Faculdade de Tecnologia de Americana "Ministro Ralph Biasi"

Curso Superior de Tecnologia em Segurança da Informação

Bruno Eleodoro Roza

Marco Antonio Gonçalves Pegoraro

Classificador De Phishing Utilizando Algoritmo de Naive Bayes

Americana, SP

2020

Faculdade de Tecnologia de Americana "Ministro Ralph Biasi"

Curso Superior de Tecnologia em Segurança da Informação

Bruno Eleodoro Roza

Marco Antonio Gonçalves Pegoraro

Classificador De Phishing Utilizando Algoritmo de Naive Bayes

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Segurança da Informação, sob a orientação Professor Me. Henri Alves Godoy.

Área de concentração: Ciência da Computação

Americana, SP.

2020

FICHA CATALOGRÁFICA – Biblioteca Fatec Americana - CEETEPS
Dados Internacionais de Catalogação-na-fonte

R797c ROZA, Bruno Eleodoro

Classificador de phishing utilizando algoritmo de Naive Bayes. / Bruno Eleodoro Roza, Marco Antonio Gonçalves Pegoraro. – Americana, 2020.

37f.

Monografia (Curso Superior de Tecnologia em Segurança da Informação) - - Faculdade de Tecnologia de Americana – Centro Estadual de Educação Tecnológica Paula Souza

Orientador: Prof. Ms. Henri Alves Godoy

1 Inteligência artificial 2. Aprendizado de máquina Segurança em sistemas de informação I. PEGORARO, Marco Antonio Gonçalves II. GODOY, Henri Alves III. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana

CDU: 007.52

Bruno Eleodoro Roza

Marco Antonio Gonçalves Pegoraro

Classificador De Phishing Utilizando Algoritmo de Naive Bayes

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Tecnólogo em Segurança da Informação pelo CEETEPS/Faculdade de Tecnologia – FATEC/ Americana.

Área de concentração: Ciência da Computação.

Americana, de julho de 2020.

Banca Examinadora:

_____
Professor Me. Henri Alves Godoy
Mestre
Fatec de Americana – Ministro Ralph Biasi

_____
José Martins Junior
Mestre
Fatec de Americana – Ministro Ralph Biasi

_____
Cesar Augusto Della Piazza
Mestre
Fatec de Americana – Ministro Ralph Biasi

AGRADECIMENTOS

Agradecimentos ao nosso orientador Henri Alves Godoy, amigos e família que nos apoiaram na entrega do trabalho de graduação.

DEDICATÓRIA

A nossa família e amigos que nos apoiaram tanto no decorrer do curso quanto a entrega do trabalho de graduação.

RESUMO

O presente trabalho tem como objetivo abordar a aplicação na prática do aprendizado de máquina para analisar textos eletrônicos com teor de fraude, em outras palavras, o *phishing*. Textos com o objetivo de enganar as pessoas para obter informações confidenciais são um método antigo que era praticado fora do meio digital, e trazia grandes prejuízos para as empresas, já hoje em dia com os meios digitais, essa propagação de mensagens falsas aumentou drasticamente e vem causando ainda mais prejuízos para pessoas físicas e pessoas jurídicas. Para demonstrar a eficiência do classificador, foi feito uma coleta de diversos *e-mails* falsos do catálogo de fraudes, e também foi utilizado o algoritmo de Naive Bayes para fazer o treinamento do classificador. Para realizar o estudo de caso e obter os resultados do experimento, foi utilizado a linguagem de programação *JavaScript* para construir uma aplicação de linha de comando, que realiza o treinamento do classificador e também faz a leitura de textos, fornecendo como saída a porcentagem de cada texto ser considerado uma fraude ou não. Por fim na conclusão, foi explicado o que é necessário para obter resultados ainda melhores e como combater com maior eficácia esse problema que vem causando prejuízos de milhões de dólares no mundo todo.

Palavras Chave: Aprendizado de máquina; fraude; Segurança;

ABSTRACT

The present work will demonstrate a command line application with machine learning to analyze electronic texts with fraud content, in other words, phishing. Texts with the objective of deceiving people to obtain confidential information are an old method that was practiced outside the digital environment, and brought great damage to companies, even today with all the technology, this spread of false messages has increased dramatically and comes causing even more damage to individuals and companies. To demonstrate the efficiency of the classifier, a collection of several fake emails from the "catalogo de fraudes" was made, and the Naive Bayes algorithm was also used to train the classifier. To perform the case study and obtain the results of the experiment, it was used the JavaScript programming language to build a command line application, which performs the classifier training and also reads texts, providing the percentage of each text as an output. be considered a fraud or not. Finally, in the conclusion, it is explained what is needed to obtain even better results and how to combat more effectively this problem that has caused millions of dollars in losses worldwide.

Keywords: *Machine learning; fraud; Security;*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

É inevitável de que com o avanço da difusão da internet no mundo, diversas preocupações a respeito da segurança viriam a tona, uma dessas preocupações é o disseminamento de mensagens falsas com o objetivo de roubar dados de vítimas para, em seguida, aplicar golpes financeiros, esse tipo de golpe é comumente chamado de *phishing*, essa expressão surgiu da palavra "*fishing*", que no inglês, significa "pescaria", ou seja, os criminosos utilizam esta técnica para "pescar" os dados das vítimas que "mordem o anzol" lançado pelo *phisher* ("pescador"), nome que é dado a quem executa um *phishing* (SIGNIFICADOS, 2014).

Esta monografia tem como objetivo, fazer uma introdução sobre o conceito do *phishing*, será demonstrado quais foram os primeiros relatos dessa prática de golpe, como ela se situa hoje, quais são os principais tipos de golpes via *e-mail*. Por fim, será demonstrado como o algoritmo de Naive Bayes foi utilizado para treinar uma inteligência artificial capaz de detectar, a partir de um texto, a porcentagem do texto ser, de fato, uma mensagem verdadeira, ou ser uma tentativa de *phishing*.

No capítulo dois, será explicado a história e o surgimento do *phishing*, como ele surgiu em 1995 no AOL[1] e desde então, a prática foi cada vez mais utilizada por criminosos virtuais.

No capítulo três, será demonstrado a teoria por trás do algoritmo de Naive Bayes, e como ele pode ser utilizado para realizar a classificação de probabilidade de uma palavra ser verdadeira ou falsa baseada em um modelo pré-treinado, e como será feito um algoritmo na programação capaz de realizar os cálculos necessários para a metodologia.

No capítulo quatro, será mostrado como foi realizado a coleta de dados de fraudes utilizando o portal “Catálogo de Fraudes” e, em seguida, o desenvolvimento de um *script*[2] utilizado para realizar o treinamento do algoritmo baseados nas fraudes coletadas para, por fim, utilizar o resultado do treinamento para classificar diferentes mensagens para coletar a porcentagem delas serem fraudes ou não.

Por fim, no quinto capítulo, é mostrado uma conclusão baseada no resultado dos testes do algoritmo, além de um resumo do projeto e sugestões que poderiam ser utilizadas para trabalhos futuros.

[1] America Online, provedor de internet muito popular na década de 90
[2] Programa geralmente feio para automatizar tarefas

2. HISTÓRIA DO PHISHING E SUAS FORMAS

Fraudes são uma categoria de crimes que sempre existiram na história da civilização moderna, fraudes podem ser classificados como estelionato segundo o Art. 171 da Lei nº 2848, de 7 de dezembro de 1940 (Brasil, 1940): “Obter, para si ou para outrem, vantagem ilícita, em prejuízo alheio, induzindo ou mantendo alguém em erro, mediante artifício, ardil, ou qualquer outro meio fraudulento”. Com a chegada da internet como é conhecida hoje, seria inevitável que criminosos assim começassem a surgir no meio virtual, os chamados cibercriminosos.

Segundo o site phishing.org, fraudes na internet sempre existiram, porém o termo *phishing* começou a se popularizar devido a crackers que, em 1995, utilizavam de técnicas para roubar senhas de usuários do principal provedor de internet dos estados unidos na época AOL, para, em seguida, utilizar algoritmos para criar números de cartões aleatórios. Embora os números de cartões acertados fossem mínimos, já foi o suficiente para causar bastante dano, programas especiais como o AOHell também era utilizado para facilitar o processo.

A AOL conseguiu parar os ataques no mesmo ano utilizando técnicas de validações de cartões de crédito mais avançadas para prevenir o uso aleatório de números de cartões de crédito.

Embora os ataques de cartões de créditos falsos tenham acabado, os *Phishers* continuaram suas práticas se passando por funcionários da AOL, para enviar mensagens instantâneas e *e-mails* para usuários do sistema com o intuito de obter informações do faturamento dos usuários, logo, muitas pessoas caiam pois não existia nada parecido com esse golpe anteriormente.

2.1 O que é *phishing*

Segundo o Site oficial do *Avast*[3], *Phishing* pode ser definido como "uma maneira desonesta que cibercriminosos usam para enganar você a revelar informações pessoais,

[3] Família de softwares antivírus

como senhas ou cartão de crédito, CPF e número de contas bancárias.". Ou seja, *phishing* é um tipo de ataque realizado por cibercriminosos com o intuito não de atacar sistemas da computação, e sim, as pessoas que estão utilizando esses sistemas, utilizando táticas de engenharia social.

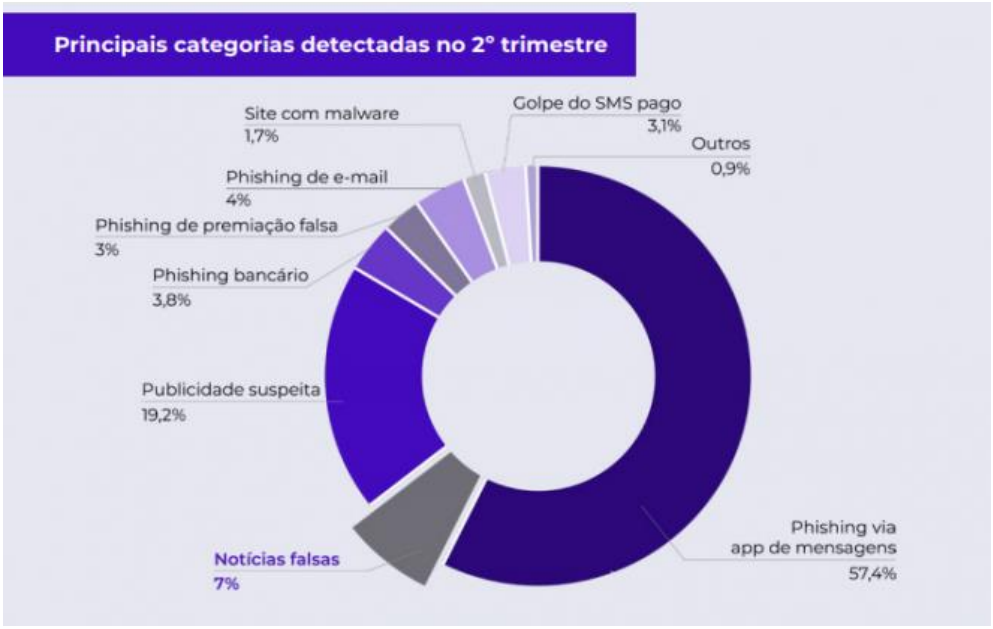
De acordo com o site do Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil, Os cibercriminosos utilizam de diversas maneiras para fazer esses ataques aos usuários de sistema, como por exemplo, Páginas falsas de comércio eletrônico ou Internet Banking, Páginas falsas de comércio eletrônico ou Internet Banking, Mensagens contendo formulários, Mensagens contendo links para códigos maliciosos, Solicitação de recadastramento, etc.

O Canaltech[4], realizou uma matéria em 2018 demonstrando que os ataques de *phishing* quase que dobraram em relação ao ano passado, a pesquisa foi realizada pelo laboratório de cibersegurança da *PSafe*[5]. O estudo também mostra os principais ataques que ocorrem, sendo o principal deles o *phishing* via app de mensagens, seguido por publicidades suspeitas e notícias falsas.

É possível observar na Figura 1 o gráfico contendo a porcentagem das principais fraudes detectadas no segundo trimestre, em primeiro lugar fica *phishing* via app de mensageria com 57,4%, em segundo e terceiro lugar, está publicidade suspeita e notícias falsas totalizando, respectivamente, 19,2% e 7%.

[4] Portal de notícias relacionadas a tecnologia
[5] Startup de origem brasileira que desenvolve aplicativos da categoria ferramentas para telefones celulares

Figura 1 - Principais categorias detectadas no 2º trimestre



Fonte: Yahoo. Disponível em https://br.financas.yahoo.com/noticias/n%C3%BAmero-ataques-cibern%C3%A9ticos-no-brasil-112000302.html. Acesso em: 4 jun 2020

De acordo com o portal de notícias Canaltech, “os golpistas escolhem os seus alvos e definem qual o objetivo do ataque”, ou seja, fazem uma pesquisa para saber quais são o público alvo mais suscetível a cair no golpe e, a partir daí, iniciam o ataque. Segundo o Canaltech, os golpes de *phishing* geralmente tem como objetivo roubar “dados pessoais, dados bancários, criar contas em nome da vítima, transferir dinheiro para uma outra conta bancária ou diversos outros tipos de fraudes.”

2.2 Quais são as formas de *phishing*

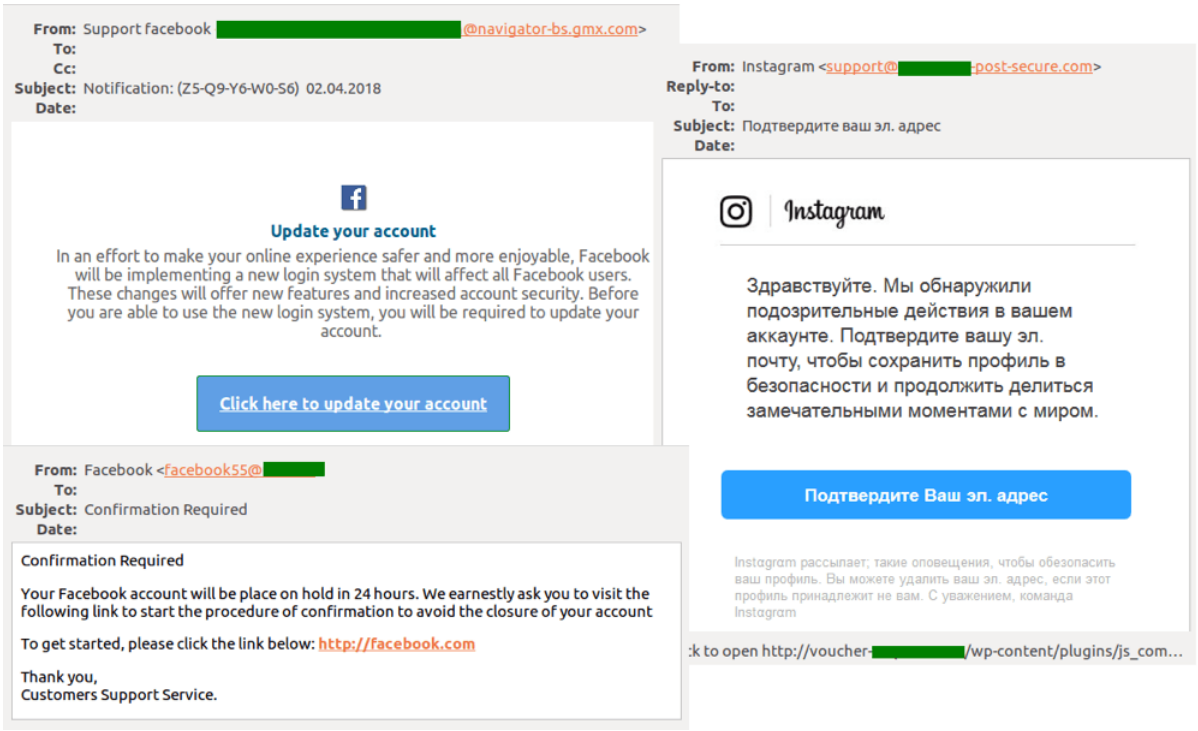
Como visto anteriormente, existem diversas maneiras nas quais criminosos se utilizam para aplicar golpes na internet, o blog Kaspersky Daily fez uma lista com os 5 principais golpes de *phishing* mais comuns, dentre elas estão: Notificações falsas de mídias sociais, *Banking phishing*, Notificações falsas de serviços ou lojas populares, Notificações falsas de serviços de *e-mail* e Golpe do “Príncipe Nigeriano”.

É importante ressaltar que além de correio eletrônico, as mensagens podem ser enviadas para as vítimas através de outras maneiras, como por exemplo “por

meio de sites, *malwares*, VoIP, ou aplicativos de mensagens instantâneas” (Canaltech, 2020).

Notificações falsas podem chegar de diversas maneiras, uma delas é via um link enviado por correio eletrônico no qual a vítima, ao clicar, é redirecionada para uma página falsa que se passa por uma página de *login* de uma rede social, como o *Facebook* ou *Twitter*. Ao digitar os dados, a vítima então é redirecionada para o site verdadeiro, porém, o *login* e senha digitados são enviados para um servidor do golpista onde ele terá acesso ao perfil do usuário. Na Figura 2 é mostrado três exemplos de notificações falsas por redes sociais.

Figura 2 - Notificações falsas de redes sociais



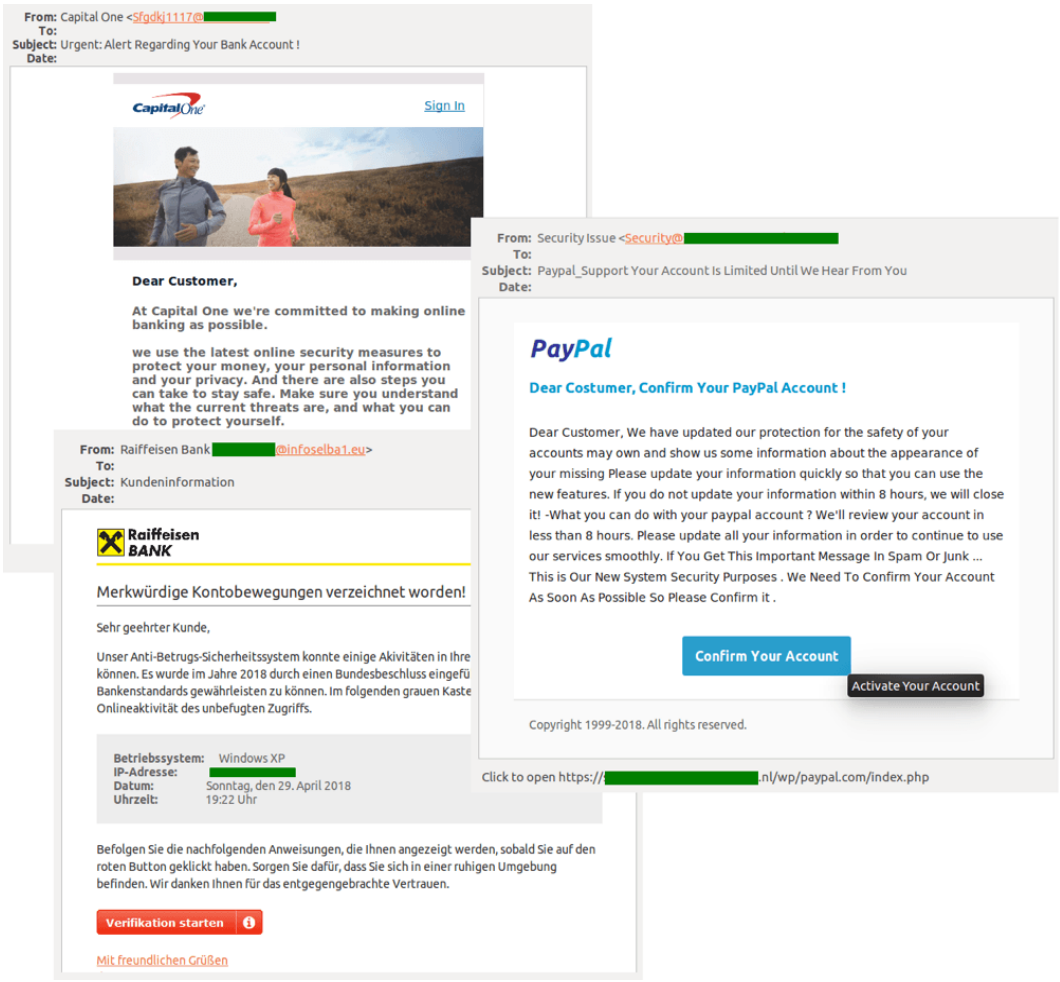
Fonte: Kaspersky. Disponível em: https://www.kaspersky.com.br/blog/phishing-spam-hooks/11174. Acesso em: 4 jun 2020

O *phishing* financeiro é utilizado quando o golpista busca obter informações sobre os dados bancários da vítima, é a modalidade de *phishing* mais utilizada hoje em dia, geralmente o conteúdo das mensagens incluem temas como “bloqueio de conta” ou “atividade suspeita”.

Ao abrir a página, a vítima encontra então, campos para a inserção dos dados do cartão de crédito, assim que os dados são inseridos, os criminosos então utilizam desses dados para realizar compras, transferências ou saque do dinheiro da conta da

vítima. Na Figura 3, é possível observar três exemplos de notificações de instituições financeiras.

Figura 3 - Notificações de instituições financeiras



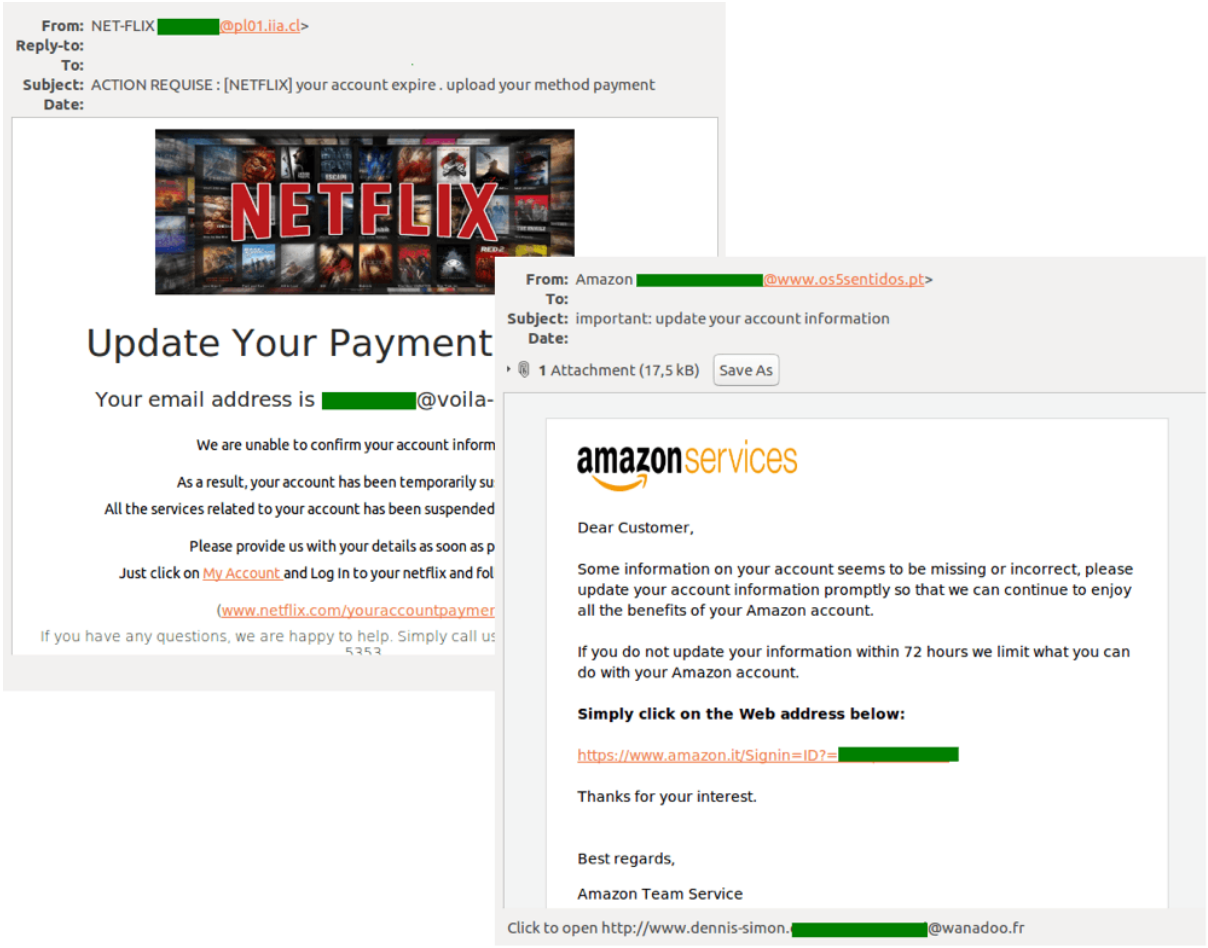
Fonte: Kaspersky. Disponível em: https://www.kaspersky.com.br/blog/phishing-spam-hooks/11174. Acesso em: 4 jun 2020

Segundo o Kaspersky, notificações falsas de serviços tem o mesmo princípio dos golpes anteriores, porém o golpe tenta se passar por serviços comuns bastante utilizados pelas pessoas, como Netflix[6], Amazon[7], Mercado Livre[8], etc. A página falsa geralmente diz para atualizar as opções de pagamento da vítima induzindo a mesma a colocar os dados do cartão na página. Na Figura 4, é possível observar dois exemplos de

[6] Empresa provedora de serviços de vídeos digitais.
[7] Empresa de tecnologia focada em comércio electrónico, computação em nuvem, streaming digital e inteligência artificial.
[8] Empresa argentina de tecnologia que oferece soluções de comércio eletrônico.

norificações falsas de serviços bem conhecidos, um sendo da Netflix e outro sendo da Amazon.

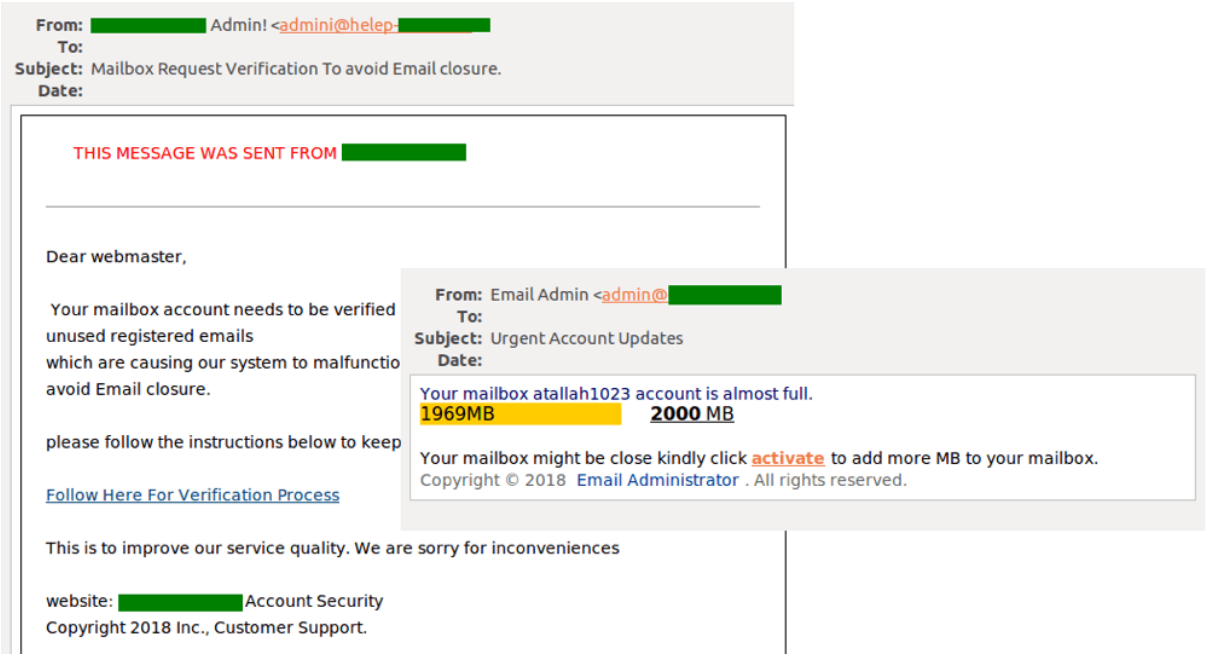
Figura 4 - Exemplo de notificações falsas recebidas via correio eletrônico



Fonte: Kaspersky. Disponível em: https://www.kaspersky.com.br/blog/phishing-spam-hooks/11174. Acesso em: 4 jun. 2020

Segundo o Kaspersky, notificações falsas de serviços de *e-mail* é um tipo de golpe aplicado por criminosos para obter o nome de usuário e a senha do mesmo para um determinado serviço de correio eletrônico, a página falsa geralmente fala de um armazenamento cheio ou então sobre um possível aumento no tamanho do armazenamento da conta da vítima por um preço atrativo e pouco suspeito. Na Figura 5, é mostrado dois exemplos de notificações falsas de correio eletrônico.

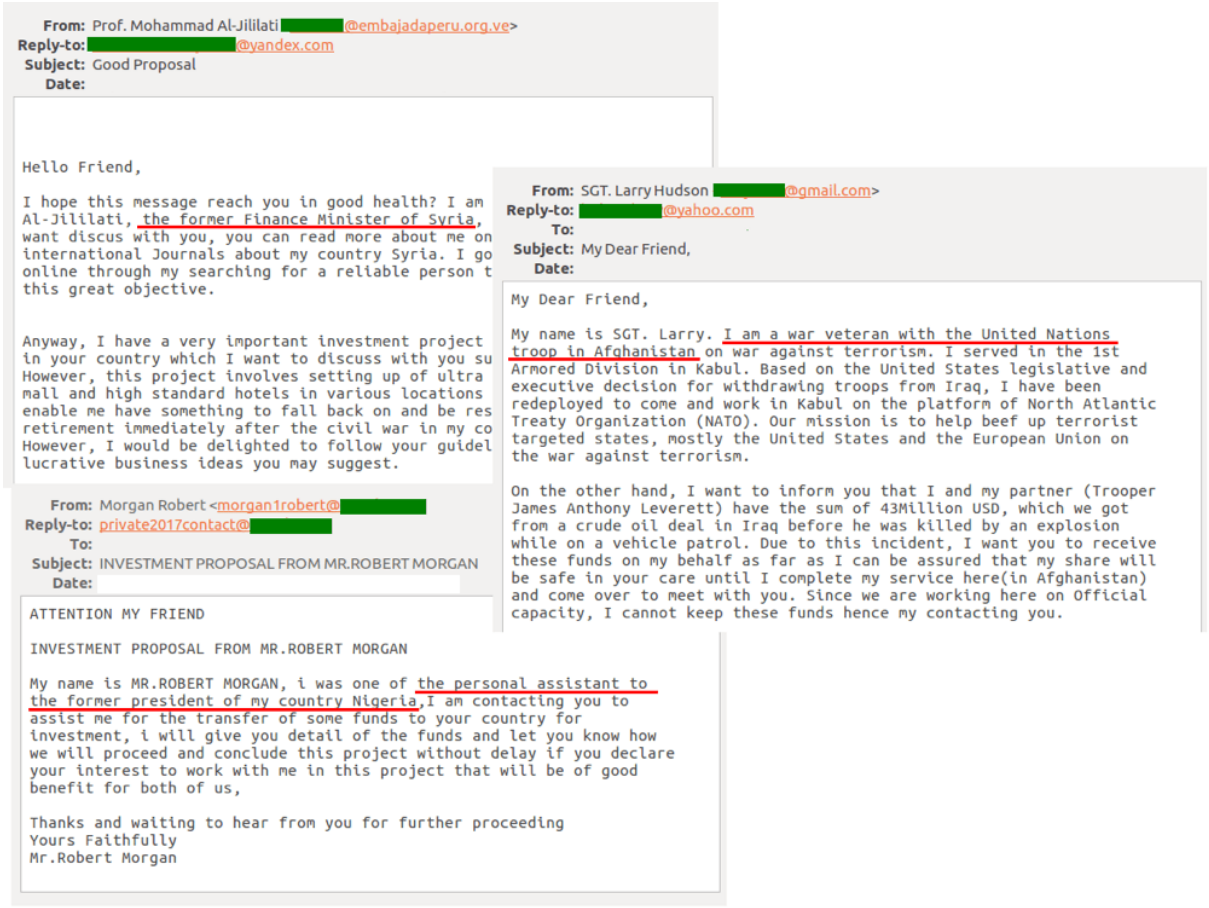
Figura 5 - Exemplo de notificação falsa de correio eletrônico



Fonte: Kaspersky. Disponível em: https://www.kaspersky.com.br/blog/phishing-spam-hooks/11174. Acesso em: 4 jun. 2020

Segundo o Kaspersky, o golpe do “príncipe nigeriano” é uma das modalidades de spam mais antigas, o golpe consiste em um *e-mail* vindo de um suposto parente próximo ou advogado representante de um milionário falecido, com a promessa de dinheiro em retorno caso a vítima envie dinheiro antecipado para um suposto “custo burocrático”, ou até mesmo, pedindo dados da vítima como documentos, dados bancários, etc (CERT.BR, 2020). Na Figura 6, é mostrado três exemplos do golpe do “príncipe nigeriano”, um se passando por um ministro da síria, o segundo se passando por um veterano do exército dos estados unidos e, por fim, um se passando por um assistente do presidente da Nigéria.

Figura 6 - Exemplo do golpe do príncipe nigeriano

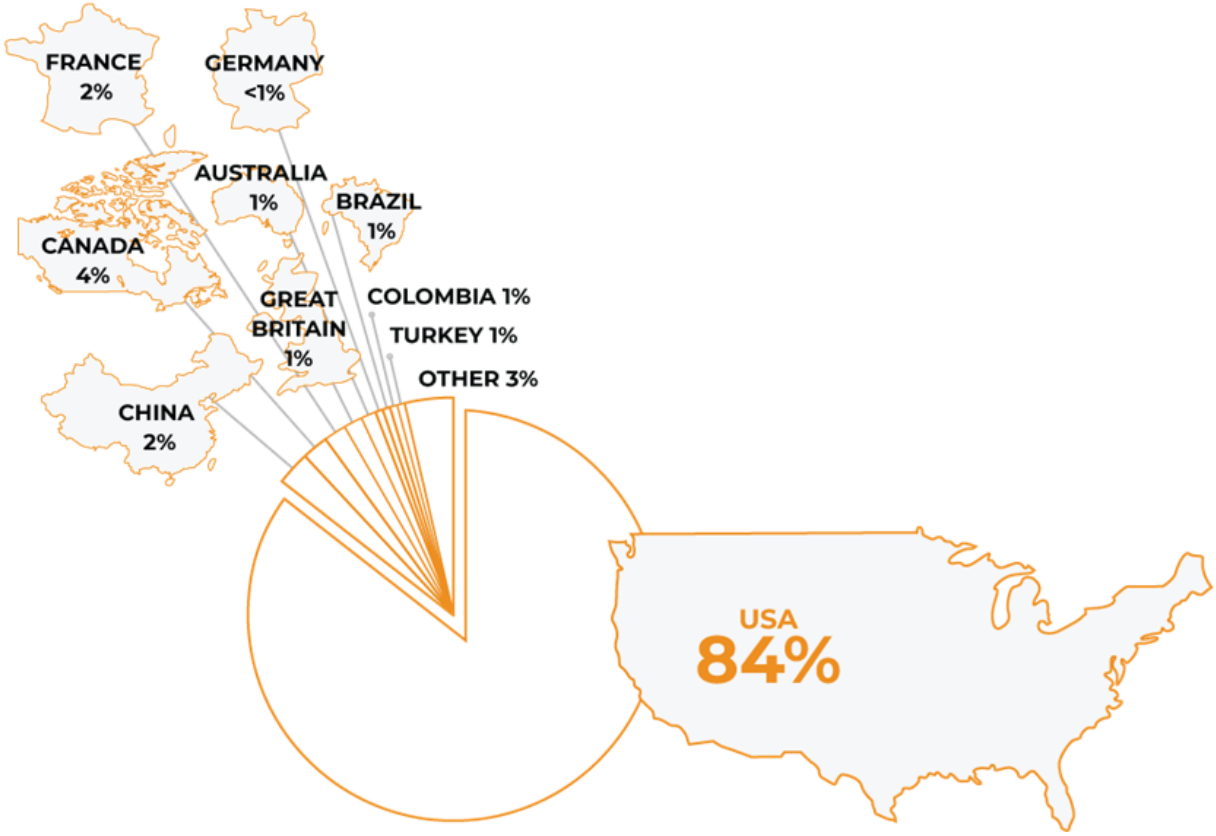


Fonte: Kaspersky. Disponível em: https://www.kaspersky.com.br/blog/phishing-spam-hooks/11174. Acesso em: 4 jun. 2020

2.3 Países mais afetados

Um estudo feito pelo site Phishlabs aponta que “os Estados Unidos foi, de novo, e previsivelmente será o principal país no quesito de alvo de ataques *phishing*, totalizando 85% dos ataques comparados com o resto do mundo” (PHISHLABS, 2019, tradução nossa). O segundo país mais afetado é o Canadá com 4% do total de ataques realizados no mundo, seguido por França com 2%, demonstrado na Figura 7.

Figura 7 - Gráfico demonstrando os países mais afetados por ataques *phishing*, os estados unidos, seguem em primeiro lugar (Foto: Phishlabs)



Fonte: Phishlabs. Disponível em https://info.phishlabs.com/blog/top-targeted-countries-by-phishing-attacks-2019. Acesso em: 5 jun. 2020

2.4 Impacto

É notável que o *phishing* tem um impacto muito negativo na vida da vítima, desde dados roubados ou até prejuízos monetários, segundo o site elpescador, um comunicado do FBI consta que:

> Esses golpes, segundo o comunicado, já teriam custado cerca de US$ 2,3 bilhões a empresas nos últimos três anos. Entre outubro de 2013 e fevereiro de 2016 autoridades dos EUA registraram mais de 17 mil queixas de phishing corporativo; desde janeiro de 2015 houve um aumento de 270% nos casos. (elpescador, 2016)

Segundo a ONG de Segurança da Informação Safernet, estima que fraudes bancárias através da Internet causem prejuízo anual de aproximadamente R$ 1 bilhão, apenas no Brasil.

No primeiro trimestre de 2019, o pico do número de spams ocorreu em março (56,3%). A participação média do spam no tráfego de *e-mail* mundial foi de 56%, porcentagem 4% superior do que no primeiro trimestre de 2018 (Security Report, 2019). ou seja, mais da metade da porcentagem das mensagens de correios eletrônicos são causadas por spam, isso, segundo o CERT.BR, pode ocasionar diversos problemas.

Por ter que se preocupar se a mensagem recebida por *e-mail* é ou não um golpe, a vítima acaba perdendo, além do tempo, a possibilidade excluir uma mensagem que não é spam, ocasionando assim, um transtorno para a vítima, ocasionando perda de mensagens importantes.

Grande parte dos spams são enviados para conjuntos aleatórios de pessoas, grande parte dessas mensagens podem ser conteúdo considerado ofensivo ou impróprio (CERT.BR).

Existe um impacto bastante no quesito de não recebimento de *e-mails*, como é explicado a seguir pelo portal de notícias Terra:

> Boa parte dos provedores de internet limita o tamanho da caixa postal do usuário no seu servidor. Caso o número de spams recebidos seja muito grande o usuário corre o risco de ter sua caixa postal lotada com mensagens não solicitadas. Se isto ocorrer, o usuário não conseguirá mais receber *e-mails* e, até que possa liberar espaço em sua caixa postal, todas as mensagens recebidas serão devolvidas ao remetente (TERRA).

Para combater os *phishings* por *e-mail*, diversas empresas possuem filtros de spam, caso esses filtros estejam mal configurados, os funcionários correm o risco de ter mensagens legítimas filtradas como spam, causando assim, um transtorno indesejado na empresa, como dito pelo CERT.BR.

Além de todos os impactos previamente citados, ainda existe, por fim, o prejuízo financeiro caso uma pessoa, sendo ela física ou jurídica, terá caso ela venha a cair em um golpe, dependendo dos dados que o golpista tenha acesso, os transtornos financeiros poderão ser grandes.

2.5 Engenharia social

O termo engenharia social ficou mais conhecido em 1990, através de um famoso hacker chamado Kevin Mitnick. Esse termo designa para práticas utilizadas a fim de se

obter informações sigilosas ou importantes de empresas, pessoas e sistemas de informação, explorando a confiança das pessoas para enganá-las. Pode-se também definir engenharia social como a arte de manipular pessoas a fim de contornar dispositivos de segurança ou construir métodos e estratégias para ludibriar pessoas, utilizando informações cedidas por elas de maneira a ganhar a confiança delas para obter informações. (SILVA, E., 2008).

Ou seja, além de todos os utensílios dos hackers utilizados para realizar um roubo virtual, também é utilizado a engenharia social, que é a prática de enganar a pessoa das mais variadas formas para que ela te de a informação, não sendo necessária obter ela através da invasão de sistemas.

O engenheiro social ataca o elo mais fraco da segurança, que é o próprio ser humano, entendendo os comportamentos da vítima, sabendo sua função em uma empresa, o que ela faz, qual sua rotina, etc. justamente para que no final, consiga ter o poder de convencimento para que ela dê as informações para o criminoso.

3. APRENDIZADO DE MÁQUINA UTILIZANDO O ALGORITMO DE NAIVE BAYES

Segundo MITCHELL (1997), o estudo do aprendizado de máquina (AM) faz parte de um campo da Inteligência Artificial, onde o computador se torna capaz de “aprender” sobre uma tarefa especifica, obtendo uma quantidade enorme de dados para realizar o treinamento do modelo.

Esses treinamentos se referem ao processo de transformação dos dados obtidos para formatos que o computador consiga interpretar, para isso é utilizado técnicas e algoritmos que através de cálculos matemáticos convertem textos e imagens em números, que são as unidades que os computadores atualmente conseguem interpretar.

Existem duas técnicas conhecidas e muito utilizadas atualmente, uma delas é a técnica de aprendizado supervisionado, onde os dados coletados que serão interpretados pelo computador já estão previamente trabalhados e com o seu respectivo rótulo, ou seja, já foi feito um agrupamento do que cada conjunto de dados representa.

A técnica não supervisionada, torna o computador responsável por fazer essa separação e rotulação dos dados da melhor forma possível, esse processo é chamado de *clusterização* dos dados e possui outros tipos de algoritmos que podem ser utilizados para fazer esse agrupamento, utilizando outras fórmulas matemáticas para isso.

3.1. Naive Bayes

Sabendo das técnicas de aprendizado, este trabalho faz o uso da técnica supervisionada, onde cada informação coletada foi rotulada previamente e para que o computador fosse capaz de aplicar e entender esses dados em um modelo estatístico e probabilístico, utilizando o algoritmo Naive Bayes.

Esta técnica foi baseada no teorema de Bayes, criada pelo matemático Thomas Bayes, ele é muito utilizado quando existe uma grande quantidade de dados e por sua simplicidade de aplicação em sistemas complexos obtendo resultados relativamente bons.

A equação está representada na Figura 8, onde existe os seguintes elementos.

Figura 8 - Fórmula de Naive Bayes

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

Fonte: O autor (2020)

- P(B|A): probabilidade do evento B acontecer uma vez que o evento A ocorreu.
- P(A): probabilidade do evento A ocorrer.
- P(B): probabilidade do evento B ocorrer.

3.2 Funcionamento do Naive Bayes

Primeiramente, deve ser classificado o modelo de dados, no caso, será demonstrado como na tabela 1.

Tabela 1 - Modelo de dados utilizados

Palavra	Classe
Cachorro	Positivo
Amor	Negativo
Ruim	Negativo
Amor	Positivo
Amor	Positivo
Cachorro	Positivo
Casa	Positivo
Gato	Positivo
Gato	Negativo
Gato	Negativo
Amor	Negativo
Casa	Positivo
Amor	Positivo

Fonte: O autor (2020)

Logo após a classificação das palavras e suas respectivas características, será criada uma outra tabela onde será analisado a frequência de cada palavra com sua classe, como mostra a tabela 2.

Tabela 2 - Frequência de cada palavra.

Palavra	Positiva	Negativa
Cachorro	2	
Amor	3	2
Ruim	1	
Gato	1	3
Casa	2	
Total	7	6

Fonte: O autor (2020)

A partir desses dados, será criado uma tabela com a respectiva probabilidade de cada palavra. Como demonstrado na tabela 3.

Tabela 3 - Tabela com o cálculo da probabilidade de cada palavra

Palavra	Positiva	Negativa	Cálculo
Cachorro	2		2/13 = 0.15
Amor	3	2	5/13 = 0.38
Ruim	1		1/13 = 0.076
Gato	1	3	4/13 = 0.30
Casa	2		2/13 = 0.15
Total	7	6	(7/13)+(6/13)=0.46

Fonte: O autor (2020)

Nesse exemplo, será verificado a probabilidade da palavra “Amor” ser positiva ou negativa.

1. P(positivo|'Amor') = P('Amor'|positivo) * P(positivo) / P('Amor')
2. P(negativo|'Amor') = P('Amor'|negativo) * P(negativo) / P('Amor')

Para a primeira equação, será calculado:

- P('Amor'|positivo) = 3/7 = **0.42**.

- P(positivo) = 7/13 = **0.53**.
- P('Amor') = 5/13 = **0.38**.

Para a segunda equação, será calculado:

- P('Amor'|negativo) = 2/6 = **0.33**.
- P(negativo) = 6/13 = **0.46**.
- P('Amor') = 5/13 = **0.38**.

Por fim, o cálculo final é obtido através da resolução da equação abaixo:

1. P(positivo|'Amor') = 0.42 * 0.53 / 0.38 = **0.58**.

2. P(negativo|'Amor') = 0.33 * 0.46 / 0.38 = **0.39**.

Ou seja, no final é provado que a probabilidade da palavra “Amor” ser positiva é maior que a probabilidade de ser negativa, já que o resultado dos cálculos para positivo foi maior do que os cálculos para negativo.

Esse exemplo foi realizado somente com base em uma palavra, no caso de um *phishing,* existe um conjunto de palavras, mas o cálculo é o mesmo, já que o algoritmo implementado pelo *framework* vai fazer a somatória de palavras suspeitas de serem *phishing* e calcular a porcentagem final do texto.

4. METODOLOGIA

Para que o classificador seja capaz de identificar padrões em textos parecidos com *phishing*, será preciso treinar esse modelo antes.

O treinamento do modelo foi feito utilizando uma base de fraudes chamada "catálogo de fraudes", ele foi lançado em 2008 e desde então vem coletando e salvando em sua base de dados diversas fraudes.

O catálogo de fraudes é mantido pela RNP[9], é um portal no qual é armazenado diversos tipos diferentes de fraudes, no qual a empresa que o mantem está sempre tentando manter atualizado através de denuncias que as próprias pessoas podem fazer através do *e-mail* phishing@cais.rnp.br.

O site atualmente conta com mais de doze mil fraudes cadastradas disponíveis para consulta, faz aproximadamente a triagem de quinze mil mensagens por mês, além de cadastrar cem novos exemplos de fraudes por mês (RNP, 2020).

Para coletar os dados do catálogo de fraudes, foi necessário desenvolver um script de "*scraping*", que foi usado para fazer requisições para o website e obter os detalhes de cada fraude encontrada.

4.1 Obtendo os dados

Foi desenvolvido um script em *JavaScript* para o *framework NodeJS* que foi responsável por fazer a coleta das fraudes do catálogo de fraudes através de requisições *HTTP*[10].

Cada página presente no catálogo de fraudes possui um indicador na URL, como na Figura 9:

[9] Rede nacional de ensino e pesquisa
[10] Hypertext Transfer Protocol

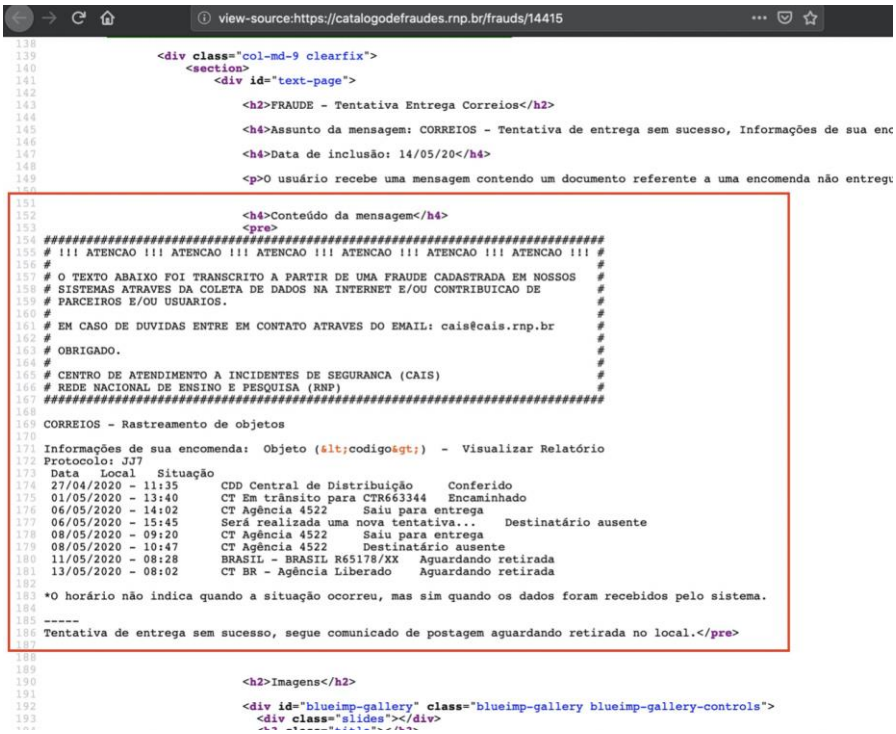
Figura 9 - URL do catálogo de fraudes

Fonte: O autor (2020)

O *script* recebe um limite de páginas que ele deve percorrer e então ele inicia sua execução obtendo o conteúdo da página HTML[11] de cada fraude presente no catálogo, uma vez obtido os dados, é preciso converter as informações para uma estrutura JSON[12].

O conteúdo da fraude que foi obtido pelo *script*, como mostrado na Figura 10, também vem em um formato que não é o ideal para o propósito do projeto. O conteúdo que vem junto ao corpo da mensagem, inclui *tags* HTML que não são necessárias, por isso será necessário realizar uma limpeza nesses dados antes de trabalhar com eles.

Figura 10 - Exemplo de conteúdo de uma fraude



Fonte: O autor (2020)

[11] HyperText Markup Language
[12] *JavaScript Object Notation*

Para resolver esse problema, existem funções do *JavaScript* que fazem "quebras" do texto em partes que podem decidir utilizar ou não.

Figura 11 - Código da função *main* que obtém os dados e também faz uma limpeza nos mesmos

```
const fs = require('fs');
const request = require("request");

async function main() {
    let i = process.env.START;
    let limite = process.env.LIMITE;
    var res = [];
    var options = {
        method: 'GET',
        url: '',
        headers:
        {
            'Postman-Token': 'be84052a-092a-41fb-b676-4dd3812f0d5f',
            'cache-control': 'no-cache',
            'Upgrade-Insecure-Requests': '1',
            Cookie: '__cfduid=d3d40e7d2d07730249ba83823037f01e21586815948',
            Connection: 'keep-alive',
            'Accept-Language': 'pt-BR,pt;q=0.8,en-US;q=0.5,en;q=0.3',
            Accept: 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
        }
    };

    while (i < limite) {
        console.log('while', i)
        options.url = `https://catalogodefraudes.rnp.br/frauds/${i}`
        let response = await doRequest(options)

        var title = ""
        var description = ""

        if (response.includes('<h4>Assunto da mensagem:')) {
            title = response.split('<h4>Assunto da mensagem:')
            title = title[1].split('</h4>')[0].trim()
        }
        if (response.includes('################################################################################')) {
            description = response.split('################################################################################\n\n')
            description = description[1].split('</pre>')[0]
        }

        if (title != "") {
            var res = await read(`./fraudes${limite}.json`)
            res = JSON.parse(res)
            res.push({
                title: title,
                description: description
            })
            await save(`./fraudes${limite}.json`, JSON.stringify(res))
        }
        i++;
    }


    // fs.writeFileSync('./fraudes.json', JSON.stringify(res))
    process.exit(0)
}
```

Fonte: O autor (2020)

O código demonstrado na Figura 11 não funcionaria sem as funções demonstradas na Figura 12, que são responsáveis por fazer o a requisição HTTP para obter a fraude, a função de ler o arquivo do disco e também a função de salvar os dados no arquivo.

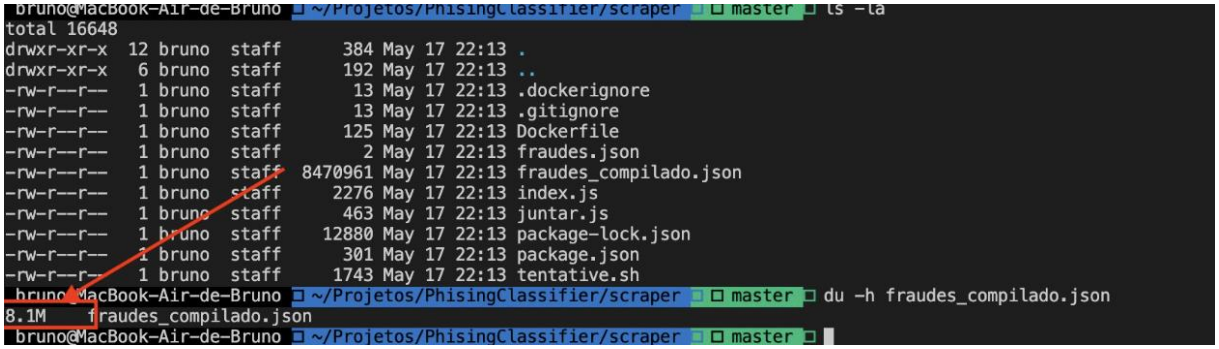
Figura 12 - Funções utilizadas para interagir com o disco e fazer a requisição

```
function doRequest(options) {
    return new Promise((resolve, reject) => {
        request(options, function (error, response, body) {
            if (error) throw new Error(error);
            // console.log(body);
            resolve(body)
        });
    })
}

function save(filename, contents) {
    return new Promise((resolve, reject) => {
        fs.writeFileSync(filename, contents)
        resolve(true)
    })
}

function read(filename) {
    return new Promise((resolve, reject) => {
        if (fs.existsSync(filename)) {
            resolve(fs.readFileSync(filename))
        } else {
            fs.writeFileSync(filename, '[]')
            resolve(fs.readFileSync(filename))
        }

    })
}
```

Fonte: O autor (2020)

Após o *script* ser executado, será gerado um arquivo “fraudes_compilado.json” com pouco mais de quatorze mil fraudes em um formato facilmente interpretado. A Figura 13 demonstra o tamanho do arquivo gerado pelo *script*, no caso, 8.1MB de memória.

Figura 13 - Compilado de mais de 14 mil fraudes



Fonte: O autor (2020)

4.2 Treinando o modelo

Uma vez com os dados compilados em formato JSON, um novo script será executado para fazer a leitura e treinar o modelo baseado no algoritmo de Naive Bayes. O *script* pode ser visto na Figura 14.

Figura 14 - Script utilizado para treinar o modelo

```
const natural = require('natural')
const classifier = new natural.BayesClassifier()
const fs = require('fs')

function main() {

    let i = 0
    let contents = JSON.parse(fs.readFileSync('./fraudes_compilado.json'))

    while (i < contents.length) {
        classifier.addDocument(contents[i].description, 'fraude')
        i++
    }

    classifier.train()

    classifier.save('./fraudes_model.json', () => {
        process.exit(0)
    });


}

main()
```

Fonte: O autor (2020)

É possível observar também que esse código exporta o nosso modelo em um arquivo JSON, é preciso salvar esse modelo treinado em disco para ter uma inicialização mais rápida sem ter a necessidade de ficar treinando o modelo toda vez que o *script* for executado.

4.3 Classificando texto

Tendo o modelo treinado e salvo dentro de um arquivo, agora é possível fazer o carregamento dele na memória e então classificar os textos para obter a probabilidade de ser um *phishing* ou não.

Nesse projeto, é necessário que o texto que será analisado pelo script esteja presente dentro de um arquivo chamado "fraude.txt".

A Figura 15 mostra o script responsável pela leitura do arquivo de modelo para o algoritmo e classificação da mensagem.

Figura 15 - Script responsável pela leitura e classificação do texto

```
const natural = require('natural')
const classifier = new natural.BayesClassifier()
const fs = require('fs')

natural.BayesClassifier.load('./fraudes_model.json', null, (err, classifier) => {
    let data = fs.readFileSync('./fraude.txt').toString()
    data = data.trim()
    let porcentage = parseFloat(classifier.getClassifications(data)[0].value.toString().substring(0,
3))
    porcentage = porcentage * 10
    porcentage = porcentage + "%"
    console.table([{ description: data.substring(0, 80) + "...", porcentage: porcentage }])
})
```

Fonte: O autor (2020)

4.4 Resultados

A ideia do trabalho é demonstrar a aplicação de aprendizado de máquina em um campo da segurança digital, que é o *phishing.* Utilizando como base de dados o catálogo de fraudes, foi obtido acesso a uma fonte confiável de dados e a uma quantidade relativamente alta para o exemplo pratico que o trabalho se propõe a mostrar.

Para validar se o *script* é realmente capaz de identificar padrões de *phishing*, foi realizado uma coleta de dados de mensagens de texto de cada participante do trabalho, e depois o classificador foi executado para obter a porcentagem de ser uma fraude em cada mensagem, como mostra na Tabela 4.

Tabela 4 - Resultados Obtidos

Texto	**Porcentagem**
"BANCO DO BRASIL: VC TEM 251.310 PTS QUE VENCEM DIA 22-04. CLIQUE P/ #RESGATARLIVELO: u.to/jQgkGA"	66%
"LEONICE,temos uma oportunidade ideal para você reorganizar sua pendencia com o Itau,ligue 08009403440 http://bit.ly/395BHIS"	25%
"A sua fatura NET acaba de chegar e para visualizar todos os detalhes basta abrir o documento anexado neste e-mail."	80%
"Identificamos uma cobrança indevida para email por parte da ¨Grupo Dimed | Panvel Farmáciasëm uma de nossas auditorias contratadas, porém não tivemos sucesso em reembolsar para o seu cartão de débito/crédito."	57%
"Você está recebendo anexa a sua conta Vivo Móvel com vencimento em 27/04/2020."	14%

Fonte: O autor (2020)

Fazendo uma relação da porcentagem dos resultados dividida pela quantidade de textos, foi possível calcular uma média de 48.4%, que significa que o modelo dentre essas 5 amostras aleatórias teve uma eficiência baixa.

Nesse estudo de caso realizado no projeto, o modelo foi treinado com base em 13268 fraudes coletadas do catalogo de fraudes, para que a eficiência do algoritmo seja maior e mais precisa, o número de dados recomendado para classificadores como este, é em torno de cem mil exemplos de *phishing* para que seja possível aumentar a qualidade da classificação dos textos.

Este foi um estudo de caso simples e apenas com objetivo de demonstrar a utilização do algoritmo, ataques de *phishing* possuem níveis de complexidade que vão desde *e-mails* generalizados para atingir qualquer pessoa, até ataques bem mais elaborados com o objetivo de obter a informação da vítima com textos e imagens voltados para uma pessoa específica. Portanto o classificador mostrado nesse trabalho consegue realizar verificações simples nos textos, o que faz com que ele não obtenha um resultado muito eficaz na classificação de ataques *phishing* específicos.

É recomendado que a quantidade de dados coletados seja ainda maior, e também não leve somente em consideração o corpo da mensagem, mas também o remetente do *e-mail*, título, imagens entre outras características que identificam uma mensagem maliciosa ou não.

5. CONCLUSÃO

Este trabalho apresentou os principais conceitos a respeito de *phishing* e golpes na internet no geral com o objetivo de fazer uma introdução ao algoritmo de Naive Bayes e como ele pode ser utilizado para realizar um classificador de golpes na internet.

Foi efetuada a análise utilizando uma biblioteca escrita na linguagem *Javascript* que implementa os conceitos do algoritmo de Naive Bayes, utilizando essa biblioteca, foi realizada uma coleta de dados utilizando o portal de catálogos de fraudes, para que esses dados fossem utilizados para treinar o algoritmo para realizar a detecção de fraudes.

Utilizando os dados coletados para treinar o algoritmo, foi possível então enviar textos para o algoritmo para que o mesmo devolvesse a porcentagem da probabilidade de a mensagem ser verdadeira, ou ser realmente um golpe.

Sugere-se para trabalhos futuros, encontrar melhores maneiras de treinar o algoritmo de Naive Bayes para o mesmo realizar uma detecção de mensagens *phishing* mais eficiente e confiável, como, por exemplo, uma base maior para treinar o algoritmo com mais dados e separar em diversos algoritmos, um para cada tipo de fraude, como por exemplo, *e-mail*, SMS, etc. Sugere-se também, disponibilizar uma API para verificação de diversos tipos de mensagem e apontar quais são golpes e quais são reais, se possível também, uma interface web ou mobile que, faça acesso a essa API de maneira rápida e fácil.

6. REFERÊNCIAS

BELCIC, Ivan. O guia essencial sobre phishing: Como funciona e como se proteger. Avast, 5 de fevereiro de 2020. Disponível em: https://www.avast.com/pt-br/c-phishing. Acesso em: 2 jun. 2020.

Brasil é o País com mais usuários atacados por mensagens de phishing. Security Report. Disponível em: https://www.securityreport.com.br/overview/brasil-e-o-pais-com-mais-usuarios-atacados-por-mensagens-de-phishing. Acesso em: 6 jun. 2020.

CATÁLOGO DE FRAUDES, Disponível em: https://catalogodefraudes.rnp.br. Acesso em: 7 jun. 2020

CATÁLOGO DE FRAUDES, Disponível em: https://www.rnp.br/sistema-rnp/cais/catalogo-de-fraudes. Acesso em: 10 jun. 2020

Golpes na Internet. Cert.br. Seção: Fraude de antecipação de recursos (Advance fee fraud). Disponível em: https://cartilha.cert.br/golpes. Acesso em: 4 jun. 2020.

History of Phishing. phishing.org. Seção: Phishing Attacks Begin. Disponível em: https://www.phishing.org/history-of-phishing. Acesso em: 4 jun. 2020.

History of Phishing. phishing.org. Seção: Phishing's America Online Origins. Disponível em: https://www.phishing.org/history-of-phishing. Acesso em: 4 jun. 2020.

O que é Phishing?. Canaltech. Disponível em: https://canaltech.com.br/hacker/O-que-e-Phishing-Scam. Acesso em: 4 jun. 2020.

O que é Phishing?. Canaltech. Disponível em: https://canaltech.com.br/seguranca/O-que-e-Phishing. Acesso em: 4 jun. 2020.

Quais são os problemas que o spam pode causar para um usuário da internet?. Terra. Disponível em: https://duvidas.terra.com.br/duvidas/549/quais-sao-os-problemas-que-o-spam-pode-causar-para-um-usuario-da-internet. Acesso em: 6 jun. 2020.

RAY, Sunil. 6 passos fáceis para aprender o algoritmo Naive Bayes (com o código em Python). Disponível em: https://www.vooo.pro/insights/6-passos-faceis-para-aprender-o-algoritmo-naive-bayes-com-o-codigo-em-python/. Acesso em: 7 jun. 2020

RODRIGUES, Renato. Brasil é o 4º país mais atacado por malware financeiro em 2019. Kaspersky Daily, 16 de abril de 2020. Disponível em: https://www.kaspersky.com.br/blog/brasil-atacado-malware-financeiro-2019-pesquisa/14894. Acesso em: 4 jun. 2020.

SALES, Robson. Fraudes virtuais em sites bancários dão prejuízo de R$ 1 bilhão. Tech Tudo, 19 de janeiro de 2012. Disponível em: https://www.techtudo.com.br/noticias/noticia/2012/01/fraudes-virtuais-em-sites-bancarios-dao-prejuizo-de-r-1-bilhao.html. Acesso em: 20 jun. 2020.

SANTANA, Felipe. Café com Código 14#: Como Funciona o Algoritmo Naive Bayes. Disponível em: https://minerandodados.com.br/naive-bayes-machine-learning. Acesso em: 7 jun. 2020

Significado de Phishing. Significados. Disponível em: https://www.significados.com.br/phishing. Acesso em: 7 jun. 2020

SILVA, Elaine M. da. Cuidado com a engenharia social: Saiba dos cuidados necessários para não cair nas armadilhas dos engenheiros sociais. [S.l.:s.n.], 2008. Disponível em: http://www.baixaki.com.br/info/1078-cuidado-coma-engenharia-social.htm. Acesso em: 6 jun. 2020

VERGELIS, Maria. Fraude online: conheça os 5 golpes mais comuns. Kaspersky Daily, 16 de abril de 2020. Disponível em: https://www.kaspersky.com.br/blog/phishing-spam-hooks/11174. Acesso em: 4 jun. 2020.

VOLKMAN, Elliot. These Are the Top Most Targeted Countries by Phishing Attacks. Phishlabs, 23 mai. 2019. Disponível em: https://info.phishlabs.com/blog/top-targeted-countries-by-phishing-attacks-2019. Acesso em: 5 jun. 2020.

WAKKA, Wagner. Número de ataques cibernéticos no Brasil quase que dobrou em 2018. Canaltech, 7 de agosto de 2018. Disponível em: https://canaltech.com.br/seguranca/numero-de-ataques-ciberneticos-no-brasil-quase-que-dobrou-em-2018-119600. Acesso em: 2 jun. 2020.