Detecção de eventos em fóruns da Dark Web e da Surface Web usando algoritmos de aprendizado de máquina

Rodrigo Bernabé Silveira

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

Uberlândia
2026

Rodrigo Bernabé Silveira

Detecção de eventos em fóruns da Dark Web e da Surface Web usando algoritmos de aprendizado de máquina

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: Inteligência Artificial

Orientador: Paulo Henrique Ribeiro Gabriel
Coorientador: Rodrigo Sanches Miani

Uberlândia
2026

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

S587 Silveira, Rodrigo Bernabé, 1985-
2026 Detecção de eventos em fóruns da Dark Web e Surface Web usando aprendizado de máquina [recurso eletrônico] / Rodrigo Bernabé Silveira. - 2026.

Orientador: Paulo Henrique Ribeiro Gabriel.
Coorientador: Rodrigo Sanches Miani.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2026.207
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Gabriel, Paulo Henrique Ribeiro,1984-, (Orient.). II. Miani, Rodrigo Sanches,1983-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

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, 07/2026, PPGCO				
Data:	26 de Fevereiro de 2026	Hora de início:	09:00	Hora de encerramento:	11:15
Matrícula do Discente:	12322CCP024				
Nome do Discente:	Rodrigo Bernabé Silveira				
Título do Trabalho:	Detecção de eventos em fóruns da Dark Web e da Surface Web usando algoritmos de aprendizado de máquina				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	Identificação de ameaças de segurança usando mineração de dados de redes sociais e darkweb. 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: Rodrigo Sanches Miani (Coorientador) - FACOM/UFU, Diego Nunes Molinos - FACOM/UFU, Bruno Bogaz Zarpelão- Departamento de Computação – CCE-UEL e Paulo Henrique Ribeiro Gabriel - FACOM/UFU, orientador do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Bruno Bogaz Zarpelão e o aluno(a) participou da cidade de Londrina/PR. Os outros membros da banca participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Paulo Henrique Ribeiro Gabriel, apresentou a Comissão Examinadora e o(a) candidato(a), agradeceu a presença do público, e concedeu ao(à) 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 27/02/2026, às 09:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Bogaz Zarpelão**, **Usuário Externo**, em 27/02/2026, às 12:02, 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 **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 27/02/2026, às 14:37, 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 Nunes Molinos**, **Professor(a) do Magistério Superior**, em 09/03/2026, às 13:42, 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 **7018069** e o código CRC **25C2DC8E**.

À Joelma, Bolinha e ao Leão, vocês são a minha inspiração.

Agradecimentos

Tenho muito a agradecer a todos aqueles que contribuíram para a conclusão desta etapa na minha vida. Tenho certeza de que não conseguiria sozinho.

Agradeço primeiramente a Deus, que me concedeu a vida e a capacidade de aprendizagem.

Agradeço à minha esposa, Joelma, que sempre esteve ao meu lado e sempre incentivando para que eu nunca desistisse.

Agradeço ao meu irmão, Luciano, que sempre me incentivou e acreditou em mim.

Agradeço também a todos os professores da pós-graduação por compartilharem o conhecimento de vocês comigo.

Por fim, um agradecimento especial ao meu orientador, Professor Paulo Henrique Ribeiro Gabriel e ao meu coorientador, Professor Rodrigo Sanches Miani, que me incentivaram e apoiaram em todas as etapas deste trabalho.

"Não seria uma maneira nobre e esclarecida gastar o nosso breve tempo sob o sol tentando entender o universo e como viemos a despertar nele?"

(Richard Dawkins)

Resumo

A detecção de publicações maliciosas em fóruns da *Dark Web* e da *Surface Web* é amplamente utilizada no campo da cibersegurança. A análise dessas publicações pode auxiliar no desenvolvimento de melhores Sistemas de detecção de intrusões (Intrusion Detection Systems) (IDSs), na identificação de ataques cibernéticos, na descoberta de novos tipos de ameaças e em outros benefícios. Muitas formas de ataques podem ser proativamente prevenidas ao se pesquisar em publicações de fóruns por tópicos relacionados à cibersegurança. Para alcançar esse objetivo, é necessário identificar eventos emergentes em publicações com novos tópicos que possam conter palavras-chave de interesse, sob a perspectiva da cibersegurança. Este trabalho objetiva propor um algoritmo para a busca de novos eventos e analisar publicações de fóruns da *Dark Web* e da *Surface Web* ao longo do tempo. Eventos novos são tópicos emergentes de interesse no domínio da cibersegurança. Para avaliar a relevância dos eventos detectados, um classificador de publicações maliciosas foi empregado para categorizar as publicações como de alta, média ou baixa relevância. O algoritmo proposto demonstrou seu potencial ao identificar 26.071 novos eventos a partir de um conjunto de 30.584 publicações analisadas, dentre os quais aproximadamente 4% foram classificados como de alta relevância.

Palavras-chave: Aprendizado de Máquina, Cibersegurança, Detecção de Eventos, *Dark Web*, *Surface Web*, Fóruns Online, Processamento de Linguagem Natural.

Abstract

The detection of malicious posts in Dark Web and Surface Web forums is widely utilized in the field of digital security. Analyzing these posts can aid in the development of better IDSs, the identification of cyber attacks, the discovery of new types of threats, and other benefits. Many forms of attacks can be proactively prevented by searching forum posts for topics related to digital security. To achieve this, it is necessary to identify emerging events in posts with new topics that may contain keywords of interest from a digital security perspective. This work aims to propose an algorithm for detecting novel events and analyzing publications from Dark Web and Surface Web forums over time. Novel events are defined as emerging topics of interest within the cybersecurity domain. To assess the significance of the detected events, a malicious post classifier was employed to categorize posts as having high, medium, or low relevance. The proposed algorithm demonstrated its potential by identifying 26.071 new events from a dataset of 30.584 analyzed posts, among which approximately 4% were classified as highly relevant.

Keywords: Machine Learning, Cibersecurity, Events Detection, Dark Web, Surface Web, Online Forums, Natural Language Processing.

Lista de ilustrações

Lista de tabelas

Lista de siglas

CTI Inteligência Contra Ameaças Cibernéticas (Cyber Threat Intelligence)

DDoS Negação de Serviço Distribuída (Distributed Denial-of-Service)

PLN Processamento de Linguagem Natural

LDA Alocação Latente de Dirichlet (Latent Dirichlet Allocation)

TF-IDF Frequência de Termo-Inverso da Frequência de Documento (Term Frequency-Inverse Document Frequency)

SJ Similaridade de *Jaccard*

SC Similaridade do *Cosseno*

IDSs Sistemas de detecção de intrusões (Intrusion Detection Systems)

NER Reconhecimento de Entidade Nomeada (Named Entity Recognition)

DBSCAN Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído (Density-based spatial clustering of applications with noise)

HTML Linguagem de Marcação de Hipertexto (HyperText Markup Language)

IoCs Indicadores de Comprometimento (Indicators of Compromise)

CSV Valores separados por vírgula (Comma-separated values)

SBERT Representações de Codificadores Bidirecionais de Transformadores para Sentenças (*Sentence Bidirectional Encoder Representations from Transformers*)

URL Localizador Uniforme de Recursos (Uniform Resource Locator)

HDBSCAN Agrupamento Espacial Hierárquico Baseado em Densidade de Aplicações com Ruído (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

PLMs Modelos de linguagem pré-treinados (Pretrained Language Models)

IPFS Sistema de Arquivos Interplanetário (*InterPlanetary File System*)

OPSEC Segurança de Operações (*Operations Security*)

HTTP Protocolo de Transferência de Hipertexto (*HyperText Transfer Protocol*)

HTTPS Protocolo de Transferência de Hipertexto Seguro (*HyperText Transfer Protocol Secure*)

Sumário

Capítulo 1

Introdução

Desde o surgimento da Internet, o número de pessoas conectadas à rede cresce significativamente. Atividades que antes eram feitas de forma presencial passaram a ser feitas de forma digital, utilizando a Internet (PRIYADARSHINI et al., 2021). Algumas dessas atividades incluem transações financeiras, reuniões, aulas e interação social. Junto às facilidades proporcionadas pela Internet, vieram também os problemas com ataques cibernéticos, que têm aumentado a cada ano (RAMIREZ, 2025), trazendo vários desafios para a área de cibersegurança. Dentre os problemas causados, destacam-se os ataques cibernéticos.

Segundo Asbas e Tuzlukaya (2022), ataques cibernéticos são tentativas de cibercriminosos de obter acesso autorizado a computadores ou sistemas de computadores com o objetivo de roubar, expor ou até mesmo destruir informações. Os ataques cibernéticos também podem ter o objetivo de desabilitar redes e infraestruturas inteiras de computadores. Como formas de ataques, podemos citar: uso de engenharia social, vírus, métodos de *phishing*, Negação de Serviço Distribuída (Distributed Denial-of-Service) (DDoS), entre outros. Esses ataques são geralmente discutidos em fóruns da *Dark Web* (NUNES et al., 2016) ou sites da *Surface Web* como o *X* (antigo *Twitter*) e *Facebook* (SABOTTKE; SUCIU; DUMITRAS, 2015). Alguns ataques foram discutidos antes de acontecerem, a exemplo do *Mirai*, quando uma publicação falando sobre o ataque foi postada em 2016 (SAPIENZA et al., 2017). O mesmo ocorreu com o ataque ao sistema do *Equifax*: alguns meses antes, foi postado um exemplo do ataque na plataforma *GitHub* (APUTUNN, 2023). Esses casos tornam os fóruns e redes sociais ambientes ideais para buscar informações sobre ataques cibernéticos. Assim, ao se fazer uma análise dessas informações, é possível conhecer diversos tipos de ataques diferentes, seus alvos, suas metodologias, tecnologias utilizadas, entre outros dados importantes para a cibersegurança.

Existem diversos trabalhos que utilizaram essas duas fontes de dados (*Dark Web* e *Surface Web*) para busca de publicações que possam conter informações sobre ataques cibernéticos (AMPEL et al., 2020; KOLOVEAS et al., 2019; HUANG et al., 2021; SAPIENZA et al., 2018). Por exemplo, Ampel et al. (2020) criaram um sistema baseado em *transfer*

learning (ZHUANG et al., 2015) para coletar e categorizar automaticamente códigos-fonte encontrados em fóruns de cibersegurança. Essa classificação, realizada pelo sistema, utilizou dados de fóruns da *Dark Web* e repositórios públicos, o que resultou na construção de Inteligência Contra Ameaças Cibernéticas (Cyber Threat Intelligence)s (CTIs) proativas e mais direcionadas. Por sua vez, Koloveas et al. (2019) propuseram uma arquitetura nova de busca e classificação de dados tanto na *Surface Web* quanto na *Dark Web* utilizando aprendizado de máquina supervisionado. Por outro lado, Huang et al. (2021) propuseram um novo sistema chamado *HackerRank* que automaticamente busca e classifica *hackers* em fóruns. O sistema utiliza análise de redes sociais e de conteúdo, e a classificação é feita por meio do algoritmo *PageRank* baseado em topicos (HAVELIWALA, 2003). No trabalho de Sapienza et al. (2018), os autores criaram um sistema chamado *DISCOVER* que busca em fóruns, *blogs* e redes sociais palavras que sinalizem possíveis ataques cibernéticos.

Esses trabalhos ilustram a importância da análise de informações coletadas nesses ambientes (*Dark Web* e *Surface Web*) na sua aplicação direta em estratégias de cibersegurança. Apesar disso, nenhum destes trabalhos fez uma comparação entre as fontes para entender as particularidades de cada uma. Essa comparação é importante porque permite construir sistemas de cibersegurança mais direcionados aos tipos de ataque que se deseja prevenir. Tal abordagem também ajuda na construção de ferramentas de CTI mais assertivas. Portanto, a motivação principal deste trabalho é oferecer o conhecimento necessário para direcionar, de forma mais assertiva, a coleta e a análise de dados, otimizando os sistemas de cibersegurança contra ameaças emergentes.

1.1 Objetivos e Hipótese

Este trabalho tem como objetivo a busca e a análise de eventos de ameaças cibernéticas em duas fontes distintas: fóruns da *Surface Web* e da *Dark Web*. Propõe-se uma investigação comparativa e quantitativa para determinar quais informações críticas podem ser extraídas de cada ambiente. A hipótese que sustenta este trabalho é que a utilização de técnicas de Processamento de Linguagem Natural (PLN), como a representação vetorial de textos e o calculo de similaridade semântica, permite não apenas detectar novos eventos de ataques cibernéticos, mas também diferenciá-los e correlacioná-los em relação à sua origem.

Para a busca, foi desenvolvido um método baseado no trabalho de Bose et al. (2020) que utiliza algoritmos de agrupamento e PLN para identificar o surgimento de eventos novos. Tais eventos referem-se a assuntos emergentes que surgem ao longo do tempo. A análise desses eventos encontrados justifica-se pela necessidade de compreender a correlação entre as fontes e o nível de periculosidade dos dados encontrados. Essa análise também permite identificar se uma ameaça é exclusiva de um ambiente ou se apresenta propagação entre eles. Além disso, foi aplicado um método desenvolvido por Jesus Filho (2024) para

a classificação dos eventos em relação à sua significância para a cibersegurança.

1.2 Organização da Dissertação

Esta dissertação está organizada da seguinte forma: O Capítulo 2 apresenta a fundamentação teórica e trabalhos relacionados. No Capítulo 3, é detalhada a proposta metodológica, incluindo todas as suas etapas. No Capítulo 4, são apresentados e discutidos os resultados obtidos e, por fim, o Capítulo 5 apresenta as conclusões e contribuições deste trabalho.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os conceitos essenciais para o desenvolvimento do projeto, seguida da descrição dos principais trabalhos correlatos ao tema. A Seção 2.1 dá uma visão geral sobre *Surface Web* e *Dark Web*. A Seção 2.2 fala sobre mineração de texto e PLN. Já a Seção 2.3 fala sobre aprendizado não supervisionado. A Seção 2.4 apresenta métricas de similaridade e avaliação. E por fim, a Seção 2.5 mostra os trabalhos relacionados.

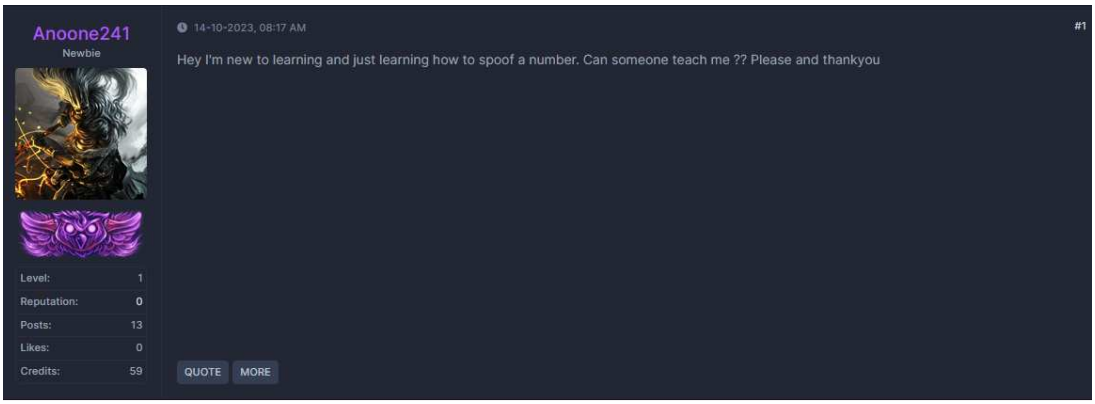
2.1 *Surface Web* e *Dark Web*

Segundo Liu et al. (2020), *Surface Web* é todo site que os usuários podem acessar por meio de navegadores convencionais, utilizando protocolos padrao como Protocolo de Transferência de Hipertexto (*HyperText Transfer Protocol*) (HTTP) ou sua versão segura Protocolo de Transferência de Hipertexto Seguro (*HyperText Transfer Protocol Secure*) (HTTPS), sem a necessidade de ferramentas de anonimato ou credenciais específicas. Apesar de corresponder a somente 0,03% de todos os sites da *Internet* (MEAD; AGARWAL, 2022), na *Surface Web* são encontradas as principais redes sociais como o *Facebook*, *X/Twitter*, *Instagram*, etc. Esses são os sites mais utilizados no cotidiano de usuários comuns. Todos os sites indexados por mecanismos de busca convencionais (*e.g.*, Google, Bing e Yahoo!) são encontrados na *Surface Web*. Por concentrar a maior parte do tráfego das redes sociais, a *Surface Web* é uma rica fonte de informações que podem ser úteis para o treinamento de sistemas de cibersegurança.

Já a *Dark Web* é composta por páginas que não são indexadas pelos buscadores. De acordo com Mead e Agarwal (2022), a *Dark Web* corresponde a cerca de 97% de todas as páginas web existentes. Nela é possível encontrar vários fóruns *hackers*, ou seja, fóruns de discussão online onde são postadas mensagens relacionadas a ferramentas de *hacking*, técnicas de ataques, códigos-fonte e outro (BENJAMIN et al., 2015). Esses fóruns são locais onde vazamentos de dados e o ensino de técnicas de *hacking* são amplamente divulgados, sendo assim um ótimo ambiente para coleta de dados para análise e melhoria de sistemas de cibersegurança também. Alguns exemplos de publicações extraídas do

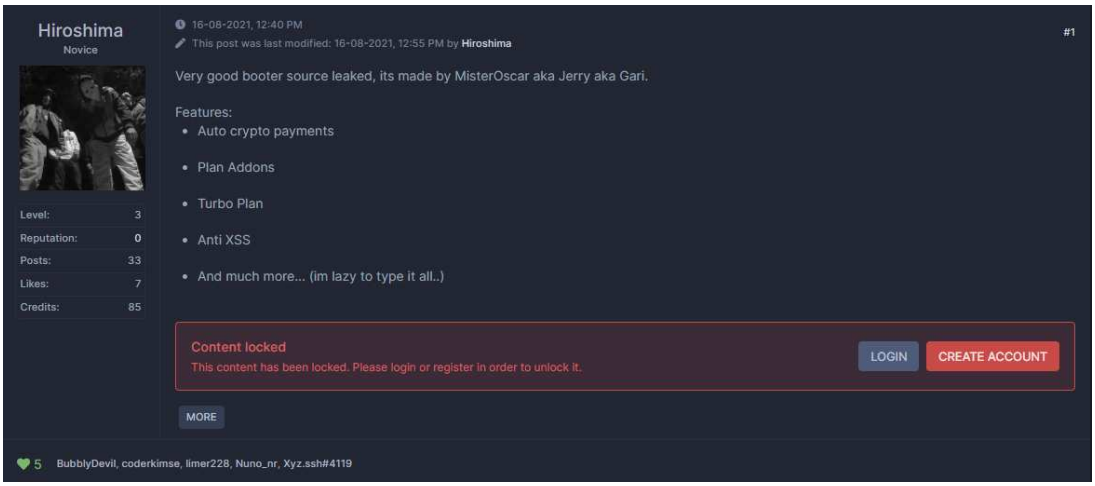
fórum *Nulledbb* que são de interesse deste trabalho podem ser vistos nas Figuras 1 e 2. Ambas mostram publicações com conteúdo que diz respeito a ataques cibernéticos.

Figura 1 – Publicação que mostra um usuário perguntando sobre técnicas de falsificação de números de telefone.



Fonte: Autoria própria.

Figura 2 – Publicação com divulgação de *software* usado em *hacking*.



Fonte: Autoria própria.

2.2 Mineração de texto e processamento de linguagem natural

A mineração de texto é uma variação da área de mineração de dados que busca padrões em grandes volumes de dados textuais (GAIKWAD; CHAUGULE; PATIL, 2014). A mineração de textos lida com dados não estruturados, ou seja, informações que não apresentam um formato predefinido. São utilizadas frequentemente diversas fontes textuais, como *e-mails*, arquivos em formato PDF e publicações em redes sociais, entre outros

ambientes. Cada uma dessas fontes de dados possui seus próprios padrões. Para a extração de informação útil desses dados, é necessário realizar um pré-processamento do texto, utilizando técnicas PLN. A seguir, serão descritas algumas etapas de pré-processamento comumente empregadas em tarefas de mineração de texto (DENNY; SPIRLING, 2018):

1. **Pontuação:**
   Muitas vezes os caracteres especiais não são importantes para a análise do texto, portanto são removidos caracteres como: $, %, &, etc. Também são removidas as etiquetas (*tags*) em Linguagem de Marcação de Hipertexto (HyperText Markup Language) (HTML) (como `<h>`, `<p>`, `<body>`, etc) e espaços em branco dos textos.

2. **Números:**
   Os dígitos numéricos são removidos do texto caso não sejam relevantes para a análise, por exemplo, números de telefone, datas ou valores monetários.

3. **Letras minúsculas:**
   Todas as palavras do texto são transformadas em minúsculas. Isso se faz necessário pois o algoritmo utilizado para processar os textos pode diferenciar as palavras "Elefante" e "elefante" por exemplo, e isso trará um comportamento errôneo no processamento.

4. ***Stemming*:**
   A técnica de *stemming* ou redução ao radical, consiste em reduzir as palavras à sua forma mais básica através da remoção de seus sufixos e prefixos. Por exemplo: As palavras: **festa**, **festejar** e **festeiro** possuem o mesmo radical: **festa**. Por possuir o mesmo radical, estas três palavras possuem o mesmo significado semântico e por isso são todas substituídas pelo seu radical, pois se não forem substituídas elas serão tratadas como palavras com significado diferente no texto.

5. **Remoção de *Stopwords*:**
   Stopwords ou palavras irrelevantes são palavras de alta frequência em um texto cuja exclusão não irá afetar semanticamente a frase. Cada língua possui seu próprio conjunto de *stopwords*; no português, podem ser citados os exemplos: **é**, **de**, **das**, **em**, **eu**, etc. A remoção desses termos de alta frequência é importante para diminuir o tempo de processamento dos algoritmos.

Após o pré-processamento dos dados, o texto estará pronto para a extração de características e de informação útil, principal objetivo da mineração de texto. Para isso, são utilizadas técnicas de PLN, como a modelagem de tópicos utilizando Alocação Latente de Dirichlet (Latent Dirichlet Allocation) (LDA), análise de frequência das palavras utilizando o cálculo de Frequência de Termo-Inverso da Frequência de Documento (Term Frequency-Inverse Document Frequency) (TF-IDF) e extração de palavras-chave utilizando *TextRank*.

LDA é um modelo probabilístico que extrai tópicos de uma coleção de textos (KIM; GIL, 2019). Segundo Blei, Ng e Jordan (2003), o modelo assume que documentos são misturas de tópicos latentes, sendo cada tópico definido por uma distribuição entre as palavras. O algoritmo recebe como entrada uma matriz termo-documento (geralmente representada por uma *Bag-of-Words*) e o hiperparâmetro K, que define a quantidade de tópicos a serem extraídos. Esta abordagem dispensa o treinamento prévio, uma vez que utiliza aprendizado não supervisionado através de análise estatística, como a amostragem de Gibbs, para examinar cada palavra do texto e definir sua probabilidade de pertencer a um tópico.

Ao tomar como exemplo o seguinte texto: "O algoritmo de aprendizado de máquina foi usado para analisar a massa e a farinha. O modelo processou os dados e verificou se a receita tinha a temperatura certa. As redes neurais ajudaram a classificar a qualidade do pão no forno.", e após a aplicação do LDA para encontrar dois tópicos ($K = 2$), o resultado dos tópicos pode ser visualizado na Tabela 1. Com base nesses resultados, um pesquisador humano poderia definir uma palavra mais genérica como tópico para cada resultado, sugerindo, por exemplo, "Inteligência Artificial" para o primeiro tópico e "Cozinha" para o segundo.

Tabela 1 – Exemplo de aplicação do LDA

Tópico	**Principais Palavras**
1	algoritmo, máquina, analisar, modelo, dados, redes neurais
2	massa, farinha, receita, temperatura, pão, forno

Fonte: Autoria própria.

O TF-IDF é uma técnica estatística que calcula a frequência de uma palavra em um grupo de documentos. A partir deste cálculo, infere-se que uma palavra com alta frequência em um documento, mas baixa frequência nos demais, é considerada importante para o texto (KIM; GIL, 2019). O termo TF (*Term Frequency*) calcula a frequência de cada palavra em um documento. Para o cálculo do TF, é necessário dividir a quantidade de vezes que determinada palavra aparece no texto pela quantidade total de palavras deste documento, conforme a Equação (1), na qual p é a palavra e D representa a coleção de documentos.

$$TF(p) = \frac{quantidade(p)}{quantidade(D)} \tag{1}$$

Já o termo IDF (*Inverse Document Frequency*) calcula a frequência de uma palavra entre todos os documentos. Esse cálculo indica se uma palavra é rara ou comum no grupo de documentos. Uma palavra com valor de IDF alto é, portanto, considerada rara e, consequentemente, relevante para o documento que a contém. O cálculo do IDF se dá pelo logaritmo da divisão entre a quantidade total de documentos e a quantidade de documentos que possuem a palavra em análise, conforme a Equação (2), na qual p é a

palavra, D é a coleção de documentos e Dp são os documentos que possuem a palavra p. O cálculo total do TF-IDF corresponde ao resultado da multiplicação entre os resultados do TF e do IDF.

$$IDF(p) = \log\left(\frac{quantidade(D)}{quantidade(Dp)}\right) \tag{2}$$

O *TextRank* é um algoritmo baseado no *PageRank* (PAGE et al., 1998) do Google, que visa extrair palavras relevantes de um texto (MIHALCEA; TARAU, 2004). O objetivo do algoritmo é construir um grafo com as palavras do texto e, em seguida, definir uma pontuação de relevância para cada termo, com base na pontuação de seus vizinhos.

Duas palavras são consideradas ligadas se coocorrem (aparecem próximas) no texto. Essa distância é definida pelo usuário por meio do tamanho de uma janela de coocorrência. Por exemplo, se o tamanho da janela for definido como 3, cada palavra será ligada às palavras que estiverem dentro desse limite. Depois de todas as ligações estabelecidas, o algoritmo calcula uma pontuação para cada palavra baseada em sua conectividade e na relevância dos termos ligados a ela. As palavras com a maior pontuação são consideradas as palavras-chave do texto.

2.2.1 Vetorização de Textos

Vetorização de textos é a transformação de dados textuais em representações numéricas, em formato de vetores, conhecidos como *embeddings*. Esse processo é amplamente empregado em PLN em tarefas como classificação de texto e análise de sentimentos, por exemplo. Essa transformação visa não apenas converter as palavras em um formato que os computadores entendam, mas também preservar sua representação semântica. Segundo Birunda e Devi (2021), as palavras são vetorizadas de forma que, no espaço vetorial, palavras que compartilham similaridade semântica terão vetores mais próximos (distância menor). Por exemplo, as palavras “computador” e “máquina” estarão mais próximas no espaço vetorial do que as palavras “computador” e “futebol”.

Ainda segundo Birunda e Devi (2021), as técnicas de vetorização de textos podem ser categorizadas em três principais tipos:

1. **Vetorização baseada em frequência:**
   Essa técnica analisa a frequência com que a palavra aparece no texto, permitindo identificar termos raros e os mais frequentes. Essa técnica é empregada no algoritmo TF-IDF.

2. **Vetorização estática:**
   Essa técnica gera os vetores por meio da análise de contexto das palavras. Palavras que aparecem no mesmo contexto tendem a estar próximas no espaço vetorial. Ela é utilizada em algoritmos como *Word2Vec*, *Glove* e *FastText*, por exemplo. A

vetorização é considerada estática porque, após a geração dos vetores, eles permanecem imutáveis, independentemente do contexto de uso. Isso significa que, mesmo que a palavra seja utilizada em um contexto diferente, ela terá o mesmo significado aprendido pelo algoritmo.

3. **Vetorização contextualizada:**
   Nesta abordagem, o contexto de cada palavra é levado em consideração e o vetor correspondente varia de acordo com a sua ocorrência na sentença. Diferentemente da vetorização estática, modelos baseados em arquiteturas modernas (como *Transformers*) utilizam mecanismos de atenção para analisar a relação de uma palavra com todas as outras da frase simultaneamente. Isso permite resolver problemas de polissemia: por exemplo, a palavra "banco" terá uma representação vetorial distinta na frase "O banco da praça está quebrado" (sentido de assento) em comparação à frase "O banco aprovou o empréstimo" (sentido de instituição financeira). Essa capacidade de gerar representações dinâmicas torna esses modelos superiores em tarefas que exigem compreensão profunda da semântica, sendo usada em modelos como ELMo, GPT-2 e BERT.

2.3 Aprendizado não supervisionado

O aprendizado não supervisionado é uma técnica de aprendizado de máquina cujos algoritmos recebem apenas os dados de entrada, sem informações sobre as saídas desejadas (GHAHRAMANI, 2004). Dessa forma, o algoritmo precisa descobrir e absorver os padrões presentes nos dados de entrada por conta própria para categorizar ou rotular os dados. O objetivo do algoritmo é aprender as associações entre os dados sem a necessidade de um "professor" que ensine tais relações. O resultado de um algoritmo de aprendizado não supervisionado auxilia na compreensão das ligações entre os dados, o que otimiza a tomada de decisões acerca do problema que se está tentando resolver. Os principais exemplos de algoritmos de aprendizado não supervisionado, juntamente com a redução de dimensionalidade, são os algoritmos de agrupamento (GHAHRAMANI, 2004), cujos conceitos são abordados a seguir.

O agrupamento é uma técnica de aprendizado não supervisionado na qual o algoritmo descobre as associações entre os dados por meio de cálculo de distâncias entre os dados para aprender a relação entre eles. Visto que o cálculo de similaridade depende da mensuração de distâncias entre os pontos, os dados são geralmente convertidos para uma representação numérica vetorial. No entanto, destaca-se que métricas de distância também podem ser aplicadas diretamente sobre outras estruturas, como cadeias de caracteres ou grafos, dependendo da natureza do algoritmo e da métrica escolhida. A métrica de distância mais utilizada nos algoritmos de agrupamento é a euclidiana (Equação (3)), mas também podem ser usadas outras distâncias como a distância de Manhattan (Equa-

ção (4)) e distância de Minkowski (Equação (5)), que generaliza as outras duas. Nessas três equações, os termos representam o seguinte: x e y são os pontos entre os quais se calcula a distância; k representa a k-ésima coordenada do ponto e r representa a ordem ou grau de distância entre os pontos.

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{3}$$

$$d(x,y) = \sum_{k=1}^{n}|x_k - y_k| \tag{4}$$

$$d(x,y) = \left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r} \tag{5}$$

O objetivo primário de algoritmos de agrupamento é organizar os dados em conjuntos chamados *clusters*. Cada *cluster* contém dados que são semelhantes entre si, ou seja, possuem uma distância média pequena (coesão) entre eles e uma distância média grande (separação) em relação aos dados dos demais grupos. A qualidade do agrupamento é determinada pela maximização da separação e pela minimização da coesão. Um exemplo de agrupamento pode ser visto na Figura 3, em que observamos os pontos vermelhos e azuis. Eles foram separados em dois grupos, porque os pontos vermelhos possuem características em comum, ou seja, a distância euclidiana média entre eles é pequena e é grande em relação a qualquer ponto azul da base de dados. Analogamente, os dados azuis possuem uma distância média pequena entre si e uma grande distância média entre os pontos vermelhos.

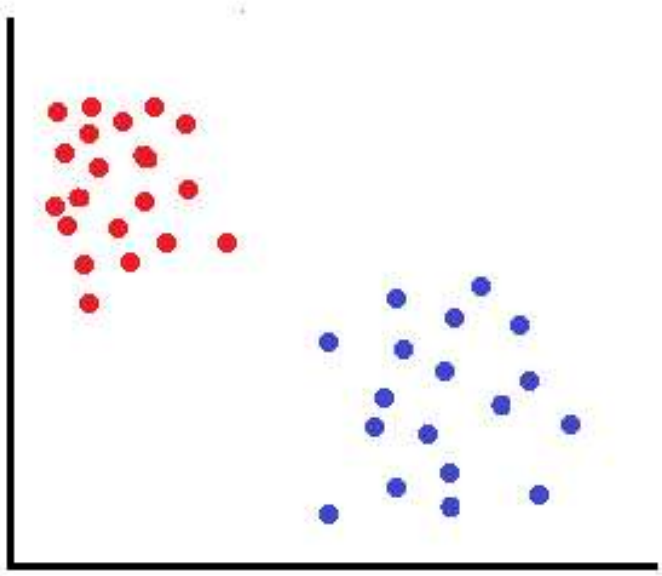
Um algoritmo de agrupamento bastante usado na literatura devido à sua simplicidade é o *k-means* (TANOVIĆ; VRANJKOVIĆ, 2024). O objetivo do *k-means* é determinar, de forma iterativa, os pontos chamados centroides. Os centroides são pontos no espaço de dados nos quais os pontos mais próximos a eles possuem uma distância média menor do que a distância em relação a qualquer outro centroide. Consequentemente, todos os pontos próximos a um centroide formam um *cluster*. No *k-means* é necessário informar inicialmente a quantidade de *clusters* e o algoritmo gera essa mesma quantidade informada de centroides de forma aleatória. Todos os passos do algoritmo *k-means* podem ser vistos a seguir (PATEL; THAKRAL, 2016):

1. Definir a posição inicial dos centroides.

2. Atribuir cada ponto de dados ao *cluster* mais próximo por meio do cálculo de distâncias.

3. Recalcular a posição de cada centroide para o valor médio de todos os pontos de dados que se enquadram nesse *cluster*.

Figura 3 – Exemplo de agrupamento.



Fonte: Autoria própria.

4. Repetir as etapas 2 a 3 até que todos os objetos "permaneçam atribuídos" ao mesmo *cluster* (ou até que a mudança de centroides seja inferior a um limiar).

Pela sua simplicidade, esse algoritmo apresenta algumas desvantagens, incluindo: a necessidade de informar a priori a quantidade de *clusters* que se deseja procurar; a sensibilidade a dados fora do padrão (*outliers*); a dependência da forma de inicialização dos centroides originais; e, devido ao modo como são calculadas as distâncias, apenas é possível encontrar *clusters* com forma hiperesférica.

2.4 Métricas de Similaridade e Avaliação

Durante a busca de eventos e avaliações dos resultados, serão utilizadas algumas métricas de similaridade para uma avaliação estatística dos dados. Essas métricas serão explicadas a seguir.

2.4.1 Similaridade de Cosseno

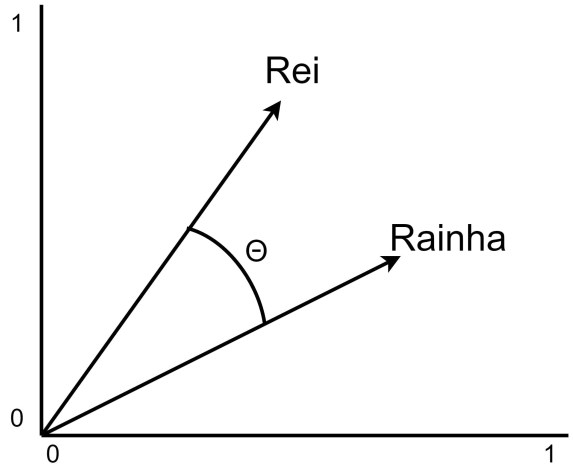
A Similaridade do *Cosseno* (SC) é uma métrica de similaridade entre dois vetores que utiliza o cálculo do cosseno do ângulo entre eles. Nessa métrica, quanto mais o valor do cosseno se aproxima de 1,0, maior é a semelhança entre os dois vetores comparados. A

Equação (6) mostra como o cálculo é realizado, sendo A e B os vetores a serem comparados.

$$sim(A,B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{6}$$

No contexto de PLN, a similaridade por cosseno começou a ser utilizada com o trabalho de Salton, Wong e Yang (1975) para medir a similaridade entre duas palavras no espaço vetorial, como mostra a Figura 4, em que são mostrados os vetores das palavras "rei" e "rainha" e o cosseno do ângulo entre eles. No trabalho atual, a SC também foi calculada utilizando vetores gerados semanticamente pelo algoritmo *FastText*. Essa abordagem foi adotada para permitir uma análise mais completa dos dados, englobando tanto a similaridade estatística quanto a semântica.

Figura 4 – Exemplo de diferença de cosseno entre duas palavras.



Fonte: Autoria própria.

2.4.2 Similaridade de Jaccard

A Similaridade de *Jaccard* (SJ) é uma métrica utilizada para quantificar a similaridade entre dois conjuntos de dados (NIWATTANAKUL et al., 2013). O resultado do cálculo é obtido a partir da razão entre a interseção dos dois conjuntos (o total de elementos em comum) e a união entre os dois conjuntos (os elementos únicos no total). Portanto, quanto maior a sobreposição entre os dois conjuntos, mais próximo de 1,0 será o resultado. A equação pode ser vista na Equação (7), na qual A e B representam conjuntos de palavras. No contexto da PLN, ela mostra o quão similares são dois conjuntos de textos.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

2.4.3 Similaridade de *Dice*

Assim como a SJ, a similaridade de *Dice* também é uma métrica baseada em conjunto, usada para medir o quanto dois textos se sobrepõem. Essa métrica dá um peso duplo a intersecção, se tornando, menos sensível à variação do tamanho dos textos que estão sendo comparados no cálculo (DICE, 1945; SØRENSEN, 1948). A equação de *Dice* pode ser vista na Equação (8), na qual A e B representam conjuntos de palavras.

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{8}$$

2.4.4 Coeficiente de sobreposição (*Overlap Coefficient*)

O coeficiente de sobreposição é uma métrica que ajuda a entender o quanto um conjunto menor está contido em um conjunto maior. É uma métrica muito usada para saber o quanto um texto foi reutilizado em outro. A fórmula pode ser vista na Equação (9), na qual A e B representam conjuntos de palavras.

$$\text{Sobreposição}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{9}$$

2.5 Trabalhos Relacionados

É possível encontrar diversos trabalhos na literatura que usam dados coletados em redes sociais e fóruns da *Dark Web*. A seguir, são apresentados alguns desses trabalhos em ordem cronológica e separados por fontes de dados.

2.5.1 *Surface Web* como fonte de dados

Lee et al. (2017) propuseram o *Sec-Buzzer*, um serviço online para mineração de tópicos emergentes em cibersegurança, a partir de postagens coletados automaticamente no *X/Twitter* de especialistas da área. O sistema identificou com sucesso os ataques *KeyRanger* e *Triada*, embora desconsidere dados provenientes da *Dark Web*, os quais poderiam aprimorar sua eficácia. De forma semelhante, Le Sceller et al. (2017) desenvolveram o *SONAR*, sistema que detecta eventos de cibersegurança no *X/Twitter* com base em palavras-chave atualizadas dinamicamente durante a coleta. Essa ferramenta identificou incidentes como o vazamento de dados do *Yahoo* e o ataque à *Dyn* em 2016. Apesar dos bons resultados, ambas as abordagens estão restritas à *Surface Web*.

Li et al. (2017) propuseram uma abordagem para a classificação de eventos emergentes com base em dados do *X/Twitter*, utilizando categorias semânticas como nomes próprios, *hashtags*, menções, localizações e links. O método apresentou um ganho de 1% nas métricas *NMI* e *B-Cubed*. No entanto, ao focar em eventos genéricos, a abordagem pode

negligenciar aspectos específicos de incidentes de cibersegurança. De maneira complementar, Shi et al. (2017) introduziram o modelo HEE (*Hot Event Evolution*), que incorpora a evolução do interesse dos usuários na detecção de eventos. Utilizando 126.995 postagens do *X/Twitter*, os experimentos demonstraram uma elevação no desempenho do modelo de 66,58% para 83,75% com a exclusão de eventos de baixa popularidade. Apesar dos avanços, o estudo também se restringe a eventos genéricos e desconsidera dados provenientes da *Dark Web*.

Troudi et al. (2018) argumentam que a limitação a uma única fonte de dados pode comprometer a eficácia na detecção de eventos. Para mitigar esse problema, empregaram dados provenientes do *X/Twitter*, *Google Plus* e *YouTube*, processados com o *framework* Hadoop. O método alcançou uma medida-f global de 0,76, enquanto os resultados individuais por fonte foram inferiores: 0,48, 0,32 e 0,38, respectivamente. Apesar da diversidade de fontes, todas pertencem exclusivamente à *Surface Web*.

Já o estudo de Hasan, Orgun e Schwitter (2019) aborda o desafio da detecção de eventos em tempo real, destacando o alto custo computacional devido ao grande volume de dados. Para mitigar esse problema, os autores propuseram o sistema *TwitterNew+*, que utiliza clusterização para reduzir o custo da análise. Nos experimentos, o sistema processou 17 milhões de *tweets* em 212 minutos, alcançando uma taxa de 1.336 *tweets* por segundo, enquanto a média de criação de *tweets* na plataforma era de 6.000 por segundo. Apesar do bom desempenho, *TwitterNew+* foi projetado exclusivamente para dados do *X/Twitter*, mas sua abordagem pode ser adaptada para outras bases.

Bose et al. (2020) propuseram um método para identificar eventos emergentes ou em desenvolvimento relacionados à cibersegurança, a partir da análise de postagens coletadas no *X/Twitter*. A abordagem baseia-se no agrupamento por similaridade e na análise da evolução temporal de palavras-chave extraídas dos textos. Com 21.368 postagens, o método obteve 93,75% de precisão e uma taxa de verdadeiro positivo de 75%. Apesar do desempenho expressivo, o estudo restringe-se a dados da *Surface Web*, cujas palavras-chave podem divergir das utilizadas na *Dark Web*.

Os experimentos conduzidos por Ayan et al. (2022) exploraram a busca de eventos de cibersegurança por meio de diferentes métodos de representação vetorial e algoritmos de aprendizado de máquina. O estudo testou TF-IDF, *Word2Vec* e *Representações de Codificadores Bidirecionais de Transformadores para Sentenças (Sentence Bidirectional Encoder Representations from Transformers) (SBERT)*, juntamente com cinco técnicas de classificação, incluindo *BERT*, Regressão Logística, *SVM*, *Random Forest* e redes neurais profundas. A análise foi realizada em 15.800 postagens do *X/Twitter*, extraídos de 17 contas ligadas à cibersegurança. Entre as abordagens avaliadas, a combinação de *BERT* com *SBERT* obteve os melhores resultados. No entanto, o estudo se restringiu a dados da *Surface Web*, sem aprofundar a análise dos eventos encontrados ou considerar informações provenientes da *Dark Web*.

Dale, McClanahan e Li (2023) desenvolveram um sistema para identificar eventos a partir de postagens no *X/Twitter*, correlacionando dados por meio da extração de nomes de entidades e formando agrupamentos de textos sobre diferentes eventos. Os testes foram realizados com 320.351 postagens provenientes da base *Sentiment-140* e 119 contas relacionadas à cibersegurança. O sistema atingiu 98,6% de acurácia e *AUC-ROC* na detecção de eventos, além de 88,55% de precisão no reconhecimento de entidades e 84,41% de acurácia na organização dos dados. Apesar dos excelentes resultados, o estudo se restringiu a dados da *Surface Web*, sem considerar informações da *Dark Web*.

Por fim, o estudo de Cui et al. (2025) propôs um novo método para criar representações vetoriais de textos voltadas à busca de eventos de cibersegurança. Chamado *Tweezer*, o método constrói um grafo conectando textos às suas entidades de cibersegurança, aproximando textos relacionados dentro da estrutura. Nos testes, foram analisados 1.566 postagens sobre 138 eventos de cibersegurança distintos, e o sistema demonstrou desempenho superior em comparação com TF-IDF, *Word2Vec* e *Modelos de linguagem pré-treinados (Pretrained Language Models) (PLMs)*. Apesar do potencial da abordagem, ela ainda não foi expandida para análise de dados da *Dark Web*.

2.5.2 *Dark Web* como fonte de dados

Poucos trabalhos têm sido desenvolvidos focando em dados extraídos exclusivamente de fóruns da *Dark Web*. Nesse contexto, merece destaque o modelo de identificação de tendências de tópicos em fóruns de cibersegurança, desenvolvido por Hughes et al. (2020).

Nesse trabalho, os autores realizaram a análise de variações linguísticas ao longo do tempo, utilizando 42 milhões de postagens do *HackForums*. Foram comparadas matrizes TF-IDF geradas em diferentes janelas temporais, extraindo os 10 termos mais relevantes. A avaliação realizada por três anotadores humanos alcançou uma concordância de 0,833, segundo coeficiente alfa de Krippendorff (KRIPPENDORFF, 1970). Embora os resultados tenham sido satisfatórios, o método não foi avaliados com dados da *Surface Web*, cuja dinâmica lexical pode diferir substancialmente.

2.5.3 *Dark Web* e *Surface Web* como fonte de dados

Sapienza et al. (2018) desenvolveram o sistema *Discover*, que identifica eventos de ameaças cibernéticas a partir de palavras-chave extraídas de mídias sociais, blogs de cibersegurança e fóruns da *Dark Web*. A análise, realizada com dados coletados entre setembro de 2016 e janeiro de 2017, demonstrou 84% de precisão com dados do *X/Twitter* e 59% com blogs de cibersegurança, além de detectar menções ao ataque *Wannacry* antes de sua ocorrência. Contudo, o estudo não diferenciou explicitamente os termos provenientes da *Dark Web* e da *Surface Web*.

2.5.4 Considerações Finais

Embora os trabalhos supracitados tenham apresentado bons resultados, a maioria se concentra em apenas um tipo de base de dados, desconsiderando as diferentes características dos eventos presentes em bases de dados variadas.

Este trabalho, por sua vez, propõe abordar essa lacuna por meio de métodos recentes de representação de dados (*SBERT*) além de garantir a reprodutibilidade do método ao compartilhar o conjunto de dados utilizado. A Tabela 2 sumariza os trabalhos relacionados e os compara a partir das seguintes características: métodos usados para a identificação de eventos, representação de texto, pré-processamento, base de dados e disponibilização dos dados.

Tabela 2 – Visão comparativa dos trabalhos relacionados e deste estudo

Trabalho	**Identificação de Eventos**				**Representação de Texto**		**Pré-processamento**	**Fonte**	**Disp. Dados**[1]	**Escopo**
	SC	K-Means	*Text Rank*	NER	TF-IDF	*SBERT*				
Este Trabalho	✓	✓	✓	✗	✗	✓	Remoção de *stopwords*, URLs e caracteres especiais.	Fóruns *hackers* da *Surface* e *Dark Web*	✓	Cibersegurança
(LEE et al., 2017)	✓	✓	✗	✗	✗	✗	Não informado.	X/Twitter e blogs de cibersegurança	✗	Cibersegurança
(LE SCELLER et al., 2017)	✗	✗	✗	✗	✓	✗	Remoção de *stopwords* e caracteres especiais.	X/Twitter	✗	Cibersegurança
(LI et al., 2017)	✓	✗	✗	✗	✓	✗	Remoção de postagens em outras línguas e *spam*.	X/Twitter	✗	Detecção de eventos
(SHI et al., 2017)	✓	✗	✗	✗	✓	✗	Remoção de *stopwords*.	X/Twitter	✗	Evolução de eventos
(SAPIENZA et al., 2018)	✗	✗	✗	✗	✗	✗	Remoção de URLs, símbolos e números.	X/Twitter, blogs e *Dark Web*	✗	Cibersegurança
(TROUDI et al., 2018)	✗	✗	✗	✓	✗	✗	Normalização, filtragem e lematização.	X/Twitter, Google+ e YouTube	✗	Eventos em Big Data
(HASAN et al., 2019)	✓	✗	✗	✗	✓	✗	Remoção de *spam*, URLs e *stopwords*.	X/Twitter	✗	Detecção de eventos

Fonte: Autoria própria.

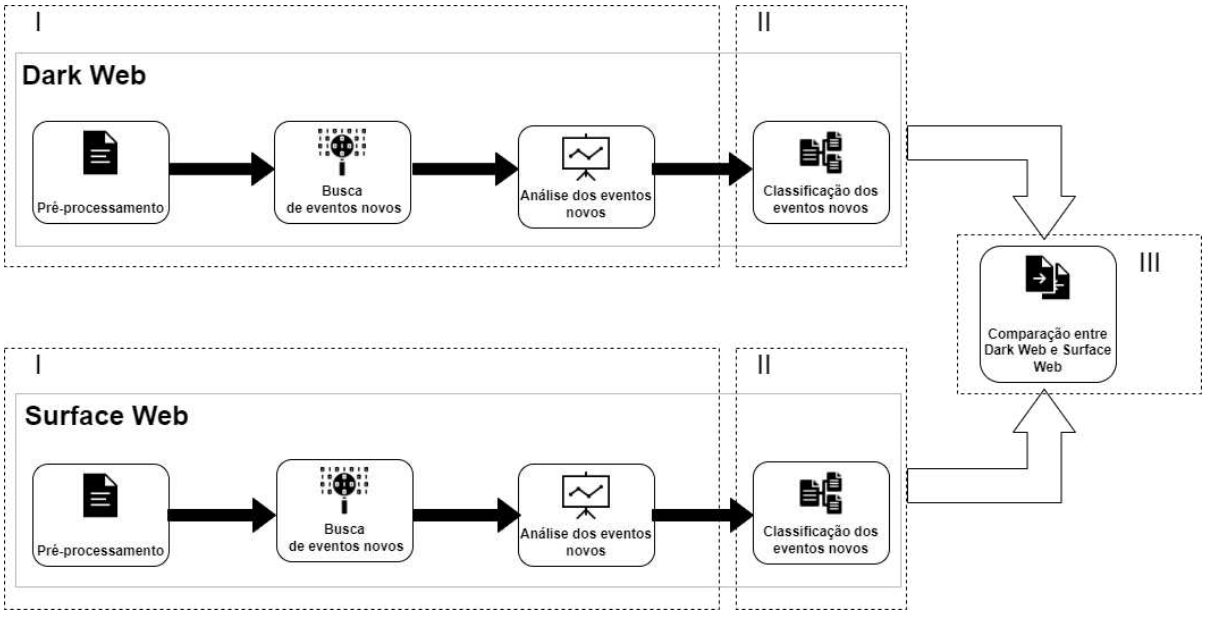
[1]Indica se os dados utilizados na pesquisa foram disponibilizados pelos autores.

Capítulo 3

Desenvolvimento

Este trabalho tem como objetivo a detecção de eventos emergentes em fóruns da *Dark Web* e da *Surface Web*, bem como a realização de uma comparação entre os eventos encontrados, visando distinguir os tipos de ocorrências oriundas de cada fonte. Neste capítulo, são descritas as etapas do desenvolvimento desta pesquisa, visando cumprir o objetivo proposto. Todas as etapas estão ilustradas, de maneira geral, na Figura 5 e apresentadas nas seções seguintes. A Seção 3.1 (Etapa 1) aborda a busca de eventos; a Seção 3.2 (Etapa 2) descreve a classificação dos eventos novos; e, por fim, a Seção 3.3 (Etapa 3) apresenta a comparação entre os eventos novos encontrados nas duas bases (*Dark Web* e *Surface Web*).

Figura 5 – Visão geral das etapas do método proposto.



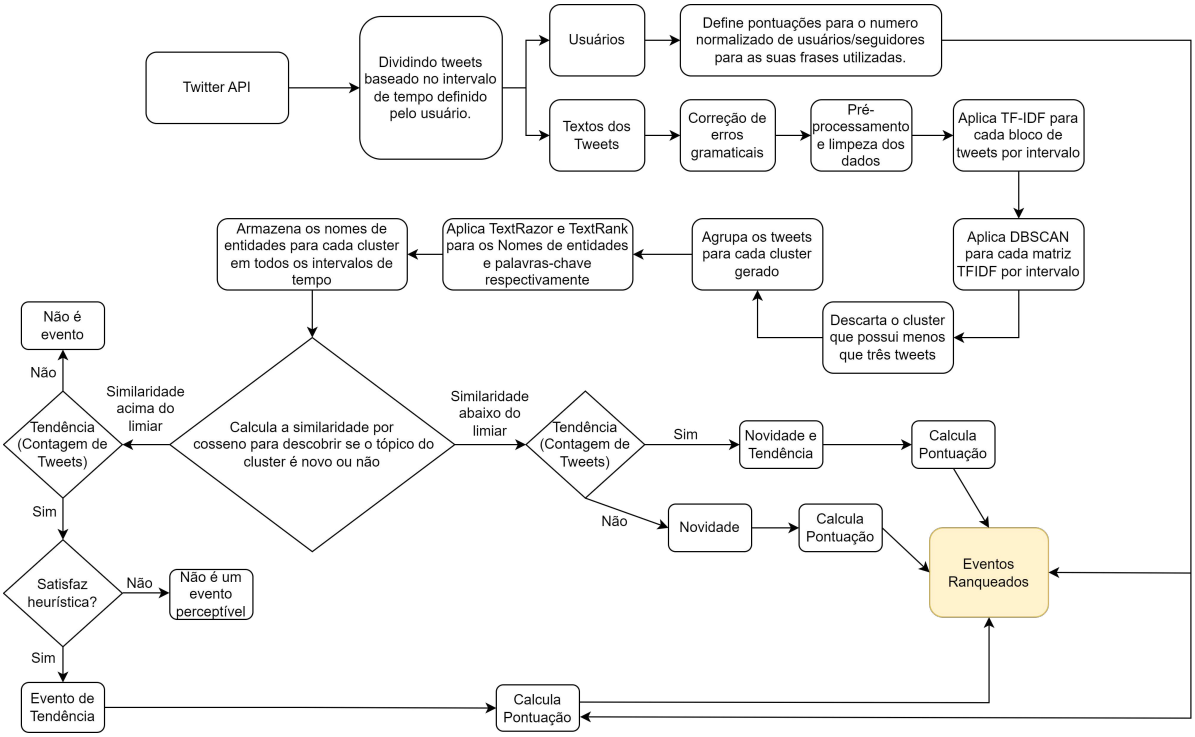
Fonte: Autoria própria.

3.1 Etapa I – Busca de eventos novos

Nesta etapa, foi desenvolvido um método baseado no trabalho de Bose et al. (2020) (Seção 2.5) para buscar eventos novos em publicações da *Surface Web* e da *Dark Web*. No trabalho original, os autores criaram uma forma de identificar automaticamente eventos novos sobre ameaça cibernética em publicações da plataforma X (antigo Twitter). Identificar eventos novos, nesse contexto, significa reconhecer eventos que não apresentem alta similaridade com outros temas já encontrados em períodos anteriores.

Na Figura 6 é possível ver todo o processo realizado por Bose et al. (2020). Os autores coletaram 21.368 *tweets* entre os dias 06 e 09 de setembro de 2018. Após a coleta de dados, foi feita uma separação dos *tweets* de acordo com o intervalo de tempo definido pelo usuário. Após essa separação, é feito o pré-processamento dos dados, seguido da aplicação de TF-IDF (MEADOW et al., 2007) para a vetorização dos dados. Finalmente, realiza-se um agrupamento dos dados utilizando o algoritmo Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído (Density-based spatial clustering of applications with noise) (DBSCAN), proposto por Ester et al. (1996). Após o agrupamento realizado, foram extraídas entidades via Reconhecimento de Entidade Nomeada (Named Entity Recognition) (NER) e palavras-chave do texto utilizando técnicas de PLN.

Figura 6 – Fluxograma do trabalho utilizado como base desta pesquisa.



Fonte: Adaptado de Bose et al. (2020).

Estas entidades e palavras-chave foram utilizadas para comparação entre os *tweets* de um intervalo de tempo em relação ao intervalo de tempo anterior para determinar se um

evento é novo ou não. Após encontrar os eventos novos, eles foram ranqueados baseando-se na pontuação das palavras-chave encontradas e no número de *tweets* do evento.

No presente trabalho, foi desenvolvida uma versão simplificada do método de Bose et al. (2020), uma vez que o objetivo é apenas detectar eventos novos e depois utilizá-los para comparação entre as bases de dados. As principais diferenças entre o presente trabalho e o trabalho de Bose et al. (2020) residem na busca e no processamento dos eventos. Primeiramente, utilizou-se o algoritmo *SBERT* para a vetorização dos textos ao invés do uso do TF-IDF utilizado no trabalho base. Além disso, as fontes utilizadas neste trabalho são postagens de fóruns *hackers* da *Dark Web* e *Surface Web* e não somente postagens do *X/Twitter* conforme utilizado em Bose et al. (2020). Por fim, foi utilizado o algoritmo *k-means* para agrupamento dos dados ao invés do DBSCAN utilizado originalmente.

Para a busca dos eventos , foram utilizados três fóruns da *Dark Web*: *Hidden Answers*, *Deep Answer* e *Raddle*; e quatro fóruns da *Surface Web*: *Hackonology*, *Legitcarders*, *Niflheim* e *Nulledbb*, todos em língua inglesa. Optou-se por utilizar esses fóruns, pois os seus dados já haviam sido disponibilizados por Jesus Filho (2024). Nesta etapa do método proposto, também são aplicadas técnicas padrão de pré-processamento para problemas de PLN, como limpeza dos dados e remoção de *stopwords*. Para a remoção de *stopwords*, foi utilizada a biblioteca *nltk* (BIRD; LOPER; KLEIN, 2009) da linguagem de programação *Python*.

Após a etapa de pré-processamento, foi aplicado o algoritmo para busca de eventos novos. Esse algoritmo auxilia na identificação dos tipos de eventos que surgem nas publicações ao longo do tempo. As etapas de busca de eventos novos são mostradas a seguir e todas elas foram aplicadas em cada uma das fontes de dados disponíveis:

1. **Agrupamento semanal dos dados:**
   Uma vez que os eventos serão detectados ao longo do tempo, os dados foram organizados em ordem temporal crescente. Posteriormente, realizou-se um agrupamento temporal dos dados em intervalos semanais a fim de tornar possível a comparação dos temas da semana corrente com a semana anterior permitindo, assim, identificar a emergência de novos eventos.

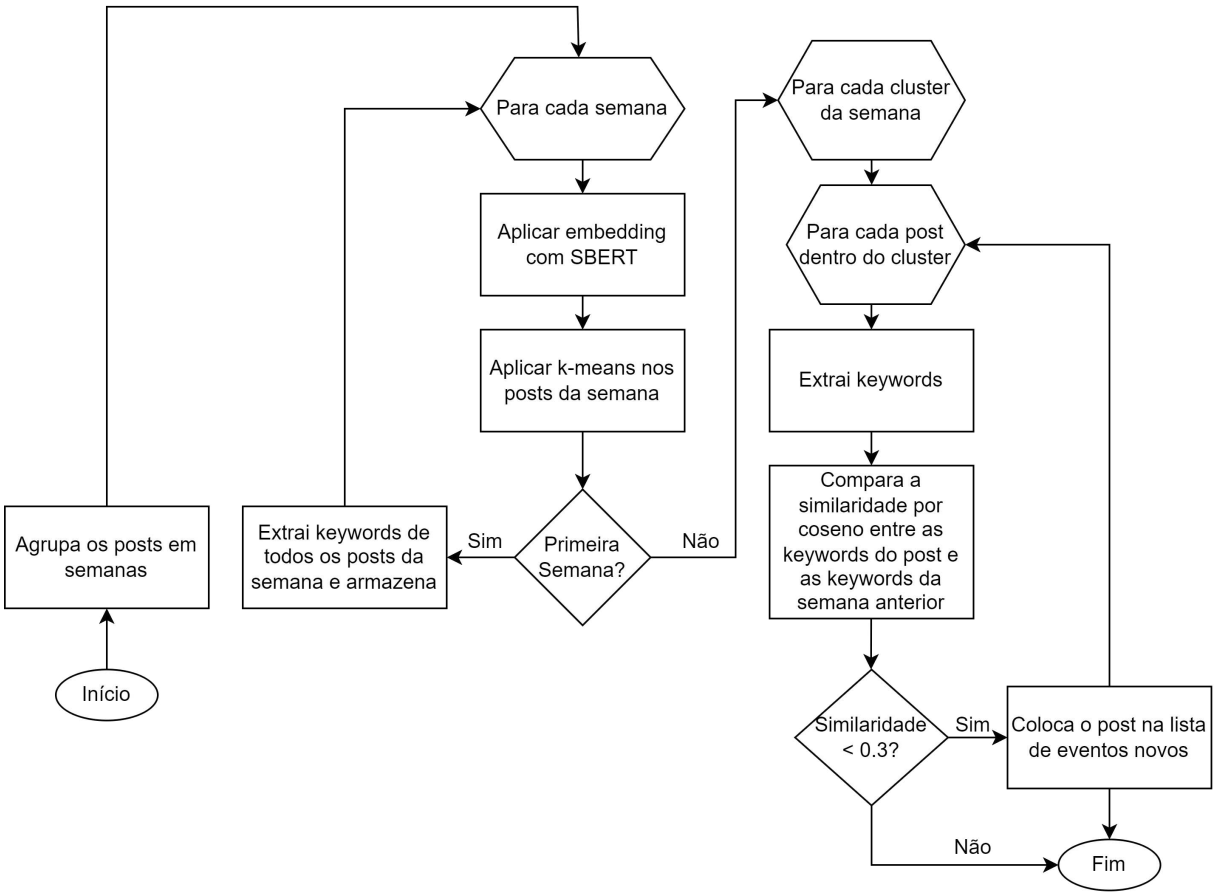
2. **Agrupamento *k-means*:**
   Para cada semana definida no tópico anterior, foi aplicado o algoritmo de agrupamento *k-means*. Esse algoritmo foi escolhido devido à sua baixa complexidade computacional e eficiência no processamento de grandes volumes de dados e também por ser amplamente utilizado no contexto de PLN. Como algumas semanas podem possuir menos de três publicações, nestes casos, o valor de k corresponderá ao número exato de publicações da semana. Para viabilizar o agrupamento, é preciso transformar os dados textuais em representações numéricas (vetores). Foi utilizado o SBERT para essa vetorização.

3. **Extração e comparação de palavras-chave:**
   Para realizar a comparação entre as publicações e identificar os eventos novos, foram utilizadas as palavras-chave extraídas do conteúdo de cada registro. Para a extração de palavras-chave, foi utilizada a biblioteca *Summa* do *Python*. Essa biblioteca modela o texto como um grafo direcionado, onde os nós representam as palavras e as arestas representam a coocorrência dentro de uma janela deslizante. A relevância de cada termo é calculada de forma iterativa, inspirada no algoritmo *PageRank* (PAGE et al., 1998), priorizando palavras que possuem maior conectividade com outros termos semanticamente importantes no grafo. Visto que a primeira semana não possui um período anterior para comparação, as palavras-chave extraídas de todos os agrupamentos foram armazenadas. Estas foram, então, empregadas na comparação com as semanas subsequentes.

Após essas etapas, a comparação entre as palavras-chave foi realizada através da SC, utilizando um limiar definido em 0,3. Para ser possível a comparação, todos os textos foram vetorizados através do algoritmo *SBERT*. Se o valor da SC entre as palavras-chave de uma publicação da semana corrente e as da semana anterior for inferior a esse limiar, a publicação é considerada como contendo um tema emergente e, portanto, é classificada como um novo evento. Ao final da análise de todas as semanas, será gerada uma lista consolidada com todas as publicações identificadas como eventos novos. Toda a etapa de verificação de eventos novos pode ser visualizada na Figura 7, como um fluxograma, e também em formato de pseudocódigo no Algoritmo 1. Para entender melhor e caracterizar os eventos novos encontrados, foram feitos alguns cálculos de similaridade mensalmente entre eles utilizando os algoritmos a seguir. Estes algoritmos foram escolhidos, pois mostram as diferenças léxicas e semânticas dos textos e são computacionalmente eficientes.

- ***Jaccard*:** Mede a sobreposição léxica absoluta entre os conjuntos de termos.
- ***Dice*:** Pondera a similaridade dando maior peso aos termos em comum.
- ***Overlap*:** Verifica se um evento é subconjunto ou parte de um evento anterior.
- **Cosseno (vetorização utilizando *SBERT*):** Avalia a similaridade semântica baseada no contexto global das sentenças.
- **Cosseno (vetorização utilizando *FastText*):** Garante a comparação de termos mesmo com variações morfológicas ou erros de digitação.

Figura 7 – Etapas do processo de busca por eventos novos.



Fonte: Autoria própria.

Algoritmo 1 Algoritmo de detecção de eventos

Entrada: Dados ordenados por data e agrupados por semana
Saída: $k = 3$

```
palavrasChaveSemana ← []
palavrasChavePublicacao ← []
limiar ← 0.3
Para semana = 1 até N faça
    corpus ← publicações da semana
    X ← aplicaVetorizacao(corpus)
    agrupamentos ← aplicaKMeans(X, k)
    Se primeiraSemana então
        palavrasChaveSemana ← definePalavrasChave(agrupamentos)
        continue
    fim se
    Para grupo = 1 até tamanho(agrupamentos) faça
        Para publicacao = 1 até tamanho(grupo) faça
            palavrasChavePublicacao ← definePalavrasChave(publicacao)
            Se similaridadeCosseno(palavrasChavePublicacao, palavrasChaveSemana) <
            limiar então
                evento novo detectado
            fim se
        fim para
    fim para
    guardaPalavrasChaveSemana(palavrasChaveSemana) {guarda as palavras chave da
    semana corrente}
fim para
```

3.2 Etapa II – Classificação dos eventos novos

Nesta etapa da pesquisa, é realizada a classificação dos eventos novos identificados na etapa anterior (Seção 3.1). Essa classificação baseia-se no conteúdo da publicação, separando-o entre material malicioso e não malicioso. Para a classificação das publicações, foi empregado o algoritmo proposto por Jesus Filho (2024) discutido a seguir.

Jesus Filho (2024) propôs um método para classificar publicações como maliciosas ou não maliciosas. No referido trabalho, uma publicação é classificada como maliciosa se contiver Indicadores de Comprometimento (Indicators of Compromise) (IoCs) e palavras-chave relacionadas à cibersegurança, tais como: CPF, *hacking*, vírus, DDoS, etc. O algoritmo foi treinado com publicações coletadas de dois fóruns abertos da *Dark Web*: *Hidden Answers* e *Deep Answers*. O modelo de classificação utilizou como algoritmo de treinamento o *LightGBM* (*Light Gradient-Boosting Machine*) mostrou alta eficiência, apresentando uma precisão de 97%. A classificação final empregou a probabilidade fornecida pela saída do algoritmo, sendo que as publicações com probabilidade igual ou superior a 50% foram classificadas como relevantes (maliciosas), e aquelas com probabilidade inferior a 50% foram classificadas como irrelevantes (não maliciosas).

Também foi desenvolvida uma versão onde as publicações foram classificadas em faixas de relevância, sendo elas: baixa, média ou alta. Para essa categorização, foram utilizadas as seguintes faixas percentuais: inferior a 30%, entre 30% e 70% e superior a 70%, respectivamente. Neste trabalho, especificamente, foi adotada essa abordagem por faixas para a classificação. Tal classificação é importante pois possibilita determinar a relevância dos eventos emergentes sob a perspectiva da cibersegurança.

3.3 Etapa III – Comparação entre dados da *Surface Web* e da *Dark Web*

Nesta etapa, são realizadas diversas comparações estatísticas, utilizando métodos de categorização de textos, como LDA e agrupamento hierárquico entre os eventos encontrados na Etapa II (Seção 3.1) para entender a relação entre os tipos de eventos encontrados nas duas fontes de dados: *Surface Web* e *Dark Web*. A aplicação do LDA foi realizada utilizando a biblioteca *gensim* (ŘEHŮŘEK; SOJKA, 2010) e considerando apenas os eventos classificados com relevância média e alta. Para a validação do LDA, são utilizados *perplexity* e Índice de Coerência que são medidas comuns para se avaliar a eficácia de um algoritmo de modelagem de tópicos.

Também são realizadas comparações estatísticas utilizando SC entre os eventos encontrados nas duas bases de dados. Essa análise foi conduzida também de forma separada para cada tipo de classificação dos eventos, comparando-se, por exemplo, os eventos de baixa relevância da *Surface Web* com os da *Dark Web*. Além disso, foi feita uma compa-

ração mensal utilizando SC entre os eventos das duas bases para descobrir como a média e desvio padrão do resultado do SC varia com o tempo. Essa comparação também foi feita separadamente para cada tipo de classificação dos eventos.

Capítulo 4

Experimentos e Análise dos Resultados

Neste capítulo, são apresentados os resultados obtidos a partir da execução das etapas descritas no Capítulo 3. A Seção 4.1 apresenta os conjuntos de dados utilizados. A Seção 4.2 descreve o processo de busca de eventos emergentes, enquanto a Seção 4.5 expõe as análises realizadas sobre essas ocorrências. A Seção 4.4 aborda a classificação feita nos novos eventos e, por fim, a Seção 4.5 detalha as comparações entre os eventos encontrados na *Surface Web* e na *Dark Web*.

4.1 Conjuntos de Dados

Neste trabalho, foram utilizadas postagens de três fóruns da *Dark Web* (*Hidden Answers, Deep Answer e Raddle*) e quatro da *Surface Web* (*Hackonology, Legitcarders, Niflheim e Nulledbb*). Todas as postagens utilizadas estão em língua inglesa (JESUS FILHO, 2024).

A base de dados utilizada tem um total de 35.693 publicações da *Surface Web* entre as datas 15 de janeiro de 2015 a 08 de novembro de 2023; e 2.579 da *Dark Web* entre as datas 07 de maio de 2017 a 18 de fevereiro de 2024. A distribuição de quantidade de publicações por ano em cada uma da base de dados pode ser vista na Tabela 3.

Ambas as bases de dados estão no formato de Valores separados por vírgula (Comma-separated values) (CSV) e possuem os seguintes campos:

- "id": Campo de identificação única;
- "title": Título da publicação;
- "category": Categoria da publicação, sendo que cada fórum define as suas próprias categorias;
- "content": Conteúdo da publicação;
- "created_at": Data e hora da publicação.

Tabela 3 – Quantidade de publicações por ano na base de dados utilizada.

Surface Web		***Dark Web***	
Ano	**Quantidade de publicações**	**Ano**	**Quantidade de publicações**
2015	1.049	2017	123
2016	538	2018	135
2017	1.499	2019	76
2018	2.305	2020	116
2019	1.963	2021	159
2020	4.252	2022	752
2021	9.732	2023	1.145
2022	9.257	2024	73
2023	5.098		
Total	**35.693**	**Total**	**2.579**

Fonte: Autoria própria.

Para otimizar a análise, o campo "content" inclui o conteúdo principal da publicação, juntamente com todas as respostas e comentários subsequentes. Um exemplo de publicação pode ser visto na Tabela 4.

Tabela 4 – Exemplo de publicação

id	title	category	content	created_at
5578	Just thinking....	Other	Does the hidden answers admins still on? I couldnt able to see them constantly replying as before. More or less they vanished suddenly ;-; health check up fr. They might be no more or underground due to some reason. Who knows except them but yes, their activity has been significantly reduced recently.	2023-12-14 04:26:40

Fonte: Autoria própria.

4.2 Busca de eventos novos

Conforme apresentado na Seção 3.1, a primeira etapa deste trabalho é a busca de eventos novos nas publicações. Para identificar os eventos, primeiramente foi preciso realizar o pré-processamento dos dados. Na etapa de pré-processamento, foram removidos os seguintes elementos: URLs, *tags* HTML, hiperlinks do tipo *onion*, nomes de usuários,

caracteres especiais e *stopwords*. Adicionalmente, foi necessário remover dos textos a frase "*To view the content, you need to Sign In or Register*". Esta remoção justifica-se pelo fato de o trecho estar presente em diversas publicações coletadas e não apresentar relevância para a análise do presente trabalho.

Após o pré-processamento, foram executadas as seguintes tarefas: *i*) organização dos dados em semanas; *ii*) agrupamento dos dados; e *iii*) extração e comparação de palavras-chave. Estas etapas foram aplicadas a cada fonte de dados e são detalhadas a seguir. Todos os experimentos foram feitos utilizando um notebook que possui processador i7 de 2.20GHZ, placa de vídeo GTX 1050 Ti e 32 GB de memória.

4.2.1 Organização e agrupamento dos dados

Visto que as bases originais apresentavam dados começando em períodos diferentes, (**2015** para a *Surface Web* e **2017** para a *Dark web*), optou-se por usar somente dados a partir de 01 de janeiro de 2020, pois assim é possível garantir a comparabilidade estatística entre os dados utilizados. Com esse ajuste, as novas quantidades de dados foram 28.339 e 2.245 para *Surface Web* e *Dark Web* respectivamente. Esses dados representam um total de 197 semanas analisadas na *Dark Web* e 202 na *Surface Web*.

Visto que os eventos serão detectados ao longo do tempo, os dados são primeiramente ordenados em ordem crescente com base no campo "created_at" (Seção 4.1). Em seguida, os dados são agrupados semanalmente, permitindo a comparação das palavras-chave de uma semana com a semana anterior, visando identificar novos eventos emergentes. A manipulação dos dados foi realizada utilizando a biblioteca *Pandas* do *Python*, sendo que o método *Grouper* da própria biblioteca foi empregado para o agrupamento semanal. O agrupamento foi executado semanalmente, visto que este intervalo demonstrou resultados superiores aos obtidos no teste inicial com periodicidade mensal.

Para cada semana, o algoritmo de agrupamento *k-means* é aplicado às publicações, utilizando inicialmente o valor $k = 3$. Porém, como podem existir semanas com menos de três publicações, o número k será igual ao número de publicações da semana, nesses casos. O valor $k = 3$ foi utilizado por apresentar a melhor interpretabilidade semântica nos testes iniciais e também para manter o rigor de agrupamento mínimo proposto por Bose et al. (2020). Para viabilizar essa etapa, as publicações foram transformadas em vetores numéricos. O método *SBERT* (REIMERS; GUREVYCH, 2019) foi escolhido para esta vetorização devido aos seus resultados superiores e por sua capacidade de gerar vetores que levam em conta o contexto semântico dos dados. Como entrada, o modelo *SBERT* recebe sequências de textos que são processadas por uma arquitetura baseada em *Transformers*. A busca de eventos também foi realizada utilizando a vetorização através do TF-IDF para fins de comparação.

O agrupamento dos dados foi realizado de modo que as publicações com conteúdo similar fossem alocadas no mesmo grupo, facilitando a análise. Adicionalmente, foram

realizados experimentos de agrupamento utilizando o algoritmo *Agrupamento Espacial Hierárquico Baseado em Densidade de Aplicações com Ruído (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (HDBSCAN)*, um método de agrupamento baseado em densidade (CAMPELLO; MOULAVI; SANDER, 2013). Contudo, esse método apresentou alto tempo de execução devido ao grande volume de dados utilizado neste trabalho, tornando-se impraticável.

4.2.2 Extração de palavras-chave e comparação

A partir de todos os grupos gerados com as publicações da semana corrente, foi realizada a extração das palavras-chave. Para isso, utilizou-se a biblioteca *Summa* (BARRIOS et al., 2016), que extrai palavras-chave por meio de um processo que envolve a vetorização dos dados, seguida pela construção de um grafo (que mostra as conexões entre as palavras) e a subsequente classificação dessas palavras utilizando o algoritmo *TextRank*. As palavras com as pontuações mais altas são, então, retornadas como as palavras-chave mais relevantes. Para comparar os textos e identificar novos eventos, foram utilizadas as palavras-chave extraídas do conteúdo de cada publicação.

Como a primeira semana não possui uma semana anterior para comparação, as palavras-chave extraídas de todos os grupos dessa semana são armazenadas para uso posterior. Posteriormente, para cada semana subsequente, o agrupamento dos dados é realizado e, para cada grupo identificado na semana corrente, as palavras-chave são extraídas de cada publicação. Estas são então comparadas com a lista de palavras-chave da semana anterior.

Para essa comparação, é utilizada a SC, com um limiar de 0,3. Neste trabalho, foi adotado o valor de 0,3 por questões de balanceamento: valores abaixo deste limiar fazem com que publicações com pequenas variações de escrita ou com poucas palavras distintas sejam consideradas novos eventos. Por outro lado, valores acima de 0,3 resultavam na classificação de muitas publicações como eventos distintos, mesmo quando elas abordavam o mesmo assunto.

Dessa forma, o valor de 0,3 foi adotado por representar um equilíbrio entre sensibilidade e especificidade na detecção de eventos. Se a SC entre as palavras-chave de uma publicação da semana corrente e as da semana anterior estiver abaixo do limiar, a publicação é considerada como abordando um novo tópico e, portanto, é classificada como um novo evento. Ao final da análise de todas as semanas, é gerada uma lista contendo todas as publicações identificadas como novos eventos.

Após a aplicação de todos esses passos, foram detectadas 1.835 publicações consideradas como novos eventos nos dados da *Dark Web* e 24.236 publicações nos dados da *Surface Web*. Esses resultados foram obtidos utilizando o algoritmo *SBERT* para a vetorização. Em contraste, para a vetorização via TF-IDF, as quantidades encontradas foram de 129 eventos na *Dark Web* e 2.346 na *Surface Web*. A quantidade de eventos identificados, o total de publicações analisadas, o método de vetorização empregado e o tempo médio

de processamento estão sumarizados na Tabela 5. Cada evento encontrado é equivalente a uma publicação da base de dados. Alguns exemplos com a data, categoria e texto da publicação de eventos encontrados na *Surface Web* e na *Dark Web* podem ser visualizados nas Tabelas 6 e 7, respectivamente.

Tabela 5 – Quantidade de eventos novos encontrados e taxa de detecção.

	Dark Web	**Surface Web**
Eventos encontrados (*SBERT*)	**1.835**	**24.236**
Eventos encontrados (TF-IDF)	129	2.346
Publicações analisadas	2.245	28.339
Taxa de detecção (*SBERT*)	**81,73%**	**85,52%**
Tempo médio de processamento (*SBERT*)	4 minutos	83 minutos

Fonte: Autoria própria.

Tabela 6 – Exemplo de eventos encontrados na *Surface Web*

Data	**Categoria**	**Texto**
06/01/2020	Dumps	Thanks again for all the hard work, keep it up dude That'a a badass share man Really gotta check this out as soon as possible.. Thanks in advance . DOWNLOAD
12/01/2020	Leak Miscellaneous	Does anyone wish to share this insane set? https://www.unrealengine.com/marketplace...c-district it would be super awesome of you, thank you <3
10/09/2021	Accounts	i ty thanks bro thanc you do
25/01/2022	Ways to make money	Please note, if you want to make a deal with this user, that it is blocked. Hello, I've mostly been lurking around here trying different methods and then finally i figured out a magic bullet! Today I want to share with everyone how to make quick and easy money in 2022. Without investing. #1 website & App ||Without investment || Make daily $50 Daily Easy task - No skill needed Signup bonus 5$ + $3 app update bonus Join here. Minimum withdraw 50$ Proof attached 100% legit - NO SCAM.
26/03/2023	Leaks - Cracked Programs & Tools	VirusTotal: https://www.virustotal.com/gui/****** Link download: http://bit.ly/****** Unzip Password is 1

Fonte: Autoria própria.

Tabela 7 – Exemplo de eventos encontrados na *Dark Web*

Data	**Categoria**	**Texto**
28/01/2020	N/A	Does anyone here use Silent Circle? Can they be trusted for the most part or is it pretty much a waste of money?
29/08/2021	Technology	Find ISO file then boot to flash etc... As the comment of above, you need to use the commonly known Rufus and use the flash. if you have Windows 10 and THE CORRECT REQUIREMENTS, you can upgrade it by just downloading an the update.
09/09/2021	Technology	If you know any alternative/or any cool forum abt cyber security etc answer this question, all links are welcomeanswer this question, all links are welcome. Thanks. Only spanish foro Hackplayers idk http://foro.hackplayers.com/
29/01/2022	Other	what do you think about doing a redit post about this forum in the english version? Fucking Good! Great idea, I fully support it. That works, but this process will be naturaly and slowly. What would that be?? Sounds good. We just need to keep posting the link wherever we can, sooner or later ppl will find H.A. We already gave the first steps! Reddit? Mate it is a forum and not a reddit
28/01/2023	Legit or Shit?	on the DW? you cannot. If you can buy on surface is better. There's a lot of scam here and because of the dark web is focused on privacy you never know if site is secure. Because if some site says that site is secure, that can be scam too. Absolutely agree with ***** only possible way could be either search youtube or surface web for some authentic dark web sites or invest small amount before making some big move.

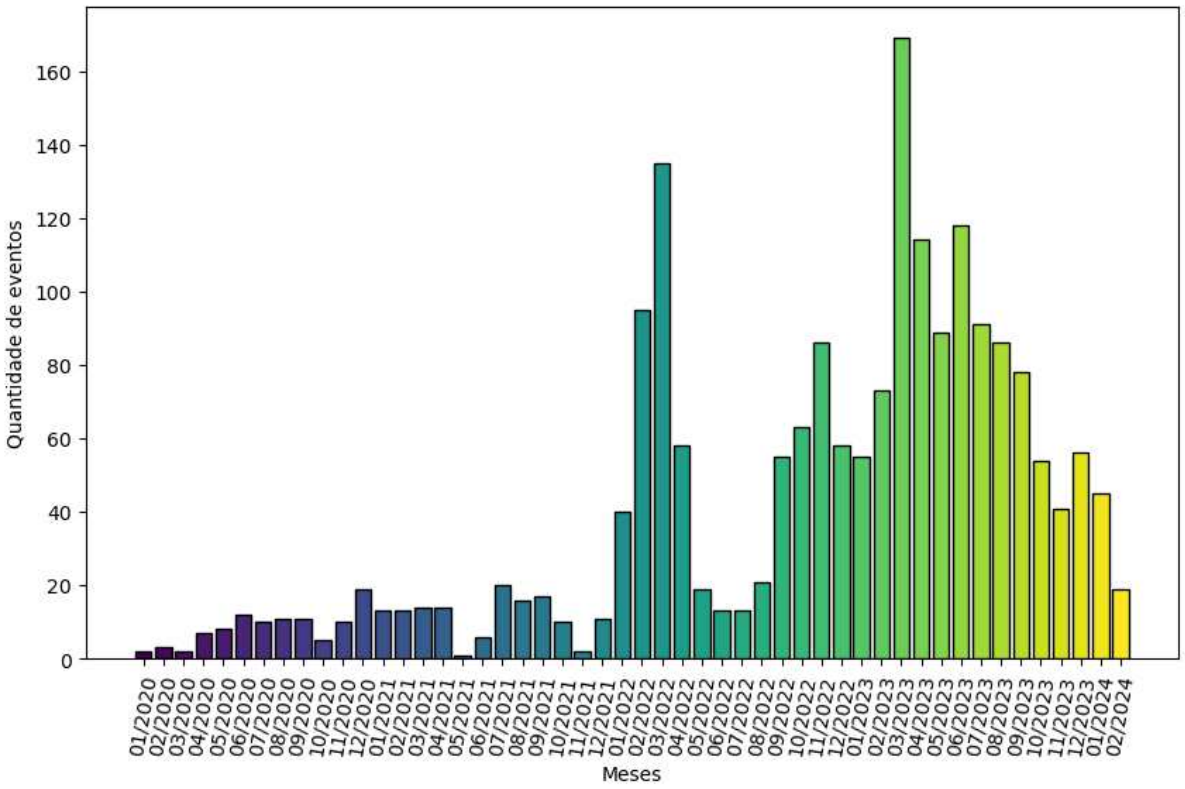
Fonte: Autoria própria.

4.3 Análise dos Resultados

Após a identificação dos eventos, foram gerados gráficos para a análise exploratória dos resultados. O objetivo desta etapa foi detalhar o perfil dos eventos encontrados, facilitando a visualização de sua distribuição e relevância em ambos os ambientes.

As Figuras 8 e 9 exibem a distribuição de frequência dos eventos novos detectados mensalmente em ambas as bases de dados. Apesar do algoritmo de busca de eventos organizar semanalmente os dados, optou-se por gerar alguns gráficos agrupados mensalmente para oferecer uma melhor legibilidade visual. Conforme pode ser observado na Figura 9, a *Surface Web* apresenta um crescimento gradual na quantidade de eventos seguido de uma diminuição gradual, esse comportamento demonstra uma menor volatilidade em relação aos dados da *Dark Web* na Figura 8 onde os dados apresentam um crescimento abrupto na quantidade de eventos precedida por períodos de baixa atividade.

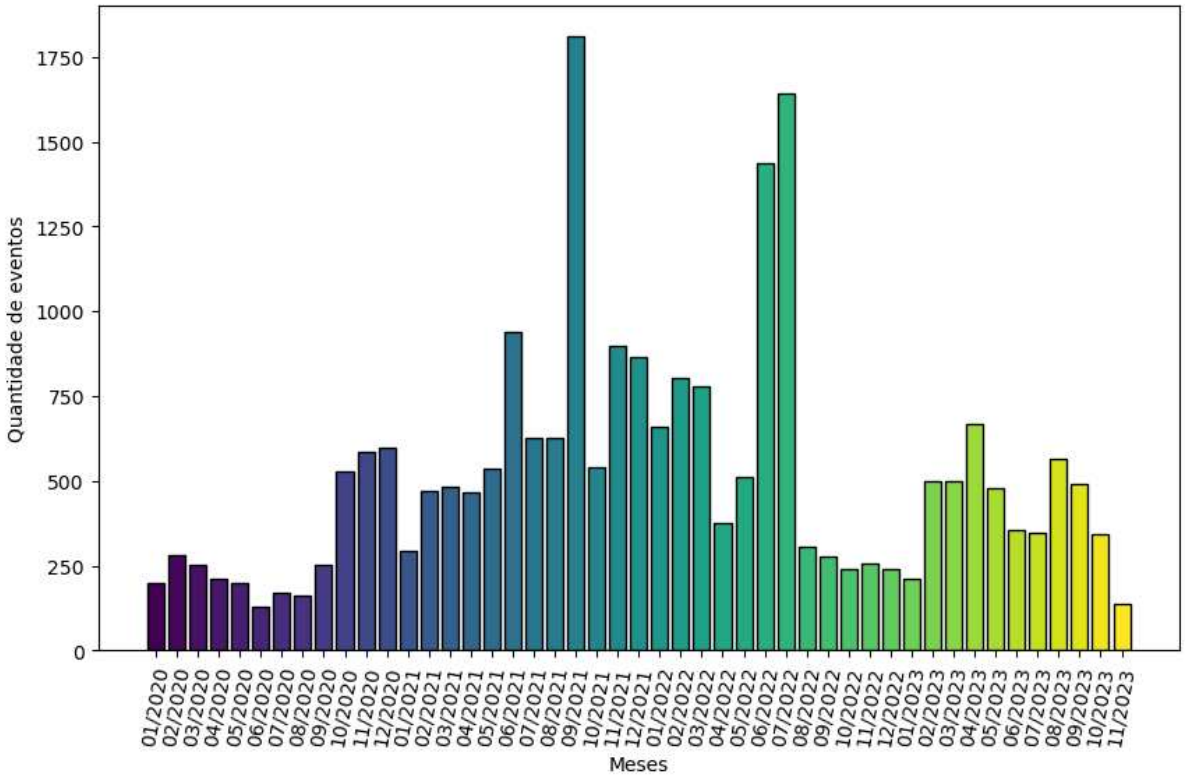
Figura 8 – Distribuição dos eventos encontrados mensalmente na *Dark Web*.



Fonte: Autoria própria.

A Figura 10 apresenta as principais categorias de eventos emergentes na *Dark Web*. Essas categorias são derivadas diretamente dos metadados dos eventos identificados. Observa-se uma concentração em categorias que indicam o propósito final da atividade maliciosa, como "*hacking*", "*money*", e "*darknet and tor*". As categorias "*other knowledge and information*" e "*links*" sugerem a troca de recursos e informações que servem como insumos para atividades ilícitas. Finalmente, a presença de "*politics*", "*health*" e "*science*

Figura 9 – Distribuição das quantidades de eventos encontrados mensalmente na *Surface Web*.



Fonte: Autoria própria.

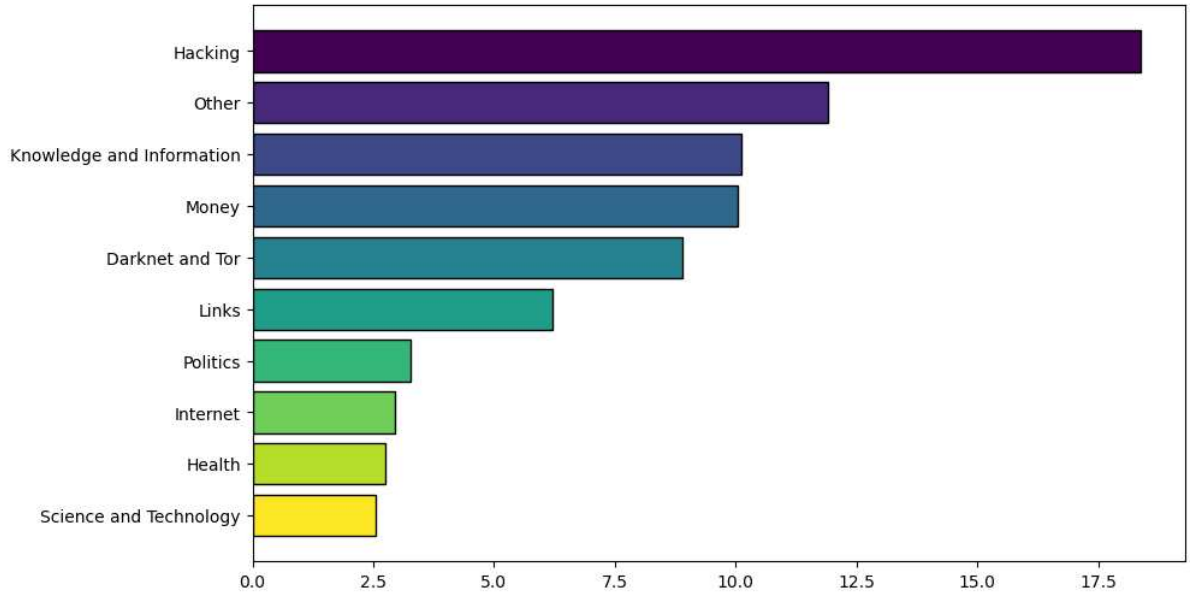
and technology" reforça que o escopo de interesse no ambiente é amplo, mas ainda assim voltado para discussões que possam ser monetizadas ou utilizadas para fins de evasão e cibersegurança.

Em contraste, a Figura 11 detalha a distribuição das categorias na *Surface Web*. O foco neste ambiente é notavelmente diferente, concentrando-se em recursos, métodos e ferramentas de execução. As categorias mais proeminentes são "*dumps*", "*leak miscellaneous*" e "*accounts*". A presença de categorias como "*openbullet*" e "*silverbullet*" (que são nomes de ferramentas de verificação de credenciais) e "*tutorials, guides etc*" indica que o conteúdo emergente na *Surface Web* está mais voltado para o compartilhamento de materiais, vazamentos de dados brutos e instruções sobre como conduzir a fraude, em vez da discussão teórica vista na *Dark Web*.

Para avaliar a novidade dos eventos encontrados em ambas as bases, foi realizada uma análise de similaridade comparando o conjunto de eventos novos em um determinado mês N com o conjunto de eventos novos do mês anterior $(N-1)$. Essa comparação foi executada utilizando cinco métricas distintas, conforme detalhado na Tabela 8. Os resultados da aplicação de cada métrica podem ser vistos nas Figuras 12 e 13 e o pseudocódigo da aplicação das métricas pode ser visto no Algoritmo 2.

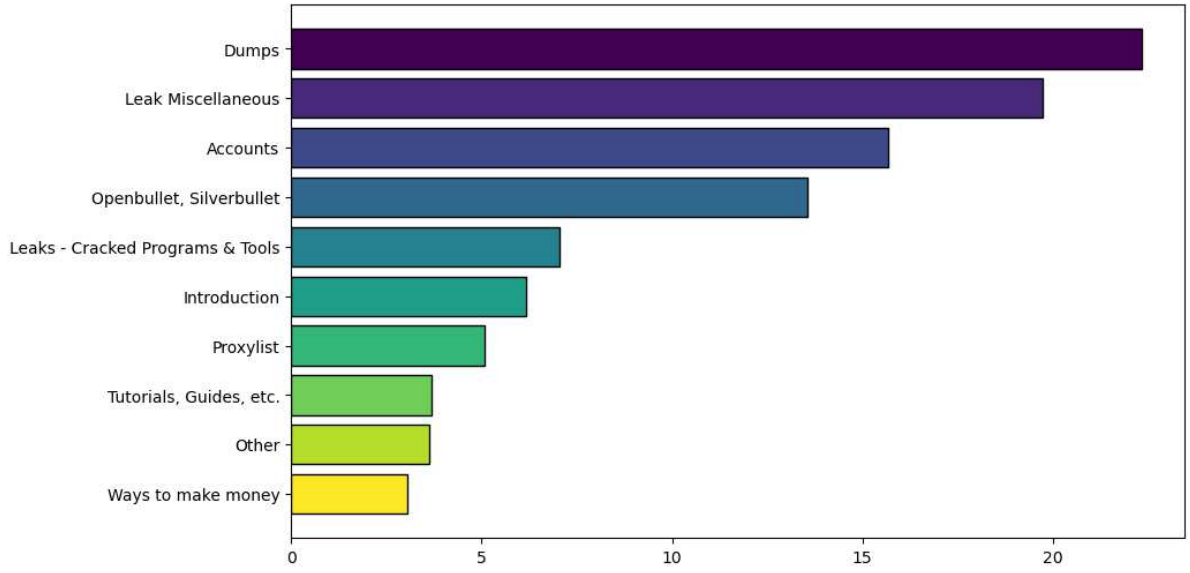
A Figura 12 mostra que o gráfico da similaridade por Cosseno (*FastText*) apresenta

Figura 10 – As dez categorias com maior porcentagem de ocorrência nos eventos novos da *Dark Web*.



Fonte: Autoria própria.

Figura 11 – As dez categorias com maior porcentagem de ocorrência nos eventos novos da *Surface Web*.



Fonte: Autoria própria.

Algoritmo 2 Algoritmo para cálculo de métricas nos eventos encontrados

Entrada: Conjunto de dados estruturado ($df\_eventos$), Modelo $SBERT$ ($modelSBert$)
Saída: Lista de eventos e métricas de similaridade ($resultado\_semelhanca$)

$dfs\_mes \leftarrow$ Agrupar $df\_eventos$ por periodicidade mensal
$posts\_referencia \leftarrow \emptyset$
$resultado\_semelhanca \leftarrow \emptyset$
for all df_atual **in** dfs_mes **faça**
  $palavras \leftarrow$ Extrair tokens de texto de df_atual
  Se índice do grupo $== 0$ **então**
    $posts\_referencia \leftarrow palavras$
    $resultado\_semelhanca.append(0.1, 0.1, 0.1, 0.1, 0.1, df\_atual)$
    continue
  fim se
  Se $palavras$ está vazio **então**
    $cs, js, di, ov, ft \leftarrow 0$
  senão
    $cs \leftarrow modelSBert.similarity(posts\_referencia, palavras)$
    $js \leftarrow Jaccard\_Similarity(posts\_referencia, palavras)$
    $di \leftarrow similaridade\_dice(posts\_referencia, palavras)$
    $ov \leftarrow overlap\_coefficient(posts\_referencia, palavras)$
    $ft \leftarrow cosseno\_fasttext(posts\_referencia, palavras)$
    $posts\_referencia \leftarrow palavras$
  fim se
  Se tamanho de $df\_atual > 0$ **então**
    $resultado\_semelhanca.append(cs, js, di, ov, ft, df\_atual)$
  fim se
fim para

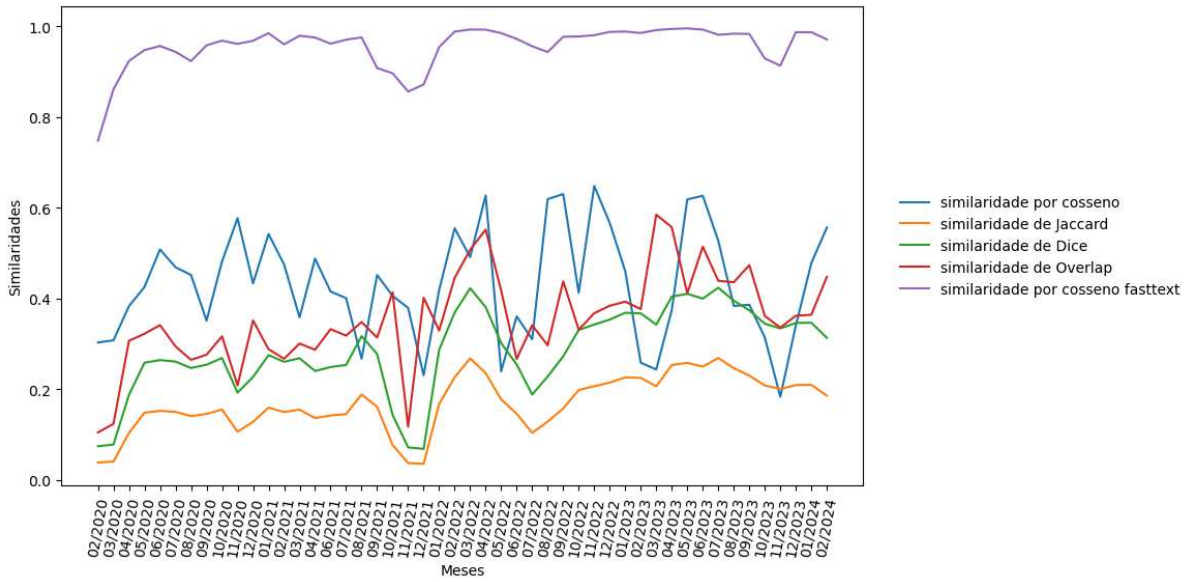
Tabela 8 – Métricas da comparação de similaridades.

Métrica	**Tipo**	**Foco da similaridade**
Jaccard	Baseada em Conjunto	Intersecção dividida pela união
Dice	Baseada em Conjunto	Intersecção normalizada dando mais peso à intersecção.
Overlap	Baseada em Conjunto	Intersecção dividida pelo menor conjunto.
Cosseno (*SBERT*)	Baseada em *Embeddings*	Similaridade semântica.
Cosseno (*FastText*)	Baseada em *Embeddings*	Similaridade semântica.

Fonte: Autoria própria.

valores consistentemente elevados e com baixa variância ao longo do período. Isso indica que os temas do conteúdo dos eventos novos da *Dark Web* não mudam drasticamente de um mês para o outro. Ou seja, mesmo que os assuntos específicos mudem, os temas centrais (fraude, venda de credenciais, discussão de *malware*, etc.) são mantidos. Um exemplo pode ser visto com os dois eventos encontrados em janeiro de 2020 e os três encontrados em fevereiro de 2020. Em janeiro, as discussões concentraram-se entre o uso de *cash* e *cryptocurrency* para a preservação da privacidade financeira. No mês subsequente, os assuntos específicos migraram para ferramentas de desenvolvimento (*React Native*) e protocolos de rede descentralizada (*Sistema de Arquivos Interplanetário (InterPlanetary File System) (IPFS)*). Contudo, o tema central, focado em soberania tecnológica e *Segurança de Operações (Operations Security) (OPSEC)*, permaneceu igual. Além disso, também é possível notar que as curvas das similaridades de *Jaccard*, *Dice* e *Overlap* estão agrupadas na parte inferior do gráfico (valores entre 0,05 e 0,5), o que mostra que a sobreposição de palavras-chave entre as publicações de um mês e o mês anterior são muito baixas. Esse resultado sugere que os usuários estão constantemente mudando os termos utilizados, mesmo que estejam falando sobre o mesmo tópico.

Figura 12 – Comparação de similaridade mensal de eventos novos da *Dark Web* sendo o eixo X representando os meses e o eixo Y o score de similaridade.



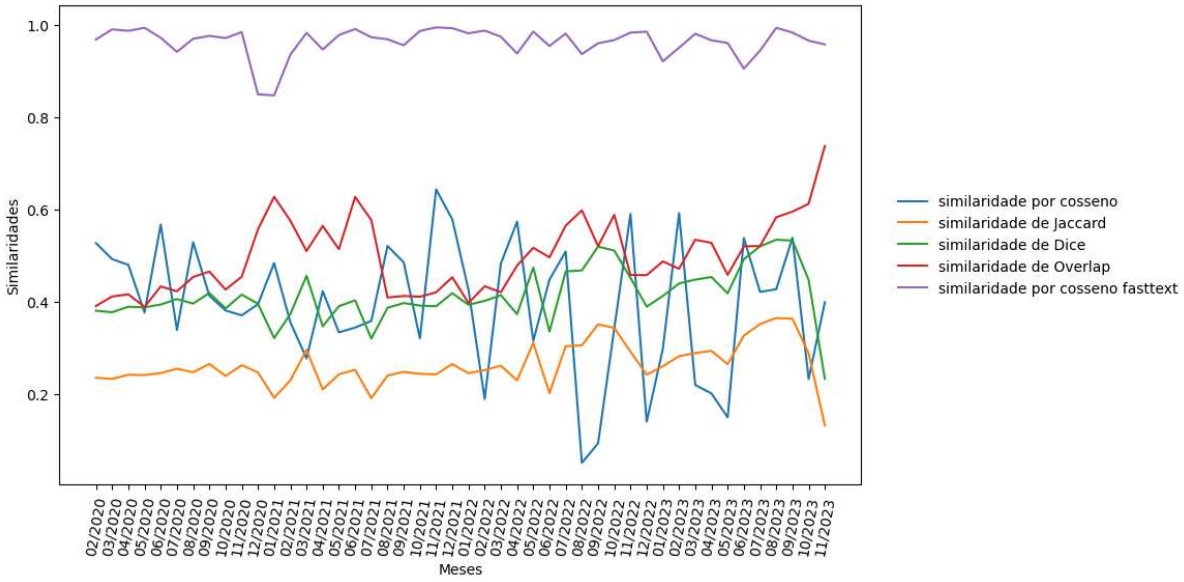
Fonte: Autoria própria.

De maneira similar, na Figura 13 é possível observar o comportamento consistente da curva da similaridade de Cosseno (*FastText*), indicando que os temas centrais de discussão continuam estáveis. A curva da similaridade por Cosseno (*SBERT*) mostra uma volatilidade maior do que a mostrada nos dados da *Surface Web*.

Dessa forma, a análise realizada nesta seção permitiu caracterizar os dados de ambos os ambientes. Conclui-se que a distribuição das categorias (como Tecnologia, *Hacking* e

vazamentos de dados) apresenta uma consistência que define o perfil de cada ambiente: enquanto a *Dark Web* mantém um foco em anonimato e execução de atividades ilícitas, a *Surface Web* caracteriza-se pela maior divulgação de vazamentos de dados, muitas vezes fruto das atividades exercidas na *Dark Web*.

Figura 13 – Comparação de similaridade mensal de eventos novos da *Surface Web* sendo o eixo X representando os meses e o eixo Y o score de similaridade.



Fonte: Autoria própria.

4.4 Classificação dos eventos encontrados

Conforme detalhado na Seção 3.2, a classificação dos eventos encontrados em ambas as bases (*Surface Web* e *Dark Web*) foi conduzida para determinar sua natureza maliciosa. Para essa tarefa, foi utilizado apenas o conjunto de eventos identificados via *SBERT*, pois os eventos identificados utilizando esse tipo de vetorização apresenta uma melhor qualidade semântica. Como vetorização de entrada para o classificador, o TF-IDF foi empregado, visto que o algoritmo de classificação criado por (JESUS FILHO, 2024) foi treinado especificamente com essa representação vetorial. A Tabela 9 mostra os resultados da classificação. Cerca de 4% dos eventos da *Surface Web* e 6% da *Dark Web* foram classificados com alta relevância; isto é, eles apresentam IoCs (como URLs, *hashes* e endereços IP) em seus conteúdos, conforme detalhado na Seção 3.2. Isso demonstra que o método proposto pode auxiliar na identificação de eventos de interesse da cibersegurança e tem potencial para melhorar sistemas como Sistemas de detecção de intrusões (Intrusion Detection Systems) (IDSs). Alguns exemplos de publicações classificadas como relevância alta na *Surface Web* e na *Dark Web* podem ser vistos nas Tabelas 10 e 11. Além disso, as

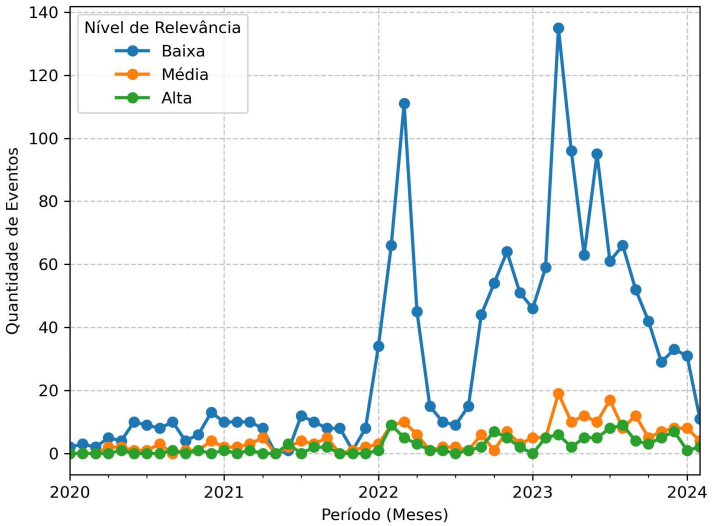
Figuras 14 e 15 mostram a distribuição das classificações ao longo do tempo, agrupadas mensalmente.

Tabela 9 – Distribuição percentual da relevância dos eventos coletados

	Surface Web		**Dark Web**	
Nível de Relevância	**Qtd.**	%	**Qtd.**	%
Alta	985	4,06	111	6,05
Média	3.552	14,66	225	12,26
Baixa	19.699	81,28	1.499	81,69
Total	**24.236**	**100,00**	**1.835**	**100,00**

Fonte: Autoria própria.

Figura 14 – Distribuição mensal das classificações dos eventos encontrados na *Dark Web*.



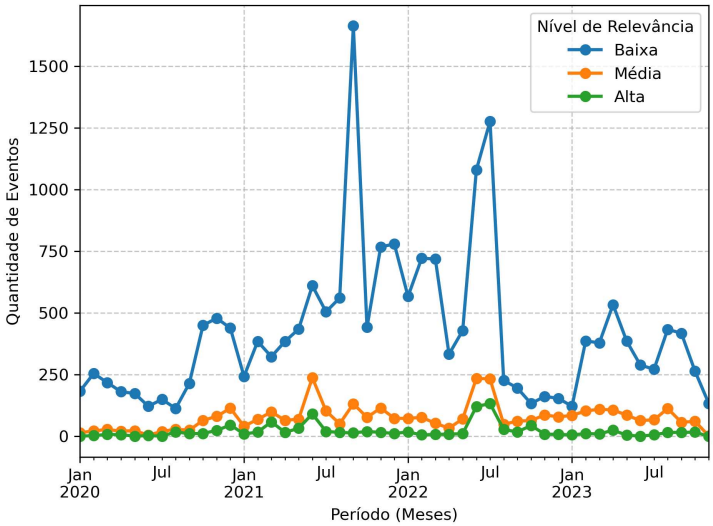
Fonte: Autoria própria.

Tabela 10 – Exemplo de eventos encontrados na *Surface Web* classificados com Relevância Alta

Data	**Categoria**	**Texto**
07/09/2023	Exploits	Feature this webshells backdoor :- Bypass 403 Bypass 404 Bypass 500 Linux Exploit Suggester Backdoor Destroyer Auto Root >Pwnkit Lock File Lock Shell Add UserName >ROOT EDITION Add UNLOCK SHELL Add Backconnect Add Hash identifier Add adminer Add Cpanel RESET Wordpress add admin user Price: 50 USD my private contact:
26/04/2020	Exploits	A very good exploit or backdoor, but its bad that the menu is not translated, its very inconvenient, really, I figured it out))) 29-04-2020, 11:33 PMRiNale Wrote: A very good exploit or backdoor, but its bad that the menu is not translated, its very inconvenient, really, I figured it out))) how did you get it to work? i can't seem to get it working This menu very useful but don't understand this language but nice. Simple Russian Exploit/Backdoor menu Pastebin The menu is not translated. The link in this hidden content has been reported as down 0 times this month. 1 times in total
29/07/2022	Tools	NSDECODER is a automated website malware detection tools. It can be used to decode and analyze weather the URL exist malware. Also, NSDECODER will analyze which vulnerability been exploit and the original source address of malware. Functionality: l Automated analyze and detect website malware. l Plenty of vulnerabilities. l Plenty of vulnerabilities. l Log export support HTML and TXT format. l Log export support HTML and TXT format. l Deeply analyze Javascript. l Deeply analyze Javascript. Download the NSDECODER : nsdecoder_gui_v1.0.zip
29/03/2020	Exploits	the thing is deleted can you update please I have been looking for this forever I remember using this a few years ago and got a new computer and have not been able to find it since Its basic backdoor menu tries to inflict backdoors looks like this :https://imgur.com/a/***** After injection write bd_menu to console to open menu pastebin
26/10/2020	Construction Web	I don't know any backdoor scanners, but a decent static code analyzer should do the trick formost backdoors. I never tried one but maybe you find something usefulhere. Dobt think there is any sites that can search for backdoors. Not sure if this is the right place but.. Is there a backdoor scanner (not for WordPress) anyone could recommend??

Fonte: Autoria própria.

Figura 15 – Distribuição mensal das classificações dos eventos encontrados na *Surface Web*.



Fonte: Autoria própria.

Tabela 11 – Exemplo de eventos encontrados na *Dark Web* classificados com Relevância Alta

Data	**Categoria**	**Texto**
28/07/2023	Knowledge and Information	Does anyone know a site to find leaked databases? **********/ you can find here link list of ransomware gangs that leak shit tons of data about companies and more google.com -_-
28/02/2022	N/A	has anyone confirmed it's legit? aka downloaded and read it? In general, everything vx-underground is archiving has been verified. They're archiving exploit, virus code and other stuff like that. But I personally didn't download it. Conti ransomware group previously put out a message siding with the Russian government. Today a Conti member has begun leaking data with the message "**** the Russian government, Glory to Ukraine!" You can download the leaked Conti data here: https://********/
22/08/2023	Hacking	i want to learn how to hack and how to dox but idk how, can you guys link me something great for doxxing especially Install Kali linux, read the man pages of packages you can start learning about network then computer science then use forums here they're the first source for actual hacking not that bullshit on youtube because trust me you"ll get in trouble in seconds after you do something illegal ,those platforms dont give hacking information only if they're sure they let a backdoor to find you.
08/08/2023	Hacking	I will get into the system but I'm afraid they would be able to trace the activities back to me. I'm using cobalt strike X2, I was able to log into, now I will download the second payload in order to insert a backdoor into the system
08/08/2023	Hacking	I have 3 versions of **** ****, the most recent ones, and a ransomware attack, all set up with targets and possible ones. All the details can be discussed privately Let's make money hell yea!

Fonte: Autoria própria.

4.5 Comparação entre os eventos da *Surface Web* e *Dark Web*

Para entender estatisticamente as relações entre os eventos encontrados na *Surface Web* e na *Dark Web*, foram feitas duas comparações, apresentadas a seguir, bem como a aplicação do LDA.

4.5.1 Comparação utilizando o cálculo de Similaridade de Cosseno

Essa comparação foi aplicada utilizando-se todos os eventos encontrados em cada base. Os eventos foram vetorizados e agregados em um vetor representativo médio (centróide) para cada base. Para a comparação das bases, não foi realizada uma análise pareada (evento a evento), somente base a base. Esta análise foi conduzida em quatro formas distintas:

1. Comparação utilizando **todos os eventos** (sem filtragem por classificação).
2. Comparação entre eventos classificados como **baixa relevância**.
3. Comparação entre eventos classificados como **média relevância**.
4. Comparação entre eventos classificados como **alta relevância**.

Os resultados detalhados de cada uma das formas de comparação propostas (itens 1 a 4) estão sumarizados na Tabela 12.

Tabela 12 – Resultado da comparação de SC, separado por relevância dos eventos.

Tipo	**Resultado**
SC Geral	$\sim$ 26,60%
SC (Relevância Baixa)	$\sim$ 26,60%
SC (Relevância Média)	$\sim$ 41,62%
SC (Relevância Alta)	$\sim$ 40,97%

Fonte: Autoria própria.

Os resultados apresentados na Tabela 12 detalham o grau de similaridade entre os eventos detectados na *Surface Web* e na *Dark Web* , divididos por faixas de relevância. Para este cálculo, utilizou-se a SC. Neste contexto, um valor próximo a 100% indica que o vocabulário dos eventos nos dois ambientes é semanticamente idêntico, enquanto valores baixos sugerem temáticas distintas. A porcentagem apresentada nos resultados é a multiplicação direta do resultado do SC multiplicado pelo valor 100.

Observa-se que a similaridade para a categoria 'Geral' foi de 26,6%. Isso significa que, ao considerar os dados sem filtros de relevância, existe uma sobreposição moderada de

textos entre os ambientes. Este resultado é fortemente influenciado pela classe de 'Baixa Relevância', que compõe mais de 80% dos dados em ambas as bases. Por outro lado, ao isolar eventos de 'Alta' e 'Média' relevância, nota-se uma proximidade semântica mais acentuada nos textos. Em todas essas comparações foi somente utilizado a métrica de SC, pois a métrica de SJ possui uma sensibilidade grande quando utilizado em textos de tamanhos diferentes.

4.5.2 Análise da Comparação Mensal entre os eventos da *Surface Web* e da *Dark Web*

Nessa análise, também foi feita uma comparação entre os eventos encontrados utilizando a SC, porém, foi feita de forma mensal para descobrir como a similaridade semântica varia mensalmente. Para começar, os dados foram agrupados mensalmente, porém, somente utilizando os anos de 2020 a 2023, pois não foram encontrados eventos novos com os dados de 2024 na *Surface Web*, conforme mostrado anteriormente (Tabela 3).

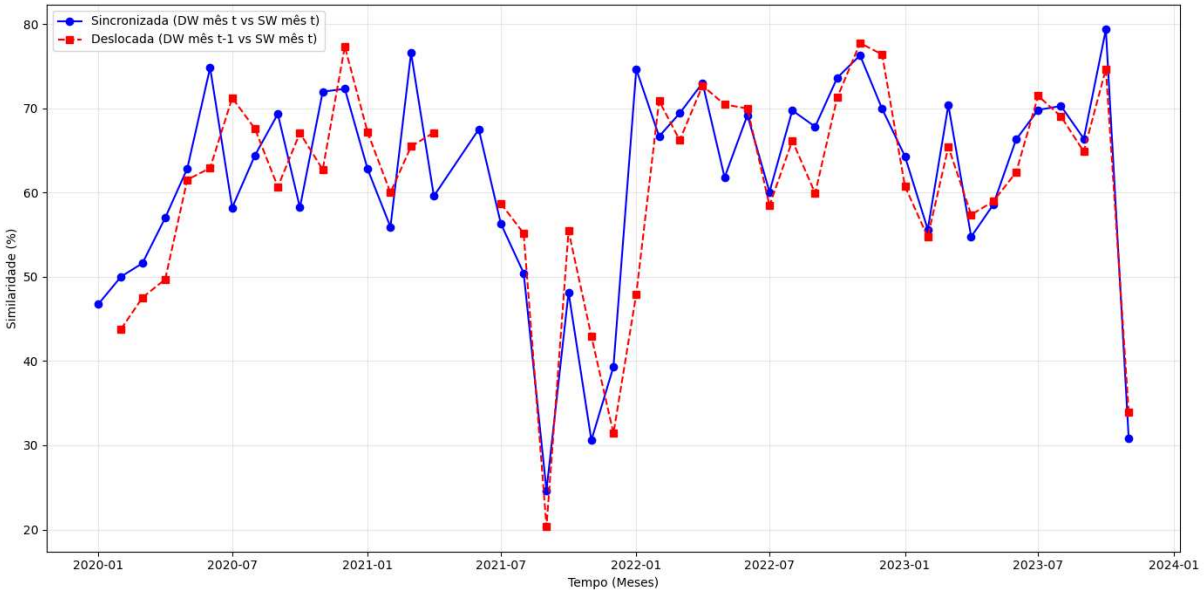
Após isso, foram comparados os dados de janeiro de 2020 da *Surface Web* com os de janeiro de 2020 da *Dark Web* e assim sucessivamente até dezembro de 2023. Os meses onde alguma das duas bases não possuíam dados não foram computados. Além dessa comparação mês a mês entre os dados, também foi feita uma análise com defasagem temporal. Essa análise com defasagem busca entender se as publicações encontradas na *Dark Web* precedem as da *Surface Web*. Para essa comparação utilizaram-se os eventos de um mês da *Dark Web* e comparou-se com os eventos da *Surface Web* do mês posterior. O resultado das duas análises (mês a mês e com defasagem de um mês) pode ser visto na Figura 16.

A análise da Figura 16 revela pontos em que a similaridade deslocada (linha vermelha) supera a similaridade síncrona (linha azul). Nestes períodos, a semelhança semântica entre os eventos da *Dark Web* do mês anterior $(t-1)$ e os dados da *Surface Web* do mês vigente (t) é superior à observada no comparativo em tempo real. Como exemplo, observa-se que em dezembro de 2020 houve uma similaridade deslocada maior, indicando que os temas discutidos em novembro daquele ano na *Dark Web* aparecem também na *Surface Web* no mês subsequente. Para exemplificar, em novembro de 2020, identificou-se na *Dark Web* uma discussão sobre a obtenção de contas no serviço *Riseup* e métodos de acesso anônimo via redes públicas. No mês seguinte, em dezembro de 2020, observou-se na *Surface Web* a publicação de guias educativos detalhando métodos de envio de e-mails anônimos e proteção de identidade.

4.5.3 Aplicação do LDA

Para descobrir os temas mais discutidos nos eventos encontrados classificados com maior relevância, foi realizado a aplicação da LDA em cada uma das bases (*Surface web*

Figura 16 – Cálculo de SC entre os eventos encontrados mês a mês e com deslocamento de um mês entre as bases.



Fonte: Autoria própria.

e *Dark Web*). Para esse cálculo, considerou-se todos os eventos classificados com alta e média relevância. Como mencionado na Seção 3.3, foi empregada a biblioteca *gensim*, utilizando cinco tópicos como saída do algoritmo. O número cinco foi escolhido, pois foram feitos testes com dois, cinco, seis, dez e vinte tópicos, e o índice de coerência foi praticamente o mesmo em todos os testes e optou-se por usar cinco tópicos para mantê-los mais abrangentes e interpretáveis.

O resultado da aplicação da LDA para os dados da *Surface Web* e *Dark Web* utilizando somente eventos classificados em relevância Alta e Média podem ser vistos nas Tabelas 13 e 14 respectivamente. Foram utilizadas somente as 30 primeiras palavras de cada tópico. Para validar a qualidade da aplicação do LDA, foram calculados os coeficientes de *perplexity* e índice de coerência de cada resultado. O índice de coerência foi calculado utilizando também a biblioteca *gensim*. Os resultados dos coeficientes podem ser vistos na Tabela 15.

A análise comparativa revela que o modelo LDA apresentou maior consistência semântica na *Surface Web* ($C_v = 0{,}46$) em relação à *Dark Web* ($C_v = 0{,}29$), com perplexidades próximas a -8. Na *Surface Web*, os tópicos são mais estruturados e técnicos, focando em vulnerabilidades de sistemas operacionais (Tópico 1), ferramentas de automação e validação de e-mails/SMTP (Tópico 2), além de discussões geopolíticas e ferramentas corporativas associadas a incidentes (Tópico 3). Por outro lado, a *Dark Web* exibe maior foco na comercialização de dados ilícitos, como documentos forjados (Tópico 4), contas invadidas e acesso a infraestruturas de nuvem (AWS).

Tabela 13 – Tópicos dos eventos novos classificados com relevância alta.

Surface Web	
Tópico	Principais Palavras
1	bank, carding, email, hacking, logs, card, sell, paypal, buy, vzlom, method, hack, cvv, good, credit, account, telegram, money, transfer, smtp, cashout, free, cash, fresh, per, methods, get, whatsapp, balance, name
2	content, download, login, link, file, register, com, password, hidden, click, thanks, view, leaked, locked, files, thank, text, virustotal, expand, wordpress, need, please, create, use, leak, virus, get, times, new, leaks
3	tools, scanner, malware, programs, keylogger, use, virus, system, hacking, antivirus, hack, trojan, anti, computer, sql, viruses, ddos, security, exploits, free, keyloggers, cracking, exploit, doxing, botnet, pack, tool, wifi, rat, apk
4	data, account, security, information, also, one, phone, access, password, use, address, social, hackers, hacking, email, accounts, attack, like, phishing, hacker, software, number, secure, time, facebook, need, may, online, personal, using
5	using, use, code, server, windows, file, vulnerability, network, password, tool, click, system, web, type, security, target, hacking, open, user, command, exploit, list, used, install, run, step, following, need, script, get
Dark Web	
Tópico	Principais Palavras
1	privacy, check, browser, use, data, fingerprint, security, domain, apps, google, web, always, test, testing, tor, without, pages, services, also, using, tool, tools, list, avoid, unique, website, helps, detect, users, system
2	use, tools, find, database, com, kali, thank, sql, hacking, need, sites, install, could, cracking, day, file, lot, github, anyone, know, work, looking, free, nice, data, tool, parrot, iso, answering, vmware
3	want, website, need, also, info, keepass, hacking, use, access, like, anyone, find, change, help, add, way, really, look, control, devices, create, information, know, thanks, without, click, got, content, would, device
4	one, images, registration, data, without, see, internet, files, like, file, type, text, supports, imgur, videos, linux, want, used, new, hacking, although, real, many, since, level, even, looking, found, companies, online
5	open, hacking, bluetooth, use, try, injection, hack, website, like, access, ports, brute, ssh, force, tool, get, data, password, tcp, sql, want, knowledge, learn, need, ddos, version, exploit, port, device, good

Fonte: Autoria própria.

Tabela 14 – Tópicos dos eventos novos classificados com relevância média.

Surface Web	
Tópico	Principais Palavras
1	windows, download, password, link, virustotal, unzip, use, code, bit, file, checker, get, data, tools, microsoft, email, files, proxy, features, office, home, mail, access, system, new, version, support, account, click, language
2	buy, online, passport, license, driver, registered, real, card, driving, german, passports, bank, cvv, contact, name, database, number, cards, email, balance, dumps, per, pin, account, full, sale, fullz, telegram, credit, sell
3	fake, documents, information, security, data, use, card, drivers, new, materials, address, portuguese, time, used, application, order, mac, one, client, system, mustang, service, without, number, verification, turkish, everything, network, specialists, also
4	content, register, login, hidden, view, thanks, checker, capture, click, need, expand, accounts, base, account, gif, said, locked, lgodlhaqab, braa, please, image, data, text, cannot, sign, quoted, thank, steam, order, leak
5	win, pro, linux, suite, software, studio, professional, design, build, visual, enterprise, cad, bentley, synopsys, edition, plus, systems, bit, engineering, embedded, siemens, analysis, designer, cadence, development, integrated, edge, autocad, data, ram
Dark Web	
Tópico	Principais Palavras
1	would, court, building, police, sites, one, even, make, network, hacking, using, could, future, devices, kali, equipment, data, tvn, computers, get, several, help, like, need, personal, school, want, two, central, hidden
2	like, use, know, find, way, data, need, phone, one, someone, get, free, also, want, people, using, name, even, good, think, many, search, would, make, used, information, google, something, help, without
3	command, value, mining, list, key, echoln, let, data, malware, also, code, set, use, created, two, file, make, one, map, string, tool, like, open, via, commands, returns, block, hardware, contain, ransomware
4	data, meta, new, aws, exploit, learning, port, ion, identify, big, create, according, adults, open, like, machine, systems, ftp, ports, messages, emails, key, including, people, message, try, voyager, using, privacy, children
5	buy, use, also, fake, online, time, documents, account, want, real, linux, images, cards, sharex, first, get, instagram, certificates, passport, certificate, second, card, work, file, license, long, free, much, little, service

Fonte: Autoria própria.

Tabela 15 – Comparação de métricas do modelo LDA: *Surface Web* vs. *Dark Web*

Base de Dados	**Perplexidade**	**Índice de Coerência (C_v)**
Surface Web	-8,2198	0,4685
Dark Web	-7,4613	0,2936

Fonte: Autoria própria.

4.6 Implicações práticas e limitações

A metodologia proposta neste trabalho para a busca de eventos novos pode auxiliar na melhoria de sistemas de CTI nos seguintes pontos:

- **Monitoramento Preditivo**: Nos testes realizados, é possível notar que algumas publicações na *Dark Web* antecedem a *Surface Web* em cerca de um mês. Isso permite que empresas de CTI criem sistemas de alerta antecipado. Profissionais podem monitorar publicações na *Dark Web* para preparar defesas (atualizar assinaturas de IDS/IPS, aplicar patches) antes que o assunto se torne popular na rede aberta;
- **Vulnerabilidades Ativas**: Se uma vulnerabilidade específica apresenta um aumento de similaridade semântica entre os dois ambientes (*Surface Web* e *Dark Web*), significa que o "conhecimento"sobre como explorá-la está se espalhando. Isso justifica uma janela de manutenção emergencial para aplicar o patch;
- **Score de Antecedência**: Um produto poderia classificar ameaças com base em quão cedo elas apareceram na *Dark Web* em relação à *Surface Web*;

O trabalho proposto também apresenta algumas limitações na sua metodologia. As principais limitações são apresentadas a seguir:

- **Escopo das Fontes de Dados**: O estudo utilizou fóruns abertos. No entanto, não foram incluídos canais fechados de *Telegram*, *Discord* ou ICQ, que são muito utilizados por cibercriminosos para troca de informações e vendas diretas;
- **Barreira Linguística**: A análise concentrou-se somente em conteúdos em inglês. Ameaças originadas em fóruns russos, chineses ou de língua portuguesa podem apresentar dinâmicas diferentes que não foram capturadas;
- **Granularidade Temporal**: A análise mensal pode mascarar ataques de "Dia Zero"(Zero-day) que ocorrem e são explorados em uma escala de dias ou horas.;
- **Ruído nos Embeddings**: Embora o SBERT seja robusto, gírias técnicas muito específicas ou códigos ofuscados podem não ter sua semântica 100% capturada pela vetorização, impactando levemente os índices de similaridade;

Capítulo 5

Conclusão

Este trabalho propôs um método para a busca e análise de eventos novos em fóruns de cibersegurança. Para isso, tomou-se como base o trabalho de Bose et al. (2020), e empregaram-se técnicas de PLN e aprendizado de máquina para a busca de eventos novos. A metodologia baseada em comparação semântica de textos se mostrou eficaz ao identificar 26.071 eventos utilizando uma base de dados com 30.584 publicações. O agrupamento realizado através do algoritmo *k-means* e a vetorização através do *SBERT* trouxeram resultados relevantes na identificação de padrões emergentes em dados de fóruns da *Dark Web* e da *Surface Web*. Os resultados obtidos ao longo do período de janeiro de 2020 a novembro de 2023 confirmam a hipótese inicial e contribuem para a melhoria na construção de sistemas CTI.

A análise dos dados revelou diferenças na dinâmica das redes: enquanto a *Surface Web* apresentou um crescimento gradual e maior estabilidade dos temas, a *Dark Web* demonstrou alta volatilidade, com picos de atividade seguidos de silêncio operacional. Esses resultados confirmam a hipótese de que é possível correlacionar as bases para identificar a propagação de ameaças, contribuindo para a construção de sistemas CTI mais reativos.

Em resumo, as principais contribuições deste trabalho são listadas a seguir:

1. **Metodologia de Busca:** Desenvolvimento e validação de um fluxo de busca de novos eventos em fóruns da *Dark Web* e *Surface Web*.

2. **Análise comparativa:** Investigação da correlação e dinâmica de propagação de ameaças entre as duas bases.

3. **Base Rotulada:** Construção de uma base rotulada com eventos considerados de baixa, média ou alta relevância. Esta base está disponível publicamente no repositório Mendeley Data (SILVEIRA, 2026).

As limitações deste trabalho incluem a ausência da aplicação da metologia proposta em um conjunto diferente de dados. Além disso, a validação dos eventos encontrados

por outros especialistas poderia ser realizada para verificar se eles representam um evento real.

Como trabalhos futuros, pretende-se utilizar outras fontes de dados além de fóruns *hackers* , como, por exemplo, *Telegram* ou *Reddit*. Além disso, também é possível melhorar o tempo de processamento de dados, além de realizar experimentos com outras formas de vetorização semântica.

5.1 Contribuições em Produção Bibliográfica

Este trabalho resultou em uma publicação apresentada no *International Conference on Machine Learning and Applications* (ICMLA) em 2025:

> SILVEIRA, R. B.; MIANI, R. S.; GABRIEL, P. H. R. Event detection in dark web and surface web forums using machine learning. In: 2025 International Conference on Machine Learning and Applications. New York, NY: IEEE, 2025. p. 1064–1069.

Referências

AMPEL, B.; SAMTANI, S.; ZHU, H.; ULLMAN, S.; CHEN, H. Labeling hacker exploits for proactive cyber threat intelligence: A deep transfer learning approach. In: **2020 IEEE International Conference on Intelligence and Security Informatics (ISI)**. New York, NY: IEEE, 2020. p. 1–6.

APUTUNN. **Case Study: The Equifax Hack 2017**. 2023. Disponível em: <https://aputunn.medium.com/case-study-the-equifax-breach-2017-682f58f1b36b>.

ASBAS, C.; TUZLUKAYA, S. Cyberattack and cyberwarfare strategies for businesses. In: ÖZSUNGUR, F. (Ed.). **Conflict Management in Digital Business**. [S.l.]: Emerald Publishing Limited, 2022. p. 303–328.

AYAN, E. T.; ZENGIN, M. S.; DENIZ, G.; DURU, H. A.; BARDAK, B. Interpretable cybersecurity event detection in Turkish: A novel dataset. In: **2022 Innovations in Intelligent Systems and Applications Conference (ASYU)**. New York, NY: IEEE, 2022. p. 1–6.

BARRIOS, F.; LÓPEZ, F.; ARGERICH, L.; WACHENCHAUZER, R. Variations of the similarity function of textrank for automated summarization. **CoRR**, abs/1602.03606, 2016. Disponível em: <http://arxiv.org/abs/1602.03606>.

BENJAMIN, V.; LI, W.; HOLT, T.; CHEN, H. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In: **2015 IEEE International Conference on Intelligence and Security Informatics (ISI)**. New York, NY: IEEE, 2015. p. 85–90.

BIRD, S.; LOPER, E.; KLEIN, E. **Natural Language Processing With Python**. Santa Rosa, CA: O'Reilly, 2009. 504 p.

BIRUNDA, S. S.; DEVI, R. K. A review on word embedding techniques for text classification. In: RAJ, J. S.; ILIYASU, A. M.; BESTAK, R.; BAIG, Z. A. (Ed.). **Innovative Data Communication Technologies and Application**. Singapore: Springer-Verlag, 2021, (Lecture Notes on Data Engineering and Communications Technologies, v. 59). p. 267–281.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, jan. 2003.

BOSE, A.; BEHZADAN, V.; AGUIRRE, C.; HSU, W. H. A novel approach for detection and ranking of trendy and emerging cyber threat events in Twitter streams. In: **Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining**. New York, NY: ACM, 2020. p. 871–878.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J.; TSENG, V. S.; CAO, L.; MOTODA, H.; XU, G. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Heidelberg: Springer-Verlag, 2013, (Lecture Notes in Computer Science, v. 7819). p. 160–172.

CUI, J.; KIM, H.; JANG, E.; YIM, D.; KIM, K.; LEE, Y.; CHUNG, J.-W.; SHIN, S.; LIAO, X. Tweezers: A framework for security event detection via event attribution-centric tweet embedding. In: **Network and Distributed System Security Symposium**. Reston, VA: The Internet Society, 2025.

DALE, D.; MCCLANAHAN, K.; LI, Q. AI-based cyber event OSINT via twitter data. In: **2023 International Conference on Computing, Networking and Communications (ICNC)**. New York, NY: IEEE, 2023. p. 436–442.

DENNY, M. J.; SPIRLING, A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. **Political Analysis**, v. 26, n. 2, p. 168–189, abr. 2018.

DICE, L. R. Measures of the amount of ecologic association between species. **Ecology**, v. 26, n. 3, p. 297–302, 1945.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

GAIKWAD, S. V.; CHAUGULE, A.; PATIL, P. Text mining methods and techniques. **International Journal of Computer Applications**, v. 85, n. 17, 2014.

GHAHRAMANI, Z. Unsupervised learning. In: BOUSQUET, O.; LUXBURG, U.; RÄTSCH, G. (Ed.). **Advanced Lectures on Machine Learning**. Heidelberg: Springer-Verlag, 2004, (Lecture Notes in Computer Science, v. 3176). p. 72–112.

HASAN, M.; ORGUN, M. A.; SCHWITTER, R. Real-time event detection from the Twitter data stream using the TwitterNews+ framework. **Information Processing & Management**, v. 56, n. 3, p. 1146–1165, maio 2019. ISSN 0306-4573.

HAVELIWALA, T. H. Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search. **IEEE Transactions on Knowledge and Data Engineering**, v. 15, n. 4, jul. 2003.

HUANG, C.; GUO, Y.; GUO, W.; LI, Y. HackerRank: Identifying key hackers in underground forums. **International Journal of Distributed Sensor Networks**, v. 17, n. 5, p. 1–12, maio 2021. ISSN 1550-1329.

HUGHES, J.; AYCOCK, S.; CAINES, A.; BUTTERY, P.; HUTCHINGS, A. Detecting trending terms in cybersecurity forum discussions. In: XU, W.; RITTER, A.; BALDWIN, T.; RAHIMI, A. (Ed.). **Proceedings of the Sixth Workshop on Noisy User-generated Text**. Online: ACL, 2020. p. 107–115.

JESUS FILHO, S. A. de. **Identificação de posts maliciosos na dark web utilizando Aprendizado de Máquina Supervisionado**. 111 p. Dissertação (Mestrado) — Universidade Federal de Uberlândia, jan. 2024.

KIM, S.-W.; GIL, J.-M. Research paper classification systems based on TF-IDF and LDA schemes. **Human-centric Computing and Information Sciences**, v. 9, n. 30, p. 1–21, ago. 2019.

KOLOVEAS, P.; CHANTZIOS, T.; TRYFONOPOULOS, C.; SKIADOPOULOS, S. A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence. In: **2019 IEEE World Congress on Services**. New York, NY: IEEE, 2019. p. 3–8.

KRIPPENDORFF, K. Estimating the reliability, systematic error and random error of interval data. **Educational and Psychological Measurement**, v. 30, n. 1, p. 61–70, 1970.

LE SCELLER, Q.; KARBAB, E. B.; DEBBABI, M.; IQBAL, F. SONAR: Automatic detection of cyber security events over the Twitter stream. In: **Proceedings of the 12th International Conference on Availability, Reliability and Security**. New York, NY, USA: ACM, 2017. p. 1–11.

LEE, K.-C.; HSIEH, C.-H.; WEI, L.-J.; MAO, C.-H.; DAI, J.-H.; KUANG, Y.-T. Sec-Buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation. **Soft Computing**, v. 21, n. 11, p. 2883–2896, jun. 2017.

LI, Q.; NOURBAKHSH, A.; SHAH, S.; LIU, X. Real-time novel event detection from social media. In: **2017 IEEE 33rd International Conference on Data Engineering**. New York, NY: IEEE, 2017. p. 1129–1139.

LIU, Y.; LIN, F. Y.; AHMAD-POST, Z.; EBRAHIMI, M.; ZHANG, N.; HU, J. L.; XIN, J.; LI, W.; CHEN, H. Identifying, collecting, and monitoring personally identifiable information: From the dark web to the surface web. In: **2020 IEEE International Conference on Intelligence and Security Informatics**. New York, NY: IEEE, 2020. p. 1–6.

MEAD, E.; AGARWAL, N. Surface web vs deep web vs dark web. In: SCHINTLER, L. A.; MCNEELY, C. L. (Ed.). **Encyclopedia of Big Data**. Cham: Springer International Publishing, 2022. p. 901–905.

MEADOW, C. T.; BOYCE, B. R.; KRAFT, D. H.; BARRY, C. **Text Information Retrieval Systems**. 3. ed. London, UK: Academic Press, 2007. 371 p.

MIHALCEA, R.; TARAU, P. TextRank: Bringing order into text. In: LIN, D.; WU, D. (Ed.). **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411.

NIWATTANAKUL, S.; SINGTHONGCHAI, J.; NAENUDORN, E.; WANAPU, S. Using of Jaccard coefficient for keywords similarity. In: AO, S.-I.; CASTILLO, O.; DOUGLAS, C.; FENG, D. D.; LEE, J.-A. (Ed.). **Proceedings of the International MultiConference of Engineers and Computer Scientists**. Hong Kong: International Association of Engineers, 2013. p. 1–5.

NUNES, E.; DIAB, A.; GUNN, A.; MARIN, E.; MISHRA, V.; PALIATH, V.; ROBERTSON, J.; SHAKARIAN, J.; THART, A.; SHAKARIAN, P. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In: **2016 IEEE Conference on Intelligence and Security Informatics (ISI)**. New York, NY: IEEE, 2016. p. 7–12.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, 1998.

PATEL, K. M. A.; THAKRAL, P. The best clustering algorithms in data mining. In: **2016 International Conference on Communication and Signal Processing (ICCSP)**. New York, NY: IEEE, 2016. p. 2042–2046.

PRIYADARSHINI, I.; KUMAR, R.; SHARMA, R.; SINGH, P. K.; SATAPATHY, S. C. Identifying cyber insecurities in trustworthy space and energy sector for smart grids. **Computers & Electrical Engineering**, v. 93, jul. 2021.

RAMIREZ, S. Cybersecurity attacks statistics 2025: Trends, costs, and implications. **SQ Magazine**, out. 2025. Disponível em: <https://sqmagazine.co.uk/cybersecurity-attacks-statistics/>. Acesso em: 15 dez. 2025.

ŘEHŮŘEK, R.; SOJKA, P. Software framework for topic modelling with large corpora. In: **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**. Valletta, Malta: University of Malta, 2010. p. 45–50.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: INUI, K.; JIANG, J.; NG, X. W. V. (Ed.). **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3982–3992.

SABOTTKE, C.; SUCIU, O.; DUMITRAS, T. Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits. In: **Proceedings of the 24th USENIX Security Symposium**. USA: USENIX Association, 2015. p. 1041–1056.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, v. 18, n. 11, p. 613–620, nov. 1975.

SAPIENZA, A.; BESSI, A.; DAMODARAN, S.; SHAKARIAN, P.; LERMAN, K.; FERRARA, E. Early warnings of cyber threats in online discussions. In: **2017 IEEE International Conference on Data Mining Workshops (ICDMW)**. New York, NY: IEEE, 2017. p. 667–674.

SAPIENZA, A.; ERNALA, S. K.; BESSI, A.; LERMAN, K.; FERRARA, E. DISCOVER: Mining online chatter for emerging cyber threats. In: **Companion of the The Web Conference 2018 on The Web Conference 2018**. Switzerland: International World Wide Web Conferences Steering Committee, 2018. p. 983–990.

SHI, L.-L.; LIU, L.; WU, Y.; JIANG, L.; HARDY, J. Event detection and user interest discovering in social media datastreams. **IEEE Access**, v. 5, p. 20953–20964, 2017.

SILVEIRA, R. **Labeled Dataset of Cybersecurity Events from Dark Web and Surface Web Forums**. Mendeley Data, 2026. Disponível em: <https://data.mendeley.com/datasets/jt8yw5gdr2/1>.

SØRENSEN, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. **Biol. Skr.**, v. 5, p. 1–34, 1948.

TANOVIĆ, A.; VRANJKOVIĆ, V. Implementation of parallel k-means algorithm for image classification using OpenMP and MPI libraries. In: **2024 Zooming Innovation in Consumer Technologies Conference (ZINC)**. New York, NY: IEEE, 2024. p. 54–59.

TROUDI, A.; ZAYANI, C. A.; JAMOUSSI, S.; AMOR, I. A. B. A new mashup based method for event detection from social media. **Information Systems Frontiers**, v. 20, n. 5, p. 981–992, out. 2018.

ZHUANG, F.; CHENG, X.; LUO, P.; PAN, S. J.; HE, Q. Supervised representation learning: Transfer learning with deep autoencoders. In: YANG, Q.; WOOLDRIDGE, M. (Ed.). **Proceedings of the 24th International Conference on Artificial Intelligence**. Washington, DC: AAAI Press, 2015.