UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Anna Letycia Fernandes Reis

Análise de *posts* maliciosos na *Dark Web* usando aprendizado de máquina não supervisionado

Uberlândia, Brasil

2024

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Anna Letycia Fernandes Reis

Análise de *posts* maliciosos na *Dark Web* usando aprendizado de máquina não supervisionado

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

2024

Anna Letycia Fernandes Reis

Análise de *posts* maliciosos na *Dark Web* usando aprendizado de máquina não supervisionado

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.

Prof. Dr. Rodrigo Sanches Miani
Orientador

Prof. Dr. Paulo Henrique Ribeiro Gabriel

Prof. Dr. Bruno Augusto Nassif Travençolo

Uberlândia, Brasil
2024

Aos meus pais, Carluce e Ernando, por serem a minha maior fonte de inspiração e os pilares que sustentam os meus sonhos.

Agradecimentos

Em primeiro lugar, agradeço a Deus por Sua graça infinita sem a qual eu não estaria aqui hoje. Por iluminar meu caminho com Suas bênçãos e me guiar nos momentos de incerteza. Por cada oportunidade que me foi concedida e por nunca me deixar só.

Aos meus pais e a toda a minha família, que são meu porto seguro e alicerce. Este trabalho só foi possível graças ao amor e apoio incondicionais de vocês, que me acompanham em cada etapa da minha jornada. Tudo o que sou é reflexo do carinho, cuidado e incentivo que sempre recebi de cada um de vocês.

Ao meu namorado, Caio, pelo apoio, paciência, carinho e companheirismo durante todo o desenvolvimento deste trabalho. Por sempre me incentivar e acreditar em mim, especialmente nos momentos difíceis.

Ao meu orientador, Professor Rodrigo Sanches Miani, pela oportunidade de desenvolver este trabalho. Pela orientação sempre cuidadosa, pelo incentivo constante e pela maneira leve e acolhedora com que me conduziu ao longo deste processo.

Aos professores da FACOM e a todos que fizeram parte da minha trajetória, minha eterna gratidão. Por cada ensinamento, por cada gesto de paciência e dedicação, que me impulsionaram até aqui. Vocês foram fundamentais para o meu crescimento, não só como estudante, mas, sobretudo, como ser humano.

Resumo

Este trabalho apresenta uma análise de *posts* maliciosos extraídos de fóruns da *Dark Web* por meio de técnicas de aprendizado não supervisionado, com o objetivo de identificar as temáticas predominantes associadas a ameaças cibernéticas. Para isso, foi empregada uma metodologia baseada em algoritmos de agrupamento, como K-means, DBSCAN e KNN, além da aplicação da Análise de Tópicos Latentes (LDA) para identificar padrões temáticos latentes. Os resultados demonstraram que o algoritmo K-means se destacou ao estruturar os dados em três *clusters* principais, identificando temas predominantes como segurança de dados, busca por informações sensíveis e comunidades de *hacking*. Essa abordagem possibilitou a rotulagem e interpretação dos conteúdos com base nos padrões observados, contribuindo para a compreensão das táticas e intenções dos cibercriminosos. Como perspectivas futuras, sugere-se ampliar a base de dados para incluir ambientes como a *Deep Web*, *Surface Web* e redes sociais, além de incorporar algoritmos avançados de aprendizado profundo e ferramentas de monitoramento em tempo real, visando um aprimoramento contínuo na detecção e categorização de ameaças.

Palavras-chave: *Dark Web*, Aprendizado não supervisionado, K-means, LDA, Segurança cibernética, Análise de posts maliciosos.

Abstract

This work presents an analysis of malicious posts extracted from Dark Web forums using unsupervised learning techniques, aiming to identify the predominant themes associated with cyber threats. A methodology was employed based on clustering algorithms, such as K-means, DBSCAN, and KNN, in addition to applying Latent Dirichlet Allocation (LDA) to identify latent thematic patterns. The results demonstrated that the K-means algorithm excelled in structuring the data into three main clusters, identifying predominant themes such as data security, search for sensitive information, and hacking communities. This approach enabled the labeling and interpretation of content based on observed patterns, contributing to the understanding of cybercriminals' tactics and intentions. For future work, it is suggested to expand the dataset to include environments such as the Deep Web, Surface Web, and social networks, as well as to incorporate advanced deep learning algorithms and real-time monitoring tools, aiming for continuous improvement in threat detection and categorization.

Keywords: Dark Web, Unsupervised learning, K-means, LDA, Cybersecurity, Malicious post analysis.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CNPMI	*Cross-lingual Normalized Pointwise Mutual Information*
CTI	*Cyber Threat Intelligence*
CTM	*Correlated Topic Model*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
IA	Inteligência Artificial
IoT	*Internet of Things*
IoCs	Indicadores de Comprometimento
IP	*Internet Protocol*
KNN	*K-Nearest Neighbors*
LDA	*Latent Dirichlet Allocation*
PAM	*Pachinko Allocation Topic Model*
PCA	*Principal Component Analysis*
PLN	Processamento de Linguagem Natural
PTM	*Pseudodocument Topic Model*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TTI	*Tactical Threat Intelligence*
Umass	*University of Massachusetts*

Sumário

1 Introdução

No século XXI, diante dos constantes avanços digitais e da crescente conectividade global, houve uma mudança significativa no modo como as pessoas se comunicam, trabalham, aprendem e interagem com o mundo ao seu redor. Essa interconexão sem precedentes, como aponta Shirky (2010), reconfigura as estruturas sociais e econômicas, ao mesmo tempo em que cria novas oportunidades para a prática de atividades ilícitas. Advinda da era da tecnologia digital, essa interconexão afetou diversas áreas na vida cotidiana, principalmente com a ascensão da Internet das Coisas (IoT), a difusão de dispositivos inteligentes e o avanço de serviços online. Todavia, emergiram também desafios expressivos, como os ataques cibernéticos, que exploram vulnerabilidades e comprometem a segurança digital. Nesse contexto, a *Dark Web*, parte oculta da Internet que só pode ser acessada por meio de redes anônimas, garante um ambiente propício para a realização de atividades ilícitas, como fraudes, golpes, venda de dados pessoais e ataques cibernéticos, o que evidencia a urgente necessidade por tecnologias mais avançadas com o propósito de detectar e impedir possíveis ameaças à segurança digital.

Segundo OGlobo (2023), em uma pesquisa levantada pela Netscout, empresa líder global em soluções de cibersegurança, o aumento dos ciberataques é evidente no mundo todo e o Brasil é o principal alvo na América Latina. Com números referentes ao segundo semestre de 2022, foram registrados mais de 285 mil ataques no país durante esse período, comprovando a demanda por estratégias mais eficientes para enfrentar tais desafios eminentes. De acordo com o relatório "*X-Force Threat Intelligence Index 2023*", realizado pela IBM (2023), o roubo de dados lidera os ataques cibernéticos no Brasil, com 33% desses incidentes sendo representados por coleta de credenciais e ataques de *phishing*, o que indica interesse crescente em informações pessoais para venda na *Dark Web*. Perante o exposto, o monitoramento das atividades nesse ambiente, por meio da análise de *posts* maliciosos, contribui para a área de Inteligência de Ameaças Cibernéticas, do inglês *Cyber Threat Intelligence* (CTI), por permitir o levantamento de dados e a exploração de informações relevantes sobre táticas de ataque emergentes, possíveis alvos e tendências dos cibercriminosos. Logo, a investigação aprofundada dos fóruns da *Dark Web* auxilia no desenvolvimento de respostas mais ágeis e de contramedidas robustas, antecipadas e devidamente direcionadas, o que, por consequência, auxilia a comunidade de segurança cibernética na extração de informações de CTI.

Em concordância com o que foi declarado anteriormente, (BASHEER; ALKHATIB, 2021) ressaltam sobre a contribuição advinda da análise de conteúdo presente em plataformas *Dark Web* e a importância do aprendizado de máquina e do processamento de linguagem para o futuro da CTI, além disso, salientam que a detecção eficiente de

ameaças à segurança cibernética é crucial para antecipar e descobrir potenciais ataques antes mesmo de se concretizarem. Por sua vez, com o intuito de auxiliar pesquisadores na área da cibersegurança, (RAHMAN; HEZAVEH; WILLIAMS, 2023) abordam sobre as técnicas atuais usadas na extração de inteligência de ameaças cibernéticas através de textos, categorizando os propósitos de extração de CTI e fornecendo *insights* valiosos sobre metodologias e abordagens. Já (JESUS FILHO, 2023) desenvolve em seu trabalho um modelo de detecção de *posts* maliciosos a partir de técnicas de mineração de texto, processamento de linguagem natural (PLN) e aprendizado de máquina, com aplicação aos dados obtidos de fóruns na *Dark Web*, enquanto (NAZAH et al., 2021a) propõem um método não supervisionado para identificar e classificar esses fóruns, com foco no monitoramento de discussões e identificação de possíveis violações de dados, proporcionando informações importantes para detectar atividades suspeitas.

Primordialmente, os objetivos gerais deste trabalho consistiram na análise de postagens maliciosas na *Dark Web*, por meio de técnicas de aprendizado não supervisionado, a fim de explorar e identificar as temáticas predominantes e os diferentes tipos de ameaças cibernéticas presentes em fóruns desse ambiente. Para isso, foi utilizado um modelo de detecção capaz de identificar *posts* contendo dados representativos de atividades maliciosas. A partir desse modelo, as postagens foram agrupadas e classificadas utilizando técnicas de processamento de linguagem natural e algoritmos de agrupamento, permitindo a organização dos dados em categorias temáticas. Esse método possibilitou a compreensão das estratégias e táticas empregadas por cibercriminosos e a identificação dos tópicos predominantes. Por fim, a intenção é compartilhar os resultados obtidos com agentes de segurança para proporcionar suporte e auxiliar o desenvolvimento de defesas proativas contra ciberataques.

Diante dos desafios apresentados anteriormente, foram empregadas diversas ferramentas e tecnologias. A princípio, a linguagem de programação Python (2024) foi utilizada para manipular e analisar os dados, garantindo maior flexibilidade e eficiência. A biblioteca Scikit-learn (2024) foi usada para explorar padrões semânticos e características distintas nos *posts*, por meio de algoritmos de aprendizado de máquina e técnicas de processamento de linguagem natural. Já o uso da biblioteca NLTK (2023) (*Natural Language Toolkit*), foi voltado para o pré-processamento e análise textual, a partir da remoção de *stop words* e da lematização. Ademais, algoritmos de modelagem de tópicos, como *Latent Dirichlet Allocation* (LDA), foram empregados para identificar tópicos relevantes e compreender as principais discussões abordadas, e algoritmos de agrupamento foram aplicados para classificar *posts* automaticamente, com base em padrões de comportamento e semelhanças.

Este trabalho é composto por cinco capítulos, além desta introdução. O Capítulo 2 apresenta o referencial teórico, abordando os conceitos fundamentais para a compreensão

do estudo e revisando trabalhos relacionados. A metodologia utilizada, com suas etapas detalhadas, é apresentada no Capítulo 3. O Capítulo 4 descreve os experimentos realizados e analisa os resultados obtidos. Finalmente, o Capítulo 5 expões as conclusões do trabalho, destacando as principais contribuições e sugerindo direções para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos básicos que sustentam o desenvolvimento deste trabalho e são fundamentais para sua compreensão. As primeiras cinco seções discutem temas como segurança cibernética, *Cyber Threat Intelligence* (CTI) com exemplos, aspectos da *Dark Web* e seus fóruns, aprendizado não supervisionado e suas aplicações, além do Processamento de Linguagem Natural (PLN). Por fim, a seção 2.6 aborda e analisa os trabalho correlatos a este estudo.

2.1 Segurança Cibernética

A cibersegurança, também conhecida como segurança cibernética, abrange a organização e o uso de recursos, processos e estruturas para proteger o ciberespaço, e os sistemas que operam dentro dele, contra ameaças que possam comprometer a preservação dos direitos de propriedade, tanto em termos legais quanto práticos (CRAIGEN; DIAKUN-THIBAULT; PURSE, 2014).

À luz de Taherdoost (2022), embora a cibersegurança e a segurança da informação sejam intimamente relacionadas pelo objetivo em comum de proteger dados, a diferença entre elas reside no fato de que, enquanto a segurança da informação envolve a proteção de dados em qualquer lugar, a cibersegurança foca especificamente na proteção de sistemas e dados no ciberespaço. O presente trabalho está diretamente associado à segurança cibernética, a fim de explorar medidas de proteção e análise de informações no ambiente digital.

Para Stallings e Brown (2014), os objetivos fundamentais da segurança da informação são divididos em três princípios, conhecidos como a tríade CID: confidencialidade, integridade e disponibilidade. Esses pilares da proteção de dados e sistemas contra ameaças podem ser descritos como:

- Confidencialidade: princípio que garante a proteção dos dados contra acessos não autorizados. Apenas usuários com permissões podem acessar as informações.
- Integridade: princípio que certifica a consistência dos dados ao longo do tempo, garantindo que as informações permaneçam inalteradas e confiáveis.
- Disponibilidade: princípio que assegura que os dados estejam acessíveis e disponíveis para uso quando necessário, sem interferência ou obstrução.

Após as grandes perturbações decorrentes das mudanças sociais, econômicas e tecnológicas impulsionadas pela COVID na década de 2020, os criminosos cibernéticos aprimoraram suas táticas para tornar seus ataques ainda mais sofisticados (SUN et al., 2023). Logo, é evidente a necessidade de elaborar técnicas de cibersegurança mais avançadas que os métodos de segurança convencionais, a fim de garantir a disponibilidade, integridade e confidencialidade dos sistemas (International Telecommunication Union (ITU), 2008).

2.2 *Cyber Threat Intelligence* (CTI)

Segundo Rahman, Hezaveh e Williams (2023), embora não exista uma definição universalmente aceita, a CTI pode ser definida como qualquer informação que seja considerada relevante para a cibersegurança e que auxilie na previsão, prevenção ou defesa proativa contra ciberataques. De modo semelhante, para Shackleford (2015), o conjunto de dados coletados, avaliados e aplicados em relação a ameaças de segurança, agentes de ameaças, *softwares* maliciosos, vulnerabilidades e indicadores de comprometimento está relacionado com a CTI. A partir disso, organizações e profissionais de segurança compreendem melhor os riscos cibernéticos e podem desenvolver estratégias mais eficazes e antecipadas para esses incidentes.

Tounsi e Rais (2018) classificaram os tipos de inteligência de ameaças existentes em estratégicas, operacionais e táticas, com foco na Inteligência de Ameaças Táticas (TTI), gerada principalmente a partir de Indicadores de Comprometimento (IoCs), como por exemplo IPs suspeitos, domínios maliciosos conhecidos, tentativas de login inusitadas e tráfego incomum. Esses indicadores auxiliam na identificação de atividades suspeitas em sistemas e redes, além de conseguirem captar operações maliciosas em estágios iniciais.

Todavia, Jo, Lee e Shin (2022) apontam que, tendo em vista a constante evolução das ameaças cibernéticas, os IoCs não devem ser a única prioridade durante a elaboração de estratégias de segurança. Por exemplo, com os recentes avanços em IA, abordagens de aprendizado de máquina e inteligência artificial são algumas das possíveis aplicações para aproveitar a CTI e, consequentemente, impedir violações, conforme discutido por Ibrahim et al. (2020). Alguns exemplos desse tipo de abordagem com aplicação à CTI são: algoritmos de classificação de texto, para separar informações relevantes em documentos; algoritmos de detecção de anomalias, para encontrar comportamentos suspeitos em dados de rede; clusterização de dados, para agrupar assuntos relacionados e verificar tendências de atividades maliciosas; e análise de sentimentos, para avaliar o conteúdo emocional de postagens online e identificar possíveis ameaças.

2.3 *Dark Web*

Considerada o maior sistema global de informações, a Internet é constituída por diferentes camadas que variam em visibilidade e profundidade. Como ilustrado na Figura 1, essas camadas incluem a *Surface Web*, a *Deep Web* e a *Dark Web*. Basheer e Alkhatib (2021) descrevem essa organização da seguinte forma:

- *Surface Web*: representada na parte superior do *iceberg*, é conhecida como a camada mais superficial, composta por sites facilmente acessíveis e indexados por mecanismos de pesquisa, como o *Google*, *Bing* e *Twitter*.
- *Deep Web*: localizada abaixo da superfície, também denominada como *Web* Invisível, é a área não indexada, formada por sites não acessíveis por mecanismos de busca, apresentando conteúdos que não são necessariamente ilegais, como registros médicos, relatórios científicos e documentos legais.
- *Dark Web*: representada na base do *iceberg*, é a parte mais oculta, acessível somente por meio de *softwares* específicos. A identidade dos usuários e os seus endereços IP são escondidos por criptografia, garantindo anonimato e, por consequência, tornando o ambiente propício para práticas ilegais, como o tráfico de drogas e outros conteúdos confidenciais.

Figura 1 – Representação da *Surface Web*, *Deep Web* e *Dark Web*. (Fonte: (XYAN; WEI; JUREMI, 2020))

A metáfora do *iceberg* é comumente empregada para representar as camadas da Internet, uma vez que enfatiza que a maior parte da *web* é oculta aos usuários comuns (AKHGAR et al., 2021). Este trabalho, por sua vez, foca na análise dos dados presentes em fóruns da *Dark Web*.

Para participar de discussões sobre diversos assuntos e interagir com outros usuários, os membros da *Dark Web* acessam os chamados fóruns, que geralmente são organizados em diferentes categorias. Alguns dos fóruns mais conhecidos incluem o *Hidden Answers*, *Raddle*, *Dread* e *Deep Answers*, que oferecem espaços para discussões anônimas. Por isso, é comum que conteúdos relacionados a práticas ilícitas, como a troca de informações sobre *hacking*, tráfico de drogas e armas, além da venda de informações sensíveis, sejam encontrados nesses espaços.

Além disso, o acesso a um fórum deve ser mediante o uso de navegadores especializados, com o intuito de dificultar o rastreamento da identidade do usuário (AKHGAR et al., 2021). Toda publicação realizada em um fórum da *Dark Web*, contendo informações que indiquem um possível risco à segurança cibernética, como por exemplo uma ameaça, vulnerabilidade ou vazamento de dados, é considerada um *post* de conteúdo malicioso.

2.4 Aprendizado Não Supervisionado

De acordo com Gentleman e Carey (2008), o aprendizado de máquina é tradicionalmente dividido em duas áreas: aprendizado supervisionado, referido na literatura estatística como classificação, e aprendizado não supervisionado, referenciado como agrupamento. Ambos os tipos de aprendizagem envolvem a análise de conjunto de dados contendo observações multivariadas. Ao contrário do supervisionado, o aprendizado não supervisionado carece de uma medida clara de sucesso, dificultando a determinação da validade das inferências e, consequentemente, gerando uma grande quantidade de métodos propostos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Gentleman e Carey (2008) também acrescentam que no contexto da aprendizagem não supervisionada, por não receber rótulos predefinidos, o algoritmo é desafiado a descobrir e interpretar características e padrões nos dados por conta própria. Isso é feito a partir da divisão do conjunto de dados em grupos ou *clusters*, baseada na similaridade entre as amostras, com o propósito de fornecer uma visão clara dos objetos estudados. Alguns dos algoritmos de agrupamento mais conhecidos são: K-means, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) e *K-Nearest Neighbors* (KNN), cada um com características e aplicações específicas.

O K-means, por exemplo, é um algoritmo de agrupamento amplamente utilizado para a segmentação de dados em aprendizado de máquina. Baseado em centroides, esse algoritmo divide os dados em k *clusters*, onde k é definido pelo usuário, buscando minimizar a distância quadrada média entre os pontos de um mesmo *cluster* (ARTHUR; VASSILVITSKII, 2006). No entanto, suas limitações incluem a dificuldade em lidar com ruídos e com *clusters* de formato não circular.

A análise de *clusters* utilizando o algoritmo K-means é uma técnica poderosa

para identificar padrões em grandes conjuntos de dados textuais. Ao agrupar documentos semelhantes, o K-means identifica temas latentes e compreende a estrutura da informação. A identificação dos termos mais relevantes em cada *cluster*, frequentemente associados a centroides com maior peso, é uma etapa crucial nesse processo. Conforme Manning, Raghavan e Schütze (2008), a análise dos termos mais frequentes em cada *cluster* pode revelar os tópicos dominantes e facilitar a interpretação dos resultados.

Por outro lado, o DBSCAN agrupa as informações com base na densidade. Ele unifica os dados presentes em regiões densas em *clusters* e considera os pontos em regiões de baixa densidade como ruídos, detectando *clusters* arbitrariamente sem uma especificação prévia. A principal limitação desse algoritmo consiste na sensibilidade a parâmetros como o raio de busca (Eps) e o número mínimo de pontos (MinPts) para formar um *cluster* (ESTER et al., 1996).

Por fim, o KNN é um algoritmo de classificação e regressão, que atribui um valor a um ponto baseado nos k pontos vizinhos, ou mais próximos, dele. A classificação de um dado é definida pela classe mais frequente entre os seus vizinhos e, por isso, seu desempenho é influenciado pela distância entre os pontos e pela escolha de k (COVER; HART, 1967). A desvantagem dessa técnica é que, para grandes conjuntos de dados o algoritmo apresenta alto custo computacional, visto que precisa calcular as distâncias a cada nova previsão.

Além disso, Wong (2021) ressalta que o PCA (Análise de Componentes Principais), uma técnica de redução de dimensionalidade, é usado para pré-processar os dados antes da aplicação de métodos de agrupamento, pois facilita a visualização de padrões e relações entre os dados. Para isso, as variáveis originais são transformadas em componentes principais, o que permite reduzir a complexidade dos dados sem que informações relevantes sejam perdidas.

Para uma análise de dados mais aprofundada, o *Latent Dirichlet Allocation* (LDA) é frequentemente aplicado em técnicas de modelagem de tópicos não supervisionadas, por se tratar de um modelo probabilístico generativo que identifica tópicos latentes em um grupo de documentos (NAZAH et al., 2021b). Ele assume que cada documento é composto por uma mistura de tópicos e cada tópico uma mistura de palavras, ajustando essas relações por meio de um processo iterativo. A partir disso, obtém-se uma distribuição equilibrada dos dados, com os tópicos dominantes de cada documento e as palavras mais significativas de cada tópico. Em conjunto, o LDA e os algoritmos de agrupamento auxiliam na compreensão minuciosa e contextual das informações analisadas.

2.5 Processamento de Linguagem Natural

O Processamento de Linguagem Natural é a área de pesquisa que estuda as formas de capacitar sistemas computacionais a compreender e manipular dados textuais em linguagem natural. Uma das etapas principais do PLN envolve a conversão de textos em representações numéricas, denominadas vetores. Essa etapa permite que os computadores processem e analisem as informações textuais, para realizar atividades úteis (CHOWDHARY, 2020).

A técnica de vetorização textual transforma um conjunto de palavras em um vetor numérico, onde cada termo no texto é representado numericamente. Uma abordagem comum para essa conversão é o TF-IDF, vetorização baseada na frequência dos termos, que atribui um peso a cada palavra, multiplicando sua frequência no documento pelo inverso da sua frequência em todos os documentos (MOREIRA, 2024).

Segundo Cheng et al. (2020), o TF-IDF é amplamente adotado para a extração eficaz de palavras-chave. Para demonstrar o seu funcionamento considere o seguinte exemplo:

1. “flor rosa nasce”
2. “folha verde nasce”
3. “flor verde cai”

Primeiramente, o método identifica o vocabulário do conjunto de frases, composto pelos termos únicos, que neste caso são: “flor”, “rosa”, “nasce”, “folha”, “verde” e “cai”. Em seguida, calcula-se a frequência de cada termo em relação a cada frase, levando em consideração a raridade do termo dentro do conjunto. Quanto mais rara é uma palavra, como por exemplo “rosa”, maior é seu peso em comparação a termos mais frequentes, como “verde”.

Esse processo resulta em uma matriz de vetores, na qual cada linha representa uma frase e cada coluna corresponde a um termo. Desse modo, cada elemento da matriz contém um valor referente a importância de cada palavra na frase e no conjunto total. Desse modo, o vetor numérico resultante pode ser aplicado em algoritmos de aprendizado de máquina, a fim de realizar análises de similaridade, classificações ou agrupamentos de documentos.

2.6 Trabalhos Relacionados

O objetivo desta seção é apresentar e analisar os trabalhos correlatos ao projeto desenvolvido, os quais dedicam-se a investigar e explorar os conteúdos da *Dark Web*, por

meio de diversas abordagens. Serão destacadas as contribuições de cada um desses estudos para a pesquisa relacionada à identificação e análise de conteúdos maliciosos na *Dark Web*.

O trabalho desenvolvido por JESUS FILHO (2023) propõe explorar fóruns da *Dark Web*, a partir do desenvolvimento de um modelo computacional, a fim de idenfiticar *posts* maliciosos e detectar ataques cibernéticos, vulnerabilidades e outras ameaças. Técnicas de mineração de texto, processamento de linguagem natural e aprendizado de máquina supervisionado foram aplicadas em dados extraídos de fóruns da *Dark Web*, e rotulados com base em Indicadores de Comprometimento (IoCs), palavras-chave contextuais e análise manual. Os resultados obtidos a partir desse trabalho, indicaram que o modelo classificou com precisão *posts* que continham predominância de termos relacionados a ameaças cibernéticas como *posts* maliciosos. Nesse sentido, o trabalho contribui no âmbito da investigação e análise do conteúdo dos *posts* maliciosos, para identificar os diferentes tipos de ameaças presentes na *Dark Web*.

Por outro lado, Nazah et al. (2021a) sugerem um outro tipo de abordagem, dessa vez por um modelo não supervisionado, porém em comum objetivo de monitorar discussões em fóruns da *Dark Web* e detectar possíveis violações virtuais. Com o intuito de não depender de conteúdos rotulados, devido à natureza dos dados, os pesquisadores combinaram algoritmos de agrupamento e árvores de decisão, avaliando o modelo utilizando dados reais, métricas de desempenho e validação cruzada K-fold. Os experimentos realizados evidenciaram a eficácia do modelo desenvolvido, que foi capaz de identificar, por exemplo, que um dos temas mais discutidos foi o de informações vazadas ou violadas. Basicamente, esse trabalho fornece uma ferramenta útil para identificar padrões de comportamento suspeito, identificar tópicos relevantes e extrair informações significativas presentes na *Dark Web*, que podem ser utilizadas para análise de ameaças cibernéticas e investigações de segurança.

De modo semelhante, Basheer e Alkhatib (2024) apresentam um método voltado a analisar o conteúdo de redes sociais na *Dark Web*, com foco em fóruns relacionados a atividades maliciosas. O procedimento aplicado envolve a modelagem de tópicos com quatro algoritmos (LDA, CTM, PAM e PTM) e quatro medidas de coerência (UMass, UCI, CNPMI e CV), para avaliar a qualidade dos modelos e determinar qual a quantidade de tópicos ideal. Logo, a partir da discussão e avaliação dos resultados encontrados, foi comprovado que a metodologia de modelagem de tópicos é promissora para analisar o conteúdo de atividades ilícitas e compreender as implicações dos padrões de pensamento na *Dark Web*. Este estudo se relaciona com a questão a ser pesquisada, pois proporciona uma compreensão sobre como os *posts* da *Dark Web* estão sendo analisados, buscando extrair padrões de pensamento e identificar os principais temas e correlações entre eles.

Outrossim, ainda por Basheer e Alkhatib (2021), mas dessa vez com o objetivo de revisar estudos sobre a análise de conteúdos na *Dark Web* para Inteligência de Ameaças

Cibernéticas (CTI), os pesquisadores exploram os diferentes tipos de técnicas, abordagens e métodos, analisando e discutindo as possíveis limitações e as principais relevâncias desses estudos, a fim de produzir uma visão abrangente do assunto. A revisão inclui tanto trabalhos com análises exclusivas da *Dark Web*, quanto com análises integradas e combinadas com informações da *Surface Web*, *Deep Web* e outras instituições de segurança cibernética. Como resultado, foi possível destacar a importância da *Dark Web* como fonte de informações para CTI, que pode se beneficiar com o uso de inteligência artificial, aprendizado de máquina, processamento de linguagem natural e técnicas de ontologia, para produzir respostas proativas às ameaças cibernéticas em constante evolução. Portanto, conclui-se que este trabalho é relevante para o tema a ser pesquisado, visto que oferece uma vasta análise das ferramentas e dos resultados obtidos de estudos, que contribuíram para o avanço da pesquisa no campo de análise de conteúdos maliciosos na *Dark Web*.

3 Desenvolvimento

Este capítulo apresenta uma visão geral sobre as etapas adotadas para conduzir a pesquisa e alcançar os objetivos propostos do presente trabalho, e os detalhes sobre cada etapa desse processo.

3.1 Visão Geral

O método adotado para executar o trabalho proposto, com base em diversas alternativas observadas durante a revisão bibliográfica, está dividido entre as seguintes etapas:

1. Coleta e pré-processamento dos dados: extração de dados presentes em fóruns da *Dark Web*, com foco em *posts* potencialmente maliciosos, seguido por um processo de pré-processamento e classificação dos dados, realizados por um modelo classificador de *posts*.

2. Seleção de *posts* relevantes e vetorização TF-IDF: identificação dos *posts* maliciosos classificados com maior faixa de relevância, seguida da aplicação da vetorização TF-IDF, para converter os textos em representações numéricas.

3. Implementação de técnicas de aprendizado não supervisionado: agrupamento dos *posts* em *clusters*, por meio de algoritmos de clusterização, com base em semelhanças no conteúdo dos textos.

4. Visualização dos dados e definição do número de *clusters*: visualização da distribuição dos dados a partir de gráficos de dispersão, seguida de uma análise visual e da escolha da quantidade de grupos formados.

5. Análise e classificação dos *clusters*: identificação das palavras mais representativas de cada *cluster*, seguida da aplicação de LDA para extrair os tópicos principais, e a partir disso rotular os *clusters*, com a visão geral dos temas abordados.

6. Documentação dos resultados: apresentação e discussão dos resultados obtidos, com foco na análise sobre os tipos de atividades maliciosas presentes na *Dark Web*.

Todos os dados, códigos e implementações apresentados, bem como os demais materiais utilizados neste trabalho, estão disponíveis no endereço <https://github.com/annaletycia/malicious_posts_analysis> para consulta e acompanhamento.

3.2 Coleta e pré-processamento dos dados

A primeira etapa deste trabalho envolveu a análise de uma base de dados composta por *posts* coletados de diferentes fóruns da *Dark Web*, como o *Deep Answers*, *Hidden Answers* e *Raddle*, com conteúdos em português e inglês. Para as etapas seguintes da análise, foram selecionados apenas os *posts* em português dos fóruns *Hidden Answers* e *Deep Answers*, totalizando 11.224 *posts* para o estudo. Os *posts* selecionados fazem parte da coleta de dados realizada por JESUS FILHO (2023), por meio de *crawlers*, que percorrem e extraem dados automaticamente. Desse modo, o uso de uma base de informações vigente, composta por uma quantidade significativa de dados a respeito de atividades ilícitas e ameaças cibernéticas, possibilitou um foco maior nas etapas subsequentes.

Após a coleta, JESUS FILHO (2023) dividiu o pré-processamento dos dados em duas partes. A primeira envolveu a limpeza e a organização dos dados, além da remoção de inconsistências que comprometeriam o resultado da análise. Enquanto a segunda parte focou na eliminação de *stop words* e outros termos irrelevantes para a compreensão do conteúdo. Essa etapa foi importante para manter e considerar apenas os termos relevantes nas análises seguintes.

Enfim, os *posts* foram classificados com base na sua relevância, por um modelo de Processamento de Linguagem Natural. A rotulagem dos *posts* foi realizada por meio de duas abordagens diferentes. A primeira abordagem considerou a ocorrência simultânea de palavras-chave e IoCs, tais como e-mails, URLs e endereços IP. Enquanto a segunda abordagem, além de considerar a ocorrência de IoCs e palavras-chave, também incluiu uma análise manual, que levou em consideração o conteúdo das postagens e suas categorias (JESUS FILHO, 2023). Consequentemente, uma precisão maior nos dados processados foi garantida, de maneira que *posts* considerados relevantes foram rotulados como maliciosos e, portanto, com alto risco de envolvimento em atividades ilegais ou nocivas.

3.3 Seleção de *posts* relevantes e vetorização TF-IDF

Após a etapa de limpeza e preparação da base de dados, a fim de garantir a qualidade da análise, foi feita a seleção dos *posts* relevantes. Conforme o modelo supervisionado proposto por JESUS FILHO (2023), que visa classificar postagens com base em sua relevância, cada *post* recebeu um valor entre 0 e 1, de acordo com o grau de relevância do seu conteúdo. Esse modelo foi aplicado aos 11.224 *posts* obtidos e, a partir dos critérios estabelecidos por JESUS FILHO (2023), a prioridade recaiu sobre os 1.152 *posts* que apresentaram uma faixa de relevância superior ou igual a 0,7. Essa filtragem possibilitou um estudo mais eficaz das informações, visto que concentra nos conteúdos que realmente se destacam e têm maior probabilidade de gerar um impacto significativo nos resultados da pesquisa. Dessa forma, o trabalho utilizou dados como os apresentados na Figura 2 e

concentrou-se na análise do conteúdo dos *posts* que apresentam alta relevância. A partir disso, o processo analítico é otimizado, e as decisões são fundamentadas nos dados mais pertinentes e importantes.

Em seguida, foi realizada a vetorização por meio da técnica de TF-IDF, que converte os textos em representações numéricas, característico do processamento dos dados. Assim, quantificou-se a importância de cada termo em relação ao conjunto de documentos, o que possibilitou a identificação eficiente de padrões e relações implícitas entre diferentes *posts*, mediante a implementação de algoritmos de aprendizado não supervisionado. Portanto, após a conclusão da vetorização, os dados estão preparados para a aplicação de técnicas que auxiliam na descoberta de tendências e na segmentação de conteúdos relevantes

Ademais, funções em *Python* e a biblioteca Scikit-learn (2024), ferramentas conhecidas para manipulação e análise de dados, são utilizadas nas etapas de seleção e vetorização, devido à notável eficiência e desempenho. Como indicado na Figura 2, o *DataFrame* resultante da filtragem apresenta somente as linhas em que a coluna de "probabilidade" é maior ou igual a 0,7 e, consequentemente, de relevância alta. Finalmente, para o processo de vetorização, a coluna "*full_text*" foi extraída e convertida em uma lista de *strings*. Por meio da classe *TfidfVectorizer* da biblioteca Scikit-learn, o texto foi transformado em vetores numéricos. Optou-se por limitar o número de características a 1.000 palavras mais significativas, para equilibrar a representatividade dos dados e a eficiência do modelo, em uma tentativa de reduzir o ruído e a redundância.

	id	full_text	probabilidade	Relevância
0	7978	google drive mega virus pensado baixar livros ...	0.855241	Alta
1	7996	plataforma dados completos precisando platafor...	0.717149	Alta
2	8005	alternativas torum raidforums usava torum temp...	0.749300	Alta
3	8011	dicas anonimato queria dizer possivel ficar an...	0.734985	Alta
4	8031	links estudo hacking dia todos tempo atras nav...	0.863225	Alta
...	...	...	...	...
1147	2036	puxar dados deste cpf mim grupo telegram galer...	0.759585	Alta
1148	2038	ajuda obter dados completos inclusive redes so...	0.740496	Alta
1149	2039	poucos conseguem possivel encontrar saber onde...	0.721667	Alta
1150	2042	hackear celular partir wifi tempo atras alguma...	0.762532	Alta
1151	2066	robo preenche formularios diversos surface pre...	0.719943	Alta

1152 rows × 4 columns

Figura 2 – *DataFrame* resultante do processo de filtragem. (Fonte: a Autora)

A matriz resultante corresponde a uma representação numérica, na qual cada linha representa um *post* da base de dados selecionada, e cada coluna a uma palavra dentre as

1.000 mais significativas para o conjunto de textos. Logo, o conteúdo da matriz é composto pelos valores TF-IDF de cada termo em relação a um documento específico, fornecendo uma representação vetorial dos *posts*, onde cada valor indica a relevância de um termo na postagem.

3.4 Implementação de técnicas de aprendizado não supervisionado

Com os dados prontos e formatados devidamente, a etapa subsequente consiste na implementação de técnicas de aprendizado não supervisionado, ou seja, o agrupamento dos *posts* baseado na semelhança de seus conteúdos. Por meio dessa abordagem, é possível descobrir as relações entre os dados e identificar os diferentes temas apresentados nos textos. Para isso, é essencial o uso da biblioteca *Scikit-learn*, que possui ferramentas e algoritmos eficientes para a análise de dados, o que auxilia no esclarecimento do estudo das informações contidas nos *posts*.

Os algoritmos aplicados, K-means, KNN e DBSCAN, apresentam características e tratamentos diferentes para o agrupamento e análise dos dados. A escolha por algoritmos não supervisionados é fomentada pela vantagem da ausência da demanda de rótulos predefinidos, o que possibilita maior flexibilidade e adaptação de análise dos *posts* coletados. Nesse caso, a implementação de diversas técnicas almeja avaliar empiricamente qual método oferece o melhor desempenho no agrupamento dos dados, considerando a capacidade de lidar com a diversidade dos *clusters*, a detecção eficaz de *outliers*, e a qualidade da divisão dos grupos.

3.5 Visualização dos dados e definição do número de *clusters*

A etapa de visualização de dados é crucial para o entendimento da distribuição e disposição das informações nos *clusters* formados. Assim, de forma intuitiva, a visualização gráfica elucida como os dados se organizam no espaço vetorial e ilustra a relação de proximidade entre os grupos. Para esse fim, o uso de mecanismos de redução de dimensionalidade, como a Análise de Componentes Principais, do inglês *Principal Component Analysis* (PCA), possibilita representações bidimensionais dos vetores TF-IDF, e auxilia na exibição dos agrupamentos.

No presente trabalho, a implementação do PCA com dois componentes principais efetua a redução da dimensionalidade para duas dimensões. Consequentemente, os dados provenientes da matriz TF-IDF, de alta dimensionalidade, são projetados em um plano bidimensional, o que proporciona uma representação compreensível e simples através de um gráfico de dispersão 2D.

Outrossim, a definição do número de *clusters* não é feita com base em métodos

predefinidos, mas sim de maneira experimental. Portanto, a escolha da quantidade de agrupamentos parte da observação visual de como os dados são mais adequadamente agrupados nas projeções gráficas. Com diferentes visualizações e números variados de *clusters*, é possível constatar a configuração que resulta em uma segmentação mais clara e categórica entre as diferentes discussões de cunho malicioso presentes nos *posts*.

3.6 Análise e classificação dos *clusters*

Finalmente, com os *clusters* formados e os dados agrupados pelo algoritmo K-means, é a vez da etapa de análise e classificação, para enfim identificar os temas predominantes de cada grupo. O primeiro passo consiste na definição das palavras mais representativas de cada agrupamento, para uma perspectiva inicial do conteúdo presente neles. Para isso, os centroides gerados pelo algoritmo, que representam o centro de massa de cada *cluster* no espaço vetorial, juntamente com a matriz TF-IDF, que pondera a importância dos termos nos documentos, são combinados a fim de identificar os termos de maior valor nos centroides, conforme código exibido na Figura 3. No entanto, é importante ressaltar que a relevância de um termo não se limita ao seu peso no centroide, mas também à sua variabilidade entre os *clusters*, como já discutido por Manning, Raghavan e Schütze (2008).

```
centroids = kmeans.cluster_centers_

termos = tfidf_vectorizer.get_feature_names_out()

num_top_words = 10

for i in range(num_clusters):
    print(f"Cluster {i}:")
    top_indices = centroids[i].argsort()[-num_top_words:][::-1]
    top_termos = [termos[index] for index in top_indices]
    print(" ".join(top_termos))
    print("\n")
```

Figura 3 – Código para obter as palavras mais representativas de cada grupo. (Fonte: a Autora)

Em seguida, a implementação da técnica de LDA contribui na detecção dos tópicos latentes dentro de cada agrupamento. Essa abordagem permite uma análise mais complexa, visto que decompõe os textos em um conjunto de tópicos subjacentes e, consequentemente, evidencia as principais categorias temáticas. Logo, cada grupo de *posts* se associa a um ou mais assuntos e revela quais as discussões predominantes.

O último passo, após análise das palavras representativas e com base nos resultados da LDA, corresponde na rotulagem dos *clusters*. Logo, o rótulo designado para cada

cluster reflete uma visão geral do tema abordado nele, ou seja, resume a tendência dos conteúdos presentes. O resultado disso garante uma alternativa de categorização eficiente das atividades maliciosas existentes na *Dark Web*.

3.7 Documentação e análise dos resultados

À luz do desenvolvimento das etapas anteriores, os resultados alcançados são avaliados e explicados, com foco na interpretação dos *clusters* formados e na detecção das atividades maliciosas identificadas em cada grupo, a partir de análises visuais e conceituais.

Assim, o próximo capítulo apresenta uma compreensão clara sobre as discussões e os padrões de comportamento de cunho malicioso presentes em fóruns da *Dark Web*. Além disso, a relevância desses resultados no contexto da segurança cibernética são destacados, a fim de contribuir significativamente para futuras investigações e auxiliar na criação de estratégias de prevenção e mitigação de riscos.

4 Resultados

O propósito deste capítulo é exibir os resultados obtidos a partir dos experimentos realizados, mediante a execução das etapas descritas no Capítulo 3, além de mencionar descobertas interessantes que foram concebidas.

4.1 Base de Dados

Como mencionado no Capítulo 3, o conjunto de dados utilizado neste trabalho é composto por *posts* extraídos de dois fóruns da *Dark Web*: *Hidden Answers* e *Deep Answers*. Estes fóruns foram selecionados por conterem discussões sobre atividades ilícitas e conteúdo pertinente para análise de ameaças cibernéticas emergentes. A Tabela 1 ilustra uma visão geral do conjunto de dados, com informações sobre o número de *posts* coletados de cada fórum, o idioma em que foram escritos e a quantidade total de textos.

Tabela 1 – Especificações sobre os *posts* da base de dados extraídos de fóruns da *Dark Web*. (Fonte: a Autora)

Fórum	***Posts***	**Idioma**
Hidden Answers	10.526	Português do Brasil
Deep Answers	698	Português do Brasil
Total de *posts*:	**11.224**	

Com os dados já pré-processados e com base na classificação pré-atribuída para cada *post*, realiza-se a filtragem a fim de garantir que apenas os *posts* de alta relevância serão incluídos na análise. De modo semelhante, a Tabela 2 exibe os detalhes da nova base de dados pós-filtragem, com a quantidade total de 1.152 *posts* considerados relevantes e, logo, mais significativos e impactantes para os resultados da pesquisa.

Tabela 2 – Detalhes dos *posts* resultantes da filtragem de dados com alta relevância. (Fonte: a Autora)

Fórum	***Posts***	**Idioma**
Hidden Answers	1.056	Português do Brasil
Deep Answers	96	Português do Brasil
Total de *posts*:	**1.152**	

4.2 Definição do Número de *Clusters*

A definição do número de grupos para dividir os dados é uma etapa importante na aplicação de algoritmos de agrupamento. Esse processo envolve testar diferentes quan-

tidades de *clusters* para determinar qual configuração é apropriada para a estrutura dos dados. O objetivo principal é avaliar a eficiência dos grupos formados e identificar padrões temáticos entre os *posts*, observando como cada técnica de clusterização divide os dados e quais os temas predominantes de cada agrupamento. Para isso, é necessário realizar um comparativo, variando o número de *clusters*, para observar a distribuição e a densidade dos dados conforme cada configuração.

As Tabelas 3, 4 e 5 exibem os resultados do teste feito com o algoritmo K-means, contendo diferentes valores de k, que representa o número de *clusters* definidos para agrupar os dados. As tabelas incluem detalhes sobre a quantidade de *posts* atribuída a cada *cluster*, o que permite observar e comparar como a escolha de k afeta a distribuição dos dados. Por exemplo, na Tabela 3, com dois grupos ($k = 2$), há uma divisão onde o *cluster* 0 contém 263 *posts*, enquanto o *cluster* 1 apresenta 889, o que sugere uma distribuição desproporcional dos dados. Por outro lado, a Tabela 4 revela uma disposição mais uniforme ao definir $k = 3$, visto que o *cluster* 0 possui 657 *posts*, o *cluster* 1 conta com 259, e o *cluster* 2 contém 236, ou seja, possivelmente uma configuração que captura melhor os temas presentes em cada grupo. Por último, na Tabela 5, com $k = 4$, a distribuição se fragmenta com uma subdivisão adicional que gera um *cluster* significativamente menor (*cluster* 3).

Tabela 3 – Distribuição dos *posts* considerando dois *clusters*. (Fonte: a Autora)

Clusters	**Nº de Posts**
0	263
1	889

Tabela 4 – Distribuição dos *posts* considerando três *clusters*. (Fonte: a Autora)

Clusters	**Nº de *Posts***
0	657
1	259
2	236

Tabela 5 – Distribuição dos *posts* considerando quatro *clusters*. (Fonte: a Autora)

Clusters	**Nº de *Posts***
0	243
1	545
2	235
3	129

Com base nessas informações, a análise visual apresentada na seção 4.3 auxiliará na definição apropriada do valor de k.

4.3 Resultados Gráficos da Distribuição por Algoritmo

Para uma análise visual dos resultados dos agrupamentos, gráficos de dispersão são desenvolvidos a partir da redução da dimensionalidade por PCA, para avaliar as diferentes distribuições obtidas, mediante a implementação de algoritmos e números variados de *clusters*. Desse modo, é possível comparar o desempenho de cada técnica na organização dos dados e identificar quais temas predominam entre os *posts*.

Primeiramente, em continuação à análise dos *clusters* formados pelo agrupamento usando o algoritmo K-means, são testadas diferentes configurações de k, que geram divisões distintas e variados níveis de generalização ou especialização dos temas.

Conforme a Figura 4 indica, na configuração com dois *clusters*, há uma separação clara entre os dados, na qual os *posts* são agrupados em duas categorias amplas. Apesar da simplicidade, essa visão pode resultar em generalizações excessivas, visto que temas variados são combinados em poucos *clusters*.

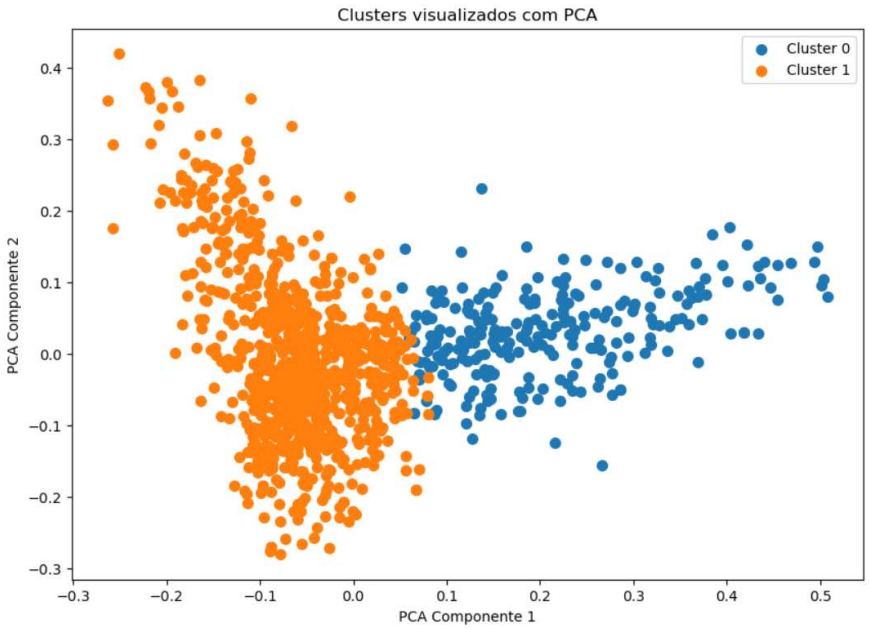

Figura 4 – Gráfico de distribuição dos *posts* usando K-means com dois grupos. (Fonte: a Autora)

Por outro lado, com três *clusters*, a separação aparenta ser mais equilibrada e clara. Na Figura 5, os grupos apresentam uma divisão nítida e organizada dos *posts*, o que indica uma boa capacidade de capturar os temas principais de cada grupo, sem fragmentar ou generalizar os agrupamentos excessivamente. Esta configuração parece ser apropriada para uma análise eficiente do conjunto de dados.

Quando o número de *clusters* é elevado para quatro, evidente na Figura 6, o modelo começa a identificar tópicos temáticos muito próximos, o que acarreta na sobreposição entre os *clusters* 0 e 3. Ou seja, esta configuração sugere que, com o aumento da quanti-

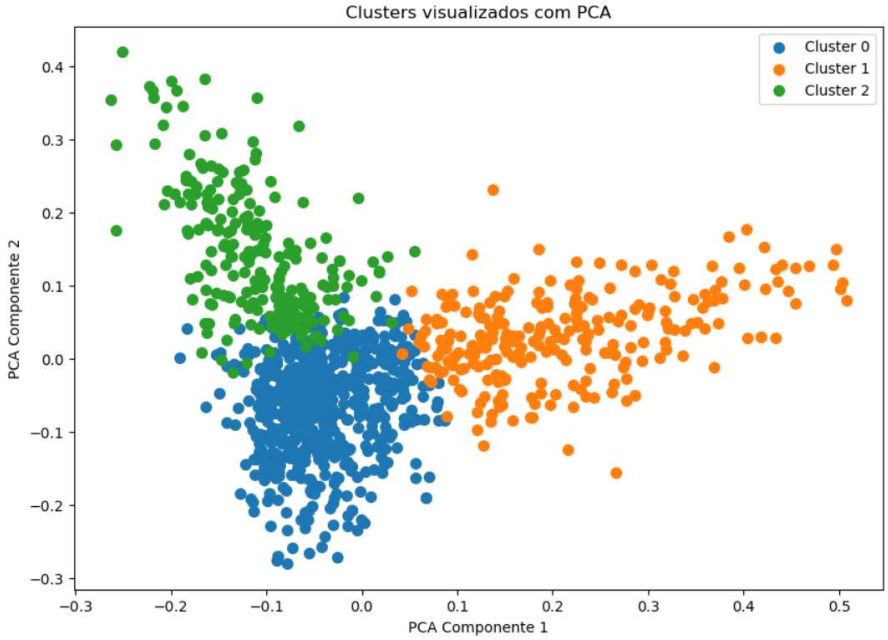

Figura 5 – Gráfico de distribuição dos *posts* usando K-means com três grupos. (Fonte: a Autora)

dade de grupos formados, a clareza na divisão dos dados reduz, o que consequentemente dificulta a interpretação dos agrupamentos.

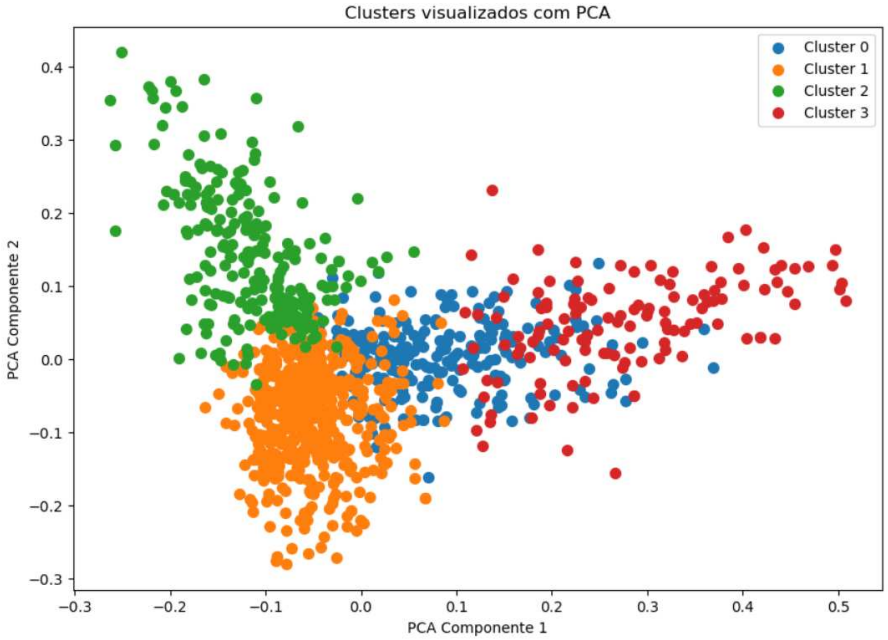

Figura 6 – Gráfico de distribuição dos *posts* usando K-means com quatro grupos. (Fonte: a Autora)

Para confirmar essa suspeita, é realizado um teste com cinco *clusters*. Porém, como presumido, a especificação demasiada dos dados compromete a clareza da visualização dos grupos de *posts* e prejudica a análise dos temas. A Figura 7 ilustra essa configuração, que

separa claramente certos grupos, como os *clusters* 1 e 4, mas que também gera certa mistura entre alguns grupos, especialmente os que se encontram no centro, como os *clusters* 0, 2 e 3.

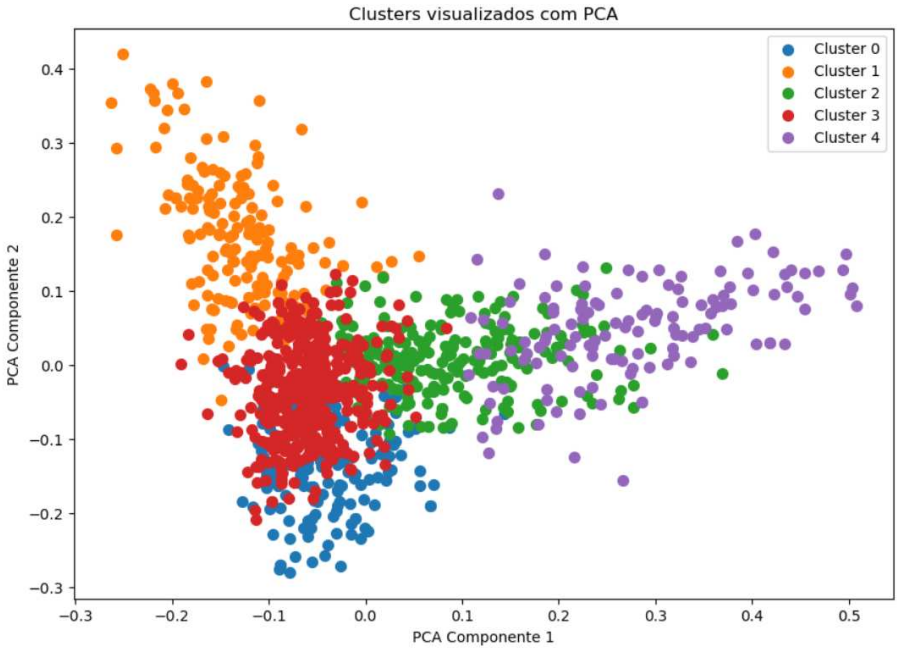

Figura 7 – Gráfico de distribuição dos *posts* usando K-means com cinco grupos. (Fonte: a Autora)

A fim de analisar e avaliar outras formas de divisão e identificação dos temas presentes nos *posts*, também foram testados os algoritmos KNN e DBSCAN, para fins de comparação em relação ao desempenho do K-means. Esse teste comparativo tem o intuito de entender como cada técnica agrupa os dados e distribui os *posts*, com base em seus diferentes tópicos temáticos.

No caso do KNN, o valor de *n_neighbors*=5 foi escolhido para garantir que o modelo capture padrões locais sem a segmentação instável que ocorre com valores muito baixos e, ao mesmo tempo, sem perder as nuances locais que geralmente acontecem com valores muito altos. A partir dessa configuração, a segmentação considera a proximidade local entre os *posts*, onde cada *post* é classificado com base nos cinco vizinhos mais próximos. Assim, a divisão se baseia nas relações diretas entre dados adjacentes, o que permite registrar padrões específicos e refletir similaridades locais com maior precisão. Desse modo, cada *post* influencia e é influenciado pelos *posts* próximos, resultando em agrupamentos que destacam as conexões entre dados.

Quando comparado ao K-means com configuração de $k = 3$, os gráficos de dispersão resultantes apresentam semelhanças no número de *clusters* e na distribuição dos *posts*. Apesar disso, o K-means se sobressai, visto que fornece um agrupamento mais uniforme e consistente, evidenciando temas de forma mais clara e coesa. Em contrapartida, o KNN,

conforme mostrado na Figura 8, revela uma segmentação menos estruturada e mais suscetível à disposição dos dados no espaço. Além disso, os pontos no gráfico apresentam um gradiente de cores, o que representa uma variação entre os *clusters*. Isso ocorre porque o KNN não estabelece centros de *clusters* bem definidos para a classificação dos dados, ao contrário do K-means, que agrupa os dados em torno de centroides, criando divisões nítidas e bem delimitadas entre os grupos.

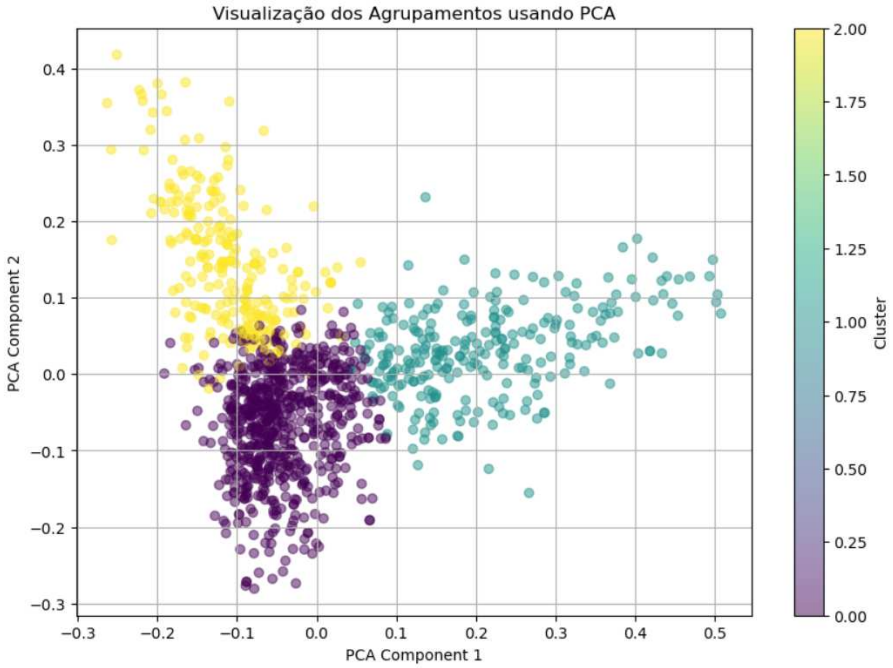

Figura 8 – Gráfico de distribuição dos *posts* usando KNN. (Fonte: a Autora)

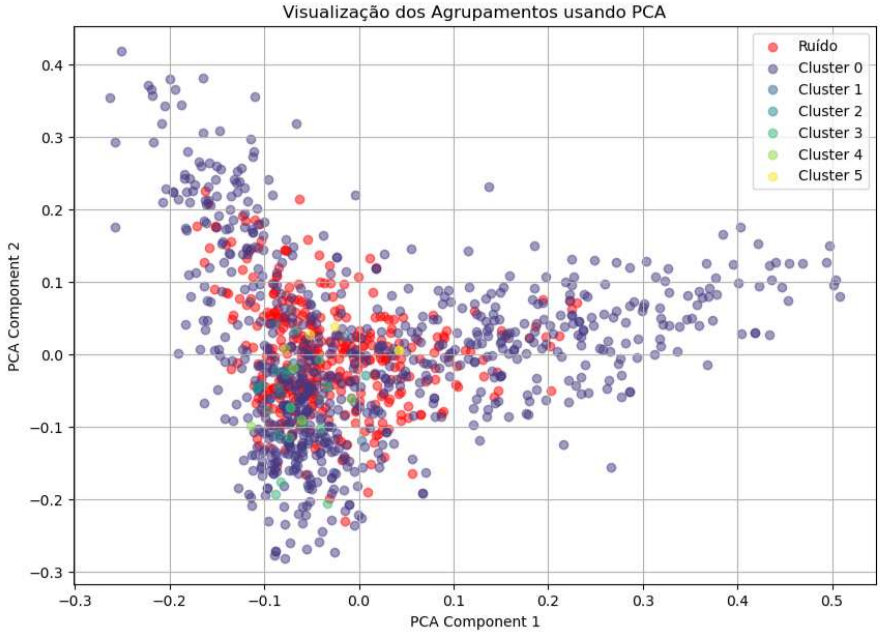

Figura 9 – Gráfico de distribuição dos *posts* usando DBSCAN. (Fonte: a Autora)

Por outro lado, o DBSCAN agrupa os *posts* com base na densidade, unindo áreas de alta concentração de dados. A configuração escolhida com $eps = 1.2$ e $min\_samples = 5$ tem como objetivo capturar áreas de alta concentração de *posts*, excluindo pontos isolados. Além disso, essa estrutura exige que cada *cluster* tenha pelo menos cinco pontos em sua vizinhança, auxiliando na classificação de pontos dispersos como ruído ou *outliers*. Ao contrário do K-means, o DBSCAN não define *clusters* de forma estrutural, o que o torna mais flexível em relação à variabilidade e ao ruído nos dados.

No gráfico de dispersão resultante, presente na Figura 9, o agrupamento pelo DBSCAN apresenta seis *clusters* sobrepostos, com dados desordenados e muitos pontos classificados como *outliers*. Comparado ao K-means, a clareza visual e a organização dos agrupamentos foram inferiores, o que reforça que, nesse caso, com a configuração $k = 3$, a abordagem usando K-means é a escolha mais eficaz para capturar os principais tópicos temáticos presentes nos *posts* da base de dados.

4.4 Rotulagem dos *Clusters*

Nesta seção, são discutidos os resultados obtidos na etapa de rotulagem dos *clusters*, a fim de identificar os temas predominantes nos grupos formados. Primeiramente, para esta etapa, foram examinadas as dez palavras mais representativas de cada *cluster*, extraídas a partir dos centros de cada grupo. Diferentes valores de k foram considerados ($k = 2$, $k = 3$ e $k = 4$) para verificar qual configuração realiza a melhor separação dos temas. O valor de k que proporcionar a divisão mais eficaz será escolhido para a etapa subsequente de análise e rotulagem dos conteúdos presentes em cada *cluster*.

A Tabela 6 demonstra que, com $k = 2$, as palavras representativas dos *clusters* obtidos exibem termos de temas misturados, o que resulta em uma baixa precisão na identificação das categorias temáticas. Já com $k = 4$, conforme exposto na Tabela 7, grupos diferentes, como o *cluster* 0 e o *cluster* 3, apresentam palavras-chave semelhantes, o que sugere uma fragmentação excessiva, onde temas próximos que podem ser combinados em um único grupo são separados.

Tabela 6 – As dez palavras mais representativas de cada *cluster*, identificadas pelo algoritmo K-means com dois grupos. (Fonte: a Autora)

Cluster	**Palavras-chave**
0	dados, cpf, telegram, número, nome, telefone, pessoa, cartão, puxar, informações
1	*hacking*, site, senha, dados, rede, hacker, hackear, obrigado, conta, acesso

Em contrapartida, a configuração de $k = 3$ permite uma coesão temática clara, sem mistura excessiva de assuntos em um único *cluster* ou fragmentação desnecessária

Tabela 7 – As dez palavras mais representativas de cada *cluster*, por meio do algoritmo K-means com quatro grupos. (Fonte: a Autora)

Cluster	**Palavras-chave**
0	dados, banco, cartao, site, vazamento, vazamentos, onde, sites, vazados, telegram
1	senha, rede, conta, site, acesso, hackear, tor, pessoa, phishing, social
2	*hacking*, curso, programacao, aprender, cursos, linux, hacker, obrigado, forum, estudar
3	cpf, numero, dados, telegram, nome, telefone, pessoa, puxar, consulta, informacoes

Tabela 8 – As dez palavras mais representativas de cada *cluster*, por meio do algoritmo K-means com três grupos. (Fonte: a Autora)

Cluster	**Palavras-chave**
0	senha, site, dados, rede, acesso, conta, hackear, pessoa, tor, phishing
1	dados, cpf, telegram, numero, nome, telefone, pessoa, cartao, puxar, informacoes
2	*hacking*, curso, programacao, aprender, hacker, cursos, forum, obrigado, links, estudar

de dados em grupos diferentes. Por isso, a Tabela 8 demonstra que essa estrutura facilita a interpretação dos *posts* presentes em cada *cluster*, e permite sugestões iniciais sobre as categorias dos tópicos principais de cada agrupamento identificado.

A seguir, com base na escolha da configuração com três *clusters*, a técnica de Análise de Tópicos Latentes (LDA) foi aplicada. A finalidade dessa aplicação é compreender a fundo os temas presentes em cada grupo, para uma identificação mais detalhada dos tópicos discutidos nos *posts*.

Figura 10 – Tópicos mais representativos por *cluster*, a partir da aplicação da LDA. (Fonte: a Autora)

Como ilustrado na Figura 10, a LDA analisa e estabelece relações entre as palavras mais frequentes em cada *cluster*. Optou-se pela definição de cinco tópicos por *cluster*, a fim de fornecer uma visão ampla das temáticas presentes, sem dividir os dados em excesso para manter a interpretação e visualização simples. Desse modo, os tópicos são enumerados de 0 a 4, onde cada linha representa um tópico específico identificado dentro de um *cluster*. Cada tópico é composto por uma sequência de palavras associadas a um peso, que indica a relevância da palavra no contexto do tópico. Assim, os temas centrais de cada *cluster* são destacados, facilitando a rotulagem e interpretação dos conteúdos com base nas palavras mais significativas e suas respectivas relevâncias.

Finalmente, após a conclusão das etapas anteriores, é possível realizar a rotulagem dos *clusters*, fundamentada na análise dos temas obtidos nos *posts* de cada grupo. Desse modo, a interpretação dos tópicos de cada *cluster*, com o objetivo de compreender os seus conteúdos específicos, será discutida nas subseções 4.4.1, 4.4.2 e 4.4.3.

4.4.1 *Cluster* 0: Segurança de Dados e Contas

Os *posts* presentes no *cluster* 0 discutem temas relacionados à segurança de dados e contas, e proteção de sites e senhas. Alguns exemplos de palavras-chave são “site”, “dados” e “senhas”, o que evidencia um foco na segurança digital. Ademais, termos como “hackear” e “conta” refletem uma preocupação com o *hacking* de contas pessoais. O *cluster* também explora tópicos de privacidade e anonimato online, apresentando palavras como “rede”, “segurança” e “tor”, ou seja, temas de proteção de identidade na rede.

4.4.2 *Cluster* 1: Dados Pessoais e Identidade

O *cluster* 1 abrange *posts* que mencionam informações pessoais e dados sensíveis, como “cpf”, “nome”, “número” e “telefone”. Focadas na coleta e uso de informações de identificação de indivíduos, as discussões abordam sobre como e onde esses dados podem ser obtidos ou compartilhados. Logo, o interesse central dos *posts* desse grupo concentra nos meios de obter acesso a informações confidenciais e na compreensão de como esses dados circulam.

4.4.3 *Cluster* 2: Educação e Comunidade em *Hacking*

Já no *cluster* 2, *posts* sobre o aprendizado e a formação de uma comunidade em *hacking* são predominantes. Esse agrupamento reúne termos-chave como “*hacking*”, “hacker”, “aprender”, “curso” e “estudar”, que indicam interesse na educação e desenvolvimento de habilidades na área de segurança cibernética e *hacking*. Além disso, aspectos de cooperação e suporte dentro de uma comunidade hacker são expressos em palavras como “fórum”, “comunidade”, “ajuda” e “obrigado”, sugerindo uma colaboração e troca de conhecimento

entre os membros. Não obstante, o interesse não se restringe apenas em conhecimento técnico, mas também cultural e ético dentro do contexto hacker. Isso fica evidente em *posts* que contêm termos como "*Anonymous*", tema relacionado a movimentos ativistas *online*, que lutam pela privacidade e liberdade de informação.

4.5 Exemplos de *posts* por *clusters*

Nesta seção, exemplos de *posts* extraídos do fórum *Hidden Answers* são apresentados para ilustrar a conformidade do conteúdo de cada *post* com os rótulos dos *clusters* definidos anteriormente. Os exemplos a seguir reforçam a classificação dos dados, alcançada mediante análise de tópicos.

A Figura 11 exibe um *post* correspondente ao *Cluster* 0, que trata sobre temas de *hacking* e segurança de dados e contas. O conteúdo presente no *post* relata sobre a experiência de um usuário hackeando um site através de uma falha de segurança, e tentando manter o controle remoto do sistema por meio de uma injeção de *malware* em um computador. Discussões sobre práticas de *hacking*, comprometimento de segurança de contas e vulnerabilidades de sistemas são temas característicos deste *cluster*, e que são evidentes no *post* devido à presença de termos como "hackeei", "*malware*" e "XSS".

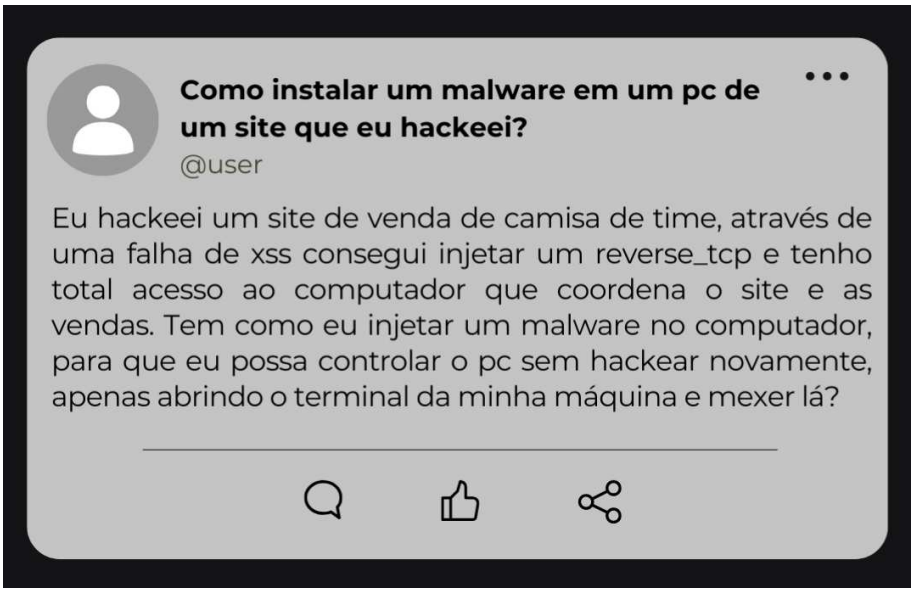

Figura 11 – Exemplo de *post* extraído da base de dados e presente no *Cluster* 0. (Fonte: a Autora)

Por outro lado, o *post* presente na Figura 12 é um exemplo de assunto abordado no *Cluster* 1, que foca em informações pessoais e dados sensíveis. Nesse caso, o usuário demonstra interesse em encontrar bancos de dados e informações confidenciais de instituições ou indivíduos de alto poder aquisitivo. Palavras como "contas bancárias", "informa-

ções pessoais" e "banco de dados" sugerem a busca por dados privados, frequentemente relacionados ao roubo e compartilhamento de informações pessoais.

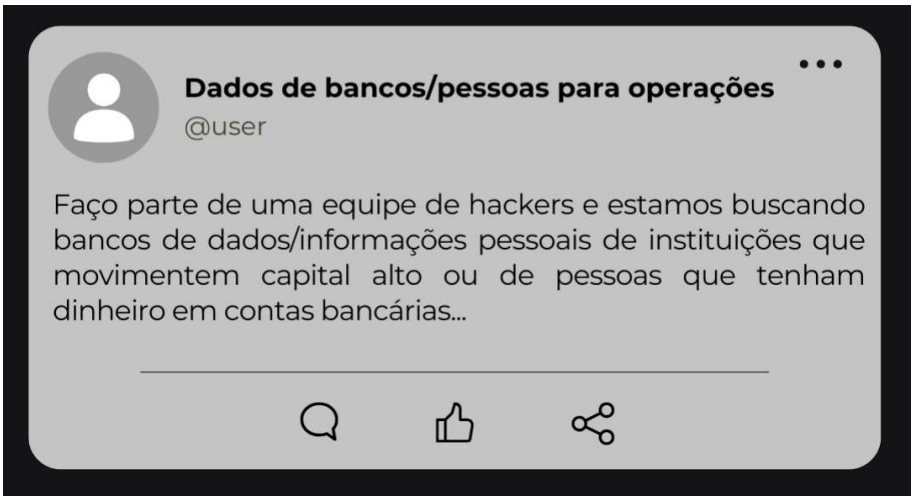

Figura 12 – Exemplo de *post* extraído da base de dados e presente no *Cluster* 1. (Fonte: a Autora)

Já a Figura 13 traz um exemplo um conteúdo que se alinha ao *Cluster* 2, focado em tópicos sobre educação e comunidade hacker. O *post* menciona a criação de um fórum voltado para a divulgação de conhecimento sobre *hacking* e tecnologias. A partir de termos como "aprender", "comunidade", "conhecimento", "*hacking*" e "fórum", é possível deduzir que o objetivo central é a construção de uma comunidade colaborativa, para troca de informações e suporte mútuo entre os integrantes.

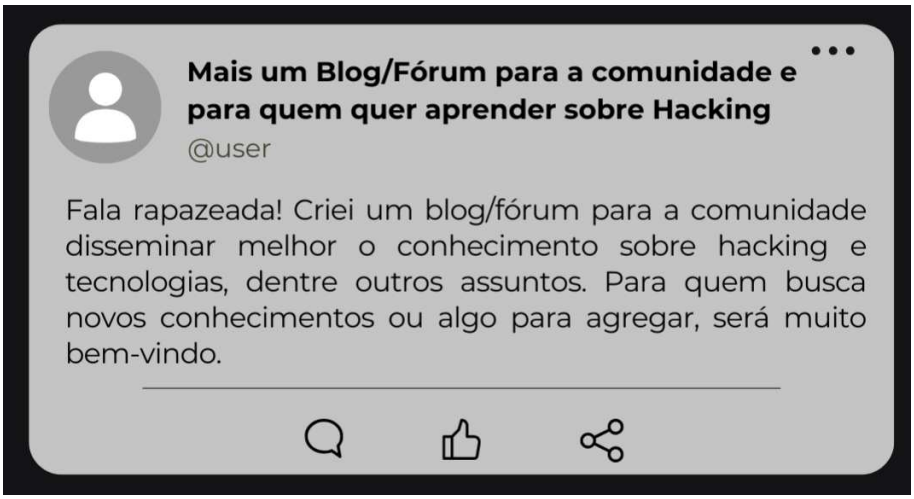

Figura 13 – Exemplo de *post* extraído da base de dados e presente no *Cluster* 2. (Fonte: a Autora)

Portanto, a partir desses exemplos, é comprovado que os *posts* de cada *cluster* representam as características e tópicos predominantes identificados em cada grupo. Isso

demonstra que os agrupamentos obtidos classificam com eficiência os temas discutidos em cada *cluster*.

5 Conclusão

Em face da constante evolução tecnológica e do aumento correspondente de ataques cibernéticos, surge a necessidade de estratégias e ferramentas eficazes para identificar, classificar e mitigar ameaças digitais, principalmente em ambientes ocultos, como a *Dark Web*, onde atividades ilícitas são recorrentes. Diante desse cenário propício a práticas ilegais, o objetivo do presente trabalho foi contribuir para o desenvolvimento de um método, por meio de técnicas de agrupamento e análise de conteúdo, para analisar e categorizar diferentes tipos de ameaças cibernéticas em *posts* extraídos da *Dark Web*.

Desse modo, foi proposta uma metodologia que permite segmentar e investigar os temas recorrentes da base de dados, extraída de dois fóruns da *Dark Web*. Para isso, foram aplicados algoritmos de agrupamento, como K-means, DBSCAN e KNN, a fim de identificar as similaridades entres os *posts* e organizá-los em grupos. Além disso, o modelo de LDA foi utilizado para detectar tópicos latentes e revelar padrões nos conteúdos. Com base nos resultados obtidos, o K-means se destacou pela sua eficácia em agrupar os dados em três *clusters* de maneira estruturada, o que possibilitou a rotulagem de cada grupo. Por meio da realização dessas etapas, conclui-se que temas como segurança de dados, busca de informações sensíveis e comunidade em *hacking*, são categorias de ameaças cibernéticas predominantes em postagens da *Dark Web*.

Para trabalhos futuros, a inclusão de novas fontes de dados provenientes de outros ambientes, como a *Deep Web*, a *Surface Web* e as redes sociais, pode auxiliar na identificação e resposta a esses riscos emergentes com antecedência. Além disso, a implementação de algoritmos mais avançados de aprendizado profundo, como redes neurais especializadas em processamento de linguagem natural, e a integração de ferramentas de monitoramento em tempo real facilitarão uma categorização ainda mais precisa e detalhada das ameaças cibernéticas, oferecendo atualizações contínuas e automáticas.

Referências

AKHGAR, B.; GERCKE, M.; VROCHIDIS, S.; GIBSON, H. **Dark Web Investigation**. Springer, 2021. 3–26 p. Disponível em: <https://doi.org/10.1007/978-3-030-55343-2>. Citado 2 vezes nas páginas 16 e 17.

ARTHUR, D.; VASSILVITSKII, S. **k-means++: The advantages of careful seeding**. [S.l.], 2006. Citado na página 17.

BASHEER, R.; ALKHATIB, B. Threats from the dark: a review over dark web investigation research for cyber threat intelligence. **Journal of Computer Networks and Communications**, Hindawi Limited, v. 2021, p. 1–21, 2021. Disponível em: <https://doi.org/10.1155/2021/1302999>. Citado 3 vezes nas páginas 11, 16 e 20.

_____. Conceptualizing discussions on the dark web: An empirical topic modeling approach. **Complexity**, Hindawi, v. 2024, p. 2775236, 2024. ISSN 1076-2787. Disponível em: <https://doi.org/10.1155/2024/2775236>. Citado na página 20.

CHENG, L.; YANG, Y.; ZHAO, K.; GAO, Z. Research and improvement of tf-idf algorithm based on information theory. In: LIU, Q.; MISIR, M.; WANG, X.; LIU, W. (Ed.). **The 8th International Conference on Computer Engineering and Networks (CENet2018)**. Cham: Springer International Publishing, 2020. p. 608–616. ISBN 978-3-030-14680-1. Citado na página 19.

CHOWDHARY, K. R. Natural language processing. In: **Fundamentals of Artificial Intelligence**. New Delhi: Springer India, 2020. p. 603–649. ISBN 978-81-322-3972-7. Disponível em: <https://doi.org/10.1007/978-81-322-3972-7_19>. Citado na página 19.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 18.

CRAIGEN, D.; DIAKUN-THIBAULT, N.; PURSE, R. Defining cybersecurity. **Technology innovation management review**, v. 4, n. 10, p. 13–20, 2014. Citado na página 14.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Citado na página 18.

GENTLEMAN, R.; CAREY, V. J. Unsupervised machine learning. In: **Bioconductor Case Studies**. New York, NY: Springer New York, 2008. p. 123–157. ISBN 978-0-387-77240-0. Disponível em: <https://doi.org/10.1007/978-0-387-77240-0_10>. Citado na página 17.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. Springer, 2009. 485–585 p. Disponível em: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>. Citado na página 17.

IBM. **Threat Intelligence**. 2023. <https://www.ibm.com/reports/threat-intelligence>. [Acesso em; 24/02/2024]. Citado na página 11.

IBRAHIM, A.; THIRUVADY, D.; SCHNEIDER, J.-G.; ABDELRAZEK, M. The challenges of leveraging threat intelligence to stop data breaches. **Frontiers in Computer Science**, v. 2, p. 1–10, 2020. ISSN 2624-9898. Disponível em: <https://www.frontiersin.org/articles/10.3389/fcomp.2020.00036>. Citado na página 15.

International Telecommunication Union (ITU). **Overview of Cybersecurity**. Geneva, 2008. Disponível em: <http://www.itu.int/rec/T-REC-X.1205-200804-I/en>. Citado na página 15.

JESUS FILHO, S. A. d. **Identificação de posts maliciosos na dark web utilizando Aprendizado de Máquina Supervisionado**. Dissertação (Mestrado) — UFU, 2023. Disponível em: <https://doi.org/10.14393/ufu.di.2023.8127>. Citado 3 vezes nas páginas 12, 20 e 23.

JO, H.; LEE, Y.; SHIN, S. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. **Computers & Security**, Elsevier, v. 120, p. 102763, 2022. Disponível em: <https://doi.org/10.1016/j.cose.2022.102763>. Citado na página 15.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. [S.l.]: MIT Press, 2008. 109–135 p. Citado 2 vezes nas páginas 18 e 26.

MOREIRA, V. P. Recuperação de informação. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. BPLN, 2024. book chapter 19. ISBN 978-65-00-95750-1. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-ir/cap-ir.html>. Citado na página 19.

NAZAH, S.; HUDA, S.; ABAWAJY, J. H.; HASSAN, M. M. An unsupervised model for identifying and characterizing dark web forums. **IEEE Access**, v. 9, p. 112871–112892, 2021. Disponível em: <https://ieeexplore.ieee.org/document/9509025>. Citado 2 vezes nas páginas 12 e 20.

_____. An unsupervised model for identifying and characterizing dark web forums. **IEEE Access**, v. 9, p. 112871–112892, 2021. Citado na página 18.

NLTK. **Natural Language Toolkit**. 2023. <https://www.nltk.org/>. [Online; accessed 27-Fevereiro-2024]. Citado na página 12.

OGLOBO. **Brasil é o maior alvo de ataques cibernéticos na América Latina**. 2023. <https://oglobo.globo.com/google/amp/economia/tecnologia/noticia/2023/06/brasil-e-o-maior-alvo-de-ataques-ciberneticos-na-america-latina-veja-ranking.ghtml>. [Acesso em; 24/02/2024]. Citado na página 11.

PYTHON. **Welcome to Python**. 2024. <https://www.python.org/>. [Online; accessed 27-Fevereiro-2024]. Citado na página 12.

RAHMAN, M. R.; HEZAVEH, R. M.; WILLIAMS, L. What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 55, n. 12, p. 16–36, mar 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3571726>. Citado 2 vezes nas páginas 12 e 15.

Scikit-learn. **scikit-learn: Machine Learning in Python**. 2024. <https://scikit-learn.org/stable/>. [Online; accessed 27-Fevereiro-2024]. Citado 2 vezes nas páginas 12 e 24.

SHACKLEFORD, D. Who's using cyberthreat intelligence and how? **SANS Inst.**, North Bethesda, MD, USA, p. 3–12, 2015. Citado na página 15.

SHIRKY, C. **Cognitive Surplus: Creativity and Generosity in a Connected Age**. [S.l.]: Penguin Press, 2010. 242 p. Citado na página 11.

STALLINGS, W.; BROWN, L. **Segurança de computadores. Princípios e Práticas**. 2. ed. Rio de Janeiro: Elsevier Editora, 2014. 7–11 p. Citado na página 14.

SUN, N.; DING, M.; JIANG, J.; XU, W.; MO, X.; TAI, Y.; ZHANG, J. Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. **IEEE Communications Surveys & Tutorials**, v. 25, n. 3, p. 1748–1774, 2023. Citado na página 15.

TAHERDOOST, H. Cybersecurity vs. information security. **Procedia Computer Science**, v. 215, p. 483–487, 2022. ISSN 1877-0509. 4th International Conference on Innovative Data Communication Technology and Application. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050922021214>. Citado na página 14.

TOUNSI, W.; RAIS, H. A survey on technical threat intelligence in the age of sophisticated cyber attacks. **Computers & Security**, Elsevier, v. 72, p. 212–233, 2018. Citado na página 15.

WONG, P. C. Unsupervised machine learning. In: DAVIER, A. A. von; MISLEVY, R. J.; HAO, J. (Ed.). **Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python**. Cham: Springer International Publishing, 2021. p. 173–193. ISBN 978-3-030-74394-9. Disponível em: <https://doi.org/10.1007/978-3-030-74394-9_10>. Citado na página 18.

XYAN, C. W.; WEI, J. L.; JUREMI, J. An exploration into the dark web. **Journal of Applied Technology and Innovation**, v. 4, n. 4, p. 10–13, 2020. Citado na página 16.