UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DANIEL DOS SANTOS SALLES

PLATAFORMA DE DECISÃO MISTA, IA E HUMANA, PARA DETECÇÃO DE FRAUDES EM E-MAILS

CURITIBA

2025

DANIEL DOS SANTOS SALLES

PLATAFORMA DE DECISÃO MISTA, IA E HUMANA, PARA DETECÇÃO DE FRAUDES EM E-MAILS

DECISION PLATFORM FOR AI AND HUMAN COLLABORATION IN EMAIL FRAUD DETECTION

Trabalho de conclusão de curso de Graduação apresentada como requisito para obtenção do título de Bacharel em Engenharia da Computação da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Alexandre Graeml

CURITIBA
2025

DANIEL DOS SANTOS SALLES

PLATAFORMA DE DECISÃO MISTA, IA E HUMANA, PARA DETECÇÃO DE FRAUDES EM E-MAILS

Trabalho de conclusão de curso de Graduação apresentado como requisito parcial para obtenção do título de Bacharel em Engenharia da Computação da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 21/Fevereiro/2025

_____

Alexandre Reis Graeml
Doutorado
Universidade Tecnológica Federal do Paraná

_____

Daniel Fernando Pigatto
Doutorado
Universidade Tecnológica Federal do Paraná

_____

Rita Cristina Galarraga Berardi
Doutorado
Universidade Tecnológica Federal do Paraná

CURITIBA
2025

Aos professores do curso de Engenharia da Computação que me forneceram todas as bases necessárias para a realização deste trabalho, agradeço com profunda admiração pelo vosso profissionalismo.

AGRADECIMENTOS

Agradeço especialmente aos meus pais, que me tornaram resoluto no caminho de concluir a graduação, independente das enormes dificuldades impostas pelo caminho.

Agradeço à minha parceira e luz do meu caminho, Mírian Ferreira.

Agradeço ao meu orientador Prof. Alexandre Graeml, pelos conselhos e total solicitude.

Agradeço, por fim, a todos os servidores e funcionários, do Reitor aos estagiários, que tornam a universidade tecnológica uma instituição plena.

“O mistério da cooperação, afinal, é que Olson estava correto: é racional aproveitar-se dos outros, mas ainda assim a cooperação, em pequena e larga escala, permeia qualquer sociedade sã. [...] Pois existe algo irredutível em seus corações e isso marca a diferença entre *sociedade*, de um lado, e apenas *pessoas vivendo juntas* do outro.”

Surowiecki (2005, cap. 6).

RESUMO

A detecção de fraudes digitais é um trabalho crítico tanto para as pessoas quanto para o mundo corporativo. A falta de acurácia na detecção de fraudes em e-mails pode gerar grandes impactos na segurança digital em operações de negócios. Os métodos tradicionais para constatar fraudes em um e-mail implementam longas árvores de decisão e modelos de *machine learning* que acarretam perda de tempo e processamento desnecessário, além da pouca efetividade. Este trabalho avalia a efetividade de complementar a Inteligência Artificial com a Inteligência Coletiva, adotando-se grupos mistos (Inteligência humana + IA) na resolução de problemas. Para isso, foi elaborada uma plataforma digital. No âmbito técnico foi desenvolvido um padrão arquitetônico para a plataforma, com base no qual foi realizada uma prova de conceito. A discussão dos resultados da aplicação desta prova de conceito em comunidades digitais é apresentada para subsidiar a conclusão sobre a melhoria de desempenho obtida a partir da utilização de grupos mistos de humanos e IA para a tomada de decisão.

Palavras-chave: Inteligência coletiva; dados confiáveis; fraude no e-mail; sabedoria das multidões; inteligência artificial.

ABSTRACT

Digital fraud detection is a critical task for both individuals and businesses. Inaccuracies in detecting fraud in emails can have significant impacts on digital security in business operations. Traditional methods for identifying email fraud involve lengthy decision trees and machine learning models, leading to wasted time and unnecessary processing, with limited effectiveness. This work aims to complement Artificial Intelligence with Collective Intelligence, demonstrating the effectiveness of using mixed groups (Human Intelligence + AI) to develop a digital platform. The technical scope will present an architectural design proposal for this platform along with a proof of concept, including a brief study of the results from applying this proof of concept within digital communities.

Keywords: Collective intelligence; reliable data; e-mail fraud; wisdom of crowd; artificial intelligence.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
IA	Inteligência Artificial
ML	*Machine Learning*
IC	Inteligência Coletiva
LLM	*Large Language Model*
UTFPR	Universidade Tecnológica Federal do Paraná

LISTA DE SÍMBOLOS

T	Temperatura
σ	Desvio padrão
F1	F-Score
r	Fator de correlação
n	Número de indivíduos distintos

SUMÁRIO

1 INTRODUÇÃO

A falta de acurácia na detecção de fraudes em correios eletrônicos pode gerar grandes impactos na segurança digital em operações de negócios, tanto em cenários em que e-mails fraudulentos são classificados como não fraudulentos quanto em cenários em que e-mails seguros são classificados erroneamente como fraudulentos. Este trabalho se insere como uma tentativa multidisciplinar de abordar o conceito de inteligência coletiva de forma tecnológica, entendendo os fatores que interferem no processo decisório de um grupo inteligente de indivíduos e como isso pode ser aplicado e medido para melhorar a qualidade das decisões, independentemente de esses grupos serem compostos por indivíduos humanos ou agentes de inteligência artificial.

Novos conceitos de inteligência artificial serão adotados em conjunto com conhecimentos de inteligência coletiva com a finalidade de criar e avaliar um procedimento superior de classificação de e-mails fraudulentos.

Para tanto, serão estabelecidas medidas de inteligência coletiva para comparar o seu desempenho com o obtido a partir da aplicação da inteligência individual dos indivíduos ou agentes de inteligência artificial, isoladamente. Por meio da plataforma digital que foi desenvolvida pretende-se avaliar se é possível melhorar a qualidade da detecção de e-mails fraudulentos adotando-se princípios de Inteligência Coletiva e Inteligência Artificial, concomitantemente.

1.1 CONTEXTUALIZAÇÃO

As Tecnologias de Informação e Comunicação (TICs) tiveram vertiginoso crescimento quanto às fraudes digitais que a sua introdução possibilitou. Dentre essas fraudes, os e-mails perniciosos se destacam pelos danos que acarretam quando ocorrem em ambientes profissionais ou sociais. Entre as fraudes dos e-mails destaca-se o *phishing*, que tem por objetivo enganar os usuários e conseguir dados sensíveis para a realização de transações criminosas, o que traz grande prejuízo para os indivíduos e organizações.

Para manter a sensação de segurança dos usuários, muitas organizações assumem os prejuízos resultantes de transações fraudulentas, repassando os

custos ao conjunto de usuários e clientes (CAMPOS, 2022), o que onera toda a sociedade em uma escala crescente, tanto quantitativa como qualitativamente.

O *Anti-Phishing Working Group* (APWG) é uma coalizão internacional de equipes de combate ao crime cibernético. Ela relata que houve 877.536 ataques por *phishing* no segundo trimestre de 2024. Houve, ainda, um aumento de US$ 89.520, acima do trimestre anterior por transferência eletrônica solicitada, onde o cliente voluntariamente envia quantias de dinheiro a um estelionatário por alguma forma de convencimento. Calcula-se que, globalmente, o *phishing* cause perdas na casa dos bilhões de dólares a cada ano, com tendência de crescimento deste tipo de ataque.

Segundo um relatório da Symantec (2023), os e-mails fraudulentos são o principal meio adotado pelos golpistas para realizarem seus crimes, utilizando-se de técnicas aprimoradas de engenharia social, que dificultam a identificação da sua intenção maliciosa. As ações dos criminosos não poupam sequer as instituições de ensino.

A UTFPR obrigou-se a lançar um comunicado, como visto na Figura 1, para que seus integrantes se prevenissem do ataque por um e-mail falso, que tentava se passar pela equipe de suporte técnico da UTFPR para furtar informações e dados (UTFPR, 2020).

Ademais, com o crescente automatismo e a competência dos fraudadores de adaptar com celeridade suas táticas, os modelos tradicionais de detecção de fraudes se mostram insuficientes para lidar com a sua evolução, o que, torna urgente o desenvolvimento de soluções fortes que integrem a Inteligência Artificial (IA) a outras técnicas modernas, para melhorar a qualidade da análise de grandes volumes de dados, na tentativa de identificar padrões com precisão.

A Inteligência Coletiva (IC) é uma potencial alternativa, que se aproveita do conhecimento e da experiência humanos para refinar e melhorar a precisão na detecção de fraudes. A colaboração homem x máquina, que pode ser obtida a partir da utilização de IA e IC concomitantemente, pode representar uma forma eficiente de enfrentar os desafios de fraudes que caracterizam o momento atual.

Figura 1 Comunicado de e-mail fraudulento emitido pela UTFPR

(2020)

1.2 OBJETIVOS DO TRABALHO

O objetivo principal deste trabalho é desenvolver uma plataforma inovadora para a detecção de e-mails fraudulentos, que integre duas abordagens complementares: a Inteligência Coletiva (IC) e a Inteligência Artificial (IA). A combinação dessas metodologias visa a criar um sistema robusto e adaptável que possa identificar e-mails fraudulentos com maior precisão e eficiência, superando as limitações das soluções existentes atualmente.

1.2.1 Objetivos Específicos

- **Implementar Algoritmos de IA**: Desenvolver e treinar modelos de IA usando técnicas avançadas de *machine learning* e processamento de linguagem natural (NLP) para analisar o conteúdo dos e-mails e identificar sinais de fraude. Segundo Goodfellow *et al.* (2016), a aplicação de técnicas de *deep learning* pode melhorar a capacidade de detectar fraudes sofisticadas e personalizadas.

- **Desenvolver um Sistema de Integração para IC**: Criar um módulo que permita aos usuários humanos interagir com a plataforma, fornecendo *feedback* sobre e-mails rotulados como suspeitos pela IA. A integração

de IC e IA permitirá que o sistema aproveite o conhecimento coletivo para aprimorar a precisão dos algoritmos, conforme sugerido por Kamar, Hacker e Horvitz (2012).

- **Avaliar a Performance da Plataforma**: Implementar um mecanismo de avaliação contínua para medir a eficácia da plataforma em detectar e-mails fraudulentos. O desempenho será avaliado com base em métricas como precisão, *recall* e taxa de falsos positivos/negativos, conforme descrito por Liem e Lykourentzou (2020). A análise dos resultados permitirá ajustes e melhorias contínuas no sistema.
- **Garantir Usabilidade e Efetividade**: Desenvolver uma interface amigável e eficiente para que os usuários possam facilmente interagir com a plataforma e fornecer *feedback*. A usabilidade é crucial para garantir que os usuários possam contribuir efetivamente para a detecção de fraudes, o que por sua vez aumenta a qualidade dos dados utilizados para treinar os modelos de IA.

1.2.2 Impacto Esperado:

Ao combinar IC e IA, a plataforma proposta visa não apenas a melhorar a precisão na detecção de e-mails fraudulentos, mas também a oferecer uma solução que possa se adaptar rapidamente às novas táticas de ataque. Existe a expectativa de que a sinergia entre os algoritmos de IA e o conhecimento humano permita o desenvolvimento de uma abordagem mais abrangente e eficaz, proporcionando um nível de segurança digital mais elevado para os usuários e empresas.

1.3 JUSTIFICATIVA

A crescente sofisticação das fraudes cibernéticas, especialmente por meio de e-mails fraudulentos, tem gerado grandes preocupações tanto para empresas quanto para usuários finais. E-mails de *phishing*, que se utilizam de engenharia social, e *spear-phishing*, que além de engenharia social ainda alavanca informações específicas da vítima, representam uma das principais ameaças à segurança digital, com a capacidade de comprometer informações

sensíveis, causar perdas financeiras significativas e danificar a reputação das empresas (HARVARD,2023).

Segundo o Relatório de Investigações de Violação de Dados da Verizon (2023), mais de 90% dos ataques cibernéticos começam com um e-mail fraudulento, o que ressalta a urgência de desenvolver soluções eficazes para mitigar esse risco.

A relevância de uma plataforma que combine Inteligência Coletiva (IC) e Inteligência Artificial (IA) é evidente em um cenário no qual as ameaças estão em constante evolução e se tornam cada vez mais difíceis de detectar por meio de métodos tradicionais. Enquanto os sistemas convencionais de detecção, como filtros de *spam* baseados em regras, conseguem bloquear e-mails maliciosos em grande escala, eles não são suficientemente adaptáveis para lidar com fraudes mais sofisticadas e direcionadas, como observado por Goodfellow, Bengio e Courville (2016).

As soluções baseadas unicamente em IA podem falhar ao não conseguir captar a nuance dos ataques mais personalizados. Assim, a combinação da IC com IA pode representar uma abordagem promissora para superar essas limitações.

Para as empresas, essa plataforma pode oferecer uma forma de proteger seus sistemas e dados sensíveis, minimizando as chances de violações de segurança que podem resultar em danos financeiros e reputacionais.

A integração de IA pode permitir a análise de grandes volumes de e-mails em tempo real, enquanto a contribuição da IC, por meio do *feedback* humano, pode ajudar a identificar novos tipos de ameaças, que não foram previamente registrados.

Malone *et al.* (2010) afirmam que a colaboração entre humanos e máquinas pode aumentar significativamente a capacidade de resolver problemas complexos, como a detecção de fraudes. Esse fator é particularmente importante no contexto empresarial, onde a segurança da informação é crítica para o sucesso e continuidade dos negócios.

Para os usuários finais, a plataforma pode representar uma camada adicional de proteção. Ataques de *phishing* geralmente visam o elo mais fraco da cadeia de segurança – o usuário.

Com a contribuição da IC, os usuários podem fornecer *feedback* sobre e-mails que consideram suspeitos, ajudando a refinar os modelos de IA e, ao mesmo tempo, se tornando parte ativa do processo de proteção.

Essa colaboração não apenas pode aumentar a eficácia do sistema, mas também empoderar os usuários, incentivando uma maior conscientização sobre as ameaças cibernéticas.

A abordagem colaborativa reduz significativamente a chance de falsos negativos, onde e-mails fraudulentos passam despercebidos, e falsos positivos, onde e-mails legítimos são erroneamente rotulados como fraudulentos (Liem & Lykourentzou, 2020).

Além disso, o impacto dessa solução pode ir além da detecção de fraudes em tempo real. Ao incorporar a IC ao processo de detecção, o sistema será capaz de aprender e evoluir continuamente, adaptando-se às novas estratégias de ataque que os fraudadores desenvolvem. Essa adaptabilidade pode ajudar no enfrentamento das ameaças cibernéticas modernas, tornando a plataforma uma ferramenta eficaz tanto no curto quanto no longo prazo.

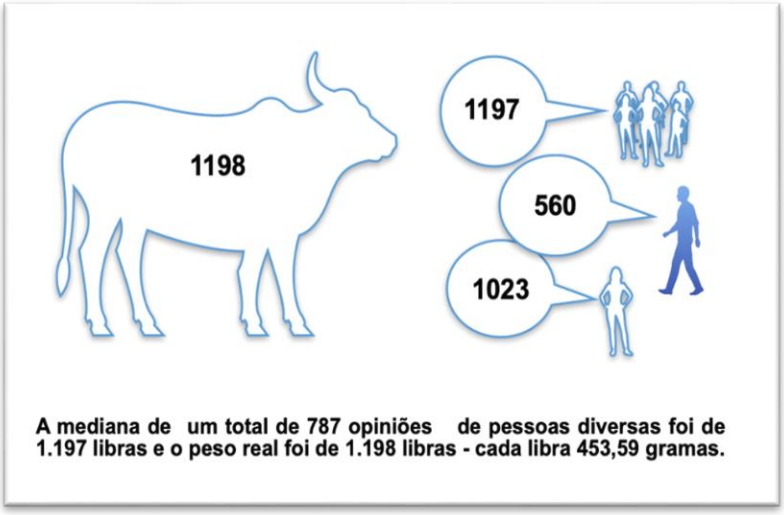
Em resumo, a combinação de IA e IC na detecção de e-mails fraudulentos não apenas poderá oferecer uma solução mais eficiente para os problemas atuais de segurança, mas também pode vir a promover um sistema adaptativo que beneficie tanto empresas quanto usuários finais, garantindo maior proteção e segurança digital.

2 REVISÃO BIBLIOGRÁFICA

2.1 INTELIGÊNCIA COLETIVA

O estatístico e cientista britânico James Galton, em 1906, realizou um experimento interessante. Deixou a sua casa na cidade em Plymouth e dirigiu-se a uma feira agropecuária que fazia uma exposição de animais, no oeste da Inglaterra, com o objetivo de corroborar seus estudos no quesito qualidade dos seres em relação à criação e reprodução (GALTON, 1907). Surowiecki (2005) explica que “a reprodução era importante para Galton porque ele acreditava que apenas poucas pessoas tinham caraterísticas necessárias para manter as sociedades saudáveis” (Surowiecki, 2005, p. 10). Galton já havia feito outros experimentos os quais o deixaram cético quanto a inteligência da pessoa média, concluindo que “a estupidez e a teimosia de muitos homens e mulheres era tão grande que dificilmente mereciam crédito” (Surowiecki, 2005, p. 11). Assim ele montou um experimento ao ar livre, usando o aparato da exposição e uma competição que acontecia no evento. A competição era para saber o peso de um boi gordo depois de preparado (abatido e limpo). As pessoas compravam o bilhete e o preenchiam com o nome e o palpite de quanto seria o peso do bovino. Os palpites mais certeiros receberiam prêmios, o que levou 800 pessoas a participarem da competição. Participaram pessoas especializadas, como açougueiros e agropecuaristas, todavia o maior volume era do povo que não tinha qualquer qualificação técnica para subsidiar o seu palpite, ao menos não individualmente.

Figura 2: Peso real do boi (de Galton) x mediana do coletivo



Fonte: autoria própria (baseado em Surowiecki (2005)

Galton então compilou os dados e ficou surpreendido que a mediana destes, praticamente, acertou o peso do boi, conforme mostrado na Figura 2 (Surowiecki, 2005, p. 12). Graeml ressaltou a honestidade de Galton na divulgação dos resultados da sua pesquisa, uma vez que ela provou que o estatístico estava equivocado em suas preconcepções. Os resultados refutaram as ideias iniciais de Galton e seu entendimento que a democracia não era uma forma adequada de governo (Graeml, 2024, min 15’20”), proporcionando um campo factual e promissor para o presente estudo, que busca explorar a eficácia da Inteligência Coletiva (IC) no apoio à decisão sobre se uma mensagem de e-mail é ou não fraudulenta.

Uma definição interessante da IC foi feita por Lévy (2002, p. 26): "uma inteligência distribuída por toda parte, constantemente valorizada, coordenada em tempo real, e que resulta em uma mobilização efetiva das competências". Para Graeml (2024), definir bem a IC não é fácil, uma vez que, além dos aspectos cognitivos, coletivos, de independência etc., tem-se ainda, os aspectos não tangíveis (Graeml, 2024, min 26"21’). Ele prefere uma definição mais flexível de IC, proposta por Malone (2010), para quem “IC é fazer coisas juntos, que parecem representar um padrão inteligente”. Mais importante do que a inteligência de cada um dos indivíduos é a existência de alguma organização para definir o que está acontecendo (Graeml, 2024, min 26"54’).

Malone (2010) considera que basta parecer inteligente a atuação de um coletivo para que já se possa tratar de inteligência coletiva.

O experimento de Galton (1907) foi um marco inicial para que se entendesse a IC, pois avaliar o peso de um boi não é uma atividade difícil. No entanto, a IC pode ser utilizada em uma ampla gama de questões. Por mais difícil que seja um problema, a complexidade não deve representar um obstáculo.

> A maioria de nós não tem a capacidade – e o desejo – de fazer cálculos sofisticados de custo-benefício. Em vez de insistirmos em encontrar a melhor decisão possível, muitas vezes aceitaremos aquela que parece boa o suficiente. E muitas vezes deixamos que a emoção afete o nosso julgamento. No entanto, apesar de todas estas limitações, quando os nossos julgamentos imperfeitos são agregados da forma correta, a nossa inteligência coletiva é muitas vezes excelente (Surowiecki, 2024, p. 27).

Surowiecki (2005) faz uma ressalva quando diz que a inteligência coletiva pode não florescer se não houver diversidade e independência dos diversos envolvidos. Para ele, quanto mais desacordo houver nas percepções inidividuais (menos consenso e compromisso) melhores serão os resultados obtidos do coletivo (Surowiecki, 2005, p.16-17). Ele ainda tenta justificar a IC dizendo que a média da maioria das coisas é a mediocridade, porém, como se fossemos programados (geneticamente), o ser humano é inteligente coletivamente (Surowiecki, 2005, p. 28).

Inteligência coletiva refere-se à capacidade de grupos de agentes (humanos ou não) trabalharem juntos de maneira eficaz para resolver problemas, tomar decisões ou realizar tarefas complexas. Quando pessoas com diferentes habilidades, conhecimentos e perspectivas colaboram, podem alcançar resultados que vão além das capacidades individuais de cada membro do grupo (Page, 2008).

Essa inteligência pode ser observada em uma variedade de contextos, desde equipes de trabalho e projetos de pesquisa até comunidades *online* e processos democráticos. Ela pode ser facilitada por meio de estruturas organizacionais adequadas, tecnologias de comunicação eficazes e uma cultura que valorize a colaboração e a diversidade de ideias.

Ao entender melhor os princípios por trás da inteligência coletiva, pode-se aproveitar todo o potencial das interações sociais para enfrentar desafios complexos e alcançar objetivos comuns.

A IC em comunidades da internet representa um fenômeno moderno que se manifesta por meio da colaboração e da troca de informações entre os usuários *online (PIÉRRE, 2022)*.

Crowdsourcing é um tipo de inteligência coletiva (Graeml, 2024, 35’09”). O termo *Crowdsourcing* se refere à prática de se conectar a grupos de indivíduos externos buscando que eles ajudem a resolver problemas de forma eficiente (GRAEML, 2019).

Existem várias plataformas *online* de *crowdsourcing* que se utilizam de conceitos de inteligência coletiva, incluindo fóruns de discussão, redes sociais, *wikis* e *sites* de perguntas e respostas. Nestes espaços, os usuários contribuem com suas opiniões, experiências, *insights* e recursos, formando um vasto *pool* de conhecimento compartilhado.

Um dos exemplos mais notáveis de inteligência coletiva na internet é a Wikipedia, uma enciclopédia *online* colaborativa, em que milhares de voluntários de todo o mundo contribuem para criar e editar artigos em uma ampla gama de tópicos. Apesar de não haver um controle centralizado sobre o conteúdo, a Wikipedia conseguiu se tornar uma fonte de informação confiável para milhões de pessoas em todo o mundo, com um sistema sofisticado de autogovernança (Forte *et al.*, 2014).

Nesse contexto, é gigante o desafio de criar mecanismos (aplicativos ou plataformas) que provejam incentivos para que haja um engajamento de indivíduos que contribuam com sua inteligência individual para se obter um resultado a partir do esforço coletivo que seja melhor. De acordo com Malone (*apud* Graeml, 2024, 49’48”), para o pleno triunfo da IC é essencial ser capaz de motivar as pessoas a participar da multidão.

Portanto, a IC pode ser utilizada para identificar padrões que escapam aos sistemas automatizados na detecção de fraudes em e-mail. Por meio da participação do ser humano, é possível obter resultados que não podem ser capturados por algoritmos.

Em plataformas de detecção de e-mails perniciosos, a IC permite pode ser utilizada para que os usuários marquem e-mails suspeitos, fornecendo *feedback* valioso para melhorar os sistemas de filtragem. Conforme apontado por Malone *et al*. (2010), a IC pode compensar as limitações do processamento informatizado, principalmente em situações que demandem julgamento subjetivo e interpretação do contexto.

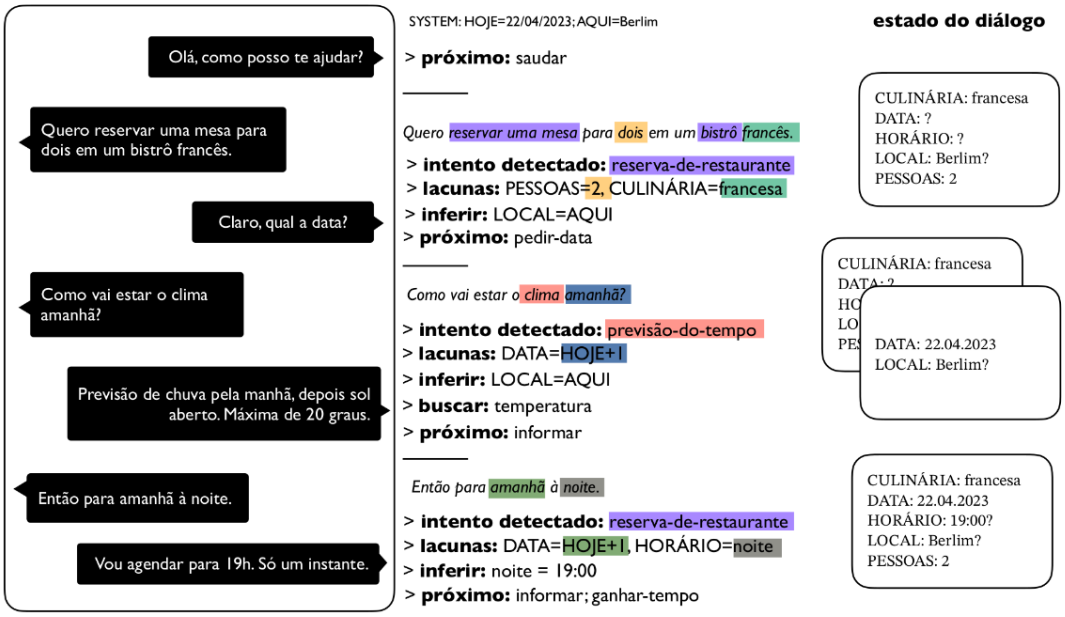
2.2 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) refere-se à criação de sistemas computacionais capazes de realizar tarefas que normalmente requerem inteligência humana, como reconhecimento de padrões, aprendizado e tomada de decisão.

No contexto da detecção de fraudes, a IA, especialmente por meio de técnicas de *machine learning* e *deep learning*, tem se mostrado eficaz na análise de grandes volumes de dados e na identificação de anomalias que indicam possíveis fraudes.

Algoritmos de IA são particularmente úteis porque conseguem "aprender" com grandes conjuntos de dados e identificar padrões que não são facilmente perceptíveis por seres humanos. De acordo com Goodfellow, Bengio e Courville (2016), a capacidade da IA de detectar fraudes de forma equiparável a um agente humano está intimamente ligada ao uso de modelos supervisionados, que podem ser treinados com base em e-mails classificados previamente como fraudulentos ou legítimos, e aos avanços em Processamento de Linguagem Natural (NLP), que permitem a análise semântica dos conteúdos dos e-mails. Isso faz com que a IA seja extremamente eficiente na detecção de fraudes em tempo real, mesmo em grandes volumes de e-mails. Na figura 2 temos um diagrama representativo do processo de análise semântica de linguagem natural como é feito por um LLM

Figura 2 Ilustração exemplificando os processos envolvidos no processamento de linguagem natural



Fonte: Caseli e Nunes (2024).

O desenvolvimento de diversos modelos como Google Gemini, GooseAI, Claude etc., disponibilizados para uso geral por meio de APIs públicas, como produtos (*NLP-as-a-service*) possibilita o uso de processamento de linguagens sem o *overhead* de treinamento e provisionamento dos modelos.

3 METODOLOGIA

3.1 DESENVOLVIMENTO DA PLATAFORMA

O desenvolvimento da plataforma teve como foco a integração de IC e IA para otimizar a detecção de e-mails fraudulentos. A IA foi empregada para analisar grandes volumes de dados e identificar padrões de comportamento associados a fraudes.

De acordo com Goodfellow, Bengio e Courville (2016), a IA pode processar e aprender com vastos conjuntos de dados para detectar padrões que os métodos tradicionais podem não capturar. A implementação de algoritmos de *machine learning* e *deep learning* permitiu à plataforma realizar a triagem inicial dos e-mails com alta eficiência.

A IC foi utilizada para complementar a IA ao incorporar o julgamento humano na revisão e validação dos resultados gerados pelos algoritmos.

Os usuários da plataforma foram envolvidos na rotulagem de e-mails suspeitos e no fornecimento de *feedback*, a ser utilizado para refinar os modelos de IA e melhorar continuamente a detecção de fraudes.

3.2 AMOSTRA DE FRAUDES

Como fonte de dados referentes a e-mails fraudulentos, foi utilizado um *dataset* público de 17538 e-mails reais rotulados como fraudulentos ou não.

O *dataset* dispunha de duas colunas: [Email text], que continha todo o conteúdo do e-mail, incluindo metadados (Remetente, destinatátio, etc...) e [Email type], que tinha um de dois valores possíveis: *Safe email* para e-mails seguros e *Fraudulent email* para e-mails fraudulentos. A segunda coluna foi usada como rótulo.

Tabela 1 Exemplo de e-mail e rótulo extraídos do *dataset*

Email text	*Email type*
“*Skip> I'll try a checkout into a new directory...Which didn't help. At the very least I think that means it's time for Bed...*	Safe Email

Skip”	

Tendo em vista que os usuários da plataforma seriam lusófonos, os emails do *dataset* foram traduzidos para o português, procurando-se manter a fidelidade ao conteúdo textual original de cada e-mail.

Os dados foram processados em duas etapas:

I. Quantização dos rótulos
II. Tradução do texto para o português

De todo o *corpus* de e-mails foram selecionados 8 e-mails.

3.2.1 Quantização dos rótulos

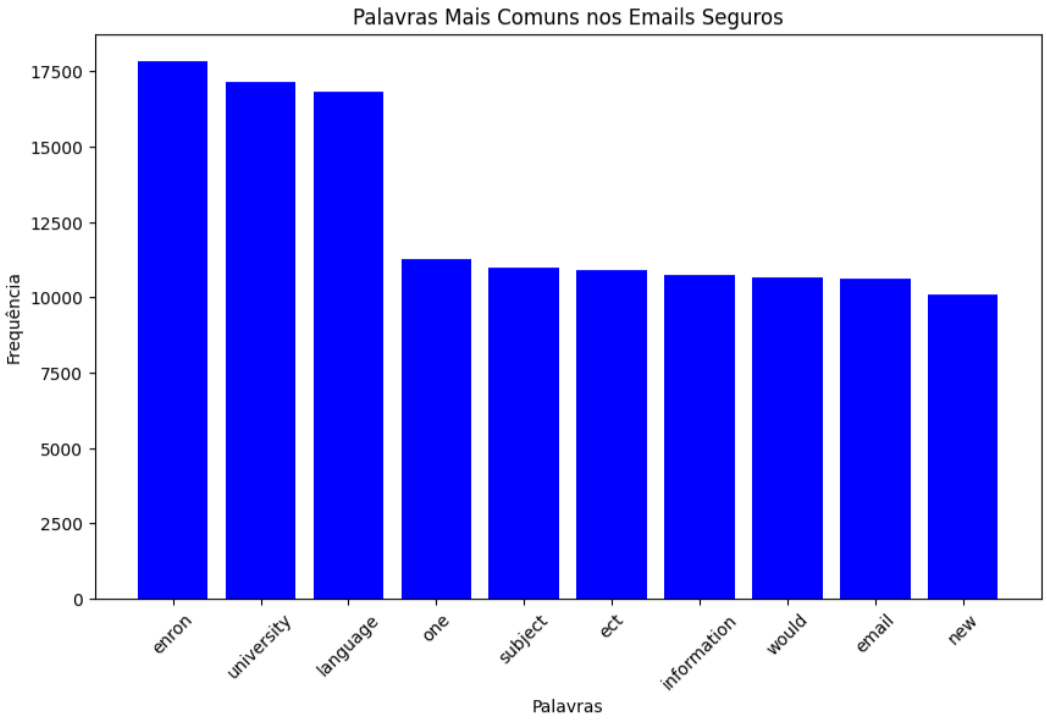
Os rótulos foram categorizados com base na sua *probabilidade de serem fraudulentos*. O valor obtido foi definido entre 0 e 1, sendo 0 equivalente a “Definitivamente seguro” e 1 equivalente a “Definitivamente fraudulento”.

Valores mais próximos de 0.5 denotavam e-mails de maior ambiguidade, ou seja, mais difíceis de se definir como falsos ou verdadeiros.

O trabalho de quantização foi feito com o auxílio de métodos de processamento textual, criando pesos de probabilidade de acordo com as palavras e suas frequências associadas.

Podemos ver, por exemplo, este gráfico de frequência de palavras encontrado no conjunto de dados dos e-mails seguros, denotando diferença perceptível de conteúdo e palavras-chave (ver a Figura 3).

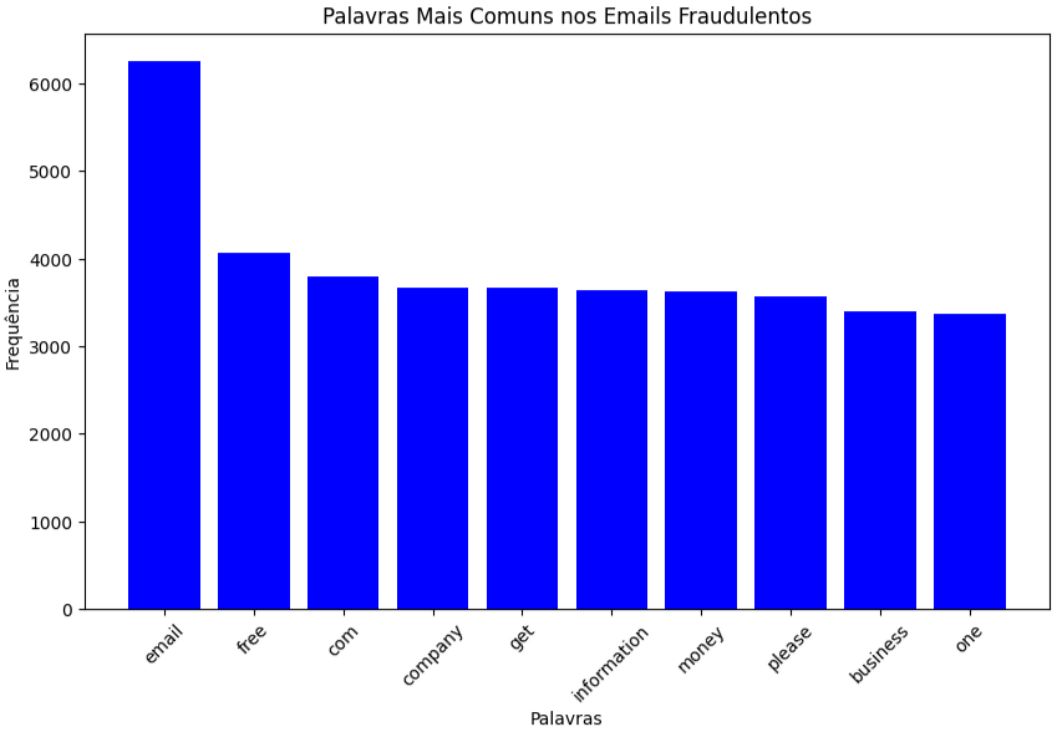
Figura 3 Gráfico de palavras mais comuns nos e-mails seguros extraídos do *dataset*



Fonte: Elaboração própria

A Figura 4 apresenta o mesmo gráfico, mas gerado para os e-mails rotulados como fraudulentos:

Figura 4 Gráfico de palavras mais comuns nos e-mails fraudulentos extraídos do *dataset*



Fonte: Elaboração própria

A diferença é notável, não só pelo teor dos e-mails fraudulentos, que abordam geralmente questões financeiras, tática comum de *phishing* (Kim et al., 2017), mas também pelo fato de que os e-mails seguros são majoritariamente universitários ou corporativos. Por meio dessa associação conseguiu-se treinar um modelo perceptron capaz de classificar corretamente e quantizar os e-mails com alta precisão.

3.2.2 Tradução do texto para o português

A tradução do conteúdo textual foi feito com o auxílio do modelo de linguagem GPT-4o. O resultado foi revisado e recebeu ajustes mínimos para manter fidelidade ao *significado* dos e-mails originais, enquanto ainda mantendo a autonomia do modelo de IA como ferramenta de tradução.

Os e-mails escolhidos, assim como sua tradução, são apresentados no apêndice A.

Os impactos e resultados referentes à tradução serão analisados posteriormente.

3.3 OBTENÇÃO DE DADOS DA INTELIGÊNCIA COLETIVA

Os dados de usuários foram coletados de forma a se assemelhar a um “teste de personalidade”. Neste modelo foi prometido aos usuários participantes que, depois de realizaram a categorização dos emails, eles receberiam uma análise do “tipo de personalidade”, com base na análise realizada dos e-mails. O intuito de se oferecer essa análise foi aumentar o engajamento dos usuários, ao entregar um serviço de autoconhecimento como recompensa pelo auxílio na tarefa.

Como se pretendia trabalhar com agentes de duas naturezas distintas (Humanos e IA), eles foram separados em duas classes de coleta de dados:

1. Agentes humanos (*Crowd Agents*), para os quais foi oferecido o incentivo motivacional (análise de personalidade) para obter o engajamento.

2. Agentes autônomos (*Auto Agents*), que não precisaram de estratégia de motivação, uma vez que foram codificados para atender a solicitações humanas.

3.3.1 Agentes Humanos (*Crowd Agents)*

Os efeitos da contribuição humana coletiva e seus resultados são muito destacados na literatura. O sucesso em obter a plenitude da IC está na diversidade, independência e até, na disparidade (desacordo) dos indivíduos do grupo (Surowiecki, 2005, p. 16). Por isso, em um primeiro momento foi necessária uma estratégia para obter estes indivíduos e ter qualidade nos dados do grupo.

Especificamente, buscou-se explorar a “sabedoria das massas” emergente em diversos grupos, o que foi feito por meio de um formulário breve, que permitia aos agentes preencherem os seguintes campos:

a. A estimativa de chance de cada e-mail ser fraudulento (em uma escala de 0 a 100, sendo 0 indicativo de o e-mail ser certamente seguro e 100 certamente fraudulento).

b. Comentários sobre cada e-mail que pudessem elucidar o motivo da decisão. Esta informação era de preenchimento opcional.

c. Dados pessoais do participante: Nome, Endereço de e-mail, Sexo e Idade.

d. Duração de preenchimento do formulário em segundos.

Estes dados juntos foram utilizados para categorizar os usuários e, posteriormente, analisar sua performance como indivíduos e grupos.

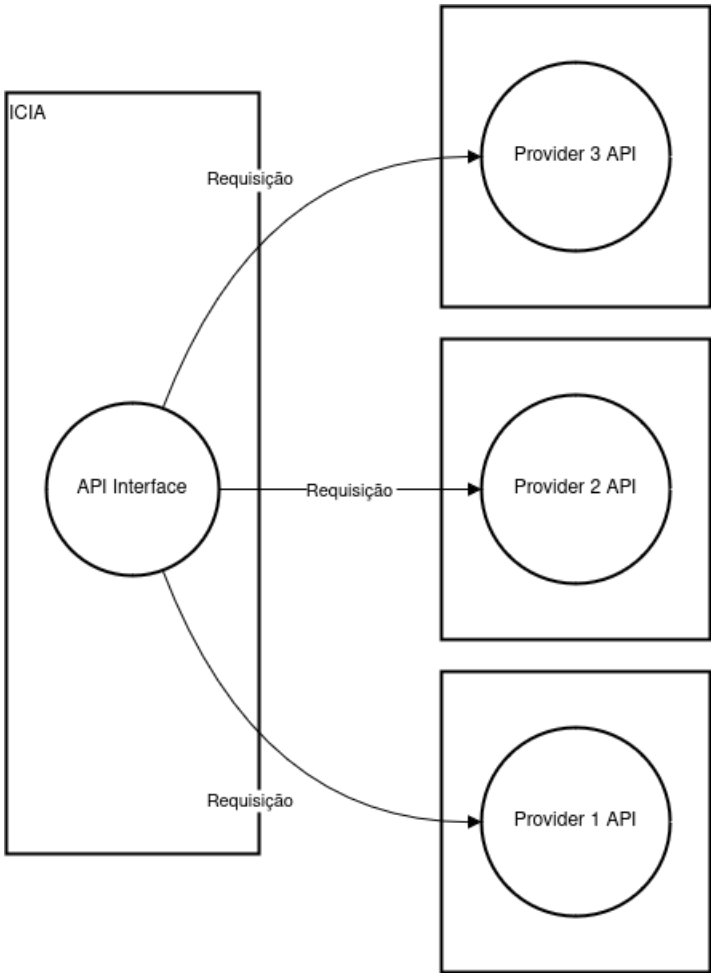
A plataforma com o formulário foi hospedada publicamente e o *link* compartilhado por meio de diversos canais digitais com o intuito de maximizar a diversidade dos participantes.

3.3.2 Agentes autônomos (*Auto Agents*)

Um agente autônomo é cada instância discreta de coleta de dados de uma API de IA. O entendimento de agente individual é tênue, pois da maneira que são distribuídos os modelos de IA por suas respectivas APIs, não se pode distinguir um “Indíviduo de IA”. Ao invés disso, é mais correto se referir a estes como “sistemas de IA” (OECD, 2024).

Dessa forma, ao desconstruir a oferta de IA a partir de suas APIs como sistemas *caixa-preta,* pode-se discretizar cada indivíduo como uma série de entradas igualmente configuradas sobre o questionário.

Figura 5 Diagrama demonstrando a interface genérica de APIs



Fonte: Elaboração própria

Das configurações relevantes, foram utilizados três níveis de configurações hierárquicas citados abaixo: Provedor, Modelo e Conjunto de entrada.

1. Provedor de API
    a. Modelo de IA
        i. Conjunto de entrada (*Inputs)*
        ii. Conjunto de entrada (*Inputs)*
    b. Modelo de IA
        i. Conjunto de entrada (*Inputs)*

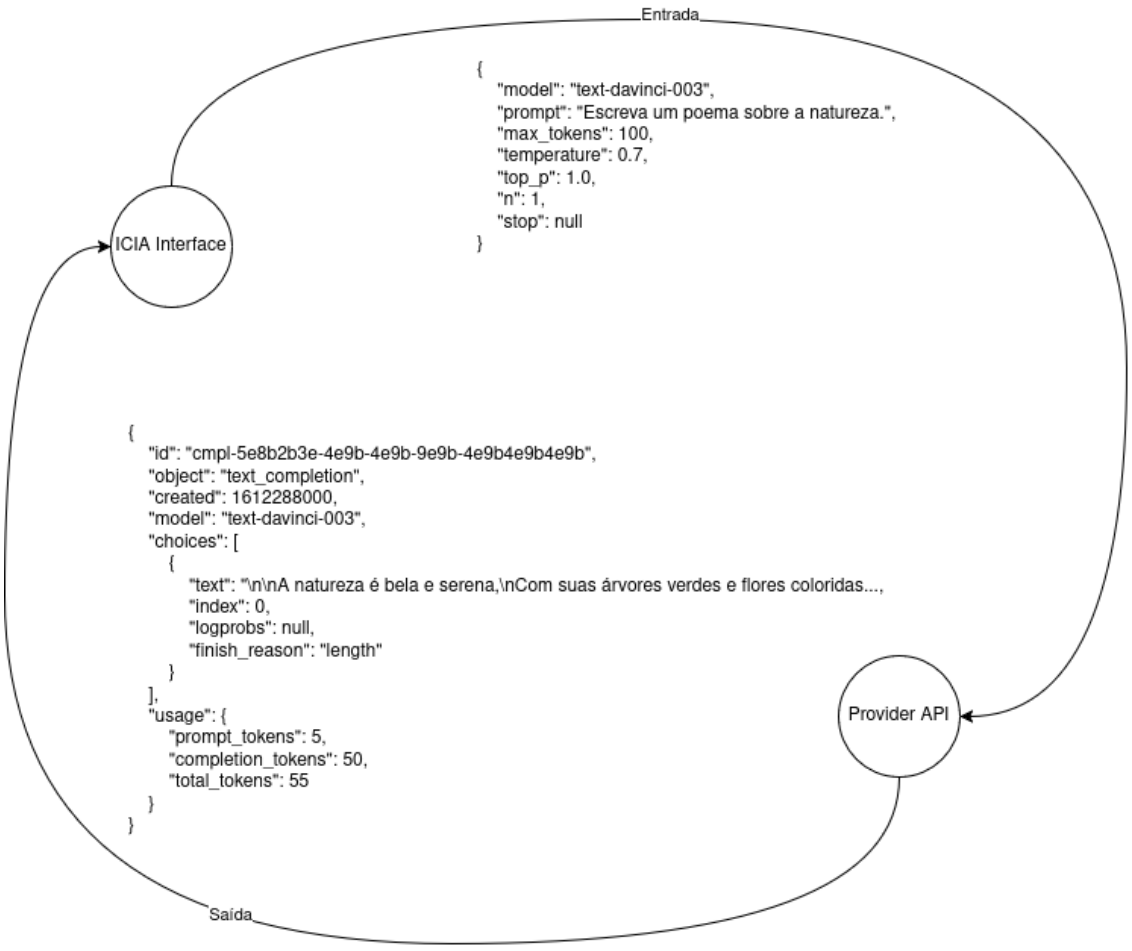
Dessa forma foi possível criar diversos indivíduos com características diferentes, advindo de um único modelo, apenas alterando os *inputs*.

Essa tripla qualificação da diversidade das respostas autônomas implica que se esperava obter uma variância das respostas muito maior do que nas respostas de agentes humanos, ainda que dispondo-se de controle total dos perfis individuais, permitindo a análise exploratória de diferentes *Auto Agents*.

A Figura 6 demonstra como é feita a comunicação entre indivíduos de IA e os provedores de IA.

Idealmente, desejava-se que um *Auto Agent* conseguisse ser abstraído e tratado exatamente como se tratasse de um agente humano, de forma que, na etapa da fusão das respostas com as dos agentes humanos, não fosse necessário considerar outros fatores extrínsecos ao *Auto Agent*.

Figura 6 Diagrama demonstrando a comunicação de indivíduos ICIA e APIs de provedores de IA



Fonte: Elaboração própria

3.4 ANÁLISE E CLASSIFICAÇÃO DOS AGENTES

Para cada agente há um conjunto único de dados referentes às respostas numéricas de probabilidade de fraude para cada pergunta.

Consolidando estes resultados, foi possível obter dois resultados importantes.

I. Indicador de performance (*Score*)
II. Posição no *cluster* de personalidades

3.4.1 Cálculo do indicador de performance (*Score*)

O indicador de performance tem o intuito de ser uma categorização geral do quão próximo o agente foi de ser um classificador ideal.

Considerando a regressão: respostas do agente X resposta ideal, o *score* representa a soma dos valores de correlação R multiplicada por 100.

Sendo assim, para N classificações, um agente podia ter um intervalo possível de *score* entre:

$$0 – 100*N$$

3.4.2 Cálculo da posição no *cluster* de personalidades

Com o intuito de promover *feedback* aos agentes humanos e melhorar a compreensão da interpretação das tendências pessoais de cada agente, foi criado um mapa de sete personalidades:

1. Super preditor
    a. Muito acima da média, sem erros e pouco desvio do preditor ideal.
2. Bom preditor
    a. Preditor moderado, bastantes acertos e pouco desvio do ideal.
3. Preditor paranoico
    a. Preditor com maior sensibilidade a fraudes, bastantes acertos mas desvio do ideal.

4. Preditor ingênuo
    a. Preditor com baixa sensibilidade a fraudes, bastantes acertos mas desvio do ideal.
5. Preditor ruim
    a. Preditor com muitos erros, grande desvio do ideal
6. Preditor aleatório
    a. Preditor cuja análise segue padrão aleatório. Pouca certeza e grande desvio do ideal
7. Preditor consistentemente ruim
    a. Preditor cuja análise parece ser completamente oposta ao ideal. Desvio máximo do ideal

É importante salientar que o intuito desta classificação foi a análise do indivíduo.

3.5 FUSÃO DOS DADOS COLETADOS E ANÁLISE DA SABEDORIA DAS MASSAS

A fusão dos resultados se deu de forma distribuída e imparcial. Ou seja, foram considerados todos os indivíduos e não houve ponderação dos votos.

Os demais dados coletados serviram apenas para fins de estudo sobre a diversidade, não sendo utilizados de forma alguma para a análise da performance individual.

Como as respostas da avaliação são todas quantitativas e moduladas entre 0 e 1, pode-se facilmente fundir as respostas por meio da média das respostas de todos os agentes.

$$A = \frac{(a_1 + a_2 + \cdots + a_n)}{n}$$

3.5.1 Abstrações de indivíduos e grupos

Um ponto importante na análise da inteligência coletiva é a compreensão da independência das partições. Em suma, quando se utiliza o método das médias, o resultado da computação final não é afetado pela ordem das partições do grupo inteiro.

Isso permite usar grupos inteiros como abstrações para indivíduos. Por exemplo, resumindo a análise de todos os Agentes Autônomos de uma mesma identidade como um único indivíduo.

3.6 CROWDSOURCING

Vista a necessidade do emprego efetivo de sabedoria das massas para a resolução do problema em mão, foram adotados métodos consagrados de *crowdsourcing*.

A tarefa proposta consistiu, essencialmente, de uma microtarefa de jogo casual, utilizando as seguintes propriedades para aumentar o engajamento (Vianna *et al*., 2022):

1. Gamificação da tarefa:
    a. Sistema de pontuação baseada na performance
    b. Sistema de tabela de pontuação global
2. Democratização da tarefa
    a. Minimizar o tempo de interação com a tarefa, simplificando os comandos e objetivos
    b. Redução da carga de trabalho do indivíduo
3. *Feedback* imediato
    a. Computação da pontuação de cada usuário e retorno de sua personalidade de fraude assim que concluída a participação.

A adoção de técnicas de *crowdsourcing* foi motivada pela expectativa de aumento da adesão dos agentes humanos e consequente aumento da sua diversidade.

Além disso, pela natureza de *crowdsourcing*, procurou-se entregar o teste com um caráter “viral”, a fim de maximizar o compartilhamento e engajamento dos participantes.

Para esse fim, foram criadas campanhas nas seguintes redes sociais:

1. Reddit
2. Tik-Tok
3. LinkedIn
4. Whatsapp

Além destas campanhas, também foi promovido o compartilhamento entre usuários usando-se as APIs de compartilhamento das seguintes redes:

1. Twitter (atual ‘X’)
2. Facebook
3. Whatsapp
4. E-mail

Desta forma, todo usuário podia compartilhar seu resultado e convidar novos contribuintes para participar da avaliação dos e-mails potencialmente fraudulentos na plataforma, alavancando ainda mais a viralização da plataforma.

3.7 COLETA DE DADOS

O desafio posto envolvia engajar um número suficiente de pessoas para verificar se e-mails eram fraudulentos ou não, a partir de um esforço de *crowdsourcing* envolvendo agentes humanos, a princípio. Apenas em um estágio posterior buscou-se envolver agentes não humanos.

Para isso, foi criada uma aplicação com botões de seleção de 1 a 10, onde 1 representava "não fraudulento" e 10 "muito fraudulento". Para incentivar a participação, o sistema foi projetado para oferecer *feedback* aos participantes sobre o quão bons preditores eles eram, classificando-os em grupos com base em suas respostas, oferecendo-lhes em contrapartida o resultado de um teste de perfil pessoal. Assim:

- ***Super* (Super preditor)** - Participante que acerta confiantemente todas as classificações ou a grande maioria;
- ***Good* (Bom)** - Agente que acerta em torno de 70% dos casos, *Recall* e precisão equiparáveis;
- ***Naive* (Ingênuo)** - Assim chamado por ter baixa sensibilidade à fraude. Baixo *Recall* mas boa precisão;
- ***Paranoid* (Paranoico)-** Assim chamado por ter alta sensibilidade à fraude. Alto *Recall* mas precisão mediana;

- ***Bad* (Ruim)** – Agente que acerta em torno de 30% das classificações. Baixo F1;
- ***Random* (Aleatório) –** Agente com comportamente ambíguo e baixa confiança. Baixa performance associada à falta de motivação na análise;
- ***Evil* (Maligno)** - Agente com comportamento destrutivo, alta confiança e próximo ao 0% de acerto. Essencialmente o oposto de um super preditor.

Essas classificações foram criadas com o intuito de motivar as pessoas a participarem. É claro que ninguém quer ser visto como ingênuo ou inversor de valores, o que traz um aspecto de jogo para a participação e pode motivar um participante a desafiar seus amigos, para ver se eles conseguem resultado melhor.

3.8 USO DA IA PARA SELEÇÃO DE FRAUDES NOS EMAILS

Usando o Perceptron, não conseguimos uma verificação completa dos e-mails. Isso significa que aspectos como endereços e a contextualização (como o momento político, econômico, cultural etc.) não são considerados. Outra limitação do Perceptron está relacionada à semântica das palavras, incluindo verbos, adjetivos, e outros tipos gramaticais, o que impede a divisão em temas e a análise etimológica das palavras (ou "stemização").

Pelo fato de utilizar um dicionário, o Perceptron tem dificuldade em distinguir palavras que utilizam gírias ou abreviações, como "p4g4r" em vez de "pagar". Também enfrenta dificuldades com erros de ortografia, o que impede a correta diferenciação de um e-mail fraudulento. Esta dificuldade no treinamento e necessidade constante de aprimoramento das técnicas de ML são limitações inerentes ao método de classificação de e-mails.

Utilizando agentes humanos e LLMs, porém, pode-se alavancar a capacidade natural destes agentes de interpretar textos ricamente, entendendo nuances e utilizando da avançada intuição inerente aos seres humanos.

3.8.1 Análise dos Participantes

Nesta etapa, o foco é alavancar o potencial da inteligência coletiva e da inteligência mista (IA e inteligência humana). Uma observação importante é que, embora o conhecimento individual seja limitado, os viéses individuais podem

neutralizar uns aos outros e a média de um grupo pode fornecer uma decisão bastante balizada.

Se uma pessoa não consegue identificar um e-mail fraudulento, outra pode conseguir. A inteligência coletiva deve ser independente, com as pessoas colaborando para agregar resultados, contribuindo de forma individual para um resultado coletivo melhor.

O objetivo não era construir um conhecimento sobre o que constitui uma fraude, mas sim responder rapidamente, com base no senso comum, se uma determinada situação envolve fraude ou não.

Depois de se ter os preditores agrupados e compilados, os mesmos testes foram realizados com a inteligência artificial.

Depois de agrupar e classificar os participantes humanos, foi realizada uma avaliação de se eles eram mais paranoicos, indecisos, errôneos ou ingênuos. Em seguida, o coletivo dos agentes de IA foi inserido, procurando-se avaliar se, juntos, humanos e IA produziam melhores resultados.

4 DESENVOLVIMENTO

Para o trabalho de coleta de dados de agentes humanos, coleta de dados dos agentes autônomos e análise dos dados, foi projetada uma plataforma digital customizada com diversas ferramentas próprias para este intuito.

4.1 PLATAFORMA ICIA

A plataforma ICIA (Inteligência coletiva/Inteligência artificial) foi desenhada com o propósito de dispor de uma camada de compatibilidade entre toda a coleta de dados entre *crowd agents* e *auto agentes* (ver a Figura 7).

Figura 7 Diagrama dos componentes para a aplicação ICIA

Fonte: Elaboração própria

A disposição dos componentes permitiu dar ênfase às seguintes funcionalidades:

1. *Website* e *Landing Page*
2. Sistema de enquetes
3. Sistema de criação de agentes autônomos
4. Escalabilidade e segurança

4.1.1 *Website* ICIA

Com o intuito de criar uma área atrativa para *crowdsourcing,* foi criado um *website* de fácil navegação e que trazia diversas informações sobre o projeto a fim de incentivar pessoas a contribuírem com respostas de qualidade.

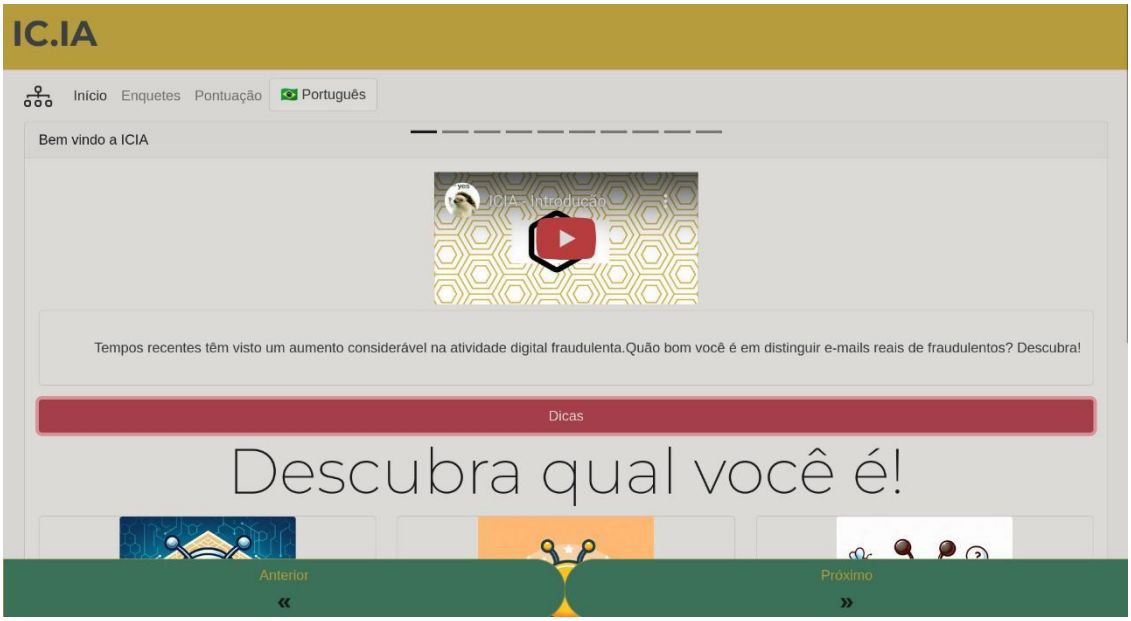
Projetar um *website* de enquetes, que utiliza cores e formas para atrair a atenção do usuário e minimizar frustrações, exige a aplicação de princípios de *design* centrados no usuário. Abaixo, os principais princípios a serem seguidos:

4.1.1.1 Uso estratégico das cores

Na disciplina de experiência do usuário (UX), entendemos a paleta de cor de um website como um elemento central da apreciação e compreensão do artefato pelo usuário, influenciando na cooperação e atraindo mais usuários (HASAN et. Al., 2024)

- **Psicologia das cores:** Paleta de cores reflete o propósito do *site*, além do tema de “abelhas”, animais conhecidos por sua inteligência coletiva sofisticada.
    - **Verde:** Transmite confiança e calma.
    - **Laranja e amarelo:** Passam energia e descontração.
    - **Vermelho:** Chama a atenção.
- **Hierarquia visual:** Uso de cores contrastantes para destacar elementos importantes, como botões de envio ou chamadas para ação. A figura 8 mostra o emprego desta técnica na página de apresentação da enquete.

Figura 8 Tela de apresentação da enquete ICIA



Fonte: Elaboração própria

- **Consistência:** A paleta de cores é baseada em cores naturais de abelhas, trazendo o conceito de inteligência coletiva dos animais como metáfora visual consistente para evitar confusão e reforçar a identidade visual.
- **Acessibilidade:** Contraste entre texto e fundo para atender às diretrizes de acessibilidade (WCAG).

4.1.1.2 Identidade visual

A identidade visual abrange além dos componentes também as emoções, garantindo que se alinhem com os objetivos de negócios (HASAN et. Al., 2024)

- **Botões e ícones:** Uso de formas arredondadas para transmitir acessibilidade e segurança. Botões devem ser grandes o suficiente para cliques precisos, especialmente em dispositivos móveis.

Agrupamento visual: Caixas são utilizadas para organizar visualmente perguntas, opções e seções, evitando sobrecarga cognitiva.

- **Efeito de foco:** Destaque elementos interativos ao passar o cursor ou clicar, como mudanças sutis de cor ou sombreamento.
- **Interface limpa:** Distrações são minimizadas, evitando excesso de texto ou elementos gráficos. Foco no conteúdo principal, que são as enquetes.

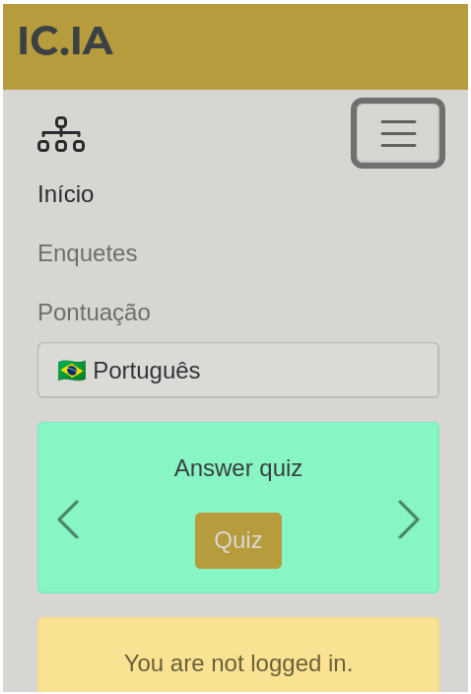
- **Direcionamento claro:** Usuário é guiado através de microscopias claras, como “Escolha uma opção” ou “Próximo passo”.
- ***Feedback* visual:**
    - O progresso da enquete é informado por meio de uma barra de progresso.
    - Ações são confirmadas, como “Resposta enviada com sucesso!”.

4.1.1.3 Responsividade e adaptabilidade

Responsividade é uma característica chave para que uma aplicação seja confortável de usar e prenda a atenção do usuário até o fim.

- **Layout responsivo:** O *site* deve se ajustar a diferentes tamanhos de tela, desde *smartphones* até *desktops*. A figura 9 demonstra como a aplicação se conforma ao tamanho de tela de um dispositivo móvel.

Figura 9 Demonstração da responsividade da interface em aparelhos móveis



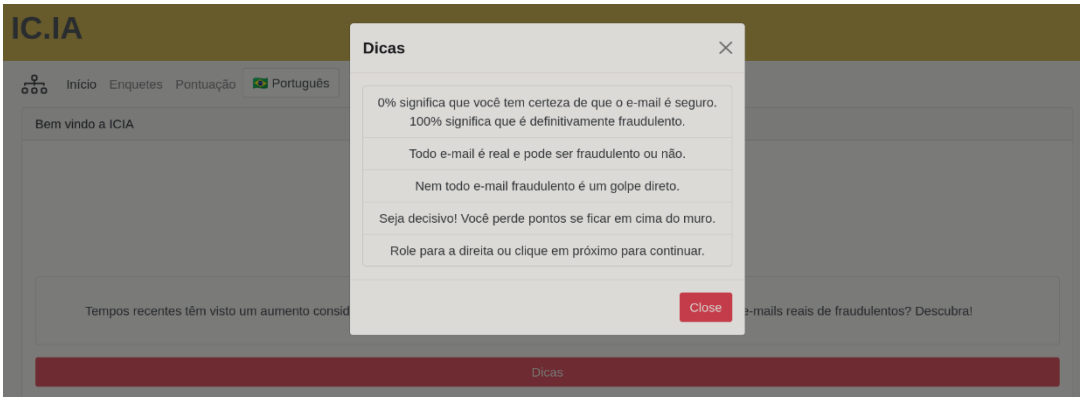
- **Toques otimizados:** Em dispositivos móveis, elementos clicáveis são espaçados adequadamente para evitar cliques acidentais.

4.1.1.4 Gamificação e interatividade

Mecanismos de gamificação permitem que o usuário se sinta mais motivado para começar e completar as tarefas apresentadas.

- **Instruções:** As instruções são detalhadas e as regras intuitivas, favorecendo a resposta rápida do usuário e minimizando frustrações.

Figura 10 Página de instruções e dicas para o preenchimento da enquete



Fonte: Elaboração própria

- **Elementos engajantes:** Os tipos de personalidades são incorporados por ilustrações de abelhinhas, engajando o trabalho de autoidentificação dos usuários.
- **Recompensas visuais:** A página de respostas é deliberadamente construída para satisfazer o usuário e promover o reenvio da enquete para colegas, maximizando o potencial viral. A figura 11 demonstra como as cores e imagens servem para maximizar emoções positivos

Figura 11 Diagrama demonstrando o fluxo de atividades do usuário

Fonte: Elaboração própria

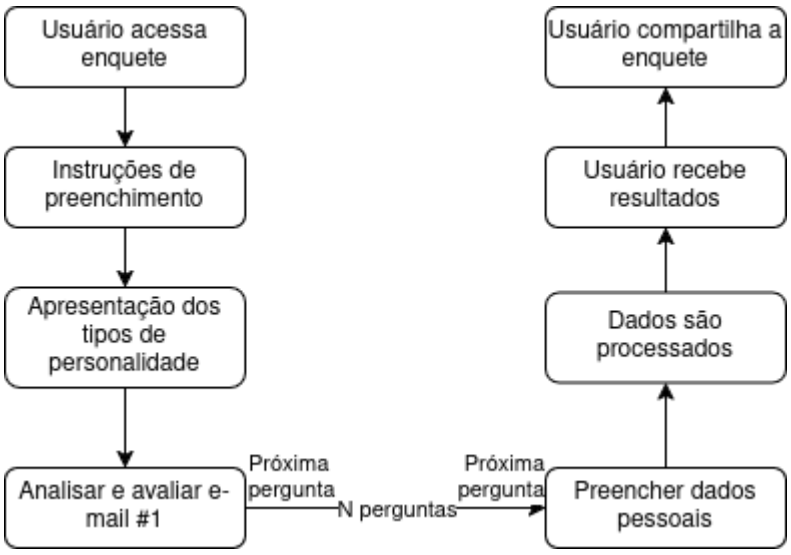
4.1.2 Sistema de enquetes

A fim de criar e editar as enquetes, foi criado um sistema modular de enquetes e grupos de enquetes. Cada grupo de enquetes tinha o propósito de agrupar um número qualquer de pares Perguntas x Respostas e fazer a apresentação para o usuário.

O sistema de enquetes serviu de componente para receber as respostas de todos os agentes, tanto humanos quanto autônomos.

O módulo basilar envolveu um sistema de formulário que agrega todas as perguntas do conjunto em uma única página interativa. A figura 12 demonstra as tarefas envolvidas por parte do usuário no preenchimento da enquete.

Figura 12 Diagrama demonstrando o fluxo de atividades do usuário



Fonte: Elaboração própria

Em essência, procurou-se criar uma experiência orgânica que preenchesse os três requisitos de um *crowdsourcing* efetivo:

1. Existência de incentivo para sua conclusão (testar seu próprio conhecimento)
2. Simplicidade e rapidez no preenchimento do formulário
3. Exibição de *feedback* instantâneo, após a conclusão do preenchimento

A enquete foi desenhada para ser completada em pouco tempo e permitir que o usuário tivesse conhecimento em toda etapa da porção necessária para completá-la.

4.1.3 Sistema de configuração de agentes autônomos

Para a interação com os agentes autônomos foi criado um fluxo próprio, com a intenção de garantir a estabilidade, segurança e precisão de todas as chamadas aos provedores de IA.

O fluxo de interação com os provedores foi dividido em duas etapas:

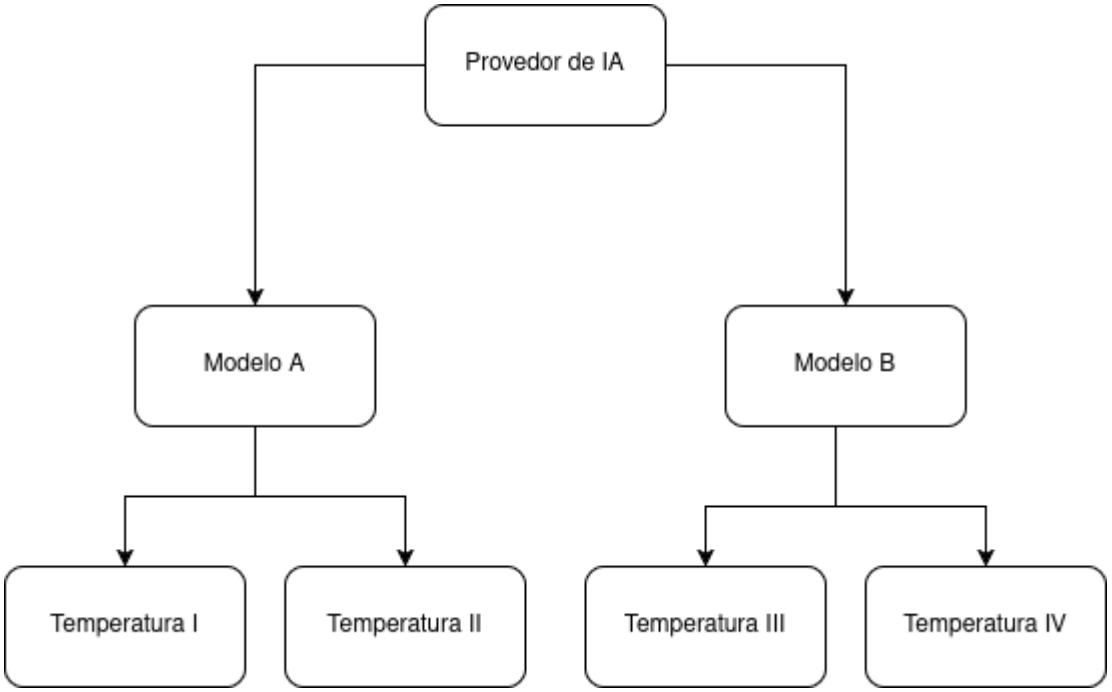
1. Configuração e testagem;
2. Criação dos agentes.

A etapa de configuração e testagem foi feita de forma supervisionada, iniciando-se com a configuração de um provedor de IA e pelo menos dois conjuntos diferentes de parâmetros de chamada.

A primeira configuração envolvia os detalhes de conexão ao provedor:

- *Endpoint* da API disponibilizada pelo provedor de IA
- *Token* de API fornecido pelo provedor de IA

Figura 13 Relação hierárquica entre as configurações



Fonte: elaboração própria

Como exemplo, o diagrama da Figura 13 permite 4 diferentes configurações de agentes.

Cada soma de temperatura e modelo resulta em um perfil de indivíduo de forma análoga às peculiaridades dos agentes humanos. Assim, foi possível criar, de forma artificial, um conceito de “Diversidade” para um mesmo modelo de IA.

Todas as chamadas realizadas utilizaram o modelo OpenAI, feito por meio de um objeto do tipo *json*[1] com formato padronizado entre diferentes modelos (OPENAI, 2024):

As configurações ideais foram atingidas por meio do processo de teste, onde eram realizadas chamadas de um determinado indivíduo com um *prompt* de teste. O sucesso do teste era determinado por dois fatores:

1. O agente autônomo respondia apenas um valor numérico entre 0 e 100 para que este valor fosse categorizado de forma adequada.
2. O agente autônomo precisava demonstrar um viés claro e não responder um valor aleatório para uma determinada pergunta.

[1] Acrônimo para *Javascript Object Notation,* uma formatação em texto de um objeto.

O resultado da execução dos testes e seu significado para cada um dos agentes será discutido mais adiante.

A etapa da criação de agentes autônomos ocorreu após a devida configuração e testagem. Tendo encontrado a combinação ideal para cada indivíduo, foi possível criar uma quantidade determinada de agentes autônomos para responder a determinada pergunta e gravar seu resultado ao lado dos outros agentes.

A opção de criar qualquer número de agentes, de forma automática, permitiu que os efeitos da sabedoria das massas surjissem quase que imediatamente, convergindo para a melhor resposta esperada do grupo.

4.1.4 Escalabilidade e segurança.

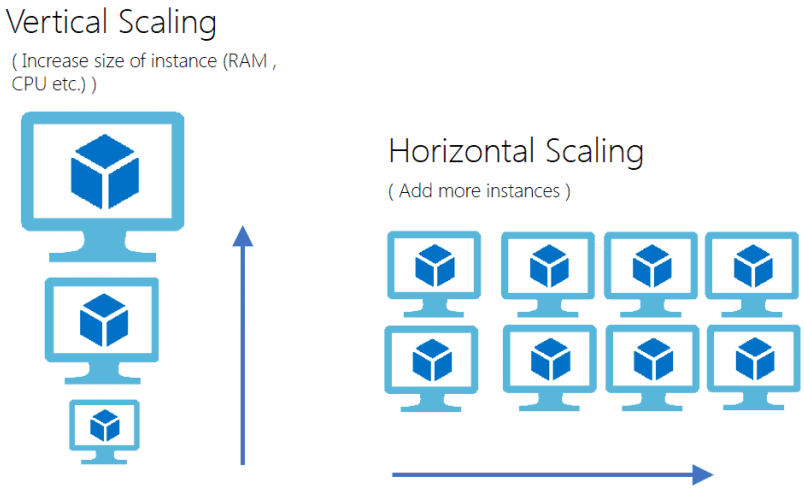
Com o intuito de habilitar o uso da plataforma como um produto condizente com sua natureza de alto tráfego de usuários e dados sensíveis, foram utilizadas ferramentas e práticas condizentes com as recomendações de engenharia.

A escalabilidade da aplicação abrangeu tanto sua capacidade de concorrência e modularidade, permitindo que pudesse ter sua usabilidade estendida para um número progressivamente maior de usuários e provedores de recursos.

Para maximizar o potencial de concorrência, a aplicação foi hospedada em uma *cloud* pública, em um recurso de *pool* de *containers* conectados a uma instância elástica de banco de dados. Essa configuração permitiria, em tese, o processamento de até 10.000 usuários simultâneos, várias vezes maior que o valor máximo esperado em qualquer momento.

Com o intuito de minimizar custos, o *pool* foi configurado para horizontalizar, ou seja, aumentar automaticamente o seu número de instâncias, de 1 até 10 instâncias. A figura 14 demonstra o processo de *scaling* realizado pelas instâncias da aplicação

Figura 14 Diagrama exemplificando o trabalho de *scaling* realizado pelo provedor *cloud*



Fonte: Microsoft Azure

Para o trabalho de modularização, a aplicação foi separada em módulos responsáveis individualmente por um domínio dentro do sistema:

1. Aplicação base (navegação e comunicação com o cliente)
2. Aplicação de autenticação
3. Aplicação de sessões de usuário
4. Aplicação de enquetes
5. Aplicação de agentes autônomos

Com essa organização se tornou possível reutilizar componentes e expandir a funcionalidade do sistema com maior facilidade.

4.2 A ENQUETE

Com o intuito de explorar as capacidades dos grupos, foram selecionadas 8 perguntas dentre as processadas do *dataset*. Estas perguntas estão apresentadas no apêndice A.

Todos os agentes receberam as exatas mesmas perguntas na mesma ordem e formatação.

4.3 ATIVAÇÃO DE AGENTES

O trabalho de aquisição de agentes se tornou fundamental não só no âmbito prático de se necessitar de uma massa considerável de indivíduos para se poder dispor da sabedoria das massas, mas também porque a natureza dos agentes iria impactar, de uma forma ou de outra, o resultado final. Tanto os agentes humanos quanto os agentes autônomos foram escolhidos de forma a apresentarem uma quantidade satisfatória de divergência e vieses.

4.3.1 Ativação dos agentes autônomos

A seleção e exploração dos agentes autônomos foi um trabalho consideravelmente mais simples, pois existe uma quantidade reduzida no mercado atualmente de provedores de IA que atendem aos requisitos esperados:

1. APIs com suporte ao modelo OpenAI de requisições
2. Diferentes modelos de IA

Desta forma, a seleção dos provedores de IA foi a seguinte:

1. OpenAI
2. Google Gemini
3. Maritaca IA

4.3.2 Comparativos entre os provedores selecionados

A tabela 2 oferece uma visão geral das características mais importantes de cada modelo.

1. OpenAI

   A empresa OpenAI foi a pioneira dentre as provedoras de serviços de LLM conversacional. Inventora dos modelos GPT-X (*Generative pre-trained transformer*). Oferece uma gama de diferentes modelos e produtos.

2. Gemini

   Gemini é o modelo multimodal de IA da empresa Google, capaz de compreender e gerar conteúdo multimodal (vídeos, imagens, voz e texto). Oferece um grande leque de diferentes produtos.

3. Maritaca

   Empresa Brasileira que oferece dois modelos de IA conversacional treinados especificamente em conteúdos em português e nativos do Brasil.

Tabela 2 Comparativo dos indivíduos autônomos

Provedor	Modelo	Preço	Saída	Câmbio
OpenAI	GPT-4o	U$2,50	U$2,50	R$15,46
OpenAI	GPT-4o-mini	U$0,15	U$0,60	R$0,93
Gemini	Gemini-flash	U$0.0375	U$0.15	R$0,23
Maritaca	Sabiá-3	R$5,00	R$10,00	R$5,00
Maritaca	Sabiazinho-3	R$1,00	R$3,00	R$1,00

Fonte: Elaboração própria

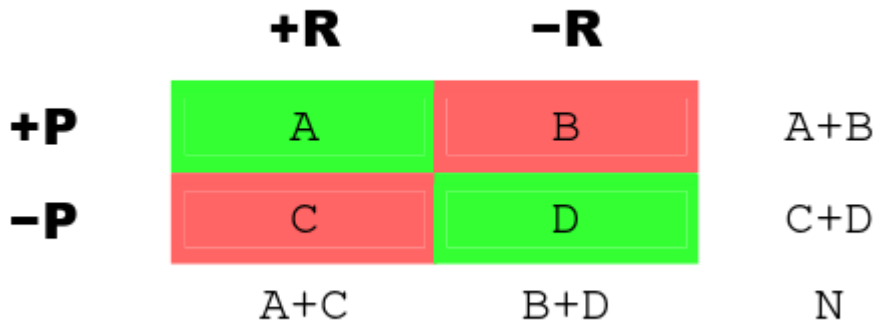
5 RESULTADOS

A análise da performance dos indivíduos foi feita com o auxílio de métodos e ferramentas analíticas consagradas, de modo a se encontrar os fatores mais importantes na construção de uma comunidade inteligente, capaz de resolver problemas de fraude.

5.1 OS PARÂMETROS DE ANÁLISE

O trabalho de interpretar a performance dos agentes como classificadores pede o uso de métodos consagrados na literatura para modelos de ML (Powers, 2010;Tharwat, 2020).

Figura 15 Exemplo de matriz de confusão



Fonte: Powers (2010, p. 2)

Da matriz apresentada na figura 15 obtém-se quatro quadrantes codificados em cores. Os quadrantes verdes representam as classificações acertadas e os vermelhos representam classificações errôneas, sendo:

A. *True positives* (Verdadeiros positivos)
B. *False positives* (Falsos positivos)
C. *False negatives* (Falsos negativos)
D. *True negatives* (Verdadeiros negativos)

Deste universo pode-se equacionar os valores dos parâmetros de Precisão e *Recall* da seguinte forma:

$$\text{Recall} = \frac{A}{(A+C)}$$

$$\text{Precisão} = \frac{A}{(A+B)}$$

Ambos os parâmetros representam proporções, sendo 0 seu mínimo valor e 1.0 seu maior valor possível. Por exemplo, um valor de *Recall* de 0.9 implica que o modelo é capaz de identificar com sucesso 90% das fraudes presentes na enquete de e-mail. Enquanto um valor de precisão de 0.9 implica que, de todos os modelos determinados como fraude, 90% eram de fato fraudulentos.

A interpretação destes valores, intuitivamente, proporciona uma ideia dos vieses dos agentes, pois um modelo com alto *Recall* implica em uma maior sensibilidade para detectar fraudes, enquanto uma alta precisão implica confiança em discernir e-mails seguros. Para concatenar esses parâmetros em uma única métrica, utiliza-se o F1 (Ou F-Score), que é a média harmônica entre o parâmetro de *Recall* e Precisão.

Desta forma pode-se usar o F1 como parâmetro principal de performance para os agentes.

5.2 ANÁLISE DO PRÉ-PROCESSAMENTO DO *DATASET*

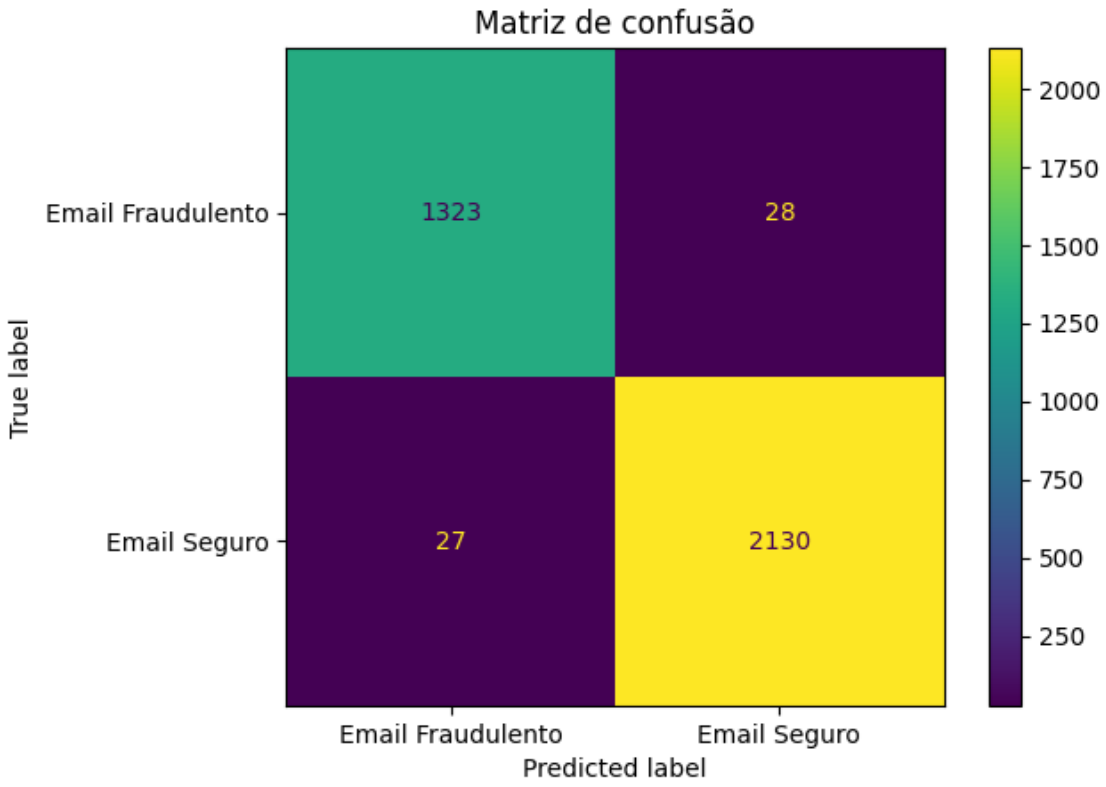
O *dataset* foi processado utilizado um algoritmo de ML conhecido como Perceptron multi-camadas classificador (*MLP Classifier*) e implementado pela biblioteca scikit-learn (Scikit-Learn, 2025). O *dataset* foi dividido entre 80% para treino e 20% para testagem. A Tabela 3 apresenta o resultado do classificador para o *dataset* de teste (n=3508)

Tabela 3 Resultados do modelo MLP para o *dataset* analisado

	Precisão	*Recall*	*F1-Score*	# amostras
Fraude	0.98	0.98	0.98	1351
Seguro	0.99	0.99	0.99	2157

Fonte: Elaboração próprio

Figura 16 Matriz de confusão para o modelo MLP



Fonte: Elaboração própria

A análise da matriz de confusão (ver a Figura 16) permite observar que houve um número bastante reduzido de erros. Observa-se uma tendência de ingenuidade (menor sensibilidade a fraudes), pois têm-se a proporção quase duas vezes maior de falsos negativos que de falsos positivos (0.21% contra 0.12%). Apesar disso, o classificador ainda se demonstra excelente, de modo geral, obtendo uma pontuação F1 de 0.98.

O resultado do classificador se demonstra satisfatório, porém ainda se mostra vulnerável a confusão. Sendo assim, procurou-se justamente os e-mails mais difíceis de serem classificados, dentre estes:

1. E-mail rotulados de forma errônea (Falsos positivos e falsos negativos);
2. E-mails rotulados com baixo intervalo de confiança.

Estes e-mails, junto com seu rótulo e intervalo de confiança, são descritos no apêndice A.

5.3 ANÁLISE DOS AGENTES

A proporção de agentes (contados por indivíduos) foi bastante desproporcional entre os agentes humanos e autônomos. Este efeito era esperado, considerando a velocidade com que se utiliza de agentes autônomos e sua escalabilidade elevada em comparação aos agentes humanos (ver a Figura 17).

Figura 17 Proporção entre agentes autônomos e humanos

Fonte: Elaboração própria

5.3.1 Análise dos componentes de personalidade

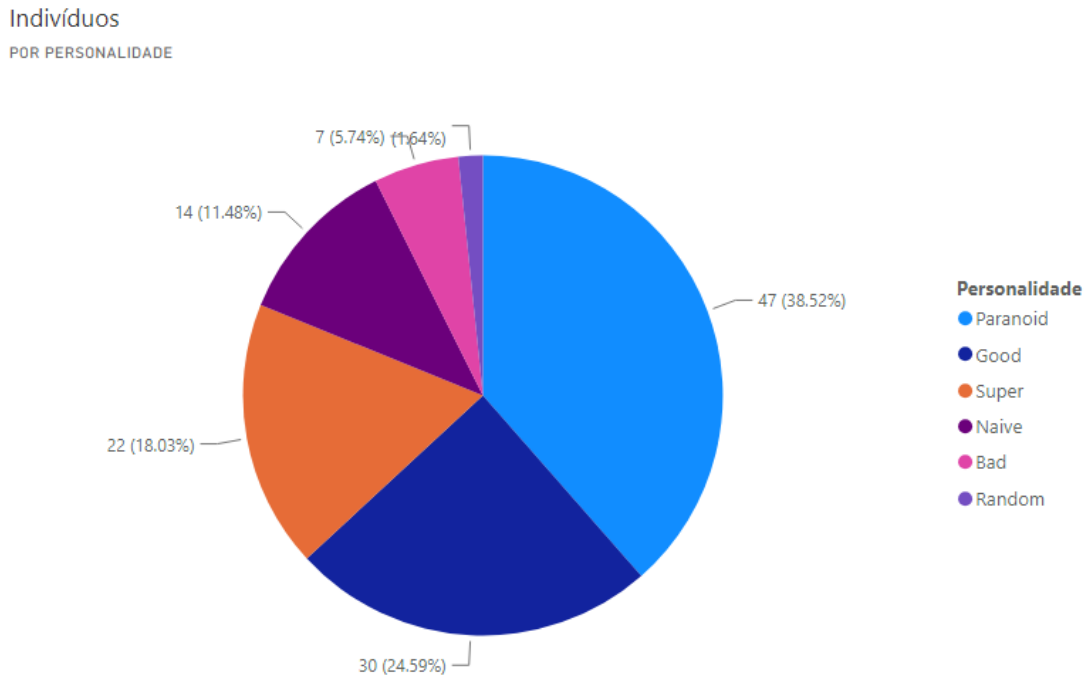
Através dos componentes de personalidade calculados usando o método de *p-clusters*, pôde-se vislumbrar as tendências dos indivíduos e populações. A personalidade de cada indivíduo foi determinada através do processo de distância mínima a cada cluster de personalidade, sendo assim, cada indivíduo foi classificado com apenas uma personalidade principal.

O conceito de personalidade para um agente autônomo deve ser abstraído, assim como foi feito, anteriormente, no trabalho de construção de indivíduos autônomos, pois não há uma definição exata de indivíduo no âmbito de modelos de IA. Ainda assim, encontram-se padrões emergentes que se assemelham a traços de personalidade na interação com LLMs. Sabe-se que estes modelos têm grande potencial de mimetização do comportamento humano (Takao *et al*., 2024).

5.3.1.1 Personalidades dos agentes humanos

A análise de personalidade para agentes humanos permite que entendamos os padrões de comportamento do contribuinte médio.

Figura 18 Proporção de personalidades entre os agentes humanos



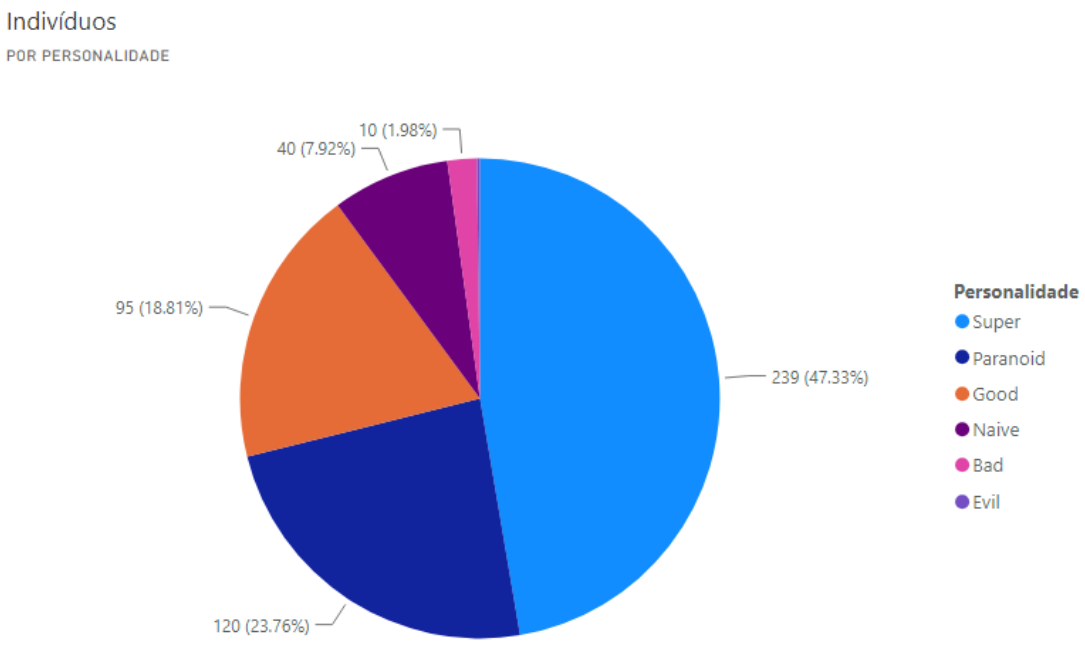
Fonte: Elaboração própria

Conforme mostra a Figura 18, a proporção de paranoicos parece elevada em relação a pesquisas similares utilizando agentes humanos como classificadores de fraude, como a realizada por Schwabe e Almendra (2009), tendo sido constatada uma proporção maior de falsos negativos (baixo *Recall*). Pode-se notar também que o número de agentes classificados como *random* foi baixa, indicando que houve esforço genuíno pela maioria dos participantes em classificar os e-mails corretamente. Os agentes classificados como *bad* ou *evil* também foram pouco frequentes, evidenciando que agentes humanos possuem discernimento para realizar avaliações normalmente boas.

5.3.2 Personalidade dos agentes autônomos

Similar à análise realizada com indivíduos humanos, nos esforçamos para que o processo de cálculo das distâncias de personalidade e leitura sejam uniformes entre os dois tipos de agentes a fim de compreender quão as análises podem ser traduzidas para um contexto de agente autônomo.

Figura 19 Proporção de personalidade entre os agentes autônomos



Fonte: Elaboração própria

A Figura 19 mostra um número muito mais significativo de *super predictors* entre os agentes autônomos, indicando um potencial analítico superior destes com relação aos humanos.

O alto grau de paranoia entre os agentes autônomos parece equiparável ao encontrado no grupo humano (23.76% x 24.59%), o que sugere que os modelos de IA possam herdar comportamentos e emoções humanas, como aponta a pesquisa elaborada por Serapio-García et al. (2023).

5.3.3 Respostas qualitativas

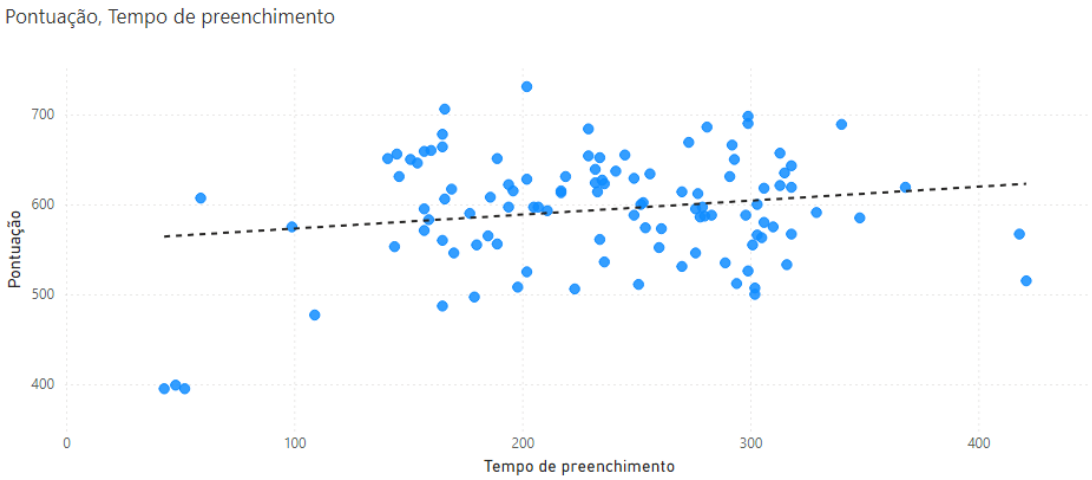
Dada a possibilidade de os agentes humanos terem justificado suas estimativas de forma discursiva, obteve-se alguma informação adicional para ajudar a compreender a maneira como os agentes analisam cada questão. Essas respostas, que não foram muitas (apenas 3 respondentes fizeram uso dessa possiblidade de prestar informações adicionais), apresentam-se listadas no apêndice B.

5.3.4 Desempenho dos agentes humanos

A coleta dos dados ocorreu por meio de *website* público no período entre os meses de setembro e outubro de 2024. Obteve-se a participação de 123 voluntários.

Apesar da capacidade preditiva da inteligência coletiva de agentes humanos ser comprovada em diversos cenários, o experimento de campo demonstrou resultados igualmente satisfatórios nesta pesquisa.

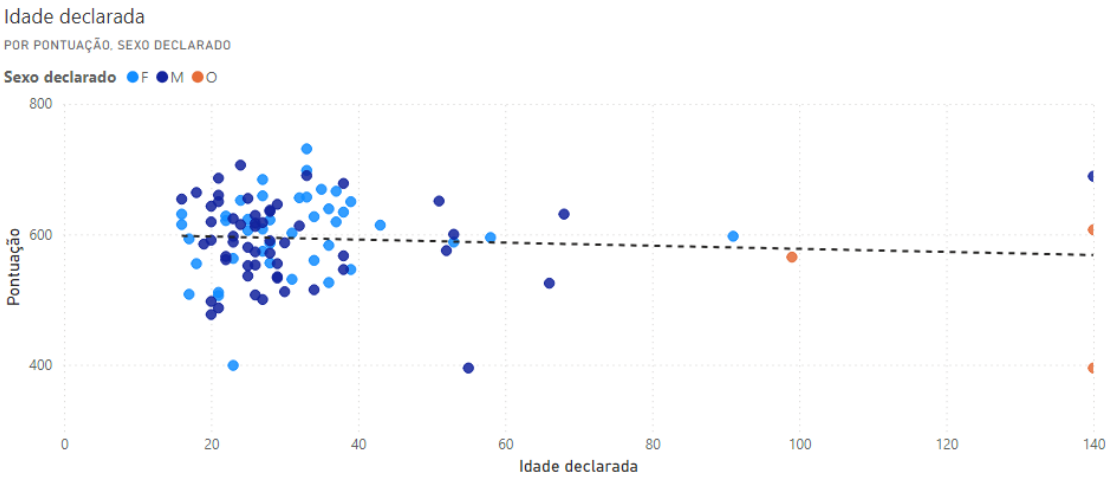
Figura 20 Distribuição de pontos por tempo de preenchimento n = 123



Fonte: Elaboração própria

Pode-se perceber, a partir da análise da Figura 20 que há uma leve correlação entre o tempo necessário para responder as enquetes e a pontuação obtida (r=1.2), indicando que quem dedicou mais tempo obteve um resultado melhor do que aqueles que se desvenciliaram da tarefa de ler e avaliar os e-mails mais rapidamente.

Figura 21 Distribuição de pontos por idade declarada



Fonte: Elaboração Própria

A Figura 21 mostra que, embora tenha se tentado obter uma amostra diversa, em função de gênero e idade, ela acabou incluindo muito mais jovens. A falta de representatividade dos mais idosos impede que se faça qualquer inferência sobre eventuais diferenças na capacidade preditiva de agentes humanos em função da idade. Visualmente, parece não haver diferença nessa capacidade em função do gênero.

A tabela a seguir apresenta os resultados de desempenho para o agregado de agentes humanos.

Tabela 4 Médias comparativas do resultado dos agentes humanos

Sexo	Recall médio	Precisão média	F1 médio	Pontuação média	σ	Maior pontuação individual
Masculino	0.82	0.63	0.70	581.27	67.09	706
Feminino	0.82	0.64	0.70	600.90	56.68	731
Outro	0.78	0.71	0.73	498.13	98.09	607
Grupo	0.82	0.64	0.70	583.52	70.19	731

Fonte: Elaboração própria

Para as respostas ponderadas de cada grupo (a soma de todas as respostas dividida pela quantidade de indivíduos), conseguimos os dados a seguir que demonstram a efetividade da sabedoria das massas.

Tabela 5 Aglomerados dos agentes humanos

Sexo	Recall médio	Precisão média	F1 médio	Pontuação coletiva
Masculino	1.00	1.00	1.00	765.28
Feminino	1.00	1.00	1.00	769.22
Outro	1.00	0.78	0.87	751.73

Grupo	1.00	1.00	1.00	786.24

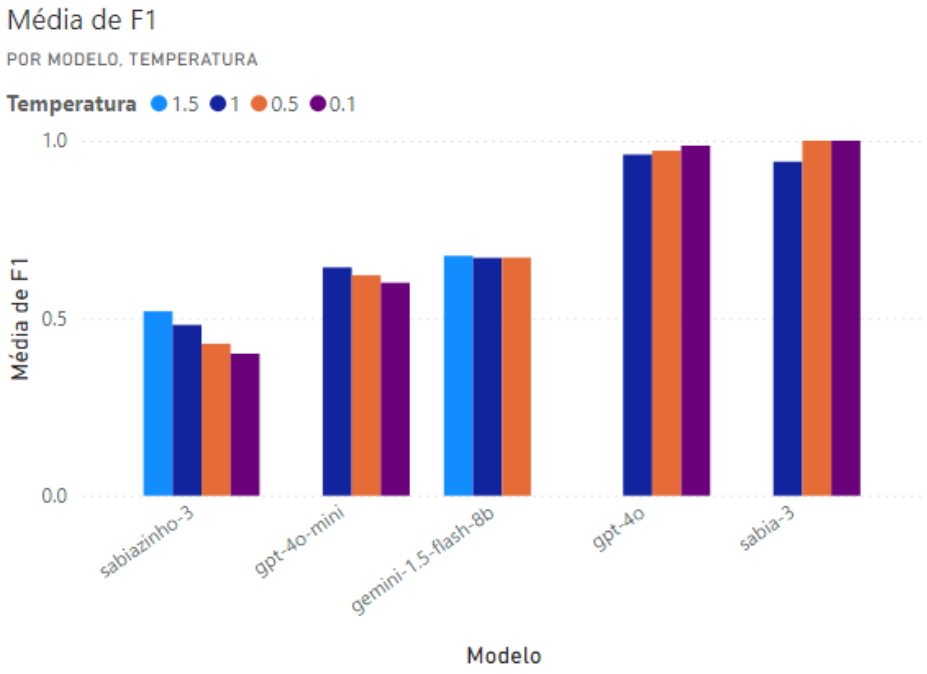
Fonte: Elaboração própria

Percebe-se, a partir da análise da Tabela 5, que há pouca diferença na performance dos coletivos masculino e feminino, indicando que o gênero não impacta na performance dos coletivos.

5.3.4 Desempenho dos agentes autônomos

A divisão “demográfica” dos indivíduos autônomos se fez pelos grupos a seguir.

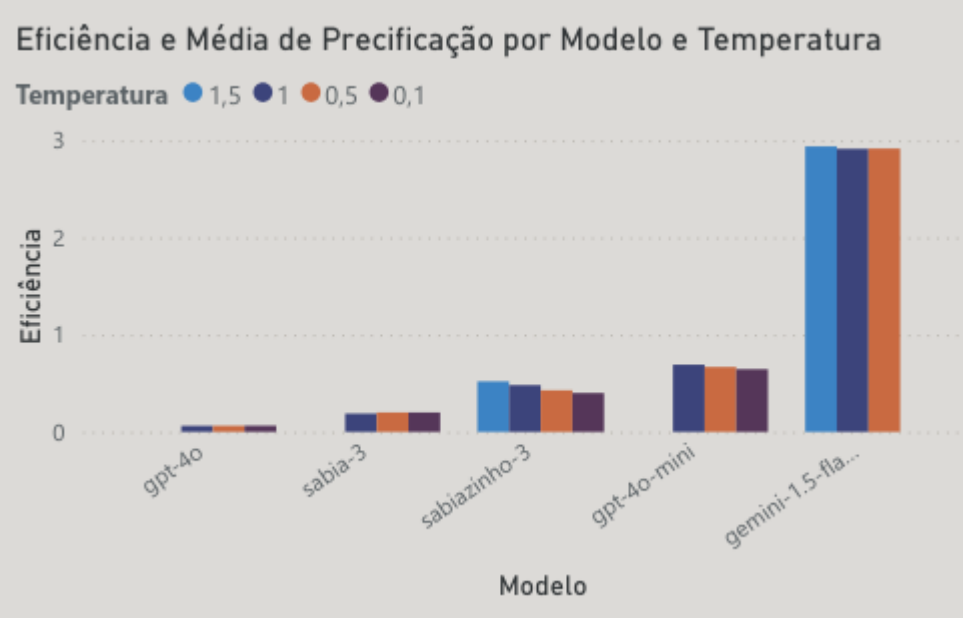
Figura 22 Relação entre pontuação F1 por cada conjunto de indivíduos autônomos



Fonte: Elaboração própria

A escolha das temperaturas foi feita de modo a explorar o comportamento do modelo específico mediante as configurações disponíveis. Observou-se que os modelos gpt-4o, sabia-3 e gpt-4o-mini não tiveram comportamento estável (pontuação equivalente a zero) com temperatura 1.5, assim como o modelo gemini-1.5-flash-8b com temperatura 0.1

Figura 23 Eficiência de preço para cada indivíduo autônomo

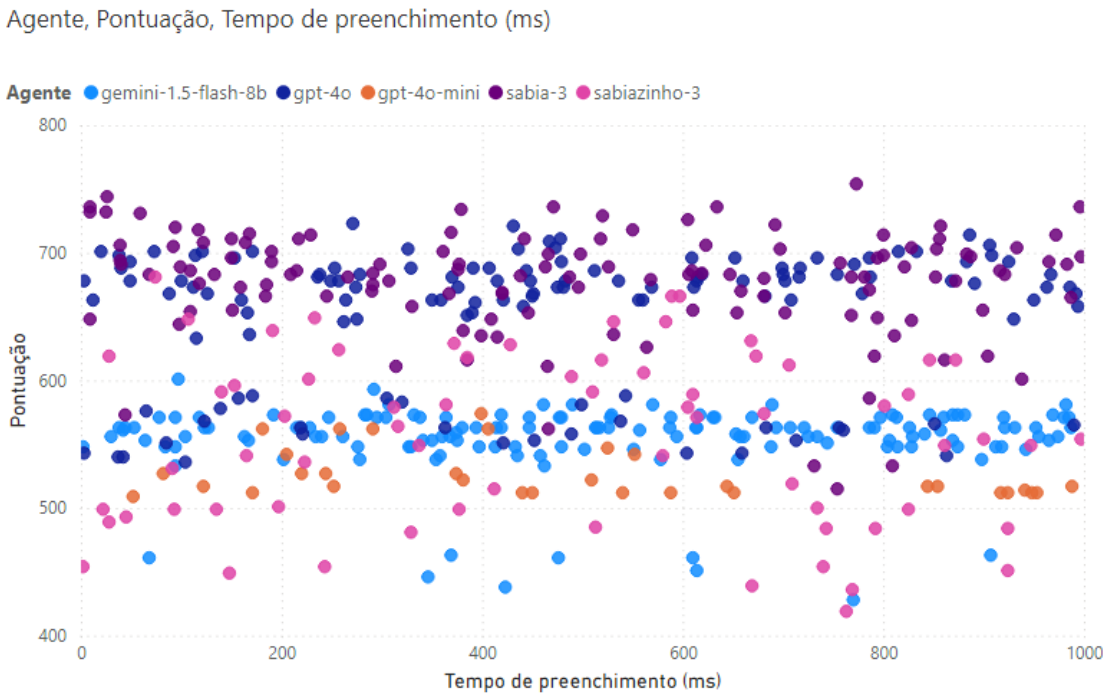


Fonte: Elaboração própria

A eficiência é um cálculo simples entre o valor F1 médio e o preço em reais por milhão de *tokens* necessário para o uso do modelo.

Para compreendermos a relação entre os indivíduos, tempo de preenchimento e pontuação, o gráfico da figura 24 foi elaborado.

Figura 24 *Scatter* de todos os indivíduos autônomos por modelo e tempo de preenchimento



Fonte: Elaboração própria

A Figura 24 permite ver claramente que não há correlação entre o tempo e a performance obtida para as análises realizadas por agentes autônomos, considerando que a maior parte do tempo é composto por latência de rede, que não impacta na qualidade da análise.

A tabela a seguir mostra os dados de desempenho para cada indivíduo, agregado por modelo.

Tabela 6 Médias comparativas dos resultados dos agentes autônomos

Modelo	Recall	Precisão	f1	Pontuação	σ	Maior pontuação individual
Flash	1.00	0.50	0.67	553.9	28.74	601
Mini	1.00	0.45	0.62	526,56	18.83	574
Sabiazin	0.46	0.53	0.48	556.25	66,96	681
Sabia-3	1.00	1.00	1.00	678.05	41,96	754
GPT-4o	0.95	1.00	0.97	652.88	52.72	723
Grupo	0.90	0.72	0.78	598,56	82,89	754

Fonte: Elaboração própria

É notável que o modelo Flash apresenta o único caso em que sua melhor pontuação individual foi superior à atuação do grupo, possivelmente por causa do comportamento errático causado pelas alucinações. De todo modo, mesmo

que um agente individual possa obter resultado superior ao obtivo pelo coletivo, ainda assim, é impossível se antecipar que agente será esse, de modo que continua, mesmo nesse caso, sendo uma estratégia mais adequada optar pela “sabedoria da multidão”.

A tabela a seguir mostra o resultado das médias ponderados das respostas de cada indivíduo agregado por modelo, nos permitindo analisar a efetividade da sabedoria das massas no contexto de agentes autônomos.

Tabela 7 Resultados dos agentes autônomos coletivamente

Modelo	*Recall*	Precisão	F1	Pontuação coletiva
Flash	1.00	0.67	0.75	594.32
Mini	1.00	0.50	0.67	622.16
Sabiazin	0.67	0.625	0.64	687.84
Sabia-3	1.00	1.00	1.00	769.52
GPT-4o	1.00	1.00	1.00	781.12
Grupo	1.00	0.875	0.93	756.96

Fonte: Elaboração própria

5.3.5 Análise da performance coletiva

Tabela 8 A performance do coletivo de coletivos (n = 602)

Recall	Precisão	F1	Pontuação	Maior pontuação
1.00	1.00	1.00	759.92	754

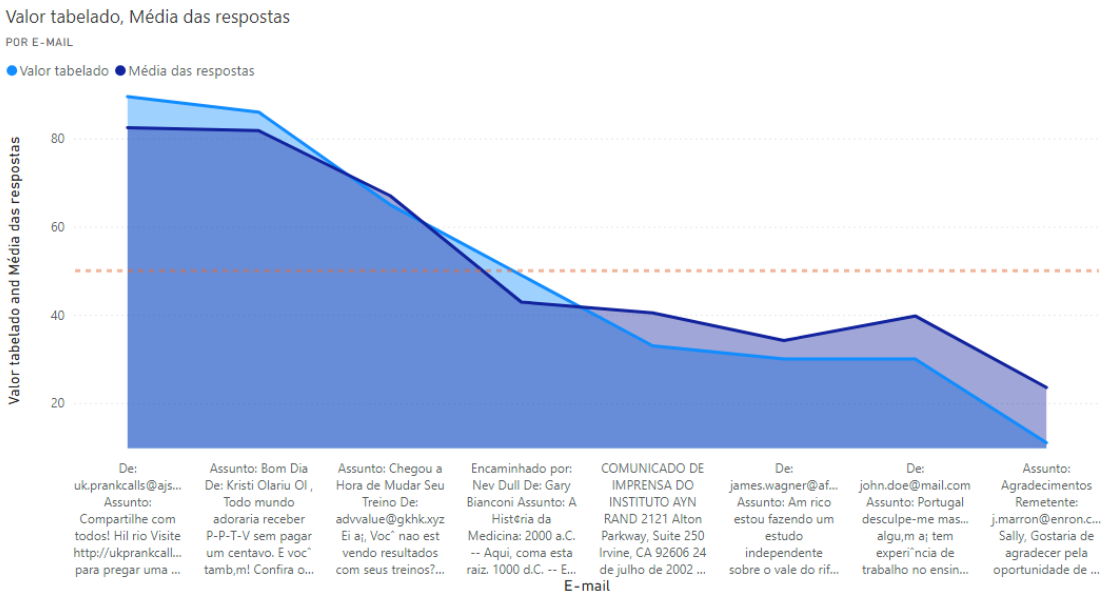
Fonte: Elaboração própria

Pode-se perceber, a partir da análise da Tabela 8, que o resultado do coletivo foi, apesar da performance surpreendente, pior que o coletivo de agentes humanos. O melhor coletivo dentre todos, porém foi o coletivo dos indivíduos autônomos que usava o modelo GPT-4o, indicando que houve diminuição da performance no coletivo dos agentes autônomos no processo de agregação de dados. Este fenômeno provavelmente decorreu de alucinações sendo agregadas ao conjunto por modelos de menor performance.

Análise da performance individual em cada resposta

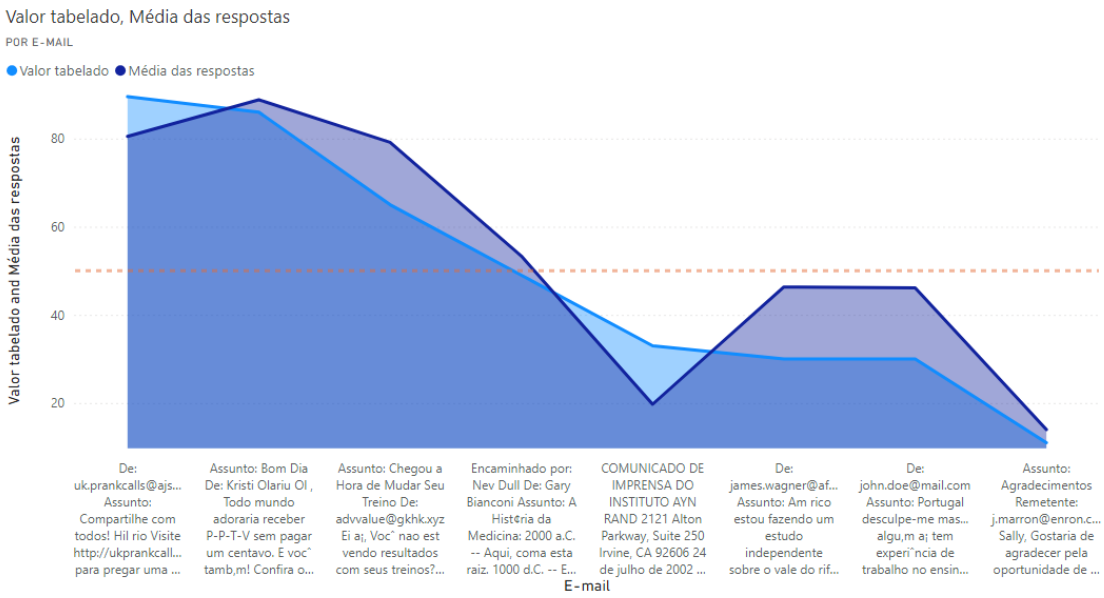
O gráfico da Figura 25 compara os valores atingidos a partir da média das respostas individuais com o valor tabelado. A linha vermelha representa o ponto 50%, ou seja, o ponto de ambiguidade na classificação do e-mail. Tudo acima da linha é classificado como fraude e abaixo, como sendo um e-mail seguro.

Figura 25 Diferenças entre as respostas esperadas e a performance para o coletivo de agentes humanos



Conforme apresenta a figura 26, os agentes humanos tiveram, coletivamente, uma performance muito mais assemelhada ao valor tabelado (Apêndice A), porém ainda existe diferença notável especialmente nos e-mails considerados mais seguros: esse fato se deve provavelmente à suspeita constante que os agentes colocam sobre o teor do e-mail, aumentando o viés “Paranoico”.

Figura 26 Diferenças entre as respostas esperadas e a performance para o coletivo de agentes autônomos



Fonte: Elaboração própria

Pode-se ver que os agentes autônomos tiveram maior diferença do valor tabelado (Apêndice A), ou seja, um desempenho pior, nas perguntas 1, 3, 5 e 7.

É notável que o coletivo de humanos não marcou nenhum e-mail como tendo menos que 23% de chance de ser fraude, enquanto o coletivo de agentes autônomos teve mais votos de confiança, marcando os e-mails mais confiáveis como apenas 10% de chance de ser fraudulento.

6 DISCUSSÃO

A plataforma se demonstrou uma ferramenta poderosa no trabalho de análise e classificação de fraudes. Os resultados demonstraram o potencial da colaboração para a resolução dos mais variados problemas.

6.1 BENEFÍCIOS DA PLATAFORMA

Houve diversas dificuldades em agregar os grupos de agentes, especialmente em se tratando do tipo de problema, que exigia muita intuição e experiência prévia dos agentes. O uso de LLMs permitiu a reutilização dos mesmos passos de resolução de diversos problemas analíticos, sendo eles relacionados a fraudes, no caso em pauta, mas podendo se relacionar a quaisquer outros assuntos que possam se beneficiar da inteligência coletiva para obter uma melhor resposta.

6.2 LIMITAÇÕES DA PLATAFORMA

Apesar da precisão elevada e alto grau de informação adquirido pelas enquetes, há problemas que impossibilitariam o uso do protótipo criado como um produto final.

1. O tempo de análise para os agentes humanos é longo, podendo demorar até dias para se ter um número bom de respostas de qualidade.
2. Mesmo para agentes autônomos o tempo é elevado comparado a algoritmos especializados na detecção de fraudes de e-mail.
3. O custo de aquisição de usuários é elevado pelo alto tráfego necessário e custo com anúncios.

As limitações apontadas exigem que a solução continue a ser aperfeiçoada para o objetivo específico de tratar de fraudes em e-mails, mas o trabalho permitiu confirmar que a sabedoria das massas não é um tópico de interesse apenas para melhores decisões por parte de humanos. Ficou evidente que IAs também pode se beneficiar de IC para reduzir o risco e os efeitos negativos da alucinação de agentes individuais. Aliás, estabelecer níveis de temperatura mais elevados pode contribuir para que IAs testem mais

possibilidades, ou seja, alucinem mais, para poder chegar a conclusões ainda mais sóbrias.

6.3 FUTURAS MELHORIAS

Na intenção da construção de um produto viável no futuro, foram identificados diversos pontos que, sozinhos ou em conjunto, podem elevar o *design* até um nível de usabilidade real.

6.3.1 Interface

Para os agentes humanos, a usabilidade da plataforma é um problema constante e alguns *feedbacks* de usuários sugerem que existem oportunidadesde melhoria, como por exemplo as instruções, o *feedback* visual instantâneo e a responsividade da página.

6.3.2 A adesão humana

A diversidade, como ponto central para a inteligência coletiva, implica que a atração de usuários para a plataforma deve acontecer de forma mais orgânica a fim de maximizar os diferentes pontos de vista. A disparidade entre o número de homens e mulheres demonstra que a fatia demográfica escolhida não estava perfeitamente distribuída, além da maior concentração de respondentes jovens.

6.3.3 Alucinação em agentes autônomos

O problema de alucinação é constante quando se trata de LLMs. O trabalho de agregar os dados de diferentes indivíduos foi um esforço para interseccionar os pontos de vista e minimizar este impacto. Ainda assim, os modelos de menor performance contribuíram negativamente para o coletivo de agentes autônomos. Este problema pode ser resolvido por meio de uma avaliação de alucinação como proposto por Wei *et al*. (2024), que se utiliza de um algoritmo de “Julgamento” de modelos propensos a alucinar, a fim de impedir que as alucinações impactem a resposta do grupo.

7 CONCLUSÃO

A capacidade da IA de identificar fraudes como um ser humano também implica sua capacidade de criar novos esquemas de fraude indetectáveis a seres humanos. Essa consciência do potencial destrutivo da inteligência artificial deve ser central aos desenvolvedores de IA e às empresas financiando os grandes modelos.

Os fenômenos únicos observados tanto no comportamento emergente dos grupos de agentes humanos e de agentes autônomos demonstram grande intersecção na maneira com que os humanos e os modelos de IA interpretam texto e abstraem conhecimento.

Os resultados permitiram concluir que o fenômeno de sabedoria das massas pode não só ser estendido para grupos de modelos de IA como também pode colaborar em um só ente coletivo com a capacidade de resolução de problemas complexos com performance superior, completamente sem programação prévia ou treinamento específico.

O aumento de diferentes modelos de IA treinados com *datasets* diferentes e compostos de arquiteturas radicalmente diferentes possibilitará a criação de grupos de agentes cada vez mais diversos, possibilitando que se desafiem os vieses inerentes ao próprio treinamento de modelos de IA.

A participação de humanos na geração de uma IC mista, envolvendo humanos e não humanos, proporcionou o melhor resultado nos testes realizados. Assim, pode-se imaginar que ainda há muito espaço para a intuição humana, mesmo ao se tratar da sabedoria das massas de IA. Ela pode ser utilizada para alavancar o potencial da intuição entre os agentes autônomos. Tarefas que exigem interpretação complexa vão poder ser realizadas tanto por máquinas quanto por humanos nos próximos anos. Ao mesmo tempo em que os humanos devem se preocupar com a competição gerada pelas máquinas, também podem aproveitar as possibilidades de colaboração com a IA, na construção do nosso futuro compartilhado.

REFERÊNCIAS

APWG. Phishing Activity Trends Report. 2023. Disponível em: https://apwg.org. Acesso em: 20 jul. 2024.

ALMENDRA, V.; SCHWABE, D. Fraud Detection by Human Agents: A Pilot Study. In: DI NOIA, T.; BUCCAFURRI, F. (eds). *E-Commerce and Web Technologies. EC-Web 2009.* Lecture Notes in Computer Science, v. 5692. Springer, Berlin, Heidelberg, 2009. https://doi.org/10.1007/978-3-642-03964-5_28.

BENGIO, Y.; GOODFELLOW, I.; COURVILLE, A. Deep learning. v. 1. Cambridge, MA, USA: MIT Press, 2017.

CAMPOS, Bianca Mariana Migliorini de. Uso de grafos no auxílio de fraudes em cartões de crédito. *Repositório Institucional da UTFPR – RIUT*, Curitiba, 2022. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/31715. Acesso em: 12 jun. 2024.

ENGEL, D.; WOOLLEY, A. W. The wisdom of select crowds. *Journal of Personality and Social Psychology*, v. 105, n. 2, p. 212–229, 2013.

FORTE, A.; LARCO, V.; BRUCKMAN, A. Decentralization in Wikipedia Governance. *Journal of Management Information Systems*, v. 26, n. 1, p. 49–72, 2009. https://doi.org/10.2753/MIS0742-1222260103.

GALTON, F. Vox Populi. *Nature*, v. 75, p. 450–451, 1907. https://doi.org/10.1038/075450a0.

HARVARD BUSINESS REVIEW. The Devastating Business Impacts of a Cyber Breach. 2023. Disponível em: https://hbr.org/2023/05/the-devastating-business-impacts-of-a-cyber-breach. Acesso em: 8 jan. 2025.

HASAN, Tsabbita Isnina; SILALAHI, Christine Irene; RUMAGIT, Reinert Yosua; PRATAMA, Galih Dea. UI/UX Design Impact on E-Commerce Attracting Users. *Procedia Computer Science*, v. 245, p. 1075-1082, 2024. DOI: 10.1016/j.procs.2024.10.336. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050924031442. Acesso em: 31 jan. 2025.

HONG, L.; PAGE, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, v. 101, n. 46, p. 16385–16389, 2004.

KIM, Hyunsoo; LEE, Jiyoon; KIM, Heejo; KIM, Huy. Whom do you trust? Analyzing online phishing attack vectors. *Computers & Security*, Amsterdam, v. 68, p. 144-157, jun. 2017.

LÉVY, Pierre. A inteligência coletiva: por uma antropologia do ciberespaço. São Paulo: Coleção Folha Grandes Nomes do Pensamento, 2015.

LAZER, D.; PENTLAND, A.; ADAMIC, L.; ARAL, S.; BARABÁSI, A. L.; BREWER, D.; ... VAN ALSTYNE, M. Computational social science. *Science*, v. 323, n. 5915, p. 721-723, 2009.

MALONE, Thomas W.; LAUBACHER, Robert; YATES, Michael S. The collective intelligence genome. *MIT Sloan Management Review*, v. 51, n. 3, p. 21, 2010.

MURPHY, Gregory B.; TOCHER, Neil. Gender differences in the effectiveness of online trust-building information cues: An empirical examination. *The Journal of High Technology Management Research*, v. 22, n. 1, p. 26-35, 2011.

OECD. Explanatory memorandum on the updated OECD definition of an AI system. *OECD Artificial Intelligence Papers*, n. 8, 2024. OECD Publishing, Paris. https://doi.org/10.1787/623da898-en.

PAGE, Scott E. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press, 2008.

POWERS, David M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, v. 2, n. 1, p. 37-63, 2011. Disponível em: https://arxiv.org/abs/2010.16061. Acesso em: 15 jan. 2025.

SCIKIT-LEARN. sklearn.neural_network.MLPClassifier. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Acesso em: 12 jan. 2025.

SERAPIO-GARCÍA, Greg; SAFDARI, Mustafa; CREPY, Clément; SUN, Luning; FITZ, Stephen; ROMERO, Peter; ABDULHAI, Marwa; FAUST, Aleksandra; MATARIĆ, Maja. Personality traits in large language models. *arXiv*, n. 2307.00184, 21 set. 2023. Disponível em: https://arxiv.org/pdf/2307.00184. Acesso em: 14 jan. 2025.

SUROWIECKI, James. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Anchor, 2005.

SUBHADEEP, Chakraborty. Phishing Email Detection [Data set]. Kaggle, 2023. https://doi.org/10.34740/KAGGLE/DSV/6090437.

SYMANTEC. Annual Threat Report. 2023. Disponível em: https://symantec.com. Acesso em: 20 jul. 2024.

TAKATA, Ryosuke; MASUMORI, Atsushi; IKEGAMI, Takashi. Spontaneous Emergence of Agent Individuality through Social Interactions in LLM-Based Communities. *arXiv preprint*, arXiv:2411.03252v1, 2024. Disponível em: https://arxiv.org/abs/2411.03252v1. Acesso em: 15 jan. 2025.

THARWAT, Alaa. Classification assessment methods. *Applied Computing and Informatics*, v. 17, n. 1, p. 168-192, 2021. Disponível em: https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html. Acesso em: 15 jan. 2025.

UTFPR. Comunicado sobre e-mails phishing, 2020. Disponível em: https://www.utfpr.edu.br/servidores/site/mural/comunicado-alerta-de-email-fraudulento. Acesso em: ago. 2024.

VIANNA, Fernando; GRAEML, Alexandre; PEINADO, Jurandir. An Aggregate Taxonomy for Crowdsourcing Platforms, their Characteristics, and Intents. *BAR - Brazilian Administration Review*, 2022.

WEI, Jiaheng et al. Measuring and reducing LLM hallucination without gold-standard answers. *arXiv*, n. 2402.10412, 6 jun. 2024. Disponível em: https://arxiv.org/pdf/2402.10412. Acesso em: 14 jan. 2025.

APÊNDICE A

A seleção dos e-mails foi uma tarefa manual, exigindo encontrar um conjunto de e-mails que pudessem explorar as abordagens de cada agente, como também mantendo o número limitado para minimizar a frustração dos agentes humanos.

As características consideradas para a escolha foi:

1. **Intervalo de confiança**: E-mails que representassem fraudes óbvias para a maioria dos agentes, assim como e-mails sem o menor indício de fraude.
2. **Tamanho do e-mail:** E-mails de diferentes comprimentos para entender como afetaria a performance dos agentes
3. **Tema:** E-mails de diferentes remetentes em diferentes contextos

Abaixo, a relação dos e-mails:

Rótulo real	**Previsão**	**Confiança**	**Texto original**	**Texto traduzido**
Safe Email	Safe Email	0.8023623499673551	re : portugal excuse me but . . . does anyone out there have experience of working in higher education in portugal - if so please get in touch . i need to talk to you ! !	De: john.doe@mail.com Assunto: Portugal Desculpe-me, mas... alguém aí tem experiência de trabalho no ensino superior em Portugal? Se tiver, entre em contato. Preciso falar com você!
Fraudulent Email	Fraudulent Email	0.9998033813959987	play a hilarious phone prank wind up your mates today ! please visit http : / / ukprankcalls . com	De: uk.prankcalls@ajskfk2.com Assunto: Compartilhe com todos! Hilário Visite http://ukprankcalls.com para pregar uma peça hilária nos seus amigos!
Safe Email	Safe Email	0.8095970285328419	re : amharic i am doing independent study on the rift valley of africa . amharic is a dialect spoken in that area , primarily eithiopia . i am trying to assertain what certain words would be in that language . for example , - - lion - - death - - baby - - water - - man - - woman - - family . any help would be appreciated ed . wagner	De: james.wagner@afu.edu.com Assunto: Amárico Estou fazendo um estudo independente sobre o Vale do Rift, na África. Amárico é um dialeto falado naquela área, principalmente na Etiópia. Estou tentando identificar como seriam certas palavras nessa língua. Por exemplo: Leão Morte Bebê Água Homem Mulher Família Qualquer ajuda será muito apreciada.

Rótulo real	Previsão	Confiança	Texto original	Texto traduzido
				Ed. Wagner
Fraudulent Email	Fraudulent Email	0.8657149642723103	re : good day , everybody will love to get p - p - t - v and pay not a cent . so will you , check the below web address , copy it and paste in your browser . the web address is : check 4 choices . com once who don ' t like such mails , plz . add slash and ' r ' to above address . and plz . give upt 10 days . i am missing working right now . . was michael enjoying running early last month ? . get back to you later , kristi olariu	Assunto: Bom Dia De: Kristi Olariu Olá, Todo mundo adoraria receber P-P-T-V sem pagar um centavo, e você também! Confira o endereço web abaixo: check4choices.com Copie e cole no seu navegador para acessar. Se você não deseja receber este tipo de e-mail, adicione uma barra e a letra "r" ao endereço acima. Por favor, aguarde até 10 dias para processar sua solicitação. Retornarei em breve. Atenciosamente, Kristi Olariu
Safe Email	Safe Email	0.890086637231321	global operations forum 2000 sally , i wanted to thank you for the opportunity to organize your global operations forum . i enjoyed working with you and patti and meeting the other individuals that keep this organization innovated . i hope you will keep me in mind should you need assistance with a similar function in the future . best of luck with your continued success . sincerely , julissa marron	Assunto: Agradecimentos Remetente: j.marron@enron.com Sally, Gostaria de agradecer pela oportunidade de organizar o seu Fórum de Operações Globais. Foi um prazer trabalhar com você e com a Patti, além de conhecer as outras pessoas que fazem parte dessa organização inovadora. Espero que você se lembre de mim caso precise de ajuda com uma função semelhante no futuro. Desejo boa sorte e sucesso contínuo. Sinceramente, Julissa Marron
Safe Email	Safe Email	0.5102856602634909	Forwarded-by: Nev Dull Forwarded-by: Gary Bianconi The History of Medicine: 2000 B.C. --	Encaminhado por: Nev Dull De: Gary Bianconi Assunto: A História da Medicina 2000 a.C. – Aqui, coma esta raiz.

Rótulo real	Previsão	Confiança	Texto original	Texto traduzido
			Here, eat this root. 1000 A.D. -- That root is heathen. Say this prayer. 1850 A.D. -- That prayer is superstition. Drink this potion. 1940 A.D. -- That potion is snake oil. Swallow this pill. 1985 A.D. -- That pill is ineffective. Take this antibiotic. 2000 A.D. -- That antibiotic is artificial. Here, eat this root.	1000 d.C. – Esta raiz é paga. Faça esta oração. 1850 d.C. – Esta oração é superstição. Beba esta poção. 1940 d.C. – Esta poção é óleo de cobra. Engula esta pílula. 1985 d.C. – Esta pílula é ineficaz. Tome este antibiótico. 2000 d.C. – Este antibiótico é artificial. Aqui, coma esta raiz.
Fraudulent Email	Safe Email	0.6551220056287638	no more work out ' s heya , are you not seeing any results from your workout ' s ? do you feel like your workout ' s are boring ? are you serious about getting back into shape ? if so , then it ' s time for that change ! http : / / therein . advantagesandvalue . com / h	Segue a mensagem corrigida: Assunto: Chegou a Hora de Mudar Seu Treino De: advvalue@gkhk.xyz Ei aí, Você não está vendo resultados com seus treinos? Seus treinos estão ficando entediantes? Está realmente sério(a) em voltar à forma? Se sim, então é hora de fazer uma mudança! Visite: http://therein.advantagesandvalue.com/h Atenciosamente,
Safe Email	Fraudulent Email	0.6642362205315443	PRESS RELEASE FROM THE AYN RAND INSTITUTE 2121 Alton Parkway, Suite 250 Irvine CA 92606July 24, 2002 FOR IMMEDIATE RELEASEGOVERNMENT REGULATION IS KILLING THE STOCK MARKET IRVINE, CA -- The steep decline in the stock market is being fueled by investors' realization that increasing government constraints on corporate America will harm business and the economy--not help them, said Yaron Brook, executive director of the Ayn Rand Institute. "Instead of	COMUNICADO DE IMPRENSA DO INSTITUTO AYN RAND 2121 Alton Parkway, Suite 250 Irvine, CA 92606 24 de julho de 2002 PARA DIVULGAÇÃO IMEDIATA A REGULAMENTAÇÃO GOVERNAMENTAL ESTÁ MATANDO O MERCADO DE AÇÕES IRVINE, CA – A forte queda no mercado de ações está sendo alimentada pela percepção dos investidores de que o aumento das restrições governamentais sobre as corporações americanas prejudicará os negócios e a economia, em vez de ajudá-los, afirmou Yaron Brook, diretor executivo do Instituto Ayn Rand.

Rótulo real	Previsão	Confiança	Texto original	Texto traduzido
			launching a witch-hunt against CEOs and rushing to give the government wider powers over business, we should get rid of the complex and contradictory regulations that encourage bad accounting and prevent shareholders from acting in their own interest." "The common explanation that 'greed' is to blame makes no sense--the abuses in companies like Enron and WorldCom were not exercises in self-interest, but in self-destruction. "In an unfettered free market the desire for profit is satisfied by honest, long-range, rational behavior: by innovating, by hiring the best employees, by selling	"Em vez de lançar uma caça às bruxas contra CEOs e apressar-se em dar ao governo mais poderes sobre as empresas, deveríamos eliminar as regulamentações complexas e contraditórias que incentivam a contabilidade ruim e impedem os acionistas de agirem em seu próprio interesse." "A explicação comum de que a ganância é a culpada não faz sentido – os abusos em empresas como Enron e WorldCom não foram exercícios de interesse próprio, mas sim de autodestruição." "Em um mercado livre e desimpedido, o desejo de lucro é satisfeito por um comportamento honesto, de longo prazo e racional: inovando, contratando os melhores funcionários, vendendo produtos de qualidade e fornecendo i

APÊNDICE B

Tabela das respostas discursivas dadas pelos agentes humanos:

ID do Agente	Ordem da pergunta	Resposta discursiva	Resposta quantitativa
13	1	A forma nada formal de falar, pra mim, nao me leva a desconfiar que seja frauda, pode ser amigos se falando	30 (Seguro)
13	2	Essa Url de procedˆncia duvidoso, cadˆ o s ali do final?	100 (Fraude)
27	2	E-mail com letras esquisitas e link esquisito, also HTTP	100 (Fraude)
27	1	E-mail gen,rico e pedindo ação urgente/imedi ata	100 (Fraude)
27	4	Golpe	50 (Ambíguo)
35	2	Ukprankcalls usando host externo	100 (Fraude)
35	4	Apoio … sensibilidade	100 (Fraude)
35	3	Falta de pontuação correta, capitaliza‡ao errada em algumas letras	70 (Fraude)
35	1	John Doe , o Sem Nome na cultura em inglês	100 (Fraude)
35	2	Pirataria	100 (Fraude)
35	7	Spam	100 (Fraude)