UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Gustavo Guimarães Reis

Análise de incidentes de segurança utilizando dados do Twitter

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Gustavo Guimarães Reis

Análise de incidentes de segurança utilizando dados do Twitter

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Rodrigo Sanches Miani

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2023

Gustavo Guimarães Reis

Análise de incidentes de segurança utilizando dados do Twitter

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 23 de junho de 2023:

Prof. Dr. Rodrigo Sanches Miani
Orientador

Uberlândia, Brasil
2023

Dedico aos meus pais, que me ofereceram todo suporte e apoio durante minha jornada. Devido aos seus intensos esforços e sacrifícios estou concluindo mais uma etapa da minha vida

Agradecimentos

À minha mãe, que sempre foi o maior exemplo de disciplina que tive na minha vida, e por me incentivar constantemente a dar prioridade aos estudos. Ao meu pai, meu exemplo diário de honestidade, esforço e dedicação, responsável por me encorajar habitualmente à prática da excelência. Aos grandes amigos que fiz durante os árduos anos de faculdade, compartilhando as aflições entre as incontáveis semanas de provas e entregas de trabalhos. À minha namorada que me motivou durante todo o período da realização deste trabalho e que me deu forças para não desistir mesmo eu estando em muitos momentos ausente com os afazeres da faculdade. E por fim e não menos importante ao meu orientador, Rodrigo Sanches Miani, pois sem a sua colaboração eu não teria conseguido me adaptar a tantas dificuldades e empecilhos que surgiram durante esta jornada. Serei eternamente grato a ele por me incitar o exercício da pesquisa e por me mostrar a beleza em compartilhar o conhecimento com os meus semelhantes.

"A educação tem raízes amargas, mas os seus frutos são doces" (Aristóteles)

Resumo

Em uma sociedade cada vez mais conectada o papel da Segurança da Informação se torna imprescindível. É de suma importância a descoberta e pesquisa de novos meios para se investigar e analisar incidentes de segurança. Nesse contexto, as Redes Sociais se mostram como uma ótima ferramenta para extração de dados referentes às interações de indivíduos ao se eclodir um evento de cibersegurança. Em especial, o *Twitter*, por meio da interação principal via mensagens de texto curtas, apresenta-se como um escolha tangível para realizar a investigação de incidentes. Por meio da manipulação das postagens de usuários pode-se extrair informações valiosas a respeito dos eventos. Este trabalho, portanto, visa sugerir um processo para a coleta, visualização, armazenamento e análise de dados extraído do *Twitter*. Por conseguinte, será realizado a sugestão de ferramentas simples, *open source* e que facilitem as tarefas de um cientista de dados. Ademais, foi exemplificado tal metodologia por meio da investigação, coleta e posterior análise de um estudo de caso envolvendo um vazamento de dados da rede social *Facebook*, bem como os resultados das análises gráficas que apontaram uma alta quantidade de postagens nos dias subsequentes após a divulgação do incidente pelos grandes meios de comunicação.

Palavras-chave: Segurança da Informação, Vazamento de dados, Twitter, Análise de dados, Python

Lista de ilustrações

Lista de abreviaturas e siglas

DoS	*Denial of Service*
FIPS	*Federal Information Processing Standards*
NIST	*National Institute of Standards and Technology*
NIS	*Network and Information Systems Cooperation Group*
ENISA	*European Union Agency for Network and Information Security*
API	*Application Programming Interface*
HTTP	*Hypertext Transfer Protocol*
HTML	*HyperText Markup Language*
SGBD	Sistema Gerenciador de Banco de Dados
SQL	*Structured Query Language*
NoSQL	*Not only SQL*
PLN	Processamento de Linguagem Natural
JSON	*JavaScript Object Notation*
BSON	*Binary JavaScript Object Notation*
URI	*Uniform Resource Identifier*
DRI	*Digital Rights Ireland*

Sumário

1 Introdução

A segurança da informação desempenha um papel crucial e de extrema importância na sociedade atual. Com o avanço contínuo da tecnologia e a digitalização das interações sociais, a proteção dos dados e informações pessoais tornou-se uma necessidade imprescindível. Desde transações financeiras até informações confidenciais de empresas e governos, a segurança da informação é essencial para garantir a privacidade, a integridade e a disponibilidade dos dados.

Em um mundo conectado, onde a troca de informações ocorre em um ritmo acelerado, ameaças significativas, tais como *hackers*, criminosos cibernéticos e outras entidades maliciosas estão buscando constantemente informações valiosas e vulnerabilidades de softwares. Segundo a consultoria de gestão alemã *Roland Berger*, em 2021 foram registrados 9,1 milhões de ocorrências relacionadas a crimes cibernéticos no Brasil. Além disso, a empresa apontou que o país é o quinto que mais sofre pela falta de cibersegurança no mundo (REPORT, 2022).

Por outro lado, as redes sociais consolidaram seu papel significativo na sociedade moderna, possuindo uma importância abrangente em vários aspectos. Desde o auxílio à comunicação e conexão de pessoas até à facilitação ao acesso à informação, democratizando, de certa forma, o compartilhamento de conhecimento e aprendizado entre os indivíduos.

Com o intuito de ressaltar a relação entre essas duas entidades o artigo de Santos et al. (2013) apresenta um método de extração de notificações de segurança a partir de mensagens postadas em redes sociais. Os autores desenvolveram uma metodologia para extrair e evidenciar alertas e problemas de segurança em ambientes de redes de computadores por meio da coleta e do processamento de mensagens postadas no *Twitter*, usando, principalmente técnicas de mineração de dados otimizadas com heurísticas de aumento de relevância.

Além disso, no trabalho Ritter et al. (2015) é proposto um modelo de aprendizagem de máquina fracamente supervisionado de extração de eventos a partir de dados provindos do *Twitter*. A proposta consiste na criação de um extrator de eventos que possa identificar rapidamente eventos emergentes de segurança, especificamente sequestro de contas, vazamento de dados e ataques de negação de serviço *DoS - Denial of Service*. Para tal propósito, foi necessário a criação de uma *bag of tweets* contendo exemplos de amostras de textos de postagens da rede social que faziam referência a estas categorias de incidentes.

Portanto, é notório como as redes sociais podem auxiliar na identificação e predição de incidentes de segurança. Dessa forma, evidencia-se uma necessidade de se procurar

meios e alternativas para lidar com a coleta e a análise de tais dados. Neste contexto, o presente trabalho procura sugerir um processo sistemático para se realizar todos os passos envolvendo tal problemática, explorando ferramentais que possam auxiliar e agilizar o trabalho de manipulação dos dados.

1.1 Objetivos

O objetivo geral deste trabalho é criar uma metodologia para a coleta e análise de dados sobre postagens de usuários que participem de discussões a cerca de incidentes de segurança provenientes de redes sociais, especificamente do *Twitter*. Considerando o desenvolvimento do trabalho e o objetivo geral apresentado, destacam-se os seguintes objetivos específicos:

- Exemplificar o passo a passo de um processo de coleta e análise de dados em um estudo de caso real, envolvendo um vazamento de dados de usuários do *Facebook*;
- Sugerir a utilização de ferramentas simples, *open source* e de alto nível para facilitar e agilizar os procedimentos de coleta, armazenamento e análise de dados;
- Realizar uma breve análise sobre o comportamento dos usuários ao engajarem em discussões a cerca de um incidente de segurança, especificamente de um caso de vazamento de dados;
- Investigar o impacto e a relevância das postagens sobre incidentes de segurança no Twitter;
- Disponibilizar o código-fonte e base de dados criados para exemplificar a metodologia.

1.2 Justificativa

Atualmente, os dados se tornaram uma das mais valiosas matérias-primas do mundo e, indubitavelmente são de fundamental importância para a sociedade, tanto no que tange à pesquisa científica e avanço do conhecimento, quanto para a segurança e detecção de fraudes. Por meio da análise de padrões, por exemplo, é possível identificar atividades suspeitas, prevenir ataques cibernéticos, mitigar riscos e proteger informações confidenciais.

O presente trabalho visa realçar a participação das redes sociais como fonte de dados valiosas sobre o comportamento de usuários, mostrando-se como uma ferramenta fundamental para a detecção e prevenção de ataques cibernéticos. A partir de um método definido e claro para a coleta e análise de dados provindos de tais fontes, a criação de

modelos de aprendizagem de máquina que reconheçam padrões em postagens torna-se factível para auxiliar a predição e detecção de incidentes de segurança.

1.3 Organização do Trabalho

Os próximos Capítulos estão divididos da seguinte forma.

Capítulo 2 - Referencial Teórico: Fundamentação teórica base para desenvolvimento do trabalho. Apresenta alguns conceitos e ferramentas de desenvolvimento, ademais discorre brevemente sobre alguns trabalhos relacionados.

Capítulo 3 - Desenvolvimento: Descreve as etapas necessárias para o desenvolvimento da metodologia, em conjunto com a exemplificação de um estudo de caso.

Capítulo 4 - Resultados: Descreve os resultados alcançados ao realizar as etapas da metodologia, além dos entregáveis que foram produzidos durante a realização do trabalho.

Capítulo 5 - Conclusão: Por fim, na conclusão, além de um breve resumo do trabalho, destacando as suas principais contribuições, é proposto algumas sugestões para trabalhos futuros.

2 Referencial Teórico

Serão definidos, neste capítulo do trabalho, algumas ferramentas e conceitos fundamentais para o seu melhor entendimento, além disso, serão apresentados trabalhos correlatos.

2.1 Conceitos Fundamentais

Essa seção da fundamentação teórica irá apresentar alguns conceitos fundamentais sobre segurança da informação, mineração de texto, *APIs* e *Web Scraping*, que são de suma importância no contexto deste trabalho.

2.1.1 Segurança da Informação

O papel da Segurança da informação é assegurar a continuidade dos negócios e minimizar os danos causados a eles através da limitação dos impactos dos incidentes de segurança (SOLMS, 1998).

Dentre os conceitos existentes sobres os objetivos fundamentais da segurança, indubitavelmente, a tríade CIA do acrônimo inglês: *confidenciability, integrity and availability* são um dos mais relevantes. Definidos pelo padrões FIPS (*Federal Information Processing Standards*) 199 da NIST (*National Institute of Standards and Techonology*) (FIPS, 2004) como:

- Confidencialidade: preservar restrições sobre acesso e divulgação de informação, incluindo meios para proteger a privacidade de indivíduos e informações privadas.
- Integridade: prevenir-se contra a modificação ou destruição impropria de informação, incluindo a autenticidade dela.
- Disponibilidade: assegurar acesso e uso rápido e confiável da informação.

Esses três itens são considerados objetivos a serem alcançados para se ter uma segurança bem estabelecida de dados e sistemas de informação. Dessa forma, quaisquer eventos que violem os requisitos citados são considerados ataques (STALLINGS, 2014), podendo ser catalogados como incidentes de segurança. Destarte, um **incidente de segurança** pode ser definido como qualquer evento adverso, confirmado ou sob suspeita, relacionado à segurança de sistemas de computação ou redes de computadores (CERT.BR, 2022). Por outro lado, Gualberto et al. (2013) definem incidente de segurança como um

evento que tem alta probabilidade de impactar o negócio e segurança da organização. E esses surgem a partir de explorações de vulnerabilidades de segurança. Dentre as vulnerabilidades, falhas de projetos, na implementação ou configuração de programas, ou seja, qualquer condição que possa ser explorada em um ataque (CERT.BR, 2022).

Conforme documento do NIS (2018) - *Network and Information Systems Cooperation Group*, no qual propõem uma taxonomia simples e de alto nível para classificar incidentes de segurança tanto de nível político quanto estratégico, pode-se subdividir a classificação em duas partes fundamentais:

- **natureza do incidente**: o que causou o acidente - falhas de sistemas, fenômeno natural, erros humanos, ações maliciosas; potencial da ameaça - alta média, baixa.
- **impacto do incidente**: quais setores foram afetados - energia, transporte, bancário, saúde, infraestrutura digital, serviços digitais, governo, entre outros; severidade do impacto - vermelho, amarelo, verde, branco - o que significa impactos respectivamente - muito grandes, grandes, pequenos e inexistentes; o panorama para as próximas 6 horas - melhorar, estabilizar e piorar.

Rotulagem	**Exemplos**
Conteúdo Abusivo	Spam, Discurso Nocivo, Violência Sexual e infantil
Código Malicioso	Virus, Worm, Trojan, Sypware, Dialler, Rootkit
Coleta de Informações	Scanning, Sniffing, Engenharia Social
Tentativa de Intrusão	Exploração de Vulnerabilidade, Tentativa de login, Novas formas de ataques
Intrusão	Comprometimento de contas, Comprometimento da aplicação, Bots
Disponibilidade	Denial of Service (DoS/ DDoS), Sabotagem, Interrupção
Segurança do Conteúdo da Informação	Acesso não autorizado, Modificação não autorizada
Fraude	Phishing, Identidade falsa (masquerade), Propriedade intelectual (copyright), Utilização não autorizada de recursos

Tabela 1 – Referência e Classificação Taxonômica de incidentes de segurança

Além disso, pode-se utilizar uma taxonomia técnica para rotular de forma mais específica a natureza do incidente. A Tabela 1 contém as principais rotulagens técnicas

desenvolvida pela ENISA - *European Union Agency for Network and Information Security* (ENISA, 2018).

Dessa forma, é possível unir as duas taxonomias e criar descrições claras e objetivas sobre os incidentes contendo todas ou algumas características, seguindo o padrão: [<causa>(<causa-técnica>, <causa-técnica-específica>)<severidade-da-ameaça>; <setor-impactado>, <severidade-do-impacto>, <panorama>].

Segue abaixo alguns exemplos:

- ação-maliciosa (código malicioso, trojan), médio; serviços-digitais, amarelo, estável;
- falha-de-sistema (disponibilidade), médio; infraestrutura-digital, amarelo, melhorando.

2.1.2 Mineração de texto e PLN

A mineração de dados, ou *data mining* é uma parte fundamental da Análise de Dados e uma das disciplinas mais importante no campo da Ciência de Dados. Nela, técnicas analíticas são utilizadas para encontrar informações úteis em conjunto de dados para as organizações.

A mineração de texto pode ser definida como o processo de descoberta e extração de conhecimento não trivial de textos livres e não estruturados, podendo ter objetivos variados como a obtenção de informações específicas, análise de sentimentos, classificação ou agrupamentos de documentos (KAO; POTEET, 2007). Ao contrário da mineração de dados tradicional, que parte do princípio de que as informações estão armazenadas de forma estruturada, a mineração de texto possui um grande esforço de pré-processamento focado na extração de características pelo uso do Processamento de Linguagem Natural (PLN). (SOUSA, 2020).

O PLN é o processo de se extrair representações estruturadas significativas de textos livres e não estruturados por meio da análise de linguagem natural. Tipicamente, o PLN utiliza conceitos linguísticos como classes gramaticais, estrutura gramatical e conjunto lexical para lidar com conceitos de difícil abstração para máquinas, como anáforas e ambiguidades (KAO; POTEET, 2007).

Dois conceitos importantes no contexto da mineração de texto são **documentos** e **coleções de documentos**. Um documento é uma unidade básica de informação textual que se deseja analisar, por exemplo, uma página web, um artigo científico ou, como presente neste trabalho, um *tweet*. Coleções de documentos é um conjunto de documentos dos quais se deseja extrair algum padrão ou relação.

2.1.3 API

Uma *API - Application Programing Interface* é um conjunto de regras e protocolos que permitem que diferentes *softwares* se comuniquem e interajam entre si. Em outras palavras, é uma interface que define como os componentes de software devem interagir e se comunicar.

Geralmente, uma *API* é um componente de software que permite o intercambio de dados e funcionalidades entre diferentes sistemas. Ela define os métodos, a formatação dos dados e as regras que os desenvolvedores devem seguir para interagir com determinado serviço.

Há vários tipos de *APIs*, tais como as de serviços *Web*, que permitem que aplicativos acessem recursos via protocolos como o *HTTP - Hypertext Transfer Protocol*; as *APIs* de sistemas operacionais, que possibilitam aplicativos utilitários acessarem recursos do sistema, como gerenciamento de arquivos e acesso a dispositivos e as *APIs* de bibliotecas, que fornecem funções e classes específicas para serem utilizadas em um determinado ambiente de programação. Estas últimas serão empregadas no presente trabalho para realizar a coleta dos dados.

As APIs desempenham um papel fundamental na integração de sistemas e na criação de aplicativos e serviços mais poderosos e versáteis. Por meio delas desenvolvedores integram novos componentes de aplicações a uma arquitetura preexistente permitindo que profissionais utilizem funcionalidades e recursos de outros softwares de forma fácil e padronizada (HAT, 2022).

2.1.4 Web Scraping

O *scraping* é um processo de extrair dados de *websites* de forma automatizada, percorrendo o código-fonte de uma página *web* e analisando a sua estrutura *HTML - HyperText Markup Language*. Nesta análise, é identificado os elementos relevantes, como texto, imagens, links, dentre outros (KHDER, 2021).

Tal processo pode ser utilizado para uma variedade de finalidades, como coletar dados para análise de mercado, monitorar preços e informações de produto em lojas online, extrair dados de redes sociais, acompanhar notícias e atualizações de sites.

Existem várias técnicas e ferramentas disponíveis para realizar o *scraping*. Alguns *softwares* são desenvolvidos especificamente para este propósito, enquanto outros utilizam bibliotecas ou *frameworks* que facilitam a coleta de dados. Neste trabalho, portanto, será utilizada uma biblioteca que realiza este procedimento, facilitando a extração de dados referentes às postagens de usuários do *Twitter*.

2.2 Ferramentas

Nessa seção, será apresentado algumas ferramentas, como *Python*, *Pandas*, *Snscrape*, *Seaborn* e *MongoDB* utilizadas na metodologia proposta.

2.2.1 Python

Python é uma linguagem de programação de alto nível, interpretada e de propósito geral. Ela foi criada por Guido van Rossum e lançada pela primeira vez em 1991. A linguagem se destaca por sua sintaxe clara e legível, o que facilita a leitura e o entendimento do código. Além disso, ela é de código aberto e multiplataforma, ou seja, pode-se executá-la em diversos sistemas operacionais, como *Windows*, *macOS* e *Linux* (FOUDANTION, 2023).

A ênfase da ferramenta é proporcionar legibilidade ao código, encorajando a escrita de programas com uma estrutura clara e concisa. Um exemplo disso é a indentação (recuo) para delimitar blocos de código de mesmo escopo, ao contrário de outras linguagens que utilizam chaves ou palavras-chave especiais.

Python se apresenta como uma linguagem de programação muito versátil, com uma variedade de usos. Entretanto, as suas principais aplicações estão relacionadas com a análise de dados, aprendizado de máquina e automação de tarefas. Ademais, ela possui uma vasta biblioteca padrão, o que oferece uma série de módulos e funções prontas para uso, além de contar com uma extensa comunidade de desenvolvedores. Portanto, apresenta-se como uma linguagem ideal para o escopo deste trabalho.

2.2.2 Twitter API e Snscrape

Criado em 2003 por Jack Dorsey, Noah Glass, Biz Stone e Evan Williams, o Twitter é uma rede social gratuita que funciona como um serviço de *microblogging*. Conforme o *website* oficial da empresa, a plataforma é um serviço para amigos, familiares e colegas de trabalho se comunicarem e se manterem conectados a partir da troca rápida e frequente de mensagens (TWITTER, 2023).

As postagens dos usuários são conhecidas como *Tweets*, nos quais podem conter fotos, vídeos, links e textos de no máximo 280 caracteres. Tais mensagens são postadas em seu perfil e enviadas para seus seguidores, podendo ser pesquisáveis por outros usuários.

A empresa disponibiliza uma API para que desenvolvedores e interessados possam interagir com o sistema. Nela contêm diversos *endpoints* que permitem o atendimento a solicitações diversas. Neste contexto o *scraper* para serviços de redes sociais *snscrape* faz a utilização de tal API, facilitando o trabalho de busca por postagens (ARCHIVIST, 2022). A biblioteca permite a inserção de uma *string* que servirá como modelo de busca, isto

Figura 1 – Exemplo de um *tweet* feito por um usuário

é, dentre um determinado período especificado todos os *tweets* públicos que conterem o texto da consulta serão selecionados.

Para coletar a postagem da Figura 1, poder-se-ia especificar uma *string* de consulta como: "freedom"ou "died for freedom", e determinar um período entre 25/05/2023 a 30/05/2023. Como a postagem está pública para todos os usuários, foi realizada no dia 29 de maio e contém a *string* especificada, ela seria coletada com sucesso. Mais detalhes do funcionamento da biblioteca serão tratados no Capítulo 3.

2.2.3 Pandas

O *pandas* é uma das mais populares bibliotecas de análise de dados em *Python*. Ela fornece estruturas de dados de alto desempenho e fáceis de usar, bem como ferramentas para manipulação e análise de dados (MCKINNEY, 2010).

O pandas é amplamente utilizado em ciência de dados, análise financeira, pesquisa acadêmica e muitas outras áreas onde a manipulação e análise de dados são essenciais. Ele oferece duas estruturas principais de dados: *Series* e *DataFrames*.

- Series: estrutura de dados unidimensional semelhante a um *array* ou uma coluna em uma planilha. Cada elemento em uma Series possui um rótulo, chamado índice. Tal estrutura é bastante interessante para se trabalhar com conjunto de dados unidimensionais;
- DataFrame: estrutura de dados bidimensional semelhante a uma tabela ou planilha. Ele organiza os dados em linhas e colunas, onde cada coluna pode conter um tipo de dado diferente. Tal estrutura fornece uma maneira poderosa de se trabalhar com dados tabulares.

Além dessas estruturas de dados, o *pandas* oferece uma ampla gama de funcionalidades para manipulação, limpeza, transformação e análise de dados. Ele permite que você carregue dados de várias fontes, execute operações estatísticas, faça agregações, filtre e selecione dados. Em resumo, apresenta-se como uma ótima opção de ferramenta para se trabalhar com a análise de dados em *Python* de forma eficiente e intuitiva.

2.2.4 Matplotlib e seaborn

Matplotlib e *Seaborn* são bibliotecas populares para visualização de dados em *Python*. Ambas são amplamente utilizadas na análise de dados, visualização e criação de gráficos.

Matplotlib é uma ferramenta mais antiga e estabelecida em seu ecossistema. Ela fornece uma ampla gama de recursos para a criação de gráficos estáticos, incluindo os gráficos de linhas, de dispersão, de barras, de pizza, histogramas, entre outros. Além disso, ela permite a personalização de todos os aspectos dos gráficos, como títulos, rótulos dos eixos, estilos, cores, legendas e anotações (HUNTER, 2007).

Seaborn é uma ferramenta baseada em *Matplolib*. Ela é projetada para trabalhar em conjunto com outras bibliotecas de análise de dados, como o *Pandas*, e simplificar o processo de criação de gráficos estatísticos atraentes. Ademais, a biblioteca fornece uma *API* de alto nível para elaboração de inúmeros gráficos que possuem um estilo visual agradável e relativamente fácil de serem codificados (WASKOM, 2021).

Portanto, a primeira se mostra como uma biblioteca poderosa e flexível enquanto a segunda é mais conveniente e menos verbosa. Neste trabalho, ambas serão utilizadas em conjunto para criar visualizações eficazes de dados com poucas linhas de código.

2.2.5 Banco de Dados NoSQL e MongoDB

Conforme descrição da ORACLE (2022), um banco de dados é uma coleção organizada de informações, ou dados estruturados, controlados geralmente por um Sistema de Gerenciamento de Banco de Dados (SGBD).

O objetivo principal de um banco de dados é fornecer uma solução para o gerenciamento de informações, possibilitando a coleta, armazenamento, manipulação e recuperação de dados de forma eficaz e segura. Destarte, sua utilização reduz ou até mesmo elimina a necessidade de armazenar dados em arquivos independentes, mitigando a redundância, inconsistência e dificuldade de acesso aos dados. Além disso, os bancos de dados fornecem mecanismos de segurança para proteger os dados contra acesso não autorizado e perda de informações.

Esses sistemas são baseados em modelos, nos quais definem a estrutura e como os dados são organizados e representandos. Nesse contexto, o **modelo relacional** se apresenta como o mais comum. Nele, utiliza-se **tabelas**, **relacionamentos** e **consultas** baseadas em linguagens como o SQL (*Structured Query Language*) para manipular e extrair informações.

Por outro lado, os **banco de dados não relacionais** (NoSQL - *not only SQL*) foram projetados para atender a requisitos específicos de armazenamento e recuperação

de dados em aplicações modernas que exigem escalabilidade horizontal, flexibilidade de esquema e alto desempenho. Tais requisitos são atendidos mediante o relaxamento de algumas restrições de consistência de dados. Dentre os tipos mais comuns estão:

- **Chave-valor:** esses bancos de dados armazenam dados como pares de chave-valor simples. Eles são altamente escaláveis e eficientes para operações de leitura e gravação rápidas;
- **Documento:** esses bancos de dados armazenam dados em formato de documentos semiestruturados, geralmente usando o formato JSON (JavaScript Object Notation) ou BSON (Binary JSON);
- **Grafo:** esses bancos de dados são projetados para armazenar e consultar dados altamente conectados, como redes sociais ou sistemas de recomendação. Eles permitem modelar relacionamentos complexos entre os dados;
- **Colunar:** esses bancos de dados armazenam dados em formato de colunas, o que é especialmente útil para consultas analíticas e agregações.

Nesse contexto, o presente trabalho optou por utilizar banco de dados não relacionais, pois eles permitem armazenar documentos com diferentes estruturas o que possibilita uma flexibilidade maior quando se tem um contexto de dados com diferentes campos e atributos, o que pode ocorrer neste ecossistema de coleta de dados de redes sociais. Além disso, foi usado especificamente o banco **MongoDB**.

O *Mongo* caracteriza-se como um sistema de gerenciamento de banco de dados, orientado a documentos, desenvolvido para lidar com a necessidade crescente de armazenamento e processamento de grandes volumes de dados não estruturados. Nele os dados são organizadas em *collections*, as quais são estruturas semelhantes às tabelas, porém sem possuir um esquema fixo. Cada registro no banco é conhecido como *document*, estruturas BSON (Binary JSON) que podem variar em termos de campos e estrutura (MongoDB, Inc., 2022b).

A ferramenta possui várias vantagens como a alta disponibilidade, escalabilidade horizontal, replicação automática e balanceamento de carga. Ademais, suporta consultas complexas, índices e recursos de agregação para análise de dados.

O Mongo possui uma interface de linha de comando, bem como *drivers* e bibliotecas para várias linguagens de programação, o que facilita a integração com diferentes aplicativos e ecossistemas de desenvolvimento. Entretanto, para hospedar e gerenciar o banco criado pelo presente trabalho será utilizado o **MongoDB Atlas**. Uma solução em nuvem fornecida pela MongoDB Inc, que simplifica o processo de implantação, escalabilidade e gerenciamento do banco de dados (MongoDB, Inc., 2022a) .

A ferramenta permite a criação e configuração e dimensionamento simplificado de *clusters* fornecendo uma interface gráfica intuitiva e amigável. Além disso, inclui compatibilidade com vários provedores de nuvem, tais como *AWS (Amazon Web Services), Azure (Microsoft Azure) e GCP (Google Cloud Plataform)*. Mais detalhes sobre a utilização tando do Mongo quanto da plataforma MongoDB Altas serão tratadas no Capítulo 3.

2.3 Trabalhos Correlatos

Essa seção da fundamentação teórica irá apresentar trabalhos que tiveram como objetivo a criação de bases de dados e análise de eventos relacionados a segurança cibernética, bem como a utilização do Twitter para análise de comportamentos de usuários e predição de incidentes de segurança. Ademais, salientar as suas deficiências em comum, ressaltando a contribuição que este trabalho quer fornecer para a problemática apontada.

Paulo (2021) utiliza conceitos da análise de dados sobre a mesma base pública sobre incidentes de segurança que o presente trabalho utilizou, a *Hackmageddon*. Em sua monografia, os dados foram transformados em cinco séries temporais, uma para todos os incidentes e quatro para cada tipo de incidente observado, a saber: crime cibernético, espionagem cibernética, guerra cibernética e *Hacktivismo*. A partir destas, foram estimados modelos *ARIMA*(Modelo Auto-Regressivo Integrado de Médias Móveis) para que se atingisse o objetivo de realizar previsões sobre tais incidentes.

Sousa (2020) em sua dissertação de mestrado utilizou métodos de aprendizado de máquina para obter informações sobre vulnerabilidades já divulgadas, procurando identificar quais delas possuem maior probabilidade de serem exploradas. Especificamente, desejou-se extrair conhecimento de discussões realizadas no *Twitter* sobre o assunto a fim de superar o desempenho de classificadores que se apoiam apenas em informações sobre a severidade de uma vulnerabilidade. Da mesma forma que o presente trabalho, (SOUSA, 2020) disponibiliza um conjunto de dados brutos e pré-processados de rótulos e do *Twitter*, à comunidade para futuros trabalhos, focando-se especialmente em *exploits* e vulnerabilidades, diferentemente deste trabalho que pretende observar incidentes de segurança, especificamente vazamento de dados.

Bose et al. (2019) apresentam uma abordagem de aprendizado de máquina não supervisionado para extrair informações textuais do Twitter com o intuito de detectar potenciais ameaças cibernéticas. Tais eventos foram catalogados como *novel* (inexistente até o momento) e *developing* (marcado com similaridade com outro evento previamente detectado). O artigo também provê um método para classificá-los conforme a sua importância baseados, dentre outros fatores, em palavras-chave e na influência do usuário que postou a mensagem.

Sceller et al. (2017) propõem o SONAR: um *framework* que detecta, geolocalização, categoriza e monitora eventos de ciber segurança a partir de dados minerados do Twitter quase em tempo real. Além disso, a ferramenta possui a capacidade de autoaprendizado automático, a qual descobre novas terminologias utilizadas no contexto da área de segurança da informação para formar novas palavras-chave que posteriormente serão utilizadas para encontrar novos eventos. O software pode prover a profissionais uma visão global do que está atualmente acontecendo no cenário de segurança da informação. Além disso, para a área forense pode proporcionar uma observação cronológica de ameaças cibernéticas presentes em redes sociais e permitir a identificação de suspeitos. Ademais, pode ajudar times de segurança a protegerem suas organizações de vulnerabilidades recentemente descobertas.

Altalhi e Gutub (2020) revisam e comparam os trabalhos relevantes que fazem uso de dados do *Twitter* para extrair informação sobre ataques cibernéticos atuais e eminentes. As comparações da pesquisa são baseadas em seis diferentes fatores de efetividade, tais como: escopo de detecção, técnica de extração, complexidade de algoritmos, sumarização da informação, escalabilidade sobre o tempo e medidas de desempenho, sempre tendo em vista a análise de ganho nas contribuições de predição. Tais trabalhos são: SYNAPSE (ALVES et al., 2019), DeepNN (DIONÍSIO et al., 2019), DataFreq (RODRIGUEZ; OKAMURA, 2019), CyerberTwitter (MITTAL et al., 2016), Text-mining (SAPIENZA et al., 2017) e SONAR (SCELLER et al., 2017). Embora, nenhum dos estudos tenha atingido a melhor pontuação em todos os fatores observados, a pesquisa conclui que, dentre os trabalhos o SYNAPSE foi o que trouxe a melhor média das características. Além disso, a pesquisa propõem múltiplas melhorias para aumentar a eficácia desses trabalhos que mineram informações sobre segurança da informação utilizando redes sociais como o *Twitter*.

Destarte, é evidente a relevância das redes sociais para a investigação de incidentes de segurança. Contudo, os trabalhos investigados não fornecem detalhes de implementação nem sequer as ferramentas utilizadas para coletar, tratar e armazenar os dados. Tal cenário pode desencorajar novos pesquisadores que estejam à procura de um método para realizar estas tarefas. Desse modo, o presente trabalho tem o intuito de disponibilizar um processo para elucidar os passos de se trabalhar com os dados.

3 Desenvolvimento

Esse capítulo será dedicado a enunciar uma visão geral sobre o processo sugerido e as etapas tomadas para o desenvolvimento do trabalho em um estudo de caso.

Cnforme citado no capítulo 3

A Figura 2 ilustra de forma simplificada as etapas da metodologia proposta por este trabalho.

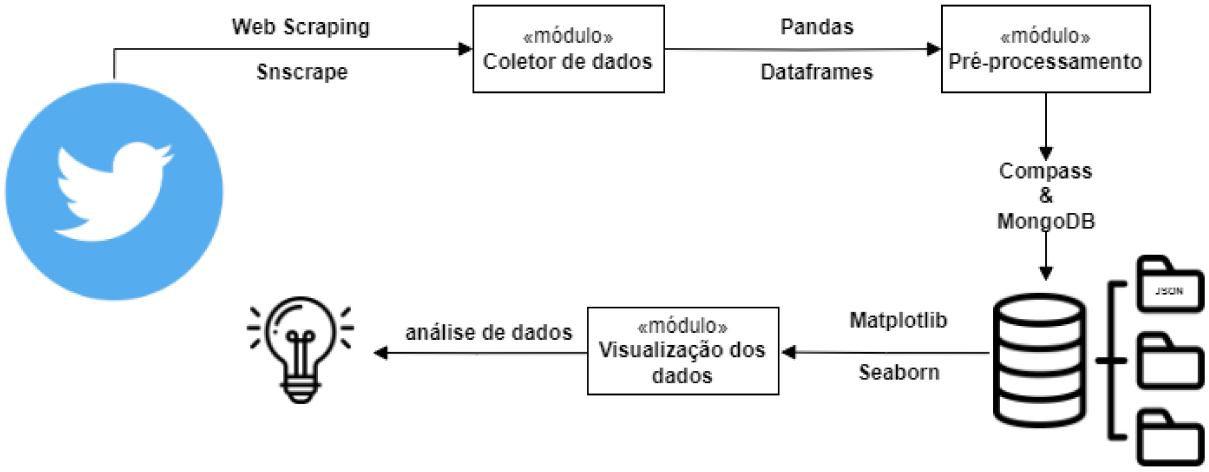

Figura 2 – Metodologia proposta

O trabalho foi dividido e realizado nas seguintes etapas:

- **Seção 3.1** Investigação de estudos de caso - descreve como foi selecionado o estudo de caso observado no presente trabalho.
- **Seção 3.2** Coleta e pré-processamento do dados - descreve como os dados foram extraídos e quais operações foram realizadas antes da persistência;
- **Seção 3.3** Persistência dos dados - descreve o procedimento para realizar a persistência das amostas em um banco de dados;
- **Seção 3.4** Análise dos dados - descreve quais análises foram realizadas;
- **Capítulo 4** Visualização Gráfica e resultados - após todas as etapas, detalhe-se como foi feita a elaboração dos gráficos e o resultado que se obteve.
- **Seção 4.1** Apresentação dos entregáveis - códigos e *scripts* criados neste trabalho com o link para o seu repositório no *GitHub*, bem como o acesso à base de dados criada.

3.1 Investigação de estudos de caso

Para a escolha de um candidato para o estudo de caso foi utilizada o projeto *Hackmageddon* de Paolo Passeri (PASSERI, 2023), especialista em *cloud security* com vinte anos de experiência na área de segurança da informação. Neste trabalho, o autor coleta a cada duas semanas todas os incidentes de segurança mais relevantes que ocorreram e os compila em categorias distintas.

Uma destas categorias são as violações de dados ou vazamento de dados, nas quais dados sensíveis, protegidos ou confidenciais são copiados, transmitidos, vistos, roubados ou utilizados por um indivíduo não autorizado. O presente trabalho se interessou em investigar especificamente esses tipos de incidentes.

O projeto de Passeri reuniu os principais vazamentos ocorridos no ano de 2021 e para realizar a seleção de um incidente levou-se em consideração os requisitos abaixo:

- Evento de escala global no qual um expressivo número de usuários foram afetados;
- Vazamento de dados de usuários ou clientes de grandes empresas;
- Incidente no qual ocorreu o engajamento dos usuários do *Twitter*.

Por conseguinte, o incidente escolhido foi o vazamento de dados de 533 milhões de usuários da rede social *Facebook* conforme mostra a Figura 3. Embora a empresa não tenha identificado a data precisa do ocorrido, em 03/04/2021, tais dados foram publicados em um fórum de acesso público de crimes cibernéticos.

De acordo com algumas amostras verificadas pelo portal de notícias *The Record*, além do número de celular, o vazamento incluiu dados de perfis, tais como nome, gênero, e-mail, localização geográfica, dados relacionados ao trabalho dos usuários, dentre outros (CIMPANU, 2021).

Figura 3 – Manchete da reportagem do The Record

Em suma, estima-se que mais de 8 milhões de brasileiros foram afetados e usuários de mais 105 países foram envolvidos. Além disso, cabe ressaltar que até o número de celular de um dos fundadores da plataforma, *Mark Zuckerberg* estava entre os dados vazados. Em entrevista, o *Facebook* confirmou o incidente e afirmou que o mesmo ocorrera em 2019 (ABRAMS, 2021). Nesse ano, o invasor abusou de uma vulnerabilidade de um recurso de importação de contatos que permitiu a correspondência de números de telefone aleatórios a contas de usuários específicas.

Desde então, acredita-se que tais dados estavam sendo vendidos online, bem como foram utilizados para a criação de um *Bot* do *Telegram* lançado em janeiro de 2021 que permitia que qualquer um retornasse o número de celular e detalhes da conta de usuários da rede social por uma pequena (CIMPANU, 2021).

Em 03/04/2021, tais dados entraram em domínio público, possibilitando que cibercriminosos menos experientes utilizassem-no para, por exemplo, a criação de *Spams* de *emails*, tentativas de extorsão, ameaças, assédio e demais crimes contra os usuários da plataforma.

3.2 Coleta de dados

Para a coleta de dados da plataforma *Twitter*, foi utilizado a ferramenta *snscrape* (ARCHIVIST, 2022), que permite extrair *tweets* entre determinado período. Embora a empresa disponibilize a própria *API*, não foi possível consumir os *endpoints* diretamente, pois quando o trabalho foi desenvolvido a empresa já não mais permitia que novos desenvolvedores as acessassem gratuitamente.

Além disso, fora utilizadas as estruturas de *DataFrames* da biblioteca **pandas** para facilitar manipulações nos campos dos *tweets* e operações simples de pré-processamento envolvendo principalmente as datas das postagens dos usuários. Ademais, tais estruturas foram essenciais para facilitar que as amostras fossem salvas temporariamente em arquivos do tipo *JSON* para que posteriormente pudessem ser transferidas para o banco de dados.

Toda a lógica envolvendo esta etapa foi divida em dois arquivos distintos: **main.py** e **querybuilder.py**. O primeiro *script*, conforme mostra a Listagem 3.1, é responsável por fazer uma interação com o programador via CLI - Command Line Interface, recebendo dados importantes para realizar o *scrapping* das postagens.

```
1  import snscrape.modules.twitter as sntwitter
2  import pandas as pd
3  import query_builder as qb
4
5  tweets_list = []
6
7  text = input("Consulta a ser realizada: ")
8  since = input("Data de inicio das buscas: ") #formato YYYY-MM-DD
9  until = input("Data de encerramento das buscas: ")
10 count = int(input("Quantos tweets voce deseja buscar: "))
11
12 file_name = f"tweets-facebook/uma-semana/{text}_{since}_{until}.json"
13
14 query = qb.query_builder(text, since, until)
```

Listing 3.1 – Arquivo main.py Parte 1

Após a importação dos pacotes declararam-se algumas variáveis. Segue abaixo uma breve descrição de cada uma:

- **tweets_list**: lista responsável por armazenar em memória os objetos *Tweet* retornados pelo *scraper*.
- **text**: *string* de busca para selecionar os *tweets*.
- **since**: data de início das buscas no formato *YYY-MM-DD*.
- **unitl**: data final das buscas no formato *YYY-MM-DD*.
- **count**: quantidade de *tweets* que se deseja retornar.
- **file_name**: o caminho do diretório para armazenar os arquivos e padrão que foi utilizado para a nomeação.
- **query**: recebe o retorno da função presente dentro do arquivo *query_builder.py*, no qual construí a consulta que será utilizada pelo *snscrape* para selecionar os *tweets*.

Conforme apresentado na Listagem 3.2, a função *query_builder* foi utilizada para concatenar na ordem correta os termos da busca e inserir os operadores de *since* e *until*, bem como verificar se as datas de início e término das buscas foram digitadas e como tratá-las caso estiverem vazias.

```
1  import datetime
2
3  def query_builder(text, since, until):
4
5      query = text
6
7      if since == ' ':
8          since = datetime.datetime.strftime(datetime,
9                                             datetime.strptime(until,
10                                            "%Y-%m-%d") -
11                                            datetime.time.delta(days=7),
12                                            "%Y-%m-%d")
13     query += f" since:{since}"
14
15     if until == ' ':
16         until = datetime.datetime.strftime(datetime.date.today(),
17                                            '%Y-%m-%d')
18     query += f" until:{until}"
19
20     return query
```

Listing 3.2 – Arquivo query_builder.py

Cabe ressaltar que após efetuar inúmeros testes com o *scraper* ficou evidente que consultas menores e que utilizassem termos não técnicos e mais genéricos resultavam em uma revocação maior. Como a princípio o foco do presente trabalho era a realização de um sucinta análise dessas postagens, optou-se por essa estratégia. Entretanto, é fundamental mencionar que consultas com termos específicos sobre segurança da informação ou mesmo termos técnicos como *data breach* conseguem gerar resultados mais precisos, porém bem menos numerosos.

Desse modo, foi manuseada a consulta mais simples possível que pudesse representar o incidente com menos termos e que usuários comuns estariam utilizando para se referir ao incidente. Portanto, a consulta: *Facebook data leak* foi escolhida como o parâmetro para as buscas.

```
for i,tweet in enumerate(
    sntwitter.TwitterSearchScraper(query).get_items()):
    if i>= count:
        break
    tweet.date = tweet.date.strftime("%d/%m/%Y")
    tweets_list.append([tweet.date, tweet.id, tweet.rawContent,
                        tweet.user.username, tweet.url])
```

Listing 3.3 – Arquivo main.py Parte 2

Posteriormente, conforme a Listagem 3.3, criou-se uma estrutura de repetição para iterarmos *count* vezes sobre os objetos retornados pela consulta, por meio da expressão *for in* e da função *built-in enumerate*. Dessa forma, foi utilizado o objeto *TwitterSearchScraper* para se conectar a *API* do *Twitter* realizando a busca por postagens que possua a consulta especificada como parâmetro de seu método construtor.

Além disso, o método *get_items()* é responsável por retornar os objetos *Tweet* que armazenam em seus atributos os dados referentes as postagens, tais como: o conteúdo, quantidade de curtidas, *retweets* e repostas que a postagem teve, bem como outros campos relevantes.

Ademais, modificou-se a formatação das datas por meio da função *strftime()* nativas de objetos do tipo *Date* para o padrão utilizado em território brasileiro. Por último, acrescentou-se os valores dos campos **data**, **id**, **conteúdo**, **nome do usuário** e **URL** referentes a postagem na lista *tweets_list*.

```
tweets_df = pd.DataFrame(tweets_list, columns = ["Date",
                                                 "TweetId",
                                                 "TweetContent",
                                                 "Username",
                                                 "URL"])

result = tweets_df.to_json(path_or_buf = file_name,
                           orient = "records",
```

```
9                                   force_ascii = False, l
10                                  ines = True)
```

Listing 3.4 – Arquivo main.py Parte3

Por conseguinte, conforme mostrado na Listagem 3.4, foi criado um **Dataframe** com as colunas representando os valores citados anteriormente. Finalmente, para encerrar a etapa de coleta dos dados, converteu-se essa estrutura para *JSON*, adicionando-a em um arquivo nomeado conforme o valor da variável *file_name*.

Cabe ressaltar, que além da formatação das datas para o padrão brasileiro (dd/mm /yyyy) e a exclusão de alguns campos que não se apresentavam como relevantes para o presente trabalho, não foram realizadas outras operações de preprocessamento ou afins. Portanto, o conteúdo das postagens foram persistidas sem nenhuma alteração.

3.3 Persistência dos dados

Para realizar a persistência dos dados, primeiramente, foi necessário criar um projeto na plataforma do Atlas. Tais projetos permitem que se criem *clusters* de instâncias do banco utilizando diversos provedores de *cloud*.

Após realizar o *deploy* sob a estrutura da *AWS* as demais interações foram efetuadas pela interface gráfica do *MongoDB Compass*. Dessa forma, realizou-se a conexão com o banco por meio de uma *URI - Uniform Resource Identifier*. A Figura 4 descreve tal processo.

Posteriormente, foi criado um banco de dados de nome **Facebook**, bem como as *collections* representando os diferentes períodos, nos quais foi realizado as coletas de dados (antes e depois do incidente - 03/04/2021). A Figura 5 corrobora tal afirmação.

Enfim, para adicionar os dados nas coleções utilizando a *Compass* é foi necessário simplesmente selecionar o arquivo *JSON* que continha os dados coletados na etapa anterior. Com os dados adicionados, a plataforma permite que se visualize, consulte ou efetue operações de agregação nos documentos. A Figura 7 exemplifica uma coleção contendo aproximadamente 2700 postagens.

3.4 Análise dos dados

Cabe salientar que foram efetuadas dois tipos de análise. Por um lado, procurou-se verificar a relevância de *tweets* que ocorreram antes e depois da data do ocorrido. Dessa forma, o conteúdo das postagens foram investigados, na qual se buscou aquelas com mais relevância, segundo a percepção do autor, ademais procurou-se entender o conteúdo de

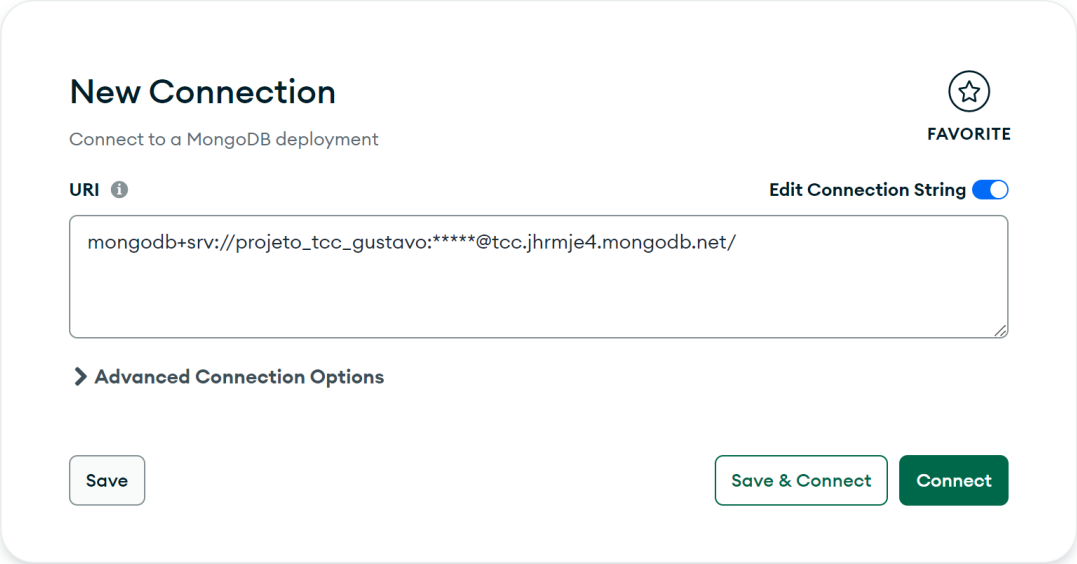

Figura 4 – Realizando a conexão com o MongoDB

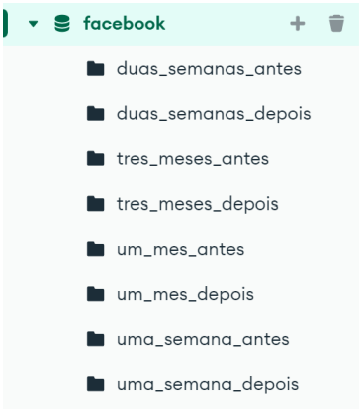

Figura 5 – *Collections* criadas no *MongoDB Compass*

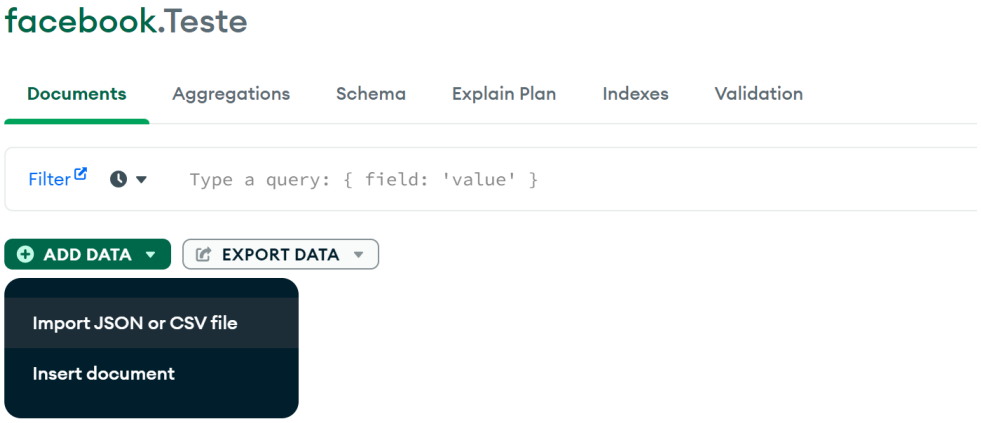

Figura 6 – *Upload* dos dados nas *Collections* utilizando o *MongoDB Compass*

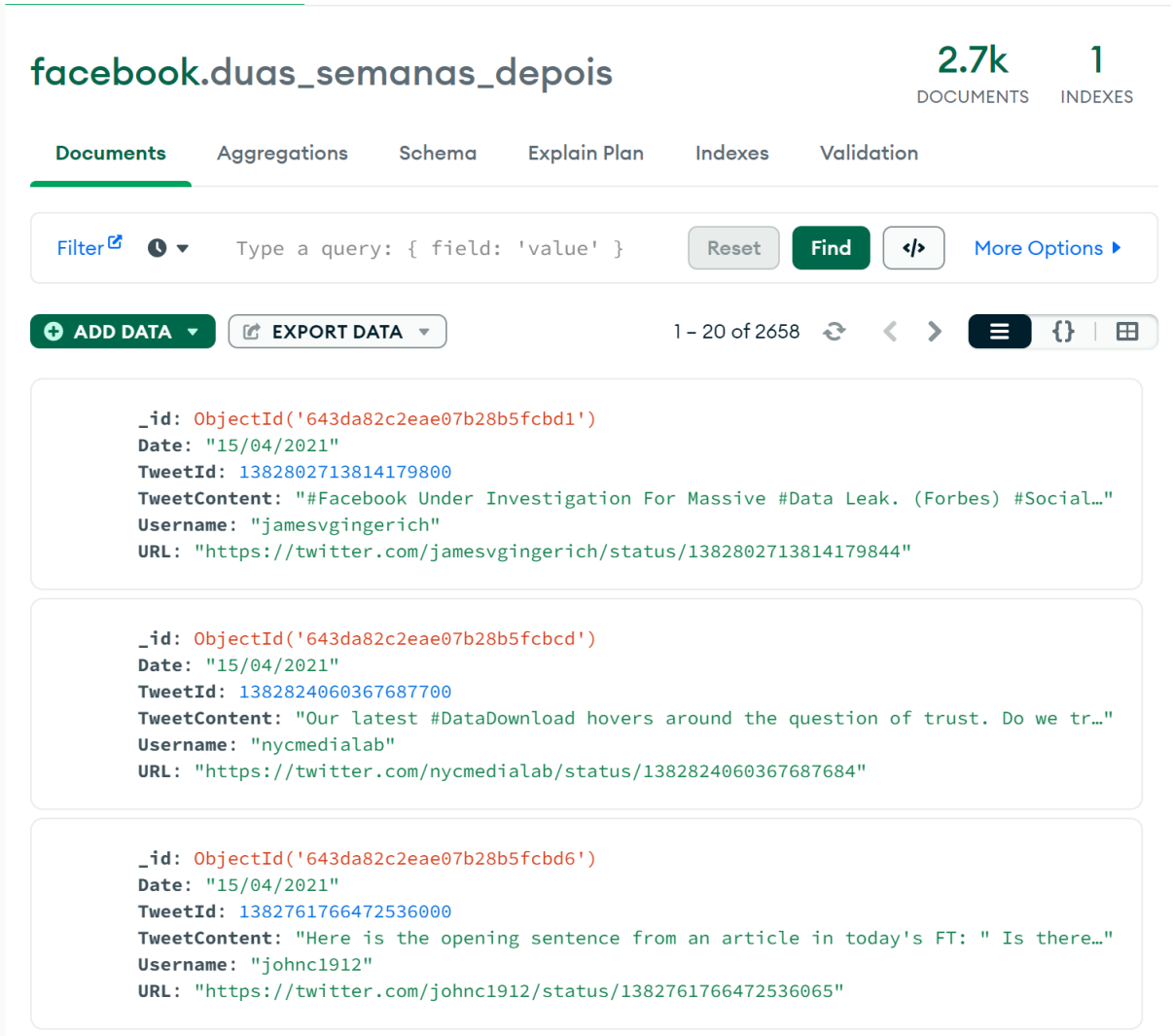

Figura 7 – Exemplos de documentos armazenado em uma coleção

cada uma detalhadamente. Além disso, foi efetuada análises gráficas sobre a quantidade de postagens que ocorram durante todo tempo especificado na etapa de coleta dos dados.

Para realizar a etapa de análise gráfica dos dados foi utilizado as bibliotecas: *seaborn* e *matplotlib*, ambas para facilitar a criação e manipulação de gráficos. Em suma, todos os gráficos observam a quantidade de *tweets* postados em diferentes datas e períodos. Desse modo, foi possível identificar o padrão de comportamento de usuários quando se ocorre um incidente de segurança, ou seja, quanto tempo as discussões ficam em alta, e quando ocorre o maior número de postagens.

O código dos gráficos criados estão dentro do arquivo *graphs.ipynb*, conforme mostrado na Listagem 3.5. Logo em seguida há uma breve explicação sobre o mesmo.

Após realizar as importações das bibliotecas necessárias e configurar a estilização de fundo, os *JSON* foram inseridos num *DataFrame*. Em seguida, assegurou-se que a data estivesse em formato simples para realizar as operações. Dessa forma, foi criado uma

coluna nova na qual foram colocadas o mês de ocorrência da postagem.

Portanto, o eixo x do gráfico representou os três meses antes da data do incidente e o eixo y representou a quantidade de *tweets*. Desse modo um gráfico do tipo **histograma** será desenhado, e suas barras verticais incrementaram conforme a quantidade de postagens para determinado mês. Demais gráficos e suas devidas análises serão realizadas na próxima seção deste trabalho.

```
1
2  import pandas as pd
3  import seaborn as sns
4  import matplotlib.pyplot as plt
5
6  # Tres meses antes do ocorrido:
7  sns.set_theme(style="whitegrid")
8  tweets = pd.read_json("tweets/facebook/tres-meses/facebook data
       leak_2021-01-02_2021-04-02.json",
9                        lines=True, nrows=94, convert_dates=False)
10
11
12 tweets["Date"] = pd.to_datetime(tweets.Date, format="%d/%m/%Y")
13 tweets["Month"] = tweets["Date"].dt.month
14
15
16 f, ax = plt.subplots(figsize=(15, 10))
17
18 sns.countplot(x= tweets["Month"])
19 ax.set_xlabel("Meses", size = 12 )
20 ax.set_ylabel("Quantidade de Tweets" , size = 12 )
21 ax.set_xticklabels(["Janeiro", "Fevereiro", "Marco", "Abril"])
22
23 for p in ax.patches:
24     ax.annotate(int(p.get_height()), (p.get_x()+0.05, p.get_height()+20)
       , fontsize = 12)
```

Listing 3.5 – Exemplo para se trabalhar com as bibliotecas de gráficos

4 Resultados - Estudo de caso

Ao todo, foram coletados 3330 *tweets* públicos referentes às discussões sobre o incidente de segurança da rede social *Facebook* ocorrido em 03/04/2021. Conforme explicitado na Etapa 3.4 (Análise dos dados) foram investigados qualitativamente todas as 94 amostras coletadas entre o dia 06/01 ao dia 01/04 por meio da consulta **Facebook data leak**.

Conforme observado na Figura 8, no mês de janeiro ocorrera aproximadamente 70% das postagens coletadas. Analisando a Figura 9, é possível verificar dois valores máximos, ocorrendo respectivamente nos dias 12 e 27. Isso ocorreu, pois até o dia 14 do referido mês, os usuários estavam discutindo sobre outro incidente envolvendo a mesma empresa, bem como outras redes sociais: *Instagram* e *LinkedIn*. Desse modo, essa situação permitiu que se fosse descoberto outros incidentes de segurança relacionados com essas redes sociais.

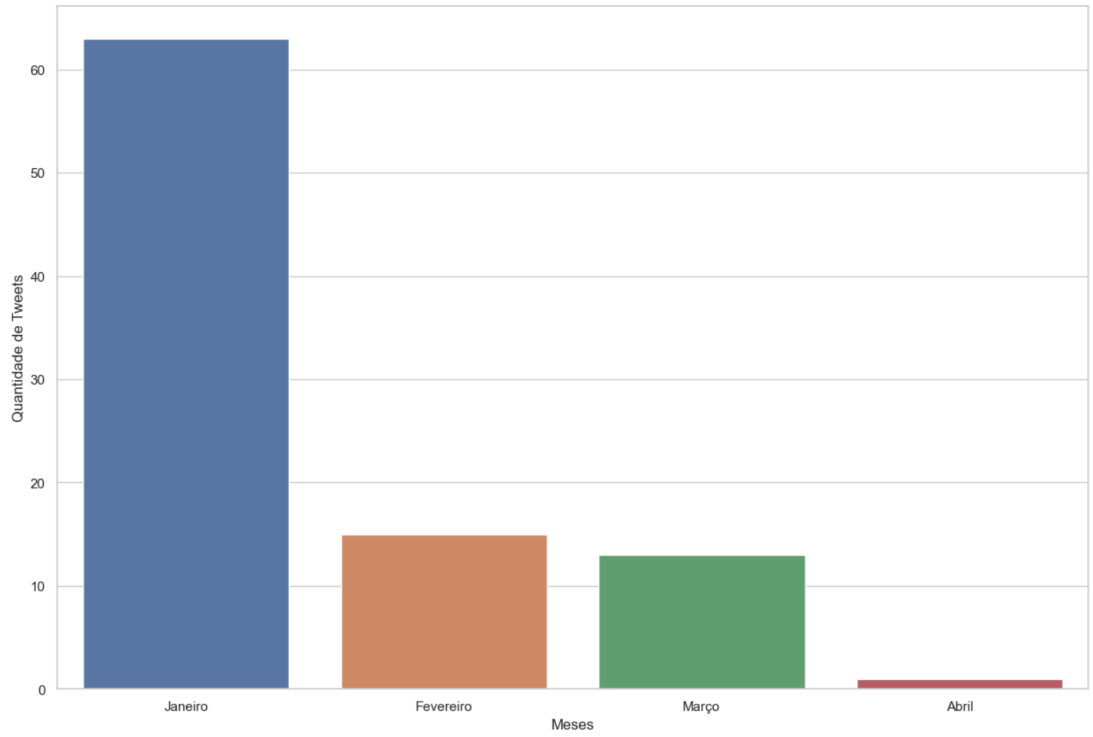

Figura 8 – Quantidade de *tweets* postados nos 3 meses antes da data do incidente

Entretanto, no dia 15, como mostra a Figura 10 ocorreu a primeira postagem realmente relacionada ao estudo de caso. O usuário **nashrafeeg**, que aparentemente teve acesso ao conteúdo do vazamento, comentou e marcou outra postagem que ocorrera no dia anterior, possivelmente com conteúdo sensível, tendo em vista que a própria rede social se assegurou de censurá-lo.

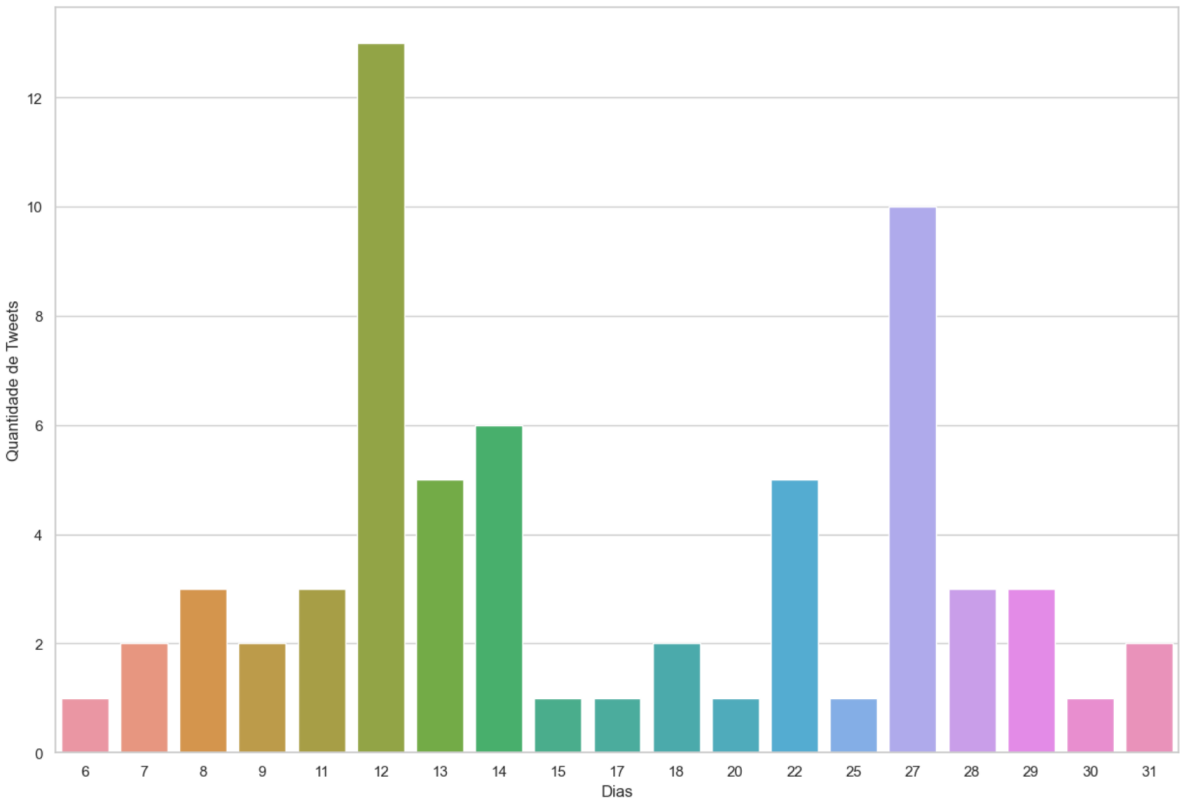

Figura 9 – Quantidade de *tweets* postados por dia no mês de janeiro de 2021

Figura 10 – Primeiras postagens sobre o incidente entre o dia 14 e 15

Possivelmente, a conta foi suspensa e o conteúdo da postagem ocultado, pois o usuário violou as regras de uso da rede social. Desse modo, a Figura 11 ratifica que as primeiras postagens referentes ao incidente ocorreram por volta do dia 14/01, corroborando

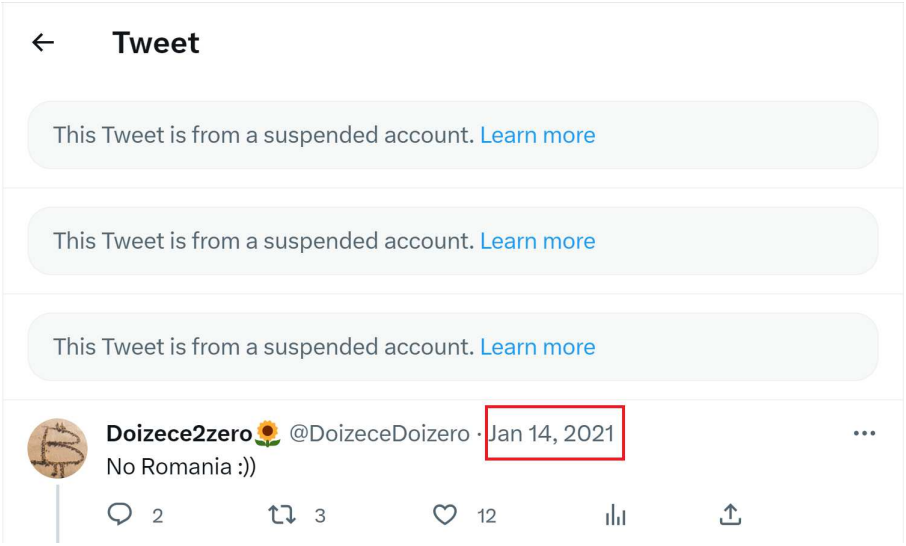

Figura 11 – Tweet censurado do dia 14/01

a informação que o **vazamento já estava sendo compartilhado e vendido** antes do dia 03/04 conforme as informações de Cimpanu (2021).

Além disso, analisando qualitativamente o segundo topo da Figura 9, observou-se que a partir do dia 26 de janeiro, a maioria das postagens continham uma menção ao mesmo *hiperlink* (FRIPP, 2021), no qual redirecionava para uma reportagem feita por uma programa de rádio americano chamado *The Kim Komando Show*. Tal programa apresenta-se como mídia especializada na área de tecnologia, divulgando semanalmente dicas de segurança e as últimas tendências tecnológicas.

Portanto, foi possível extrair informações relevantes sobre o incidente por meio da análise de tais discussões, tais como, o **acesso antecipado ao vazamento**, a descoberta de **usuários que possivelmente tiveram contato com o vazamento** e também a **identificação de uma fonte de informação confiável** a respeito do ocorrido.

Entretanto, cabe salientar que 97% dos *tweets* foram postados após a data do incidente, ou seja, entre o dia 03/04 ao dia 02/07, conforme mostrado na Figura 12. Dentre estes, cerca de 96% foram publicados no mesmo mês de abril. Ao observar a Figura 13, concluiu-se que 2231 *tweets*, aproximadamente 72% foram postados uma semana após a data do incidente. Desse modo, é possível concluir que os usuários realizaram suas discussões sobre um **breve período** de tempo. Neste, as postagens ocorrem de maneira frequente e há um forte engajamento e descontentamento com a empresa responsável.

Primeiramente, cabe salientar que não foi possível descobrir nenhuma informação adicional e relevante nos 129 *tweets* analisados qualitativamente nos meses de maio, junho e julho. Embora tenha-se encontrando amostras relevantes, todas as informações contidas a respeito do incidente já se havia encontrado no mês do evento. Todavia, após analisar algumas amostras referentes ao mês de abril (aproximadamente 20%) foi possível

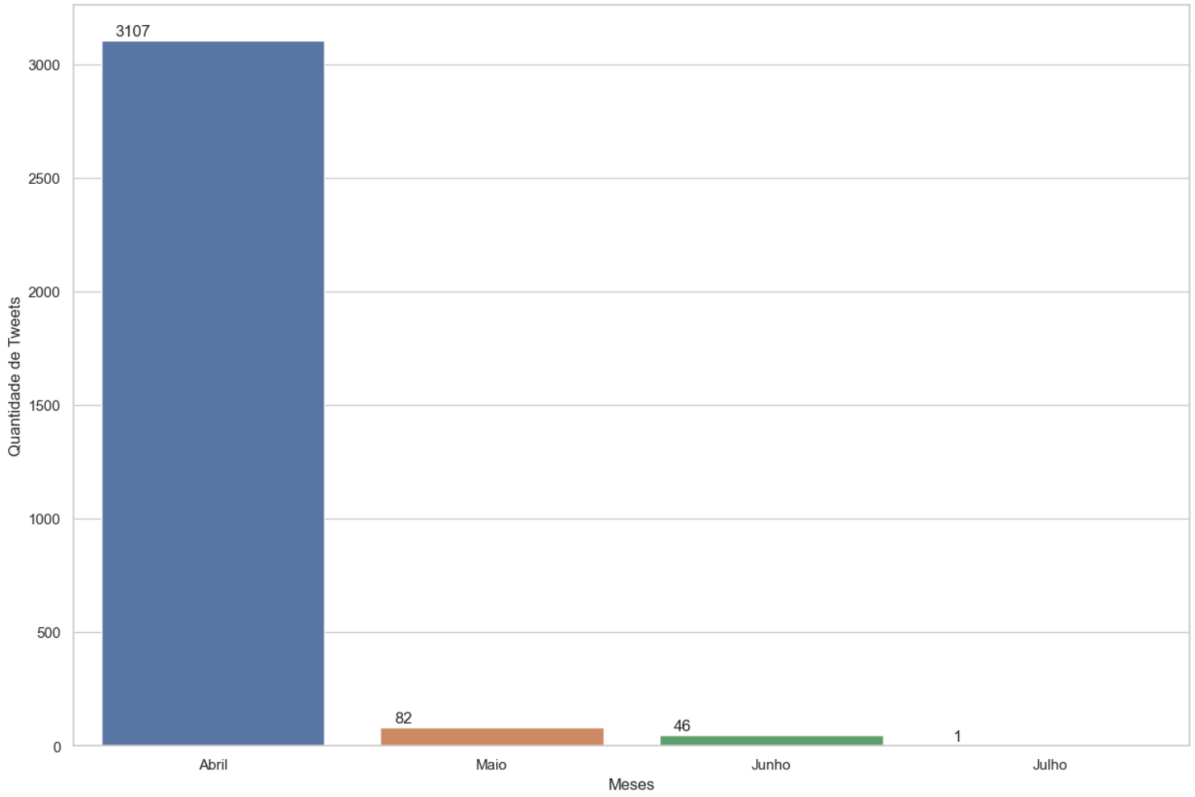

Figura 12 – Quantidade de *tweets* postados depois da data do incidente

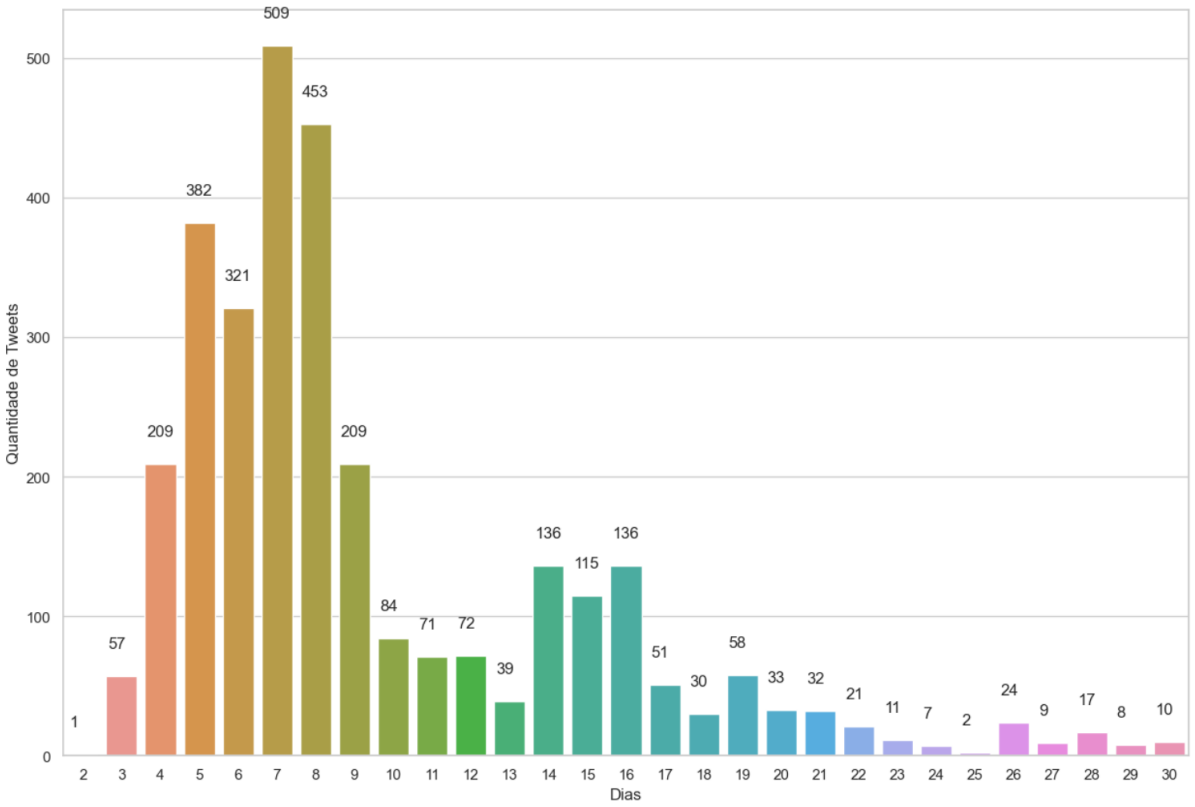

Figura 13 – Quantidade de *tweets* postados no mês de abril

identificar os seguintes pontos relevantes:

- Foi descoberto que Alon Gal, CTO da empresa especializada em cibercrimes, Hudson Rock, foi o primeiro indivíduo a entrar em contato com o vazamento e publicá-lo na rede social, por meio da conta, atualmente banida "*UnderTheBreach*"conforme informações de Holmes (2021);
- Conforme a Figura 14, constatou-se o que a primeira pessoa relacionada à empresa que deu um parecer a respeito do ocorrido foi Liz Bourgeois, Chefe de Comunicações da empresa *Meta*, empresa que é atualmente proprietária do *Facebook*;

Figura 14 – Primeira comunicação por parte da empresa em 03/04

- O usuário Zlatan Vano compartilhou uma *dashboard* no dia 04/04 referente ao incidente muito interessante, que permitiu a visualização da grandeza do impacto gerado pelo vazamento (VANO, 2021).
- Muitos usuários compartilharam *hiperlinks* para ferramentas que verificam se o *e-mail* ou número do usuário estava dentro do vazamento, ou de demais incidentes. Dentre estes, as principais foram a criada por Troy Hunt, have i been pwned? e a the news each day, criada por David Johnstone;
- No dia 06/04, Mike Clark, Diretor da Gestão de Produtos da *Meta* se pronunciou a respeito do ocorrido. Ele revelou que o vazamento ocorreu devido a um *scraping* feito na rede social em setembro de 2019 (CLARK, 2021).

- No mesmo dia, a reportagem de Aaron Shaffer do *The Washington Post* em conjunto com a pesquisadora de cibersegurança Tonya Riley apresentou o posicionamento de autoridades políticas e os desdobramentos legais do incidente, bem como a opinião de diversos especialistas em segurança da informação (SCHAFFER, 2021);

- No dia 08/04, a consultoria de segurança *ReliaQuest* divulgou uma explicação do caso e sua evolução desde 2019. Na postagem do *blog* foi informado que os dados começaram a ser negociados em fóruns em 2020 no valor de 25.000 dólares, conforme mostra a Figura 15. Além disso, foi explicado que os cibercriminosos vão revendendo os dados do vazamento e diminuindo os preços em cada transação e que eventualmente usuários de fóruns da *deep* e *dark web* os compartilham de graça para ganharem reputação e notoriedade nesse ecossistema (RELIAQUEST, 2021). Tal situação é a provável para explicar a divulgação do dia 03/04, na qual o usuário do fórum *RaidForums* compartilhou por um custo insignificante de aproximadamente 2,25 dólares, possibilitando que, em cinco dias, cerca de 4800 membros tivessem contato com os dados conforme mostrado na Figura 16;

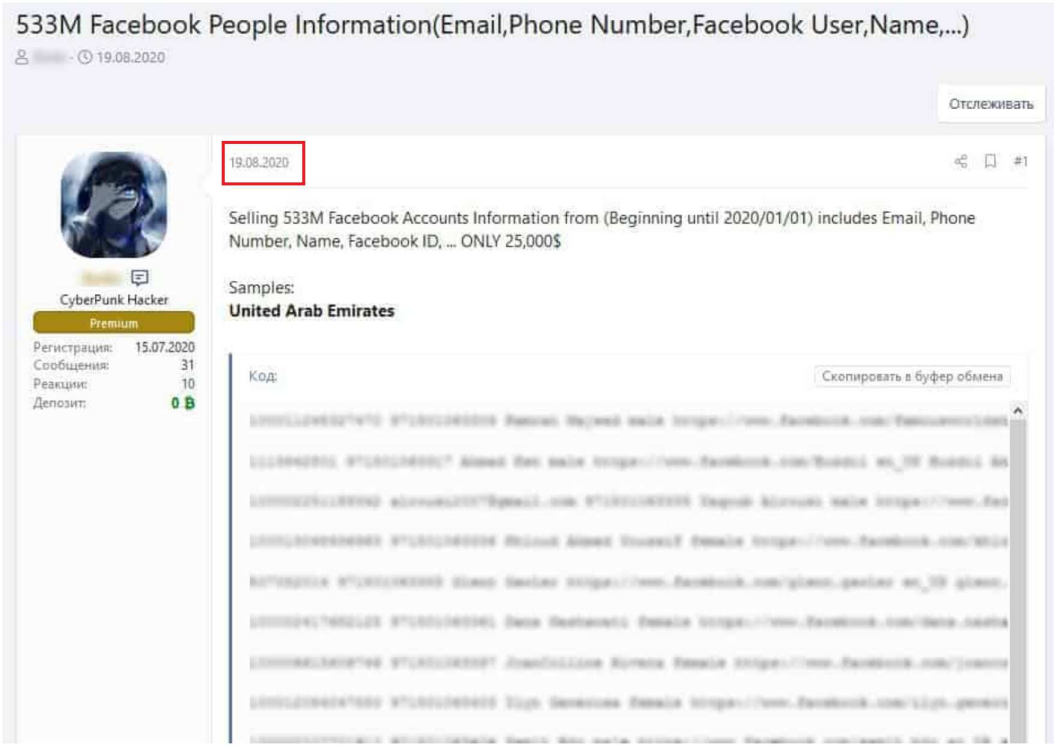

Figura 15 – *Hacker* vendendo os dados do vazamento no fórum XXS

- Entre os dias 14, 15 e 16 surgiram notícias relatando as primeiras medidas judicias contra empresa, o que justifica um leve aumento na quantidade de *tweets* no gráfico da Figura 13. Dentre essas notícias, a da *BBC news* informou que a DRI - *Digital Rights Ireland* estava preparando uma ação contra a empresa em nome dos cidadãos europeus afetados pelo vazamento (BBC, 2021).

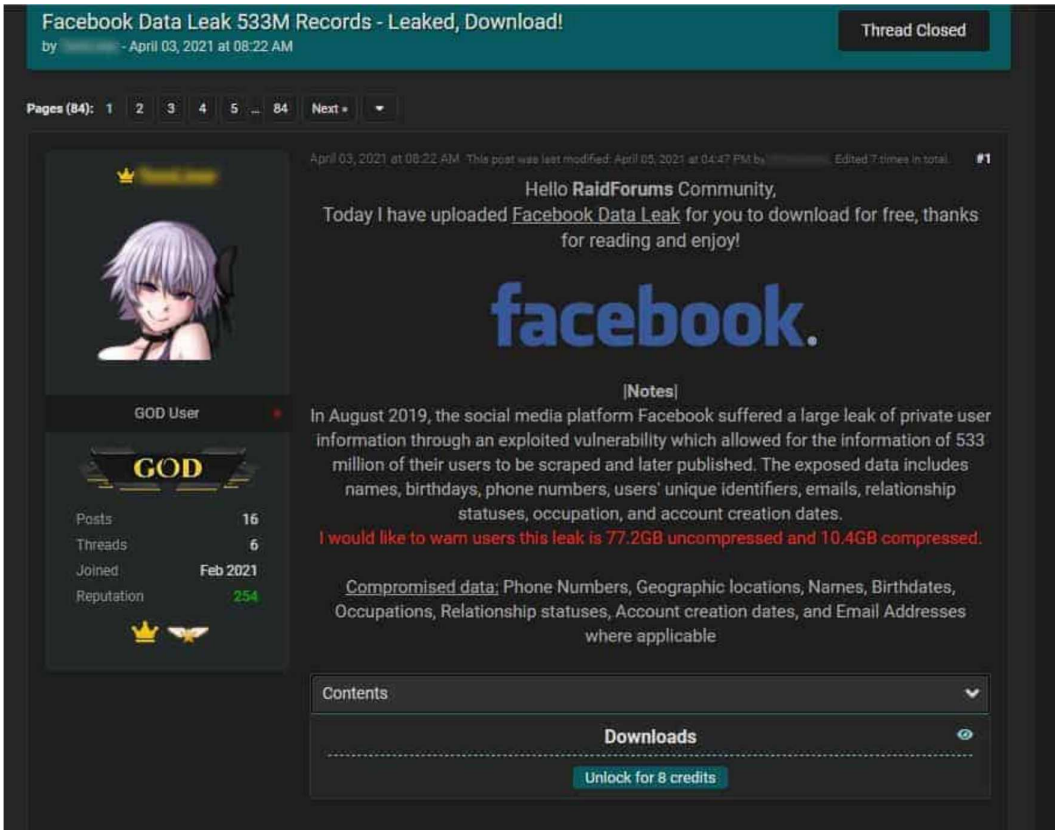

Figura 16 – Postagem que ocasionou todo o escândalo do dia 03/04

Em resumo, é importante salientar que a maioria dos usuários compartilham os mesmos *hiperlinks* para *websites* que noticiaram o incidente. Dentre eles estão: O'Sullivan (2021) da *CNN Business*, o Holmes (2021) da *Business Insider*, o Newman (2021) da *Wired*, o Culliford (2021), da *Reuters* e o Jee (2021) da *MIT Tecnology Review*.

Além disso, o aspecto mais relevante é o fato que a rede social permite a visualização da evolução das notícias sobre o incidente. Portanto, mostrando-se como uma ótima ferramenta para se descobrir informações que se complementam temporalmente, bem como os desdobramentos do ocorrido.

4.1 Entregáveis

Todos os códigos e arquivos *JSON* gerados pelo presente trabalho se encontram no repositório do GitHub utilizado pelo autor do trabalho. Além disso, para acessar os dados presentes no banco, têm-se as seguintes opções de *string* de conexão mostradas nas Figuras 17 e 18. Para saber a senha do banco de dados, bem como realizar o acesso via *Compass*, favor entrar em contato em <gustavoguimaraesreis@ufu.br>.

```
mongodb+srv://projeto_tcc_gustavo:<password>@tcc.jhrmje4.mongodb.net/?
retryWrites=true&w=majority
```

Figura 17 – Acesso via *driver Node.js*

```
mongosh "mongodb+srv://tcc.jhrmje4.mongodb.net/" --apiVersion 1 --username
projeto_tcc_gustavo
```

Figura 18 – Acesso via *MongoShell*

5 Conclusão

Este trabalho de conclusão de curso teve como objetivo sugerir um processo para se trabalhar com dados coletados em redes sociais sobre incidentes de segurança. Por conseguinte, foi revisado conceitos fundamentais presentes nesse ecossistema, assim como proposto ferramentas simples e *open source* para facilitar a realização de determinadas etapas.

Nesse contexto, o *Twitter* foi escolhido como a fonte de dados devido ao histórico de engajamento dos usuários em discussões que cernem segurança da informação tanto por pessoas técnicas e com conhecimento sobre cibersegurança quanto por pessoas que de certa forma tiveram contato com o conteúdo do vazamento. Além disso, inúmeros canais de distribuição de notícias utilizam a plataforma para divulgar as suas publicações, o que se mostrou útil para a descoberta de detalhes sobre o incidente bem como descobrir novos incidentes que aconteceram paralelamente ao estudo de caso observado.

Um objetivo secundário era explorar a linguagem de programação *Python* para lidar com problemas envolvendo a coleta, visualização e análise de dados. Como conjecturado, ela se apresentou como uma forte ferramenta e uma ótima escolha para se solucionar problemas neste escopo. Isso advém, tanto pela simplicidade do código, possuindo uma sintaxe clara e objetiva, quanto pela enorme comunidade de desenvolvedores que criam constantemente *frameworks* e bibliotecas robustas para auxiliarem as tarefas comuns do cotidiano de um cientista de dados.

Ademais, realizou-se uma sucinta análise sobre o comportamento dos usuários do *Twitter*. Embora, tenha-se investigado apenas um estudo de caso, é razoável inferir que ao eclodir um incidente de segurança de grande escala, a maior densidade de postagens referentes ao evento ocorrerá nos dias subsequentes à da data de sua divulgação, por parte dos grandes meios de comunicação ou pela própria empresa. Em contrapartida, foi constatada a possibilidade de extrair informações valiosas antes da data de divulgação do evento, principalmente nos casos onde um usuário possa ter entrado em contato com o vazamento em fóruns da *DeepWeb* e *DarkWeb*, antes que a notícia chegasse para os canais de comunicação mais formais. Além disso, ficou evidente como a rede social possibilita a análise de reportagens relevantes publicadas após a data do incidente pelos diversos *websites*. Tais informações foram muito úteis para complementar o conhecimento acerca do ocorrido, bem como observar os seus desdobramentos com o passar dos dias. Todavia, o autor recomenda o prosseguimento de futuras pesquisas para um melhor e mais preciso entendimento sobre o comportamento dos usuários em redes sociais, tendo em vista que este não era o foco primordial do presente trabalho.

5.1 Contribuições

Como principais contribuições deste trabalho, destacam-se:

1. Criação de uma base de dados *NoSQL* composta por 3330 *tweets*. Além disso, a base foi dividida em coleções que representam diferentes períodos temporais;
2. Destacar a capacidade da linguagem *python* e suas bibliotecas auxiliares para lidar com a análise dos dados;
3. Ressaltar a importância das redes sociais como mecanismo de análise de incidentes de segurança;
4. Análise das postagens de usuários que antecederam a data do incidente, confirmando a possibilidade de se extrair informações valiosas sobre o ocorrido;
5. Análise do comportamento das discussões presentes em redes sociais ao eclodir um incidente de segurança, atestando a alta densidade de postagens em um curto período, bem como a postagem frequente de *hiperlinks* para notícias que auxiliam a complementar as informações acerca do incidente e os desdobramentos ao decorrer do tempo.

5.2 Trabalhos futuros

Neste trabalho, a coleta de dados foi realizada por meio da biblioteca *snscrape*, na qual utilizava diretamente a API disponibilizada pelo *Twitter*. Entretanto, o acesso gratuito da mesma fora interrompida no final de janeiro de 2023, o que prejudicou demasiada a coleta dos dados. Portanto, uma sugestão para trabalhos futuros é a descoberta de novas ferramentas que permitem o *scrapping* de *tweets*.

Caso, a empresa não reverta a decisão, recomenda-se a pesquisa por meio da coleta de dados proveniente de outras redes sociais, tais como: *Reedit* e *Facebook*, bem como fóruns especializados em tecnologia e segurança da informação.

Além disso, sugere-se que demais pesquisas referentes ao comportamento de usuários se expanda para análise específica de incidentes que ocorreram em território brasileiro. Dessa forma, será possível comparar e analisar dados de discussões vindas de diferentes países, possibilitando encontrar similaridades e divergências geográficas em seu comportamento.

Referências

MongoDB, Inc. *Get Started with Atlas*. New York, United States, 2022. Acessado em: 05/11/2022. Disponível em: <https://www.mongodb.com/docs/atlas/getting-started/>. Citado na página 20.

MongoDB, Inc. *What is MongoDB?* New York, United States, 2022. Acessado em: 05/11/2022. Disponível em: <https://www.mongodb.com/docs/manual/>. Citado na página 20.

ABRAMS, L. **533 million Facebook users' phone numbers leaked on hacker forum**. bleepingcomputer.com, 2021. Acessado em: 14/011/2022. Disponível em: <https://www.bleepingcomputer.com/news/security/533-million-facebook-users-phone-numbers-leaked-on-hacker-forum/>. Citado na página 24.

ALTALHI, S.; GUTUB, A. A survey on predictions of cyber-attacks utilizing real-time twitter tracing recognition. **Journal of Ambient Intelligence and Humanized Computing**, 2020. Citado na página 22.

ALVES, F.; BETTINI, A.; FERREIRA, P. M.; BESSANI, A. Processing tweets for cybersecurity threat awareness. Lisboa, Portugal, 2019. Citado na página 22.

ARCHIVIST, J. A. **Snscrape**. JustAnotherArchivist, 2022. Acessado em: 25/10/2022. Disponível em: <https://github.com/JustAnotherArchivist/snscrape>. Citado 2 vezes nas páginas 17 e 25.

BBC, N. **Facebook faces mass legal action over data leak**. bbc.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.bbc.com/news/technology-56772772>. Citado na página 37.

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: **International Conference on Advances in Social Networks Analysis and Mining**. Manhattan, Kansas: IEE/ACM, 2019. (66506), p. 207–216. Citado na página 21.

CERT.BR. **Cartilha de segurança para a internet**. Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, MD 20899-8900, 2022. Acessado em: 20/12/2022. Disponível em: <https://cartilha.cert.br/>. Citado 2 vezes nas páginas 13 e 14.

CIMPANU, C. **Phone numbers for 533 million Facebook users leaked on hacking forum**. The Record from Recorded Future News, 2021. Disponível em: <https://therecord.media/phone-numbers-for-533-million-facebook-users-leaked-on-hacking-forum>. Citado 3 vezes nas páginas 24, 25 e 34.

CLARK, M. **The Facts on News Reports About Facebook Data**. about.fb.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://about.fb.com/news/2021/04/facts-on-news-reports-about-facebook-data/>. Citado na página 36.

CULLIFORD, E. **Facebook does not plan to notify half-billion users affected by data leak**. reuters.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.reuters.com/article/us-facebook-data-leak/facebook-does-not-plan-to-notify-half-billion-users-affected-by-data-leak-idUSKBN2BU2ZY>. Citado na página 38.

DIONÍSIO, N.; ALVES, F.; FERREIRA, P. M.; BESSANI, A. Cyberthreat detection from twitter using deep neural networks. **CoRR**, abs/1904.01127, 2019. Disponível em: <http://arxiv.org/abs/1904.01127>. Citado na página 22.

ENISA. **Reference Incident Classification Taxonomy**. ENISA, 2018. Acessado em: 31/05/2023. Disponível em: <https://www.enisa.europa.eu/publications/reference-incident-classification-taxonomy>. Citado na página 15.

FIPS. **Standards for Security Categorization of Federal Information and Information Systems**. Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, MD 20899-8900, 2004. Disponível em: <https://nvlpubs.nist.gov/nistpubs/fips/nist.fips.199.pdf>. Citado na página 13.

FOUDANTION, P. S. **Welcome to Python**. Python Software Foudantion, 2023. Acessado em: 12/06/2023. Disponível em: <https://www.python.org/>. Citado na página 17.

FRIPP, C. **Facebook data leak: 500M user phone numbers for sale online**. KOMANDO.COM, 2021. Acessado em: 10/06/2023. Disponível em: <https://www.komando.com/security-privacy/facebook-data-leak/775660/>. Citado na página 34.

GUALBERTO Éder S.; JR, R. T. de S.; DEUS, F. E. G. de; DUQUE, C. G. Proposição de uma ontologia de apoio à gestão de riscos de segurança da informação. Revista Brasileira de Sistemas de Informação, Brasília, Distrito Federal, 2013. Citado na página 13.

HAT, I. R. **What is an API**. redhat.com, 2022. Acessado em: 11/06/2023. Disponível em: <https://www.redhat.com/pt-br/topics/api/what-are-application-programming-interfaces>. Citado na página 16.

HOLMES, A. **533 million Facebook users' phone numbers and personal data have been leaked online**. businessinsider.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.businessinsider.com/stolen-data-of-533-million-facebook-users-leaked-online-2021-4>. Citado 2 vezes nas páginas 36 e 38.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 19.

JEE, C. **What you need to know about the Facebook data leak**. technologyreview.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.technologyreview.com/2021/04/07/1021892/facebook-data-leak/>. Citado na página 38.

KAO, A.; POTEET, S. R. **Natural Language Processing and Text Mining**. Bellevue, USA: Springer-Verlag London Limited, 2007. Citado na página 15.

KHDER, M. A. Web scraping or web crawling: State of art, techniques, approaches and application. IW3C2, Amã, Jordânia, v. 13, n. 3, 2021. Citado na página 16.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). **Proceedings of the 9th Python in Science Conference**. Scipy Conference 2010: Scipy, 2010. p. 56 – 61. Citado na página 18.

MITTAL, S.; DAS, P. K.; MULWAD, V.; JOSHI, A.; FININ, T. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In: **2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. San Francisco, CA, USA: IEEE Press, 2016. p. 860–867. Citado na página 22.

NEWMAN, L. H. **What Really Caused Facebook's 500M-User Data Leak?** wired.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.wired.com/story/facebook-data-leak-500-million-users-phone-numbers/?utm_medium=twitter&utm_source=dlvr.it>. Citado na página 38.

NIS, C. G. T. **Cybersecurity Incident Taxonomy**. NIS Cooperation Group, 2018. Acessado em: 12/02/2023. Disponível em: <https://ec.europa.eu/information_society/newsroom/image/document/2018-30/cybersecurity_incident_taxonomy_00CD828C-F851-AFC4-0B1B416696B5F710_53646.pdf>. Citado na página 14.

ORACLE. **What is a database?** 2022. Acessado em: 18/12/2022. Disponível em: <https://www.oracle.com/database/what-is-database/>. Citado na página 19.

O'SULLIVAN, D. **Half a billion Facebook users' information posted on hacking website, cyber experts say**. edition.cnn.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://edition.cnn.com/2021/04/04/tech/facebook-user-info-leaked/index.html>. Citado na página 38.

PASSERI, P. **hackmageddon**. Paolo Passeri, 2023. Acessado em: 15/03/2023. Disponível em: <https://www.hackmageddon.com/>. Citado na página 24.

PAULO, J. de F. Análise de incidentes de segurança usando séries temporais e modelos arima. Uberlândia, Minas Gerais, 2021. Citado na página 21.

RELIAQUEST. **The Facebook Data Leak Explained**. reliaquest.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.reliaquest.com/blog/the-facebook-data-leak-explained/>. Citado na página 37.

REPORT, S. **Brasil foi o 5º país com mais ataques cibernéticos em 2021**. Security Report, 2022. Acessado em 09/11/2022. Disponível em: <https://www.securityreport.com.br/overview/brasil-foi-o-5o-pais-com-mais-ataques-ciberneticos-em-2021/#.ZFrr6I3MLMU>. Citado na página 10.

RITTER, A.; CASEY, W.; WRIGHT, E.; MITCHELL, T. Weakly supervised extraction of computer security events from twitter. IW3C2, Florence, Italy, 2015. Citado na página 10.

RODRIGUEZ, A.; OKAMURA, K. Generating real time cyber situational awareness information through social media data mining. In: **2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)**. Milwaukee, WI, USA: IEEE, 2019. v. 2, p. 502–507. Citado na página 22.

SANTOS, L. A. F.; CAMPIOLO, R.; GEROSA, M. A.; BATISTA, D. M. Detecção de alertas de segurança em redes de computadores usando redes sociais. 31o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, Brasília, Distrito Federal, 2013. Citado na página 10.

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 Orleans, LA, USA: IEEE, 2017. p. 667–674. Citado na página 22.

SCELLER, Q. L.; 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: Association for Computing Machinery, 2017. (ARES '17). ISBN 9781450352574. Disponível em: <https://doi.org/10.1145/3098954.3098992>. Citado na página 22.

SCHAFFER, A. **The Cybersecurity 202: A massive Facebook breach underscores limits to current data breach notification laws**. washingtonpost.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://www.washingtonpost.com/politics/2021/04/06/cybersecurity-202-massive-facebook-breach-underscores-limits-current-data-breach-notification-laws>. Citado na página 37.

SOLMS, R. von. Information security management (3): the code of practice for information security management (bs 7799). **Inf. Manag. Comput. Secur.**, v. 6, p. 224–225, 1998. Citado na página 13.

SOUSA, D. A. de. **Descoberta de Exploits Usando Dados da Rede Social Twitter**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Uberlândia, 2020. Citado 2 vezes nas páginas 15 e 21.

STALLINGS, W. **Criptografica e segurança de redes princípios e práticas**. São Paulo, Brasi: Pearson Education, 2014. Citado na página 13.

TWITTER, I. **New user FAQ**. Twitter help center, 2023. Acessado em: 15/01/2023. Disponível em: <https://help.twitter.com/en/resources/new-user-faq>. Citado na página 17.

VANO, Z. **Facebook Personal Profiles Data Leak [April 2021]**. lookerstudio.google.com, 2021. Acessado em: 15/06/2021. Disponível em: <https://lookerstudio.google.com/reporting/afa08373-621e-4e45-990e-bd631fd3b27a/page/Cn9AC>. Citado na página 36.

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>. Citado na página 19.