FACULDADE DE TECNOLOGIA DE AMERICANA "Ministro Ralph Biasi"
Curso Superior de Tecnologia em Segurança da Informação

Arthur Germano

Victor Ferreira Franco

EXPERIMENTAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA CONTENÇÃO DE PHISHING.

Americana, SP

2024

FACULDADE DE TECNOLOGIA DE AMERICANA "Ministro Ralph Biasi"
Curso Superior de Tecnologia em Segurança da Informação

Arthur Germano

Victor Ferreira Franco

EXPERIMENTAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA CONTENÇÃO DE PHISHING.

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Segurança da Informação sob a orientação do Prof.(a) Dr. Henri Alves de Godoy

Área de concentração: Estudos Avançados em Segurança da Informação.

Americana, SP

2024

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

FRANCO, Victor Ferreira

Experimentação do uso de inteligência artificial generativa para contenção de phishing. / Victor Ferreira Franco, Arthur Germano – Americana, 2024.

50f.

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

Orientador: Prof. Dr. Henri Alves de Godoy

1. Análise de dados 2. Inteligência artificial 3. Segurança em sistemas de informação. I. FRANCO, Victor Ferreira, II. GERMANO, Arthur III. GODOY, Henri Alves de IV. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana Ministro Ralph Biasi

CDU: 681516
007.52
681.518.5

Elaborada pelo autor por meio de sistema automático gerador de ficha catalográfica da Fatec de Americana Ministro Ralph Biasi.

Arthur Germano
Victor Ferreira Franco

EXPERIMENTAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA CONTENÇÃO DE PHISHING.

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Tecnólogo em Segurança da Informação pelo Centro Paula Souza – Faculdade de Tecnologia de Americana – Ministro Ralph Biasi.
Área de concentração: Estudos Avançados em Segurança da Informação.

Americana, 03 de dezembro de 2024.

Banca Examinadora:

Henri Alves de Godoy
Doutor
Faculdade de Tecnologia de Americana

Marcus Vinicius Lahr Giraldi
Especialista
Faculdade de Tecnologia de Americana

Benedito Aparecido Cruz
Mestre
Faculdade de Tecnologia de Americana

Dedicamos este trabalho a todos que ajudaram a construir nossa formação acadêmica. Aos nossos pais, pelo seu amor incondicional, apoio e encorajamento constante. Aos nossos professores, pela paciência e dedicação, que foram fundamentais na nossa jornada de aprendizado. E, finalmente, a nós mesmos, pelo esforço e determinação para superar todos os obstáculos que encontramos ao longo deste caminho.

Nós, os autores, gostaríamos de expressar nossa profunda gratidão a todos que contribuíram para a realização deste trabalho.
Agradecemos ao nosso orientador pela paciência, dedicação e conhecimento compartilhado. Seu apoio e orientação foram inestimáveis.
Por fim, agradecemos às nossas famílias, pelo apoio durante o percurso de escrita e pesquisa. Sua fé em nós foi uma fonte constante de força.

RESUMO

Este trabalho tem como objetivo identificar a capacidade de detecção de ciberataques, como o phishing, por meio da inteligência artificial e avaliar sua eficácia na prática. Pretende-se promover um conhecimento aprofundado sobre o tema e desenvolver de maneira sólida um estudo dos benefícios associados à inteligência artificial para a mitigação de um dos principais tipos de ataques cibernéticos. A construção desta monografia se baseia por meio de uma pesquisa exploratória dos temas para fundamentação teórica e na realização de experimentos com ferramentas atuais, como o ChatGPT, Copilot e o Gemini, em diversos cenários, com casos reais. Através deste estudo, busca-se contribuir para o campo da segurança cibernética, fornecendo percepções valiosas sobre a aplicação da inteligência artificial na prevenção e combate a ciberataques. Os resultados alcançados indicam que, embora a detecção de e-mails de phishing tenha atingido taxas satisfatórias, a identificação de e-mails legítimos ainda apresenta desafios, com taxas medianas. Isso demonstra a necessidade de aprimoramentos, especialmente no reconhecimento de falsos positivos, sugerindo que técnicas mais avançadas, como o aprendizado de máquina, podem aumentar a precisão das ferramentas no futuro.

Palavras-chave: Ciberataques; Mitigação; Segurança cibernética, Engenharia Social.

ABSTRACT

This work aims to identify the detection capability of cyberattacks, such as phishing, through artificial intelligence and evaluate its practical effectiveness. The goal is to promote an in-depth understanding of the subject and develop a solid study of the benefits associated with artificial intelligence for mitigating one of the main types of cyberattacks. The construction of this thesis is based on exploratory research of the topics for theoretical foundation and the conduct of experiments with current tools, such as ChatGPT, Copilot and Gemini, in various scenarios with real cases. Through this study, we seek to contribute to the field of cybersecurity by providing valuable insights into the application of artificial intelligence in the prevention and combat of cyberattacks. The results indicate that, although the detection of phishing emails has reached satisfactory rates, the identification of legitimate emails still presents challenges, with median rates. This demonstrates the need for improvements, especially in the recognition of false positives, suggesting that more advanced techniques, such as machine learning, may increase the accuracy of the tools in the future.

Keywords: Cyberattacks; Mitigation; Cybersecurity, Social Engineering.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

CEOs: Chief Executive Officers.

CFOs: Chief Financial Officers.

COOs: Chief Operating Officers.

DDoS: Distributed Denial of Service.

IA: Inteligência Artificial.

LLMs: Large Language Models.

ML: Machine Learning.

PREP: Prompt Rule Explicit Parameters.

NRP: Rede Nacional de Ensino e Pesquisa.

SMS: Short Message Service.

SOC: Security Operations Center.

SQL: Structured Query Language.

SUMÁRIO

INTRODUÇÃO

Na década de 1990, a tecnologia avançou de maneira exponencial, transformando profundamente a forma como vivemos e interagimos. Essa transformação resultou em uma dependência crescente do mundo digital, especialmente com a chegada da Inteligência Artificial (IA). A IA tem automatizado processos manuais que seriam desafiadores para sistemas convencionais, incluindo a geração de textos, imagens, áudios e a realização de pesquisas rápidas.

Contudo, essa dependência expôs indivíduos, organizações e infraestruturas a uma gama de ameaças cibernéticas, com o crescente destaque de dados pessoais e informações confidenciais. Nesse cenário, a IA se destaca como uma poderosa ferramenta para combater o aumento de ataques. Entre esses ataques, o *phishing* é um dos mais conhecidos, caracterizado pela coleta de informações pessoais e confidenciais, onde o atacante se passa por uma instituição legítima, como bancos e empresas de comércio, utilizando canais de comunicação como e-mails, plataformas de vendas online e redes sociais.

A IA generativa, além de gerar textos, tem a capacidade de analisá-los em diferentes contextos e cenários, gerando relatórios abrangentes sobre todos os pontos observados, incluindo a veracidade de uma informação e possíveis erros ortográficos, que são comuns em tentativas de *phishing*.

Ademais, essa tecnologia desempenha um papel crucial no monitoramento contínuo de sistemas, mesmo na ausência de supervisão humana. Por exemplo, em uma rede de computadores gerenciada por IA, algoritmos avançados podem detectar e neutralizar ameaças em tempo real, protegendo assim os usuários e seus dados.

Essa ampliação da temática de defesas impulsionadas pela IA foi a motivadora para a realização deste trabalho, com grande relevância para os autores e contribuindo de maneira significativa para o campo da segurança da informação.

Este estudo proporciona uma perspectiva atualizada, demonstrando como essa poderosa ferramenta pode ser empregada para fins adversos. A IA, além de ser uma aliada no progresso tecnológico e na otimização de processos, também pode ser utilizada para fortalecer a segurança da informação. Assim, entender os benefícios associados a essa tecnologia é fundamental, pois permite criar métodos eficazes de mitigação de riscos e fortalecer as defesas contra ameaças cibernéticas.

A disponibilidade de ferramentas gratuitas e sofisticadas como ChatGPT, Copilot e Gemini viabiliza a realização deste estudo. Devido aos recursos poderosos que essas plataformas oferecem, é possível realizar pesquisas abrangentes e complexas.

O ChatGPT e o Gemini, por exemplo, são capazes de gerar textos coerentes e contextuais, tornando-se ferramentas valiosas para simular cenários de *phishing* e entender como esses ataques podem ser estruturados. Da mesma forma, o Copilot pode ser utilizado para analisar e aprimorar códigos, o que pode ser útil para entender as técnicas de programação usadas em ataques cibernéticos.

O objetivo principal deste trabalho é realizar um experimento detalhado com IA, com o intuito de compreender sua capacidade de detecção de *phishing*. Os objetivos específicos desta monografia incluem a realização de experimentos com casos reais e a identificação da eficácia dessas ferramentas.

A hipótese principal é que seja possível neutralizar cerca de 85% dos ataques de *phishing* por meio de análises conduzidas por IA, identificando os aspectos minuciosos e evidenciando a natureza maliciosa do ataque.

O caminho metodológico deste trabalho envolve uma pesquisa exploratória, baseada em revisões bibliográficas sobre os temas Inteligência Artificial, Ciberataques, *Phishing* e Engenharia Social. Além disso, será realizada uma experimentação quantitativa com testes reais, onde as informações coletadas serão analisadas.

A monografia está organizada em quatro capítulos: o capítulo I aborda os fundamentos teóricos, o capítulo II trata da metodologia de pesquisa, o capítulo III descreve o esquema de experimentação, definindo como serão realizados os testes, e o capítulo IV apresenta os resultados, análise e discussão dos dados obtidos.

1 FUNDAMENTAÇÃO TEÓRICA

1.1 Inteligência Artificial

A IA pode ser entendida como uma tecnologia composta por uma ampla estrutura de hardwares e softwares, capaz de fazer escolhas autonomamente com base em dados fornecidos e no reconhecimento de padrões, com o objetivo de apoiar os seres humanos na tomada de decisões (Morais *et al.*, 2020). Sendo uma ferramenta de enorme potencial, a IA tem sido alvo de intensa pesquisa e desenvolvimento. Nos últimos anos, grandes empresas de tecnologia têm lançado suas próprias IAs, integrando-as aos seus sistemas e serviços e impulsionando seu desenvolvimento.

De maneira mais clara, a IA abrange dispositivos e sistemas que buscam imitar a capacidade cognitiva humana, podendo ser aplicados nas mais diversas tarefas. Sua capacidade mais notável é a evolução na assimilação e utilização de informações previamente processadas (Barbosa e Portes, 2023).

Para compreender como essas tecnologias aprendem, analisam e tomam decisões sozinhas ao lidar com dados fornecidos, é importante destacar dois conceitos essenciais. O primeiro é o *Machine Learning* (ML), ou aprendizado de máquina, que permite a uma máquina aprender por conta própria e fazer escolhas de forma independente, utilizando a análise de dados e a identificação de padrões. Em conjunto com o ML, há o *Deep Learning* (aprendizado profundo), uma subárea do *Machine Learning* que se destaca por sua capacidade ampliada de aprendizado, utilizando redes neurais complexas baseadas na estrutura do cérebro humano, o que permite maior profundidade e complexidade na análise de dados (Barbosa e Portes, 2023).

O impacto dessa inovação tem sido um marco na sociedade, impulsionando o interesse e a utilização dessa tecnologia para fins pessoais, indo além dos meios acadêmicos e de pesquisa (Machado *et al***.**, 2023). Uma das mais recentes e amplamente utilizadas extensões é a IA Generativa, cujo principal foco é simular a engenhosidade humana na criação de conteúdos diversos, como a produção de textos e até mesmo vídeos animados com áudio, com base nos cenários e informações fornecidas (Timpone e Guidi, 2023).

1.2 Ciberataques

Um ato malicioso realizado por meio de computadores ou sistemas de comunicação, com o fim de obter vantagem, se apropriando indevidamente de dados ou comprometendo a segurança de sistemas, é conhecido como um ciberataque. Esses ataques podem envolver roubo, manipulação e destruição de dados, muitas vezes sensíveis e críticos para sistemas e indivíduos. Além disso, os ciberataques podem ser direcionados a uma variedade de alvos, incluindo corporações, sistemas de informação, bancos de dados, órgãos governamentais, serviços online e até mesmo dispositivos pessoais. Vale ressaltar que os autores desses ataques sempre têm uma motivação por trás de cada ato, frequentemente se reunindo e se organizando de forma coordenada para atingir seus objetivos, podendo causar graves danos e comprometendo a segurança e a privacidade de seus alvos (Zeferino, 2020).

Tendo em vista que, após um ciberataque, a vítima pode sofrer diversos tipos de prejuízos significativos, como:

- Danos financeiros”, que envolvem a maior parte dos ataques, gerando altos custos associados à reparação de sistemas.
- " Equipamentos de infraestrutura de afetados”, gerando a paralisação de sistemas críticos.
- “Multas governamentais” relativas a leis de proteção de dados.

Além dos impactos puramente financeiros, pode haver uma perda na reputação de uma empresa, o que pode resultar na saída de clientes importantes, perda de vantagens competitivas e, posteriormente, em uma baixa demanda no mercado. Com isso, podem se desenvolver ainda mais lesões financeiras de longo prazo, já que será necessário um investimento considerável para recuperar a confiança do mercado (Oliveira, 2023).

Os ciberataques podem se manifestar de várias formas, permitindo que as vítimas sofram diferentes tipos de ataques. Um exemplo é o uso de *malwares* (software mal-intencionado), que se disfarçam como aplicativos e arquivos confiáveis para infectar uma rede de computadores, transmitindo códigos maliciosos como vírus, *worms* e *trojans*. Esses códigos enfraquecem e interrompem os processos de uma infraestrutura, consequentemente abrindo brechas para o acesso de outros hackers não autorizados à rede. Além desse tipo de ataque, outros métodos incluem ataques

DDoS, injeção de SQL e o uso de scripts maliciosos em sites, entre outros (Microsoft, 2024).

É importante também ter conhecimento sobre *ransomware* e *phishing* para melhor compreensão do trabalho. *Ransomware* é um tipo específico de *malware* que se concentra em bloquear dispositivos e sequestrar as informações de determinado sistema. O invasor, buscando benefício próprio, pode exigir, por meio de chantagem, um pagamento para a recuperação dos dados (Oliveira, 2022). Já o *phishing* consiste em "pescar" vítimas por meio do envio de e-mails enganosos, nos quais o atacante se passa por indivíduos ou empresas confiáveis (Microsoft, 2024). Além disso, outros tipos de ataque podem ser combinados com o *phishing*, como a utilização de *malware*, *ransomware* e outros meios dentro de um e-mail malicioso.

1.3 Phishing

Conforme Santos e Gonzaga (2024), a prática conhecida como "*phishing*" é uma analogia perfeita para a pesca. Da mesma forma que os pescadores lançam suas redes para capturar peixes, os criminosos virtuais lançam suas iscas na internet, com a esperança de atrair vítimas. A origem do termo "*phishing*" deriva da palavra em inglês "*fishing*", que significa pesca.

No *phishing*, os atacantes lançam iscas na forma de e-mails falsos, sites clonados, mensagens em redes sociais ou Serviço de Mensagens Curtas (SMS), com o objetivo de enganar as vítimas para que revelem informações pessoais ou confidenciais (Piovesan; Silva; Sousa; Turibus, 2019).

Os ataques de *phishing* também são uma forma de engenharia social, e seu sucesso depende principalmente da exploração de características humanas. Hackers mal-intencionados usam essas vulnerabilidades para criar intervenções eficazes (Souza; Tanaka, 2023). Isso torna o *phishing* uma ameaça significativa, já que se aproveita de falhas humanas, em vez de falhas tecnológicas de segurança.

Essa técnica utiliza textos apelativos para convencer suas vítimas. A mensagem, seja um alerta de segurança, uma oferta irresistível ou uma suposta atualização de senha, é estrategicamente elaborada para mexer com as emoções, induzindo uma ação imediata, sem que haja tempo para questionar a veracidade (Santos; Gonzaga, 2024).

A detecção desses ataques vem se tornando cada vez mais difícil, conforme apontado por Guedes e Moreira (2023). Os atacantes utilizam técnicas para esconder suas ações, tornando-as menos visíveis aos sistemas de segurança. Além disso, eles mudam frequentemente de hospedagem e domínio, dificultando o rastreamento. Em algumas situações, não há padrões claros, e as técnicas de *phishing* estão em constante evolução, o que dificulta ainda mais sua detecção.

Existem várias formas de *phishing*, classificadas de acordo com a maneira como são realizadas e o tipo de informação que buscam obter. São elas:

- ***Scam***: termo usado para descrever golpes que visam obter dados financeiros. Conforme explicado por Salviano, Santos e Silva (2022), esses golpes envolvem a indução da vítima a fornecer informações pessoais, como número de conta bancária, dados de cartão de crédito e senhas, por meio de links e arquivos mal-intencionados.

- ***Blind Phishing***: é uma das formas mais conhecidas e aplicadas. Caracteriza-se pelo envio massivo de e-mails com conteúdo aleatório, contando apenas com a sorte de que os receptores caiam nas armadilhas. Esse tipo de ataque é responsável por muitas vítimas (Salviano; Santos; Silva, 2022).

- ***Spear Phishing***: conforme a IBM, é um tipo de ataque que se concentra em um indivíduo ou grupo específico dentro de uma organização, visando enganar os alvos para que revelem informações privadas.

- ***Clone Phishing***: técnica que envolve a clonagem de sites. Nesse tipo de ataque, são criados sites falsos que imitam os originais, fazendo com que os usuários acreditem estar no site legítimo e, assim, forneçam suas informações confidenciais (Salviano; Santos; Silva, 2022).

- ***Whaling***: ataque direcionado a alvos corporativos de alto escalão, como CEOs, CFOs e COOs. Esses indivíduos, muitas vezes referidos como "baleias", possuem a capacidade de autorizar grandes transações financeiras ou divulgar informações sensíveis (IBM).

- ***Vishing***: técnica de *phishing* que utiliza serviços de telefonia. De acordo com a ESET (2023), essa modalidade explora chamadas telefônicas ou via internet para enganar as vítimas e obter informações pessoais.

- ***Smishing***: forma de *phishing* que utiliza mensagens de texto (SMS) falsas para enganar as vítimas. O termo "*smishing*" é uma combinação de "SMS" com "*phishing*" (IBM).

- **Engenharia Social**: uma estratégia que se baseia na manipulação da psicologia humana, ao invés de exploração de falhas tecnológicas, para obter acesso a sistemas ou dados (Salviano; Santos; Silva, 2022).

Os oito tipos identificados foram organizados em uma quadro, indicando os locais onde ocorrem e seus respectivos alvos de ataque, conforme demonstrado no Quadro 1.

Quadro 1 – Tipos de Phishing

Tipo de phishing	Onde ocorre	Principais focos
Scam	Links e arquivos maliciosos	Usuário final
Blind Phishing	Links e arquivos maliciosos com disparos feitos aleatoriamente	Usuário final
Spear Phishing	Links e arquivos maliciosos	Usuário final (funcionários públicos, clientes de empresas) Empresas e bancos
Clone Phishing	Sites clonados	Usuário final
Whaling	Links e arquivos maliciosos	Usuário de alto poder executivo
Vishing	Ligações telefônicas e SMS	Usuário final
Pharming	Link redirecionando para um site falso	Usuário final Empresas

Smishing	SMS	Usuário final
Engenharia Social	Ataques interpessoais	Usuário final

Fonte: Salviano; Santos; Silva (2022).

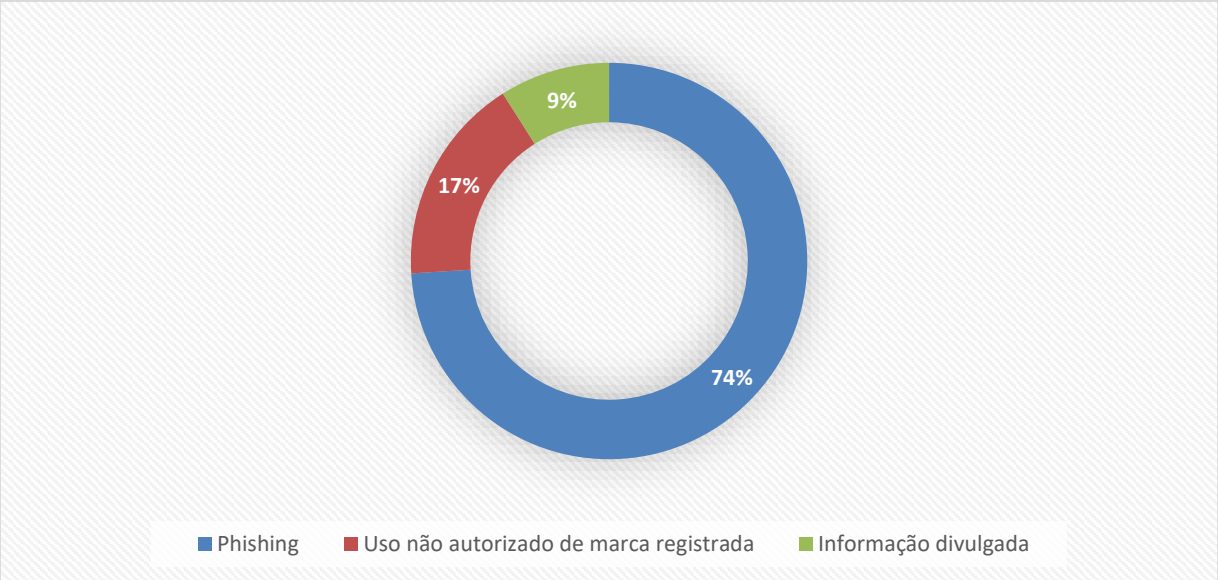
Dada a persistência desses ataques e a recente onda de golpes digitais que se intensificaram durante a pandemia e continuam até hoje, torna-se cada vez mais importante conduzir pesquisas voltadas para o desenvolvimento de estratégias de defesa contra o *phishing* (Souza; Tanaka, 2023).

As tentativas de *phishing* vêm crescendo exponencialmente. Segundo Romer (2023), houve um aumento alarmante desse tipo de ataque, especialmente na América Latina, onde os incidentes cresceram 617% em um ano, totalizando 286 milhões de tentativas de *phishing*.

O Brasil, sendo o país mais populoso da América Latina, tornou-se o principal alvo, registrando 134 milhões de tentativas de *phishing* nos últimos doze meses, um aumento de cinco vezes em relação às 25 milhões de tentativas registradas entre 2021 e 2022 (Romer, 2023).

De acordo com um relatório da Appgate (2023), elaborado por seus analistas do Security Operations Center (SOC), 74% dos incidentes de fraudes observados foram resultado de ataques de *phishing*, tornando-se o tipo de ataque mais comum entre os cibercriminosos, como apresentado no Gráfico 1.

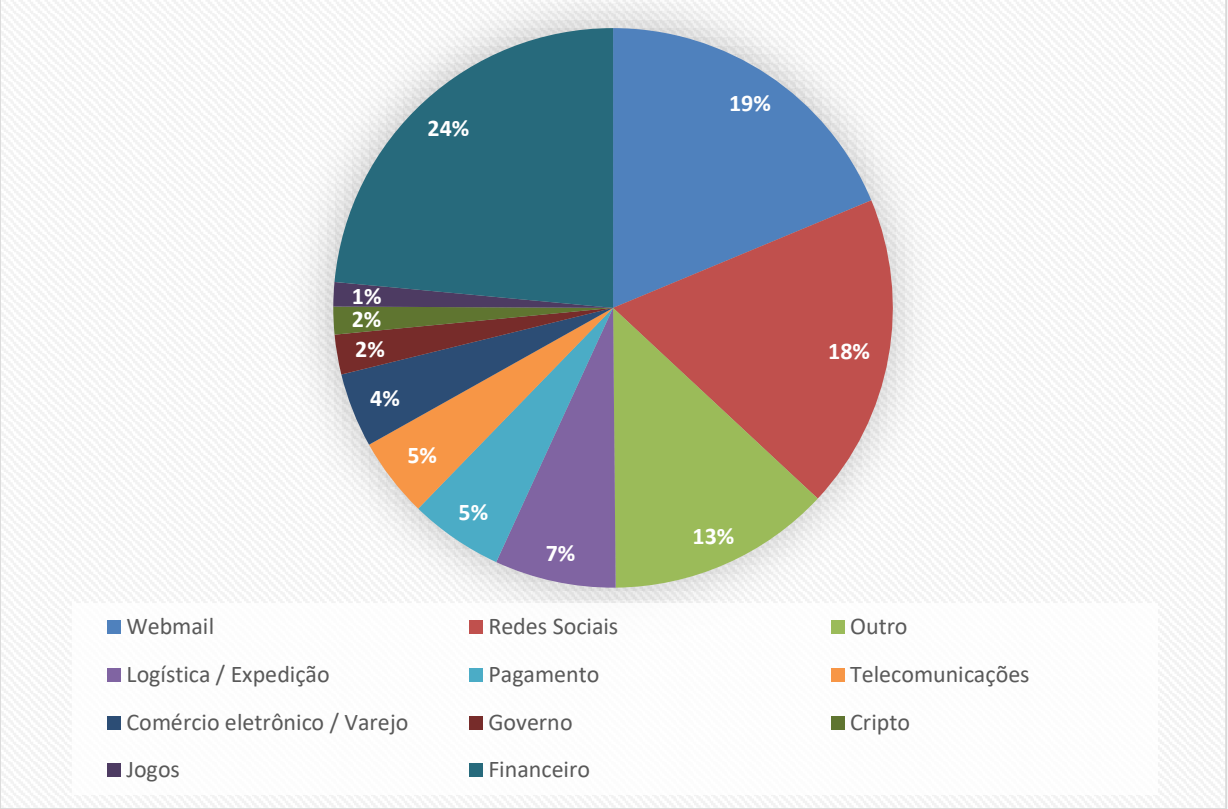
Gráfico 1 – Os Três Principais Tipos de Ataques em 2023



Fonte: Appgate (2024). Adaptado pelos autores (2024).

A APGW (2023) divulgou um relatório que identificou 1.624.144 ataques de *phishing*, um número recorde em comparação com suas análises anteriores. Os dados foram organizados por setores, destacando o setor financeiro como o mais afetado, representando 23,5% dos ataques. O Gráfico 2 apresenta essas informações de forma mais clara, ilustrando a distribuição dos ataques por setor.

Gráfico 2 – Indústrias Mais Visada Para Ataques no 1º Trimestre de 2023



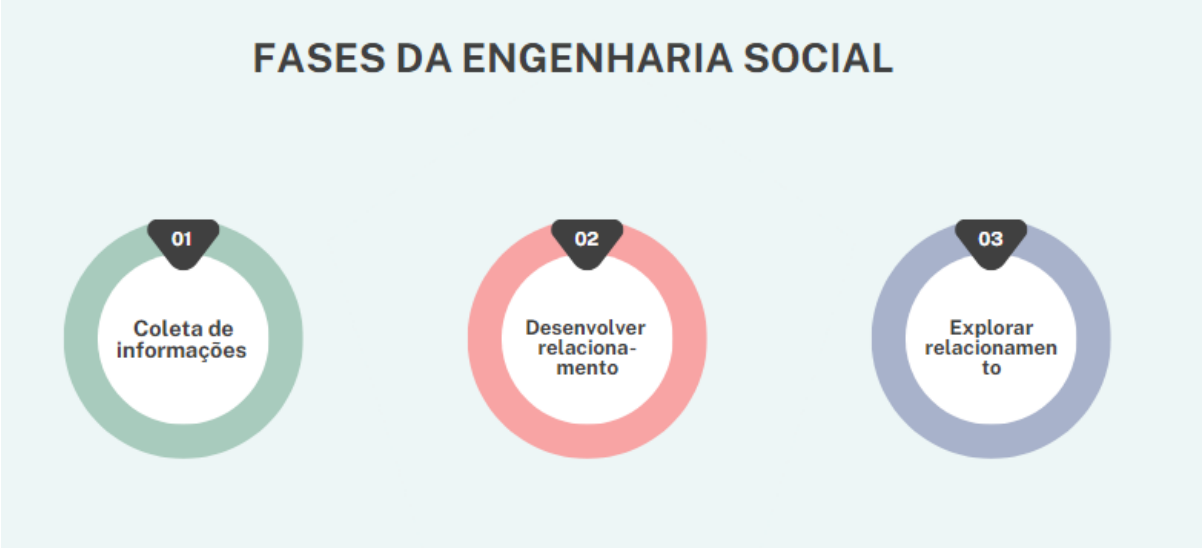
Fonte: APGW (2023). Adaptado pelos autores (2024).

1.4 Técnicas de Engenharia Social

O conceito de engenharia social baseia-se na exploração da psicologia humana por cibercriminosos, que aproveitam a ingenuidade, a falta de cuidado e a falta de conscientização das pessoas sobre como se proteger no meio digital. Esses criminosos manipulam suas vítimas para que revelem informações confidenciais, utilizando ações prejudiciais e links falsos (Fernandes, 2023).

De acordo com Henriques (2017), o ciclo da engenharia social divide-se em três fases: coleta de informações, desenvolvimento e exploração do relacionamento conforme representado na Figura 1.

Figura 1 – Fases da Engenharia Social



Fonte: Heniques (2017). Adaptado pelos autores (2024).

Na primeira etapa, identifica-se as fontes para obter as informações. Aqui, aplica-se o conceito de "*Footprinting*", que é o processo de observar e coletar informações sobre um alvo com a intenção de encontrar uma maneira de comprometer sua segurança.

Na segunda etapa, para desenvolver uma relação, é necessário criar confiança entre as partes. Utilizam-se, então, informações privilegiadas, citando pessoas conhecidas da vítima e particularidades, sempre mantendo uma imagem de boas intenções. Para o sucesso do engenheiro social, é necessário usar a técnica conhecida como "*Rapport*", que é a capacidade de construir um vínculo com alguém.

A terceira etapa é a exploração do relacionamento. O atacante induz um ataque emocional na vítima e inicia o processo de investigação ou elicitação para obter informações necessárias. Após a exploração, avalia-se o estado emocional do alvo para evitar que perceba o ataque.

Os impactos globais desses métodos incluem o roubo de propriedade intelectual, perda financeira e danos à reputação, usados para fins maliciosos, como extorsão ou chantagem. Além disso, as vítimas são persuadidas a clicar em links falsos oferecendo bens, prêmios e outros bônus (a famosa técnica de *phishing*), o que pode ter consequências negativas, como a instalação de malware nos dispositivos.

De acordo com Rabelo (2023), fica evidente que os criminosos utilizam princípios psicológicos para manipular suas vítimas. Esse fenômeno pode ser explicado pela abordagem do psicólogo Burrhus Frederic Skinner, que desenvolveu quatro conceitos aplicáveis à interpretação das estratégias de ataques de engenharia social:

- **Reforço positivo**: ocorre quando um comportamento é seguido por uma recompensa, aumentando a probabilidade de que seja repetido no futuro. No contexto da engenharia social, isso acontece quando os criminosos oferecem recompensas às vítimas com o objetivo de obter seus dados.

- **Reforço negativo**: ocorre quando um comportamento resulta na remoção de algo desagradável, como acordos sobre dívidas.

- **Punição positiva**: acontece quando um comportamento é seguido por uma consequência desagradável. Engenheiros sociais podem usar táticas como chantagem emocional ou ameaças, informando à vítima que a punição cessará caso ela coopere.

- **Punição negativa**: acontece quando um comportamento leva à remoção de algo desejável. Entre as estratégias comuns, destaca-se a ameaça de retirar algo valioso da vítima, como informações pessoais, caso ela não coopere. A vítima pode evitar a punição ao seguir as instruções do atacante.

Portanto, a engenharia social é uma estratégia de manipulação utilizada há muitos anos, representando uma ameaça significativa à segurança da informação, particularmente por seu impacto direto no aspecto emocional das vítimas.

2 METODOLOGIA

O método de pesquisa exploratória foi essencial para a fundamentação teórica deste trabalho, assim como os experimentos com as ferramentas atuais, tais como ChatGPT, Copilot e Gemini. A pesquisa exploratória permitiu um aprofundamento teórico nos temas de IA, ciberataques e *phishing*, enquanto os experimentos possibilitaram aplicações práticas. As ferramentas foram escolhidas por sua relevância e utilidade para o estudo.

O objetivo deste trabalho é ampliar o entendimento das defesas potencializadas pela IA, contribuindo significativamente para a área de segurança da informação. A pesquisa foca na exploração e compreensão das tecnologias mais recentes neste campo, além de oferecer uma visão atual de como a IA, um recurso poderoso para o avanço tecnológico e a otimização de processos, pode ser empregada para fins de segurança da informação.

Para obter um melhor entendimento dos temas abordados, recorremos a uma variedade de trabalhos acadêmicos, que nos permitiram realizar uma análise crítica e comparativa das diversas ideias e abordagens de cada autor.

Nosso trabalho consistiu na realização de testes e experimentos para validar nossas hipóteses. Os resultados obtidos foram utilizados na construção da conclusão final, aumentando a precisão e a confiabilidade deste estudo. Os instrumentos utilizados incluíram anotações relevantes, observações registradas e relatórios, o que elevou o nível de detalhamento do trabalho.

O tempo total para a realização do estudo foi de aproximadamente 10 meses, desde a condução das pesquisas até a realização dos experimentos. A equipe, composta por dois alunos do curso de Segurança da Informação, teve suas responsabilidades estrategicamente divididas, de modo que todos os membros tivessem uma compreensão global das atividades.

As amostras para análise de dados foram geradas com base nos experimentos e posteriormente selecionadas e organizadas conforme sua relevância para este trabalho. Para facilitar a compreensão dos dados, as informações foram apresentadas por meio de gráficos, tabelas e imagens.

A qualidade das informações foi garantida por uma abordagem quantitativa, que permitiu uma análise com detalhamento nos dados coletados nos experimentos.

2.1 Caracterização de Pesquisa.

O trabalho foi desenvolvido com o objetivo de ser uma pesquisa experimental, na qual buscamos, por meio de testes, expandir o conhecimento sobre o tema escolhido. Segundo Gil (2018), a pesquisa experimental é utilizada para investigar as influências geradas por ações sobre um alvo de estudo, em que o pesquisador precisa ser ativo, manipulando as variáveis e os métodos de controle, o que possibilita a criação de cenários específicos e uma inspeção minuciosa dos efeitos resultantes. Gil (2018) também destaca que esse método é amplamente respeitado na comunidade científica, sendo amplamente utilizado para comprovações de pesquisa.

2.2 Procedimentos para coleta e análise de dados.

Os métodos de coleta de dados foram realizados em várias etapas, para assegurar que os experimentos ocorressem de forma satisfatória:

- **Escolha dos equipamentos e ferramentas:** Inicialmente, foram definidos os equipamentos a serem utilizados (computadores com acesso à internet) e as ferramentas necessárias para os experimentos (IAs selecionadas).

- **Definição dos cenários de teste:** Foram elaborados cenários específicos para os testes de detecção de *phishing*, sempre com a maior proximidade possível da realidade.

- **Configuração do ambiente de teste:** A ambientação foi realizada dentro das Inteligências Artificiais utilizadas. Os cenários definidos foram fornecidos às IAs, bem como o papel que cada IA deveria desempenhar nos testes.

- **Execução dos experimentos:** Após a configuração, foram entregues mensagens potencialmente de *phishing* para análise de detecção.

- **Registro de dados:** Com o decorrer dos testes, os dados obtidos foram documentados e armazenados para posterior análise e conclusão.

A análise de dados foi realizada em etapas, para garantir a precisão e a relevância dos resultados:

- **Organização dos dados:** Os dados coletados foram devidamente categorizados por tipo de ataque, taxa de sucesso, taxa de falha, entre outros. Esse processo garantiu uma visualização clara dos resultados.

- **Análise quantitativa:** Com os resultados organizados, foi realizada uma análise quantitativa, utilizando métricas como taxas de sucesso e falha na detecção de *phishing*. Esses dados foram transformados em números e porcentagens, permitindo a identificação de padrões estatísticos que auxiliaram na compreensão objetiva dos resultados.

- **Conclusão**: Após a análise minuciosa dos dados, foram elaboradas as conclusões sobre os experimentos, destacando os testes mais relevantes para serem incluídos no trabalho.

2.2.1 Ambiente de coleta de dados.

Foi realizada uma pesquisa de laboratório para o desenvolvimento dos experimentos. Os computadores dos dois participantes do projeto serviram como ambiente de experimentação. Essa técnica permitiu o controle das variáveis e a obtenção de resultados precisos. Os testes foram conduzidos em um ambiente computacional planejado, utilizando mensagens de *phishing* reais previamente selecionadas, o que possibilitou a obtenção de dados sobre as respostas de detecção das Inteligências Artificiais utilizadas.

2.2.2 Técnicas para coleta de dados

O estudo empregou várias técnicas para garantir a precisão dos resultados. As principais técnicas utilizadas foram:

- **Anotações relevantes:** Durante a pesquisa, foram feitas anotações dos pontos principais, fundamentais para o desenvolvimento das referências teóricas.

- **Observação registrada:** Eventos e comportamentos foram documentados para análise posterior. Esta abordagem foi essencial após os experimentos.

- **Relatórios:** Os relatórios foram utilizados para registrar oficialmente as descobertas e os resultados da pesquisa, proporcionando uma compreensão completa e organizada dos dados.

Essas técnicas foram escolhidas por sua capacidade de fornecer uma visão detalhada do estudo.

2.2.3 Natureza da análise de dados

Este projeto é caracterizado por uma análise quantitativa, na qual foram observados vários casos para compor uma análise estatística. Segundo Suhete (2023), essa análise se concentra em dados que podem ser medidos e expressos numericamente, coletados por métodos padronizados, como questionários e observações sistemáticas. Esses dados são quantificados e analisados por meio de técnicas estatísticas, como média, desvio de padrão e correlação.

O principal objetivo é utilizar métodos estatísticos e ferramentas matemáticas para identificar padrões e tendências, resultando em informações quantificáveis. O tamanho da amostra, seja pequeno ou grande, é secundário, desde que a análise seja capaz de gerar novas informações e responder ao problema de pesquisa (Coelho, 2023).

3 ESQUEMA DE EXPERIMENTAÇÃO

Neste capítulo, será detalhado o formato da experimentação, incluindo a preparação do ambiente de testes, a escolha das ferramentas de IA generativa utilizadas e a metodologia de coleta de casos reais de *phishing*, de forma a explicar os passos.

3.1 Preparação de ambiente de testes

Para preparar nosso ambiente de testes, utilizamos mecanismos de *prompt* (comando/instrução). Inicialmente, será introduzido esse comando para, em seguida, realizarmos os demais testes.

Aplicamos o conceito de "alfabetização de *prompts*" (*prompt literacy*), segundo Jacobs e Fisher (2023), onde o domínio do uso de *prompts* aumenta a eficácia e a capacidade de resposta da IA. Um dos modelos de *prompt* utilizados é o "PREP", apresentado por Fitzpatrick (2023), que é uma maneira de preparar a IA para obter o melhor resultado. O modelo consiste em quatro etapas:

- **(P) *Prompt*:** Nesta etapa, o condutor informa à IA os próximos passos e seu objetivo.

- **(R) *Rule*:** O objetivo aqui é apresentar as regras de modo que a IA compreenda seu papel e o que precisa fazer.

- **(E) *Explicit*:** Nesta etapa, as instruções são inseridas, sendo o núcleo do comando, onde toda a explicação detalhada é feita.

- **(P) *Parameters*:** Por fim, são apresentados parâmetros claros para que a IA execute a tarefa.

O comando apresentado no Quadro 2 foi utilizado em todos os testes para padronizar as respostas e obter resultados mais consistentes.

Quadro 2 – Prompt Para Identificação de Phishing

Prompt**:** Identificação de Phishing Estou realizando testes para identificar a capacidade de detecção de phishing por inteligência artificial. Para isso, colocarei no final da mensagem um texto ou imagem para que você verifique se trata-se de phishing ou não. Para realizar essa identificação, você pode se basear em casos já ocorridos e utilizar a probabilidade em determinados contextos. Priorize informações de fontes confiáveis e com maior rigor. Preciso que determine a probabilidade do conteúdo apresentado ser um phishing, apontando quais trechos validam sua resposta. Também é importante uma explicação detalhada sobre sua análise, com justificativas claras. Organize sua resposta de forma clara e estruturada, destacando o percentual de probabilidade, explicando cada trecho analisado e apresentando suas justificativas. Texto/Imagem: ""

Fonte: Elaborado pelos autores (2024).

3.2 Ferramentas utilizadas para execução dos testes

Para a realização dos experimentos, optamos por utilizar três ferramentas de IA generativa: o ChatGPT da OpenAI, o Copilot da Microsoft e o Gemini da Google. A escolha dessas ferramentas visa oferecer uma análise mais abrangente dos resultados obtidos, sem a intenção de realizar uma comparação direta entre elas. Ao utilizá-las em conjunto, buscamos examinar e contrastar suas respostas e desempenhos em cenários de *phishing*.

Essa abordagem permite calcular uma média dos resultados, proporcionando conclusões mais robustas e confiáveis sobre a eficácia das ferramentas na detecção de ciberataques. Além disso, ao combinar as ferramentas, conseguimos minimizar

possíveis limitações individuais, aumentando a precisão e a validade dos resultados da pesquisa.

As ferramentas têm o mesmo princípio de funcionamento e se assemelham por utilizarem modelos de linguagem similares sendo o *Large Language Models* (LLMs). Segundo uma publicação da UOL (2023), o ChatGPT aprende através de repetições e utiliza algoritmos de redes neurais para processar grandes volumes de dados e gerar respostas. Quando um usuário faz uma pergunta, o modelo avalia a entrada e responde com base em probabilidades, buscando fornecer a resposta mais apropriada ao questionamento.

O que difere entre as ferramentas são os dados disponíveis. A Equipe de Comunicação da Solo Network (2024) afirma que as respostas estão diretamente ligadas ao conjunto de dados ao qual elas têm acesso. Enquanto o ChatGPT não possui acesso aos dados internos de uma organização, o Copilot, quando integrado ao ambiente Microsoft 365 e outras fontes de dados, pode oferecer respostas mais detalhadas e contextualizadas.

Mais semelhante ao ChatGPT, temos o Gemini. Sua semelhança reside nos dados utilizados para aprendizado. Como relata Teixeira (2023), o Google Gemini foi desenvolvido com base em um vasto conjunto de dados de texto e código, permitindo que ele produza textos, traduza idiomas, crie diversos tipos de conteúdo e responda a perguntas de maneira informativa.

3.3 Coleta de casos reais de *phishing* e e-mails legítimos

A utilização de casos reais foi fundamental para garantir maior veracidade e confiabilidade nos testes. A inclusão desses exemplos permitiu uma análise mais precisa e fundamentada do desempenho na detecção de *phishing* com IA. Os casos foram selecionados a partir de um catálogo público de fraudes, pertencente à Rede Nacional de Ensino e Pesquisa (RNP), garantindo uma amostra diversificada e representativa das ameaças atuais.

Além disso, foi realizada a filtragem de e-mails sem a aplicação de *phishing*, ou seja, e-mails reais com procedência legítima, a fim de verificar casos de falso positivo. A coleta foi feita manualmente, separando e-mails da nossa caixa de entrada e de amigos, nos quais havia assuntos mais sensíveis e recorrentes.

Essa abordagem não apenas fortalece a validade dos resultados obtidos, mas também possibilita a identificação de possíveis lacunas nos sistemas de IA em situações do mundo real. Ao utilizar um catálogo reconhecido e amplamente utilizado, conseguimos criar cenários de teste que refletem as condições enfrentadas por usuários comuns, proporcionando uma visão mais clara da eficácia das ferramentas de detecção de *phishing*.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

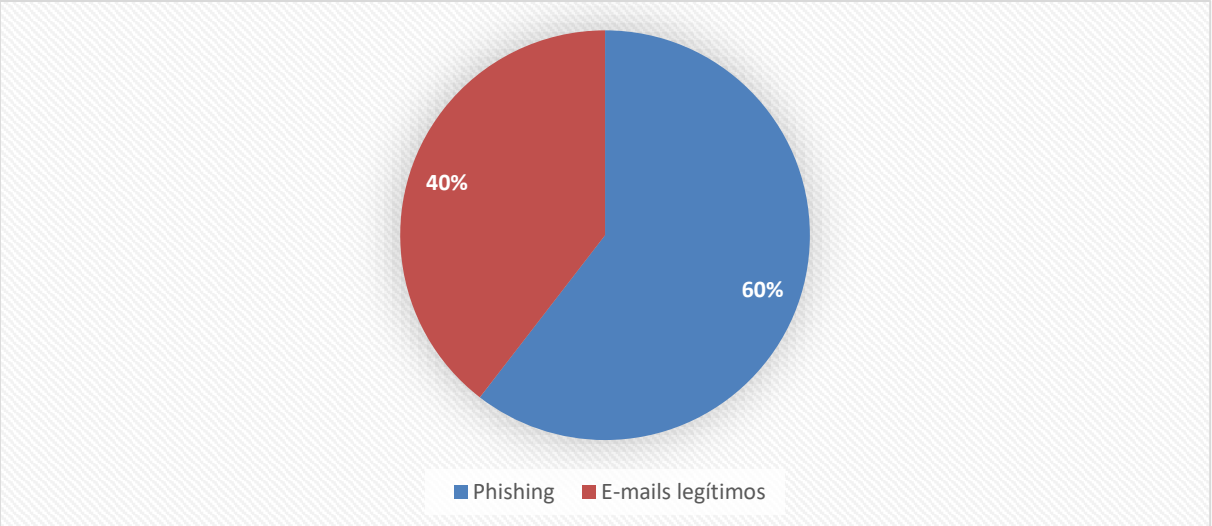
Neste capítulo, serão apresentados os experimentos realizados para testar a capacidade da inteligência artificial generativa de prevenir e combater ataques de *phishing*. O objetivo foi utilizar as ferramentas ChatGPT, Copilot e Gemini em um cenário real de ciberataque, verificando se esses ataques podem ser detectados e prevenidos em tempo real.

A realização desses experimentos foi essencial para validar as ideias e suposições discutidas na revisão da literatura. Além disso, os experimentos fornecem informações sobre a aplicabilidade e a eficácia dessas ferramentas em ambientes práticos, o que é fundamental para sua integração em esquemas de defesa cibernética. Portanto, busca-se, por meio desses testes, identificar as limitações, os desafios e as potencialidades das tecnologias de IA generativa no enfrentamento das ameaças cibernéticas.

4.1 Quantificação e Categorias dos Casos de Teste

Para a realização dos testes, foram coletados, ao todo, 129 casos, incluindo 78 casos de *phishing* e 51 e-mails legítimos. Como mostrado no Gráfico 3, é possível compreender a proporção entre os diferentes tipos de casos coletados.

Gráfico 3 – Casos Coletados Para Realização dos Testes

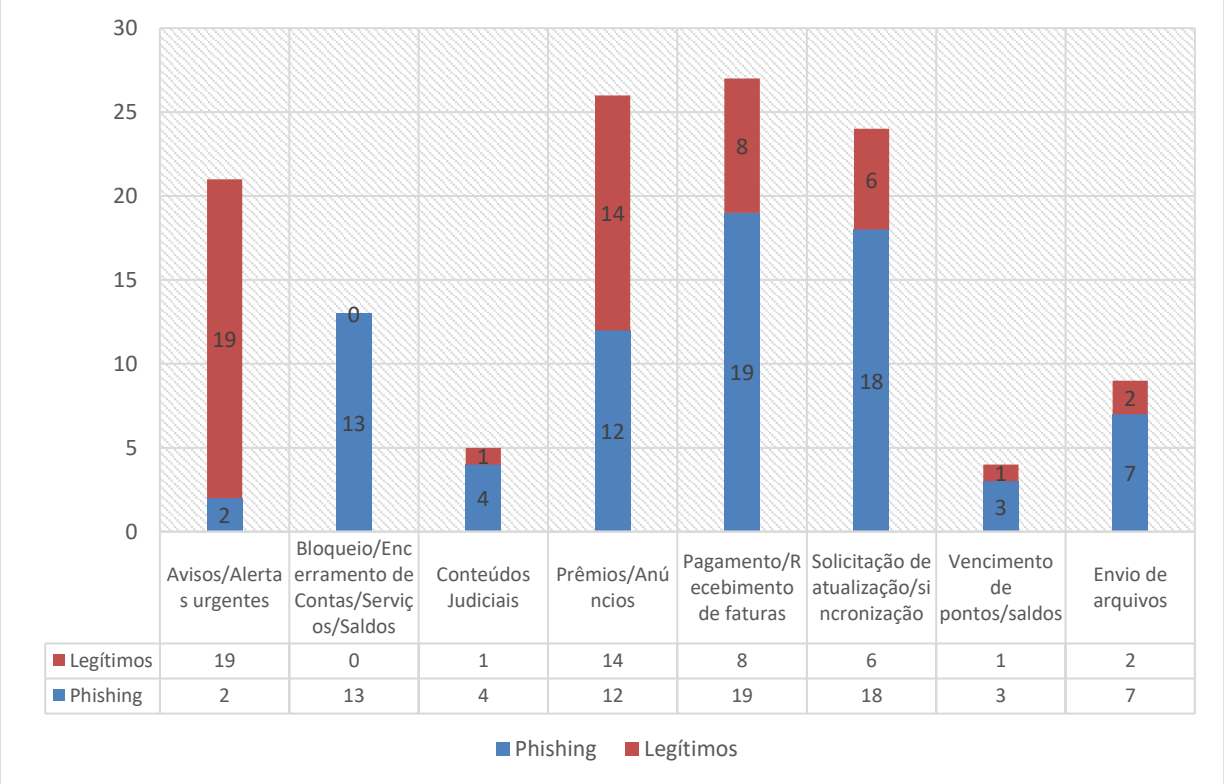


Fonte: Elaborado pelos autores (2024).

Dentre todos os casos analisados, organizamos os e-mails atribuindo-lhes uma categoria específica. Para isso, realizamos uma análise prévia do conteúdo e os separamos em 8 tipos distintos: “Avisos/Alertas urgentes”, “Bloqueio/Encerramento de Contas/Serviços/Saldos”, “Conteúdos Judiciais”, “Prêmios/Anúncios”, “Pagamento/Recebimento de Faturas”, “Solicitação de Atualização/Sincronização”, “Vencimento de Pontos/Saldos” e “Envio de Arquivos”.

O objetivo dessa categorização foi observar se o conteúdo dos e-mails poderia influenciar nos resultados dos testes, possibilitando uma compreensão sobre como diferentes tipos de e-mails afetam a detecção de *phishing*. A quantidade de e-mails em cada categoria pode ser visualizada no Gráfico 4.

Gráfico 4 – Casos Coletados Distinguidos por Assuntos



	Avisos/Alertas urgentes	Bloqueio/Encerramento de Contas/Serviços/Saldos	Conteúdos Judiciais	Prêmios/Anúncios	Pagamento/Recebimento de faturas	Solicitação de atualização/sincronização	Vencimento de pontos/saldos	Envio de arquivos
Legítimos	19	0	1	14	8	6	1	2
Phishing	2	13	4	12	19	18	3	7

Fonte: Elaborado pelos autores (2024).

4.2 Resultados obtidos dos testes

Com base nas métricas obtidas, este capítulo está organizado da seguinte forma: inicialmente, serão apresentados os casos de *phishing*, com destaque para os

resultados e as médias, separados por ferramenta de IA e, posteriormente, de forma unificada. Em seguida, serão expostos os resultados referentes aos e-mails legítimos, seguindo a mesma estrutura adotada para os casos de *phishing*.

4.2.1 Resultados obtidos dos e-mails de *phishing*

A partir dos 78 relatos coletados, cada um foi testado individualmente em cada ferramenta de IA, utilizando-se o mesmo prompt, totalizando 234 testes. Ao analisar os resultados obtidos pelo ChatGPT, conforme demonstrado na Tabela 1, observou-se uma pontuação média de 80,8 pontos, resultante de um total de 6.305 pontos, divididos pelo número total de casos.

Posteriormente, foram analisados os resultados obtidos pela ferramenta Copilot. Conforme apresentado na Tabela 1, a pontuação total alcançada foi de 6.585, resultando em uma média de 84,42 pontos, após a divisão pelo número total de casos testados.

Por fim, na análise do desempenho da ferramenta Gemini, constatou-se que a pontuação total foi de 6.360, resultando em uma média de 81,53 pontos, conforme demonstrado na Tabela 1.

Após a análise individual de cada ferramenta de IA, foi realizada a unificação dos resultados para a obtenção de uma média geral. Somando-se as pontuações obtidas por todas as ferramentas, obteve-se um total combinado de 19.250 pontos, que, ao ser dividido pelo número total de testes (234), resultou em uma média geral de 82,26 pontos.

Tabela 1 – Resultados dos Casos de Phishing.

Casos	**Pontuação**			**Categoria**
	ChatGPT	**Copilot**	**Gemini**	
Caso 01	90	90	90	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 02	85	90	95	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 03	75	70	95	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 04	80	70	85	Solicitação de atualização/sincronização

Caso 05	90	70	30	Conteúdos Judiciais
Caso 06	95	70	80	Envio de arquivos
Caso 07	90	95	95	Vencimento de pontos/saldos
Caso 08	70	10	5	Conteúdos Judiciais
Caso 09	90	80	85	Pagamento/Recebimento de faturas
Caso 10	70	75	90	Pagamento/Recebimento de faturas
Caso 11	70	80	80	Prêmios/Anúncios
Caso 12	75	90	80	Pagamento/Recebimento de faturas
Caso 13	90	80	95	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 14	85	90	90	Envio de arquivos
Caso 15	85	90	90	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 16	90	80	90	Avisos/Alertas urgentes
Caso 17	85	75	90	Vencimento de pontos/saldos
Caso 18	90	90	95	Solicitação de atualização/sincronização
Caso 19	70	90	80	Pagamento/Recebimento de faturas
Caso 20	95	90	95	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 21	60	60	90	Pagamento/Recebimento de faturas
Caso 22	85	90	95	Pagamento/Recebimento de faturas
Caso 23	20	10	5	Envio de arquivos
Caso 24	70	60	90	Pagamento/Recebimento de faturas
Caso 25	70	70	80	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 26	75	85	80	Avisos/Alertas urgentes
Caso 27	90	90	95	Pagamento/Recebimento de faturas
Caso 28	85	80	60	Pagamento/Recebimento de faturas
Caso 29	95	95	90	Conteúdos Judiciais
Caso 30	80	90	80	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 31	90	90	95	Envio de arquivos
Caso 32	85	85	85	Prêmios/Anúncios

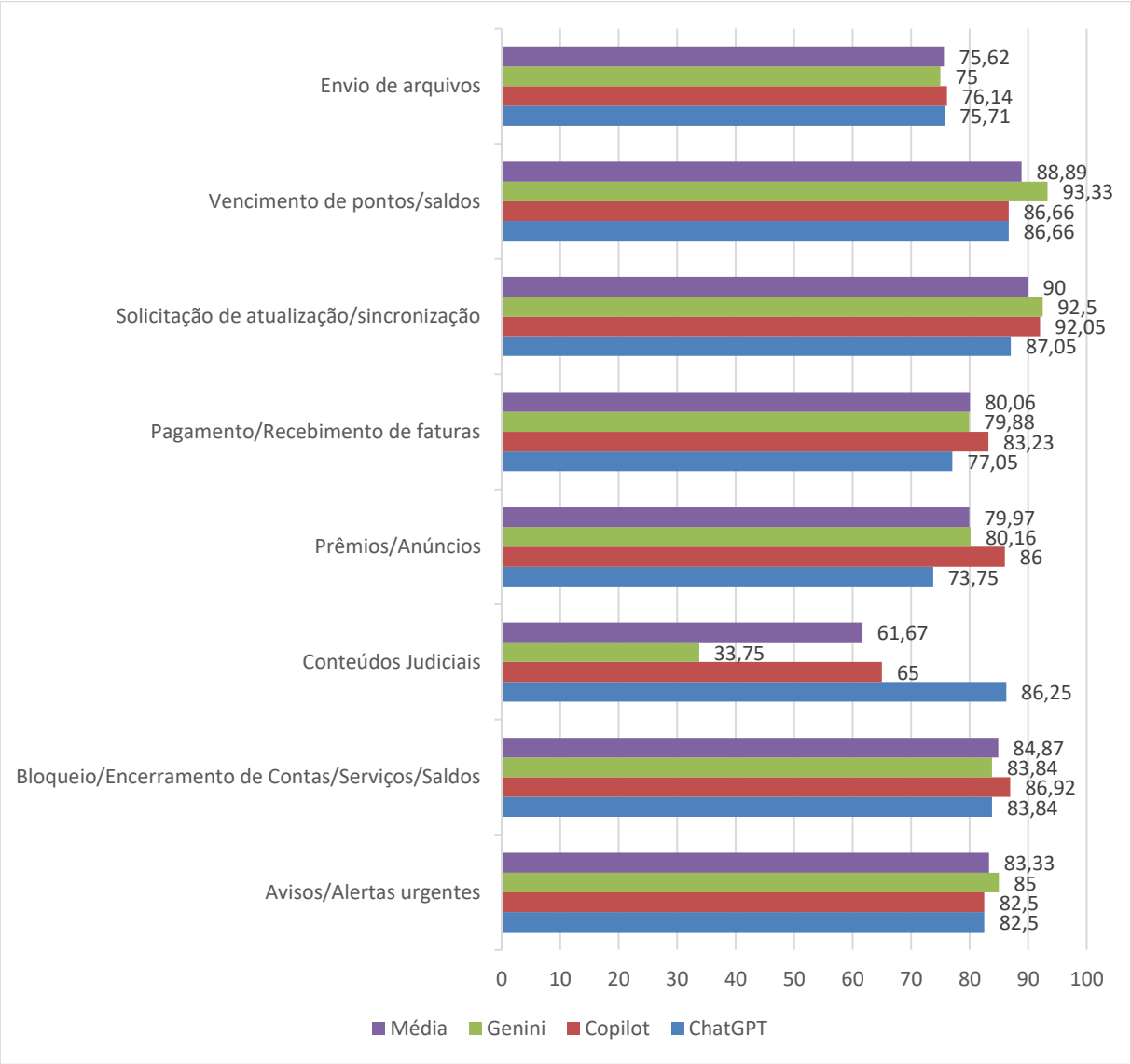
Caso 33	70	75	85	Prêmios/Anúncios
Caso 34	90	95	10	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 35	80	85	95	Solicitação de atualização/sincronização
Caso 36	85	90	95	Vencimento de pontos/saldos
Caso 37	90	90	90	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 38	85	85	80	Envio de arquivos
Caso 39	70	95	80	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 40	90	95	90	Solicitação de atualização/sincronização
Caso 41	20	85	2	Prêmios/Anúncios
Caso 42	90	90	80	Pagamento/Recebimento de faturas
Caso 43	40	70	80	Pagamento/Recebimento de faturas
Caso 44	80	95	95	Solicitação de atualização/sincronização
Caso 45	75	90	80	Prêmios/Anúncios
Caso 46	80	90	95	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 47	90	85	10	Conteúdos Judiciais
Caso 48	40	80	90	Prêmios/Anúncios
Caso 49	70	85	3	Pagamento/Recebimento de faturas
Caso 50	80	90	80	Prêmios/Anúncios
Caso 51	90	95	95	Solicitação de atualização/sincronização
Caso 52	60	90	80	Envio de arquivos
Caso 53	90	95	95	Solicitação de atualização/sincronização
Caso 54	95	98	95	Envio de arquivos
Caso 55	90	97	90	Prêmios/Anúncios
Caso 56	80	85	95	Pagamento/Recebimento de faturas
Caso 57	90	90	95	Bloqueio/Encerramento de Contas/Serviços/Saldos
Caso 58	95	95	95	Solicitação de atualização/sincronização
Caso 59	90	90	95	Solicitação de atualização/sincronização
Caso 60	95	90	95	Prêmios/Anúncios

Caso 61	80	85	95	Solicitação de atualização/sincronização
Caso 62	85	90	95	Solicitação de atualização/sincronização
Caso 63	90	80	80	Pagamento/Recebimento de faturas
Caso 64	85	85	90	Prêmios/Anúncios
Caso 65	90	85	90	Prêmios/Anúncios
Caso 66	95	95	95	Solicitação de atualização/sincronização
Caso 67	85	90	95	Solicitação de atualização/sincronização
Caso 68	85	95	95	Solicitação de atualização/sincronização
Caso 69	80	95	90	Solicitação de atualização/sincronização
Caso 70	85	90	95	Prêmios/Anúncios
Caso 71	90	85	80	Pagamento/Recebimento de faturas
Caso 72	80	85	80	Solicitação de atualização/sincronização
Caso 73	70	90	95	Pagamento/Recebimento de faturas
Caso 74	90	95	85	Solicitação de atualização/sincronização
Caso 75	95	95	95	Solicitação de atualização/sincronização
Caso 76	70	90	80	Pagamento/Recebimento de faturas
Caso 77	90	90	90	Pagamento/Recebimento de faturas
Caso 78	85	90	85	Pagamento/Recebimento de faturas

Fonte: Elaborado pelos autores (2024).

Para uma análise específica, o Gráfico 5 apresenta as perspectivas das pontuações médias dos casos de *phishing*, separados por categoria. Essa abordagem permite observar como cada categoria de ataque foi identificada pelas ferramentas de IA, proporcionando uma visão sobre a eficácia da detecção em diferentes tipos.

Gráfico 5 – Pontuações Médias dos Casos de Phishing Organizado por Categoria.



Fonte: Elaborado pelos autores (2024).

4.2.2 Resultados obtidos dos e-mails legítimos

Para a análise dos casos legítimos, foram coletados 51 e-mails, testados em três ferramentas distintas, totalizando 153 testes. A seleção desses e-mails teve como objetivo identificar possíveis ocorrências de falsos positivos, ou seja, situações em que a ferramenta identificou incorretamente um e-mail legítimo como *phishing*. Essa etapa foi fundamental para avaliar a precisão das ferramentas na distinção entre ameaças reais e comunicações autênticas.

Os resultados de cada ferramenta foram analisados individualmente. No caso da ferramenta ChatGPT, conforme demonstrado na Tabela 2, a pontuação total foi de 2.720 pontos, resultando em uma média de 53,33 pontos por teste.

Em seguida, a ferramenta Copilot apresentou um desempenho superior, alcançando uma pontuação total de 3.145 pontos, com uma média de 61,66 pontos por teste.

Por fim, os e-mails foram submetidos à ferramenta Gemini, que registrou uma pontuação total de 2.010 pontos, resultando em uma média de 39,31 pontos por teste.

Ao calcular a média das três ferramentas, somando-se as pontuações totais (2.720 + 3.145 + 2.010 = 7.875 pontos) e dividindo-se pelo número total de testes (153), obteve-se uma média geral de 51,47 pontos. Esse valor reflete o desempenho médio das ferramentas na identificação de falsos positivos em e-mails legítimos.

Tabela 2 – Resultados dos Casos de e-mails legítimos.

Casos	**Pontuação**			**Categoria**
	ChatGPT	**Copilot**	**Gemini**	
Caso 01	60	70	20	Solicitação de atualização/sincronização
Caso 02	15	40	30	Prêmios/Anúncios
Caso 03	65	40	60	Prêmios/Anúncios
Caso 04	30	20	50	Solicitação de atualização/sincronização
Caso 05	80	85	90	Avisos/Alertas urgentes
Caso 06	50	50	5	Avisos/Alertas urgentes
Caso 07	80	85	80	Solicitação de atualização/sincronização
Caso 08	10	10	5	Prêmios/Anúncios
Caso 09	20	70	10	Avisos/Alertas urgentes
Caso 10	60	65	50	Avisos/Alertas urgentes
Caso 11	5	20	10	Solicitação de atualização/sincronização
Caso 12	15	10	20	Avisos/Alertas urgentes
Caso 13	20	60	60	Avisos/Alertas urgentes
Caso 14	30	60	80	Prêmios/Anúncios
Caso 15	10	40	20	Avisos/Alertas urgentes
Caso 16	10	70	20	Pagamento/Recebimento de faturas
Caso 17	60	65	10	Avisos/Alertas urgentes

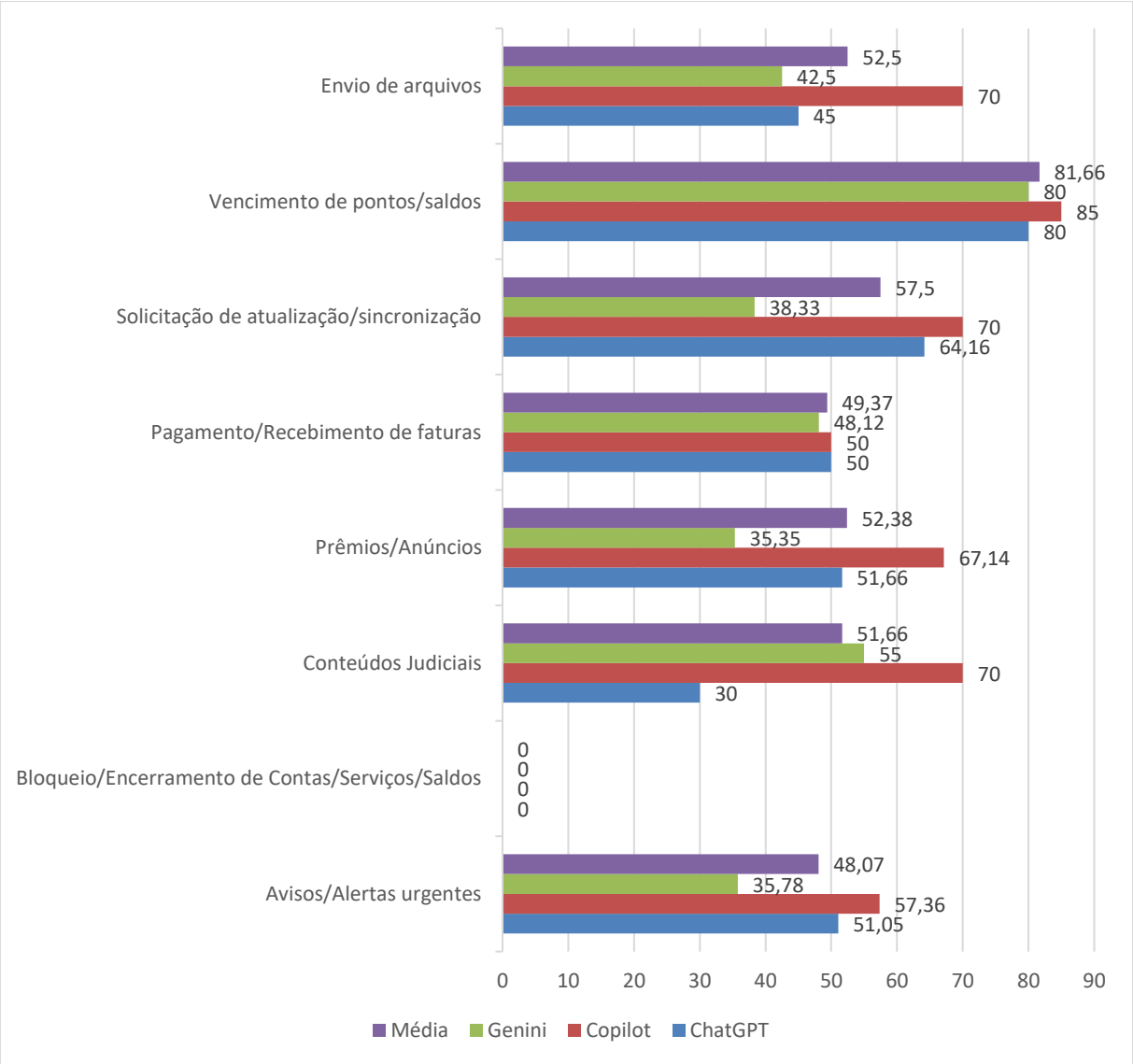
Caso 18	70	70	75	Avisos/Alertas urgentes
Caso 19	30	70	55	Prêmios/Anúncios
Caso 20	70	60	45	Prêmios/Anúncios
Caso 21	80	75	30	Avisos/Alertas urgentes
Caso 22	70	65	25	Avisos/Alertas urgentes
Caso 23	85	85	10	Prêmios/Anúncios
Caso 24	80	70	80	Avisos/Alertas urgentes
Caso 25	65	70	40	Avisos/Alertas urgentes
Caso 26	60	70	5	Pagamento/Recebimento de faturas
Caso 27	80	70	10	Solicitação de atualização/sincronização
Caso 28	25	30	5	Avisos/Alertas urgentes
Caso 29	70	70	10	Prêmios/Anúncios
Caso 30	70	70	80	Prêmios/Anúncios
Caso 31	75	70	5	Solicitação de atualização/sincronização
Caso 32	75	85	80	Pagamento/Recebimento de faturas
Caso 33	85	80	80	Solicitação de atualização/sincronização
Caso 34	75	70	20	Prêmios/Anúncios
Caso 35	80	70	80	Prêmios/Anúncios
Caso 36	70	70	20	Solicitação de atualização/sincronização
Caso 37	70	70	50	Avisos/Alertas urgentes
Caso 38	85	85	35	Avisos/Alertas urgentes
Caso 39	70	75	50	Solicitação de atualização/sincronização
Caso 40	65	40	20	Prêmios/Anúncios
Caso 41	75	70	80	Avisos/Alertas urgentes
Caso 42	70	70	70	Avisos/Alertas urgentes
Caso 43	70	70	10	Solicitação de atualização/sincronização
Caso 44	20	50	50	Avisos/Alertas urgentes
Caso 45	70	70	80	Avisos/Alertas urgentes
Caso 46	30	60	10	Prêmios/Anúncios
Caso 47	20	55	80	Avisos/Alertas urgentes
Caso 48	20	60	10	Pagamento/Recebimento de faturas
Caso 49	40	85	5	Avisos/Alertas urgentes
Caso 50	70	75	80	Avisos/Alertas urgentes

Caso 51	10	30	5	Prêmios/Anúncios

Fonte: Elaborado pelos autores (2024).

Da mesma forma como apresentado nos relatos de *phishing*, organizamos os e-mails legítimos em suas respectivas categorias, seguindo a mesma classificação anterior. O Gráfico 6 apresenta as médias de desempenho de cada ferramenta, proporcionando uma análise mais prática e visual dos resultados.

Gráfico 6 – Pontuações Médias dos Casos Legítimos Organizado por Categoria.



Fonte: Elaborado pelos autores (2024).

4.3 Análise dos resultados

Observando os dados obtidos, foi possível identificar os principais pontos positivos e negativos das ferramentas utilizadas (ChatGPT, Copilot e Gemini). Com base nos gráficos 5 e 6, é factível examinar individualmente o desempenho de cada uma, considerando as categorias de e-mails.

No caso do ChatGPT, em situações de phishing, a ferramenta alcançou uma média de 80,8 pontos. Seu melhor desempenho foi na categoria "Solicitação de atualização/sincronização", com 87,05 pontos, enquanto a pior pontuação ocorreu na categoria "Prêmios/Anúncios", com 73,75 pontos. Isso demonstra que o ChatGPT manteve uma constância razoável na detecção de *phishing* em todas as categorias. No entanto, ao analisar os e-mails legítimos para identificar falsos positivos, a ferramenta obteve uma média de 53,33 pontos. O melhor desempenho ocorreu na categoria "Conteúdos Judiciais", com apenas 30 pontos, enquanto o pior percentual foi na categoria "Vencimento de pontos/saldos", com 80 pontos. Esses resultados indicam que o ChatGPT apresentou um desempenho inverso nos casos de *phishing* e falsos positivos.

Para o Copilot, a média em casos de *phishing* foi de 84,42 pontos. Seu maior desempenho foi na categoria "Solicitação de atualização/sincronização", com 92,05 pontos, e o menor resultado foi em "Conteúdos Judiciais", com 65 pontos. Nos testes com e-mails legítimos, a média do Copilot foi de 61,66 pontos. A melhor categoria foi "Pagamento/Recebimento de faturas", com 50 pontos, e a pior pontuação foi em "Vencimento de pontos/saldos", com 85 pontos. Embora eficaz para *phishing*, o Copilot apresentou dificuldades em lidar com e-mails legítimos.

O Gemini alcançou uma média de 81,53 pontos em casos de *phishing*. Sua melhor pontuação foi na categoria "Vencimento de pontos/saldos", com 93,33 pontos, e o desempenho mais baixo foi na categoria "Conteúdos Judiciais", com 33,75 pontos. Nos e-mails legítimos, o Gemini obteve sua melhor pontuação na categoria "Prêmios/Anúncios", com 35,35 pontos, e o pior desempenho na categoria "Vencimento de pontos/saldos", com 80 pontos. O Gemini se mostrou eficaz em quase todas as categorias de e-mails verdadeiros, com exceção de algumas situações específicas.

De modo geral, o ChatGPT e o Copilot tiveram boas médias na detecção de *phishing*, mas foram inconsistentes na identificação de e-mails legítimos,

apresentando uma alta taxa de falsos positivos. O Gemini, por sua vez, demonstrou um desempenho equilibrado em ambos os testes, exceto em casos de *phishing* jurídicos e e-mails legítimos envolvendo vencimento de pontos e saldos.

Com base nos resultados obtidos, é possível avaliar a viabilidade de utilizar essas ferramentas para neutralizar ataques de *phishing*. Para responder à questão central deste estudo — "É possível detectar e-mails de *phishing* com IA?" —, pode-se afirmar que sim, a detecção é possível. Entretanto, outros fatores devem ser considerados, como a eficácia dessas ferramentas.

A eficácia, neste contexto, refere-se à capacidade das IAs de atingir seus objetivos, ou seja, distinguir com precisão e-mails de *phishing* e legítimos. Para este estudo, foi estabelecido que a IA classifica um e-mail como *phishing* quando sua pontuação é igual ou superior a 60 pontos.

De acordo com os dados apresentados na Tabela 3, a taxa de acerto na detecção de *phishing* foi de 94,87%, um resultado excelente. No entanto, a taxa de acerto para identificar e-mails legítimos foi de 66,67%, o que, apesar de inferior, ainda permite que a IA seja considerada funcional para essa tarefa.

Ao considerar o desempenho geral das IAs, levando em conta tanto *phishing* quanto e-mails legítimos, a média de acerto foi de 80,77%. Esse valor reflete uma boa capacidade de detecção, embora evidencie a necessidade de melhorias na identificação de e-mails legítimos, onde ainda ocorre uma taxa significativa de erros.

Tabela 3 – Comparação Entre Resultados dos Casos Analisados.

E-mails	Quantidade	Resultado < 60	Resultado >= 60
Phishing	78	4	74
Legítimos	51	34	17

Fonte: Elaborado pelos autores (2024).

Portanto, conclui-se que as ferramentas de IA são capazes de identificar e-mails e tomar decisões automáticas sobre eles. No entanto, é necessário aprimorar a detecção de e-mails legítimos, cujo desempenho se mostrou menos eficaz. A adoção de técnicas de aprendizado de máquina (*Machine Learning* - ML) poderia potencialmente melhorar os resultados, uma vez que essas tecnologias são projetadas e treinadas para otimizar o reconhecimento e a classificação de padrões complexos, como os observados em e-mails de *phishing*.

5 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi testar ferramentas de IA na detecção de *phishing*. A hipótese inicial previa uma eficácia de 85% na neutralização desses ataques, que foi superada, com uma taxa de acerto de 94,87%. Todos os objetivos específicos foram atingidos, sendo os experimentos conduzidos com sucesso e permitindo uma análise detalhada da eficiência das IAs.

Ferramentas gratuitas baseadas em IA, como ChatGPT, Copilot e Gemini, foram utilizadas onde foi utilizado um prompt para a tarefa de detecção de *phishing*. Diversos relatos foram coletados e aplicados individualmente em cada ferramenta. Ao final, um relatório detalhado foi gerado, destacando os resultados obtidos e as observações descritas ao longo do estudo.

Conclui-se que o uso dessas ferramentas foi eficaz na identificação de e-mails de *phishing*. Contudo, durante o processo, foram observadas dificuldades em reconhecer e-mails legítimos, resultando em uma taxa de 66,67% de acertos nesse tipo de detecção, o que, embora satisfatório, ainda está abaixo do ideal. Esse percentual evidencia uma tendência das ferramentas a apresentar falsos positivos, algo que merece atenção em futuros aprimoramentos.

Apesar dos resultados satisfatórios, algumas dificuldades foram enfrentadas durante a execução deste projeto. O primeiro desafio foi reunir uma amostra robusta de casos de *phishing*, e o segundo foi o esforço demandado pela execução manual dos 387 testes, o que consumiu considerável tempo e recursos.

Para aprimorar futuros estudos na área, algumas sugestões de melhorias foram identificadas:

1. **Uso de técnicas de *machine learning* (ML)**: A implementação de técnicas avançadas de ML pode fornecer instruções mais detalhadas para distinguir e-mails legítimos de *phishing*. A criação de uma base de conhecimento abrangente e consistente, composta por padrões suspeitos observados regularmente, bem como um catálogo de e-mails de remetentes confiáveis, pode contribuir significativamente para aprimorar a acurácia das ferramentas.
2. **Incorporação de mais detalhes nos dados de teste**: Para melhorar os resultados, é recomendável incluir informações adicionais, como o endereço do remetente e elementos visuais (como imagens no corpo do e-mail). Esses

dados adicionais podem ajudar a IA a detectar nuances que atualmente são ignoradas, além de aumentar e padronizar a quantidade de casos por categoria de e-mail, otimizando os testes.

Essas melhorias sugerem que o desenvolvimento de IAs para detecção de *phishing* pode ser potencializado com o uso de técnicas mais sofisticadas e um conjunto de dados mais robusto, levando a uma detecção mais precisa e eficiente tanto de e-mails fraudulentos quanto de mensagens legítimas.

REFERÊNCIAS

APPGATE. **Fraud Beat Annual Report**: A comprehensive analysis of electronic fraud trends and threats throughout 2023. 19 mar. 2024. Disponível em: https://www.appgate.com/resources/ebooks/fraud-beat-annual-report. Acesso em: 21 abr. 2024.

APWG. **Unifying the Global Response to Cybercrime**. Phishing activity trends report. 1º trimestre de 2023. 02 nov. 2023. Disponível em: https://apwg.org/trendsreports/. Acesso em: 21 abr. 2024.

BARBOSA, Lucia Martins; PORTES, Luiza Alves Ferreira. **Inteligência artificial**. Revista Tecnologia Educacional, Rio de Janeiro, n. 236, p. 16-27, 2023. Acesso em: 08 abr. 2024.

COELHO, Beatriz. **"Quanto?" — como fazer uma análise quantitativa dos dados?** 10 maio. 2023. Disponível em: https://blog.mettzer.com/analise-quantitativa/. Acesso em: 29 ago. 2024.

DE MORAIS, Diogo Martins Gonçalves *et al*. **O conceito de inteligência artificial usado no mercado de softwares, na educação tecnológica e na literatura científica**. Educação Profissional e Tecnológica em Revista, v. 4, n. 2, p. 98-109, 2020. Acesso em: 07 abr. 2024.

ESET. **Vishing: o que é, como identificar e se proteger?** 18 set. 2023. Disponível em: https://www.eset.com/br/artigos/vishing/#:~:text=O%20Vishing%20%C3%A9%20uma%20pr%C3%A1tica,dados%20pessoais%20das%20suas%20v%C3%ADtimas/. Acesso em: 22 abr. 2024.

FERNANDES, Oerton. **O Impacto Global da Engenharia Social e a Importância das Camadas de Proteção.** 20 jul. 2023. Disponível em: https://www.linkedin.com/pulse/o-impacto-global-da-engenharia-social-e-import%C3%A2ncia-das-msc-com-or/. Acesso em: 18 jun. 2024.

FITZPATRICK, Dan. **AI Won't Replace Teachers, but it will replace teachers who don't use AI.**, 27 fev. 2023. Disponível em: https://www.linkedin.com/pulse/ai-wont-replace-teachers-who-dont-use-dan-fitzpatrick/. Acesso em: 07 jul. 2024.

GIL, A. C. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2018. Acesso em: 01 mai. 2024.

GUEDES, Ronan de Paula; MOREIRA, **Junior. Uma análise das tecnologias de detecção e mitigação na identificação de páginas de phishing.** Instituto Federal Triângulo Mineiro: 10º EnPE - Encontro de Ensino, Pesquisa e Extensão. v. 10 n. 1. 22 nov. 2023. Disponível em: http://enpe.ptc.iftm.edu.br/index.php/enpe/article/view/339. Acesso em: 20 abr. 2024.

HENRIQUES, Francisco de Assis Fialho. **A influência da Engenharia Social no fator humano das organizações. 06 mar. 2017.** Disponível em: https://repositorio.ufpe.br/handle/123456789/25353. Acesso em: 21 jun. 2024.

IBM. **O que é smishing (phishing por SMS)?** Disponível em: https://www.ibm.com/br-pt/topics/smishing. Acesso em: 21 abr. 2024.

IBM. **O que é spear phishing?** Disponível em: https://www.ibm.com/br-pt/topics/spear-phishing. Acesso em: 21 abr. 2024.

IBM. **O que é whale phishing?** Disponível em: https://www.ibm.com/br-pt/topics/whale-phishing/. Acesso em: 21 abr. 2024.

JACOBS, Heidi Hayes; FISHER, Michael. **Prompt Literacy: A Key for AI-Based Learning.** Vol. 80, No. 9, 26 jun. 2023. Disponível em: https://ascd.org/el/articles/prompt-literacy-a-key-for-ai-based-learning. Acesso em: 20 abr. 2024.

MACHADO, Alexandre de Oliveira Bittencourt *Machine***. A inteligência artificial generativa como agente disruptor de mercado**. 2023. Acesso em: 10 abr. 2024.

MICROSOFT. **O que é um ataque cibernético?** Microsoft, 2024. Disponível em: https://www.microsoft.com/pt-br/security/business/security-101/what-is-a-cyberattack. Acesso em: 27 abr. 2024

OLIVERIA, P. **Que prejuízos um ciberataque pode gerar para uma empresa?** Untangle Brasil, 29 jul. 2023. Disponível em: https://www.untanglebrasil.com.br/que-prejuizos-um-ciberataque-pode-gerar-para-uma-empresa-2/. Acesso em: 25 abr. 2024.

PIOVESAN, Leonardo. Gubert; SILVA, Edilmárcio. Reis. Costa; SOUSA, Jakson. Ferreira de; TURIBUS, Sérgio. Noleto. **Engenharia social: Uma abordagem sobre Phishing.** Revista Científica da Faculdade de Balsas: Desenvolvimento regional. v. 10 n. 1. 03 dez. 2019. DOI: https://doi.org/10.46761/unibalsas.v10i1.94.
Disponível em: https://revista.unibalsas.edu.br/index.php/unibalsas/article/view/94.
Acesso em: 13 abr. 2024.

UOL. **Como usar o Chat GPT para acelerar a divulgação do seu negócio com inteligência artificial.** 22 maio. 2023. Disponível em: https://meunegocio.uol.com.br/blog/chatgpt/. Acesso em: 18 ago. 2024.

RABELO, **Gabriela. Engenharia Social: Como criminosos usam a Psicologia para manipular suas vítimas**. 06 set. 2024. Disponível em: https://www.dio.me/articles/engenharia-social-como-criminosos-usam-a-psicologia-para-manipular-suas-vitimas/. Acesso em: 21 jun. 2024.

ROMER, Rafael. **Ransomware recua em 2023, mas América Latina vive 'epidemia' de phishing.** IT Forum. 22 ago. 2023. Disponível em: https://itforum.com.br/noticias/ransomware-2023-epidemia-phishing/. Acesso em: 21 abr. 2024.

SALVIANO, Edgard Mesquita; SANTOS, João Pedro Ribeiro; SILVA, Matheus Almeida. **Principais tipos de ataques Phishing e mecanismos de segurança.** UNICEPLAC. 08 jul. 2022. Disponível em: https://dspace.uniceplac.edu.br/handle/123456789/1611. Acesso em 21 abr. 2024.

SANTOS, Fernando Lins; GONZAGA, José Luiz Teixeira. Segurança digital: **Ataques de Phishing e os modelos atuais de privacidade de dados**. Revista FT: Engenharias, v. 28, Edição 132, 30 mar. 2024. SSN 1678-0817 Qualis B2. DOI: 10.5281/zenodo.10897930. Disponível em: https://revistaft.com.br/seguranca-digital-ataques-de-phishing-e-os-modelos-atuais-de-privacidade-de-dados/. Acesso em: 20 abr. 2024.

SOLO NETWORK. **Microsoft Copilot X ChatGPT: entenda as diferenças.** 11 jun. 2024. Disponível em: https://solonetwork.com.br/blog-post/post/2024/06/11/microsoft-copilot-x-chatgpt-entenda-as-diferencas. Acesso em: 18 ago. 2024.

SOUZA, Leonardo Correa de; TANAKA, Simone Sawasaki. **Estudo sobre ataques de Phishing e suas técnicas de defesa.** Revista Terra & Cultura: Cadernos de Ensino e Pesquisa, [S.l.], v. 39, n. especial, p. 90-95, fev. 2023. ISSN 2596-2809. Disponível em: http://periodicos.unifil.br/index.php/Revistateste/article/view/2804. Acesso em: 18 abr. 2024.

SUHETE, Ricardo. **Qual é a diferença entre análise quantitativa e qualitativa?** 01 jun. 2023. Disponível em: https://pt.linkedin.com/pulse/29-pergunta-do-dia-qual-%C3%A9-diferen%C3%A7a-entre-an%C3%A1lise-e-ricardo-suhete/. Acesso em: 27 ago. 2024.

TEIXEIRA, Renato. **Desvendando o Google Gemini: Inteligência Artificial do Google Explorada.** 17 out. 2023. Disponível em: https://www.teixos.com.br/inteligencia-artificial-do-google/. Acesso em: 24 ago. 2024.

TIMPONE, Rich; GUIDI, Michel. **Explorando a mudança de cenário da IA. Da IA analítica à IA generativa**, p. 2023-05, 2023. Acesso em: 11 abr. 2024.

ZEFERINO, D. **O que são ciberataques, como acontecem e como prevenir?** 17 nov. 2020. Disponível em: https://www.certifiquei.com.br/ciberataques/. Acesso em: 20 abr. 2024.