 doi: https://doi.org/10.35699/2525-8036.2024.53509

DETECÇÃO DE TEXTOS GERADOS PELO CHATGPT:
CAPACIDADES, LIMITAÇÕES E APLICAÇÕES PRÁTICAS

Felipe Soares de Moares[*]

Resumo: A popularização de ferramentas baseadas em Inteligência Artificial (IA) para geração de textos tem intensificado os debates sobre a autenticidade e confiabilidade do conteúdo digital, em meio a crescentes preocupações com fraudes acadêmicas, violações de direitos autorais e desinformação. Este estudo avaliou a eficácia de cinco ferramentas de detecção (Copyleaks AI Content Detector, AI da Smodin, PlagiarismDeTector.NET AI, ZeroGPT, e AI Content Detector) na distinção entre textos escritos por humanos e textos gerados pelo ChatGPT. Os resultados indicam que as ferramentas diferem em sua precisão e confiabilidade. Enquanto o Copyleaks e a AI da Smodin demonstraram alta precisão na identificação de textos gerados por IA e por humanos, o PlagiarismDeTector.NET AI apresentou diversas inconsistências, especialmente para distinguir corretamente textos gerados por IA. O ZeroGPT e o AI Content Detector mostraram desempenho variável, destacando-se em certos cenários e enfrentando desafios em outros. Essas descobertas evidenciam a necessidade contínua de aprimoramento das ferramentas de detecção de textos gerados por IA para garantir um ambiente digital íntegro e seguro.

Palavras-chave: Inteligência artificial; ChatGPT; Detecção de IA; Segurança digital.

DETECTION OF TEXTS GENERATED BY CHATGPT:
CAPABILITIES, LIMITATIONS, AND PRACTICAL APPLICATIONS

Abstract: The popularization of Artificial Intelligence (AI) tools for text generation has intensified debates about the authenticity and reliability of digital content amid growing concerns about academic fraud, copyright violations, and misinformation. This study evaluated the effectiveness of five detection tools (Copyleaks AI Content Detector, AI by Smodin, PlagiarismDeTector.NET AI, ZeroGPT, and AI Content Detector) in distinguishing between human-written texts and texts generated by ChatGPT. The results indicate that the tools vary in their accuracy and reliability. While Copyleaks and AI by Smodin demonstrated high accuracy in identifying AI-generated and human-written texts, PlagiarismDeTector.NET AI showed several inconsistencies, especially in correctly distinguishing AI-generated texts. ZeroGPT and AI Content Detector showed variable performance, excelling in some scenarios and facing challenges in others. These findings highlight the ongoing need for improvement in AI text detection tools to ensure a secure and trustworthy digital environment.

Keywords: Artificial intelligence; ChatGPT; AI detection; Digital security.

DETECCIÓN DE TEXTOS GENERADOS POR CHATGPT:
CAPACIDADES, LIMITACIONES Y APLICACIONES PRÁCTICAS

[*] Mestrando do Programa de Pós-Graduação em Comunicação da Universidade de Brasília (UnB), Brasil. Graduado em Humanidades e em Jornalismo pela Universidade Federal do Sul da Bahia (UFSB). Pesquisador do Centro de Estudos em Comunicação, Política e Tecnologia (CTPol/UnB). ORCID: https://orcid.org/0000-0002-6585-9073. Contato: felipe.moraes2018@gmail.com.

Resumen: La popularización de las herramientas de Inteligencia Artificial (IA) para la generación de textos ha intensificado los debates sobre la autenticidad y confiabilidad del contenido digital, en medio de crecientes preocupaciones por fraudes académicos, violaciones de derechos de autor y desinformación. Este estudio evaluó la eficacia de cinco herramientas de detección (Copyleaks AI Content Detector, AI de Smodin, PlagiarismDeTector.NET AI, ZeroGPT y AI Content Detector) en distinguir entre textos escritos por humanos y textos generados por ChatGPT. Los resultados indican que las herramientas varían en su precisión y confiabilidad. Mientras que Copyleaks y AI de Smodin demostraron alta precisión en la identificación de textos generados por IA y escritos por humanos, PlagiarismDeTector.NET AI presentó varias inconsistencias, especialmente en distinguir correctamente los textos generados por IA. ZeroGPT y AI Content Detector mostraron un rendimiento variable, destacándose en ciertos escenarios y enfrentando desafíos en otros. Estos hallazgos evidencian la necesidad continua de mejorar las herramientas de detección de textos generados por IA para garantizar un entorno digital seguro y confiable.

Palabras clave: Inteligencia artificial; ChatGPT; Detección de IA; Seguridad digital.

---

1 Introdução

A recente disseminação de ferramentas de geração de textos baseadas em Inteligência Artificial (IA) tem levantado profundos debates sobre a autenticidade e confiabilidade do conteúdo digital. Modelos de linguagem (LLMs) como o GPT-4 da *OpenAI* são capazes de produzir textos muitas vezes indistinguíveis daqueles escritos por seres humanos[1]. Diante desse cenário, torna-se imperativo o desenvolvimento de estratégias eficazes para mitigar o uso indevido dessas tecnologias, especialmente no que tange à disseminação de informações falsas, assegurando um ambiente digital íntegro e seguro[2].

Neste contexto, a detecção de textos gerados por IA tem se tornado uma área crítica de pesquisa e desenvolvimento à medida que crescem as preocupações com fraudes acadêmicas, violações de direitos autorais e desinformação[3]. Em relação a este último, levantam-se questões sobre como essas ferramentas podem ser usadas para influenciar decisões políticas, comportamentais e de consumo, muitas vezes de maneira não transparente e manipulativa. Embora apresente crescente relevância, este tema ainda é pouco explorado no campo da comunicação, área na qual este autor se insere, uma vez que o desenvolvimento de

---

[1] CHAKA, Chaka. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, v. 6, n. 2, 2023.
[2] AREMU, Toluwani. Unlocking Pandora's Box: Unveiling the Elusive Realm of AI Text Detection. *SSRN*, Rochester, NY, 2023.
[3] LUCCHI, Nicola. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, v. 15, n. 3, p. 602-624, 2024.

ferramentas para detecção de conteúdos artificiais constitui um fenômeno ainda bastante recente – o que justifica a necessidade de investigação mais aprofundada nesta área.

Para enfrentar esses novos desafios, ferramentas específicas foram desenvolvidas para identificar e diferenciar textos criados por IA daqueles produzidos por humanos. Conforme observado por Huang e Sun, a popularização do *ChatGPT* aprofundou esse cenário, uma vez que a tecnologia se destaca por ser significativamente mais poderosa do que as ferramentas convencionais utilizadas para esses fins[4].

Por outro lado, a literatura que versa sobre a análise de detecção de IA ainda é bastante escassa dada a recente popularização das ferramentas de IA. Em artigo publicado no *Journal of Applied Learning & Teaching* no ano passado, o pesquisador Chaka testou a precisão de cinco ferramentas de detecção de conteúdo de IA para detectar conteúdos gerados a partir dos *chatbots ChatGPT*, *YouChat* e *Chatsonic*. Conforme o autor, "esses detectores de conteúdo de IA parecem ainda não estar completamente prontos para detectar de maneira precisa e convincente o conteúdo gerado por IA em textos produzidos por máquinas em diferentes contextos[5]".

A partir desse cenário, pergunta-se: as atuais ferramentas de detecção de textos criados por IA são eficazes? Para responder essa indagação, este artigo objetiva analisar as ferramentas *Copyleaks AI Content Detector*[6]*; AI da Smodin*[7]*; PlagiarismDeTector.NET AI*[8]*; ZeroGPT*[9]*; e AI Content Detector*[10], que supostamente são capazes de realizar a detecção de textos gerados pelo *ChatGPT* – um dos *chatbots* mais populares no Brasil e no mundo. Tem-se como objetivo específico avaliar as habilidades técnicas das ferramentas, identificar possíveis limitações e explorar suas aplicações em cenários práticos.

Para alcançar o objetivo proposto, este artigo utilizará a abordagem de pesquisa qualitativa aliada com o método exploratório. Para isso, foram desenvolvidos cenários simulados contendo textos escritos tanto por humanos quanto pelo *ChatGPT*, no intuito de avaliar a precisão e a eficácia das ferramentas na distinção entre conteúdo gerado por IA e textos produzidos manualmente.

---

[4] HUANG, Yue; SUN, Lichao. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. Pré-print. *arXiv preprint*, v. *1*, 2023.
[5] CHAKA, Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic, *cit.*, p. 102.
[6] Disponível em: https://copyleaks.com/ai-content-detector. Acesso em: 20 jun. 2024.
[7] Disponível em: https://smodin.io/pt/detector-de-conteudo-ai. Acesso em: 20 jun. 2024.
[8] Disponível em:https://plagiarismdetector.net/pt/ai-content-detector. Acesso em: 20 jun. 2024.
[9] Disponível em: https://www.zerogpt.com/. Acesso em: 20 jun. 2024.
[10] Disponível em: https://chatgpt.com/g/g-NcgzdHmZc-ai-content-detector. Acesso em: 20 jun. 2024.

2 ChatGPT: origem, funcionamento e princípios

O *ChatGPT*, desenvolvido pela *OpenAI*, emergiu da fundação da empresa em 11 de setembro de 2015, sediada em São Francisco. A missão inicial da *OpenAI*, liderada por nomes proeminentes como Elon Musk, Sam Altman, Peter Thiel, Reid Hoffman e Jessica Livingston, foi desenvolver sistemas de IA para o benefício global[11].

Baseado no modelo de aprendizado de máquina GPT (*Generative Pre-trained Transformer*), o *ChatGPT* realiza treinamento supervisionado e não supervisionado, aprendendo autonomamente nuances e padrões da linguagem a partir de vastos conjuntos de textos disponíveis na *internet*[12]. Durante esse processo, ajusta seus parâmetros internos para aprimorar a previsão das próximas palavras em um contexto dado.

O funcionamento do *ChatGPT* baseia-se em redes neurais profundas, compostas por múltiplas camadas de processamento. Cada camada é dedicada a interpretar diferentes aspectos do texto, permitindo ao modelo compreender contextos complexos e gerar respostas coerentes e contextualmente apropriadas[13]. Essas redes utilizam mecanismos de atenção, como a atenção autorregressiva, para focar nas partes mais relevantes do texto durante a geração de respostas.

Desta forma, o *ChatGPT* representa um avanço significativo na IA, exemplificando como os modelos de linguagem podem revolucionar a interação humana com a tecnologia. Enraizado na *OpenAI* e com um robusto funcionamento em redes neurais profundas, o *ChatGPT* continua a se destacar como uma ferramenta inovadora, abrindo caminho para futuras descobertas e aplicações em IA – apesar de seus riscos.

3 Procedimentos Metodológicos

A escolha do método exploratório se concretiza devido ao problema de pesquisa ainda ser bastante recente e pouco explorado pela literatura. Para Gil, a pesquisa exploratória busca aprofundar a compreensão do problema em estudo, com o objetivo de proporcionar maior familiaridade e, assim, facilitar a construção de hipóteses[14]. Esse mesmo pensamento é corroborado por Trindade e Oliveira, que afirmam que “a pesquisa exploratória costuma ser

---

[11] COUTINHO, Dimítria. 2023, o ano em que o ChatGPT mostrou a IA generativa ao mundo. *Data Privacy Brasil*, 29 de dez. 2023.

[12] MEYER, Jesse G.; URBANOWICZ, Ryan J.; MARTIN, Patrick C. N.; O’CONNOR, Karen; LI, Ruowang; PENG, Pei-Chen; BRIGHT, Tiffani J.; TATONETTI, Nicholas; WON, Kyoung Jae; GONZALEZ-HERNANDEZ, Graciela; MOORE, Jason H. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, v. 16, n. 1, 2023, p. 20.

[13] SARRION, Eric. *Exploring the Power of ChatGPT:* Applications, Techniques, and Implications. Berkeley: Apress, 2023, p. 9-16.

[14] GIL, Antonio Carlos. *Métodos e técnicas de pesquisa social*. São Paulo: Editora Atlas, 2008.

realizada quando o tema é pouco explorado, pois fornece informações preliminares sobre o objeto de pesquisa, esclarece conceitos fundamentais e promove familiaridade com o objeto de estudo[15]".

Desta forma, para viabilizar este estudo por meio da metodologia proposta, este trabalho construiu cenários simulados de textos para avaliar a precisão de cinco ferramentas disponíveis publicamente na *internet*. A seleção de quatro dessas ferramentas foi feita com base na ordem de resultados dos primeiros sites listados nas buscas realizadas no Google e Yahoo!, os dois buscadores mais usados no Brasil, ao pesquisar os termos "Detector de IA" e "Detector de ChatGPT", respectivamente. As ferramentas selecionadas foram: *Copyleaks AI Content Detector* e *AI da Smodin* (Google); *PlagiarismDeTector.NET AI* e *ZeroGPT* (Yahoo!).

Além das ferramentas mencionadas, também foi incluída a ferramenta *AI Content Detector*, desenvolvida com base na arquitetura GPT, da *OpenAI*, para detectar conteúdo gerado por IA. De acordo com a descrição da ferramenta, que pode ser acessada diretamente na plataforma do *ChatGPT*,

> o AI Content Detector é uma ferramenta sofisticada projetada para identificar e analisar textos gerados por inteligência artificial. Ele ajuda a manter a autenticidade e a integridade na escrita, distinguindo entre conteúdo gerado por IA e criado por humanos, o que é essencial para educadores e editores[16].

Nesse sentido, devido ao seu desenvolvimento baseado na tecnologia do *ChatGPT*, um dos objetos deste estudo, e suas supostas capacidades avançadas de detecção, a ferramenta foi incluída nos testes desta pesquisa.

Em relação aos textos, baseando-se na experiência de Aremu, foram selecionados três tipos textuais para serem testados: descritivo, narrativo e expositivo, sendo três deles escritos por humanos antes do surgimento das IAs, como se pode observar na Tabela 1[17].

Tabela 1 - Textos escritos por humanos antes das IAs

Tipo	Texto
Descritivo	"Ficara sentada à mesa a ler o Diário de Notícias, no seu roupão de manhã de fazenda preta, bordado a sutache, com largos botões de madrepérola; o cabelo louro um pouco

[15] TRINDADE, Alessandra Stefane C. E.; OLIVEIRA, Henry Poncio C. Inteligência artificial (IA) generativa e competência em informação: habilidades informacionais necessárias ao uso de ferramentas de IA generativa em demandas informacionais de natureza acadêmica-científica. *Perspectivas em Ciência da Informação*, v. 29, p. e-47485, 2024.
[16] SARKAR, Ratan Kumar. AI Content Detector. *ChatGPT*, 2024.
[17] AREMU, Unlocking Pandora's Box, *cit*., p. 03.

	desmanchado, com um tom seco do calor do travesseiro, enrolava-se, torcido no alto da cabeça pequenina, de perfil bonito; a sua pele tinha a brancura tenra e láctea das louras; com o cotovelo encostado à mesa acariciava a orelha, e, no movimento lento e suave dos seus dedos, dois anéis de rubis miudinhos davam cintilações escarlates" (O Primo Basílio, Eça de Queiroz).
Narrativo	"Julgou-a estúpida e egoísta, deixou-a, indignada, foi puxar a manga do vestido da mãe, desejando comunicar-se com ela. Sinha Vitória soltou uma exclamação de aborrecimentos, e, como o pirralho insistisse, deu-lhe um cascudo. Retirou-se zangado, encostou-se num esteio do alpendre, achando o mundo todo ruim e insensato. Dirigiu-se ao chiqueiro, onde os bichos bodejavam, fungando, erguendo os focinhos franzidos. Aquilo era tão engraçado que o egoísmo de Baleia e o mau humor de Sinha Vitória desapareceram. A admiração a Fabiano é que ia ficando maior" (Trecho de Vidas Secas, de Graciliano Ramos).
Expositivo	**Folha - Defender a focalização dos programas sociais é ser liberal?** "Tavares - Estive em São Paulo [depois da divulgação do documento" e tive de ouvir o dr. Delfim Netto defender a Constituinte de 1988, onde estão consagrados os direitos universais nas três áreas: saúde, assistência social e Previdência Social. Isso vinha sendo construído como políticas universais desde o tempo da ditadura, logo, não é um problema de ser conservador. É um problema de ser pateta ou de má-fé. E esse pessoal está tentando dar as rédeas da política social do governo. É evidente que os ministros da área social estão possessos, mas não vão armar uma briga com o ministro Palocci, a quem terei o prazer de, assim que for a Brasília, ir visitar para perguntar o que é aquilo. Como um documento da Fazenda fala sobre focalização?" (Entrevista da economista Maria da Conceição Tavares para a Folha de São Paulo, 2003).

Fonte: autoria própria.

E três deles escritos pelo *ChatGPT*, como pode ser observado na Tabela 2:

Tabela 2 - Textos escritos pelo ChatGPT

Tipo	Texto
Descritivo	A velha casa no topo da colina exalava uma aura de mistério. Suas janelas, embaçadas pelo tempo, revelavam cortinas desgastadas, pendendo tristemente. As paredes, outrora brancas, agora exibiam tons de cinza e verde, cobertas de musgo e hera. O telhado, de telhas partidas, parecia lutar para se manter de pé, enquanto o vento uivava entre as frestas. No jardim abandonado, flores silvestres cresciam desordenadamente, contrastando com a solidez das árvores centenárias. O cheiro de madeira úmida e folhas secas impregnava o ar. Cada detalhe sussurrava histórias de um passado distante, aguardando pacientemente para serem redescobertas (Escrito pelo ChatGPT a partir do comando "Escreva um texto de até 10 linhas no gênero textual descritivo").
Narrativo	Era uma tarde ensolarada quando João decidiu explorar a floresta atrás de sua casa. Com uma mochila cheia de lanches e uma curiosidade insaciável, ele seguiu por um trilho estreito, ladeado por árvores imponentes. Após caminhar por uma hora, avistou uma cabana abandonada. O lugar parecia ter saído de um conto de fadas, com janelas quebradas e uma porta que rangia ao menor toque. Ao entrar, encontrou um velho diário empoeirado sobre a mesa. As páginas revelavam a história de um aventureiro perdido, descrevendo tesouros escondidos e perigos inimagináveis. Fascinado, João sabia que sua vida jamais seria a mesma dali em diante (Escrito pelo ChatGPT a partir do comando "Escreva um texto de até 10 linhas no gênero textual narrativo").
Expositivo	A energia solar é uma fonte de energia renovável que utiliza a luz do sol para gerar eletricidade. Painéis solares, compostos por células fotovoltaicas, captam a radiação solar e a convertem em energia elétrica utilizável. Este processo não emite poluentes e reduz a dependência de combustíveis fósseis, contribuindo para a diminuição do impacto ambiental. Além disso, a energia solar é inesgotável e disponível em praticamente todas as regiões do planeta. Recentes avanços tecnológicos têm tornado os sistemas solares mais eficientes e acessíveis, incentivando o seu uso em residências, empresas e indústrias. Dessa forma, a energia solar se destaca como uma

	alternativa sustentável e promissora para o futuro energético global” (Escrito pelo ChatGPT a partir do comando “Escreva um texto de até 10 linhas no gênero textual expositivo”).

Fonte: autoria própria.

A incorporação de textos dos três principais gêneros textuais nos testes é essencial, uma vez que cada gênero possui características linguísticas e estilísticas distintas. Essa abordagem proporciona uma avaliação ampla e equilibrada da ferramenta, permitindo identificar suas capacidades e limitações em diferentes contextos. A partir desses cenários, pretende-se responder: i) qual é a precisão das ferramentas estudadas na detecção do conteúdo gerado integralmente pelo *ChatGPT*? e ii) qual é a precisão das ferramentas na detecção de cada tipo textual proposto?

4 Resultados e Discussões

4.1 Copyleaks AI Content Detector

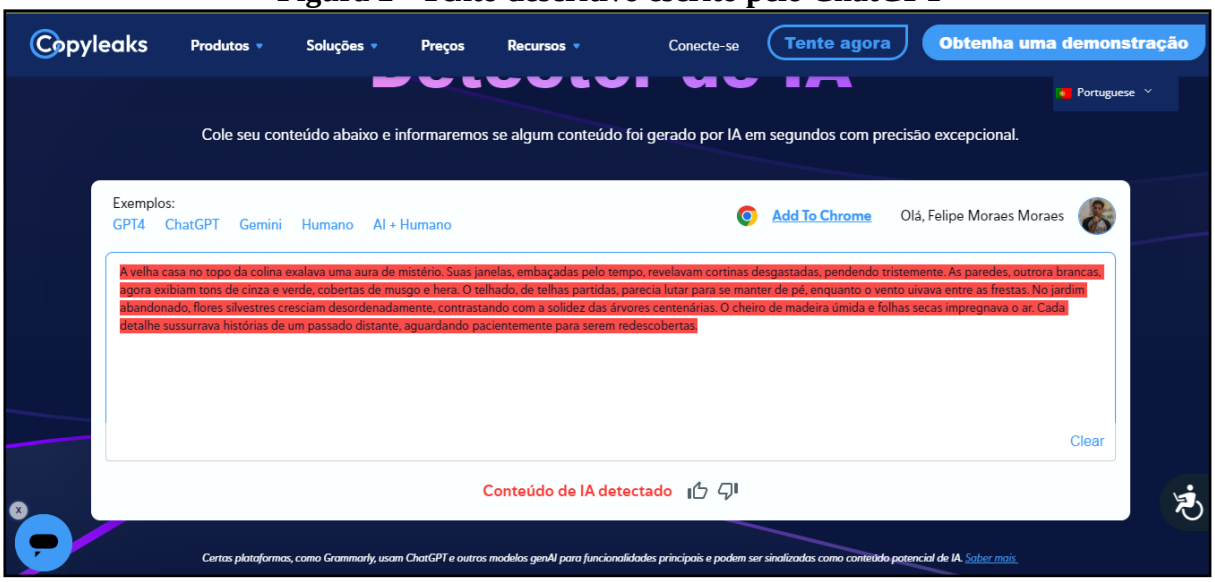
A primeira ferramenta analisada foi o *Copyleaks AI Content Detector*, na versão em português. Primeiramente, foram inseridos textos escritos por humanos nas seguintes categorias: descritivo, narrativo e expositivo. Nesta primeira fase, o aplicativo identificou corretamente que todos os textos foram escritos por humanos, como exemplificado na Figura 1, que mostra um texto descritivo retirado da obra "O Primo Basílio", de Eça de Queiroz, publicado em 1878.

Figura 1 - Texto descritivo escrito por humano

Fonte: autoria própria.

Já na segunda etapa, todos os textos escritos pelo *ChatGPT* foram também detectados, assertivamente, com algum conteúdo de IA, conforme exemplificado na Figura 2. Nesse exemplo, foi inserido um texto descritivo gerado a partir do comando "escreva um texto de até 10 linhas no gênero textual expositivo" no *ChatGPT*.

Figura 2 - Texto descritivo escrito pelo ChatGPT



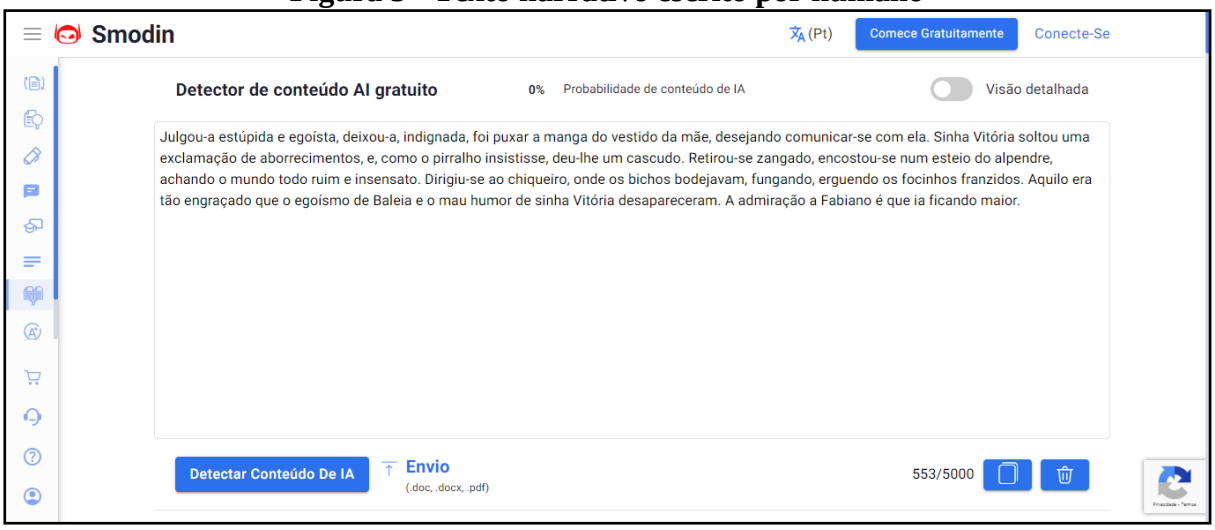
Fonte: autoria própria.

Com base nos testes realizados, conclui-se que a ferramenta *Copyleaks AI Content Detector* demonstrou alta precisão na detecção da autoria dos textos, diferenciando com sucesso os conteúdos escritos por humanos e os gerados por IA.

4.1.1 AI da Smodin

Na sequência de testes, a *AI da Smodin* seguiu os mesmos procedimentos das avaliações anteriores. Na primeira etapa, a ferramenta conseguiu identificar com precisão a origem dos textos escritos por humanos, seja ele descritivo, narrativo ou expositivo, como mostra o exemplo na Figura 3, onde foi inserido um trecho do livro "Vidas Secas", de Graciliano Ramos. Em todos os casos, a ferramenta identificou 0% de probabilidade de algum dos conteúdos escritos por humanos terem sido produzidos por IA.

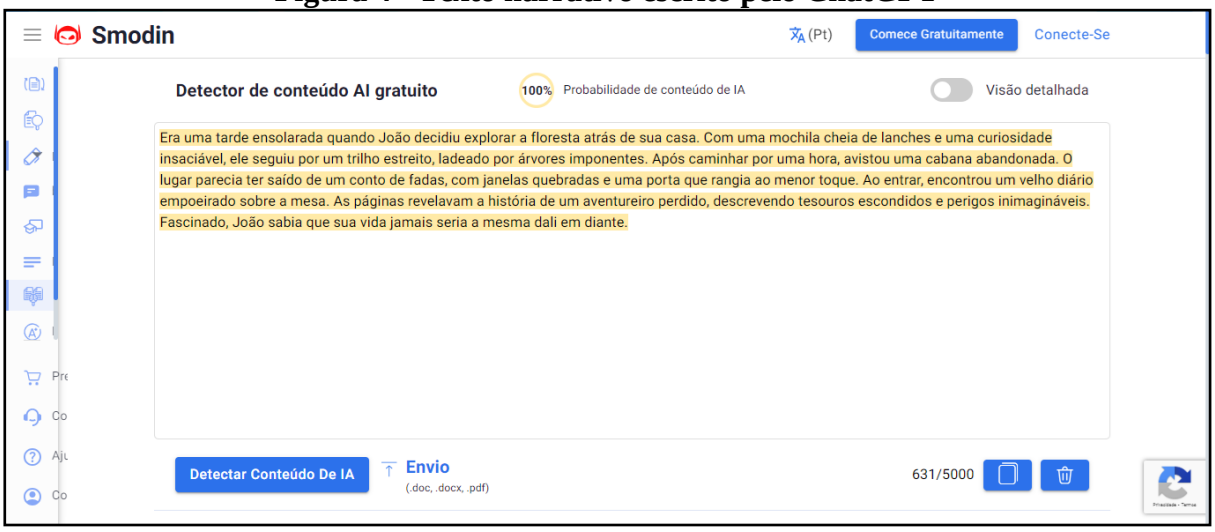
Figura 3 - Texto narrativo escrito por humano



Fonte: autoria própria.

Já na segunda etapa, todos os textos escritos pelo *ChatGPT* foram identificados com 100% de probabilidade de terem sido escritos por IA, como ilustrado na Figura 4. Nesse exemplo, foi inserido um texto narrativo escrito pelo *ChatGPT* a partir do comando “escreva um texto de até 10 linhas no gênero textual narrativo”.

Figura 4 - Texto narrativo escrito pelo ChatGPT



Fonte: autoria própria.

A partir dos testes realizados, conclui-se que a ferramenta *AI da Smodin* demonstrou ampla competência na distinção entre textos escritos por humanos e textos gerados pelo *ChatGPT*.

4.1.2 PlagiarismDeTector.NET AI

Os testes realizados com a ferramenta *PlagiarismDeTector.NET AI* também seguiram os mesmos passos das avaliações anteriores, apesar dos resultados divergentes. Na primeira etapa, na qual foram inseridos os textos escritos por seres humanos, a ferramenta identificou de forma assertiva que todos eles foram produzidos por humanos. No entanto, no teste exposto na Figura 5, ela detectou que 3% do texto foi gerado por IA – o que é falso, uma vez que o texto foi publicado séculos atrás.

Figura 5 - Texto descritivo escrito por humano

Fonte: autoria própria.

Na segunda fase de testes, os resultados foram ainda mais inconsistentes. Todos os textos escritos pelo *ChatGPT* foram classificados como produzidos por seres humanos. Apenas no teste de texto descritivo, conforme evidenciado na Figura 6, houve uma discrepância de 1%, indicando que essa porção específica foi gerada por uma IA – o que não é verdade, uma vez que todo o texto foi escrito por IA.

Figura 6 - Texto descritivo escrito pelo ChatGPT

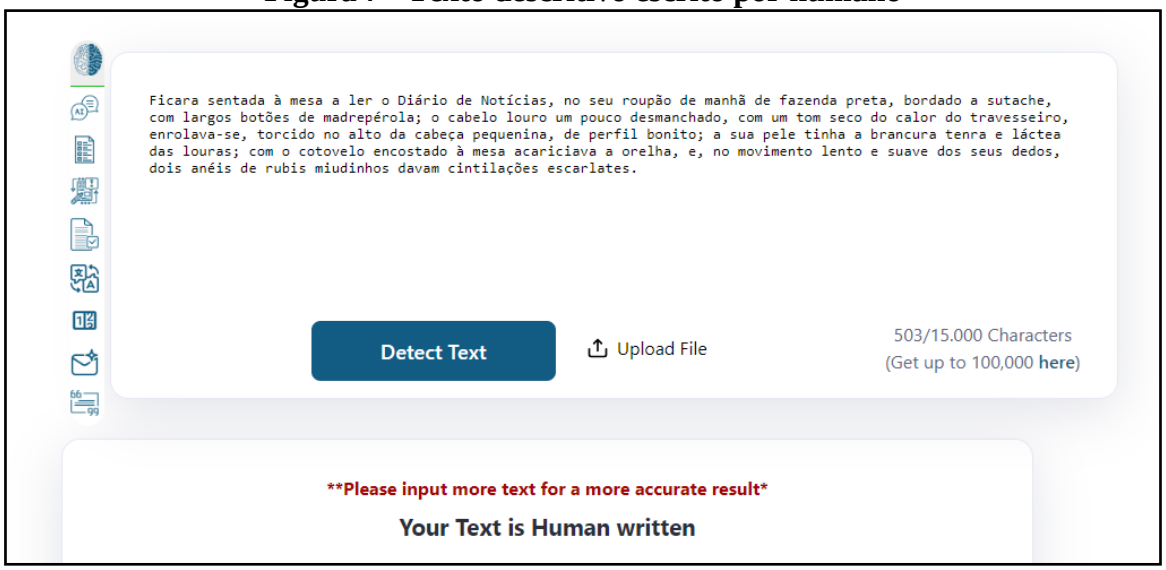
Fonte: autoria própria.

Desta forma, conclui-se que a ferramenta *PlagiarismDeTector.NET AI* demonstrou inconsistências relevantes em sua capacidade de diferenciar textos gerados por IA de textos escritos por seres humanos. Na primeira etapa, apesar de a ferramenta identificar corretamente os textos humanos na maioria dos casos, houve um falso positivo indicando que 3% de um dos textos foi gerado por IA. Na segunda fase, a ferramenta falhou em identificar corretamente os textos gerados pelo *ChatGPT*, classificando-os predominantemente como escritos por humanos, com apenas uma pequena discrepância de 1% no teste de texto descritivo. Esses resultados evidenciam limitações substanciais na precisão da ferramenta para detecção de autoria de textos, especialmente nos escritos por IA, em todos os gêneros textuais.

4.1.3 ZeroGPT

A ferramenta *ZeroGPT* confirmou com precisão que os textos escritos por humanos foram de fato produzidos por eles. No entanto, ela solicitou a inserção de mais texto para obter "um resultado mais preciso", como evidenciado na Figura 7 – mas que não impediu que desse um resultado satisfatório:

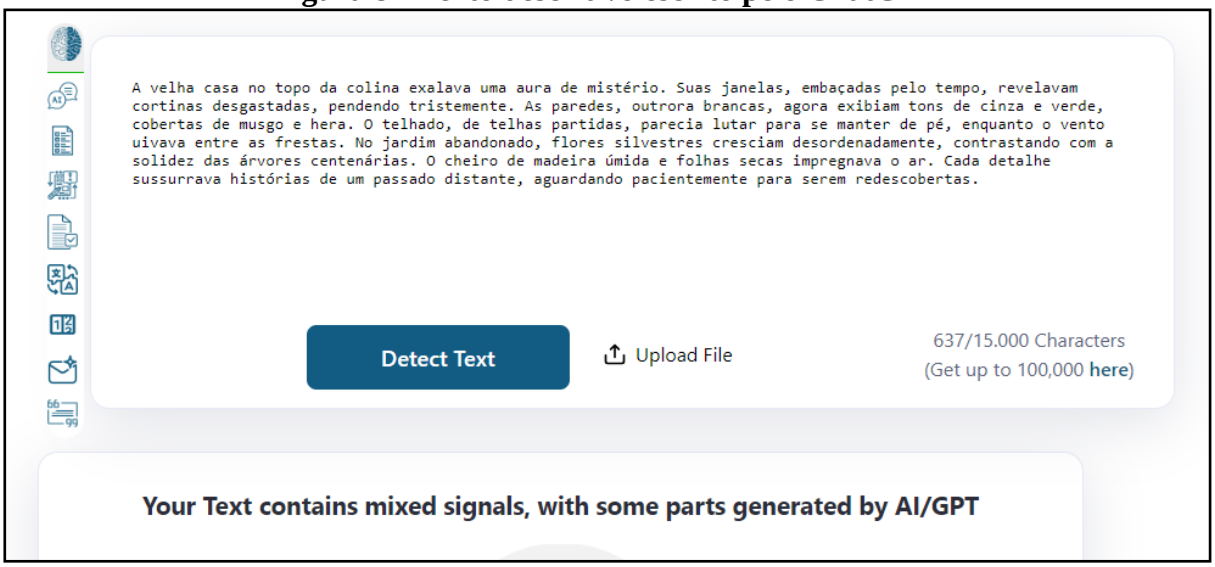
Figura 7 - Texto descritivo escrito por humano



Fonte: autoria própria.

Já na segunda fase de testes, em que foram inseridos os textos escritos pelo *ChatGPT*, a ferramenta obteve resultados variados. Em um primeiro momento, ao inserir um texto descritivo, ela afirmou que o "texto contém sinais mistos, com algumas partes geradas por AI/GPT", como mostra a Figura 8:

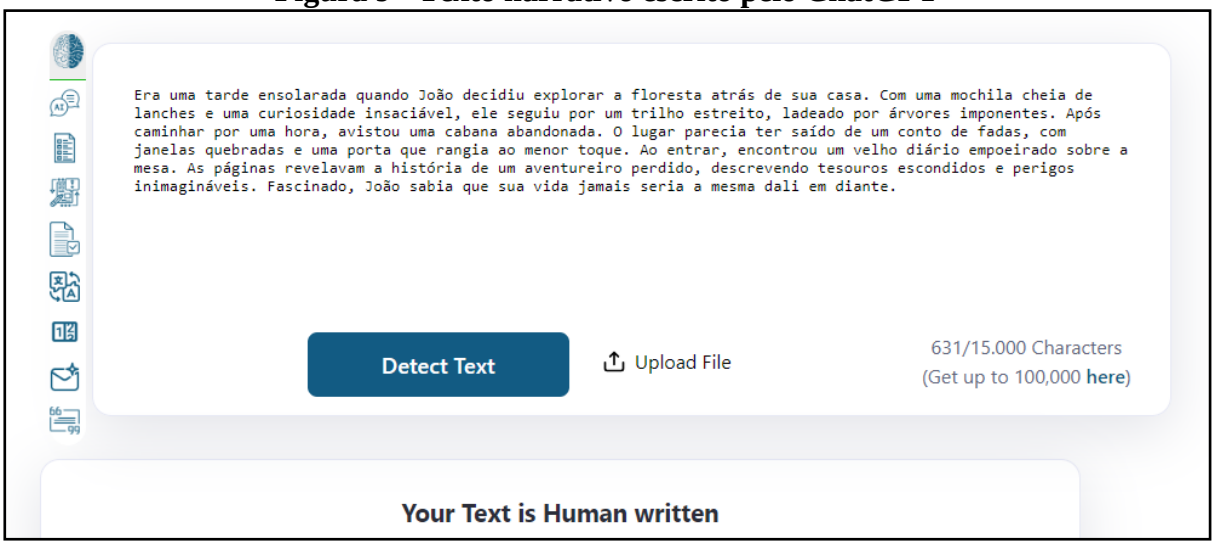
Figura 8 - Texto descritivo escrito pelo ChatGPT



Fonte: autoria própria.

Ao inserir um texto narrativo escrito pelo *ChatGPT*, a ferramenta acusou erroneamente ser escrito por um humano, como evidenciado na Figura 9:

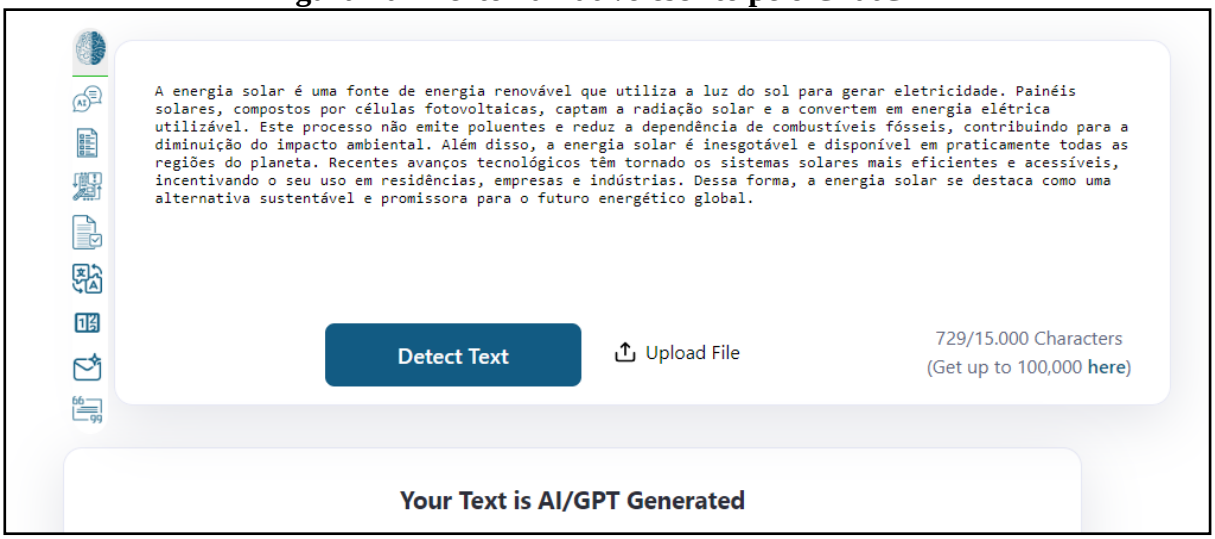
Figura 9 - Texto narrativo escrito pelo ChatGPT



Fonte: autoria própria.

E ao inserir um texto expositivo, informou corretamente que se tratava de uma produção por IA, como ilustrado na Figura 10:

Figura 10 - Texto narrativo escrito pelo ChatGPT



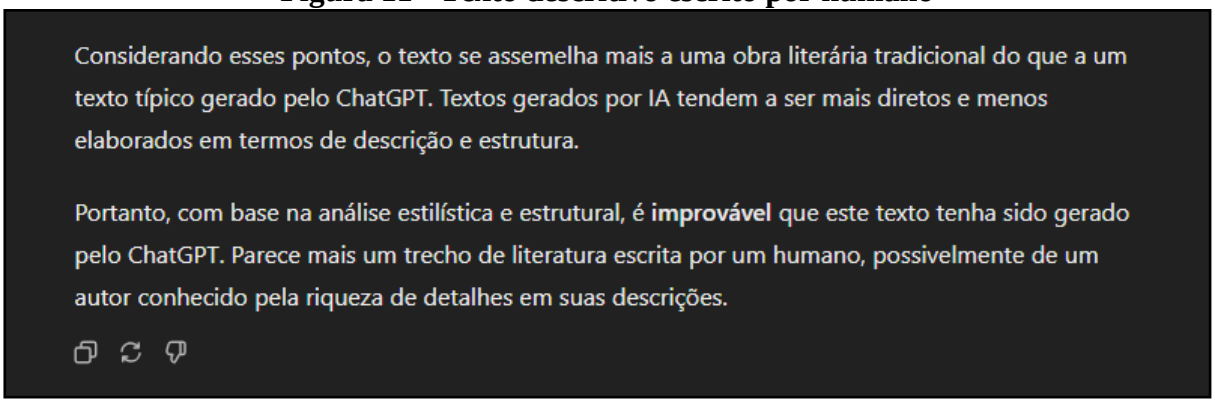
Fonte: autoria própria.

Nesse sentido, conclui-se que a ferramenta *ZeroGPT* demonstrou precisão ao confirmar que os textos escritos por humanos foram de fato produzidos por eles. No entanto, na segunda fase de testes, que envolveu textos gerados pelo *ChatGPT*, a ferramenta apresentou resultados variados, com erro ou imprecisão na maior parte dos cenários e apenas um acerto. Esses resultados destacam a necessidade de maior refinamento e precisão na avaliação de textos gerados por IA pela ferramenta.

4.1.4 AI Content Detector

Conforme descrito na metodologia, a ferramenta *AI Content Detector* é baseada na tecnologia do *ChatGPT*. Por esse motivo, ela também foi selecionada para os testes deste estudo, uma vez que compartilha a mesma metodologia do *chatbot* da *OpenAI*. Na primeira fase, a ferramenta apresentou resultados variados. Ao tentar identificar o texto descritivo escrito por humano, ela afirmou de forma correta que "é improvável que este texto tenha sido gerado pelo *ChatGPT*", como mostra a Figura 11:

Figura 11 - Texto descritivo escrito por humano


Considerando esses pontos, o texto se assemelha mais a uma obra literária tradicional do que a um texto típico gerado pelo ChatGPT. Textos gerados por IA tendem a ser mais diretos e menos elaborados em termos de descrição e estrutura.

Portanto, com base na análise estilística e estrutural, é **improvável** que este texto tenha sido gerado pelo ChatGPT. Parece mais um trecho de literatura escrita por um humano, possivelmente de um autor conhecido pela riqueza de detalhes em suas descrições.

Fonte: autoria própria.

Este mesmo resultado perpetuou ao tentar identificar o texto expositivo. Ela afirmou que "é improvável que este texto tenha sido gerado pelo *ChatGPT*. Parece ser um discurso ou uma declaração feita por uma pessoa real com conhecimento específico sobre a política brasileira". No entanto, ao tentar identificar o texto narrativo, não conseguiu fazê-lo com precisão, como indicado na Figura 12:

Figura 12 - Texto narrativo escrito por humano

Considerando esses pontos, o texto apresenta características que poderiam estar presentes em um texto gerado por ChatGPT, especialmente por sua simplicidade e clareza. No entanto, a presença de elementos emocionais e a forma de descrever ações e sentimentos também são típicos de textos literários humanos, particularmente em obras que retratam a vida cotidiana e os sentimentos de forma direta.

Dado que o texto não apresenta padrões claros de geração de IA, como repetições ou falta de coesão, e possui um estilo narrativo comum em literatura brasileira, é possível que tenha sido escrito por um humano. No entanto, não podemos descartar totalmente a possibilidade de ser gerado por ChatGPT sem mais informações contextuais.

Com base na análise, é **possível** que o texto seja gerado por ChatGPT, mas também pode ser um trecho de literatura escrita por um humano. A análise não fornece uma certeza definitiva.

Fonte: autoria própria.

Conforme demonstrado na imagem, a ferramenta afirma que "é possível que o texto seja gerado por *ChatGPT*, mas também pode ser um trecho de literatura escrita por um humano. A análise não fornece uma certeza definitiva". Em vista disso, a análise não oferece uma certeza sobre a origem do texto, deixando espaço para interpretações e tornando a ferramenta pouco confiável. Apesar disso, um aspecto interessante da ferramenta é como ela fundamenta e descreve os textos ao emitir seu veredito – inédito entre as ferramentas analisadas.

Já na segunda fase, a ferramenta também não obteve resultados integralmente satisfatórios, a começar pelo texto descritivo escrito pelo *ChatGPT*, como ilustrado na Figura 13, que ela acusa erroneamente como sendo escrito por um humano:

Figura 13 - Texto descritivo escrito pelo ChatGPT

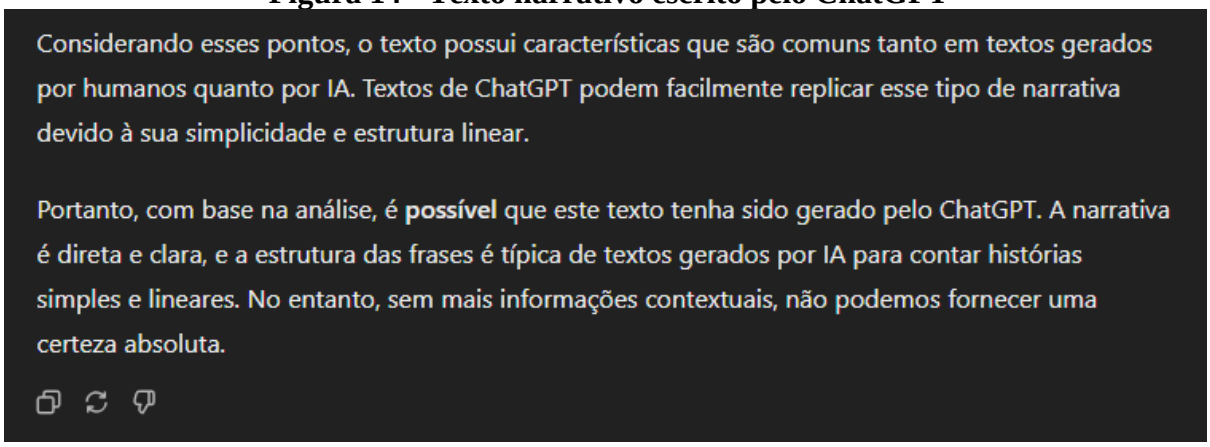
Considerando esses pontos, o texto possui características que podem estar presentes tanto em textos gerados por humanos quanto por IA. No entanto, a riqueza descritiva e a criação de uma atmosfera específica são frequentemente associadas a escritos humanos, especialmente em literatura descritiva.

Portanto, com base na análise, é **provável** que este texto tenha sido escrito por um humano. A atenção aos detalhes, a variação nas estruturas das frases e a evocação de uma atmosfera específica são indicativos de um estilo literário humano.

Fonte: autoria própria.

Nos outros dois testes, a ferramenta conseguiu identificar como escrito pelo *ChatGPT*, mas manteve algumas dúvidas, como mostra a Figura 14:

Figura 14 - Texto narrativo escrito pelo ChatGPT

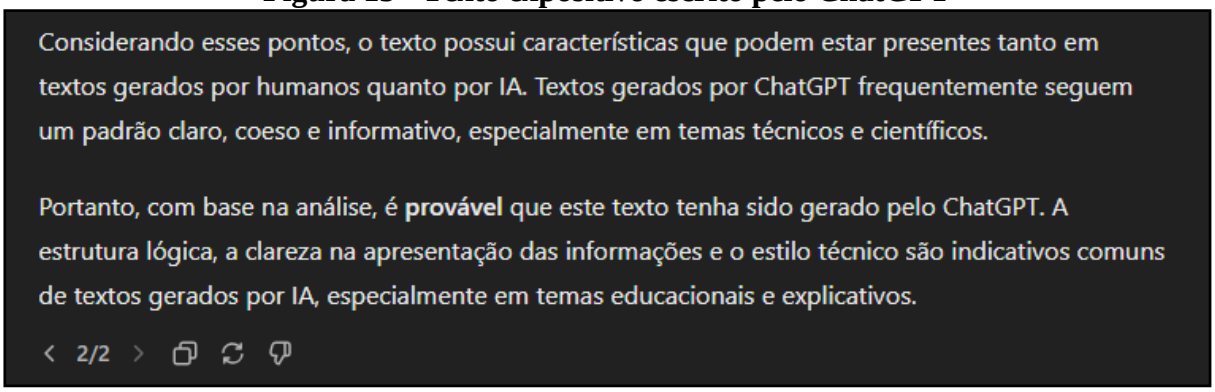

Considerando esses pontos, o texto possui características que são comuns tanto em textos gerados por humanos quanto por IA. Textos de ChatGPT podem facilmente replicar esse tipo de narrativa devido à sua simplicidade e estrutura linear.

Portanto, com base na análise, é **possível** que este texto tenha sido gerado pelo ChatGPT. A narrativa é direta e clara, e a estrutura das frases é típica de textos gerados por IA para contar histórias simples e lineares. No entanto, sem mais informações contextuais, não podemos fornecer uma certeza absoluta.

Fonte: autoria própria.

E também na Figura 15:

Figura 15 - Texto expositivo escrito pelo ChatGPT


Considerando esses pontos, o texto possui características que podem estar presentes tanto em textos gerados por humanos quanto por IA. Textos gerados por ChatGPT frequentemente seguem um padrão claro, coeso e informativo, especialmente em temas técnicos e científicos.

Portanto, com base na análise, é **provável** que este texto tenha sido gerado pelo ChatGPT. A estrutura lógica, a clareza na apresentação das informações e o estilo técnico são indicativos comuns de textos gerados por IA, especialmente em temas educacionais e explicativos.

2/2

Fonte: autoria própria.

Contudo, conclui-se que ao longo das fases de teste, a ferramenta demonstrou um desempenho variável. Em particular, ela se destacou ao identificar textos descritivos e expositivos escritos por humanos como possivelmente escrito por estes, sugerindo uma habilidade de reconhecer nuances e estilos típicos de escrita humana. Entretanto, a precisão da ferramenta foi menos consistente ao lidar com textos narrativos escritos por humanos, onde expressou incerteza sobre sua origem.

Na segunda fase, a ferramenta enfrentou desafios ao avaliar um texto descritivo gerado pelo *ChatGPT*, identificando-o erroneamente como escrito por um humano. Por outro lado, demonstrou mais eficácia ao reconhecer textos narrativos gerados pela IA, embora ainda

manifestasse hesitação em suas conclusões. Quanto aos textos expositivos, a ferramenta conseguiu indicar com maior segurança sua origem como sendo da IA.

Em suma, pode-se atestar que a ferramenta possui mais habilidade para identificar textos descritivos e expositivos escritos por humanos e textos expositivos gerados pela IA; e menos habilidade para identificar textos narrativos escritos por humanos e textos descritivos gerados pela IA. Por fim, um aspecto positivo notável foi a transparência da ferramenta ao fornecer justificativas detalhadas para suas análises, o que contribui para o entendimento das decisões tomadas durante a avaliação dos textos.

5 Considerações Finais

A análise das ferramentas de detecção de textos gerados por IA revelou uma variação significativa na eficácia de cada uma, destacando a necessidade de melhorias contínuas e refinamentos técnicos para garantir um ambiente digital confiável. Entre as ferramentas avaliadas, o *Copyleaks AI Content Detector* e a *AI da Smodin* se destacaram por sua alta precisão na distinção entre textos humanos e textos gerados pelo *ChatGPT*, demonstrando uma capacidade consistente de identificação correta em ambos os cenários testados.

Por outro lado, o *PlagiarismDeTector.NET AI* apresentou inconsistências preocupantes, falhando em detectar corretamente a origem dos textos em diversas ocasiões. A ferramenta apresentou um falso positivo na identificação de texto antigo como gerado por IA e, posteriormente, falhou em reconhecer corretamente textos produzidos pelo *ChatGPT* como sendo de origem artificial. Esse desempenho insatisfatório ressalta a importância de aprimoramentos na ferramenta para que possa ser considerada confiável e eficaz na detecção de conteúdos gerados por IA.

A ferramenta *ZeroGPT*, embora tenha mostrado precisão na identificação de textos escritos por humanos, revelou-se menos eficaz ao lidar com textos gerados pelo *ChatGPT*, indicando a necessidade de maior precisão e refinamento em suas análises. A ferramenta *AI Content Detector*, baseada na tecnologia do *ChatGPT*, apresentou um desempenho variável, mostrando competência em certos aspectos, mas falhando em outros, especialmente ao lidar com textos narrativos e descritivos gerados pela IA. Esse fenômeno pode ocorrer devido ao fato de que os textos descritivos gerados por IA podem ser confundidos com textos humanos pela consistência das descrições, enquanto os textos narrativos gerados por IA são mais facilmente identificados pela previsibilidade.

Em conclusão, a eficácia das ferramentas de detecção de textos gerados por IA varia

significativamente, com muitas delas ainda enfrentando dificuldades para realizar detecções precisas. Essa imprecisão pode facilitar a disseminação de informações falsas ou manipuladas, o que prejudica a capacidade do público de distinguir entre conteúdo confiável e não confiável, uma vez que há um risco elevado de que conteúdos falsos ou manipulados sejam apresentados como legítimos. Do mesmo modo, ao falharem em reconhecer a autenticidade de textos originais, estudantes podem ser erroneamente acusados de plágio ou de utilizar conteúdo gerado por IA, prejudicando suas avaliações acadêmicas e sua reputação.

Em última instância, é urgente implementar melhorias contínuas e desenvolvimentos técnicos que fortaleçam a capacidade das ferramentas de detectar textos gerados por IA, assegurando a autenticidade e a integridade do conteúdo digital. Para enfrentar esses desafios, é fundamental investir no aprimoramento das ferramentas de detecção, promover colaborações entre desenvolvedores, instituições acadêmicas e organizações de mídia, e aumentar a educação e a conscientização sobre as limitações dessas tecnologias, capacitando usuários a utilizá-las de forma crítica e informada.

Referências Bibliográficas

AREMU, Toluwani. Unlocking Pandora's Box: Unveiling the Elusive Realm of AI Text Detection. *SSRN*, Rochester, NY, 2023. Disponível em: http://dx.doi.org/10.2139/ssrn.4470719. Acesso em: 23 dez. 2024.

CHAKA, Chaka. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, v. 6, n. 2, 2023. Disponível em: https://doi.org/10.37074/jalt.2023.6.2.12. Acesso em: 23 dez. 2024.

COUTINHO, Dimítria. 2023, o ano em que o ChatGPT mostrou a IA generativa ao mundo. *Data Privacy Brasil*, 29 dez. 2023. Disponível em: https://www.dataprivacybr.org/documentos/2023-o-ano-em-que-o-chatgpt-mostrou-a-ia-generativa-ao-mundo/. Acesso em: 23 dez. 2024.

GIL, Antonio Carlos. *Métodos e técnicas de pesquisa social*. São Paulo: Editora Atlas, 2008.

HUANG, Yue; SUN, Lichao. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. Pré-print. *arXiv preprint*, v. 1, 2023. Disponível em: https://doi.org/10.48550/arXiv.2310.05046. Acesso em: 23 dez. 2024.

LUCCHI, Nicola. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, v. 15, n. 3, p. 602-624, 2024. Disponível em: https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965. Acesso em: 23 dez. 2024.

MEYER, Jesse G.; URBANOWICZ, Ryan J.; MARTIN, Patrick C. N.; O'CONNOR, Karen; LI, Ruowang; PENG, Pei-Chen; BRIGHT, Tiffani J.; TATONETTI, Nicholas; WON, Kyoung Jae; GONZALEZ-HERNANDEZ, Graciela; MOORE, Jason H. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, v. 16, n. 1, 2023. Disponível em: https://doi.org/10.1186/s13040-023-00339-9. Acesso em: 23 dez. 2024.

OPENAI. *OpenAI Charter*. São Francisco: 2024. Disponível em: https://openai.com/charter/. Acesso em: 23 dez. 2024.

SARKAR, Ratan Kumar. AI Content Detector. *ChatGPT*, 2024. Disponível em: https://chatgpt.com/g/g-NcgzdHmZc-ai-content-detector. Acesso em: 23 dez. 2024.

SARRION, Eric. *Exploring the Power of ChatGPT:* Applications, Techniques, and Implications. Berkeley: Apress, 2023.

TRINDADE, Alessandra Stefane C. E.; OLIVEIRA, Henry Poncio C. Inteligência artificial (IA) generativa e competência em informação: habilidades informacionais necessárias ao uso de ferramentas de IA generativa em demandas informacionais de natureza acadêmica-científica. *Perspectivas em Ciência da Informação*, v. 29, p. e-47485, 2024. Disponível em: https://doi.org/10.1590/1981-5344/47485. Acesso em: 23 dez. 2024.

Como citar este artigo: MORAES, Felipe Soares de. Detecção de textos gerados pelo ChatGPT: capacidades, limitações e aplicações práticas. *Revista de Ciências do Estado*, Belo Horizonte, v. 9, n. 2, p. 1–21, 2024.

Recebido em 16.07.2024

Publicado em 24.12.2024

CC BY 4.0 Atribuição 4.0 Internacional