PUC Minas
A B A K Ó S
Instituto de Ciências Exatas e Informática

Proposta de Ferramenta para Otimizar a Extração de Dados em Boletins de Ocorrência: Um Estudo Piloto na Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Estado do Pará*

Data Extraction Optimization Tool from Police Reports: a Case Study in the Diretoria Estadual de Combate a Crimes Cibernéticos of Polícia Civil do Estado do Pará

Sainy Gabriel Dias Rosa Antonio[1]
Felipe da Costa Giestas[2]
Fantiny Santos dos Santos[3]
Emmanuelle Pantoja Silva[4]
Renato Hidaka Torres[5]

Resumo

Motivado pelo incentivo ao desenvolvimento de soluções inovadoras para combater os crimes praticados pela internet, denominados crimes cibernéticos, este artigo apresenta a proposta de uma ferramenta para extrair dados em boletins de ocorrência de forma automática. A ferramenta foi desenvolvida a partir da teoria de linguagens formais e construção de expressões regulares. A eficiência da ferramenta foi avaliada utilizando uma base de dados rotulada e calculando os valores das métricas de precisão e revocação. Conclui-se que a ferramenta é eficiente, apresentando taxas de precisão e revocação superiores a 0.9, conforme o estudo piloto, demonstrando as possibilidades de informações e padrões que a ferramenta proporciona para a investigação policial.

Palavras-chave: Crimes cibernéticos. Vazamento de dados. Ferramenta de extração de dados. Teoria de linguagens regulares. Boletins de ocorrência.

*Submetido em 10/05/2024 - Aceito em 07/02/2025
[1]Universidade Federal do Pará, Brasil- sainygabriel08@gmail.com.
[2]Universidade Federal do Pará, Brasil- felipe.giestas@ifch.ufpa.br.
[3]Universidade Federal do Pará, Brasil- fantiny.santos@gmail.com.
[4]Universidade Federal do Pará, Brasil- emmanuellepantojas@gmail.com.
[5]Universidade Federal do Pará, Brasil- renatohidaka@ufpa.br.

Abstract

Driven by the incentive to develop innovative solutions in the fight against cybercrime, this article proposes a tool to automatically extract data from police reports. The tool was based on the foundations of formal language theory and implemented using regular expressions. The tool's performance was assessed using a labeled dataset by computing the precision and recall metrics. The pilot study demonstrated the tool's efficiency, with precision and recall rates exceeding 0.9, showcasing its potential to uncover valuable information and patterns for police investigations.

Keywords: Cybercrimes. Data leak. Data extraction tool. Theory of regular languages. Police reports.

1 INTRODUÇÃO

O avanço da computação pervasiva está contribuindo para que a população brasileira tenha acesso facilitado a informações e serviços online. Isso pode ser constatado, uma vez que, segundo dados da pesquisa nacional por amostra de domicílios (Pnad) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em 2021, a internet chegou em 90% dos domicílios do país, sendo o telefone celular o principal equipamento de acesso à internet em mais de 99% das casas (NERY; BRITTO, 2022). Ao mesmo tempo em que a computação pervasiva trouxe benefícios para a população, também trouxe um ponto que é motivo de preocupação, que é o aumento do número de crimes cometidos a partir da internet, denominados como crimes cibernéticos.

Aliado ao crescimento da conectividade à Internet e do avanço da tecnologia dos smartphones, também há o crescimento na utilização das redes sociais. O uso dessas plataformas proporciona às pessoas a oportunidade de compartilhar suas experiências e informações. Por outro lado, esse aumento constante no acesso a elas abre espaço para possíveis ataques à privacidade, o que pode resultar no comprometimento das contas dos usuários e no roubo de informações sensíveis por parte de adversários mal-intencionados (MAJEED; KHAN; HWANG, 2022). Essas práticas tornam o ambiente propício ao cometimento de crimes praticados no meio cibernético.

Com base nos dados do relatório técnico da FORTINET (2022), uma renomada organização de pesquisa e inteligência de ameaças associada ao Fortinet Training Institute, dedicada à promoção da segurança cibernética, destaca-se que, ao longo do ano de 2022, o Brasil registrou a segunda maior incidência de tentativas de ataques cibernéticos na região da América Latina e Caribe, totalizando 88,5 bilhões de incidentes. Esse dado coloca o Brasil em segundo lugar, logo atrás do México, que liderou com 156 bilhões de tentativas de ataques do mesmo gênero (FORTINET, 2022; FORTINET, 2023). Em relação aos ataques que resultam em vazamentos de dados, segundo o relatório técnico da SURFSHARK (2022), no cenário mundial, o Brasil é o décimo segundo país que mais contabilizou vazamento de dados no primeiro trimestre de 2022. Entre janeiro e março, 285 mil brasileiros ficaram expostos, vítimas de crimes cibernéticos (SURFSHARK, 2022).

Os vazamentos de dados, além de causar prejuízos psicológicos para as vítimas, uma vez que fere o princípio da dignidade da pessoa humana, também podem causar prejuízos econômicos. Segundo dados do relatório técnico da IBM SECURITY (2023), no Brasil, em 2022, o prejuízo econômico em decorrência de violação de dados foi de US$ 1,38 milhões. O relatório também destaca que não apenas no Brasil como no mundo, o custo de uma violação de dados para as organizações atingiu o seu nível máximo em comparação com os anos anteriores (IBM SECURITY, 2023).

Dada a relevância do crime cibernético no cenário nacional, o Ministério da Justiça e Segurança Pública lançou, em 2022, o Plano Tático de Combate a Crimes Cibernéticos deno-

minado PTat-C3 (BRASIL, 2022), com o objetivo de ser mais um esforço de aprimoramento da integração e articulação interinstitucional para mitigação, prevenção e repressão aos crimes cibernéticos no país. Entre as ações estratégicas previstas no PTat-C3, está a de incentivar a concepção de soluções inovadoras em segurança cibernética. Também consta que o desenvolvimento de ferramentas que aprimorem o processo de investigação de crimes cibernéticos é um avanço importante para a concepção de soluções inovadoras (BRASIL, 2022).

Dessa forma, o presente artigo tem como objetivo propor e analisar a aplicação de uma ferramenta de extração de informação, fundamentada na teoria de linguagens regulares, que apresenta como finalidade aprimorar o processo de extração e análise das informações de Boletins de Ocorrência (BO). O presente estudo justifica-se por compreender que no processo de investigação criminal, o boletim de ocorrência é um instrumento administrativo que goza de presunção de veracidade. Logo, as informações e fatos descritos nos BOs são importantes para determinar o curso das investigações.

No contexto dos crimes cibernéticos, um investigador, para realizar uma análise criminal das infrações penais registradas no boletim de ocorrência (BO), a fim de identificar o comportamento do criminoso e da vítima, necessita ler esse documento, e perceber evidências de dados que podem ser utilizados no curso da investigação, por exemplo: e-mail, chave PIX, endereço eletrônico, número de telefone, conta bancária, valor monetário, plataforma digital e número de CPF.

Para otimizar a percepção das evidências do investigador, as informações citadas podem ser extraídas automaticamente, a partir da abordagem de reconhecimento léxico utilizada na ferramenta proposta neste artigo. O processo de extração desses dados parte da hipótese de que, para cada tipo de dado a ser extraído, existe um padrão que pode ser determinado por expressões regulares. Segundo Gersting (2001), na Ciência da Computação, as expressões regulares equivalem a máquinas de estados que podem ser construídas para determinar o reconhecimento de linguagens regulares. Para testar a hipótese da pesquisa, este artigo também apresenta um estudo piloto na Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Estado do Pará.

2 MÉTODO

Este estudo trata de uma pesquisa aplicada, em que é considerado o contexto e problemas de instituições, grupos, organizações e outros. Para FLEURY, M. T. L.; WERLANG, S. R. da C. (2016, p.11), "ela está empenhada na elaboração de diagnósticos, identificação de problemas e busca de soluções. Responde a uma demanda formulada por clientes, atores sociais ou instituições". Para o problema de pesquisa foram utilizadas técnicas quantitativas, com o intuito de mensurar, por meio de técnicas estatísticas, os fenômenos estudados (GIL, 2002). E quanto aos objetivos, utilizam-se técnicas exploratórias e descritivas, a fim de se aproximar do objeto de estudo e descrever suas possibilidades (RAUPP; BEUREN, 2006).

Quanto aos procedimentos técnicos, optou-se pela combinação de procedimentos, quais foram: (A) uma pesquisa de campo, em que foram reunidas informações a partir de um grupo de participantes, os quais fazem parte do contexto analisado; e, um estudo documental, pois foram analisados documentos oficiais que não receberam tratamento analítico, que podem ser diários, jornais, ofícios (LUNETTA; GUERRA, 2023). No caso desta pesquisa, foram utilizados boletins de ocorrência, sendo, portanto, considerados de fonte primária, já que não receberam nenhuma espécie de análise.

2.1 Lócus e Participantes

Para pesquisa de campo foram convidados a participar investigadores da Polícia Civil do Pará, que tinham contato direto com os boletins de ocorrências das Delegacias Especializadas, qual seja, a Divisão de Combate a Crimes Contra Direitos Individuais por Meios Cibernéticos, Divisão de Combate a Crimes Contra Grupos Vulneráveis Praticados por Meios Cibernéticos e Divisão de Combate a Crimes Econômicos e Patrimoniais Praticados por Meios Cibernéticos, ambos correspondem à Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Estado do Pará. Os participantes foram selecionados por conveniência, ressaltando que as informações fornecidas são confidenciais e direcionadas à pesquisa.

2.2 Procedimentos de Coleta e Aspectos Éticos

A coleta de dados ocorreu de forma verbal em companhia dos investigadores, em que foram anotadas as demandas envolvidas em um processo de investigação e o relato de informações. Nesse caso, os investigadores informaram que liam os boletins de ocorrência e extraíam informações-chave utilizadas nas investigações, elementos analisados nesta pesquisa. Vale ressaltar que os investigadores aceitaram participar da pesquisa, sendo-lhes informados que a pesquisa ocorre de forma anônima e sua utilização destina-se para fins acadêmicos.

Na fase documental, os dados utilizados na pesquisa são boletins de ocorrência da Polícia Civil do estado do Pará. Devido ao critério de confidencialidade, esses boletins de ocorrência não podem ser disponibilizados para o público em geral. Os documentos coletados são do período de 2018 a 2021, em que foram utilizados 5.858 boletins de ocorrência, e um conjunto de informações que normalmente são utilizadas no curso da investigação foram extraídas utilizando a ferramenta de extração de informação proposta nesta pesquisa. As seguintes classes de informações foram extraídas: valor monetário; chave PIX; aplicativos; agência e conta bancária; nome de banco.

A partir da extração, a equipe de investigação pode produzir informações estratégicas para responder perguntas que antes só seriam possíveis a partir da leitura individual dos BOs. Ao perceber essas respostas de forma eficiente para um grande volume de boletins de ocorrência, há a possibilidade de identificação de padrões que corroboram para uma investigação preventiva.

O empenho em realizar intervenções preventivas encontra respaldo no Plano Nacional de Segurança Pública e Defesa Social (2021-2030), uma iniciativa apresentada pelo Ministério da Justiça e Segurança Pública. No âmbito da "Ação estratégica 2" desse plano, a meta é efetuar a implementação de programas e projetos que facilitem a execução de ações tanto preventivas quanto repressivas. Essas ações buscam uma articulação eficaz com outros setores, sejam eles públicos ou privados, com o propósito de contribuir para a redução de crimes e conflitos sociais (BRASIL, 2023).

Sendo assim, a partir de uma proposta de identificação de padrões nos boletins de ocorrência da Polícia Civil do Estado do Pará por meio de uma ferramenta de extração de dados, o presente artigo divide-se nas seguintes seções: Contextualização do problema, Fundamentação da expressão regular, Desenvolvimento da ferramenta, Análise da eficiência e Estudo piloto.

2.3 Procedimentos de Análise

Na análise da eficiência da ferramenta optou-se pela utilização da matriz de confusão, bem como as métricas de desempenho de precisão e revocação (TORRES; OHASHI; PESSIN, 2019). Ao utilizar as métricas de precisão e revocação, é possível observar o desempenho da ferramenta de extração de dados objetivando a minimização de dados falsos positivos e falsos negativos. Para a análise dos dados extraídos no estudo piloto, optou-se por técnicas estatísticas de análise de frequência, bem como a análise de conectividade em grafos.

2.4 Contextualização do Problema

Na Polícia Civil do Estado do Pará, os boletins de ocorrência são armazenados em banco de dados relacional. Um banco de dados relacional é formado por tabelas, o que significa que os relatos dos boletins de ocorrência são instâncias de uma coluna de alguma tabela do banco de dados. Por ser uma instância, ao utilizar a linguagem Structured Query Language (SQL) para realizar buscas de dados, é necessário especificar o filtro da busca a partir da especificação da coluna e um valor de referência. Por exemplo, considerando a coluna BO, as seguintes buscas podem ser realizadas nos boletins de ocorrência:

(A) Quais boletins de ocorrência possuem em seu relato o telefone 99222-3333?

(B) Quais boletins de ocorrência possuem em seu relato os seguintes valores monetários: [R$ 1000.00, R$ 10000.00, R$ 100000.00]?

(C) Quais boletins de ocorrência possuem chave PIX e-mail do provedor Gmail?

Contudo, devido à necessidade da especificação de um valor de referência, as seguintes buscas não podem ser realizadas:

(D) Quais boletins de ocorrência possuem algum telefone em seu relato?

(E) Quais boletins de ocorrência possuem algum valor monetário em seu relato?

(F) Quais boletins de ocorrência possuem alguma chave PIX em seu relato e quais são duplicadas?

Essas últimas buscas exemplificadas só seriam possíveis em linguagem SQL utilizando expressões regulares. A utilização de expressão regular requer um conhecimento técnico que nem sempre é do conhecimento de profissionais que não são da área da computação. Nesse sentido, visando possibilitar o acesso de informações importantes para a investigação, sem a necessidade do conhecimento técnico de expressões regulares, a ferramenta de extração de dados de boletins de ocorrência apresentada neste trabalho se justifica.

2.5 Expressão Regular

Segundo (ISIDRO, 2008), uma expressão regular é uma representação algébrica de Autômatos Finitos, os quais são utilizados para realizar o reconhecimento de linguagens regulares. Uma linguagem regular L sobre o alfabeto $\sum$ é um conjunto finito ou infinito de palavras que podem ser formadas pelo alfabeto. Ou seja, $L \subseteq \sum^*$. Uma palavra do alfabeto $\sum$ é qualquer string finita formada pela concatenação dos elementos de $\sum$. Uma palavra pode ser escrita como: $\alpha = \beta_1 \ldots \beta_k$ para $k \geq 1$ e $\beta_i \in \sum$.

As expressões regulares são uma maneira eficaz de representar padrões em cadeias de caracteres, independentemente de sua complexidade. Essa facilidade de representação torna as expressões regulares uma linguagem de entrada comum para ferramentas de pesquisa de texto. Além disso, desempenham um papel crucial nos componentes léxicos de compiladores, bem como possuem aplicações abrangentes em diversas áreas, dentre as quais: processamento de sinais, recuperação de textos, reconhecimento de escrita à mão, reconhecimento de padrões, entre outras (ISIDRO, 2008). A utilização de expressões regulares para reconhecimento das palavras de uma linguagem regular dá-se o nome de reconhecimento léxico.

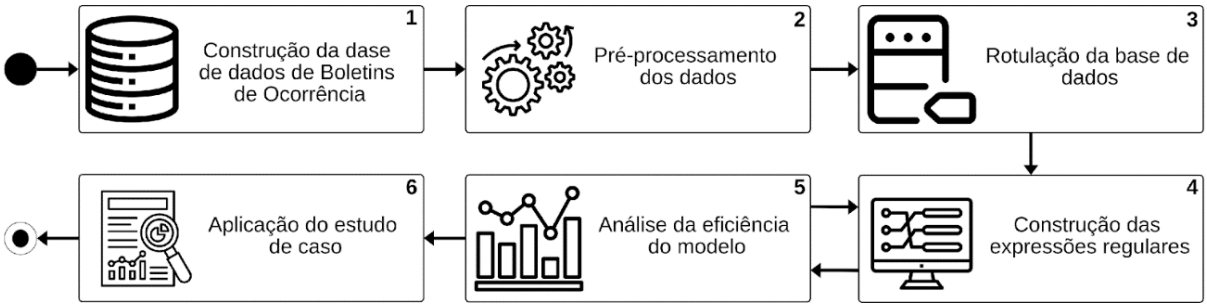
No processo de reconhecimento léxico proposto nesta pesquisa, foram definidas cinco linguagens com as palavras que compreendem: valores monetários, chaves PIX, nomes de aplicativos, agência e conta bancária e nomes de bancos.

Para cada linguagem, o objetivo consiste em reconhecer o maior número de palavras possíveis e, após a avaliação de eficiência, se a quantidade de palavras reconhecidas for satisfatória, pode-se concluir que a ferramenta desenvolvida possui desempenho aceitável.

3 DESENVOLVIMENTO DA FERRAMENTA

A construção da ferramenta de extração de dados dos boletins de ocorrência, bem como a análise da eficiência e aplicação do estudo piloto, seguiu o fluxograma de desenvolvimento, conforme apresentado na Figura 1.

Figura 1 – Fluxograma de desenvolvimento da ferramenta de extração de dados de boletins de ocorrência proposta pelos pesquisadores



Fonte: Elaboração própria (2024).

No primeiro momento, foi construída a base de dados a partir daqueles extraídos dos boletins de ocorrência acessados, após isso é feito um pré-processamento dos mesmos. Em seguida, foi realizada a rotulação da base de dados e a construção das expressões regulares. Na quinta etapa foi promovida a análise da eficiência do modelo e a sexta etapa compreende a condução do estudo piloto. As etapas 2, 3, 4, e 5 foram realizadas de forma iterativa com o objetivo de aprimorar a construção das expressões regulares.

Etapa 1: Construção da base de dados de boletins de ocorrência

A construção da base de dados se deu a partir de um convênio institucional firmado com a Polícia Civil do Estado do Pará, especificamente com a Diretoria Estadual de Combate a Crimes Cibernéticos. Na ocasião, foram disponibilizados 5.858 boletins de ocorrência, os quais são provenientes de ocorrências registradas na Divisão de Combate a Crimes Contra Direitos Individuais por Meios Cibernéticos, Divisão de Combate a Crimes Contra Grupos Vulneráveis Praticados por Meios Cibernéticos e Divisão de Combate a Crimes Econômicos e Patrimoniais Praticados por Meios Cibernéticos, no período de 2018 a 2021. Devido à existência da cláusula de confidencialidade no convênio institucional firmado com a Polícia Civil, a base de dados de boletins de ocorrência utilizada nesta pesquisa não pode ser disponibilizada publicamente para fins de reprodutibilidade do estudo.

A base de dados é formada por boletins de ocorrência em que cada $bo_i \in BD$ corresponde a um boletim de ocorrência. Por representar um texto não estruturado dos fatos relatados, cada bo_i pode ser representado por um vetor na forma $bo_i = [tk_1, tk_1, ..., tk_n]$ em que cada $tk_w \in bo_i$ é uma palavra do texto. As palavras $tk_w \in bo_i$ pertencentes a uma das linguagens definidas neste estudo devem ser reconhecidas pelas expressões regulares do modelo

computacional. A Tabela 1 sumariza as características gerais da base de dados.

Tabela 1 – Características gerais da base de dados composta pelos boletins de ocorrência disponibilizados pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023

Características Gerais	Quantidade
Quantidade de $bo_i \in BD$	5.858
Quantidade média de tk_w por bo_i	169,36
Desvio padrão de tk_w por bo_i	108,23
Quantidade máxima de tk_w por bo_i	1.696
Quantidade mínima de tk_w por bo_i	14
Quantidade de tk_w distintos na BD	49064

Fonte: Elaboração própria (2024).

Etapa 2: Pré-processamento dos dados

A fase de pré-processamento foi necessária para poder preparar o conteúdo da base de dados para o processo de extração automática de dados. O pré-processamento foi realizado em três etapas, sendo elas: (1) padronização de caracteres; (2) tokenização; e (3) remoção de caracteres indesejados. Foram realizados dois procedimentos de padronização de caracteres. O primeiro consistiu na transformação de todas as letras para letra maiúscula. Essa transformação foi necessária, uma vez que reduz o tamanho do alfabeto e permite simplificar as expressões regulares construídas na Etapa 4.

O segundo procedimento de padronização realizado refere-se ao ponto flutuante dos números com casas decimais. Foram observados boletins de ocorrência que utilizam a vírgula para determinar o ponto flutuante de números com casas decimais e outros que utilizam o ponto. Na padronização, optou-se por utilizar o ponto em todos os números de casas decimais. Essa transformação também teve como objetivo simplificar as expressões regulares construídas na Etapa 4.

Etapa 3: Rotulação da base de dados

A rotulação dos dados foi uma etapa necessária para poder realizar a avaliação da eficiência da ferramenta na Etapa 5. Foram selecionadas, aleatoriamente, 300 amostras da base de dados. Para cada amostra selecionada, realizou-se a leitura dos relatos e foram extraídos manualmente os seguintes dados: valor monetário, chave PIX, agência/ conta bancária, aplicativos e nome de banco. A Tabela 2 sumariza a quantidade de informações extraídas.

Tabela 2 – Quantidade de informações extraídas dos 300 boletins de ocorrência que foram selecionados aleatoriamente entre os disponibilizados pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023

Informações extraídas	Quantidade
Valor monetário	511
Chave PIX	84
Aplicativos	286
Agência e conta bancária	637
Banco	359

Fonte: Elaboração própria (2024).

Na etapa de avaliação da eficiência da ferramenta, os dados rotulados foram utilizados como gabarito para verificar se as extrações realizadas foram satisfatórias.

Etapa 4: Construção das expressões regulares

Para a construção das expressões regulares, foram realizadas nos boletins de ocorrência análises textuais dos padrões que caracterizam os dados de valor monetário, chave PIX, agência/conta bancária, aplicativos e nome de bancos. A partir desta análise foi possível criar padrões de máscaras para que pudessem extrair de forma eficiente os dados necessários. A extração de uma palavra implica o seu reconhecimento na linguagem regular definida. A seguir, são descritas as expressões regulares construídas para a ferramenta proposta neste trabalho. As expressões foram construídas a partir da regra sintaxe da biblioteca de expressão regular da linguagem Python (REGEX, 2024).

Valor Monetário

A expressão regular de valor monetário é composta por duas partes de captura e ambas se diferem somente pela máscara. A primeira parte extrai os valores que iniciam com um determinado padrão de máscara, enquanto a segunda captura os valores que iniciam com a máscara que contém o símbolo monetário, o $, e ambas podendo haver grupos de até 3 dígitos separados por um caractere de milhar (ponto ou vírgula), um separador de decimal (vírgula ou ponto), e, por fim, um grupo de dois dígitos.

```
RegEx=(?:(?:DE\s)\d{1,3}(?:,|\.?\d{3})*(?:,\d{2}))|(?:R?[$]\s
   ?\d{1,3}(?:,|\.?\d{3})*(?:(?:,|\.)?\d+))
```

Chave PIX

A expressão regular para a captura de chaves PIX se dá por meio de várias concatenações de máscaras e expressões de e-mail, CPF/CNPJ, número telefônico e chave aleatória.

Inicialmente, há uma expressão regular que procurará pela palavra "PIX", após isso, serão aplicadas as seguintes máscaras nos 70 índices subsequentes à captura da palavra:

A. E-mail, no qual consiste em um conjunto de caracteres alfanuméricos e especiais sucedidos por uma arroba, seguidos de outro grupo de caracteres alfanuméricos que formam o domínio, separado por um ponto e outro grupo de caracteres alfanuméricos, podendo haver várias repetições dessa sequência após o domínio;

```
RegEx=[A-Z0-9!#$%&'*+/=?^_`{|}~-]+(?:[.][A-Z0-9!#$%&'*+/=?^_
    `{|}~-]+)*@(?:[A-Z0-9](?:[A-Z0-9-]*[A-Z0-9])?[.])+[A-Z0
    -9](?:[A-Z0-9-]*[A-Z0-9])?
```

B. Número telefônico, uma sequência de nove a onze dígitos numéricos, sendo o DDD e o código de região opcionais, podendo ou não haver separação dos 4 primeiros dígitos com um traço (-) ou espaçamentos;

```
RegEx=(?:\+\d\d\s)?\(?[\d]{2}\)?\s\d?\s?[\d]{4}[-][\d
    ]{4}|(?:\+\d\d\s)[\d]{2}\s[\d]{4}\s?[\d]{4,5}
```

C. Chave aleatória é composta por um conjunto de oito caracteres alfanuméricos, em seguida um traço (-) separando de uma sequência de três conjuntos alfanuméricos com tamanho de quatro caracteres e tais conjuntos separados por um traço, logo após um traço e, por fim, um conjunto de doze caracteres alfanuméricos.

```
RegEx=[A-Z0-9]{8}-(?:[A-Z0-9]{4}-){3}[A-Z0-9]{12}
```

D. CPF/CNPJ, consiste em duas expressões concatenadas, a primeira se refere a chave de CPF que é composta por três conjuntos de três dígitos numéricos, separados por um ponto, logo após seguido de um traço, sendo o traço e o ponto opcionais, e, por fim, dois dígitos numéricos. Já a segunda expressão, no qual captura as chaves de CNPJ, é composta por dois dígitos numéricos, seguidos de dois conjuntos de três dígitos numéricos, separados por um ponto e em seguida um caractere barra (/), sendo ambos também opcionais, logo depois quatro dígitos numéricos, e finalmente um traço opcional seguido de dois dígitos.

```
RegEx=(?:[\d]{2}[.]?(?:[\d]{3}[.]?){2}[/]?[\d]{4}[-]?[\d]{2})
    |(?:(?:[\d]{3}[.]?){2}[\d]{3}[-]?[\d]{2})|(?:[*]{3}[.][\d
    ]{3}[.][\d]{3}[.]?[-][*]{2})
```

Agência e Conta Bancária

Os padrões de expressões regulares que extraem os dados de agência e conta bancária estão divididos em grupos de expressões que capturam agência e conta, que podem vir separados, ou juntos, independentemente da ordem. A expressão regular de agência consiste primeiramente por uma máscara de padrões que antecedem um grupo de até cinco dígitos, podendo haver um separador (barra ou traço) e um grupo de até dois dígitos. Assim como a regex

de agência, a expressão regular de conta bancária também é composta por outra máscara de padrões, e em seguida uma sequência de dígitos, podendo haver um separador (traço ou ponto) e outro grupo de dígitos numéricos.

```
RegEx Agência=(?:(AGENCIA|AG|AGENCIA DESTINO)[ :]*
([0-9]1,5[-]?[0-9]0,2))
RegEx Conta=(?:(CONTA|CONTA CORRENTE|C/?C|CONTA CORRENTE
PJ|CONTA DESTINO|CONTA CORRENTE N°)[ ,:]*
([0-9]+(?:[-.]?[0-9])+))
```

Aplicativos

A captura dos nomes de aplicativos digitais é feita por meio da concatenação de uma base de dados com os nomes dos principais aplicativos.

Nome de banco

A extração dos nomes de bancos é feita por meio da concatenação dos principais e mais frequentes nomes de bancos que estão armazenados em uma base de dados.

Etapa 5: Análise da eficiência do modelo

Para analisar a eficiência do modelo, foram utilizadas as 300 amostras selecionadas e rotuladas na Etapa 3. A partir desse conjunto de dados, foram construídas matrizes de confusão para cada uma das linguagens definidas, quais sejam: valor monetário, chave PIX, agência/-conta bancária, aplicativos e nome de banco.

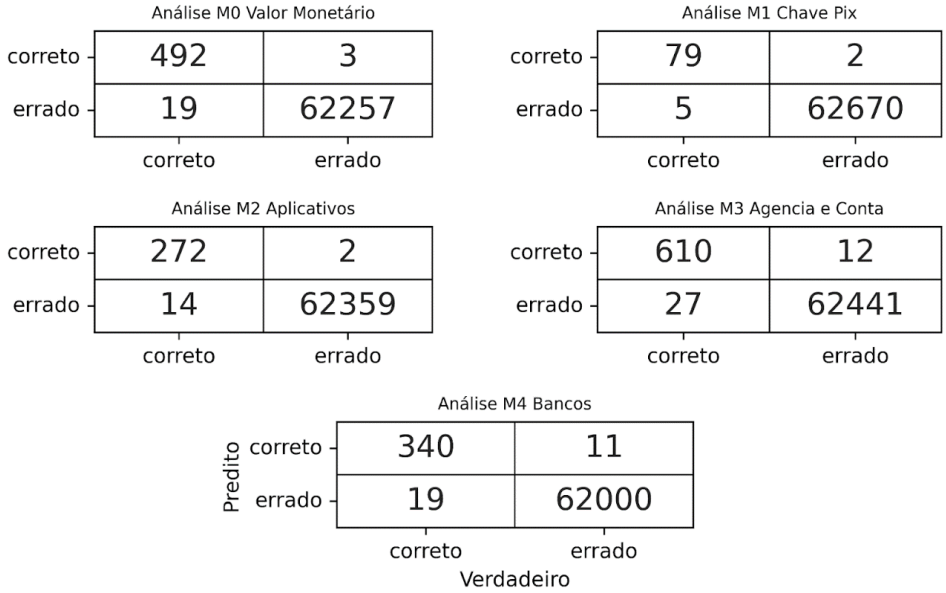
Todas as matrizes de confusão construídas foram binárias, contendo quatro células, conforme representada na Figura 2. As linhas da matriz correspondem às extrações esperadas e as colunas correspondem às extrações realizadas (ER).

Cada célula da matriz de confusão binária possui um rótulo específico de interpretação. A célula da posição linha correto e coluna correto é rotulada como True positive (TP) que corresponde aos dados que deveriam ser extraídos e realmente foram extraídos por uma ER. A célula da posição linha correto e coluna errado é rotulada como False positive (FP) que corresponde aos dados que não deveriam ser extraídos, mas foram incorretamente extraídos por uma ER. A célula da posição linha errado e coluna correto é rotulada como False negative (FN) que corresponde aos dados que deveriam ser extraídos por uma ER, mas não foram. Por fim, a célula da posição linha errado e coluna errado é rotulada como True negative (TN) que corresponde aos dados que não deveriam ser extraídos e realmente não foram.

Considerando que a ferramenta proposta tem como finalidade otimizar a leitura dos BOs, dando garantias de que os dados extraídos estão corretos, durante o processo de análise da

eficiência das expressões regulares, é importante alcançar bons índices de precisão e revocação. Porém, a maximização do índice de precisão é mais importante porque garante que nenhuma informação extraída é incorreta.

Figura 2 – Matrizes de confusão resultante da extração dos dados rotulados, a partir de uma amostra dos boletins de ocorrência disponibilizados pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023



Fonte: Elaboração própria (2024).

A Figura 2 apresenta as matrizes de confusão de cada classe de interesse. A partir dos dados apresentados, foi possível calcular a precisão e a revocação, conforme definido pela Equação 1 e 2.

$$precisao = \frac{TP}{TP + FP} \tag{1}$$

$$precisao = \frac{TP}{TP + FN} \tag{2}$$

As métricas de precisão e revocação admitem valores no intervalo fechado de 0 a 1. Quanto mais próximo de um, melhor é o valor obtido pela métrica. A precisão tem como objetivo verificar a proporção de palavras extraídas corretamente, considerando todas as palavras que foram extraídas. A precisão é calculada a partir da análise da primeira linha da matriz de confusão, analisando as células TP e FP. Já a métrica de revocação tem como objetivo verificar a proporção de palavras extraídas corretamente, considerando todas as palavras que deveriam ser extraídas. A revocação é calculada a partir da análise da primeira coluna da matriz de confusão, analisando as células TP e FN.

A Tabela 3 apresenta os valores calculados de cada métrica. Em uma análise geral, é possível constatar que a extração de dados realizada pela ferramenta é eficiente, uma vez que todas as métricas apresentaram valores superiores a 0.94. Também é possível constatar que, para todas as classes de extração de dados, a precisão da ferramenta foi um pouco melhor que a revocação. Isso permite concluir que a ferramenta, apesar de não extrair 100% dos dados esperados, minimiza o erro de falso positivo ligeiramente melhor do que o erro de falso negativo.

No contexto da investigação policial, a minimização do erro de falso positivo garantidos pela alta precisão assegura que praticamente não há dados incorretos extraídos pela ferramenta. Já a minimização do erro de falso negativo garantido pela alta revocação assegura que a ferramenta praticamente não está deixando de extrair dados corretos que deveriam ser extraídos. Portanto, pode-se concluir que a ferramenta de extração de dados em boletins de ocorrência proposta neste estudo é eficiente e pode otimizar o processo de investigação policial.

Tabela 3 – Valores das matrizes de precisão e revocação calculados de cada métrica, a partir de uma amostra dos boletins de ocorrência disponibilizados pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023

Matrizes	Precisão	Revocação
M0	0.99	0.96
M1	0.97	0.94
M2	0.99	0.95
M3	0.98	0.95
M4	0.96	0.94

Fonte: Elaboração própria (2024).

A ferramenta proposta foi desenvolvida utilizando a linguagem de programação Python e os *framework* Flet e Pandas (PYTHON, 2023; FLET, ; PANDAS, 2023). O *framework* Flet foi utilizado para a construção da interface gráfica da ferramenta e o *framework* Pandas foi utilizado para possibilitar a leitura e escrita de arquivos. A Figura 3 ilustra a ferramenta desenvolvida que está registrada como propriedade intelectual no Instituto Nacional de Propriedade Intelectual (INPI), com o número de registro BR512023002935-1.

Figura 3 – Tela da ferramenta de extração de dados – número de registro BR512023002935-1

Fonte: Instituto Nacional de Propriedade Intelectual (2023).

Etapa 6: Aplicação do Estudo de caso ou Estudo piloto

A fim de demonstrar as possibilidades de utilização da ferramenta proposta, foi realizado um estudo piloto na Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Estado do Pará. Foram utilizados 5.858 boletins de ocorrência e as seguintes perguntas foram respondidas, a partir dos dados extraídos:

(A) No período de 2018 a 2021, qual foi o prejuízo monetário de crimes contra o patrimônio?

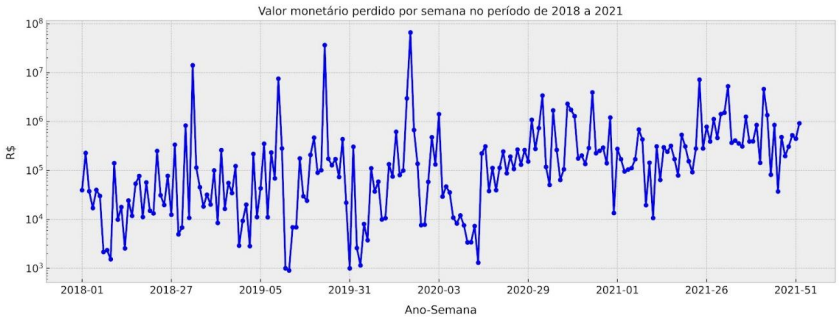
(B) No período de 2020 a 2021, quantas chaves PIX foram utilizadas e quais são duplicadas?

(C) No período de 2018 a 2021, qual foi o aplicativo mais utilizado para a aplicação de golpes virtuais?

(D) No período de 2018 a 2021, quais os principais bancos envolvidos em golpes virtuais?

Conforme as variáveis elegíveis para análise, os dados foram analisados e representados por meio de figuras. A Figura 4 ilustra a série temporal semanal do prejuízo monetário que foram relatados nos boletins de ocorrência, no período de 2018 a 2021.

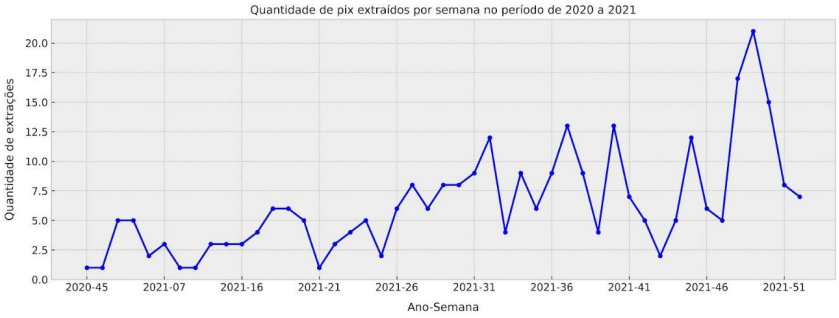
Figura 4 – Prejuízo monetário de crimes contra o patrimônio no estado do Pará, no período de 2018 a 2021



Fonte: Elaboração própria (2024), com base em dados extraídos dos BOs fornecidos pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023.

Conforme mostra a Figura 4, os dados extraídos demonstram que houve três picos em que o prejuízo monetário semanal superou a marca de dez milhões de reais (R$ 107). Também é possível notar que na maioria dos casos o prejuízo semanal supera o montante de dez mil reais. A extração de dados de valores monetários, além de permitir a análise de série temporal, também permite que o investigador classifique os crimes por prejuízo econômico. Na Figura 5, verifica-se a série temporal semanal utilizando a chave PIX de 2020 a 2021 como variáveis.

Figura 5 – Quantidade de chaves PIX utilizadas em golpes virtuais praticados no estado do Pará, nos anos de 2020 e 2021



Fonte: Elaboração própria (2023), com base em dados extraídos dos BOs fornecidos pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023.

A Figura 5 ilustra a série temporal semanal da frequência absoluta de chaves PIX utilizadas em golpes virtuais relatados nos boletins de ocorrência. Apesar da base de dados ser do período de 2018 a 2021, na análise de chave PIX foi a única em que se considerou somente os anos de 2020 e 2021, pois nos anos anteriores a chave PIX ainda não existia. Na Figura 5, é possível perceber a tendência de crescimento de golpes virtuais que utilizam chave PIX. Ao mesmo tempo em que a transação por PIX trouxe benefício para o comércio, também trouxe desafios de segurança, dada a facilidade de aplicação de golpes utilizando esta tecnologia.

Além de verificar a frequência de utilização de chaves PIX, bem como quais chaves estão sendo utilizadas, uma informação importante no curso da investigação policial refere-se

à verificação de quais boletins de ocorrência possuem a mesma chave PIX relatada. Essa informação é importante porque pode revelar padrões de aplicação de golpes virtuais. É possível perceber esse padrão a partir da representação em grafo. No grafo da Figura 6, existem dois tipos de vértices denominados como fonte e sumidouro. O vértice fonte só possui arestas de saída e representa um boletim de ocorrência. O vértice sumidouro só possui aresta de entrada e corresponde à uma chave PIX. Todos os vértices fontes que estão conectados a um vértice sumidouro estão compartilhando a mesma chave PIX.

A Figura 6 apresenta o grafo formado a partir da extração de chaves PIX realizada pela ferramenta. Para facilitar a visualização, foram considerados somente as chaves PIX compartilhadas por pelo menos dois boletins de ocorrência. Nota-se que foram formados 17 grupos provenientes do compartilhamento de chave PIX. Dentre os grupos, 15 grupos são com o número mínimo de dois vértices fonte, um grupo possui três vértices fonte e o maior grupo possui sete vértices fonte. Em cada vértice sumidouro consta a chave PIX anonimizada.

Figura 6 – Representação em grafo dos boletins de ocorrência analisados que possuem relação de chave PIX, utilizadas em golpes virtuais praticados no estado do Pará, nos anos de 2020 e 2021

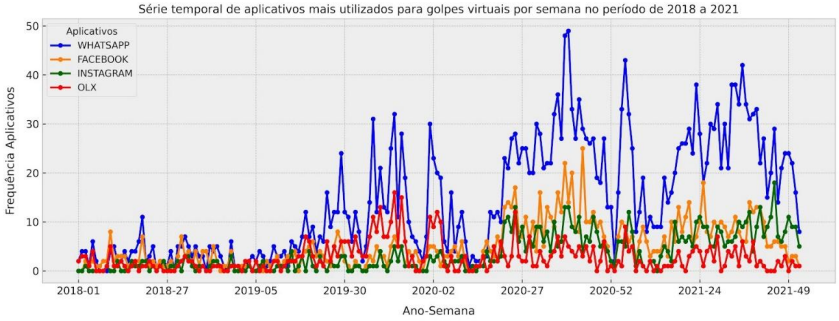
Fonte: Elaboração própria (2024), com base em dados extraídos dos BOs fornecidos pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023.

A formação e o tamanho de cada grupo revelam padrões importantes, por exemplo, o grupo formado por sete vértices fonte evidencia que sete vítimas registraram boletim de ocorrência relatando golpe virtual e que a mesma chave PIX (e-mail, número de telefone, CPF/CNPJ) foi utilizada para a aplicação do golpe. A partir desta percepção, de formação de grupos, o investigador pode identificar o modus operandi do golpe e otimizar o processo de investigação. Em seguida, a Figura 7 ilustra a série temporal semanal da frequência absoluta de aplicativos utilizados em golpes virtuais relatados em boletins de ocorrência.

De acordo com a Figura 7, partir de 2020, é possível perceber um crescimento nos golpes utilizando aplicativos, com destaque ao Whatsapp, como ferramenta social na qual mais ocorreu esse tipo de crime no Estado do Pará. Esse levantamento segue tendência global, em

que segundo pesquisa realizada pela Kaspersky Internet Security (2021) o aplicativo de troca de mensagens mais usado em golpes de mensagens falsas *(phishing)* é o Whatsapp, com 89,6% das ocorrências, a partir de dados anônimos enviados pelos clientes da Kaspersky Internet Security, voluntariamente, entre dezembro de 2020 e maio de 2021. Além disso, nesse ambiente é comum a utilização da técnica de engenharia social por meio da qual se induz a vítima a compartilhar dados pessoais confidenciais (FREITAS, 2020).

Figura 7 – Aplicativos mais utilizados para a aplicação de golpes virtuais no estado do Pará, no período de 2018 a 2021



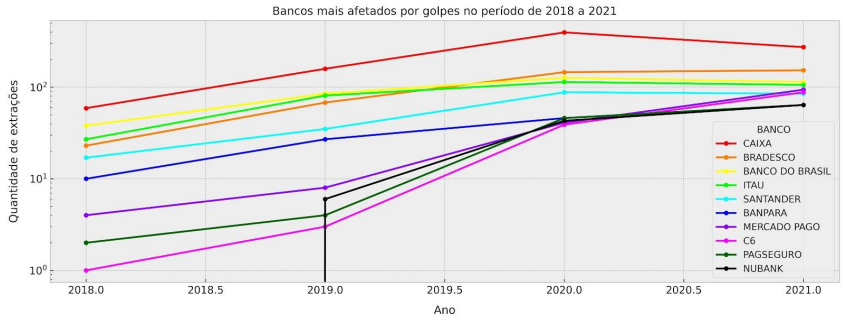
Fonte: Elaboração própria (2023), com base em dados extraídos dos BOs fornecidos pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023.

Ainda na Figura 7, no final da série temporal, é possível perceber que o Instagram se tornou o segundo aplicativo mais utilizado em práticas criminosas, no Estado do Pará. Uma possível explicação para esse crescimento pode estar no fato do Instagram ter se tornado uma plataforma também focada em anúncio e comércio. Vengga e Gunadi (2022) expõem os riscos das negociações por meio do Instagram, sem um encontro presencial. Os autores ressaltam que juridicamente, as transações de compra e venda realizadas por meio da plataforma não são seguras e/ou legalmente apropriadas, pois se trata de uma mídia social que deve ter como objetivo atender às necessidades de comunicação remota ou socialização entre outros usuários de sistemas eletrônicos.

Outra explicação para o crescimento de relatos envolvendo o Instagram pode estar no elevado número de contas roubadas. Segundo o KASPERSKY (2021) há uma tendência de *hacking* contra contas dessa rede social, promovida por cibercriminosos. De acordo com previsões do NordVPN (BARECKAS, 2022), empresa fornecedora de serviços de Rede Privada Virtual, há uma probabilidade de 25% de comprometimento de todas as contas de redes sociais, em algum momento. No que diz respeito às instituições bancárias, a Figura 8 demonstra os principais bancos envolvidos em golpes virtuais.

Com base na Figura 8, em relação às instituições bancárias e contas frequentemente associadas a golpes virtuais, destaca-se uma incidência de atividades fraudulentas direcionadas à Caixa Econômica Federal. Em sequência, aparece o Bradesco, Banco do Brasil, Itaú e Santander, configurando-se como os bancos mais propensos a essas práticas, conforme ilustrado na Figura 8. Além disso, é perceptível um aumento na prática desses crimes desde o ano de 2019.

Figura 8 – Aplicativos mais utilizados para a aplicação de golpes virtuais no estado do Pará, no período de 2018 a 2021



Fonte: Elaboração própria (2023), com base em dados extraídos dos BOs fornecidos pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023.

A Tabela 4 apresenta as respostas das quatro questões que foram elaboradas na condução deste estudo piloto. O objetivo de responder essas perguntas é demonstrar o ganho de informações que a extração da informação realizada pela ferramenta proposta pode proporcionar. A extração de dados em boletins de ocorrência ocorre pela transformação de dados não estruturados em dados estruturados. Os dados não estruturados são os relatos dos BOs dispostos em texto corrido. Os dados estruturados são tabelas formadas por linhas e colunas. No processo de extração, as colunas correspondem às classes de interesse e as linhas aos dados extraídos.

Quadro 1 – Valores das matrizes de precisão e revocação calculados de cada métrica, a partir de uma amostra dos boletins de ocorrência disponibilizados pela Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Pará, 2023

Matrizes	Resposta
No período de 2018 a 2021, qual foi o prejuízo monetário de crimes contra o patrimônio?	R$199.060.145,60
No período de 2020 a 2021, quantas chaves PIX foram utilizadas e quais são duplicadas?	513 chaves PIX e 17 duplicadas
No período de 2018 a 2021, qual foi o aplicativo mais utilizado para a aplicação de golpes virtuais?	WhatsApp
No período de 2018 a 2021, quais os principais bancos envolvidos em golpes virtuais?	Caixa, Bradesco, Banco do Brasil e Itaú

Fonte: Elaboração própria (2024).

A partir da estruturação dos dados, as perguntas como as da Tabela 4 podem ser respondidas. Conforme demonstrado no estudo piloto, a estruturação dos dados também permite a construção de gráficos para a análise de séries temporais. Analisar séries temporais é uma atividade policial importante, pois permite a construção de indicadores para conduzir atividades de monitoramento e avaliação do desempenho. Outro ganho de informação demonstrado no estudo piloto refere-se à análise de dados a partir da teoria dos grafos. Ao utilizar teoria dos grafos, foi possível perceber a relação dos boletins de ocorrência em função das chaves PIX. A observação desta relação pode direcionar a investigação policial para ler os boletins de ocorrência relacionados e identificar o modus de operação dos crimes semelhantes.

4 CONSIDERAÇÕES E REFLEXÕES FINAIS

A análise do estudo piloto realizado na Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Estado do Pará permitiu constatar que a proposta de ferramenta de extração de dados de boletins de ocorrência tem potencial para dar celeridade e aperfeiçoar a qualidade da investigação policial. A celeridade ocorre quando os investigadores deixam de realizar a extração de dados de forma manual e passam a utilizar a ferramenta, tornando o processo de extração automático. Em relação a qualidade da investigação, essa é aperfeiçoada, uma vez que os dados extraídos permitem novas análises das informações tabuladas. Foi demonstrado nesta pesquisa que as informações extraídas permitem a análise de séries temporais e a análise da relação entre boletins de ocorrência utilizando teoria dos grafos.

Uma limitação da ferramenta proposta nesta pesquisa, diz respeito à extração dos nomes

das vítimas e investigados que constam nos relatos dos BOs. Não foi possível realizar essa extração, pois a teoria de Linguagem Regular adotada nesta pesquisa não é suficiente para realizar o reconhecimento de nome próprio. Para realizar o reconhecimento de nome próprio, é necessário a utilização de algoritmos de inteligência artificial. Como a utilização de algoritmos de inteligência artificial está fora do escopo desta pesquisa, recomenda-se como trabalho futuro o desenvolvimento e avaliação do modelo de inteligência artificial para a extração de nome próprio.

Diante do exposto, a ferramenta de extração de dados desenvolvida neste estudo apresenta-se como uma solução eficiente e inovadora para otimizar o processo de investigação policial, especificamente na análise de boletins de ocorrência. A análise da eficiência do modelo, por meio das matrizes de confusão e métricas de precisão e revocação, evidencia sua capacidade em extrair informações relevantes de maneira precisa, minimizando tanto o erro de falso positivo quanto o de falso negativo.

A eficácia da ferramenta foi comprovada em um estudo piloto na Diretoria Estadual de Combate a Crimes Cibernéticos da Polícia Civil do Estado do Pará, ao analisar dados de 5.858 boletins de ocorrência. Essa análise permitiu o alcance de insights valiosos para a investigação, revelando padrões, como a relação entre boletins de ocorrência por meio da chave PIX, além de oferecer uma visão abrangente dos crimes cibernéticos no Estado por meio de representações temporais de prejuízo monetário, frequência de chaves PIX, uso de aplicativos e envolvimento de instituições bancárias.

A aplicação da ferramenta revelou a importância crescente de plataformas como o Whatsapp e Instagram nas atividades criminosas, destacando tanto a necessidade de maior vigilância quanto conscientização por parte dos usuários dessas redes, bem como a elaboração de políticas públicas no âmbito da segurança pública voltadas para o combate a essa tendência, além de maior implementação de medidas de segurança por parte dos desenvolvedores da plataforma, utilizando recursos para coibir essa prática delituosa.

O desenvolvimento da ferramenta em Python, com o suporte dos *frameworks* Flet e Pandas, não apenas facilitou a extração de dados, mas também possibilitou a criação de uma interface gráfica amigável. O registro da ferramenta no Instituto Nacional de Propriedade Intelectual (INPI) confirma sua originalidade, relevância e conformidade com os padrões nacionais exigidos.

Em síntese, a ferramenta de extração de dados de boletins de ocorrência não apenas agiliza o trabalho investigativo, mas também oferece uma abordagem inovadora para a compreensão e prevenção de crimes cibernéticos, destacando-se como uma contribuição valiosa para a área da segurança pública.

REFERÊNCIAS

BARECKAS, K. **Instagram hacked: Everything you need to know**. NordVPN, 2022. Disponível em: <https://nordvpn.com/pt-br/blog/instagram-hacked/>. Acesso em: 10 nov. 2023.

BRASIL. **Ministério da Justiça e Segurança Pública. Ministério da Justiça e Segurança Pública lança Plano Tático de Combate a Crimes Cibernéticos.** Governo Federal, 2022. Disponível em: <https://www.gov.br/mj/pt-br/assuntos/noticias/ministerio-da-justica-e-seguranca-publica-lanca-plano-tatico-de-combate-a-crimes-ciberneticos>. Acesso em: 31 out. 2023.

BRASIL. Ministério da Justiça e Segurança Pública. **Plano Tático de Combate ao Crime Cibernético (PTat-C3).** Governo Federal, 2022.

BRASIL. **Ministério da Justiça e Segurança Pública. Plano Nacional de Segurança Pública e Defesa Social (2021-2030).** Governo Federal, 2023. Disponível em: <https://www.gov.br/mj/pt-br/assuntos/noticias/ministerio-da-justica-e-seguranca-publica-lanca-plano-tatico-de-combate-a-crimes-ciberneticos>. Acesso em: 31 out. 2023.

DEMIDOVA, N. **Como contas do Instagram são roubadas?** Governo Federal, 2023. Disponível em: <https://www.kaspersky.com.br/blog/instagram-hijack/10746/>. Acesso em: 11 nov. 2023.

FLET. **The fastest way to build Flutter apps in Python.** Flet, Copyright ©. Disponível em: <https://flet.dev/>. Acesso e m: 12 nov. 2023.

FLEURY, M. T. L.; WERLANG, S. R. da C. **Pesquisa aplicada: conceitos e abordagens.** Anuário de pesquisa 2016–2017. São Paulo: Única Gráfica e Editora Ltda., 2016.

FORTINET. **Comunicados à imprensa:** Brasil sofreu mais de 88,5 bilhões de tentativas de ataques cibernéticos em 2021. Global Research Report, 2022. Disponível em: <https://www.fortinet.com/br/corporate/about-us/newsroom/press-releases/2022/fortiguard-labs-relatorio-ciberataques-brasil-2021>. Acesso em: 26 nov. 2023.

FORTINET. Training institute. **Cybersecurity Skills Gap.** 2023 Global Research Report., 2023. Disponível em: <https://www.fortinet.com/content/dam/maindam/PUBLIC/02_MARKETING/08_Report/2023-cybersecurity_skills_gap_report_final.pdf?utm_source=website&utm_medium=pr&utm_campaign=cybersecurity-skills-gap-2023>. Acesso em: 26 nov. 2023.

FREITAS, E. **Golpe da Engenharia Social com WhatsApp**. JusBrasil, 2020. Disponível em: <https://www.jusbrasil.com.br/noticias/golpe-da-engenharia-social-com-whatsapp/1178033046>. Acesso em: 22 nov. 2023.

GERSTING, J. L. **Fundamentos matemáticos para a ciência da computação**. 4. ed. Rio de Janeiro: LTC, 2001.

GIL, A. C. **Como elaborar projetos de pesquisa**, 4. ed. São Paulo: Atlas. 2002.

IBM SECURITY. **Cost of a Data Breach Report 2023**. **IBM Security**, 2023. Disponível em: <https://cyberalberta.ca/system/files/cyberalberta-coi-cost-of-a-data-breach.pdf>. Acesso em: 26 nov. 2023.

ISIDRO, C. R. G. **Uma abordagem quântica para o uso de expressões regulares.** 2008. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Campina Grande, Campina Grande, Paraíba, 2008.

KASPERSKY. **Kaspersky:** quase 90% das mensagens fraudulentas são disseminadas via WhatsApp. 2021. Disponível em: <https://www.kaspersky.com.br/about/press-releases/2021_kaspersky-quase-90-das-mensagens-fraudulentas-sao-disseminadas-via-whatsapp>. Acesso em: 30 out. 2022.

LUNETTA, A. D.; GUERRA, R. Metodologia da pesquisa científica e acadêmica. **Revista OWL (OWL Journal)-Revista Interdisciplinar de Ensino e Educação**, v. 1, n. 2, 2023.

MAJEED, A.; KHAN, S.; HWANG, S. O. A comprehensive analysis of privacy-preserving solutions developed for online social networks. **Electronics**, v. 11, n. 13, p. 1931, 2022.

NERY, C.; BRITTO, V. Internet já é acessível em 90,0% dos domicílios do país em 2021. **Agência IBGE Notícias**, 2022. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/34954-internet-ja-e-acessivel-em-90-0-dos-domicilios-do-pais-em-2021>. Acesso em: 30 jun. 2023.

PANDAS. **Pandas – Python Data Analysis**. Pandas, 2023. Disponível em: <https://pandas.pydata.org/>. Acesso em: 12 nov. 2023.

PYTHON. **Aplicativos para Python**. Python, 2023. Disponível em: <https://www.python.org/about/apps/>. Acesso em: 12 nov. 2023.

RAUPP, F. M.; BEUREN, I. M. Como elaborar trabalhos monográficos em contabilidade. Metodologia da pesquisa aplicável às Ciências Sociais. In: I. M. Beuren (Ed.) **Como Elaborar Trabalhos Monográficos em Contabilidade:** Teoria e prática. 3. ed, p. 76–97, 2006.

REGEX. **Regular expression operations**. Python, v. 3.13.0, 2024. Disponível em: <https://docs.python.org/3/library/re.html>. Acesso em: 28 nov. 2024.

SURFSHARK. **Data breach statistics by country:** first quarter of 2022. Surfshark, 2022. Disponível em: <https://surfshark.com/blog/data-breach-statistics-by-country>. Acesso em: 26 out. 2023.

TORRES, R.; OHASHI, O.; PESSIN, G. A machine-learning approach to distinguish passengers and drivers reading while driving. **SENSORS**, v. 19, p. 3174, 2019.

VENGGA, V.; GUNADI, A. **Liability of Instagram Social Media Platform as an Advertising Service Provider in Case of Online Shop Fraud** In: 3rd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2021). Atlantis Press, p. 712–716, 2022.