Evolução de ameaças em fóruns da *Dark Web* e *Surface Web*: um estudo baseado em modelagem de tópicos e séries temporais

Miguel Henrique de Brito Pereira[1], **Sebastião Alves de Jesus Filho**[1], **Paulo Henrique Ribeiro Gabriel**[1], **Rodrigo Sanches Miani**[1]

[1]Faculdade de Computação - FACOM – Universidade Federal de Uberlândia (UFU)
Av. João Naves de Ávila, nº 2121, Santa Mônica, Uberlândia – MG – Brasil

{miguelbrito,sebastiao,phrg,miani}@ufu.br

Abstract. *This work investigates the temporal evolution of discussions about cyber threats in Dark Web and Surface Web forums between 2015 and 2024, aiming to identify trends, seasonal patterns, and differences between these environments. By analyzing over 52,000 posts using text preprocessing and Latent Dirichlet Allocation (LDA) topic modeling, the study identifies key trends, seasonal patterns, differences between environments, and dynamics within online communities. The analysis showed that Surface Web forums exhibited high topic variability. In contrast, the Portuguese-speaking Dark Web demonstrated a predominance of personal data commercialization, while the English-speaking Dark Web consistently maintained technical offensive topics, such as phishing and malware creation.*

Resumo. *Este trabalho 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. Ao analisar mais de 52.000 postagens utilizando processamento de linguagem natural e modelagem de tópicos com Latent Dirichlet Allocation (LDA), o estudo revela tendências-chave, padrões sazonais, diferenças entre os ambientes e dinâmicas dentro das comunidades online. A análise revelou que fóruns da Surface Web apresentaram alta variabilidade de tópicos. Por outro lado, a Dark Web em português demonstrou predominância na comercialização de dados pessoais, enquanto a Dark Web em inglês manteve tópicos técnicos ofensivos, como phishing e criação de malware, com frequência contínua.*

1. Introdução

A popularização da Internet trouxe consigo um aumento significativo das ameaças cibernéticas, que evoluíram tanto em sofisticação quanto em alcance [Labs 2024]. Fóruns online, incluindo a *Dark Web* e a *Surface Web*, tornaram-se ambientes onde atores maliciosos compartilham informações sobre vulnerabilidades, ataques, ferramentas de exploração e dados vazados[Avanzi et al. 2023]. O monitoramento dessas discussões pode revelar padrões de comportamento de cibercriminosos, permitindo uma resposta proativa por parte de empresas e órgãos de segurança [Koloveas et al. 2021].

Nesse contexto, os cibercriminosos têm se organizado de forma cada vez mais estruturada e coordenada. De acordo com [Rahman et al. 2023], os ataques cibernéticos deixaram de ser casos isolados para se tornarem parte de um crime organizado. Um exemplo ocorreu em 2020, quando a Universidade de Utah foi alvo de um ataque de *ransomware*, resultando no roubo de informações confidenciais de estudantes e em uma perda financeira de 457 mil dólares para a instituição [Cimpanu 2020]. Esse incidente ilustra a mudança de paradigma nos ataques cibernéticos, agora conduzidos por grupos organizados. Conforme evidenciado por [Nunes et al. 2016, Cascavilla 2025], muitos desses ataques são discutidos em fóruns da *Dark Web* antes de ocorrerem. Um caso relevante, mencionado no mesmo artigo, refere-se a uma vulnerabilidade no sistema operacional *Microsoft Windows*, identificada em fevereiro de 2015. Em pouco mais de um mês, surgiram relatos indicando que um fórum na *Dark Web* estava comercializando informações relacionadas a essa vulnerabilidade específica.

Uma estratégia para se preparar e responder a tais ataques é através do conceito de *Cyber Threat Intelligence* (CTI). Segundo [Wagner et al. 2019], CTI é definido como um conjunto de informações baseadas em evidências, abrangendo contexto, mecanismos, indicadores, implicações e orientações acionáveis relacionadas a uma ameaça existente ou emergente. A aplicação dessas informações na tomada de decisão pode fortalecer a capacidade de resposta às crescentes complexidades do cenário cibernético.

Com base neste contexto, este trabalho busca investigar como as discussões sobre cibercrimes evoluíram ao longo do tempo, analisando postagens em fóruns da *Dark Web* e da *Surface Web* entre 2015 e 2024. Essa análise tem como objetivo apoiar estratégias de CTI, ao identificar tendências, comportamentos maliciosos e padrões sazonais que possam antecipar ameaças futuras. Neste trabalho, emprega-se modelagem de tópicos (LDA) e análise de séries temporais para responder às seguintes questões de pesquisa:

1. Como as discussões sobre ameaças evoluíram ao longo do tempo nesses fóruns?
2. Existem padrões sazonais ou picos de atividade correlacionados a assuntos específicos?
3. Há diferenças entre *Dark Web* e *Surface Web*?

Para a realização da análise, foram coletados 11.087 *posts* de fóruns da *Dark Web*, incluindo *Hidden Answers*, *Deep Answer* e *Raddle*. Da *Surface Web*, foram extraídos 41.248 *posts* dos fóruns *Hacknology*, *Legit Cards*, *Nifheim* e *Nulledbb*. Espera-se que os resultados deste estudo contribuam para uma melhor compreensão do uso de fóruns da *Dark Web* e da *Surface Web* como fontes de dados para CTI, além de fornecer subsídios para o desenvolvimento de aplicações, como a identificação de *posts* maliciosos e a análise de tendências.

O restante do artigo está organizado da seguinte forma: na Seção 2, são apresentados os principais trabalhos que contribuíram para o desenvolvimento desta pesquisa. A Seção 3 aborda conceitos fundamentais sobre coleta de dados, aprendizado de máquina, *Deep Web*, *Dark Web*, *Surface Web* e o modelo LDA (Latent Dirichlet Allocation), utilizado para a modelagem de tópicos. Na Seção 4, detalham-se as etapas do desenvolvimento do estudo. Os resultados obtidos são analisados na Seção 5, e, por fim, as conclusões do artigo são apresentadas na Seção 6.

2. Trabalhos Relacionados

A literatura atual apresenta estudos que exploram dados coletados em redes sociais e fóruns da *Dark Web* para identificar postagens maliciosas e prever incidentes de segurança. Esses trabalhos partem do pressuposto de que é possível detectar indícios de ataques cibernéticos nessas fontes antes de sua execução. Além disso, algumas pesquisas focam no compartilhamento de CTI, utilizando técnicas de extração de dados em redes sociais, como o *X/Twitter*, e em plataformas da *Surface Web*, *Deep Web* e *Dark Web*. A seguir, são apresentados alguns desses estudos.

O trabalho conduzido por [Rahman et al. 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, esse 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 et al. 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 semiautomatizada 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 et al. 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] apresentam uma abordagem mais abrangente ao ampliar sua área de obtenção dos 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 dedicaram 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.

Os trabalhos citados apresentam contribuições significativas na extração e mineração de dados. No entanto, nenhum desses estudos realizou uma análise temporal que identificasse tendências e diferenças entre as fontes. Além disso, os trabalhos investigados lidam com fontes de dados de somente um tipo (*Surface Web* ou *Deep Web*) enquanto o presente trabalho investiga ambas as fontes. Essas lacunas motivam o presente estudo, que se propõe a preencher essa ausência de análise temporal.

3. Fundamentação Teórica

Nesta seção, são apresentados os conceitos de *Latent Dirichlet Allocation* (LDA) (Subseção 3.1), bem como os de *Surface Web*, *Deep Web* e *Dark Web* (Subseção 3.2) e Fóruns de discussão (Subseção 3.3).

3.1. *Latent Dirichlet Allocation* (LDA)

O *Latent Dirichlet Allocation* (LDA) é uma técnica de modelagem de tópicos para descobrir os tópicos centrais e suas distribuições em um conjunto de documentos. Apresentado pela primeira vez por [Blei et al. 2003], é um dos métodos mais populares para modelagem de tópicos. O algoritmo ignora a ordem e o contexto das palavras, concentrando-se apenas na frequência com que elas aparecem e coocorrem em cada documento.

De forma geral, a aplicação do LDA envolve etapas de pré-processamento, como a remoção de palavras de parada (*stopwords*) ou de termos irrelevantes. Essa prática tem como objetivo eliminar elementos que não contribuem para o significado desejado da análise, preservando, assim, a integridade das informações. Nesse sentido, palavras como “as”, “os”, “uns”, “de”, “para”, “com” e “por” são consideradas palavras de parada (*stopwords*). Outra etapa essencial é a tokenização, que consiste em dividir o texto em palavras individuais (tokens), 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.

3.2. *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 da web, utilizando navegadores convencionais, também chamados de *browsers* [Kavallieros et al. 2021]. Esses programas permitem que os usuários interajam com documentos HTML hospedados em servidores.

Segundo [Kavallieros et al. 2021], a *Deep Web* corresponde à outra parte da *web*. De forma simplificada, é o oposto da *Surface Web*, pois seus mecanismos de busca não conseguem indexar seu conteúdo. Essa é a principal diferença entre as duas em termos de acessibilidade aos dados. Os sites na *Surface Web* são indexados para que possam ser encontrados por mecanismos de busca, enquanto a *Deep Web* não passa por esse processo. Por fim, a *Dark Web* é uma parte da *Deep Web*, mas tem uma grande diferença: não pode ser acessada por meio de navegadores convencionais. Para isso, é necessário o uso de um navegador alternativo, 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 os usuários protegerem suas identidades online.

3.3. Fóruns de discussão

Fóruns de discussão na *Dark Web* são comunidades anônimas onde os usuários discutem temas, desde tecnologia e cibersegurança a assuntos de natureza ilegal. Esses fóruns operam em redes como TOR, permitindo que os participantes mantenham o anonimato. Comunidades semelhantes também existem na *Surface Web* e na *Deep Web*, porém não há mecanismos para dificultar 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, uma *thread* serve a uma discussão específica sobre um tópico. 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 *thread*. Para postar 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.

4. Materiais e Métodos

Esta seção apresenta a metodologia adotada, com o propósito de detalhar de forma sistemática as etapas desenvolvidas ao longo deste estudo. A Figura 1 mostra a estrutura geral da proposta, que é composta por três etapas principais (I, II e III). Cada uma dessas etapas é descrita nas subseções a seguir.

4.1. Coleta de *Posts*

A primeira etapa desse processo consiste na coleta de *posts*. Para isso, foram utilizados *crawlers*, também conhecidos como *web crawlers* ou *scrapers*. Esses algoritmos têm a

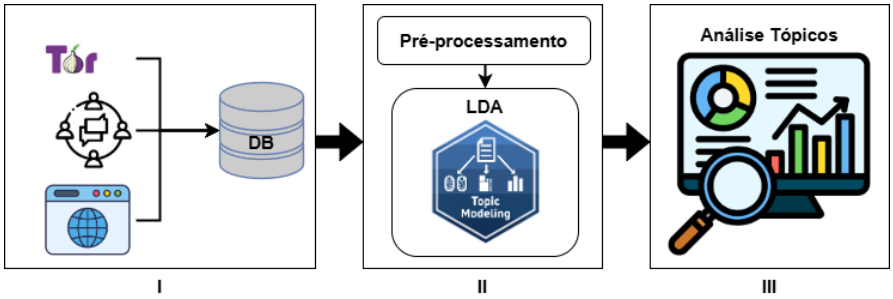

Figura 1. Etapas do desenvolvimento da pesquisa

função de vasculhar sites ou bancos de dados para diversos propósitos, como catalogar páginas, indexar conteúdo ou até mesmo extrair partes específicas [Crawly 2021].

No contexto deste trabalho, um *crawler* foi desenvolvido para examinar uma grande quantidade de páginas contendo perguntas e respostas dos fóruns da *Deep Web*, *Dark Web* e *Surface Web*. Esse sistema foi projetado para identificar e capturar as informações relevantes em tempo de execução, garantindo assim a coleta dos dados desejados. Os fóruns da *Deep web*, *Dark Web* e *Surface Web* que foram coletados por meio dos *crawlers* foram:

- **Respostas Ocultas** e **Deep Answers**: versões em inglês e português.
- **Raddle**, **Hackonology**, **Legitcarders**, **Niflheim.World** e **Nulledbb**: apenas em inglês.

A seleção dos fóruns analisados fundamentou-se na expressiva quantidade de perguntas e *posts* disponíveis, bem como em sua popularidade e reconhecimento entre os usuários, especialmente no contexto da *Dark Web*. Os endereços dos fóruns mencionados anteriormente 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. Esse processo é conduzido de forma recursiva, permitindo a identificação e exploração contínua de novas páginas interligadas por meio desses *hiperlinks* [Najork 2009].

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 pela própria biblioteca. 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
```

No final da execução do algoritmo, os dados coletados são armazenados em uma tabela específica do banco de dados, garantindo sua organização e disponibilidade para análises posteriores.

4.2. Pré-processamento

Na segunda etapa do processo, foram desenvolvidos dois módulos inter-relacionados: o módulo de pré-processamento dos dados coletados e o de modelagem de tópicos LDA. 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 dos *posts* em um único conjunto de dados; e (iv) a concatenação dos atributos de conteúdo relevantes para a modelagem, como texto principal, respostas e comentários. Ambos os módulos foram desenvolvidos em *Python*.

Foram aplicadas técnicas de mineração de textos, tais como a remoção de *stopwords* e termos irrelevantes. Essa prática tem o objetivo de eliminar palavras que não contribuem para o significado da análise, preservando assim a integridade da informação procurada. Neste sentido, palavras como “as”, “os”, “uns”, “de”, “para”, “com”, “por” foram removidas. Além das técnicas mencionadas, 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*).
- Redução de ruído: eliminação de sequências repetitivas de caracteres, como *kkkkkkkk* e *aaaaaa*.

Assim, a preparação do texto para a etapa de LDA assegura que os dados sejam processados corretamente, minimizando interferências causadas por anomalias ou ruídos textuais.

4.3. Modelagem de tópicos - LDA

A modelagem de tópicos foi executada de acordo com as seguintes etapas [Tong and Zhang 2016]: pré-processamento, treino do modelo e análise dos tópicos. A biblioteca *Gensim* [Řehůřek 2024] foi empregada para implementar o LDA sobre toda a base de dados coletada, devidamente pré-processada. A seguir, foram combinadas em um único campo textual as colunas consideradas mais relevantes para a análise do conteúdo: o título da postagem, a questão (ou corpo da postagem) e as respectivas respostas. Em seguida, o texto resultante passou por etapas de tokenização, *stemming* e normalização. 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 *post*, 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 and 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. É importante destacar que foi utilizado um modelo de aprendizado de máquina pré-treinado apresentado em [de Jesus Filho 2024] para selecionar *posts* com uma alta probabilidade de conter conteúdo malicioso. O modelo foi treinado usando o algoritmo *LightGBM* com TF-IDF. O modelo em questão atingiu acurácia de 94% em um conjunto de dados semelhante ao estudado neste trabalho[1]. 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 LDA foi aplicado em *posts* que receberam notas maiores ou iguais a $0,7$.

4.4. Análise dos Resultados

Na terceira e última etapa, foi realizada uma análise dos tópicos gerados e de sua ocorrência ao longo dos anos, com o objetivo de identificar informações relevantes em *posts* que possam auxiliar a comunidade de segurança da informação. Durante o estudo, diversas informações significativas foram extraídas, e as seguintes análises foram realizadas:

1. **Análise de atividade**: quantidade de *posts* maliciosos, identificação de variações significativas (picos) e evolução das discussões ao longo do tempo.
2. **Análise temática**: distribuição de *posts* por tópico e principais temas abordados.
3. **Comparações e aplicações**: análise das diferenças entre a *Dark Web* e a *Surface Web* considerando o tipo de atividade realizada, os temas abordados e suas aplicações na área de segurança cibernética.

Essas análises foram escolhidas para permitir uma compreensão mais aprofundada dos padrões de discussão nos fóruns analisados, bem como para identificar possíveis ameaças à segurança cibernética e tendências ao longo do tempo.

5. Resultados

Os resultados desta investigação são apresentados a seguir, com base nas três questões formuladas na introdução. As análises foram organizadas para responder: (i) como evoluíram, ao longo do tempo, as discussões sobre ameaças nos fóruns analisados; (ii) se há padrões sazonais recorrentes ou picos de atividade; e (iii) quais diferenças existem

[1] `https://github.com/sebastiaoafilho/Malicious_Posts_Identification`

entre os ambientes da *Surface Web* e *Dark Web*. No total, foram coletados 52.335 *posts* provenientes de três fóruns da *Dark Web — Hidden Answers*, *Deep Answers* e *Raddle* — e quatro da *Surface Web — Hacknology*, *Legitcarders*, *Niflheim.World* e *Nulledbb*.Esses conteúdos estão disponíveis em português e inglês. A Tabela 1 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. Todo o código usado para a implementação do método assim como o conjunto de dados usado para a validação do mesmo estão disponíveis no repositório do projeto[2].

Tabela 1. 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. 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.

Tabela 2. 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

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 2 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,

[2]https://anonymous.4open.science/r/apiMsc-4F87

optou-se por manter oito tópicos para uma visualização mais detalhada e diversificada dos temas discutidos, permitindo uma análise mais abrangente.

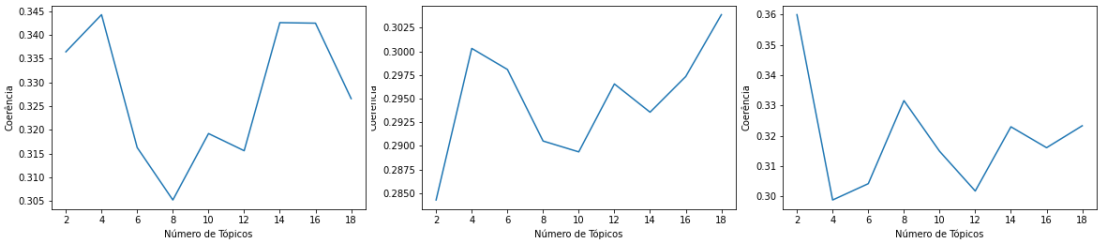

Figura 2. Relação da Coerência com número de tópicos por ambientes. *Surface Web*, *Dark Web* (PT-BR) e *Dark Web* (EN-US) respectivamente.

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 2 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.

5.2. Tendência Temporal dos Tópicos

Para compreender a evolução dos tópicos nos fóruns analisados e responder às duas primeiras questões de pesquisa — que tratam, respectivamente, das discussões sobre ameaças ao longo do tempo e da identificação de padrões sazonais ou picos de atividade —, associou-se a ocorrência dos documentos classificados em determinado tópico ao trimestre de publicação, criando séries temporais por ambiente. Essa abordagem permite identificar tendências, picos de interesse e possíveis reações a eventos externos.

A Figura 3 apresenta a evolução dos tópicos ao longo do tempo nos fóruns da *Dark Web* em português. A Tabela 3 mostra o conteúdo de cada um dos tópicos criados com o LDA. Os principais pontos identificados na análise são apresentados na Tabela 4.

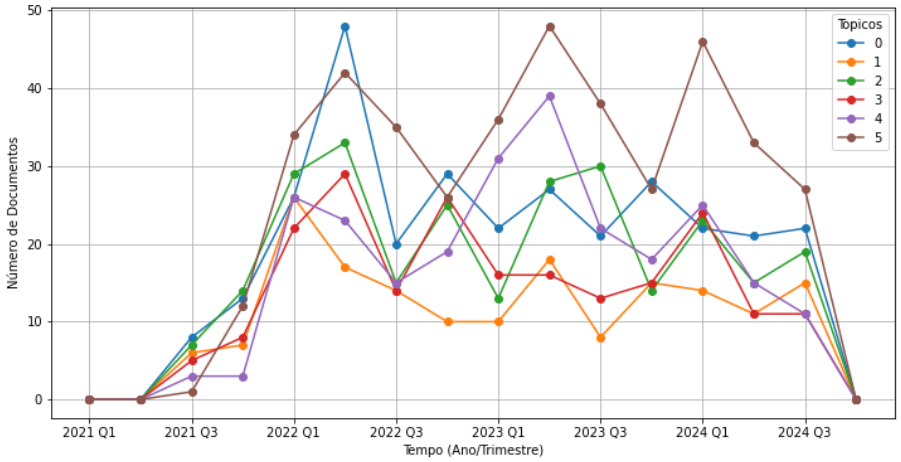

Figura 3. Tendências dos tópicos ao longo do tempo *Dark Web* em português

Tabela 3. Tópicos gerados com LDA - *Dark Web* em português

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

Tabela 4. Resumo da evolução dos tópicos ao longo do tempo da *Dark Web* em português

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.

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

A Figura 4 apresenta a evolução dos tópicos ao longo do tempo em fóruns da *Dark Web* em inglês. Todos os tópicos gerados são mostrados na Tabela 5 e os principais pontos identificados na análise são apresentados na Tabela 6.

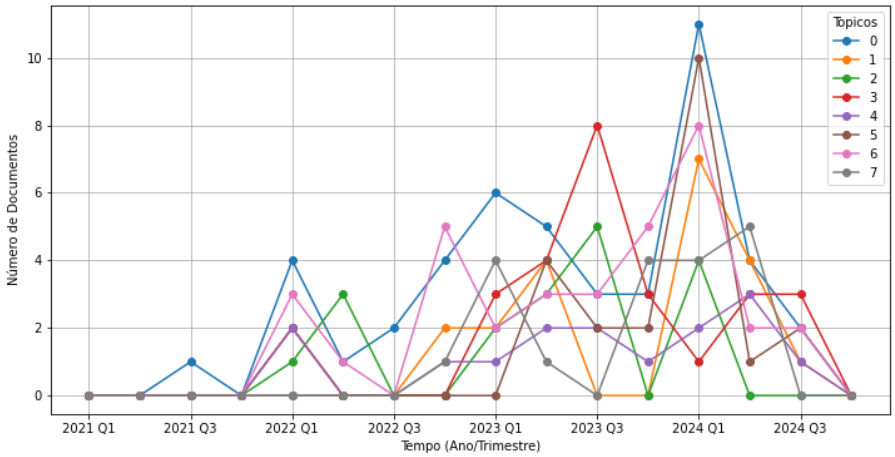

Figura 4. Tendências dos tópicos ao longo do tempo - *Dark Web* em inglês

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

Tabela 5. Tópicos gerados com LDA e suas palavras-chave - *Dark Web* em inglês

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 6. Resumo da evolução dos tópicos ao longo do tempo.

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.

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 5 apresenta a evolução dos tópicos ao longo do tempo em fóruns da *Surface Web*, e todos os tópicos gerados são mostrados na Tabela 7. Os principais pontos identificados na análise são apresentados na Tabela 8.

Tabela 7. Tópicos gerados com LDA - *Surface Web* em inglês

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

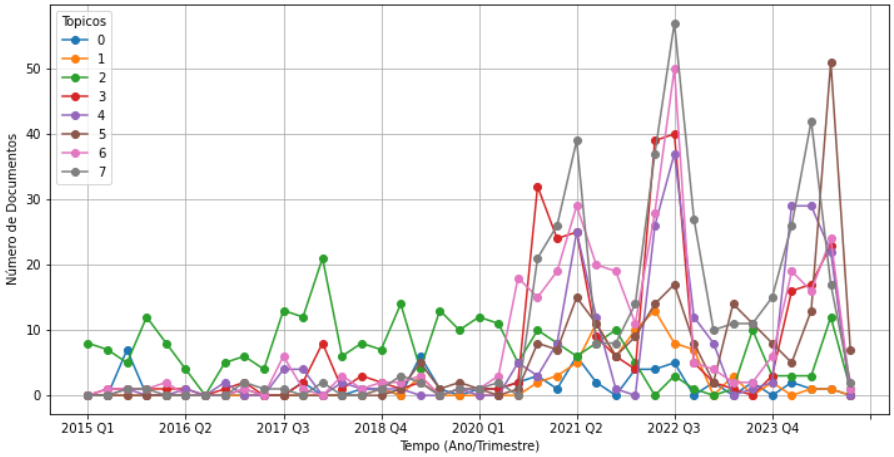

Figura 5. Tendências dos tópicos ao longo do tempo - *Surface Web* em inglês

Tabela 8. Resumo da evolução dos tópicos ao longo do tempo.

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*.

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

5.3. 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 9 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.

5.4. Comparação entre ambientes

Para responder a terceira questão da pesquisa, que busca identificar possíveis diferenças entre a *Dark Web* e a *Surface Web*, a Tabela 10 apresenta uma comparação entre os tópicos extraídos dos três ambientes analisados.

Tabela 9. Palavras mais frequentes por ambiente (Top 10)

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

Tabela 10. Comparação entre Surface Web, Dark Web PT-BR e EN-US

Elemento	Surface Web	Dark Web PT-BR	Dark Web EN-US
Tópicos mais frequentes	T6 (317 docs): Tabela 7 T7 (389 docs): Tabela 7	T0 (307 docs): Tabela 3 T5 (405 docs): Tabela 3	T0 (46 docs): Tabela 5 T6 (34 docs): Tabela 5
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, dado 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 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.

6. Conclusão

Este artigo teve como foco principal a análise da evolução temporal das discussões sobre ameaças cibernéticas na *Surface Web* e na *Dark Web*, no período de 2015 a 2024, utilizando modelagem de tópicos (LDA) e séries temporais. A base de dados foi composta por 11.087 *posts* coletados de três fóruns públicos com alta atividade na *Dark Web* e 41.248 *posts* de quatro fóruns da *Surface Web*. Em relação à primeira questão da pesquisa, observou-se que os mesmos tópicos não são discutidos de forma constante ao longo do tempo. Foram identificados picos de atividade e predominância de determinados temas em períodos específicos, como a comercialização de dados pessoais na *Dark Web* em português e a discussão sobre *phishing* na *Dark Web* em inglês. Em relação à segunda questão, observou-se padrões sazonais e picos de atividade com possível associação a

eventos externos, como vazamentos de dados e campanhas sazonais. Quanto à terceira questão, destacaram-se distinções entre os ambientes analisados. A *Surface Web* apresentou uma postura mais técnica, com discussões focadas em vulnerabilidades e ataques, enquanto a *Dark Web* em português destacou-se pela comercialização de dados pessoais e ações maliciosas.

De forma geral, 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. Entretanto, 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. Ainda assim, compreender essas variações é fundamental para a antecipação de comportamentos maliciosos e para o fortalecimento de estratégias de CTI, promovendo uma atuação mais proativa frente ao cenário de ameaças.

Como trabalho futuro, pretende-se expandir o conjunto de dados, incorporando novas plataformas e idiomas, além de explorar modelos de aprendizado mais avançados para análises aprofundadas. A integração de dados provenientes de mídias sociais e plataformas de mensagens também pode ampliar a capacidade de identificar e responder a ameaças cibernéticas em constante evolução. Adicionalmente, sugere-se a realização de uma análise comparativa entre os tópicos discutidos e dados referentes à incidência de crimes relacionados aos temas abordados, com ênfase no contexto brasileiro.

Referências

Avanzi, B., Tan, X., Taylor, G., and Wong, B. (2023). On the evolution of data breach reporting patterns and frequency in the united states: a cross-state analysis. *arXiv preprint arXiv:2310.04786*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cascavilla, G. (2025). The rise of cybercrime and cyber-threat intelligence: Perspectives and challenges from law enforcement. *IEEE Security & Privacy*, 23(1):17–26.

Cimpanu, C. (2020). University of utah pays $457,000 to ransomware gang. Acessado: 12-04-2023.

Crawly (2021). O que é crawler e como funcionam os robôs para coleta de dados. Acessado: 25-10-2024.

de Jesus Filho, S. A. (2024). Identificação de posts maliciosos na dark web utilizando aprendizado de máquina supervisionado. Dissertação de mestrado, Universidade Federal de Uberlândia, Uberlândia, Brasil. Orientador: Rodrigo Sanches Miani.

Fu, T., Abbasi, A., and Chen, H.-c. (2010). A focused crawler for dark web forums. *JASIST*, 61:1213–1231.

Hickman, L., Thapa, S., Tay, L., Cao, M., and Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146.

Kavallieros, D., Myttas, D., Kermitsis, E., Lissaris, E., Giataganas, G., and Darra, E. (2021). *Understanding the Dark Web*, pages 3–26. Springer International Publishing, Cham.

Koloveas, P., Chantzios, T., Alevizopoulou, S., Skiadopoulos, S., and Tryfonopoulos, C. (2021). intime: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. *Electronics*, 10(7).

Kühn, P., Wittorf, K., and Reuter, C. (2024). Navigating the shadows: Manual and semi-automated evaluation of the dark web for cyber threat intelligence. *IEEE Access*, 12:118903–118922.

Labs, F. (2024). Pesquisa de ameaças da fortinet descobre que os cibercriminosos estão explorando novas vulnerabilidades do setor 43% mais rápido do que no 1º semestre de 2023. Acessado: 13-04-2025.

Liakos, P., Ntoulas, A., Labrinidis, A., and Delis, A. (2015). Focused crawling for the hidden web. *World Wide Web*, 19.

Najork, M. (2009). *Web Crawler Architecture*, pages 3462–3465. Springer US, Boston, MA.

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., Robertson, J., Shakarian, J., Thart, A., and Shakarian, P. (2016). Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 7–12.

Rahman, M. R., Hezaveh, R. M., and Williams, L. (2023). What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. *ACM Comput. Surv.*, 55(12).

Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., and Ferrara, E. (2017). Early warnings of cyber threats in online discussions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 667–674.

Sarkar, S., Almukaynizi, M., Shakarian, J., and Shakarian, P. (2018). Predicting enterprise cyber incidents using social network analysis on the darkweb hacker forums.

Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., and Zhang, J. (2023). Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. *IEEE Communications Surveys & Tutorials*, 25(3):1748–1774.

Syed, S. and Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174.

Tong, Z. and Zhang, H. (2016). A text mining research based on lda topic modelling. In *International conference on computer science, engineering and information technology*, pages 201–210.

Wagner, T. D., Mahbub, K., Palomar, E., and Abdallah, A. E. (2019). Cyber threat intelligence sharing: Survey and research directions. *Computers & Security*, 87:101589.

Řehůřek, R. (2024). What is gensim. Acessado: 27-04-2025.