MINERAÇÃO DE DADOS NO TWITTER PARA VERIFICAR EXPOSIÇÃO DE INFORMAÇÕES PESSOAIS POR MEIO DE PYTHON E MYSQL

Ânderson Luís de Souza[1], Paulo João Martins[2]

Resumo: O crescimento acelerado dos sites de redes sociais nos últimos anos criou uma enorme base de dados. Dentre o grande volume de informações compartilhadas diariamente, encontram-se dados que comprometem a segurança dos usuários. Atentos a esses dados, existem pessoas mal-intencionadas e à espera de qualquer tipo de informação que possa lhes trazer vantagens sobre determinado usuário. Utilizando de técnicas de engenharia social, alguns poucos dados são suficientes para planejar ataques diretos. O propósito desse trabalho foi realizar uma análise na rede social Twitter, utilizando técnicas de Descoberta de conhecimento em base de dados para localizar dados que possam ser possíveis alvos da engenharia social. Para execução das etapas de descoberta de conhecimento foram implementados algoritmos na linguagem de programação Python e os dados foram armazenados no banco de dados relacional MySQL. Para realizar extração dos dados da rede social, foi utilizado a biblioteca *tweepy* e para classificação dos dados, a biblioteca *sklearn*. Utilizando a biblioteca *sklearn*, dois modelos foram testados: Máquinas de Vetores de Suporte e Multinomial Naive Bayes. O modelo de Máquinas de Vetores de Suporte se mostrou superior e foi selecionado para realizar a classificação da base principal. Os resultados mostraram que mais de 99% dos usuários possuem algum tipo de exposição, porém, a falta da validação de um dos campos, diminuiu a credibilidade dos resultados. Ao final, o resultado foi satisfatório, identificando-se que uma pequena parcela dos usuários compartilha dados pessoais suscetíveis a ataques de engenharia social.

Palavras-chave: Análise de dados. Engenharia social. Descoberta de conhecimento em base de dados. MySQL. Python. Dados pessoais. Twitter.

[1] Ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC), andersonluis@unesc.net
[2] Ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC), pjm@unesc.net

ABSTRACT: The rapid growth of social networking sites in recent years has created a huge database. Among the large volume of information that is shared daily, there is data that compromises the security of users. Aware of this data, there are malicious people prepared to take advantage of these users. Using social engineering techniques, some data is enough to plan direct attacks. The purpose of this work is to perform an analysis on the social network Twitter, using knowledge discovery in databases (KDD) techniques to locate data that may be possible targets of social engineering. To execute the KDD steps, algorithms were implemented in the Python programming language and the data were stored in the MySQL relational database. To extract data from the social network, the tweepy library was used and, for data classification, the sklearn library. Using the sklearn library, two models were tested: Support Vector Machines and Naive Bayes Multinomials. The Support Vector Machines model proved to be superior and was selected to perform the classification of the main base. The results showed that more than 99% of users have some type of exposure, however, the lack of validation of one of the fields reduced the credibility of the results. In the end, the result was satisfactory, identifying that a small portion of users share personal data susceptible to social engineering attacks.

Keywords: Data analysis. Social engineering. Discovery of knowledge in databases. MySQL. Python. Personal data. Twitter

1 INTRODUÇÃO

Ao longo do tempo, os sites de redes sociais tornaram-se cada vez mais populares. O volume de informações compartilhadas, tem chamado a atenção de usuários com intenções maliciosas, que procuram por formas de obter vantagens sobre outros. Periodicamente é noticiado pela imprensa que algum indivíduo teve prejuízo por conta de sua exposição em sites de redes sociais. Muitos desses indivíduos tiveram seus perfis monitorados e acabam sofrendo um ataque de engenharia social.

Engenharia social é o ato de persuadir as pessoas ao ponto de convencê-las de que algo, ou alguém, na verdade não é aquilo que aparenta ser. Essa técnica não está totalmente ligada à tecnologia, e, tem como objetivo influenciar as pessoas

para obter informações, com as quais, será possível conseguir vantagens sobre um outro indivíduo ou organização (MITNICK; SIMON, 2002).

Segundo o Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (2012, p. 20), esses indivíduos maliciosos usam as informações que os usuários postam em sites de redes sociais, como nome, sobrenome, data de nascimento, locais que frequenta e até mesmo números de documentos, para planejar ataques eficientes.

Carpim (2014) em sua monografia, descreve que as pessoas precisam divulgar o que está ocorrendo em sua vida. Na maioria das vezes a exposição é para exibir uma vida perfeita e feliz, exibindo apenas os momentos alegres. Muitas vezes, nessa necessidade de expor uma boa vida diante das outras pessoas, alguns protocolos de segurança acabam sendo ignorados e, muitas vezes, informações importantes de cunho pessoal acabam sendo compartilhadas.

Com o avanço da tecnologia, os armazenamentos foram ficando cada vez maiores e a quantidade de dados armazenados era muito grande para os tipos de análise convencionais que havia no início da história da tecnologia da informação (CASTRO, 2016). Os dados armazenados, quando extraídos, possuem capacidade de gerar conhecimento para tomada de decisão, prevendo padrões que podem mudar o direcionamento dos negócios nas mais variadas áreas (GOLDSCHMIDT; PASSOS, 2015).

Devido a necessidade de analisar grandes bases de dados, surgiu a Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases – KDD), popularmente conhecida como mineração de dados. Apesar do termo mineração de dados ser mais comum, ele é na verdade uma das etapas do KDD (CASTRO, 2016).

KDD é um processo de várias etapas operacionais, que requer interação humana e possíveis repetições, a fim de refinar os dados obtidos. Antes de iniciar a mineração de dados, é necessário capturar, tratar e organizar os dados, para que eles estejam preparados para serem minerados. Essa etapa é chamada de pré-processamento. A busca por conhecimento acontece na etapa de mineração de dados, onde são aplicados algoritmos inteligentes sobre os dados pré processados. Por último, o conhecimento é tratado na etapa de pós-processamento, onde são avaliadas as utilidades do conhecimento obtido (GOLDSCHMIDT; PASSOS, 2015).

Partindo do princípio de que os engenheiros sociais procuram dados suficientes para transformar em conhecimento e que os sites de redes sociais são grandes repositórios de dados, pode-se então, aplicar técnicas computacionais, capazes de extrair as informações e mostrar o quanto os indivíduos que frequentam sites de redes sociais estão expondo informações que comprometam sua segurança.

A análise de redes sociais, têm chamado a atenção de pesquisadores nos últimos anos, devido a grande quantidade de dados disponíveis na rede e o conhecimento que pode ser extraído (RECUERO, 2017).

Ainda, segundo Recuero (2017), existem diversas formas de coletar dados de redes sociais, mas primeiramente é preciso decidir com qual rede se está trabalhando e, se será uma pesquisa quantitativa ou qualitativa. Qualitativa é quando a pesquisa é feita por meio de entrevistas com os membros de uma rede, já a quantitativa, que faz parte do objetivo deste trabalho, é feita de forma automática por meio de ferramentas de extração de dados.

Nessa pesquisa, será analisado a rede social Twitter, utilizando ferramentas de KDD. Devido ao grande número de bibliotecas dedicadas a extração, análise, classificação e métricas de KDD, será utilizada a linguagem de programação Python e, para armazenar os dados, um sistema gerenciador de banco de dados. Nesta pesquisa, optou-se pelo MySQL Community 8.0. Para uma melhor organização das etapas do projeto, os seguintes objetivos específicos foram elaborados: compreender o conceito de engenharia social, compreender o conceito redes sociais e análise de sites de redes sociais, extrair informações da rede social Twitter por meio de ferramentas gratuitas, comparar os modelos SVM e MNB que estão presentes nas bibliotecas do Python e Utilizar Python e MySQL para realizar a descoberta do conhecimento. O objetivo geral desse projeto é mostrar se, com as ferramentas propostas, é possível encontrar e determinar o quanto os usuários do Twitter estão expostos a um possível ataque de engenharia social.

Durante o levantamento bibliográfico, poucos trabalhos encontrados traziam mineração de dados e engenharia social no mesmo material. Porém, pesquisas que envolvem aplicação de KDD sobre dados da rede social Twitter são encontrados ligados a outros domínios de aplicação. Algoritmos de mineração não foram concebidos com tantos propósitos diferentes, mas podem ser usados de várias maneiras em diversas áreas do conhecimento (BERRY, 2004, tradução nossa). O trabalho de pesquisa de Lansley, Mouton, Kapetanakis e Polatidis (2020) e publicado

pelo *Journal of Information and Telecommunication* é o que mais se aproxima do domínio de aplicação aqui proposto. Os autores utilizaram Processamento de Linguagem Natural em conjunto com Redes Neurais Artificiais para detectar ataques de engenharia social em diálogos por texto. Dados da rede social Twitter foram utilizados como experimento de avaliação. A conclusão dos autores é que o método proposto identifica ataques de engenharia social com uma precisão muito alta.

O trabalho escrito por Peplow, Thomas e AlShehhi (2021) e publicado na *International Journal of Environmental Research and Public Health* propõe analisar dados da rede social Twitter para identificar a localização aproximada das reclamações por poluição sonora nos Emirados Árabes Unidos, utilizando linguagem Python. A conclusão é que foi possível verificar a localização aproximada de onde os usuários do site de redes sociais mais postam queixas sobre poluição sonora, porém, ao contrário do que os autores pensavam, o trânsito não é umas das principais queixas de poluição sonora, mas sim, a vizinhança.

Inocio Felipe da Costa (2016) em seu trabalho de conclusão de curso pela UNESC – Universidade do Extremos Sul Catarinense, teve como objetivo utilizar algoritmos em Java e a ferramenta de análise Weka em dados da rede social Twitter, para localizar textos relacionados as doenças *Dengue, Zika e Chikungunya.* Um banco de dados relacional foi utilizado para gerenciar os dados extraídos da rede social. Como resultado, concluiu-se que a maioria das menções não tinham relação pessoal com as doenças, mas se tratava de postagens com teor cômico ou campanhas de prevenção.

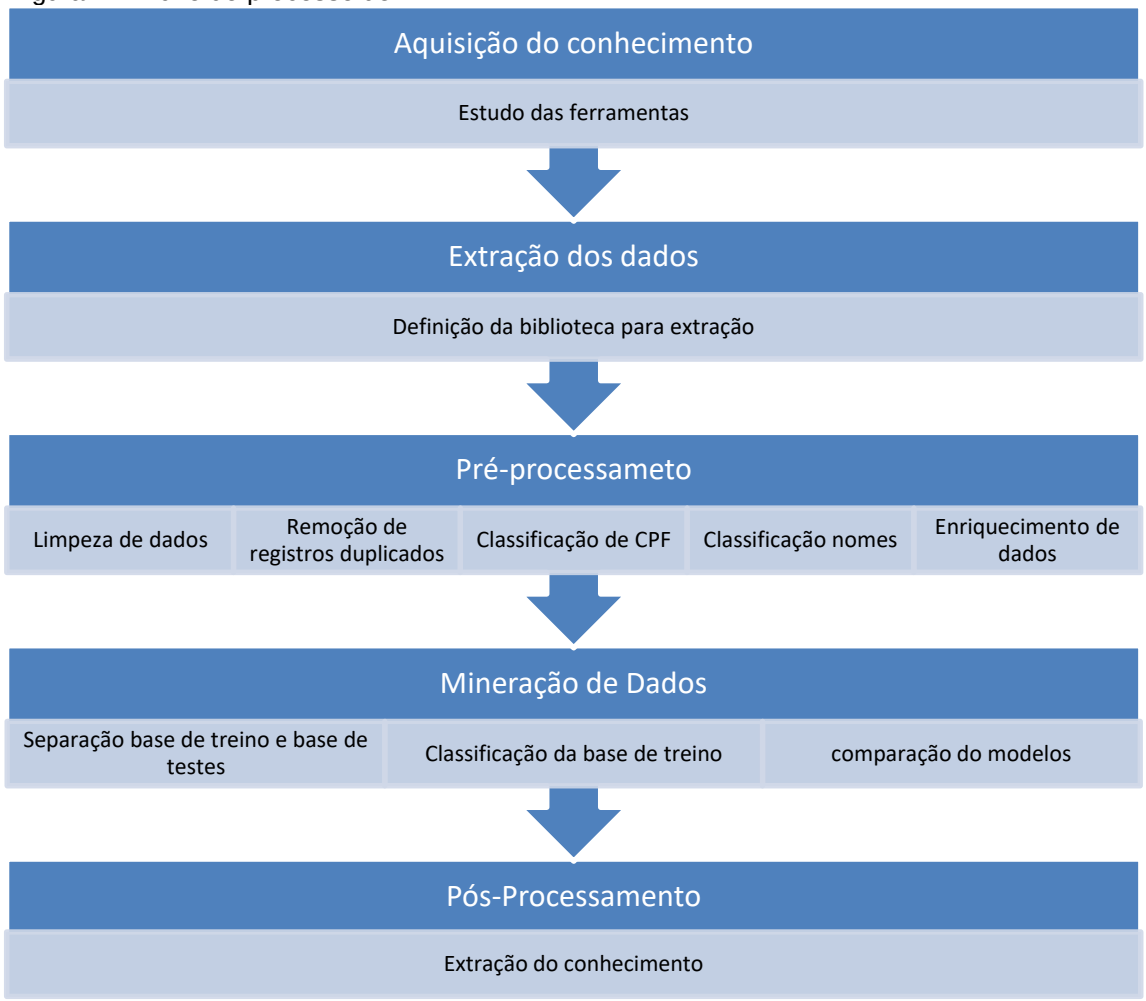
O trabalho desenvolvido por Lucas Valdati Cardoso (2014) na Universidade do Extremo Sul Catarinense – UNESC para obtenção do grau de Bacharel em Ciência da Computação tem como proposta utilizar unicamente a ferramenta Scup, para realizar todas as etapas do processo de descoberta de conhecimento. Essa ferramenta utiliza um algoritmo baseado na técnica de Máquina de Suporte Vetorial. O objetivo da pesquisa é analisar as menções aos candidatos à presidência do Brasil, no segundo turno do ano de 2014. Os resultados foram satisfatórios e o conhecimento extraído se mostrou muito próximo das pesquisas eleitorais e do resultado da eleição no segundo turno.

2 MATERIAIS E MÉTODOS

O presente trabalho consiste em uma pesquisa aplicada, de natureza secundária, bibliográfica e quanto aos objetivos ela é descritiva. Segundo Wazlawick (2021), na pesquisa descritiva busca-se obter dados mais consistentes sobre um determinado tema, mas sem a participação direta do pesquisador. Foram extraídos dados da rede social Twitter, utilizando a linguagem de programação Python e armazenados no banco de dados relacional MySQL, posteriormente submetidos a algoritmos de KDD. O Python foi escolhido para realizar as tarefas de KDD por oferecer uma grande variedade de bibliotecas voltadas para extração e mineração de dados bem como processamento de linguagem natural.

Os passos seguidos para realizar o processo de KDD foram: aquisição do conhecimento, extração da base de dados, pré-processamento, mineração de dados e pós-processamento (figura 1).

Figura 1 – Fluxo do processo de KDD.



Fonte: Elaborado pelo autor.

Na primeira etapa foi adquirido o conhecimento necessário para lidar com a linguagem Python, na sequência iniciou-se a coleta dos dados na rede social Twitter. No pré-processamento foi realizado um tratamento nos dados para que o algoritmo de mineração tenha melhores resultados. O Pós-processamento é a etapa de organização e levantamento dos resultados obtidos.

2.1 COLETA DE DADOS

Para realizar a coleta de dados na rede social Twitter é preciso ter um conta de desenvolvedor, que está disponível para qualquer pessoa que tenha uma conta na plataforma. Foi feita a solicitação, respondido alguns questionários e após a aprovação da solicitação, foi concedido acesso à API do Twitter.

Para realizar a extração, foram realizados testes em duas bibliotecas do Python: a *TwitterSearch* e a *Tweepy*. As duas traziam as informações completas em formato JSON que a API do Twitter fornece, mas, o objetivo da etapa era extrair apenas os campos que são relevantes para pesquisa. A *TwitterSearch* foi descartada pois extraiu apenas o texto resumido e, não era interessante para a pesquisa.

Os dados fornecidos pelo Twitter são no formato JSON, porém, é possível indicar qual o campo específico que se deseja extrair. Para este trabalho foram extraídos os campos:

a) verified: campo que diz se um perfil é verificado, normalmente aplicado a pessoas públicas ou empresas de grande porte;
b) geo: coordenadas geográficas que mostram a localização aproximada no momento da postagem;
c) location: campo livre onde é usuário da rede social pode indicar dados de endereço;
d) created_at: data e hora da postagem do texto;
e) screen_name: nome do usuário, é único e serve para identificar o perfil na rede social;
f) name: campo livre para em que o usuário indica o nome que aparecerá no perfil;
g) description: uma breve descrição do perfil do usuário;
h) full_text: texto completo postado pelo usuário.

A indicação campos foi intencional, pois além de trazer apenas informações relevantes para o trabalho, foi possível separar os dados em variáveis e posteriormente, no mesmo algoritmo, realizar a inserção dos dados no banco MySQL, alimentando uma tabela, onde cada variável correspondia a um campo desta tabela.

Por meio da biblioteca *Tweepy* foi possível determinar as sentenças-chave para coleta de *tweets.* Utilizando a bibliografia levantada, juntamente com notícias recentes de ataques de engenharia social, foi elaborado uma lista com cinquenta e oito sentenças, para que fosse possível ter um retorno satisfatório (quadro 1).

Quadro 1 – Sentenças-chave utilizadas para pesquisa de tweets

#criciuma	Endereço	minha casa fica	pix AND tel
analista	endereço completo	minha escola é	pix AND telefone
auxiliar	Escola	minha rua é	profissão
bairro	estou saindo	niver	relacionamento
cargo	eu trabalho	nome	rg
celular	Família	nome completo	rua
cep	Gerente	nome da mãe	sair
cpf	meu aniversário é	nome de pai	tel
cpf AND pix	meu bairro é	número de cpf	telefone
criciuma	meu cfp é	número de rg	volta
curriculo	meu endereço é	parabéns	whats
curriculum	meu filho estuda na	passeio	whatsapp
data de nascimento	meu filho estuda no	pix	zap
e-mail	meu pix é	pix AND e-mail	
email	meu rg é	pix AND email	

Fonte: Elaborado pelo autor.

Algumas sentenças possuem o operador “AND” para que as frases extraídas contenham obrigatoriamente as duas palavras. As palavras “criciuma” e “#criciuma” foram utilizadas na intenção de investigar o grau de exposição dos habitantes da região, mas devido a veiculação de notícias em rede nacional que envolviam a cidade, a pesquisa retornou *tweets* de outras regiões. O levantamento foi feito, mas usando os dados de localização digitados pelo usuário.

Foram coletados mais de novecentos mil registros entre os dias vinte e oito de agosto e seis de setembro do ano de 2021.

2.2 PRÉ-PROCESSAMENTO

De acordo com Mariano et al (2021) o cientista de dados deve analisar as inconsistências da base de dados e tomar as decisões necessárias antes de iniciar o processo de mineração. Durante o pré-processamento, a base de dados foi submetida a algoritmos de limpeza e alguns registros precisaram ser totalmente apagados. Foi realizado também um processo de enriquecimento dos dados para um melhor resultado na mineração. Esta etapa tomou a maior parte de tempo do projeto.

Inicialmente, estava planejado desenvolver um léxico para realizar o projeto utilizando Redes Neurais Artificiais e Processamento de Linguagem Natural, porém devido aos prazos, optei por trabalhar com os dados prontos que o *Twitter* já entrega.

A primeira etapa consistiu em realizar uma limpeza de dados. Foram removidos todos os caracteres diferentes de letras e números, assim como excessos de espaços em branco. Todas as letras foram convertidas para minúsculas e, no mesmo algoritmo, foram localizadas sequências de onze dígitos, em seguida, outra função verificava se essa sequência se tratava de um CPF válido e realizava uma marcação. Logo após no banco de dados MySQL, foram excluídos os registros duplicados do mesmo usuário e, aqueles que tem perfis verificados, pois é de entendimento que pessoas públicas ou empresas já têm seus dados públicos.

Por se tratar de um campo que permite ao usuário digitar qualquer palavra, um novo algoritmo precisou ser implementado para classificar os nomes. Uma base de nomes do IBGE foi utilizada, porém devido limitações de hardware, foi necessário separar apenas 10% da base, resgatar os nomes mais recorrentes e depois aplicar na base toda. No entanto, ao final do processamento, ficou entendido que por se tratar de um campo livre, pode-se haver problemas, como a pessoas escrevendo um nome fictício totalmente diferente do nome real ou, utilizar caracteres especiais na composição do nome, dificultando a busca do algoritmo. O campo local teve a mesma consideração, mas ambos prosseguiram no projeto para fins de teste de classificação.

Para o projeto, foram considerados os seguintes itens como possíveis casos de exposição de informações pessoais: CPF, coordenadas geográficas, informação no campo local, telefone, nome e aniversário.

Como o campo analisado seria apenas o texto, foi necessário enriquecer esse campo com os dados acima. Para tanto, foram criadas palavras-chave que viriam antes do texto, são elas: +cpf, +geo, +niver, +local, +nome, +tel, onde:

a) CPF: +cpf;
b) Coordenadas geográficas: +geo;
c) Informações no campo local: + local;
d) Telefone: +tel;
e) Nome: +nome;
f) Aniversário: +niver.

Dados como CPF, telefone e aniversário foram localizados diretamente dentro do texto, porém, nome, local e coordenadas geográficas são disponibilizados pela API do Twitter no momento da extração.

Ao final do pré-processamento, a base de dados foi reduzida para 792655 registros de usuários diferentes.

2.3 MINERAÇÃO DE DADOS

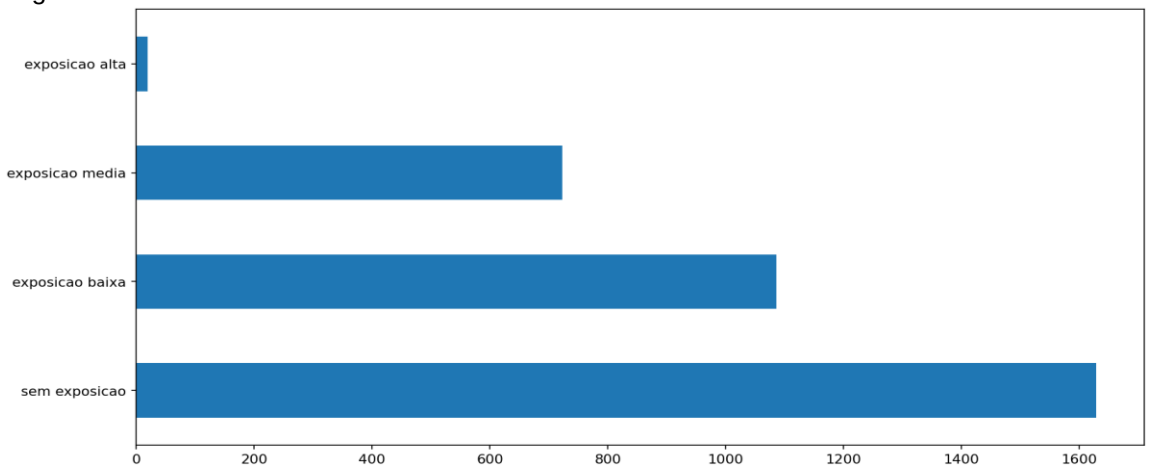
Com os dados pré-processados, iniciou-se a etapa de mineração destes dados. A intenção da mineração de dados é encontrar padrões entre os dados, de maneira que formem conhecimento para que sejam usados para tomada de decisão. Para isso são utilizados algoritmos inteligentes que trabalham de maneira automática (SASSI, 2006).

O primeiro passo foi separar uma base de dados de teste e outra de treino. Foi separado uma base de dados de 15000 registros para teste e outra com 5000 registros para treino. A base de treino foi classificada automaticamente por um algoritmo, que aplicava um peso para cada palavra-chave encontrada. Conforme a abordagem de Silva (2013), foram aplicados pesos maiores (3 pontos) para dados de cadastro (Nome e CPF) e pesos normais (2 pontos) para os demais dados. Os textos foram classificados em: exposição alta, exposição baixa, exposição média e sem exposição, onde zero é sem exposição, menor que três pontos é exposição baixa, entre dois e seis é exposição média e seis ou mais (que é soma dos dois dados cadastrais no mesmo texto) é considerado exposição alta.

Para que o algoritmo de classificação tenha um bom desempenho, os dados precisam estar balanceados. Portanto, a base de treino foi reduzida para que todas as classificações tenham números semelhantes, porém, devido ao baixo número de registros classificados como “exposição alta”, optou-se por realizar o balanceamento por proporção, deixando aproximadamente a mesma diferença entre

as classes (figura 2). Após o balanceamento, a base de treino ficou com 3457 registros.

Figura 2 – Gráfico de balanceamento da base de treino.



Fonte: Elaborado pelo autor.

O Python possui uma biblioteca chamada *sklearn* que traz em sua implementação vários métodos de mineração de dados. Para o presente trabalho, dois métodos tradicionais foram testados: *Multinomial Naive Bayes* (MNB) e *Máquinas de Vetores Suporte* (Support Vector Machine - SVM). Para treinar o algoritmo, foram entregues os textos e suas respectivas classificações na base de treino. Os resultados do método SVM foram superiores ao MNB. Enquanto MNB teve um total de 1357 falso positivos, o SVM teve apenas 33 em um total de 3457 registros. Esse resultado, junto a outras métricas definiram qual seria o algoritmo utilizado para classificar a base principal. Abaixo, a tabela de métricas dos modelos MNB (tabela 1) e SVM (tabela 2).

Tabela 1 – Métricas do Modelo MNB

	precisão	sensibilidade	especificidade	F1-score	support
Exposição Alta	100,00%	5,00%	100,00%	10,00%	19
Exposição Baixa	50,00%	69,00%	64,38%	58,00%	1086
Exposição Média	60,00%	83,00%	78,60%	69,00%	723
Sem Exposição	76,00%	45,00%	85,44%	57,00%	1629

Fonte: Elaborado pelo autor.

Tabela 2 – Métricas do Modelo SVM

	precisão	sensibilidade	especificidade	F1-score	support
Exposição Alta	100,00%	63,00%	100,00%	77,00%	19
Exposição Baixa	99,00%	100,00%	99,32%	99,00%	1086
Exposição Média	99,00%	97,00%	99,74%	98,00%	723
Sem Exposição	99,00%	100,00%	99,44%	100,00%	1629

Fonte: Elaborado pelo autor.

O método MNB teve uma acurácia de 60,25% enquanto o método SVM teve acurácia de 99,04%. A sensibilidade na classificação de “exposição alta” ficou um pouco a abaixo nos dois modelos, pois havia em baixa quantidade de registros com esta classificação. No modelo MNB a sensibilidade ficou com apenas 5%, pois havia um único registro verdadeiro positivo enquanto 18 eram falso negativos. No modelo SVM, o resultado foi um pouco melhor porque o algoritmo classificou 12 com verdadeiro positivos, diante de apenas 7 falso negativos. Outro ponto importante que ficou abaixo da média, foi a sensibilidade do modelo MNB ao classificar registros como “sem exposição”. Dos 1629 registro pré-classificados como “sem exposição”, o modelo considerou apenas 739 como verdadeiros positivos. Isso representa menos da metade dos registros.

Nos testes, o SVM mostrou-se lento durante o treino e a classificação, segundo Han e Kamber (2006), esse processo mais demorado é absolutamente normal devido sua capacidade de modelar limites complexos não lineares, o que torna o método altamente preciso (tradução nossa). Analisando as métricas obtidas, o modelo SVM foi selecionado para prosseguir com a classificação da base principal.

Um algoritmo simples foi rodado na sequência para que fosse possível uma melhor visualização das associações entre os dados extraídos. Esse algoritmo realizou uma tarefa bem simples, verificar as palavras-chave nos textos e contar quantas vezes elas apareciam juntas.

2.3 PÓS-PROCESSAMENTO

Com os dados devidamente classificados, foi iniciado a validação do conhecimento, que é um processo em que os conhecimentos extraídos são avaliados, de modo que a qualidade desses conhecimentos irá determinar se o processo de mineração será refeito ou finalizado (CASTRO, 2016).

Os textos classificados tiveram um resultado satisfatório. Utilizando o MySQL Workbench, foram realizadas algumas consultas para extrair os resultados e posteriormente, inseridos em uma planilha de textos para melhor visualização.

3 RESULTADOS E DISCUSSÃO

Por meio das ferramentas utilizadas (MySQL, Python e planilhas de texto), foi possível extrair resultados satisfatórios.

O primeiro resultado, mostra que, a maioria dos usuários (99,992%) possui algum tipo de exposição, porém, 89,59% têm exposição baixa e apenas 0,091% têm exposição considerada alta (tabela 3).

Tabela 3 – Resultados por classificação

Classificação	Quantidade	%
Exposição baixa	710152	89,592%
Exposição média	81723	10,310%
Sem exposição	61	0,008%
Exposição alta	719	0,091%
TOTAL	792655	

Fonte: Elaborado pelo autor.

Em seguida, foi feito o levantamento por ocorrência da palavra-chave. Pôde-se notar que o local, mesmo que seja um local fictício, é um dos campos que os usuários não costumam deixar em branco em seus perfis (tabela 4).

Tabela 4 – Quantidade por palavras-chave

Palavra-chave	Quantidade	%
Local	790744	99,76%
Nome	50765	6,40%
Aniversário	30376	3,83%
Telefone	4583	0,58%
CPF	1214	0,15%
Geolocalização	388	0,05%

Fonte: Elaborado pelo autor.

O nome, por ser uma identificação amigável do usuário dentro da rede social, tem uma recorrência de 99,99%, porém como nome reconhecível e comum no Brasil, foram classificados apenas 6,40%. Esse resultado mostra que a maioria dos usuários procura usar algum apelido, usar o campo para escrever alguma mensagem ou o próprio nome utilizando uma sequência de caracteres especiais que lembram o verdadeiro nome, como "*PΣDЯΘ*" por exemplo. Outro caso, é dos profissionais autônomos, como advogados, esteticistas, eletricistas, entre outros, que usam a rede social para divulgar seus trabalhos. Apesar de serem nomes completos, não se

configura como uma informação que exponha a pessoa, pois trata-se de uma forma de divulgação do serviço prestado.

Quanto a Geolocalização, percebeu-se que não é muito comum sua utilização no território brasileiro, pois representa apenas 0,05% do total de dados analisados. A maioria dos resultados encontrados são de estabelecimentos comerciais que divulgam sua localização exata para os seus clientes. Devido à baixa ocorrência a geolocalização aparece pouco relacionada com os outros dados.

Geolocalização e CPF são dados que tem maior possibilidade de serem reais, o primeiro é um campo que está preenchido, ou em branco, portanto se estiver preenchido obrigatoriamente terá as coordenadas geográficas de quem postou. Quanto ao CPF, foi feito uma verificação, porém, normalmente estão associados a pedidos de doações para beneficiar pessoas carentes via *pix*. São numerações válidas que podem ser gerados facilmente com uma breve pesquisa na Internet. Porém, tratando-se de um pedido real de doação, ou talvez uma tentativa de extorsão, é muito provável que sejam numerações reais válidas.

Para localizar registros que citavam números de telefones, foi considerado todo texto que havia ligação com pix, tinha um número de onze dígitos, porém não era uma numeração de CPF válida. Esse método não foi tão eficiente, pois para que o algoritmo classifique como telefone, ele deve ser um CPF não válido e isso pode acontecer por falha na digitação ou por algum usuário ter digitado uma numeração aleatória, sem uma lógica.

Os dados marcados com a palavra-chave *aniversário* são os mais inconsistentes, pois bastava ter as palavras niver ou aniversário no texto. Todavia, diversos registros citavam a data de aniversário direta ou indiretamente (quadro 2).

Quadro 2 – Menção direta x Menção indireta

Menção direta	**Menção indireta**
meu aniversário é dia 30 de setembro	faltam 2 dias pro meu niver e eu tô sem ânimo nenhum

Fonte: Elaborado pelo autor.

A análise feita usando dados de usuário da cidade de Criciúma trouxe 205 resultados, que corresponde a 0,026% do total. Dentre estes, 85,3% são classificados como exposição baixa, 13,2% exposição média e 1,5% exposição alta. Esses

números são bem próximo do resultado total e mostram que o há uma distribuição equilibrada nas médias pelas grandes cidades do Brasil.

Quanto as sentenças-chave usadas na extração, a que mais trouxe exposição alta foi “*niver*”, seguida de “*meu aniversário é*”. As duas juntas somam 320 registros e representam 44,5% do total de casos de exposição alta. Outro dado que chamou atenção nos casos de exposição alta foram as sentenças que que possuem a palavra “*pix*”, todas as sentenças juntas que foram classificadas como exposição alta e tinham a palavra “*pix*”, somam 227 registros, ou 31,6% do total de registros classificados como exposição alta.

No estudo de Lansley, Mouton, Kapetanakis & Polatidis (2020), a ideia assemelha-se ao presente trabalho, porém, com objetivos diferentes. Eles se comprometeram em usar processamento de linguagem natural e redes neurais artificiais, para monitorar conversas de texto em busca de frases que se configurem em um ataque de engenharia social. Enquanto o presente trabalho procura por exposição de dados pessoais que possam ser utilizados para planejar ataques de engenharia social. A princípio, um dos objetivos específicos, seria implementar processamento de linguagem natural para analisar as frases, o que tornaria os trabalhos muito mais semelhantes. Devido ao prazo, optou-se por implementar algoritmos mais simples.

A utilização do modelo SMV foi explorada também nos trabalhos de Cardoso (2014) e Costa (2016). O primeiro explorou a ferramenta SCUP, que é um software comercial popular no meio empresarial, para monitoramento das marcas comerciais nas redes sociais. A aplicação da ferramenta, no momento certo e com os termos certos, trouxe bons resultados e o trabalho teve uma boa conclusão. Costa (2016) utilizou o modelo SVM com a ferramenta WEKA. Seu trabalho teve boa execução por parte do método, porém os resultados não foram os esperados. Em ambos os trabalhos, o modelo SVM teve performance satisfatória.

4 CONCLUSÃO

Após análise dos resultados, ficou evidente que grande parte dos registros se tratava de informações com baixa credibilidade, portanto, há uma grande possibilidade destes resultados não representarem a realidade com uma margem de erro baixa. Nesse ponto, o projeto assemelha-se ao trabalho de Inocio Felipe da Costa

(2016), que também teve vários falsos positivos, porém, mesmo que a contabilização dos resultados tenha uma margem de erro alta, ficou entendido que há dados pessoais expostos.

A importância de um léxico dentro do tema de engenharia social ficou evidente após análise dos resultados da verificação de CPF e nome, pois apesar da simplicidade dos algoritmos, realizaram uma validação precisa. No trabalho de pesquisa de Lansley, Mouton, Kapetanakis & Polatidis (2020), implementou-se um léxico dentro do tema e os autores concluíram que os resultados foram satisfatórios.

Verificando a validação realizada no campo nome, percebeu-se que a falta de uma validação também nos dados do campo local, acabou classificando muito registros como exposição baixa onde deveria ser sem exposição. A falta dessa validação poderia mudar completamente os resultados.

As maiores dificuldades foram encontradas durante o pré-processamento. A cada teste realizado, entendia-se que era necessário realizar algum processo a mais para que os dados tivessem uma melhor classificação pelo algoritmo. Foi a etapa mais demorada do projeto, pois foi necessário desenvolver várias funções para deixar o texto pronto para a etapa de mineração de dados. Grande parte das funções não mostravam resultados na primeira tentativa e por muitas vezes foi necessário reformular o algoritmo. Nesse ponto foi importante ter uma base de testes e um backup da base principal.

As ferramentas escolhidas para realizar o trabalho tiveram um ótimo desempenho. A versatilidade do Python, com o seu grande número de bibliotecas voltadas para extração e aplicação de algoritmos de classificação, ajudou a realizar grande parte do projeto. A utilização do MySQL como gerenciador de banco de dados relacional foi eficiente para armazenar os dados tratados pelo Python e teve um papel fundamental na análise dos resultados. A união das ferramentas, junto aos algoritmos que foram implementados, foram fundamentais para alcançar o objetivo do projeto. Concluiu-se, portanto, que, apesar da baixa credibilidade dos resultados, foi possível encontrar dados pessoais expostos na rede social Twitter e alcançar o objetivo do projeto.

Com as experiências adquiridas ao longo do projeto, assim como os resultados obtidos, os seguintes pontos são recomendados para trabalhos futuros: desenvolver um léxico sobre o tema de engenharia social buscando por outros dados pessoais e frases que não foram explorados nesse trabalho, realizar validação do

campo local, testar novos métodos de classificação, utilizar o caractere undescore no lugar do sinal de positivo na frente das palavras chaves, pois o sinal de positivo apresenta problemas em algumas bibliotecas do Python. Considerar expandir para outros sites de redes sociais e realizar análise de imagens, áudios ou vídeos.

REFERÊNCIAS

BERRY, Michael J. A.; LINOFF, Gordon. **Data mining techniques**: for marketing, sales, and customer relationship management. Indianápolis: Wiley Publishing, 2004.

CARPIM, S. M. **A era do exibicionismo digital**: o sentido da proliferação da selfie nas redes sociais. São Paulo: Escola de Comunicações e Artes/ECA-USP, 2014.

CASTRO, Leandro Nunes de. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CENTRO DE ESTUDOS, RESPOSTA E TRATAMENTO DE INCIDENTES DE SEGURANÇA NO BRASIL. **Cartilha de segurança para Internet**. Disponível em: <http://cartilha.cert.br/livro/cartilhaseguranca-Internet.pdf>. Acesso em: 16 de maio de 2021.

COSTA, Inocio F. **Mineração De Dados Na Rede Social Twitter A Respeito De Casos Das Doenças Dengue, Zika E Chikungunya**. Trabalho de Conclusão de Curso (Graduação). Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense - UNESC. Criciúma, p. 77. 2016.

MARIANO, Diego et. al. **Data Mining**. 1. ed. Porto Alegre: Sagah, 2020.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data mining**: um guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005. 261 p.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and Techniques. San Francisco, CA, 3rd edition, Morgan Kaufmann, 2006.

LANSLEY, Merton; MOUTON, Francois; KAPETANAKIS, Stelios; POLATIDIS, Nikolaos. **SEADer++**: social engineering attack detection in online environments using machine learning. Journal of Information and Telecommunication, v.4, n. 3, p. 346, 10 de abril de 2021.

MITNICK, Kevin; SIMON, William. **A arte de enganar**. São Paulo: Pearson Education, 2003. 286 p.

PEPLOW, A.; THOMAS, J.; ALSHEHHI, A. **Noise Annoyance in the UAE**: A Twitter Case Study via a Data-Mining Approach. International Journal of Environmental Research and Public Health, v. 18, n. 4, p. 2198, 23 de fevereiro de 2021.

RECUERO, Raquel. **Introdução à análise de redes sociais**. Salvador: EDUFBA, 2017.

SASSI, Renato José. **Uma Arquitetura Híbrida para Descoberta de Conhecimento em Bases de Dados**: Teoria dos Rough Sets e Redes Neurais Artificiais Mapas Auto-organizáveis. Tese de Doutorado. Escola Técnica da Universidade de São Paulo, 2006.

SILVA, N. B. X.; ARAÚJO, W. J. de; AZEVEDO, P. M. de. **Engenharia social nas redes sociais online**: um estudo de caso sobre a exposição de informações pessoais e a necessidade de estratégias de segurança da informação. Revista Ibero-Americana de Ciência da Informação, [S. l.], v. 6, n. 2, p. 37–55, 2013. DOI: 10.26512/rici.v6.n2.2013.1782. Disponível em: <https://periodicos.unb.br/index.php/RICI/article/view/1782>. Acesso em: 02 de novembro de 2021.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. 2. ed Rio de Janeiro: Elsevier, 2021.