, amplia as possibilidades de antecipação de ameaças cibernéticas em um cenário ainda pouco explorado pela literatura (HOWARD; RUDER, 2018; ALYAFEAI; ALSHAIBANI; AHMAD, 2020).

A fim de proporcionar uma visão dos estudos revisados e destacar o posicionamento desta dissertação, a Tabela 1 sintetiza os principais trabalhos relacionados, apresentando seus objetivos, fontes de dados, idiomas e abordagens técnicas.

Por fim, é importante destacar que ainda persistem desafios relacionados à dificuldade, à disparidade e ao volume dos dados extraídos desses ambientes. Dessa forma, esta dissertação contribui não apenas para o avanço do conhecimento na área de CTI, mas também para a consolidação de um campo de pesquisa que combina a mineração de dados em ambientes digitais de alto risco à aplicação de modelos de linguagem, com potencial impacto tanto acadêmico quanto prático.

Tabela 1 – Comparativo de Trabalhos Relacionados e Proposta de Dissertação

Referência	**Escopo / Domínio**	**Objetivo**	**Fonte de Dados**	**Idioma**	**Técnicas / Modelos**
(SARKAR et al., 2018)	Cibersegurança	Prever incidentes via análise de redes sociais.	*Dark Web*	Inglês	Mineração de rede e análise de especialistas.
(BARBON; AKABANE, 2022)	Notícias	Transferência de aprendizado em notícias.	*Surface Web*	PT-BR e Inglês	BERT e DistilBERT.
(NARDE et al., 2024)	Mídias Sociais	Classificar veracidade de *posts* no *Twitter*.	*Surface Web*	PT-BR	*Transformers* (BERT, XLM-R, etc.).
(FILHO, 2024)	Cibersegurança	Identificar *posts* maliciosos na *Dark Web*.	*Dark Web*	PT-BR	*LightGBM* com TF-IDF.
Esta Dissertação	Cibersegurança	Evolução temporal e generalização de modelos de CTI.	*Dark Web* e *Surface Web*	PT-BR e Inglês	LDA e *Transfer Learning*.

Capítulo 4

Materiais e Métodos

Neste capítulo, que detalha a metodologia da dissertação, será apresentado o processo completo de identificação de postagens maliciosas na *Dark Web* e na *Surface Web*. O trabalho está estruturado em sete seções que englobam cada uma das etapas do método, conforme ilustrado na Figura 12.

O capítulo inicia-se com a Seção 4.1, que apresenta as *Fontes de Dados* e o *Escopo da Pesquisa*, delimitando a escolha dos fóruns. Em seguida, a Seção 4.2 descreve a *Coleta de Dados*, detalhando o desenvolvimento e a funcionalidade do *crawler* em *Golang* para a extração de *posts* em diferentes fontes: *Hidden Answers, Deep Answers, Raddle, Hackonology, Legitcarders, Niflheim* e *Nulledbb*, abrangendo tanto a *Surface Web* quanto a *Dark Web*. A Seção 4.3 aborda o *Armazenamento* dos dados brutos coletados, garantindo sua persistência e integridade.

A Seção 4.4 é dedicada ao *Pré-Processamento*, que engloba as técnicas de limpeza e padronização, como tokenização, remoção de palavras de parada, *stemming* e lematização. Com os dados tratados, a Seção 4.5 descreve o processo de *Classificação*, que tem como objetivo identificar se os *posts* são maliciosos ou não maliciosos.

Na sequência, a Seção 4.6 foca na *Evolução de Ameaças*, utilizando técnicas como o LDA para estudar a evolução dos tópicos e tendências temporais nas plataformas. Por fim, a Seção 4.7 detalha a *Transferência de Aprendizado*, onde a generalização do modelo de classificação é avaliada por meio de sua aplicação em conjuntos de dados externos.

Dessa forma, este capítulo apresentará uma visão detalhada das etapas, ferramentas e metodologias adotadas neste trabalho, possibilitando uma compreensão mais consistente do processo de identificação de conteúdo malicioso na *Dark Web* e *Surface Web* a partir do uso de técnicas de aprendizado de máquina. Todas as implementações e recursos desenvolvidos estão disponíveis no *GitHub*[1].

[1] <https://github.com/miguelhbrito/mscproject>

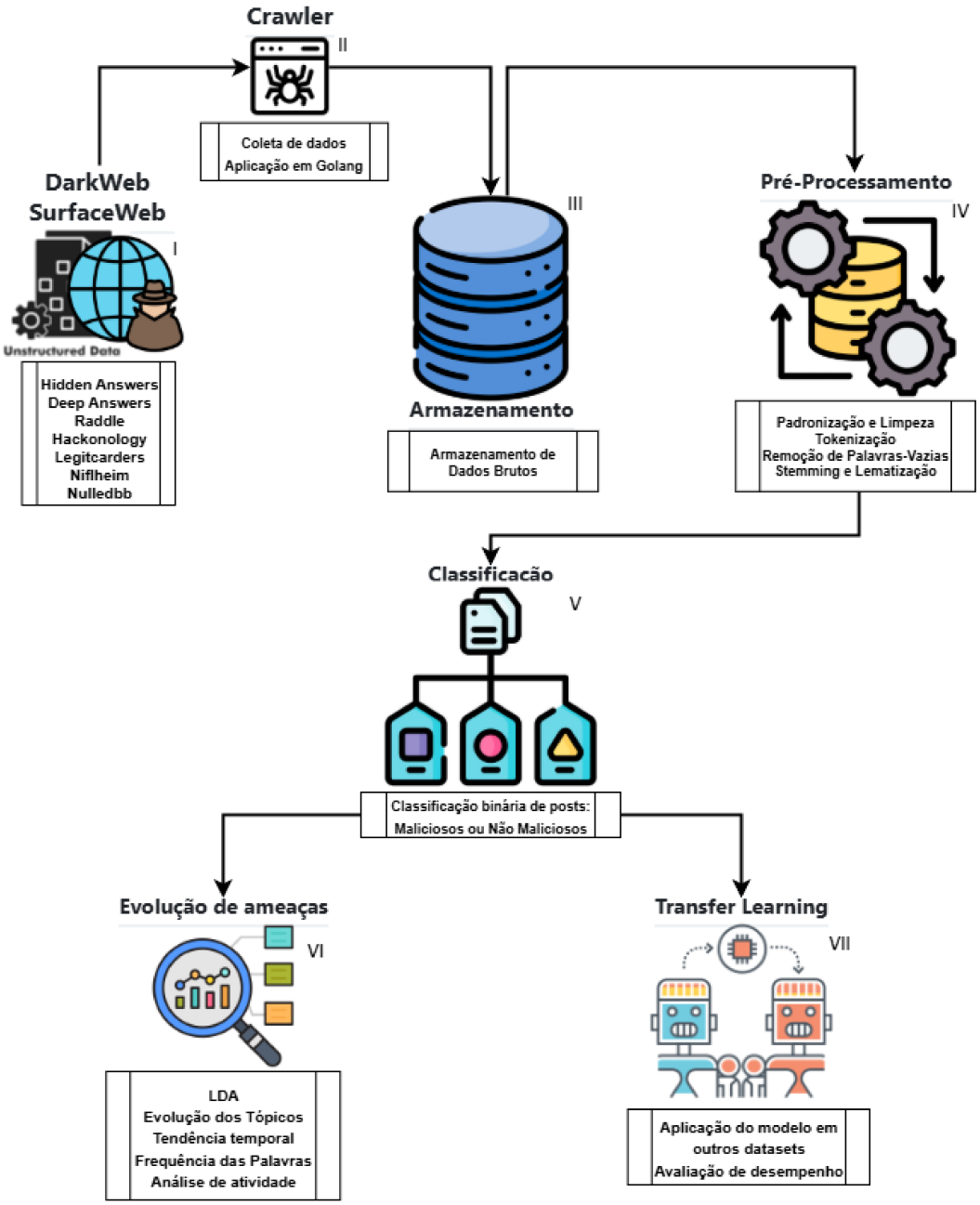

Figura 12 – Visão geral do trabalho (Fonte: O autor (2025))

4.1 Etapa I - Dados da *Surface Web* e *Dark Web*

A primeira etapa do método consiste na identificação dos dados em fóruns da *Surface Web* e da *Dark Web*, tanto em língua portuguesa quanto em língua inglesa. A seleção dos fóruns analisados foi feita manualmente e baseou-se na quantidade de perguntas e *posts* disponíveis, bem como em sua popularidade entre os usuários, especialmente no contexto da *Dark Web*.

Os endereços dos fóruns foram obtidos por meio de um indexador que, a partir de uma ou mais *URLs* fornecidas, realiza a navegação pelas páginas da *web* associadas, extraindo os *hiperlinks* nelas contidos. Para tal, utilizou-se a ferramenta *Miss Spider*,

desenvolvida em linguagem *Python* por um dos estudantes do NuSec (Núcleo de Segurança Cibernética da FACOM/UFU) e disponibilizada no repositório [2]. Esse processo ocorre de forma recursiva, possibilitando a identificação e a exploração contínua de novas páginas interligadas por meio desses *hiperlinks* (NAJORK, 2009).

Foram selecionados os seguintes fóruns da *Deep Web, Dark Web* e *Surface Web*:

- Respostas Ocultas e *Deep Answers*: versões em inglês e português, ambas pertencentes ao domínio da *Dark Web*;
- *Raddle, Hackonology, Legitcarders, Niflheim.World* e *Nulledbb*: disponíveis apenas em inglês, apenas o *Raddle* pertence ao domínio da *Dark Web*, enquanto os demais estão acessíveis na *Deep Web* ou *Surface Web*.

A escolha destes fóruns foi guiada por dois critérios principais. O primeiro critério foi a popularidade e a relevância apresentadas em pesquisas anteriores na área, indicando uma alta atividade e um volume significativo de conteúdo potencialmente malicioso. O segundo critério focou na viabilidade operacional, considerando a facilidade em encontrar e manter os endereços de acesso, sejam eles *onions* ou *URLs* diretas, funcionais e estáveis durante o período de coleta de dados.

Nesta etapa, a coleta direta dos dados ainda não foi feita. Em vez disso, definiu-se quais categorias de informações são de interesse para o estudo, de modo a orientar a etapa seguinte de extração. Assim, a Tabela 2 apresenta uma representação dos dados que foram considerados na análise.

Tabela 2 – Dados de interesse dos fóruns da *Surface Web* e *Dark Web* (Fonte: O autor (2025)).

Fóruns Hidden Answers e Deep Answers	
1	Título
2	Pergunta
3	Autor
4	Categoria
5	Tags
6	Votos
7	Respostas

Fórum Raddle	
1	Titulo
2	Post
3	Autor
4	Fórum(Categoria)
5	Votos
6	Comentarios

Foruns Hackonology, Legitcarders, Niflheim.World e Nulledbb	
1	Titulo
2	Conteúdo
3	Autor
4	Categoria
5	Respostas

[2] <https://github.com/miguelhbrito/mscproject/tree/master/miss-spider-and-crawler-master/miss_spider>

4.2 Etapa II - Coleta dos dados

A etapa seguinte da metodologia consistiu na coleta dos dados. Para isso, foram utilizados *crawler*, também conhecidos como *web crawlers* ou *scrapers*. Conforme já apresentado no Capítulo 2, os *crawlers* são ferramentas destinadas ao rastreamento e extração de informações na *web*. No contexto deste trabalho, foi desenvolvido um *crawler* específico para a coleta de dados nos fóruns selecionados, todo o código esta disponível no repositório do projeto[3].

O algoritmo proposto neste trabalho foi desenvolvido com o uso do *framework Colly*, em conjunto com a linguagem de programação *Golang*. A escolha do *Colly* justifica-se por sua elevada eficiência no rastreamento de páginas e pelo suporte nativo à concorrência proporcionado pela *Golang*, o que assegura um desempenho superior na coleta de grandes volumes de dados. O principal objetivo do algoritmo é extrair exclusivamente os trechos textuais relevantes contidos nas estruturas *div*. Para otimizar o processamento, foi empregado o paralelismo oferecido pelo próprio *framework*. O mapa contendo os caminhos das *URLs* referentes às perguntas e postagens dos fóruns está estruturado da seguinte forma:

```
URL_BASE + /index.php/ + índice
```

O valor do índice é incrementado a cada nova pergunta ou postagem feita no fórum, então, a primeira pergunta realizada no site levará o algarismo 1 no final do endereço. As páginas foram também filtradas por temas que sejam do interesse do projeto, desconsiderando perguntas ou postagens fora do escopo. O algoritmo empregado para a obtenção das perguntas e respostas tem a seguinte estrutura:

```
URL_BASE do Forum
for URL_BASE + /index.php/ + 1 to N
    if question or answer div
        question.append
for question index 1 to N
    save database
```

Para a comunicação com a *Dark Web*, foi utilizado um contêiner baseado na imagem *dperson/torproxy*[4]. Esse recurso compõe o Tor e o *Privoxy* (*proxy web* configurado para rotear através do Tor) em um contêiner *Docker*. A configuração foi realizada no arquivo docker-compose.yaml, conforme apresentado a seguir:

```
mytor:
 image: dperson/torproxy
 container_name: mytor
 ports:
```

[3] <https://github.com/miguelhbrito/mscproject>
[4] <https://github.com/dperson/torproxy>

```
    - "9050:9050"
    - "8118:8118"
  environment:
    - TORUSER=root
  networks:
    - local-network
  restart: on-failure
```

No caso, a utilização ocorreu da seguinte maneira:

```
socks5://mytor:9050
```

4.3 Etapa III - Armazenamento

Após a execução do *crawler* e a extração dos dados relevantes, os dados foram organizados em um banco de dados relacional, de modo a assegurar a integridade, consistência e a facilidade de acesso nas etapas subsequentes. Cada entrada corresponde a uma postagem ou resposta obtida nos fóruns.

Além do conteúdo textual, foram preservados metadados como identificador do fórum, autor, data de publicação e relações entre postagens, garantindo maior flexibilidade para análises posteriores.

A Figura 13 corresponde às tabelas relacionais dos fóruns *Hidden Answers* e *Deep Answers*. Já a Figura 14 apresenta a estrutura relacional referente ao fórum *Raddle*. Por fim, a Figura 15 reúne os esquemas das tabelas utilizadas para representar os fóruns *Hackonology*, *Legitcarders*, *Niflheim.World* e *Nulledbb*. Essas representações não refletem os dados brutos, mas sim as entidades, atributos e relações que compõem o repositório.

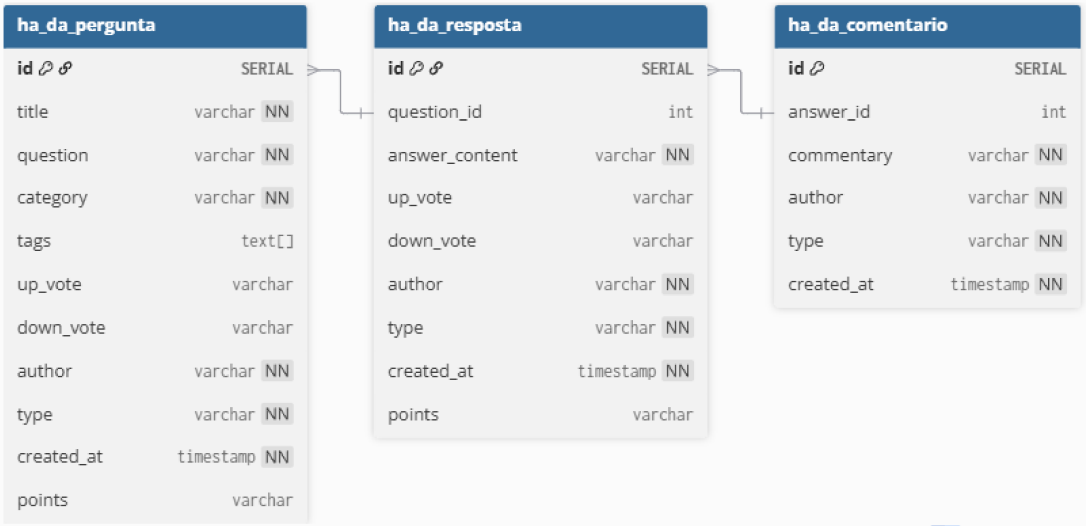

Figura 13 – Tabelas relacionais dos fóruns *Hidden Answers* e *Deep Answers* (Fonte: O autor (2025))

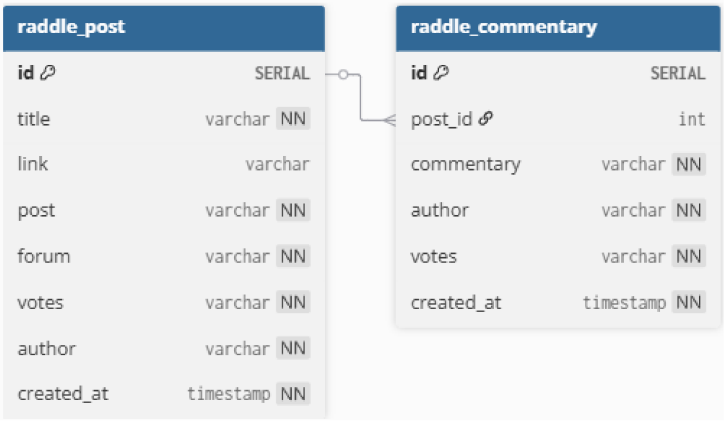

Figura 14 – Tabela relacional do fórum *Raddle* (Fonte: O autor (2025))

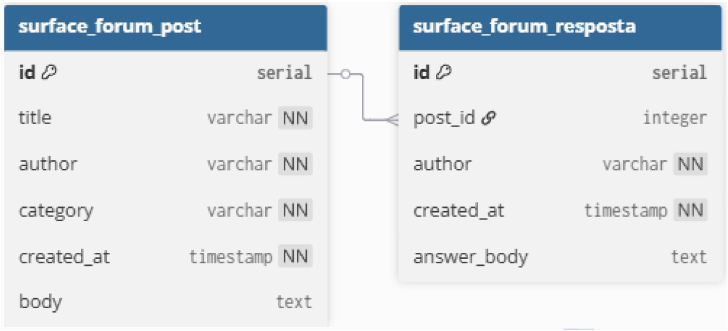

Figura 15 – Tabelas relacionais dos fóruns *Hackonology, Legitcarders, Niflheim.World* e *Nulledbb* (Fonte: O autor (2025))

O banco de dados utilizado para o armazenamento das informações foi o *PostgreSQL*, em sua versão mais recente disponível à época da pesquisa. A execução deu-se por meio de contêiner *Docker*, cuja configuração foi definida no arquivo `docker-compose.yaml`, conforme apresentado a seguir:

```
postgresdb:
    image: postgres:latest
    container_name: postgresdb_container
    ports:
      - "5432:5432"
    environment:
      - POSTGRES_DB=database
      - POSTGRES_USER=user
      - POSTGRES_PASSWORD=password
      - MAX_CONNECTIONS=300
    volumes:
      - pgdata:/var/lib/postgresql/data/
      - /logs/
    networks:
```

```
    - local-network
  restart: on-failure
```

Para o gerenciamento das migrações do banco de dados, foi utilizado o *framework golang-migrate*[5]. Foram definidos arquivos com os sufixos *up* e *down*, responsáveis por controlar a criação e a reversão das alterações, respectivamente. Além disso, utilizou-se um prefixo numérico, por exemplo, 01, para indicar a ordem e possibilitar o versionamento das migrações.

A título de ilustração, um conjunto de arquivos de migração pode assumir a seguinte forma:

```
01_create_table.up.sql
01_create_table.down.sql
02_insert_initial_data.up.sql
02_insert_initial_data.down.sql
```

4.4 Etapa IV - Pré processamento

O pré-processamento teve como objetivo limpar e organizar os dados, abrangendo tarefas relevantes para mineração de texto, como (HICKMAN et al., 2022): (i) a remoção de atributos irrelevantes para a análise, como identificadores internos ou campos nulos; (ii) a padronização dos nomes dos atributos; (iii) a unificação das postagens em um único conjunto de dados; e (iv) a concatenação dos atributos de conteúdo relevantes para a modelagem, como texto principal, respostas e comentários. Essa etapa foi desenvolvida utilizando a biblioteca *pandas* do *Python*.

Foram aplicadas técnicas de mineração de textos, tais como a remoção de *stopwords* e termos irrelevantes. Conforme já apresentado no Capítulo 2, essa prática tem o objetivo de eliminar palavras que não contribuem para o significado da análise, preservando assim a integridade da informação procurada. Além da técnica mencionada, outros métodos foram aplicados, tais como:

- ❏ Normalização do texto: conversão para letras minúsculas e remoção de acentuação.
- ❏ Limpeza estrutural: remoção de HTML, links, espaços em branco, caracteres especiais, números e identificadores (*QuestionID*, *AnswerID*, *PostID*)).
- ❏ Redução de ruído: eliminação de sequências repetitivas de caracteres, como *kkkkkkkk* e *aaaaaa*.

Essa etapa foi realizada para garantir que os dados estivessem em condições adequadas para as fases seguintes. Nos fóruns, alguns atributos apresentam nomes diferentes, embora

[5] <https://github.com/golang-migrate/migrate>

possuam o mesmo significado. Por exemplo, na Tabela 2, nos fóruns *Hidden Answers* e *Deep Answers*, aparece o atributo Pergunta; no *Raddle*, o atributo correspondente é *Post*; e, nos demais fóruns, utiliza-se Conteúdo.

Além disso, há também as respostas e os comentários, conforme mostrado na Tabela 2, que igualmente precisam ser considerados. Dessa forma, o processo incluiu a seleção dos atributos de interesse para a análise, a padronização da nomenclatura desses atributos, a consolidação de todas as instâncias de Pergunta, Conteúdo e *Post* em um único conjunto de dados e a concatenação de alguns atributos do tipo texto em um novo atributo denominado *full_text*. A Tabela 3 expõe a definição final dos atributos do conjunto de dados estabelecida nesta fase de pré-processamento.

Tabela 3 – Definição dos atributos do conjunto de dados na etapa de pré-processamento (Fonte: O autor (2025)).

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

4.5 Etapa V - Classificação

É importante destacar que foi utilizado um modelo de aprendizado de máquina pré-treinado apresentado em (FILHO, 2024) para selecionar as postagens com alta probabilidade de conter conteúdo malicioso. O modelo foi treinado usando o algoritmo *Light Gradient Boosting Machine* (LightGBM) com TF-IDF. O modelo em questão atingiu acurácia de 94% em um conjunto de dados semelhante ao estudado neste trabalho[6]. Em linhas gerais, o modelo trabalha com palavras-chave, indicadores de comprometimento (IoC) e seleção manual para gerar um valor entre 0 e 1 para textos. Quanto mais próximo de 1 maior a chance do texto possui algum tipo de conteúdo malicioso.

O processo de desenvolvimento desse modelo envolveu, inicialmente, a rotulagem dos dados, etapa na qual as postagens foram classificadas como relevantes ou não relevantes, com base em IoCs, palavras-chave contextuais e análise manual. Em seguida, realizou-se a vetorização, convertendo os textos em representações numéricas por meio de técnicas como TF e TF-IDF, o que permitiu a identificação de padrões linguísticos significativos. Com os dados vetorizados, passou-se ao aprendizado de máquina, no qual diferentes algoritmos supervisionados foram avaliados até que o LightGBM apresentasse os melhores resultados em métricas como acurácia, precisão e medida F1. Por fim, a classificação de postagens

[6] <https://github.com/sebastiaoafilho/Malicious_Posts_Identification>

foi implementada com base no modelo treinado, possibilitando a identificação automática de publicações relevantes em fóruns da *Dark Web*, auxiliando na detecção proativa de ameaças cibernéticas.

Esse fluxo metodológico possibilitou que cada *post* analisado recebesse um valor de score entre 0 e 1, indicando a probabilidade de conter conteúdo malicioso. Para fins de decisão, adotou-se uma classificação binária, obtida a partir de um limiar pré-definido: valores iguais ou superiores ao limiar são rotulados como "relevantes", enquanto valores inferiores são considerados "não relevantes". A Tabela 4 apresenta um exemplo de saída do modelo, contendo o score atribuído e a respectiva classificação binária com o limiar a 0,5.

Tabela 4 – Exemplo de saída do modelo de classificação com limiar 0.5 (Fonte: O autor (2025)

Post	**Score**	**Classificação**
Venda de dados pessoais...[...]	0,87	1 (Relevante)

4.6 Etapa VI - Evolução de ameaças

A modelagem de tópicos foi realizada de acordo com as etapas do trabalho, em especial as etapas descritas nas Seções 4.4 e 4.5 e baseada na abordagem proposta por (TONG; ZHANG, 2016). A biblioteca *Gensim* (ŘEHůřEK, 2024) foi empregada para implementar o LDA sobre toda a base de dados coletada. Para tanto, as informações das colunas consideradas mais relevantes, como o título da postagem, a questão (ou corpo da postagem) e as respectivas respostas, foram unificadas em um único campo textual, o qual foi submetido às etapas de pré-processamento já apresentadas. O LDA foi treinado uma única vez sobre o conjunto completo de dados, gerando um conjunto de tópicos. Em seguida, cada postagem foi associada ao tópico de maior probabilidade e vinculada ao respectivo ano e trimestre de publicação. Para cada trimestre, contabilizou-se a quantidade de postagens atribuídas a cada tópico, permitindo a construção das séries temporais dos temas emergentes.

O algoritmo LDA foi aplicado considerando o trimestre de cada postagem, permitindo uma análise dinâmica da evolução das discussões nos fóruns. Para cada fórum, foram testadas diferentes quantidades de tópicos para aprimorar a coerência e a segmentação dos grupos. Os resultados foram avaliados com base em métricas de Coerência de Tópicos (*Topic Coherence*), que medem a similaridade semântica entre os termos mais representativos de cada tópico, garantindo que a segmentação refletisse os principais temas emergentes em cada trimestre (SYED; SPRUIT, 2017). Esse procedimento possibilitou uma análise mais refinada da variação dos tópicos ao longo do tempo, identificando mudanças nas

discussões e tendências dentro dos fóruns analisados.

A análise da evolução temporal foi fundamentada na distinção entre três conceitos-chave, adaptados à realidade dos dados coletados. A tendência foi definida como o aumento ou a queda constante no interesse por um tema ao longo dos anos, como a mudança gradual de discussões técnicas para práticas de venda de dados. A sazonalidade refere-se a comportamentos que se repetem em épocas específicas, identificada por picos de atividade em trimestres recorrentes que aparecem tanto na *Surface Web* quanto na *Dark Web*. Na prática, os picos foram identificados como os momentos com o maior número de postagens em um único trimestre. Já o crescimento foi definido pelo aumento na quantidade de mensagens de um período para o próximo. Por fim, o ruído representa as informações irrelevantes ou repetitivas que podem atrapalhar a análise; este foi removido durante o pré-processamento e por meio da filtragem de postagens com baixa probabilidade de conteúdo malicioso, garantindo que os resultados mostrassem apenas as discussões realmente importantes.

Essa dinâmica conceitual é operacionalizada tecnicamente por meio da variação estatística das postagens, visto que , neste estudo, o que evolui é a distribuição da relevância dos tópicos ao longo do tempo. Como foi utilizado o LDA de forma estática, treinado uma única vez sobre o conjunto completo de dados, a definição semântica das palavras-chave de cada tópico permanece fixa. A evolução é demonstrada pela variação na quantidade de postagens atribuídas a cada tema em diferentes trimestres, e não por uma alteração no significado intrínseco do tópico.

4.7 Etapa VII - Transferência de Aprendizado

Nesta etapa, busca-se avaliar a aplicabilidade do modelo proposto por meio da técnica de transferência de aprendizado para verificar sua utilidade em diferentes contextos de análise. Especificamente, o modelo foi aplicado tanto na identificação de postagens maliciosas em fóruns genéricos quanto na detecção de itens relacionados à cibersegurança em mercados da *Dark Web*. A Figura 16 ilustra o processo de transferência de aprendizado aplicado nesta pesquisa. Nela, o conjunto de dados rotulado e o modelo pré-treinado, parte superior, são provenientes do trabalho (FILHO, 2024). A parte inferior da figura representa a contribuição deste estudo, onde o referido modelo é reutilizado para inferência em novos conjuntos de dados de domínios similares, abrangendo postagens em fóruns genéricos e mercados da *Dark Web*, que serão detalhados a seguir.

Para a realização dos experimentos desta etapa, foram utilizados repositórios de dados abertos e públicos. A escolha por bases de acesso livre justifica-se pelo compromisso com a transparência e a reprodutibilidade da pesquisa, permitindo que os métodos e resultados aqui apresentados possam ser validados e auditados pela comunidade acadêmica sem restrições de acesso ou licenciamento. Nesse contexto, foram selecionados os seguintes

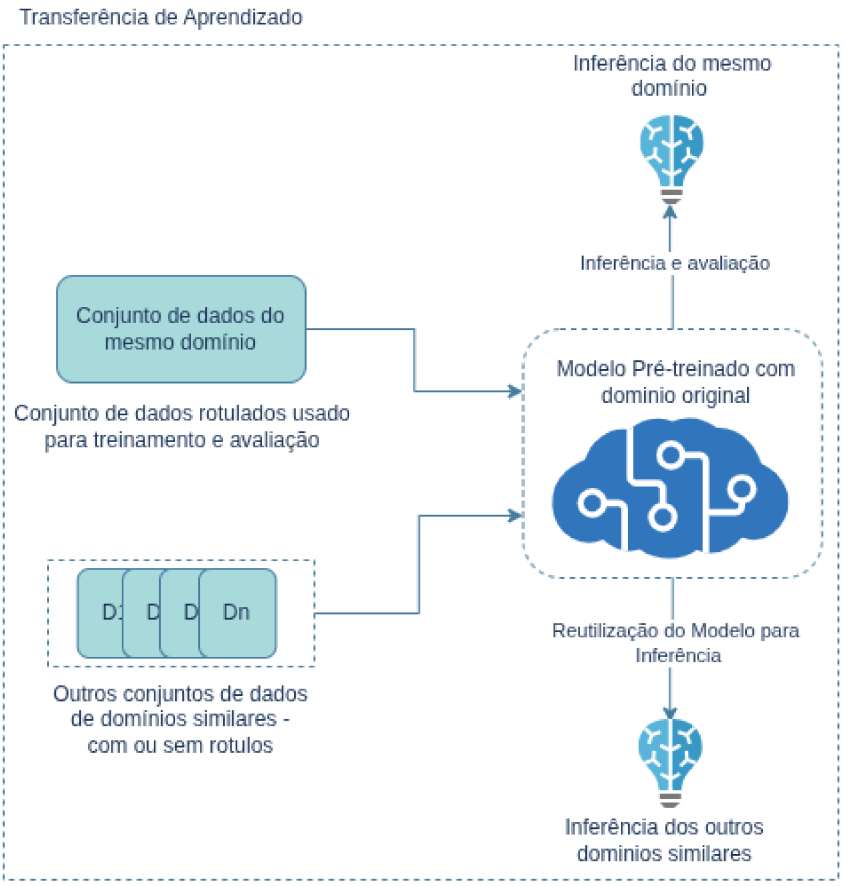

Figura 16 – Detalhamento da transferência de aprendizado (Fonte: O autor (2025))

conjuntos de dados:

- Bases 8kun e VGR [7]: Esta base consiste em uma base de texto não estruturado com discussões voltadas ao tema de videogames. Como os dados não são rotulados, o objetivo é avaliar a capacidade do modelo em identificar ameaças potenciais em dados não classificados, provenientes de domínios distintos da *web*.

- Base *Dark Topics*[8]: Este conjunto compreende mais de 3 milhões de postagens distribuídas em sete fóruns de discussão, abrangendo desde mercados criptográficos da *Dark Web* até subfóruns do *Reddit*. O conteúdo é textual e não rotulado. O propósito de sua utilização é aplicar o modelo para identificar e caracterizar conteúdos maliciosos, testando sua eficácia em um grande volume de dados originados de fóruns de mercado negro.

- Base *Darkweb Market (2021)*[9]: Este conjunto de dados contém anúncios de produtos e serviços de mercados da *Dark Web* em 2021 . Os dados possuem rótulos de categoria. Para avaliar a generalização do modelo neste domínio, foram adotadas

[7] <https://github.com/octokami/darknet_forum>
[8] <https://github.com/gayanku/darkweb_clearweb_darktopics>
[9] <https://github.com/Netsec-SJTU/darkweb-market-dataset-2021>

duas estratégias de rotulagem para estabelecer a "verdade de referência", palavra chave e LLM.

Primeiramente, buscou-se avaliar a capacidade do modelo em identificar postagens maliciosas em fóruns genéricos por meio de dois experimentos distintos, focados na transferência de aprendizado. No primeiro, aplicou-se o modelo sobre dados não rotulados provenientes de fóruns genéricos (Bases 8kun/VGR e Base *Dark Topics*), visando observar o comportamento do sistema em ambientes abertos. No segundo, o escopo foi direcionado ao mercado da *Dark Web*, utilizando-se bases de dados devidamente rotuladas para validar a precisão técnica da ferramenta.

Após a execução sobre essas bases, analisou-se a distribuição dos resultados com foco nos valores extremos. Essa análise verificou a coerência das classificações: se postagens maliciosas foram corretamente associadas a pontuações altas e postagens lícitas a valores baixos. Por fim, foram identificadas limitações e inconsistências que servirão de base para o desenvolvimento de trabalhos futuros.

Diferente da experimentação anterior com fóruns genéricos, também foi investigada a aplicabilidade do modelo proposto em um cenário que envolve o uso de dados rotulados, focado na detecção de itens de cibersegurança em mercados da *Dark Web*. Para essa etapa, utilizou-se especificamente a base de dados *Darkweb Market*. Como esse conjunto de dados não possuía rótulos prévios de classificação maliciosa, foi necessário realizar um processo de rotulagem sobre o campo categoria para estabelecer a verdade de referência, utilizando duas estratégias:

- Estratégia baseada em palavras-chave: consistiu na aplicação de um repertório de termos técnicos sobre o campo de categoria; caso o registro contivesse algum desses termos, era classificado como relacionado à cibersegurança (valor 1); caso contrário, era rotulado como não relacionado (valor 0).
- Estratégia baseada em LLM: utilizou o modelo *LLaMA 3* para realizar a classificação automatizada do mesmo campo de categoria, aproveitando sua capacidade de compreensão semântica para categorizar os itens entre relacionados (valor 1) ou não (valor 0) à cibersegurança.

A escolha pelo *LLaMA 3* justificou-se por ser um modelo de código aberto, permitindo a execução em hardware local, placa de vídeo própria. A coluna gerada por essa rotulagem foi definida como a "verdade de referência". Em seguida, as predições do modelo desenvolvido neste trabalho foram comparadas a essa referência por meio de matrizes de confusão, utilizando diferentes limiares de decisão (0,5 e 0,7) para avaliar o desempenho e a capacidade de generalização via transferência de aprendizado.

Prompt utilizado:
Você é um especialista em segurança cibernética.

Analise a seguinte categoria e responda **APENAS com o número**:
– 1 se ela estiver relacionada a segurança cibernética, *hacking*, crimes virtuais, fraudes digitais ou temas semelhantes.
– 0 caso contrário.

Para possibilitar a execução dos experimentos e o acesso ao modelo *LLaMA 3 (LLM)*, foi utilizada a configuração apresentada a seguir, implementada em um arquivo *Docker Compose*.

Listing 4.1 – Configuração em Docker Compose para execução do modelo LLaMA 3

```
version: "3.9"

services:
  ollama:
    image: ollama/ollama
    ports:
      - 11434:11434
    volumes:
      - ollama_models:/root/.ollama
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: all
              capabilities: [gpu]
    environment:
      - NVIDIA_VISIBLE_DEVICES=all

  ollama-webui:
    image: ghcr.io/ollama-webui/ollama-webui:main
    container_name: ollama-webui
    ports:
      - 3001:8080
    depends_on:
      - ollama
    environment:
      - OLLAMA_API_BASE_URL=http://ollama:11434/api
      - WEBUI_AUTH=False
```

```
volumes:
  ollama_models:
```

Fonte: Fonte: O autor (2025).

Dessa maneira, foram descritas as fontes de dados da *Surface* e *Dark Web*, o processo de coleta com *crawlers* em Go/Colly, a arquitetura de armazenamento relacional *(PostgreSQL/Docker)*, os procedimentos de pré-processamento, a estratégia de classificação e as análises de evolução de tópicos (LDA), além dos experimentos de transferência de aprendizado. Esses procedimentos garantem a reprodutibilidade metodológica do estudo. No capítulo seguinte, 5, são apresentados os resultados obtidos, incluindo métricas de desempenho do classificador, séries temporais de tópicos e evidências da capacidade de generalização observada nos experimentos de transferência de aprendizado.

Capítulo 5

Experimentos e Análise dos Resultados

Este capítulo apresenta os experimentos realizados e analisa os resultados obtidos a partir da metodologia descrita anteriormente no 4. A avaliação é organizada em duas dimensões principais.

Na primeira dimensão, a Seção 5.1, examina as evidências relacionadas à evolução de ameaças nos fóruns investigados. Para tal, avaliam-se: (i) a coerência e seleção do número de tópicos (5.1.1); (ii) a tendência temporal dos tópicos ao longo do período analisado (5.1.2); (iii) as palavras mais frequentes nos registros (5.1.3); e (iv) diferenças entre os ambientes da *Surface Web* e *Dark Web* (5.1.4). Essa análise busca identificar padrões emergentes, comportamentos recorrentes e indícios de atividade maliciosa ao longo do tempo.

Na segunda dimensão, tratada na 5.2, são apresentados os experimentos com transferência de aprendizado, a fim de avaliar a capacidade de generalização do modelo. Os resultados são estruturados em três frentes: discussão de ameaças em fóruns externos disponibilizados no repositório *Oktokami* (subseção 5.2.1), análise de tópicos em bases do *DarkTopics* (subseção 5.2.2) e detecção de itens de cibersegurança em mercados clandestinos da *Dark Web* (subseção 5.2.3).

Por fim, a seção 5.3 sintetiza os principais achados e discute suas implicações para a aplicabilidade do modelo na identificação automática de conteúdo malicioso, destacando limitações e potenciais direções para trabalhos futuros.

5.1 Evolução de Ameaças

Nesta seção, apresentam-se e discutem-se os resultados referentes a caracterização e a dinâmica das discussões nos fóruns, fornecendo o embasamento necessário para responder às Questões de Pesquisas(QP), enunciadas no 1: QP1, que investiga de que maneira os tópicos de ameaças cibernéticas evoluíram temporalmente; QP2, que busca identificar padrões sazonais ou picos de atividade correlacionados a tópicos específicos ; e QP3,

focada nas principais distinções estruturais e temáticas entre os conteúdos da *Dark Web* e da *Surface Web*.

A análise está estruturada em cinco eixos fundamentais:(i) a caracterização do conjunto de dados analisado; (ii) a coerência dos tópicos identificados nos registros; (iii) as tendências temporais observadas ao longo do período de coleta; (iv) a frequência das palavras mais recorrentes; e (v) a comparação entre diferentes ambientes investigados.

Conforme mencionado no Capítulo 4, os dados analisados neste estudo foram coletados de três fóruns da *Dark Web* — *Hidden Answers*, *Deep Answers* e *Raddle* — e quatro da *Surface Web* — *Hacknology*, *Legitcarders*, *Niflheim.World* e *Nulledbb*, em português e inglês. O processo de coleta foi realizado utilizando um *crawler* desenvolvido em linguagem *Go*, conforme já descrito.

A Tabela 5 apresenta a quantidade total de *posts* coletados, o número de *posts* classificados como contendo conteúdo malicioso, o período em que foram publicados e o idioma das mensagens nos sete fóruns que compõem a base de dados. Vale destacar que os períodos de coleta não foram uniformes devido à inatividade de alguns fóruns, à mudança frequente de endereços, comum em domínios *.onion*, e à instabilidade dos servidores, também recorrente nesses domínios. O código e o conjunto de dados utilizados encontram-se disponíveis no repositório do projeto[1].

Tabela 5 – Quantidade total de *posts* e número de conteúdos maliciosos

Fórum	**Periodo**	**Posts**	**Posts Maliciosos (Prob. Alta)**	**Idioma**
Hidden Answers	Entre 10/2021 e 06/2024	7.202	1.492	Português
Hidden Answers	Entre 01/2022 e 09/2024	2.165	183	Inglês
Deep Answers	Entre 07/2021 e 09/2023	776	115	Português
Deep Answers	Entre 01/2021 e 09/2023	94	9	Inglês
Raddle	Entre 10/2017 e 04/2024	850	53	Inglês
Hacknology	Entre 01/2015 e 12/2024	5.958	277	Inglês
Legitcarders	Entre 01/2015 e 12/2024	4.507	750	Inglês
Niflheim.World	Entre 01/2015 e 12/2024	12.272	268	Inglês
Nulledbb	Entre 01/2015 e 12/2024	18.511	445	Inglês
	Total	52.335	3.592	

5.1.1 Coerência e Escolha do Número de Tópicos

Para garantir a extração de tópicos coesos e interpretáveis, foi utilizada a métrica de coerência (*coherence score*). Essa métrica avalia o grau de similaridade entre os termos de cada tópico, refletindo sua consistência semântica. Os valores variam de 0 a 1, em que pontuações mais altas indicam maior coerência, ou seja, maior proximidade semântica entre os termos, enquanto valores mais baixos indicam pouca ou nenhuma relação entre eles.

Para facilitar a compreensão dos dados e a geração de tópicos, os fóruns foram agrupados por ambiente: *Surface Web*, *Dark Web* (PT-BR) e *Dark Web* (EN-US). A Tabela 6

[1] https://github.com/miguelhbrito/mscproject

Tabela 6 – Coerência e número de tópicos escolhidos por ambiente

Ambiente	Número de Tópicos Escolhidos	Coerência (CV)
Surface Web (EN)	8	0.3052
Dark Web (PT-BR)	6	0.2981
Dark Web (EN-US)	8	0.3316

apresenta os valores de coerência obtidos para diferentes quantidades de tópicos em cada ambiente. Observou-se que, para a *Surface Web*, embora o valor máximo tenha ocorrido com 4 tópicos, optou-se por 8 tópicos, que apresentaram coerência estável e permitiram uma melhor separação dos assuntos. Na *Dark Web* (PT-BR), valores próximos de 0,30 indicaram estabilidade, sem ganhos significativos com o aumento do número de tópicos. Já na *Dark Web* (EN-US), a coerência máxima de 0.36 foi obtida com dois tópicos; porém, optou-se por manter oito tópicos para uma visualização mais detalhada e diversificada dos temas discutidos, permitindo uma análise mais abrangente.

A qualidade desses grupos de palavras foi medida através da coerência de tópicos, testando diferentes números de tópicos para cada ambiente a fim de encontrar o equilíbrio ideal: um modelo matematicamente consistente e que, ao mesmo tempo, conseguisse separar bem os diferentes tipos de conversas. Assim, a configuração final foi escolhida por representar melhor os temas reais discutidos em cada cenário, garantindo que a divisão dos assuntos fizesse sentido para a análise.

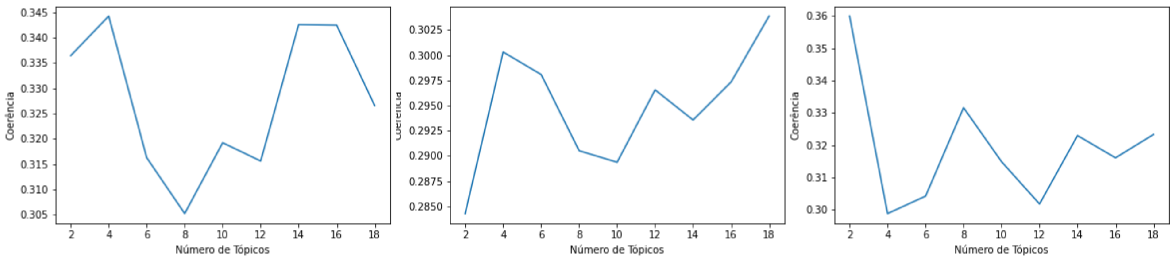

Figura 17 – Relação da Coerência com número de tópicos por ambientes. *Surface Web*, *Dark Web* (PT-BR) e *Dark Web* (EN-US) respectivamente (Fonte: O autor (2025)).

Com base nas análises de coerência, os tópicos extraídos em cada ambiente apresentam coesão suficiente para embasar as análises temporais, semânticas e comparativas desenvolvidas nas seções seguintes, as quais buscam responder às questões centrais deste estudo. A Figura 17 mostra os gráficos gerados para análise e escolha da quantidade de tópicos dos ambientes *Surface Web*, *Dark Web* (PT-BR) e *Dark Web* (EN-US) respectivamente.

5.1.2 Tendência Temporal dos Tópicos

Para compreender a evolução dos tópicos nos fóruns analisados, associou-se a ocorrência dos documentos classificados em cada tópico ao trimestre de publicação, resultando

em séries temporais por ambiente. Essa abordagem possibilita observar tendências, identificar picos de interesse e reconhecer potenciais reações a eventos externos.

É importante ressaltar que a comparabilidade semântica desses tópicos ao longo de todo o período (2015-2024) é garantida pelo treinamento global do modelo. Como o LDA foi aplicado de forma estática sobre o *corpus* completo, o significado de cada tópico permanece fixo por meio de seus índices, permitindo que a frequência de um tema em 2016 seja comparada diretamente com sua recorrência em 2023. Assim, a evolução observada refere-se à variação na distribuição e relevância das discussões, e não a uma mudança na definição dos tópicos em si.

A Figura 18 apresenta a evolução dos tópicos ao longo do tempo nos fóruns da *Dark Web* em português. A Tabela 7 mostra o conteúdo de cada um dos tópicos criados com o LDA. Os principais pontos identificados na análise são apresentados na Tabela 8.

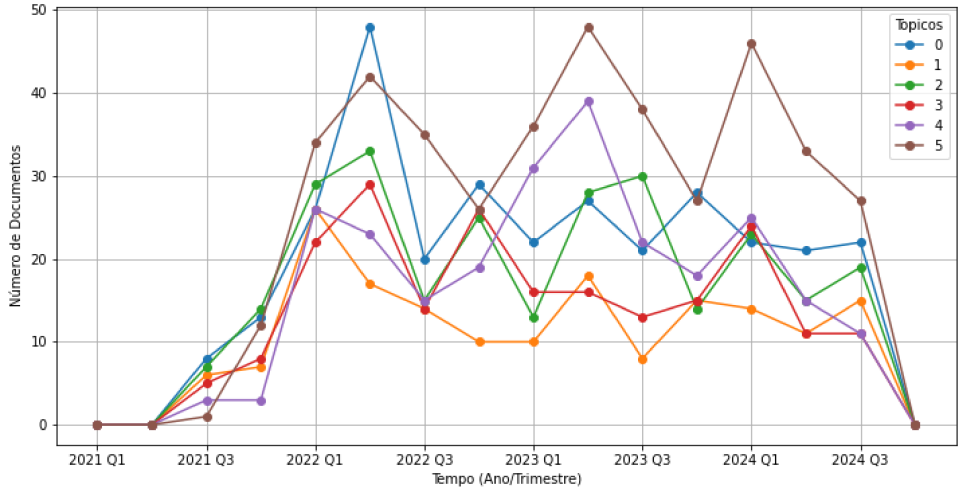

Figura 18 – Tendências dos tópicos ao longo do tempo *Dark Web* em português (Fonte: O autor (2025)).

Tabela 7 – Tópicos gerados com LDA - *Dark Web* em português (Fonte: O autor (2025)).

Tópico	**Palavras-chave**
0	site, hacking, programacao, tipo, kali, hackear, sites, seguranca, linux, aprender
1	rede, dados, wifi, informacoes, site, maquina, senha, linux, acesso, kali
2	link, virus, arquivo, tor, boa, windows, vitima, hacking, dados, curso
3	hacking, hacker, aprender, curso, dados, linux, celular, senha, cursos, links
4	pessoa, senha, site, social, pessoas, dados, faz, engenharia, facil, informacoes
5	dados, cpf, nome, conta, pessoa, telegram, numero, informacoes, banco, site

Na *Dark Web* em português, observou-se um crescimento do Tópico 5 — relacionado à comercialização de dados pessoais — entre o primeiro trimestre de 2022 e o segundo trimestre de 2023, sugerindo um aumento nas práticas ilegais durante esse período ou algum interesse por vazamento de informação. Por outro lado, o Tópico 0, de natureza

Tabela 8 – Resumo da evolução dos tópicos ao longo do tempo da *Dark Web* em português (Fonte: O autor (2025)).

Tópico 5	(*dados, cpf, nome, conta, pessoa, telegram, número, informações, banco, site*): apresenta crescimento constante do primeiro trimestre de 2022 até o segundo trimestre de 2023, depois estabiliza.
Tópico 0	(*site, hacking, programacao, tipo, kali, hackear, sites, seguranca, linux, aprender*): tem pico inicial e decresce, indicando perda de interesse.
Geral (consolidado)	Pico geral de atividade no segundo trimestre de 2023.

técnica, atingiu seu auge nos trimestres iniciais e apresentou queda ao longo do tempo, possivelmente indicando uma migração de interesse.

A Figura 19 apresenta a evolução dos tópicos ao longo do tempo em fóruns da *Dark Web* em inglês. Todos os tópicos gerados são mostrados na Tabela 9 e os principais pontos identificados na análise são apresentados na Tabela 10.

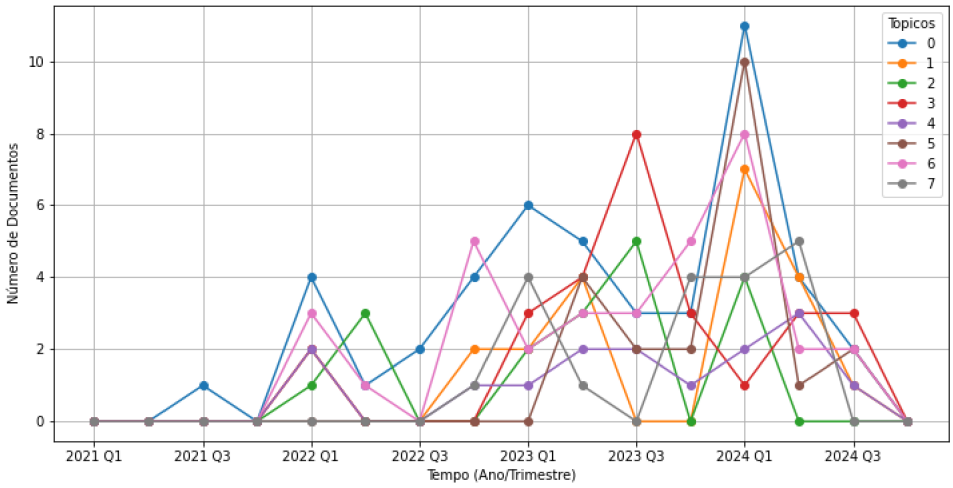

Figura 19 – Tendências dos tópicos ao longo do tempo - *Dark Web* em inglês (Fonte: O autor (2025)).

Tabela 9 – Tópicos gerados com LDA e suas palavras-chave - *Dark Web* em inglês (Fonte: O autor (2025)).

Tópico	Palavras-chave
0	phishing, email, account, brute, social, password, information, force, tools, phone
1	org, know, good, help, find, hacker, would, way, need, net
2	new, hack, github, linux, tools, hello, looking, programs, used, hacker
3	link, data, information, access, need, website, leak, good, tool
4	app, data, tool, safe, device, access, key, url, rat, leaked
5	learn, hacking, hacker, language, programming, linux, go, work, learning, start
6	learn, hacking, hack, want, need, website, social, sql, first, websites
7	contact, hacking, wifi, want, need, target, attack, telegram, learn, ddos

Tabela 10 – Resumo da evolução dos tópicos ao longo do tempo (Fonte: O autor (2025)).

Tópico 5	(*learn, hacking, hacker, language, programming, linux, go, work, learning, start*): cresce no segundo trimestre de 2023 até o primeiro trimestre de 2024, indicando foco em conteúdo educacional.
Tópico 0	(*phishing, email, account, brute, social, password, information, force, tools, phone*): mantém presença ao longo de todo o período, com picos no terceiro trimestre de 2023 e primeiro trimestre de 2024.
Geral (consolidado)	Tendência geral de aumento de *posts* em 2023 e início de 2024.

Na *Dark Web* em inglês, observa-se um crescimento expressivo nos tópicos relacionados ao aprendizado e ao desenvolvimento de habilidades técnicas (Tópico 5), refletindo a atuação de uma comunidade voltada à construção e disseminação de conhecimento ofensivo. Por outro lado, o Tópico 0, com foco em *phishing* e comercialização de credenciais, apresenta picos consistentes de atividade, alinhando-se a padrões sazonais previamente documentados, como os observados no final de ano — períodos frequentemente associados ao aumento de ataques e fraudes digitais.

O ambiente da *Surface Web* apresenta um comportamento mais instável, com picos em trimestres específicos, como o segundo trimestre de 2022 e o quarto trimestre de 2023. Os tópicos mais ativos — relacionados a ferramentas, *malware* e venda de acessos — parecem seguir relação à divulgação de ferramentas em plataformas públicas. A intensidade das discussões, no entanto, é consideravelmente menor do que nos fóruns da *Dark Web* devido à menor quantidade de postagens disponíveis antes de 2020. A Figura 20 apresenta a evolução dos tópicos ao longo do tempo em fóruns da *Surface Web*, e todos os tópicos gerados são mostrados na Tabela 11. Os principais pontos identificados na análise são apresentados na Tabela 12.

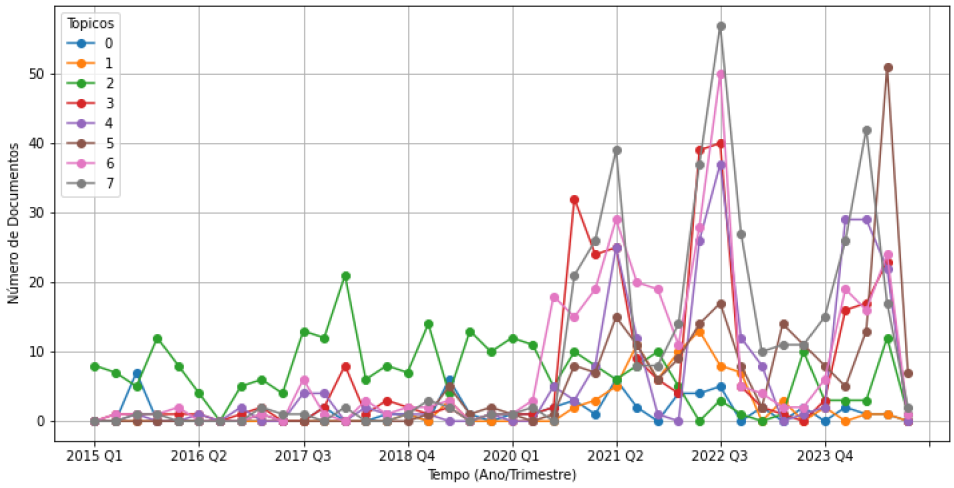

Figura 20 – Tendências dos tópicos ao longo do tempo - *Surface Web* em inglês (Fonte: O autor (2025)).

Em resumo, os três ambientes analisados demonstram padrões distintos. A *Dark*

Tabela 11 – Tópicos gerados com LDA - *Surface Web* em inglês (Fonte: O autor (2025)).

Tópico	Palavras-chave
0	onion, may, sql, exploit, tools, x, system, users, stealer, phisher
1	vzlom, hacking, hack, vkontakte, vzlomat, mail, order, zakazat, odnoklassniki, ru
2	login, nulledbb, register, locked, thread, thanks, wrote, leak, please, database
3	malware, data, server, network, attack, system, using, security, information, used
4	tool, target, using, exploit, password, github, vulnerability, code, web, linux
5	bank, email, account, card, logs, sell, us, cc, access, credit
6	password, file, download, hash, proxy, mega, free, files, using, click
7	data, world, security, attack, attacks, information, proxy, network, tools, ddos

Tabela 12 – Resumo da evolução dos tópicos ao longo do tempo *Surface Web* (Fonte: O autor (2025)).

Tópico 3	(*malware, data, server, network, attack, system, using, security, information, used*): apresenta crescimento expressivo entre o quarto trimestre de 2021 e terceiro trimestre de 2022, relacionado a malware.
Tópico 5	(*bank, email, account, card, logs, sell, us, cc, access, credit*): cresce no terceiro trimestre de 2023, com temática associada a credenciais e cartões.
Geral (consolidado)	Crescimento mais moderado, porém com picos no segundo trimestre de 2022 e quarto trimestre de 2023. Menor intensidade em comparação à *Dark Web*.

Web em português apresenta um amadurecimento dos temas, com transição de tópicos técnicos para práticas criminosas. A versão em inglês mantém uma base educacional, enquanto a *Surface Web* responde a tendências externas relacionadas a ataques, *malware* e vazamento de dados. Os picos temáticos coincidem no segundo trimestre de 2022 e no terceiro trimestre de 2023 nos três ambientes, sugerindo uma influência de eventos reais na dinâmica das discussões ou até mesmo intercâmbio de informações entre ambientes.

5.1.3 Palavras mais frequentes

A análise de frequência das palavras, ignorando as *stopwords*, nas postagens com alta relevância maliciosa, permite validar os tópicos previamente identificados por meio de LDA e revela padrões distintos de linguagem, refletindo tanto o conteúdo quanto o comportamento das comunidades. A Tabela 13 apresenta os padrões de linguagem que ajudam os tópicos discutidos na seção anterior, validando a correta extração dos temas. Enquanto a *Surface Web* mistura termos técnicos com ferramentas e senhas, a *Dark Web* brasileira foca mais na exposição de dados e ataques. Por outro lado, a *Dark Web* em inglês concentra-se em instruções e ferramentas para a aplicação de ataques ofensivos.

Tabela 13 – Palavras mais frequentes por ambiente (Fonte: O autor (2025)).

Ambiente	**Palavras mais frequentes (Top 10)**
Surface Web	hacking, vzlom, links, password, data, email, file, security, login, account
Dark Web PT-BR	dados, site, hacking, senha, pessoa, nome, conta, informacoes, acesso, cpf
Dark Web EN-US	hacking, learn, need, want, hack, tools, phishing, access, information, information

5.1.4 Comparação entre ambientes

Com o objetivo de identificar diferenças estruturais entre a *Surface Web* e a *Dark Web*, foi realizada uma análise comparativa considerando os tópicos mais frequentes, o foco em aprendizado, a presença de dados pessoais, a postura das comunidades e a variação temporal das discussões. A Tabela 14 sintetiza esses elementos, permitindo observar como cada ambiente se organiza sobre os temas específicos e como tais características refletem os perfis de uso e interação de suas comunidades.

Tabela 14 – Comparação entre Surface Web, Dark Web PT-BR e EN-US (Fonte: O autor (2025)).

Elemento	**Surface Web**	**Dark Web PT-BR**	**Dark Web EN-US**
Tópicos mais frequentes	T6 (317 docs): Tabela 11 T7 (389 docs): Tabela 11	T0 (307 docs): Tabela 7 T5 (405 docs): Tabela 7	T0 (46 docs): Tabela 9 T6 (34 docs): Tabela 9
Tópicos principais	Técnicas, ferramentas, ataques e vulnerabilidade.	Venda de dados pessoais e conteúdo ofensivo	Aprendizado, *phishing*, ferramentas e invasão.
Foco em aprendizado	Médio - diversos conteúdos instrutivos, tutoriais, guias de análise de *malware* e uso de *scanners*.	Médio - iniciante ou curioso, invadir redes sociais, manipular autenticação e sem estrutura acadêmica.	Alto - mais avançado e direto ao ponto, manipular autenticação , *phishing*, criação de *malware* ou uso de ferramentas.
Presença de dados pessoais	Média - alto compartilhamento de cartões de crédito	Muito alta - presença de dados pessoais, como nomes, senhas, contas e outras informações sensíveis	Alta - ênfase em técnicas
Postura da comunidade	Técnica – discussões sobre exploração, vulnerabilidades e tutoriais	Prática - voltada a pratica de ações maliciosas.	Técnica - porém com foco em invasão.
Analise Temporal	Alta variabilidade dos tópicos — as discussões flutuam entre vazamentos e ataques	Quantidade alta de perguntas e *posts* de tópicos relacionados à venda de dados pessoais	Tópicos técnicos como *phishing* e *hacking* se mantêm em quantidade contínua

A *Surface Web* concentra-se em conteúdos como senhas, arquivos, *malwares* e dados, com circulação de informações sensíveis, como cartões de crédito e credenciais vazadas. O foco no aprendizado é intermediário, dada a quantidade de *posts* com tutoriais e guias técnicos sobre ferramentas ofensivas e defensivas. Já a *Dark Web* brasileira destaca-se por conter dados pessoais, como *Cadastro de Pessoa Física* (CPF), nomes, números bancários e contas, indicando um ambiente voltado a ações maliciosas, com aprendizado também médio, porém desestruturado. Por outro lado, a *Dark Web* em inglês apresenta um ambiente mais técnico, com foco em *phishing*, *brute-force* e exploração de vulnerabilidades, além de um aprendizado mais avançado e prático, voltado à ofensiva. A presença de dados pessoais também é alta, embora, diferentemente da versão brasileira, o ambiente seja mais voltado ao aprendizado do que à comercialização direta. A identificação desses

picos de atividade e a evolução dos temas discutidos fornecem evidências para a resposta da QP1, evolução temporal, e da QP2, sazonalidade e picos. As disparidades observadas na Tabela 14, tanto em volume quanto em maturidade técnica das postagens, consolidam a resposta à QP3, evidenciando que a *Surface Web* e a *Dark Web* operam como ambientes distintos.

5.2 Transferência de Aprendizado

Esta seção dedica-se a responder a QP4: Qual é a eficácia da transferência de aprendizado na generalização de um modelo de classificação de posts maliciosos para domínios distintos? Para isso, busca-se avaliar a eficácia dessa técnica, que foi empregada para mensurar a capacidade de generalização do modelo de detecção de postagens maliciosas desenvolvido por Filho (2024) quando submetido a domínios distintos do seu treinamento original.

Esta seção detalha os resultados obtidos ao aplicar o modelo nos três conjuntos de dados distintos conforme detalhado na Seção 4.7:

a) **Bases 8kun e VGR**[2] visando à identificação de postagens maliciosas em fóruns.

b) **DarkTopics**[3], abrangendo discussões de tópicos diversos em ambientes da *Dark Web*.

c) **DarkMarketItems**[4], com foco na identificação de anúncios de itens de cibersegurança em mercados da *Dark Web*.

O objetivo desta seção é expor os resultados alcançados e discutir sua relevância, limitações e implicações para a aplicabilidade prática do modelo em diferentes cenários. A Tabela 15 apresenta os valores quantitativos, ou seja, o número total de postagens em cada uma das bases de dados utilizadas. Os conjuntos de dados, que incluem os textos completos, estão disponíveis no repositório[5].

Tabela 15 – Quantitativo de Postagens por Base de Dados.

Base de Dados	**Número de Postagens**
8kun	1.320
VGR	38.468
Dark Topics	1.846.564
Dark Market Items	363.843

[2] <https://github.com/octokami/darknet_forum>
[3] <https://github.com/gayanku/darkweb_clearweb_darktopics>
[4] <https://github.com/Netsec-SJTU/darkweb-market-dataset-2021>
[5] <https://github.com/miguelhbrito/mscdatasets>

5.2.1 Identificação de Postagens Maliciosas em Fóruns de Alta Informalidade (Bases 8kun e VGR)

O fórum *8kun* caracteriza-se como um *imageboard*, uma plataforma anônima de discussão em torno de postagens que combinam texto e imagem. Esse tipo de ambiente apresenta alto nível de ruído textual, expressões irônicas e uso recorrente de gírias, o que torna a identificação de postagens maliciosas uma tarefa difícil. Diante dessas particularidades, o modelo aplicado nesta pesquisa foi avaliado quanto à sua capacidade de distinguir postagens maliciosas e não maliciosas apenas a partir do conteúdo textual.

Foram analisados os resultados da aplicação do modelo sobre 1.320 postagens coletadas no *8kun*. É importante ressaltar que a base de dados não continha rótulos prévios e as probabilidades de comportamento malicioso foram integralmente atribuídas pelo modelo com base na análise dos textos. Dessa forma, as postagens foram ordenadas de acordo com os escores probabilísticos, permitindo observar tanto os exemplos mais prováveis de apresentarem comportamento malicioso quanto aqueles considerados menos relevantes nesse sentido.

No caso do fórum *8kun*, a Tabela 16 apresenta os casos com as maiores probabilidades atribuídas pelo modelo. Observa-se que, entre essas postagens, encontram-se os quatro únicos registros classificados como relevantes em toda a base. A análise das dez maiores probabilidades indica que o modelo atribuiu valores entre 0,26 e 0,55, concentrando-se nesse grupo postagens que expressam conteúdos sensíveis, como discussões sobre censura, menções a grupos de *hackers*, vazamentos de dados e referências a conflitos ideológicos ou políticos. Ressaltando que o limiar utilizado foi 0.5.

Em contrapartida, a Tabela 17 reúne os casos com as menores probabilidades, todos correspondentes a postagens não maliciosas. Variando entre 0,0044 e 0,0058, correspondem a postagens com temas neutros, como discussões sobre jogos, motores gráficos ou preferências pessoais em entretenimento digital. Nessas situações, o modelo atribuiu uma previsão binária igual a 0, indicando a ausência de comportamento malicioso.

Observa-se que o modelo apresentou uma tendência conservadora, uma vez que as probabilidades máximas não ultrapassam 0,55. Esse comportamento pode ser interpretado como coerente com o contexto de um *imageboard*, em que a linguagem é ambígua e informal. Dessa forma, o modelo evita fazer classificações precipitadas, preferindo agir com cautela para reduzir o número de erros ao identificar postagens como maliciosas.

Além disso, o modelo, neste caso, conseguiu diferenciar os contextos, associando maiores probabilidades a textos com indícios de comportamento malicioso, como *links* externos, menções a vazamentos e menores probabilidades a conteúdos descritivos ou informativos. Dado a natureza do fórum, isso sugere alta sensibilidade à termos ligados a segurança cibernética e *URLs*.

O fórum VGR pertence à *Surface Web*, o site é voltado à comunidade *gamer*, com foco em discussões sobre jogos, consoles e tendências do setor. Por se tratar de uma plataforma

Tabela 16 – Dez maiores probabilidades no fórum 8kun, sendo apenas quatro casos relevantes (Fonte: O autor (2025)).

texto	**probabilidade**	**relevância**	**previsão binária**
No, mister FBI agent. Maybe we will release our game and he will come out publicly. But, until then I want him to be able get out of this ride if he so desire. I've, probably, fucked up. It's possible to get my personal data, but not his.	0.5489	Média	1
8chan ran (albeit inconsistently) without DDOS protection when it had 10x the activity it had when cloudflare stopped hosting it. There was no reason to take the site down because cloudflare dropped them other then to censor the site to ensure they're no longer an enemy of big tech and thats exactly what they decided to do by choice [...]	0.5387	Média	1
in RE6 it was a 2-part virus. Part 1 was delivered through bottled water across the country while part 2 was an aerosol that activated it	0.5121	Média	1
I'm more inclined to believe what Jim'd say over some spud trying to say 8chan was runnung perfectly fine after getting ejected from cloudflare to people that were already both trying, and saying they intended to DDoS the site. Not that I'd let Jim shove his hand in my pants if he totally swears he's not gonna touch my dick, but yeah. [...]	0.5085	Média	1
Hey /agdg/ there's something I want to bounce by you guys, because I've been on the fence about it. I'm an anon who has undertaken the task of making a rewrite of feralpheonix's Uncommon Time, last year. As far as script, I've made a whole ton of progress on that front — about 70% done to be precise. Lately however [...]	0.4888	Média	0
Leak? This? https://URL.xyz/v/xPY4qzQMZmfqdd/weih.mp4	0.3076	Média	0
I've just been checking 8kun every couple weeks and informing people about the migration in case anyone checks the board after a long absence and doesn't know where the posters went. The site is barely functional, almost the entire userbase migrated elsewhere, and pretty much the only remaining activity is from spam [...]	0.2834	Baixa	0
I have absolutely no fucking comprehension of what happened on the 6th OY VEY REPUBLICANZ BAD CUZ JEW MEDIA SAY SOOOOOO THEY'RE NOT THE SAME AS MEEEEE WHITE NATIONALISM IS GOP Zero effort spam. Reported.	0.2798	Baixa	0
ete unspoilered porn but not spam from this faggot	0.2799	Baixa	0
Ransomware group known as Ragnar Locker leaks CAPCOM data. Don't you "hate"leaks? Lots of personal information like NDAs, sales reports and employee information is in there. All in all they stole 1 TB of data but haven't released it all. It's likely they'll sell some of it since their main motive is money. https://URL/?068vV05uS2GCgqa [...]	0.2634	Baixa	0

Tabela 17 – Dez menores probabilidades no fórum 8kun (Fonte: O autor (2025))

texto	probabilidade	relevância	previsão binária
Any game that resembles work (any game that's not fun). Too much of that now. Do this boring work, unlock this fun feature. I bought a game to have fun, not to make my life difficult. I highly recommend trainers for PC games, I doubt I'll ever play another game without one.	0.0044	Baixa	0
The real question is: How come Rockstar doesn't fix the fucking game? Can't add retarded Saints Row vehicles ruining the game's artstyle?	0.0044	Baixa	0
doomers gtfoCommit suicide, you worthless piece of shit. Prove wrong a single thing he said. Your own goddamn autistic image proves you believe the same thing. You will never fight back against the ZOG. Ever. You already know he's right.	0.0050	Baixa	0
The character and weapon levels are locked behind account levels because thats how the progression works. You get to a certain account level and it raises your world tier and makes all the enemies in the world tougher and drop better loot. If you're finding all the enemies too easy you're probably just close to having your world tier go up thats all [...]	0.0052	Baixa	0
I think it was a totally unnecessary sex scene that was clearly only there to make trannies not feel bad for not passing by giving them a manly female video game character to identify with but to call it pornography is just stupid. The agenda behind it was political not sexual.	0.0055	Baixa	0
There were many servers who even had 18+ verification and disallowed people who were not from entering, but mojang also blacklisted those too for having this mod because they wanted it to be wiped out.	0.0055	Baixa	0
I wouldn't violate the TOS though if I were this dev. Makes more sense to do it as a legit statement about their movement and if it does get censored decry it as a violation of freedom of speech. If you do it properly it makes the games journalists look like they're opposing free speech in favour of a cause that is violent. It'd give us exactly what we wanted [...]	0.0056	Baixa	0
The engine takes about 1/4 (∼150MB) of the download. That's not nearly as bad as I thought. Godot 4.0 might be more viable when its released. Yeah, it will have Vulkan support which will hopefully make it a real competitor in the 3d graphics segment. Development will probably speed up now that they got money [...]	0.0058	Baixa	0
General shit which has bled into every aspect of society lately: Women in places where they simply wouldn't realistically belong, like as the head of a raider gang, going solo in a post-apocalyptic/dystopian world, or simply as a soldier in a war which women had no major combative involvement in. Online games banning people for saying bad words or forbidden [...]	0.0058	Baixa	0
games can be a great mediun for storytelling, but when your game its more a movie than a interactive experience you just throw out all potential of the medium.	0.0058	Baixa	0

Tabela 18 – Dez maiores probabilidades atribuídas pelo modelo no fórum VGR (Fonte: O autor (2025)

Texto	**Probabilidade**	**Relevância**	**Previsão binária**
I didn't know that. Maybe they have a backdoor agreement with retail to help draw in customers during the week when business is less. Movies don't go through retail so that's why it's always released on weekends when people have more time and get that big opening night for ratings [...]	0.8997	Alta	1
Yeah I heard that too. I especially believe it after Bandersnatch. And who knows, it could have been a pitch for a game, or maybe even a hint that one could be coming. And I could see him providing a lot to a video game world.	0.8996	Alta	1
Charlie Brooker's really into video games, so I'm sure if someone pitched the idea of a game to him with Black Mirror branding he'd be able to lend a lot of creative direction to it. I'm sure the thought has crossed his mind [...]	0.8958	Alta	1
If anyone can really make The Black Mirror happen, it's definitely going to be Charlie Brooker. I think if the right pressure points are pushed, it's definitely going to happen.	0.8935	Alta	1
Hackers will never stop hacking, so try as much as possible to protect your privacy, identification and security protocols. I personally don't give out my personal information online so easily. Do you use a different email for social media to your private life? [...]	0.8628	Alta	1
Doing that works very well. I have at least two friends that used the support channel to get their accounts back, but the hacker had already messed up their resources...	0.8228	Alta	1
Nothing is 100% secure; if there is enough incentive to hack, it will be hacked. That's why I don't blame those who are skeptical about leaving any trace of their data online.	0.8020	Alta	1
I think hackers are getting a bad name these days. Most corporations are the ones creating problems with hacking and security. Most of the time identity issues happen because corporations want your identity, not the hacker's [...]	0.7857	Alta	1
Hackers will never stop hacking, so try as much as possible to protect your privacy, identification and security protocols. I personally don't give out my personal information online so easily.	0.7731	Alta	1
Yes, you are right, some owners have installed a plugin known as NoCheatPlus which prevents users from hacking, cheating or exploiting their server. Hacked clients don't work on some multiplayer servers though, but may work on others [...]	0.7703	Alta	1

pública e de grande alcance, o VGR apresentou volume de conteúdo superior ao *8kun*, totalizando 38.468 postagens.

Assim como nas demais bases, as postagens do VGR não foram previamente rotuladas. As classificações foram atribuídas pelo modelo, que estimou a probabilidade de comportamento malicioso com base no conteúdo do texto. Essa metodologia permite avaliar como o modelo reage a textos provenientes de um ambiente aberto e moderado, diferente do contexto mais hostil e anônimo característico da *Dark Web*.

A Tabela 18 apresenta as postagens com as maiores probabilidades atribuídas pelo modelo no fórum VGR. Os valores variam entre 0,77 e 0,89, indicando confiança do modelo para essas postagens. Todas elas receberam previsão binária igual a 1, com o limiar de 0,5, sendo interpretadas como potencialmente maliciosas. A análise qualitativa dos textos revela que os conteúdos de maior probabilidade tratam de temas relacionados à segurança digital, privacidade, *hackers* e vulnerabilidades de sistemas. Esses resultados sugerem que o modelo identifica a presença de termos ligados à segurança cibernética como indicativos de risco, mesmo quando usados em contextos neutros ou informativos.

Em alguns casos, o modelo errou ao dar notas altas a postagens que só avisavam sobre segurança ou falavam de falhas e invasões já conhecidas em jogos e servidores. Essa sensibilidade alta ao vocabulário técnico faz com que o modelo superestime o risco em discussões legítimas.

A Tabela 19 apresenta as postagens com as menores probabilidades atribuídas pelo modelo no fórum VGR. Os valores variam entre 0,0029 e 0,0033. Todas as postagens receberam previsão binária igual a 0, com o limiar a 0,5 como ja descrito, ou seja, foram interpretadas como não maliciosas. A análise qualitativa mostra que os textos tratam de assuntos cotidianos da comunidade *gamer*, como nostalgia por jogos antigos, preferências de jogabilidade, experiências pessoais e opiniões sobre títulos ou plataformas, sem qualquer indício de discurso agressivo, técnico ou potencialmente ofensivo.

As mensagens de baixa probabilidade refletem o caráter social do espaço, onde os usuários compartilham lembranças, opiniões e preferências pessoais. A falta de palavras relacionadas à segurança, invasões, conflitos ou ideias radicais faz com que o modelo classifique essas postagens como de baixo risco.

Por outro lado, o fato de as probabilidades serem muito parecidas entre si mostra que o modelo costuma dar valores baixos de forma uniforme a postagens com tom leve e palavras relacionadas à segurança. Isso pode significar que o modelo tem dificuldade para perceber pequenas diferenças de sentido entre mensagens neutras, tratando todas como igualmente seguras.

Em resumo, a análise das menores probabilidades no fórum VGR mostra que o modelo consegue identificar situações de conversa sem indícios de comportamento suspeito. Isso indica que o modelo mantém um equilíbrio entre identificar riscos e evitar erros, reduzindo as chances de classificar como maliciosas postagens que usam uma linguagem amigável e

Tabela 19 – Dez menores probabilidades no fórum VGR (Fonte: O autor (2025)

texto	**probabilidade**	**relevância**	**previsão binária**
Remember those days when you or your friend got a brand new game and either they or you would to each other's house just to watch each other play the single player game. Don't let this distract you from the fact that Kurenai tried using genjutsu on Itachi. [...]	0.0029	Baixa	0
Because retro games are good. Great, Amazing of some of them. And somepeople love to experience games from all eras. it's about pick-up-and-play simplicity. I love old games becasue i can just sit down and play them. No cutscenes, no story, just gameplay and a solid challenge. Sometimes I just want to play a game. [...]	0.0031	Baixa	0
That's insane. Never knew these functions existed on certain Twitch channels. That definitely would drive a lot more engagement with the viewers of such streams and would see them coming back. Just the other day a friend of mine showed me a twitch stream where you can literally play one of the new Pokemon games by entering commands in the chat. [...]	0.0031	Baixa	0
[...] Yeah - It offers more realness and a better sense of connection unlike what we are familiar with now.	0.003212016878	Baixa	0
The Blackangel mentioned in another thread having an empty feeling sometimes in games after characters change a great deal from who they were. I've had this experience as well a few times. It can be true to life, but sad too. Have others felt the same way?	0.0032	Baixa	0
Care to give examples? I'm not sure how one can be alienated by true to life changes? [...]	0.0032	Baixa	0
Do you guys like to sometimes watch others play video games instead of you? Maybe you want to watch your favorite lets players, or maybe you just want to watch a playthrough for a game you don't ever intend to play. Do you ever do anything like that? Or do you prefer to play the game yourself and experience it by your own hands? [...]	0.0032	Baixa	0
I don't like when games give you a 50/50 pick. Pick this or that. Maybe give us more options instead of two that will decide the story for you. I think having more options to choose from, expands the story and gives you more to explore each playthrough too. [...]	0.0032	Baixa	0
Well I never get surprised by it, but I do sometimes game more than I should. But I do some games that I can work online at the same time while playing them. Like TFT for example, the auto-chess game, which you can play and write articles online or such. [...]	0.0032	Baixa	0
Movies are nice but I can only sit through 2 movies at most before I'm completely bored and I've never watched TV for more than 2-3 hours at a time. We're all different though. For me, as of late I've been watching a couple movies a night. Sometimes 3. Watching with friends. [...]	0.0033	Baixa	0

sem intenção negativa.

5.2.2 Avaliação de Sensibilidade em Discussões Técnicas de Segurança (Base *Dark Topics*)

Enquanto a Seção 5.2.1 avaliou o modelo frente à linguagem informal e ao ruído de fóruns fora do escopo de segurança, esta etapa busca testar a sua capacidade de discernimento em ambientes com termos técnicos. Para isso, utilizou-se o conjunto de dados *Dark Topics*, que compila discussões de fóruns da *Surface Web* e *Dark Web* voltados a temas de segurança, privacidade e atividades cibernéticas. O objetivo aqui é verificar se o modelo consegue distinguir postagens maliciosas de conversas técnicas ou informativas sobre cibersegurança, reduzindo a ocorrência de falsos positivos.

A Tabela 20 apresenta as dez postagens com as maiores probabilidades atribuídas pelo modelo. Os valores variam entre 0,96 e 0,97, indicando confiança nas classificações positivas. Todas as postagens receberam previsão binária igual a 1 com o limiar a 0,5, sendo interpretadas como maliciosas ou com potencial de risco elevado. A análise dos textos mostra que a maior parte das mensagens discute assuntos como *backdoors*, *malwares*, vulnerabilidades de sistemas e ferramentas de invasão, muitas vezes com descrições detalhadas sobre técnicas,*softwares* e procedimentos de exploração.

Esses resultados demonstram que o modelo é altamente sensível à presença de vocabulário técnico e a expressões associadas a atividades cibernéticas ofensivas. A identificação de termos como "*trojan*", "*exploit*", "*botnet*" e "*NSA backdoor*" leva o modelo a atribuir probabilidades elevadas, o que é coerente com seu objetivo de reconhecer potenciais indícios de comportamento suspeito. Contudo, também é possível observar que parte das postagens analisadas apresenta caráter informativo, descrevendo vulnerabilidades ou práticas de segurança sem, necessariamente, indicar intenções maliciosas.

Assim, a análise sugere que o modelo tende a reagir fortemente a conteúdos com alta densidade de termos técnicos relacionados à segurança da informação, independentemente do contexto. Essa característica é desejável em tarefas de triagem ou detecção inicial de risco, pois privilegia a sensibilidade do sistema. Entretanto, em aplicações práticas, pode demandar etapas adicionais de refinamento ou validação para reduzir falsos positivos.

A Tabela 21 apresenta as dez postagens que receberam as menores probabilidades de classificação positiva pelo modelo no conjunto *darktopics*. Os valores variam entre 0.0028 e 0.0032, refletindo confiança de que essas mensagens não apresentam características associadas a comportamento malicioso. Todas as amostras receberam previsão binária igual a 0 com o limiar a 0,5.

As postagens tratam de temas como experiências de compra, atrasos em entregas, trocas entre usuários, comentários sobre reputação de vendedores e desabafos pessoais. Esse tipo de discurso é comum em comunidades, mas não contém elementos que indiquem

Tabela 20 – Dez maiores probabilidades dos foruns *DarkTopics* (Fonte: O autor (2025)

texto	**probabilidade**	**relevância**	**previsao binaria**
Wouldn't take too much resources if done illegally. Like a backdoor trojan that installs a botnet and have many computers in on the attack including ones unrelated to DNM activity	0.9798	Alta	1
It's called a 'false-positive' and most antivirus and anti-spyware jump to conclusions, you'll find they will do with a majority of programs you can get out there like key generators and software patches. Because essentially they exploit a backdoor in software (trojan like behavior), change files, delete files etc. Only they don't do any damage direct to the end user. [...]	0.9791	Alta	1
[...] Make sure you are downloading Tor from the Tor Project website and not an alternative source. Check the URL as well. Alternatively you can run a signature check or verifying sha256 sums.	0.9791	Alta	1
Could be a targeted firmware update, could be an exploit that the nsa discovers, any number of things really. Do a search for "port 32764 Backdoor". This went undiscovered for years, and was only eventually noticed by accident. Open source does not imply more or better security (unless every end user is a security expert who compiles the code themselves). [...]	0.9761	Alta	1
what you need is citadel, which is a zeus fork, and also, which, fixes a critical zeus panel vulnerability.forgive me, this is a slightly outdated version. I seem to have deleted the up-to-date citadel from Trojanforge, and now the d/l link is down.[...]	0.9753	Alta	1
Here is a message received from user : "Agora...."four dots.Registered 1 day ago the profile is the Agora rules and they are sending a message letting people know that tor is outdated. Be warned!Please do not take any of the following information as legitimate it it just a spam message!It appears that you are using an outdated version of Tor Browser to access Agora. [...]	0.9691	Alta	1
http://www.theguardian.com/world/2013/oct/04/nsa-gchq-attack-tor-network-encryptionexplains the capabilities of NSA and GCHQ,they use special backdoor browser exploit links to track tor users true browsers line IE and crome, in windows and other non open source operating systems. [...]	0.9681	Alta	1
I talked to a friend who is a network engineer and he said it was a backdoor built into the system so the nsa could exploit it go I really hate the NSA even if it turns out not to be true	0.9649	Alta	1
Nah I only operate from a phone I can dispose of. Keep different passwords for your accounts always check over the address don't click links to get therw n NEVER KEEP your eggs in one basket I've a few wallets just incase but I always empty my localbitcoin wallet ASAP to another wallet then from there to agora n whatever is left over I send to another wallet. [...]	0.9636	Alta	1
indeterminable probability that windows has a backdoor, Uncle Sam ONLY uses windows. As in, on air force networks Linux(even ones running the navy exclusively developed distro) and Mac machines are not allowed. For the only OS the government is allowed to use(make no questions, it's because they don't want to be assed with managing their own security) [...]	0.9629	Alta	1

intenção criminosa, como oferta de serviços ilegais, exploração de vulnerabilidades ou incentivo a práticas de invasão.

A ausência de termos técnicos relacionados à segurança, exploração, ideologia extremista ou anonimato contribuiu para a atribuição de probabilidades muito baixas. O vocabulário dessas postagens tende a ser mais informal e conversacional, muitas vezes expressando frustração, humor ou relatos pessoais.

Além disso, observa-se que as probabilidades são muito próximas entre si, o que indica que o modelo tende a tratar postagens neutras de forma homogênea. Esse comportamento mostra que o modelo apresenta resultados com baixa variabilidade e as classifica como mensagens que não apresentam indícios de risco. Por outro lado, o fato de as probabilidades ficarem muito próximas entre si pode indicar uma certa dificuldade do modelo em perceber pequenas diferenças entre mensagens neutras, já que todas acabam recebendo valores muito parecidos.

De forma geral, a comparação entre as postagens com maiores e menores probabilidades evidencia a coerência do modelo em suas classificações. Enquanto os textos com alta pontuação concentram-se em discussões sobre segurança ofensiva, vulnerabilidades e práticas potencialmente ilícitas, aqueles com valores mínimos refletem interações cotidianas e neutras, sem qualquer traço de comportamento suspeito. Essa diferença entre os dois extremos mostra que o modelo consegue identificar diferentes padrões de linguagem.

5.2.3 Identificação de Itens de Cibersegurança em mercados da *Dark Web*

A base de dados *DarkMarketItems* contém anúncios extraídos de mercados da *Dark Web*. Cada item é descrito por título, categoria e demais informações textuais, permitindo a identificação de potenciais produtos e serviços ligados a atividades de cibersegurança, como ferramentas de invasão, credenciais comprometidas e serviços de ataque.

O objetivo deste experimento foi avaliar a transferência de aprendizado de um modelo treinado previamente em fóruns para outros dominios, verificando sua capacidade de identificar itens relevantes para cibersegurança. Essa análise foi conduzida em dois cenários complementares a rotulação dos *posts*, conforme detalhado na Seção 4.7:

- **Cenário 1 – estratégia baseada em palavras-chave:** o campo *category* do conjunto de dados foi processado utilizando o repertório de termos técnicos definido na metodologia. Esta abordagem permitiu separar itens associados à segurança daqueles não associados, gerando uma coluna binária (*rótulo*) de verdade de referência, em que valor 1 representa associação à segurança e 0 indica não associação. Essa estratégia forneceu uma tabela de verdade fundamental para avaliar a acurácia do modelo. Para complementar, o modelo foi testado usando os limiares de decisão 0,5 e 0,7. A escolha desses valores fundamenta-se na abordagem proposta por (FILHO,

Tabela 21 – Dez menores probabilidades dos foruns *DarkTopics* (Fonte: O autor (2025)

texto	**probabilidade**	**relevância**	**previsao binaria**
Boobear is most definitely a one of a kind vendor. I placed one of his first orders (third I think), about two weeks ago, and he has been nothing but excellent since. We ran into some unforseen issues with the first delivery in the mail, and it seems it got lost (through no fault of Boobear's), [...]	0.0028	Baixa	0
[...]Thank you....but yeah I think my days of raffling are over. I do them alot in real life and in real business and they are successful. I dont know what this guys issue is but I will just have to settle with not being able to please him. I'm ok with that. Thank you again for your understanding I dont know if you are a fellow vendor but you must be on some level to understand.	0.0028	Baixa	0
so why dont you just cash them out your self?CashKing wrote:I do.It makes no logical sense to sell this when you can simply go to an ATM and withdraw cash.Maybe you don't realize it, but I'm not selling CC+PIN. I'm only providing a list of vulnerable BINs. [...]	0.0029	Baixa	0
Not only did I receive a prompt response from Boosie but I was offered an immediate reship. Having been a customer for a long time this felt great, Boosie didn't even have to bat an eye at this review but he made it right. That speaks loudly to his character. [...]	0.0030	Baixa	0
I think I knew the answer before I posted here. Just needed to clear my head a bit on this and see what others would have to say. This is a subject I just cant bring up among my friends and co-workers. They are not even aware that this exists. [...]	0.0030	Baixa	0
I deposited BTC on Thursday night and they are still in limbo. I agree they should have put up a banner as soon as they knew things were faltering but to be honest I'd rather they fix the problem than put up banners. [...]	0.0030	Baixa	0
You have to understand that this is probably his first raffle as well. So he was kind of testing the water per se. I'm sure he knows exactly how to work a rtaffle now or maybe he figured out that he doesn't ever want the hassle of doing one ever again lol. [...]	0.0031	Baixa	0
To the person who PM'd me my name. Congratulations you have found me. Now, in order to make this aspainless and as to not just waste my life for nothing, I am going to propose on offer. 15k in escrow fundsby a reputable third party. Instead of sending someone to my house to kill me, you send the funds their,and I kill myself live on webcam.[...]	0.0032	Baixa	0
they arent in any orderI see im on it now, why is that?? Never been called that to my face but it makes me laugh the power some stupid little wankers get from writing shit about people they dont know...In my opinion you and one guy who is into only guns are the least clued up in life and defo still live with mum.....[..]	0.0032	Baixa	0
they arent in any orderI see im on it now, why is that?? Never been called that to my face but it makes me laugh the power some stupid little wankers get from writing shit about people they dont know...In my opinion you and one guy who is into only guns are the least clued up in life and defo still live with mum.....[...]	0.0032	Baixa	0

2024), permitindo avaliar o modelo tanto em sua configuração padrão (0,5) quanto em um cenário de maior seletividade (0,7), essencial para a validação da acurácia sob critérios mais rigorosos.

- ❑ **Cenário 2 – Estratégia baseada em LLM:** conforme mencionado no 4, o procedimento metodológico adotado empregou o modelo *LLaMA 3 (LLM)* na rotulagem dos itens da base de dados, atribuindo valor 1 aos relacionados à cibersegurança e valor 0 aos não relacionados. A coluna resultante dessa rotulagem foi considerada como a verdade de referência para a avaliação do modelo proposto. Em seguida, foi criada uma segunda coluna contendo a classificação gerada pelo próprio modelo desenvolvido por (FILHO, 2024). As duas colunas, referência e predição, foram então comparadas por meio da matriz de confusão, que permitiu mensurar o desempenho em termos de acertos e erros. Além disso, foram conduzidos experimentos com diferentes limiares de decisão (0,5 e 0,7), possibilitando uma análise comparativa do desempenho sob distintos critérios de classificação e fornecendo evidências sobre a capacidade de generalização do modelo por meio da técnica de transferência de aprendizado.

A base de dados *Darkweb Market* contem um total de 363.843 anúncios, abrangendo diferentes categorias de produtos e serviços. A análise concentrou-se em medir a capacidade do modelo em identificar itens de cibersegurança. A análise teve como objetivo avaliar o desempenho do modelo na identificação de itens relacionados à cibersegurança, utilizando como referência tanto as classificações geradas por um modelo de linguagem (LLM) quanto aquelas definidas por meio de regras baseadas em palavras-chave.

Para o limiar de 0,5 nos experimentos de palavras-chave (Figura 21), revela um volume expressivo de verdadeiros negativos (306.936), indicando que o modelo conseguiu identificar corretamente a maioria dos itens não relacionados à segurança. Ainda assim, há um número considerável de falsos negativos (43.073), o que sugere que parte dos anúncios relevantes pode ter sido subestimada pelo modelo.

Os valores consolidados na Tabela 22 reforçam essa observação: o modelo alcançou acurácia de 0,86, com desempenho na identificação de itens não relacionados à segurança (classe 0), obtendo precisão de 0,88 e revocação de 0,97. Já para os itens de cibersegurança (classe 1), a precisão foi de 0,41 e a revocação de 0,12, resultando em uma medida-F de 0,18. Esses resultados indicam que, embora o modelo consiga identificar corretamente a maior parte dos itens neutros, não relacionados a segurança, ainda apresenta limitação na recuperação de amostras realmente relevantes.

Essa proporção mostra que o modelo foi capaz de aplicar o conhecimento adquirido em fóruns a um novo domínio, identificando expressões e padrões de linguagem associados à segurança digital, mesmo quando o vocabulário utilizado era diferente. No entanto, a baixa revocação para a classe positiva evidencia que o modelo tende a adotar uma postura

conservadora, priorizando a redução de falsos positivos em detrimento da detecção de itens relevantes.

Tabela 22 – Métricas de desempenho do modelo no conjunto *DarkMarketItems* (estratégia baseada em palavras-chave, limiar 0.5).

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
0 (não relacionado)	0.88	0.97	0.92	315.158
1 (relacionado à segurança)	0.41	0.12	0.18	48.685
Acurácia		0.86		363.843
Média Macro	0.64	0.54	0.55	363.843
Média Ponderada	0.81	0.86	0.82	363.843

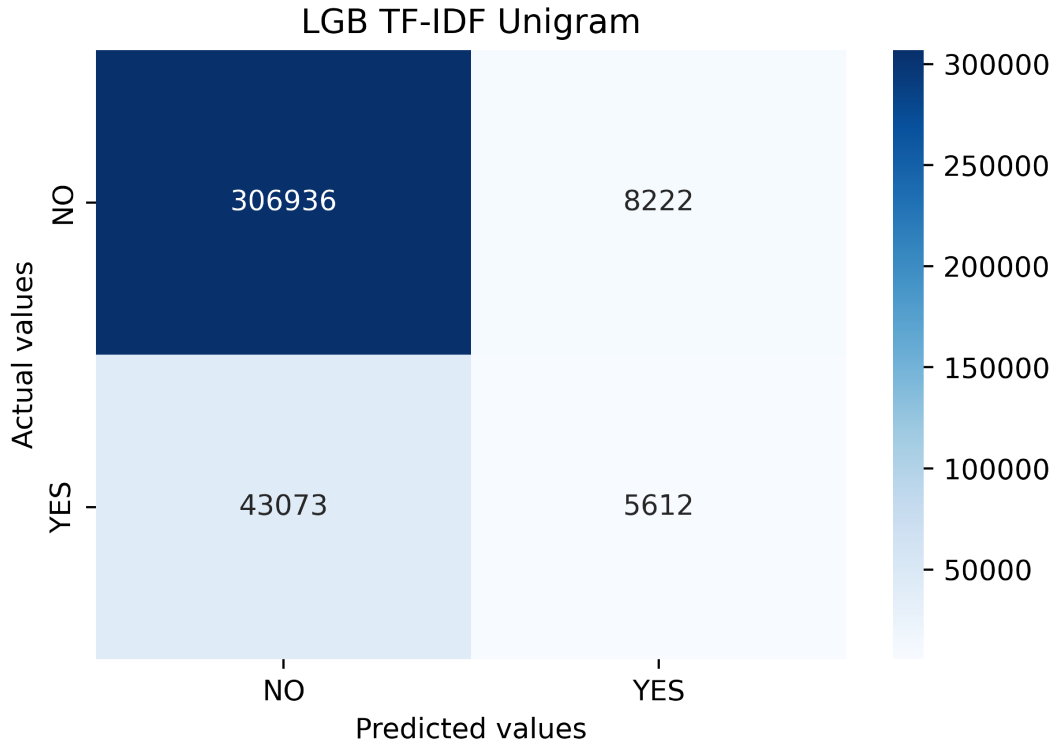

Figura 21 – Matriz de confusão, classificação de relevância no *DarkMarketItems* utilizando abordagem baseada em *keywords* com limiar 0.5 (Fonte: O autor (2025)).

Ao elevar o limiar de decisão para 0,7 (Figura 22), observa-se uma grande redução nos falsos positivos (de 8.222 para 2.517), mas também uma queda nos verdadeiros positivos (de 5.612 para 2.039). Esse resultado já era esperado, pois limiares mais altos tendem a privilegiar a precisão em vez da sensibilidade. Em outras palavras, o modelo passa a agir de forma mais cautelosa, classificando como relevantes apenas os casos em que possui maior grau de confiança.

Os resultados quantitativos apresentados na Tabela 23 confirmam essa tendência. A acurácia geral manteve-se em 0,86, indicando estabilidade global no desempenho. Para a classe não relacionada à segurança (classe 0), o modelo obteve precisão de 0,87 e revocação

de 0,99, demonstrando capacidade de reconhecer corretamente os anúncios neutros. Já para os itens de cibersegurança (classe 1), o modelo atingiu precisão de 0,45, mas com uma revocação reduzida de 0,04, resultando em uma medida-F1 de 0,08.

Tabela 23 – Métricas de desempenho do modelo no conjunto *DarkMarketItems* (estratégia baseada em palavras-chave, limiar 0.7).

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
0 (não relacionado)	0.87	0.99	0.93	315.158
1 (relacionado à segurança)	0.45	0.04	0.08	48.685
Acurácia		0.86		363.843
Média Macro	0.66	0.52	0.50	363.843
Média Ponderada	0.81	0.86	0.81	363.843

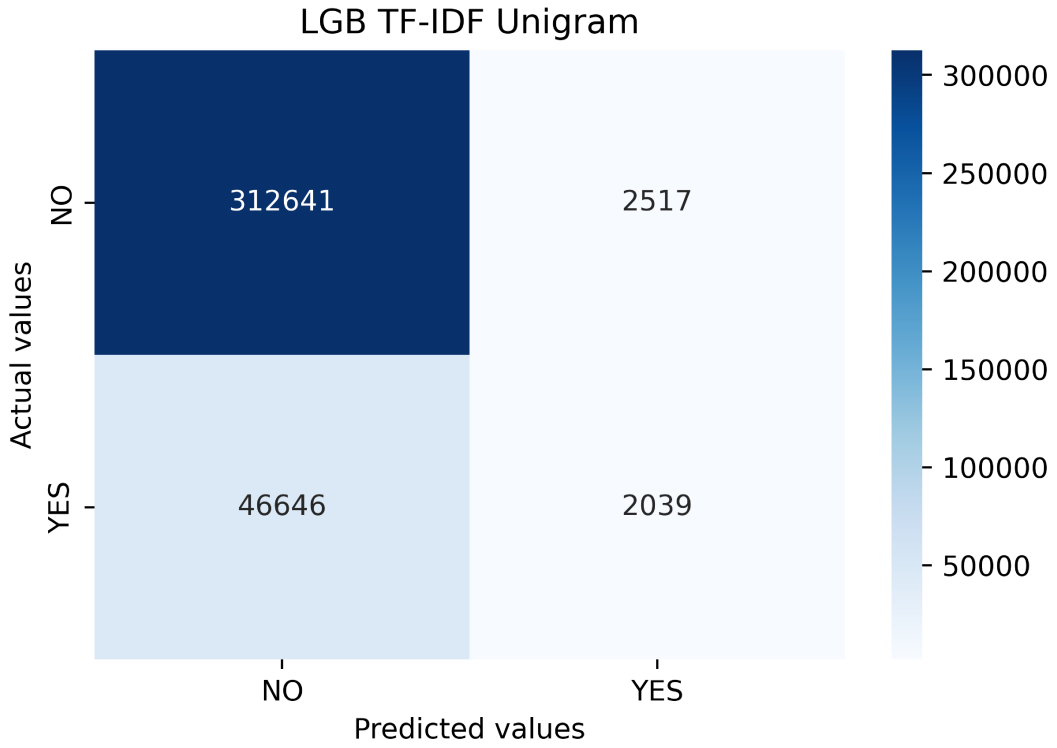

Figura 22 – Matriz de confusão, classificação de relevância no *DarkMarketItems* utilizando abordagem baseada em *keywords* com limiar 0.7 (Fonte: O autor (2025)).

Nos experimentos com o LLM como referência, a matriz de confusão, Figura 23, os resultados detalhados revelam 244.130 verdadeiros negativos e 105.879 falsos negativos, enquanto apenas 11.914 instâncias foram classificadas como verdadeiros positivos.

Os indicadores quantitativos mostram que o modelo obteve precisão de 0,70 e revocação de 0,99 para a classe não relacionada à segurança (classe 0), o que confirma sua elevada capacidade de evitar falsos positivos. Em contrapartida, para a classe relacionada à segurança (classe 1), a precisão foi de 0,86, mas o revocação caiu para 0,10, resultando em uma medida F1 de 0,18, o que indica baixa sensibilidade na detecção de itens efetivamente

relevantes. O desempenho geral atingiu uma acurácia de 0,70, com média ponderada da medida F1 igual a 0,61.

Esses resultados estão resumidos na Tabela 24, que apresenta as principais métricas de desempenho obtidas no cenário com o LLM como referência de classificação.

Tabela 24 – Métricas de desempenho do modelo no conjunto *DarkMarketItems* (LLM, limiar 0.5).

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
0 (não relacionado)	0.70	0.99	0.82	246.050
1 (relacionado à segurança)	0.86	0.10	0.18	117.793
Acurácia		0.70		363.843
Média Macro	0.78	0.55	0.50	363.843
Média Ponderada	0.75	0.70	0.61	363.843

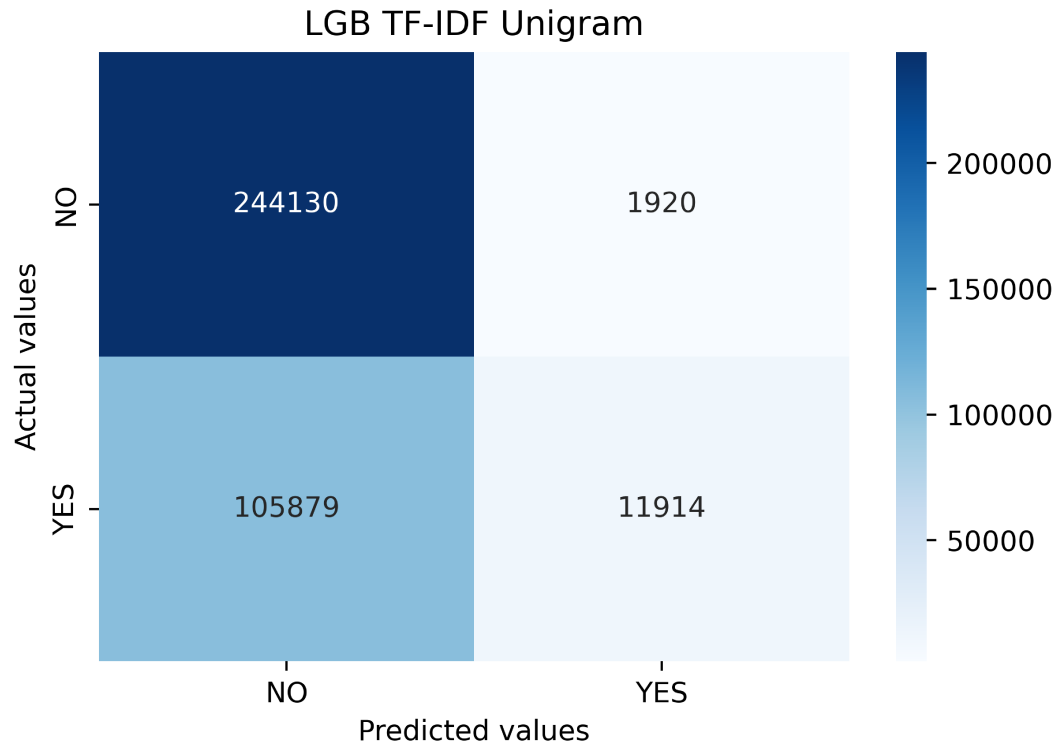

Figura 23 – Matriz de confusão, classificação de relevância no *DarkMarketItems* utilizando LLM como tabela de verdade com limiar 0.5 (Fonte: O autor (2025)).

Com o aumento do limiar para 0,7 (Figura 22), o modelo torna-se mais seletivo, apresentando redução de falsos positivos (de 1.920 para 441) e diminuição nos verdadeiros positivos (de 11.914 para 4.115). O aumento do limiar torna a classificação mais rigorosa, elevando a confiança nas previsões. Por outro lado, parte dos itens realmente relevantes deixa de ser identificada.

Os resultados quantitativos apresentados na Tabela 25 reforçam esse comportamento: para a classe não relacionada à segurança (classe 0), o modelo manteve revocação de 1,00

Tabela 25 – Métricas de desempenho do modelo no conjunto *DarkMarketItems* (LLM, limiar 0.7).

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
0 (não relacionado)	0.68	1.00	0.81	246.050
1 (relacionado à segurança)	0.90	0.03	0.07	117.793
Acurácia		0.69		363.843
Média Macro	0.79	0.52	0.44	363.843
Média Ponderada	0.75	0.69	0.57	363.843

e precisão de 0,68, indicando alta capacidade de identificar corretamente itens neutros. Por outro lado, para a classe relacionada à segurança (classe 1), a precisão atingiu 0,90, mas a revocação caiu para 0,03, resultando em uma medida-F1 de apenas 0,07. A acurácia global foi de 0,69, ligeiramente inferior à obtida com limiar 0,5. Esses valores evidenciam que, embora o modelo se torne mais rigoroso ao elevar o limiar, essa decisão reduz significativamente sua sensibilidade, limitando a detecção de itens de cibersegurança com menor grau de confiança.

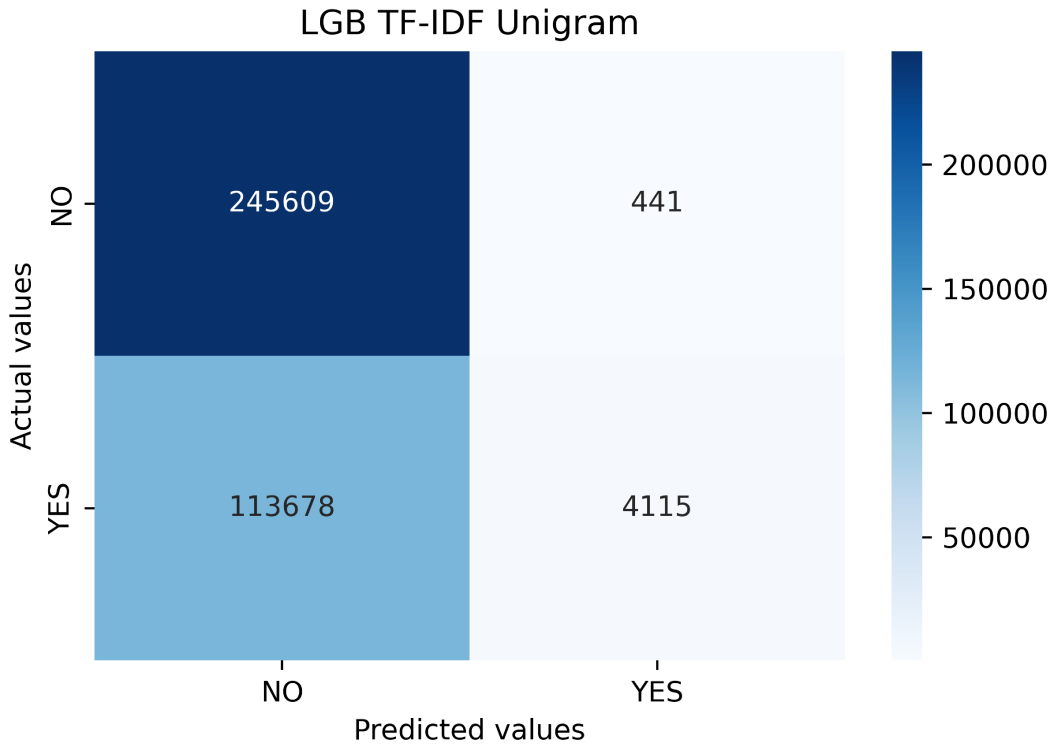

Figura 24 – Matriz de confusão, classificação de relevância no *DarkMarketItems* utilizando LLM como tabela de verdade com limiar 0.7 (Fonte: O autor (2025)).

Esses resultados sugerem que o aumento do limiar leva o modelo a um comportamento ainda mais conservador, priorizando previsões em detrimento da sensibilidade. Em outras palavras, o modelo evita classificar falsos positivos, mas passa a negligenciar parte das amostras realmente relevantes. Apesar disso, a consistência da acurácia e da precisão nas

classes majoritárias demonstra que a base de conhecimento adquirida em fóruns ainda é aplicável ao contexto de mercados da *Dark Web*, mesmo diante de variações de vocabulário e contexto textual.

A capacidade do modelo em classificar postagens em fóruns distintos e em diferentes idiomas, mesmo sem o ajuste fino em dados específicos de cada nova fonte, fornece as evidências para responder a QP4. Os resultados apresentados nesta seção confirmam que a técnica de transferência de aprendizado é eficaz para conferir ao modelo operação em ambientes diferentes. Assim, conclui-se que o modelo proposto não apenas detecta ameaças em contextos conhecidos, mas também generaliza o conhecimento adquirido para novas frentes de CTI. Ressalta-se que tal automatização atua como um sistema de suporte, sem destacar a ação humana na validação final e na interpretação das ameaças identificadas.

5.3 Discussão

A avaliação técnica da coleta fundamentou-se no uso da linguagem *Golang* e do *framework Colly* para assegurar performance e processamento concorrente , utilizando roteamento via *Tor* para viabilizar o acesso a domínios *.onion*. Todavia, reconhece-se a impossibilidade de assegurar a integridade total ou a completude temporal do conjunto de dados, uma vez que a volatilidade dos endereços e a instabilidade dos servidores na *Dark Web* impediram a coleta de períodos uniformes entre todos os fóruns. Esse viés na estrutura da coleta significa que os picos de atividade detectados podem não representar um aumento real das ameaças, mas sim apenas uma maior disponibilidade momentânea de certos fóruns em comparação a outros. Isso ocorre porque a falta de um catálogo completo e estável de endereços ativos na *Dark Web* impede o uso de métricas matemáticas para confirmar se os dados coletados são, de fato, representativos de todo o cenário. Assim, os resultados não devem ser vistos como uma linha do tempo completa ou um ciclo exato de como as ameaças evoluíram. Eles indicam apenas os temas que estavam em alta nos momentos em que os fóruns estavam acessíveis e a ferramenta de coleta conseguiu funcionar com sucesso.

Devido a essa natureza irregular da coleta, o estudo seguiu uma linha exploratória para descrever como novos temas surgem e como as discussões mudam qualitativamente ao longo do tempo. Por essa razão, não foram aplicados testes estatísticos formais, como cálculos de tendência ou sazonalidade, pois a queda constante de servidores e a mudança frequente de endereços criam falhas na linha do tempo que comprometem a precisão de tais métodos. Como esses testes exigem dados contínuos e sem interrupções para serem precisos, a instabilidade do ambiente da *Dark Web* impediria a obtenção de resultados estatisticamente confiáveis.

A análise da evolução das ameaças, em termos de tópicos e tendências, utilizando modelagem de tópicos (LDA) e séries temporais ao longo dos meses, nos fornece importantes

lições aprendidas e implicações para a prática deste estudo nesses ambientes.

Uma lição aprendida com esta análise é que os resultados observados não evidenciam de maneira conclusiva a existência de ciclos na evolução das ameaças cibernéticas, nem permitem inferir diretamente sobre o amadurecimento das comunidades, aspectos que podem ser explorados em estudos futuros. No entanto, é importante reconhecer algumas limitações deste estudo, como a dificuldade em obter endereços dos links ativos da *Dark Web*, a instabilidade dos fóruns analisados e a comunicação restrita entre seus membros, fatores que podem ter impactado a abrangência das análises.

A análise dos resultados obtidos nos diferentes domínios de aplicação (fóruns 8kun, VGR, *DarkTopics* e mercados *Dark Market Items*) fornece importantes lições aprendidas e implicações práticas sobre a generalização do modelo proposto.

A principal lição aprendida está na demonstração da eficácia da Transferência de Aprendizado em um domínio de cibersegurança e segurança digital, mesmo diante de ruídos textuais. O modelo, treinado em um tipo específico de conteúdo malicioso, mostrou-se capaz de reconhecer padrões de risco em novos contextos, como fóruns anônimos (8kun) e comunidades *gamer* (VGR), mantendo a diferenciação entre discursos neutros, com probabilidades próximas de 0, e conteúdos que utilizam vocabulário de segurança, invasão ou conflito. Adicionalmente, o modelo demonstrou alta sensibilidade ao vocabulário técnico nos conjuntos *DarkTopics* e *Dark Market Items*. Essa capacidade é muito importante para a triagem inicial, indicando que o modelo é muito bom para identificar o tema de cibersegurança em grandes volumes de dados.

Em contrapartida, essa alta sensibilidade ao vocabulário técnico revelou uma importante implicação prática: a tendência de classificar como perigosas postagens que, na verdade, não eram uma ameaça real. Conforme observado no VGR, postagens informativas ou de alerta sobre segurança foram classificadas com alta probabilidade, o que sugere que o modelo é capaz de identificar a palavra-chave, mas não a intenção. Portanto, o modelo é uma ferramenta robusta para a detecção de assuntos relevantes, mas pode exigir uma etapa posterior de refinamento, como um pós-processamento humano, para diferenciar a intenção real da postagem e reduzir os falsos positivos em ambientes não maliciosos.

Nesse contexto, a análise revela que o deslocamento entre domínios foi o fator determinante para a variação das métricas de desempenho. Para medir de forma quantitativa e evitar o risco de validação circular, que ocorreria se o próprio modelo fosse usado para confirmar suas predições em dados sem rótulos, adotaram-se referências externas independentes. No cenário 1, a validade foi confirmada por auditoria humana direta nos escores extremos, enquanto no cenário 2, utilizou-se um modelo de linguagem (LLM) como fonte externa. Essa abordagem garantiu que as métricas de precisão, revocação e medida F1 apresentadas fossem fundamentadas em critérios externos ao processo de treinamento original.

Contudo, reforça-se que a principal limitação que ainda compromete uma generalização reside na instabilidade técnica da coleta de dados. Como discutido no início desta seção, a volatilidade dos endereços *.onion* resulta em dados ruidosos e na perda de dados, como a autoria das postagens, o que impede uma análise contextual mais profunda. Para trabalhos futuros, é fundamental o desenvolvimento de *crawlers* mais resilientes e de técnicas de tratamento de ruído mais agressivas, visando capturar o contexto completo das discussões e reduzir o impacto das interrupções de serviço na qualidade.

Capítulo 6

Conclusão

Em vista das abordagens disponíveis na literatura, que incluem técnicas de mineração de dados em fóruns da *Dark* e *Surface Web*, e de trabalhos sobre modelos de Aprendizado de Máquina Supervisionado aplicados a conjuntos de dados, esta dissertação propôs aprimorar a capacidade de geração de *Cyber Threat Intelligence* (CTI). Isso foi alcançado pela integração das técnicas de mineração de dados em fóruns com a aplicação de modelos de Aprendizado de Máquina Supervisionado. O objetivo central foi avaliar a eficácia da Transferência de Aprendizado na classificação de conteúdo malicioso em novos domínios textuais, bem como analisar a evolução temporal das ameaças discutidas nesses ambientes.

A metodologia central adotada combinou a mineração de dados em ambientes não estruturados, fóruns *Dark* e *Surface Web*, com a aplicação de um modelo de aprendizado de máquina supervisionado, *LightGBM*, utilizando a codificação TF-IDF para a representação textual. O principal esforço metodológico concentrou-se na avaliação da transferência de aprendizado transdutiva, onde o modelo pré-treinado em um conjunto de dados inicial rotulado foi aplicado e validado em domínios completamente novos e não rotulados, como o fórum VGR e diversos *marketplaces*. Além disso, a modelagem de tópicos LDA foi empregada para analisar a evolução temporal e as tendências de discussões em cada ambiente.

Os resultados obtidos permitiram responder as questões de pesquisa enunciadas na introdução. No que tange à evolução temporal e aos padrões de atividade, QP1 e QP2, a análise de tópicos LDA revelou que a *Dark Web* em língua portuguesa apresenta uma transição evidente de discussões técnicas para práticas criminosas, como a venda de dados pessoais, fornecendo percepções para equipes de segurança sobre picos de interesses. No tocante às disparidades entre os dois ambientes estudados, *Surface Web* e *Dark Web* presentes na QP3, observou-se que a natureza dos fóruns dita a maturidade técnica e o foco das discussões, separando ecossistemas de aprendizado de ambientes (*Dark Web*) de comercialização de dados (*Surface Web*).

Quanto à eficácia da transferência de aprendizado QP4, o modelo demonstrou capaci-

dade de generalização ao manter um desempenho promissor na diferenciação de conteúdos, sendo capaz de isolar o vocabulário de risco em novas fontes de dados. A capacidade do modelo de generalizar o aprendizado em novos domínios estabelece uma base para a construção de sistemas de CTI proativa. Contudo, foi identificada uma alta sensibilidade ao vocabulário técnico, que, embora excelente para a triagem inicial, levou à classificação de postagens informativas ou de alerta como de alto risco, revelando a limitação do modelo em diferenciar a palavra-chave da intenção real da postagem. Nesse sentido, não se descarta a necessidade da ação humana especializada para a interpretação contextual final.

Este estudo, ao validar a Transferência de Aprendizado em ambientes multilíngues e domínios variados, contribui diretamente para a classificação de ameaças cibernéticas em textos em língua portuguesa, uma área de pesquisa ainda escassa no cenário acadêmico.

6.1 Principais Contribuições

As principais contribuições deste trabalho são listadas a seguir:

1. Conjunto de dados adicional, complementar ao apresentado em (FILHO, 2024)
2. Conjunto de dados rotulados automaticamente aplicando-se o modelo descrito em (FILHO, 2024)
3. Arcabouço Metodológico de Coleta e Mineração: Definição e implementação de um *pipeline* completo de CTI (coleta na *Dark/Surface Web*, armazenamento relacional e pré-processamento multilíngue), incluindo ferramentas de *web crawling* em Golang, otimizadas para a instabilidade de domínios *Onion*;
4. Demonstração e Validação da Transferência de Aprendizado: Comprovação da capacidade de generalização de modelos de CTI a partir de bases de dados específicas para novos domínios (8kun, VGR, Dark Topics, DarkMarketItems);
5. Análise da Evolução de Ameaças: Caracterização detalhada da evolução temporal de tópicos de cibercrime (LDA) em fóruns de língua portuguesa e inglesa, identificando a transição de discussões técnicas para práticas criminosas (venda de dados pessoais e CPF).

6.2 Trabalhos Futuros

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

- ❑ Implementar uma etapa de filtragem de falsos positivos por meio de um módulo de Processamento de Linguagem Natural avançado para diferenciar a intenção real da

postagem, informativa ou maliciosa, mitigando a alta sensibilidade do modelo ao vocabulário técnico observado.

- ❏ Expandir a base de dados temporal, coleta de *posts*, para um período mais extenso, para determinar a existência de ciclos e sazonalidades conclusivas na evolução das ameaças cibernéticas, e não apenas tendências trimestrais, aprimorando a CTI preditiva.

- ❏ Integrar a abordagem de classificação com Modelos de Linguagem de Grande Escala (LLMs) baseados em arquiteturas *Transformer* (como BERT, DistilBERT, ou XLM-R) para aproveitar seu entendimento contextual e semântico. O objetivo é aprimorar a capacidade de generalização e aumentar o *recall* para a classe positiva, que se mostrou baixo nos experimentos, sem comprometer a precisão na detecção de ameaças.

- ❏ Desenvolver estratégias automatizadas para a coleta de dados, mitigando as limitações observadas, como a instabilidade dos fóruns e a mudança frequente de endereços em domínios *onion*.

6.3 Contribuições em Produção Bibliográfica

O artigo intitulado *"Evolução de ameaças em fóruns da Dark Web e Surface Web: um estudo baseado em modelagem de tópicos e séries temporais"* (PEREIRA et al., 2025), investiga a evolução temporal das discussões sobre ameaças cibernéticas em fóruns da *Dark Web* e da *Surface Web* entre 2015 e 2024, com o objetivo de identificar tendências, padrões sazonais e diferenças entre esses ambientes. Este trabalho foi aceito e apresentado durante o *XXIII Simpósio Brasileiro de Cibersegurança* (*SBSeg 2025*), um evento promovido pela Sociedade Brasileira de Computação que ocorreu em setembro de 2025 organizado pela Pontifícia Universidade Católica do Paraná e pela Universidade Federal do Paraná em Foz do Iguaçu no Paraná.

Na ocasião, o trabalho recebeu o prêmio de *Melhor Artigo da Trilha de Sistemas*. Tal reconhecimento destaca a relevância dos resultados apresentados para a área de segurança cibernética.

Referências

ALYAFEAI, Z.; ALSHAIBANI, M. S.; AHMAD, I. **A Survey on Transfer Learning in Natural Language Processing**. 2020. Disponível em: <https://arxiv.org/abs/2007.04239>.

AVANZI, B. et al. On the evolution of data breach reporting patterns and frequency in the united states: a cross-state analysis. **arXiv preprint arXiv:2310.04786**, 2023. Disponível em: <https://arxiv.org/abs/2310.04786>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca**. Bookman Editora, 2013. ISBN 9788582600498. Disponível em: <https://books.google.com.br/books?id=YWk3AgAAQBAJ>.

BARBON, R. S.; AKABANE, A. T. Towards transfer learning techniquesmdash;bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: A case study. **Sensors**, v. 22, n. 21, 2022. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/22/21/8184>.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, n. null, p. 993–1022, mar. 2003. ISSN 1532-4435.

BURKOV, A. The hundred-page machine learning book. In: . [s.n.], 2019. Disponível em: <https://api.semanticscholar.org/CorpusID:202780305>.

CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. [S.l.]: BPLN, 2023. <https://brasileiraspln.com/livro-pln>. ISBN 978-65-00-80693-9.

CHEN, X.; LOU, X. Enhancing text classification through quantum transfer learning: A hybrid quantum-classical approach with complex kernel self-attention networks. **IEEE Access**, v. 13, p. 133882–133890, 2025.

CIMPANU, C. **University of Utah pays $457,000 to ransomware gang**. 2020. Acessado: 12-04-2023. Disponível em: <https://www.zdnet.com/article/university-of-utah-pays-457000-to-ransomware-gang/>.

CRAWLY. **O que é Crawler e como funcionam os robôs para coleta de dados**. 2021. Accessed : 2024-10-25. Disponível em: <https://www.crawly.com.br/blog/o-que-e-crawler-robos-para-coleta-de-dados>.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: <https://arxiv.org/abs/1810.04805>.

DIALANI, P. **The Future of Data Revolution will be Unstructured Data**. 2020. Accessed : 2023-01-11. Disponível em: <https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/>.

DONGES, N. **What Is Transfer Learning? Exploring the Popular Deep Learning Approach.** 2022. Accessed : 2023-11-16. Disponível em: <https://builtin.com/data-science/transfer-learning>.

FELDMAN, R.; SANGER, J. et al. **The text mining handbook: advanced approaches in analyzing unstructured data**. [S.l.]: Cambridge university press, 2007.

FILHO, S. A. de J. **Identificação de posts maliciosos na dark web utilizando Aprendizado de Máquina Supervisionado**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Uberlândia, Uberlândia, Brasil, jan 2024. Orientador: Rodrigo Sanches Miani. Disponível em: <https://repositorio.ufu.br/handle/123456789/41232>.

FU, T.; ABBASI, A.; CHEN, H.-c. A focused crawler for dark web forums. **JASIST**, v. 61, p. 1213–1231, 06 2010.

GARG, N.; SHARMA, K. Text pre-processing of multilingual for sentiment analysis based on social network data. **International Journal of Electrical and Computer Engineering (IJECE)**, v. 12, p. 776, 02 2022.

GLOBO. **JBS diz que pagou US$ 11 milhões em resgate a ataque hacker em operações nos EUA**. 2021. Accessed : 2024-02-27. Disponível em: <https://g1.globo.com/economia/noticia/2021/06/09/jbs-diz-que-pagou-11-milhoes-em-resposta-a-ataque-hacker-em-operacoes-nos-eua.ghtml>.

GOOGLE. **Como a Pesquisa Google funciona**. 2021. <https://developers.google.com/search/docs/beginner/how-search-works#indexing>. Acesso em: 2025-07-16.

GUTTMAN, B.; ROBACK, E. A. **An introduction to computer security: the NIST handbook**. [S.l.]: DIANE Publishing, 1995.

HICKMAN, L. et al. Text preprocessing for text mining in organizational research: Review and recommendations. **Organizational Research Methods**, Sage Publications Sage CA: Los Angeles, CA, v. 25, n. 1, p. 114–146, 2022.

HOWARD, J.; RUDER, S. **Universal Language Model Fine-tuning for Text Classification**. 2018. Disponível em: <https://arxiv.org/abs/1801.06146>.

IBM. **What is text mining?** 2023. Accessed : 2023-12-11. Disponível em: <https://www.ibm.com/topics/text-mining>.

JO, T. **Text Mining: Concepts, Implementation, and Big Data Challenge**. Springer Nature Switzerland, 2024. (Studies in Big Data). ISBN 9783031759765. Disponível em: <https://books.google.com.br/books?id=ewk4EQAAQBAJ>.

KAVALLIEROS, D. et al. Understanding the dark web. In: _____. **Dark Web Investigation**. Cham: Springer International Publishing, 2021. p. 3–26. ISBN 978-3-030-55343-2. Disponível em: <https://doi.org/10.1007/978-3-030-55343-2_1>.

KHYANI, D. et al. An interpretation of lemmatization and stemming in natural language processing. **Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology**, v. 22, p. 350–357, 01 2021.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. v. 14, 03 2001.

KOLOVEAS, P. et al. intime: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. **Electronics**, v. 10, n. 7, 2021. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/10/7/818>.

KONCHADY, M. **Text Mining Application Programming**. Charles River Media, 2006. (Charles River Media programming series). ISBN 9781584504603. Disponível em: <https://books.google.com.br/books?id=m9JQAAAAMAAJ>.

KOST, E. **What Caused the Uber Data Breach in 2022?** 2022. Accessed : 2023-12-04. Disponível em: <https://www.upguard.com/blog/what-caused-the-uber-data-breach>.

KOWSARI, K. et al. Text classification algorithms: A survey. **Information**, v. 10, n. 4, 2019. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/10/4/150>.

_____. Text classification algorithms: A survey. **Information**, v. 10, n. 4, 2019. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/10/4/150>.

KüHN, P.; WITTORF, K.; REUTER, C. Navigating the shadows: Manual and semi-automated evaluation of the dark web for cyber threat intelligence. **IEEE Access**, v. 12, p. 118903–118922, 2024.

LIAKOS, P. et al. Focused crawling for the hidden web. **World Wide Web**, v. 19, 05 2015.

MUREL, J. **O que é alocação latente de Dirichlet?** 2024. Accessed : 2025-03-06. Disponível em: <https://www.ibm.com/br-pt/think/topics/latent-dirichlet-allocation>.

NAJORK, M. Web crawler architecture. In: _____. **Encyclopedia of Database Systems**. Boston, MA: Springer US, 2009. p. 3462–3465. ISBN 978-0-387-39940-9. Disponível em: <https://doi.org/10.1007/978-0-387-39940-9_457>.

NARDE, W. et al. Classificação de notícias em português utilizando modelos baseados em transferência de aprendizagem e transformers. In: **Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: SBC, 2024. p. 212–216. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/31133>.

NAYAK, A. et al. Survey on pre-processing techniques for text mining. **International Journal Of Engineering And Computer Science**, v. 5, p. 2319–7242, 06 2016.

NUNES, E. et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In: **2016 IEEE Conference on Intelligence and Security Informatics (ISI)**. [S.l.: s.n.], 2016. p. 7–12.

PEREIRA, M. et al. Evolução de ameaças em fóruns da dark web e surface web: um estudo baseado em modelagem de tópicos e séries temporais. In: **Anais do XXV Simpósio Brasileiro de Cibersegurança**. Porto Alegre, RS, Brasil: SBC, 2025. p. 401–416. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbseg/article/view/36633>.

POWERS, D. M. W. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation**. 2020. Disponível em: <https://arxiv.org/abs/2010.16061>.

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, 07 2018.

RAFFEL, C. et al. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. 2023. Disponível em: <https://arxiv.org/abs/1910.10683>.

RAHMAN, M. R.; HEZAVEH, R. M.; WILLIAMS, L. What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 55, n. 12, mar 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3571726>.

SAPIENZA, A. et al. Early warnings of cyber threats in online discussions. In: **2017 IEEE International Conference on Data Mining Workshops (ICDMW)**. [S.l.: s.n.], 2017. p. 667–674.

SARKAR, S. et al. Predicting enterprise cyber incidents using social network analysis on the darkweb hacker forums. 11 2018. ISSN 24742120, 24742139. Disponível em: <https://www.jstor.org/stable/26846122>.

SHACKLEFORD, D. **Who's Using Cyberthreat Intelligence and How?** 2015. Disponível em: <https://www.sans.org/webcasts/cyberthreat-intelligence-how-1-definitions-tools-standards-99052/>. Acesso em: 2023-11-11.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing Management**, v. 45, p. 427–437, 07 2009.

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

STEDMAN, C. **DEFINITION text mining (text analytics)**. 2023. Accessed : 2023-12-11. Disponível em: <https://www.techtarget.com/searchbusinessanalytics/definition/text-mining>.

SUN, N. et al. Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. **IEEE Communications Surveys & Tutorials**, v. 25, n. 3, p. 1748–1774, 2023.

SWAN, K. Onion routing and tor. **GEO. L. TECH. REV.**, v. 1, p. 110–110, 2016.

SYED, S.; SPRUIT, M. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: **2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)**. [S.l.: s.n.], 2017. p. 165–174.

TABASSUM, A.; PATIL, D. R. R. A survey on text pre-processing & feature extraction techniques in natural language processing. In: . [s.n.], 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:235211496>.

THARWAT, A. Classification assessment methods: a detailed tutorial. 08 2018.

_____. Classification assessment methods. **Applied Computing and Informatics**, v. 17, n. 1, p. 168–192, 07 2020. ISSN 2634-1964. Disponível em: <https://doi.org/10.1016/j.aci.2018.08.003>.

The MITRE Corporation. **CVE - Frequently Asked Questions**. 2018. Accessed : 2023-10-16. Disponível em: <https://cve.mitre.org/about/faqs.html>.

TONG, Z.; ZHANG, H. A text mining research based on lda topic modelling. In: . [S.l.: s.n.], 2016. v. 6, p. 201–210.

WEISS, K.; KHOSHGOFTAAR, T.; WANG, D. A survey of transfer learning. **Journal of Big Data**, v. 3, 05 2016.

WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. **Pattern Recognition**, v. 48, n. 9, p. 2839–2846, 2015. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320315000989>.

ZHUANG, F. et al. Triplex transfer learning: Exploiting both shared and distinct concepts for text classification. **IEEE Transactions on Cybernetics**, v. 44, n. 7, p. 1191–1203, 2014.

_____. **A Comprehensive Survey on Transfer Learning**. 2020. Disponível em: <https://arxiv.org/abs/1911.02685>.

ŘEHůřEK, R. **What is Gensim**. 2024. Acessado: 27-04-2025. Disponível em: <https://radimrehurek.com/gensim/intro.html#what-is-gensim>.