UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I – CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CLEARLISON NÓBREGA DA COSTA

ANÁLISE DINÂMICA DE DETECTORES DE CONTEÚDO IA

CAMPINA GRANDE
2023

CLEARLISON NÓBREGA DA COSTA

ANÁLISE DINÂMICA DE DETECTORES DE CONTEÚDO IA

Trabalho de Conclusão de Curso apresentado ao Departamento de Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Ciência da Computação.

Área de concentração: Engenharia de Software

Orientador: Profa. Me. Ana Isabella Muniz Leite.

CAMPINA GRANDE
2023

C837a Costa, Clearlison Nobrega da.
Análise dinâmica de detectores de conteúdo IA [manuscrito] / Clearlison Nobrega da Costa. - 2023.
41 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.
"Orientação : Profa. Ma. Ana Isabella Muniz Leite, Coordenação do Curso de Computação - CCT. "

1. Inteligência artificial. 2. Testes automatizados. 3. Selenium IDE. I. Título

21. ed. CDD 006.3

Elaborada por Geovani S. de Oliveira - CRB - 4/2314

Biblioteca Central BC/UEPB

CLEARLISON NÓBREGA DA COSTA

ANÁLISE DINÂMICA DE DETECTORES DE CONTEÚDO IA

Trabalho de Conclusão de Curso de Graduação em Ciência da Computação da Universidade Estadual da Paraíba, como requisito à obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Engenharia de Software

Aprovada em: 31 / Agosto / 2023.

BANCA EXAMINADORA

Ana Isabella M. Leite

Profa. Me. Ana Isabella Muniz Leite (CCT/UEPB)
Orientador(a)

Fabio Luiz Leite Junior

Prof. Dr. Fábio Luiz Leite Jr (CCT/UEPB)
Examinador(a)

Cheyenne Ribeiro G. Isidro

Profa. Me. Cheyenne Ribeiro Guedes Isidro (CCT/UEPB)
Examinador(a)

RESUMO

É notável a evolução tecnológica a que vivenciamos atualmente, em que diversas áreas veem ganhando mais e mais inovações. Nesse caminho, podemos constatar que o processamento de linguagem natural despertou cada vez mais o interesse da sociedade em geral quando tecnologias, tal qual o ChatGPT, começaram a emergir e apresentar uma capacidade ímpar na produção de conteúdo textual. Com isso, veio à luz também uma válida preocupação, conseguir identificar quando um conteúdo foi ou não produzido por uma inteligência artificial. Isso se dá pela constatação que nós, seres humanos, temos dificuldades consideráveis de faz-lo sem nenhum auxílio. A principal ajuda que nos está sendo fornecida no presente momento, são diversas ferramentas que se dizem capazes de fazer a detecção desse tipo de conteúdo. Diante desse contexto, a presente pesquisa busca aferir a eficácia desses detectores de conteúdo de IA, através de uma análise dinâmica, capitaneada por testes automatizados com o auxílio da ferramenta Selenium IDE. Vale destacar que para a realização de tais testes, foram escolhidos alguns cenários, os quais todas as ferramentas analisadas foram expostas. Por fim, a pesquisa conseguiu constatar que detectores de conteúdo de IA terão cada vez mais importância dentro do contexto em que estamos inseridos, tais ferramentas conseguiram uma taxa de acerto de 56,7%. Levando em consideração os cenários em que as ferramentas foram expostas e tendo em vista que os mesmos tem um contexto limitado, foi notório que as ferramentas ainda não possuem a confiabilidade necessária, precisando assim evoluir um pouco mais.

Palavras-Chave: inteligência artificial; testes automatizados; Selenium IDE.

ABSTRACT

The technological advancements we are currently experiencing are remarkable, with various fields witnessing continuous innovations. In this journey, it is evident that natural language processing has increasingly captured the interest of society, especially with technologies like ChatGPT emerging and demonstrating unparalleled ability in generating textual content. Along with this progress, a valid concern has arisen: the ability to identify whether a piece of content was produced by artificial intelligence or not. This concern arises from the realization that we, as humans, struggle considerably to do so without any assistance. At present, the primary aid provided comes in the form of several tools claiming to detect this type of content. In this context, this research aims to analyze the effectiveness of these AI content detectors through a dynamic analysis, conducted via automated tests using the Selenium IDE tool. It is worth noting that for these tests, specific scenarios were selected, exposing all the analyzed tools. In the end, the research managed to determine that AI content detectors will play an increasingly important role within our current context. These tools achieved an accuracy rate of 56.7%. Considering the scenarios in which the tools were tested and the fact that these scenarios have limited contexts, it became apparent that the tools still lack the necessary reliability, indicating a need for further development.

Keywords: artificial intelligence; automated tests; Selenium IDE.

LISTA DE ILUSTRAÇÕES

Figura 1 – Metodologia da análise dinâmica 19
Figura 2 – Interface do Selenium IDE 25
Figura 3 – Taxa de acertos para textos elaborados por inteligência artificial 34
Figura 4 – Taxa de acertos para textos elaborados por humano 35
Figura 5 – Taxa de acertos para textos híbridos 36
Figura 6 – Taxa geral de acertos para texto acadêmico 36
Figura 7 – Taxa de acertos para textos de apresentação 37
Figura 8 – Taxa geral de acertos 37

LISTA DE QUADROS

Quadro 1 – Cenário 1 (Texto elaborado por uma IA) 21
Quadro 2 – Cenário 2 (Texto elaborado por humano) 22
Quadro 3 – Cenário 3 (Texto Híbrido) 23

LISTA DE TABELAS

Tabela 1 – Características dos Detectores de IA 17
Tabela 2 – Resultados do cenário 1 (Texto elaborado por inteligência artificial) 28
Tabelo 3 – Resultados do cenário 2 (Texto elaborado por humano) 29
Tabela 4 – Resultados do cenário 3 (Texto híbrido) 30

SUMÁRIO

1 INTRODUÇÃO 09
2 FUNDAMENTAÇÃO TEÓRICA 11
2.1 Análise Dinâmica 12
2.1.1 Testes exploratórios usando Selenium 13
2.1.1.1 Selenium IDE 13
2.1.1.2 Selenium WebDriver 14
2.1.1.3 Selenium Grid 14
2.2 ChatGPT 14
2.3 Detectores de Conteúdo de IA 14
2.4 Lista de Detectores de IA 16
2.4.1 AI Content Detector (Copyleaks) 16
2.4.2 AI Text Classifier 16
2.4.3 AI Writing Check 16
2.4.4 Writefull GPT 17
2.4.5 ZeroGPT 17
2.4.6 Características dos detectores de IA 17
3 METODOLOGIA 19
3.1 Definição dos Cenários de Teste 20
3.2 Execução dos testes 23
3.2.1 Método de classificação dos resultados 26
4 RESULTADOS DOS TESTES 27
4.1 Resultados dos testes por cenário 27
4.2 Análise dos resultados 31
4.2.1 Análise por ferramenta 31
4.2.2 Análise geral dos resultados 33
5 CONCLUSÃO 38
REFERÊNCIAS 40

1 INTRODUÇÃO

É inegável que vivemos uma revolução tecnológica nos últimos anos, em diversas áreas e com os mais variados objetivos, seja desde um simples sensor de presença que aciona uma lâmpada em um corredor escuro até mesmo um robô que otimiza e automatiza um serviço que por vezes era feito por diversos homens e que demandava muito esforço. Isso denota uma mudança sensível no comportamento e ações dentro da sociedade atual, como apontam Alsulaimani e Islam (2022) em seu artigo “Impact of 4ir technology and its impact on the current deployment”,

> [...] a Quarta Revolução Industrial alterará tanto o que fazemos quanto quem somos. Nosso sentimento de privacidade, nossos conceitos de propriedade, nossos hábitos de compra, a quantidade de tempo que passamos trabalhando e brincando, e como buscamos carreiras, aprimoramos nossos talentos, conhecemos pessoas e cultivamos relacionamentos serão todos impactados por isso. (Alsulaimani e Islam, 2022)

Dentre essas áreas, uma que vem ganhando mais interesse por parte dos pesquisadores, desenvolvedores e tantos outros profissionais da tecnologia é a inteligência artificial (IA). E dentro dessa área, uma tecnologia que vem sendo cada vez mais aprimorada é o processamento de linguagem natural (PLN), como muitos conhecem atualmente pela alcunha de ‘ChatGPT’, embora existam tantos outros. Esse avanço trouxe à luz a necessidade de haver meios de se identificar se algum conteúdo foi produzido por uma inteligência artificial, como afirma Muhammad Abdul-Mageed (apud Heikkilä, 2023), “Estamos em uma corrida armamentista para criar métodos de detecção que possam corresponder aos modelos mais recentes e poderosos”.

O presente estudo tem o intuito de analisar ferramentas que se dizem capazes de classificar se um conteúdo textual foi uma produção de um ser humano ou se o mesmo foi processado por uma inteligência artificial. Diante do crescente avanço tecnológico nessa linha, tornou-se fundamental compreender até que ponto é possível identificar a autoria de um texto.

Obter esses dados é de suma importância, uma vez que ferramentas de produção de textos através de inteligência artificial vem ganhando cada vez mais espaço nos dias atuais. Compreender o que pode ser melhorado dentro dessas ferramentas de detecção de conteúdo produzido por inteligência artificial pode

produzir novas versões cada vez mais eficientes na distinção de textos, uma vez que é observado que grande parte dessas ferramentas que tem esse objetivo foram desenvolvidas recentemente, o que pode indicar que existe muito espaço para melhorias, contribuindo cada vez mais com ferramentas capazes de discernir textos com mais facilidade e precisão.

2 FUNDAMENTAÇÃO TEÓRICA

Com o avanço tecnológico que se vem observando ao longo dos últimos anos, uma área amplamente explorada e aperfeiçoada é a inteligência artificial. Dentro dessa subárea da tecnologia, o processamento de linguagem natural vem ganhando cada vez mais relevância. O'Reilly (2023) aponta em seu relatório anual de tendências tecnológicas que houve um crescimento de 42% no interesse ao referido tópico.

Esse crescente avanço e interesse sobre o tema acabou trazendo à luz a importância de existirem ferramentas que possam distinguir conteúdos que foram produzidos por uma IA de conteúdos 100% autorais. Isso porque essa tarefa, para uma pessoa sem nenhum auxílio, se torna muito difícil, justamente pelo fato de o objetivo de ferramentas de produção de conteúdo serem exatamente o de conseguirem se assemelhar ao máximo a um ser humano. Os autores Brandl e Ellis (2023) mostram em sua pesquisa, na qual entrevistaram mais de 1.900 americanos, que mais de 53% não conseguiram diferenciar com precisão entre a saída do ChatGPT e a escrita humana.

Como tais ferramentas estão tendo seu uso amplamente disseminado dentro da sociedade atual, a possibilidade de identificar se textos foram produzidos de forma não natural vem ganhando cada vez mais importância, conforme destacado por McCoy (2023). Essas ferramentas de processamento de linguagem natural podem facilmente produzir textos para notícias, descrições de produtos que uma empresa queira comercializar, textos muito bem estruturados que possam ser utilizados por um aluno na produção de um trabalho escolar, discursos políticos, comentários em redes sociais. Muitas vezes, tal utilização dessas ferramentas tem por objetivo ludibriar e/ou manipular o seu público alvo. Conforme destacado por Couto (2023),

> [...] A preocupação reside na facilidade com que narrativas falsas podem ser geradas por grupos organizados de maneira automatizada, aproveitando um ambiente de interface amigável que resulta em textos aparentemente coesos e desafiadores de serem identificados como produzidos por inteligência artificial. (Couto, 2023)

Com um cenário tão propício para que esse tipo de problema seja cada vez mais observado em diversos setores da sociedade, é imprescindível que existam

ferramentas que possam ser utilizadas dentro dos contextos já citados e atestem a legitimidade e confiabilidade dos textos que consumimos diariamente. De acordo com Kirchner et al (2023),

> Reconhecemos que a identificação de textos escritos por IA tem sido um ponto importante de discussão entre educadores, e igualmente relevante é reconhecer os limites e impactos dos classificadores de texto gerados por IA na sala de aula. (Kirchner et al, 2023)

Os detectores de texto utilizam em sua análise textual a exploração de diversas características de linguagem bem específicas, tais como o vocabulário encontrado no texto, a estrutura gramatical, o estilo de escrita, entre outros. Para que isso seja possível, utilizam algoritmos de aprendizado de máquina que foram treinados em uma base de dados que disponha de diversos textos.

Ao analisar essas ferramentas de forma dinâmica, foi possível aferir quanto à eficácia, precisão e limitações das mesmas.

2.1 Análise Dinâmica

A análise dinâmica consiste numa técnica de teste de software a qual possui o objetivo de analisar o comportamento de um software dentro de um ambiente controlado e com cenários definidos previamente. Esse tipo de análise é considerado uma análise do tipo ‘caixa preta’, onde não há conhecimento prévio da estrutura interna do software que está sob análise. “O programa executável final é tratado como uma caixa preta, e os testes são planejados para mostrar se o sistema atende ou não a seus requisitos.” (Sommerville, 2011, pág. 409).

Como destaca Carvalho (2019), o principal objetivo da referida análise é o de demonstrar se o software está em conformidade com as funções e requisitos ao qual o mesmo foi planejado e desenvolvido para desempenhar. Essa conformidade pode ser medida de diversas formas, como por exemplo, I) as saídas obtidas através de uma interação com uma determinada função é realmente a esperada ou não; II) se o tempo de resposta é adequado ou está causando muita espera; III) se o comportamento apresentado após a interação com o software é o que foi determinado nas etapas que antecederam o desenvolvimento e a implantação propriamente dita desse software em um ambiente testável; IV) qual a performance e eficiência desse software como um todo.

Para que seja realizada essa análise, um ponto importante a ser realizado é a escolha e criação de cenários de teste que serão aplicados no software. Tais cenários descritos como um conjunto de entradas que foram definidas de forma prévia e que serão aplicadas no software afim de elucidar se todos os pontos citados acima serão atendidos ou não como o esperado.

Pressman e Maxim (2014, pág. 520) explana sobre a importância da realização desse tipo de testes em uma aplicação,

> [...] é necessário determinar se o software se comportará ou não de maneira que satisfaça aos usuários. O padrão Teste de cenário descreve uma técnica para exercitar o software do ponto de vista do usuário. Uma falha nesse nível indica que o software deixou de atender aos requisitos visíveis para o usuário. (Pressman e Maxim, 2014, p. 520)

2.1.1 Testes exploratórios usando Selenium

Existem diversas ferramentas de automação de teste que podem auxiliar na análise dinâmica, tais como TestComplete (ref), Cypress (ref) e Ranorex (ref), que tem a finalidade de simular a interação entre um usuário real e o software em questão. Essa possibilidade de automação traz um grande auxílio na execução dos cenários de teste, uma vez que tais ferramentas podem mapeá-los e executá-los quantas vezes for necessário para se obter a análise mais precisa. Dentre essas várias ferramentas existe o Selenium (ref).

Segundo Garcia et al (2020), o ecossistema Selenium é formado por um conjunto de ferramentas de código aberto multiplataforma, usado essencialmente na automatização de testes de aplicações web via navegador. Em geral, são testadas funcionalidades da aplicação e sua compatibilidade com navegadores diversos. Pode ser utilizado em diversos navegadores, tais como Chrome e Firefox, também consegue suportar várias linguagens de programação, como por exemplo, Python e Java. Algumas ferramentas contidas no ecossistema são: Selenium IDE, Selenium WebDriver e Selenium Grid.

2.1.1.1 Selenium IDE

Contri (2019) afirma que o Selenium IDE é uma extensão disponível para os navegadores Chrome, Firefox e Edge que se baseia na utilização dos comandos do Selenium para gravar as interações de um usuário com uma determinada aplicação

Web em um script. Tal script reproduz fielmente, de forma automatizada, a ação que foi gravada, podendo fazer isso repetidas vezes.

2.1.1.2 Selenium WebDriver

Como destacado por Contri (2019), cada navegador possui seu driver específico, e o WebDriver utiliza-se dos mesmos para realizar a automação das interações com a aplicação. É considerada a abordagem mais moderna do Selenium atualmente, pois permite que o script interaja de forma nativa com cada navegador, como qualquer usuário real faria. É recomendado para testes mais complexos, uma vez que necessita que o utilizador de uma maior familiaridade com a ferramenta em comparação com o Selenium IDE.

2.1.1.3 Selenium Grid

Consiste na utilização do WebDriver na realização de um mesmo teste de forma paralela em várias máquinas remotas. Em outras palavras, é o uso do WebDriver de uma forma potencializada e que permite que o mesmo teste seja realizado em várias máquinas diferentes ao mesmo tempo, com navegadores distintos, trazendo uma redução significativa no tempo para que o teste cubra várias configurações possíveis, como destaca Contri (2019).

2.2 ChatGPT

Segundo Deng e Lin (2022), o ChatGPT é um sistema de processamento de linguagem natural desenvolvido pela OpenAI (ref), com a finalidade de criar diálogos que se assemelham a conversas humanas, através da compreensão do contexto da conversa e da geração de respostas adequadas. Este sistema foi treinado com uma extensa base de dados de interações conversacionais.

2.3 Detectores de Conteúdo de IA

De acordo com Helfstein (2023), detectores de IA são ferramentas que foram desenvolvidas com o objetivo de fazer a distinção entre um texto que foi elaborado por um ser humano e um texto que foi elaborado por uma inteligência artificial. Para isso, ele analisa padrões textuais, com auxílio do aprendizado de máquina e processamento de linguagem natural.

Em sua maioria, essas ferramentas também utilizam o processamento de linguagem natural aliado a algoritmos de classificação que buscam por palavras ou sequência de palavras iguais ou com muita semelhança em outros textos que estão em suas bases de dados, analisa também a coesão e coerência do texto que foi colocado sob análise. Em resumo, o quão maior for a possibilidade que o texto analisado pelo detector de conteúdo possuir trechos iguais ou semelhantes a outros textos ao qual foi comparado, mais cresce a possibilidade que o texto seja uma produção de uma inteligência artificial, como destaca Pinho et al (2022).

Pode parecer que sejam apenas ferramentas de detecção de plágio, porém é justamente essa possibilidade de fazer comparações com outros textos que ajuda na detecção da utilização de uma IA para produzir tal texto. É sabido que essas ferramentas que geram textos de forma automática baseiam-se na produção do seu texto a uma vasta base de dados que possui uma quantidade imensurável de textos. Ou seja, IA’s de produção de textos, nada mais são do que ferramentas treinadas para produzir respostas baseadas em conhecimentos que foram previamente expostos a elas durante o treinamento de aprendizado de máquina.

O funcionamento desses detectores, em sua grande maioria, se baseia na exposição de um texto de entrada para que seja feita a análise e posteriormente seja apresentada uma saída que traga como resposta a informação de que o texto foi ou não uma produção de um modelo de processamento de linguagem natural. A resposta produzida varia entre cada um dos detectores, alguns trazem apenas uma resposta simples e objetiva com uma confirmação de que o texto foi ou não produzido por uma IA, outras trazem não só a resposta como também a porcentagem da probabilidade que essa resposta esteja correta, outras apresentam gráficos de quantas palavras foram classificadas como “suspeitas”. Por mais que a forma das respostas sejam variáveis, o objetivo final é o mesmo.

Alguns exemplos desses detectores de conteúdo de IA são:

- AI Content Detector;
- AI Writing Check;
- AI Text Classifier;
- ZeroGPT;
- Writefull GPT.

2.4 Lista de Detectores de IA

Nessa seção, o objetivo é apresentar a lista de detectores de conteúdo de IA que foram abordados no estudo. A apresentação de características relevantes destes detectores é de suma importância para que se haja o entendimento do motivo deles terem entrado no estudo.

2.4.1 AI Content Detector (Copyleaks)

Essa ferramenta é totalmente online[1] e alega ter a capacidade de detectar se um texto foi produzido por uma inteligência artificial, por um humano ou se o referido texto é uma combinação entre IA e humano. É uma ferramenta de funcionamento simples, onde o usuário apenas insere o texto e clica em ‘verificar’, após isso a ferramenta apresenta sua resposta. A ferramenta não aceita textos com menos de 150 caracteres.[2]

2.4.2 AI Text Classifier

Ferramenta online[3] desenvolvida pela mesma desenvolvedora do ChatGPT. Ela tem por objetivo prever a probabilidade de um texto ter sido ou não gerado por uma inteligência artificial. Suas principais limitações são de aceitar apenas textos acima de 1.000 caracteres e como sua base de dados é, em sua grande maioria, textos em inglês, a análise de textos em outros idiomas pode ter mais propensão a falhar.[2]

2.4.3 AI Writing Check

É uma ferramenta que foi desenvolvida pela Quil.org em conjunto com a CommonLit.org. não necessita instalação pois é totalmente online[4]. O principal objetivo dessa ferramenta é auxiliar os professores que queiram averiguar se seus alunos se utilizaram de inteligência artificial na geração de textos que foram entregues por eles. A ferramenta traz sua limitação do tamanho do texto contada em palavras, sendo, nesse caso, o mínimo de 100 palavras e o máximo de 400

[1] Disponível em: https://copyleaks.com/ai-content-detector
[2] Todas as informações foram coletadas nas interfaces das ferramentas conforme foram sendo testadas
[3] Disponível em: https://platform.openai.com/ai-text-classifier
[4] Disponível em: https://aiwritingcheck.org/

palavras. Não traz com precisão se existe um número limite de caracteres que é aceito.[5]

2.4.4 Writefull GPT

É uma ferramenta de uso online[6] que originalmente foi desenvolvida para a detecção de plágios em textos, porém também pode ser utilizada para identificar se um conteúdo foi gerado ou não pelo ChatGPT. Não existe indicação de limitação do tamanho do texto que será verificado pela ferramenta.[5]

2.4.5 ZeroGPT

Essa ferramenta de uso online[7] foi desenvolvida especificamente com o intuito de ser capaz de classificar se o texto checado pela mesma foi produzido por alguma ferramenta de produção de textos do ChatGPT. Essa ferramenta não determina com quantos caracteres mínimos ela consegue fazer essa checagem.[5]

2.4.6 Características dos detectores de IA

A Tabela 1, a seguir, exibe um resumo das características mais significativas das ferramentas envolvidas na pesquisa. Estas características foram coletadas a partir dos testes que foram realizados e oferecem uma compreensão profunda sobre o funcionamento das ferramentas e também destacam algumas das principais restrições relacionadas aos tipos de entradas que elas suportam:

Tabela 1 – Características dos Detectores de IA

Ferramenta	Ano	Suporta			Pago	Limitações
		GPT-2	GPT-3	ChatGPT		
AI Content Detector (Copyleaks)	2023	Não	Sim	Sim	Sim	Acima de 150 caracteres
AI Text Classifier	2023	Não	Não	Sim	Não	Acima de 1000 caracteres

[5] Todas as informações foram coletadas nas interfaces das ferramentas conforme foram sendo testadas
[6] Disponível em: https://x.writefull.com/gpt-detector
[7] Disponível em: https://www.zerogpt.com/

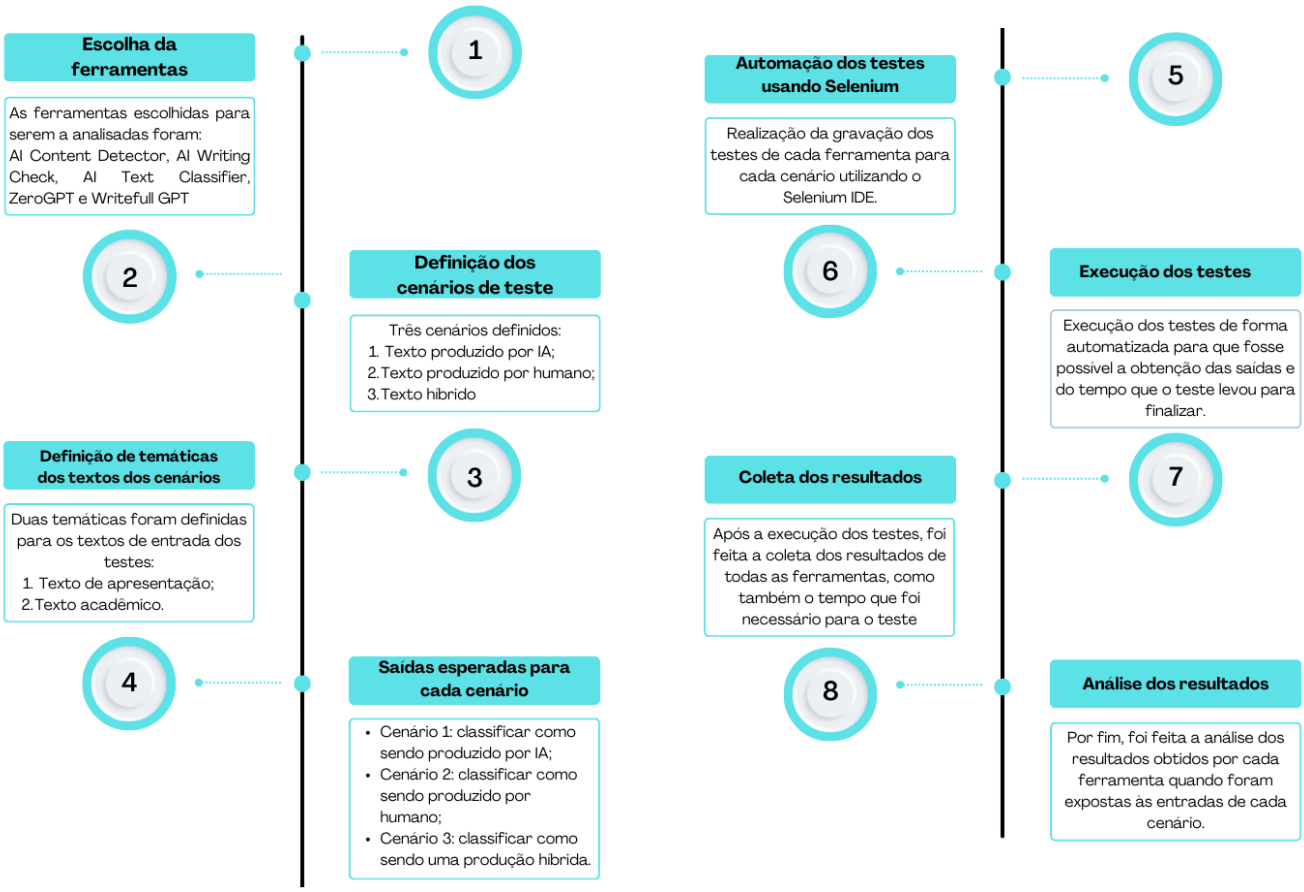
AI Writing Check	2023	Não	Não	Sim	Não	Entre 100 e 400 palavras
Writefull GPT	2023	Não	Sim	Sim	Sim	Sem informações
ZeroGPT	2023	Não	Não	Sim	Sim	Sem informações

Fonte: Elaborada pelo autor, 2023.

3 METODOLOGIA

Esta seção tem por objetivo explanar como o presente estudo foi conduzido. Será apresentado como cada etapa foi conduzida. A Figura 1 ilustra o processo de análise dinâmica que foi realizado para a obtenção dos resultados que foram analisados.

Figura 1 – Metodologia da análise dinâmica



Fonte: Elaborada pelo autor, 2023.

3.1 Definição dos Cenários de Teste

Todas as ferramentas que foram analisadas de forma dinâmica nesse estudo têm por característica receber um texto como entrada, processar essa entrada e produzir a sua resposta. Desse modo, é importante que seja feita a definição de cenários de testes que serão utilizados na condução da análise dinâmica dessas ferramentas.

Os cenários definidos para serem expostos e analisados pelas ferramentas foram três: o primeiro foi um texto completamente produzido pelo ChatGPT, versão 3.5; o segundo foi o inverso, o texto que foi produzido por um ser humano, de forma totalmente autoral, sem qualquer auxílio; e o terceiro foi a junção dos cenários anteriores, um texto que foi produzido por um ser humano e que teve trechos alterados por uma ferramenta de produção de textos baseada em inteligência artificial.

Cada cenário foi subdividido em mais dois. Essa nova divisão dentro dos cenários já mencionados diz respeito à temática do conteúdo do texto: o primeiro tema foi sobre uma apresentação de uma pessoa, ou seja, um texto com um teor mais informal; e o segundo texto abordou sobre algum assunto acadêmico. O objetivo foi verificar como a detecção dessas ferramentas se comporta com textos com estruturas tão distintas.

Outro ponto que foi definido nos referidos cenários é que o idioma utilizado em todos foi o inglês. Como o objetivo do estudo não se baseia em fazer comparativos da eficácia de cada ferramenta ante diferentes idiomas e o idioma principal de todas as ferramentas serem efetivamente o inglês, tal decisão pelo idioma inglês traz uma paridade entre elas na hora da análise.

Os quadros 1, 2 e 3, a seguir, detalham cada cenário descrito acima, bem como as entradas inseridas e as saídas esperadas:

Quadro 1 – Cenário 1 (Texto elaborado por uma IA)

Entrada 1 (Texto de Acadêmico)	Albert Einstein's theory of relativity revolutionized our understanding of the universe. The theory consists of two main pillars: the special theory of relativity, published in 1905, and the general theory of relativity, formulated in 1915. The special theory of relativity introduced the concept that the laws of physics remain the same for all observers in uniform motion, regardless of their relative velocity. It also revealed that the speed of light is constant and acts as a fundamental limit in the universe. On the other hand, the general theory of relativity expanded upon this by proposing that gravity arises from the curvature of spacetime caused by mass and energy. This groundbreaking theory provides a deeper understanding of the nature of space, time, and gravity, and its principles have been confirmed by numerous experiments and observations. The theory of relativity continues to shape our knowledge of the cosmos and has had profound implications for modern physics and our understanding of the universe at both the macroscopic and microscopic scales.
Entrada 2 (Texto de Apresentação)	Hello! I'm Emily, a passionate individual who finds joy in both the arts and sciences. The elegance and beauty of mathematics, physics, and computer science have captivated me from a young age. I chose to pursue applied mathematics and computational modeling, immersing myself in the complexities of problem-solving and discovering innovative solutions. Beyond the analytical realm, I also have a deep appreciation for artistic expression. Exploring art galleries, attending live performances, and engaging in creative writing provide me with inspiration and a means of personal expression. Nature is another source of inspiration for me, and I enjoy immersing myself in its serenity through activities such as hiking, camping, and photography. Collaboration and teamwork are integral to my approach, as I believe that diverse perspectives fuel creativity and growth. I am committed to lifelong learning and consistently seek out opportunities to expand my knowledge and skills through workshops, conferences, and online courses. Ultimately, my goal is to bridge the gap between science and creativity, harnessing their combined power to make a positive impact on the world. Thank you for taking the time to get to know me!
Saída Esperada	Classificar textos como sendo conteúdo produzido por uma inteligência artificial

Fonte: Elaborada pelo autor, 2023.

Quadro 2 – Cenário 2 (Texto elaborado por humano)

Entrada 1 (Texto de Acadêmico)	When Albert Einstein’s theory of relativity had published, only 12 person of that era could understand this theory. One day, a young journalist asked Einstein to explain his theory. Then he explained his theory with a joke. He told that, “When a man converse story with a beautiful girl for one hour, it seems to him that it’s past only one minute. And if anyone stand on stove for a minute, it seems to him that he is stand here for one hour. This is relativity”. Albert Einstein's theory of relativity is actually two separate theories: his special theory of relativity, postulated in the 1905 paper, The Electrodynamics of Moving Bodies and his theory of general relativity, an expansion of the earlier theory, published as The Foundation of the General Theory of Relativity in 1916. Einstein sought to explain situations in which Newtonian physics might fail to deal successfully with phenomena, and in so doing proposed revolutionary changes in human concepts of time, space and gravity. The special theory of relativity was based on two main postulates: first, that the speed of light is constant for all observers; and second, that observers moving at constant speeds should be subject to the same physical laws.
Entrada 2 (Texto de Apresentação)	Hello, my name is Clearlison Costa. I'm 29 years old. I am currently a student of the computer science course at the State University of Paraíba. I'm in the last period of the course and I'm writing my final thesis. The theme of my work will be on artificial intelligence content detectors, I want to see if they are really efficient in knowing whether or not a text was written with the aid of artificial intelligence. Despite not having finished the course, I already work in the area, I am currently a frontend developer at a company in Santa Catarina, working on the development of an online platform aimed at the financial sector. Previously, I worked as a backend developer on an API aimed at port control at the Port of Santos. Both jobs were performed remotely. The main programming languages that I work with or have worked with are PHP, JavaScript and TypeScript. I intend to learn new programming languages in the future, starting with Python. My objective is to always acquire more and more knowledge, our area evolves much faster than any other, we cannot accommodate ourselves.
Saída Esperada	Classificar textos como sendo conteúdo produzido por um ser humano de forma totalmente autoral.

Fonte: Elaborada pelo autor, 2023.

Quadro 3 – Cenário 3 (Texto Híbrido)

Entrada 1 (Texto de Acadêmico)	When Albert Einstein's theory of relativity was published, only a select few individuals in that era possessed the capacity to comprehend its intricacies. One memorable occasion arose when a young journalist approached Einstein, seeking an explanation of his theory. In response, Einstein resorted to the use of humor as a pedagogical tool. He shared a thought-provoking anecdote, stating, "When a man engages in conversation with a captivating woman for one hour, it appears to him as if only a single minute has elapsed. Conversely, if someone were to stand upon a hot stove for a mere minute, it would feel as though an entire hour had passed. This, my friend, is relativity.". Albert Einstein's theory of relativity encompasses two distinct propositions: his special theory of relativity, initially proposed in his 1905 publication titled "The Electrodynamics of Moving Bodies," and his general theory of relativity, an expansion of the former theory expounded upon in his 1916 work titled "The Foundation of the General Theory of Relativity.". Einstein sought to explain situations in which Newtonian physics might fail to deal successfully with phenomena, and in so doing proposed revolutionary changes in human concepts of time, space and gravity. The special theory of relativity was based on two main postulates: first, that the speed of light is constant for all observers; and second, that observers moving at constant speeds should be subject to the same physical laws.
Entrada 2 (Texto de Apresentação)	Greetings, allow me to introduce myself formally. My name is Clearlison Costa, and I am a 29-year-old individual. Presently, I am enrolled as a student in the computer science program at the esteemed State University of Paraíba. I am currently in the final stage of my academic journey, actively engaged in the composition of my final thesis. The central focus of my research revolves around the efficacy of artificial intelligence content detectors. I aim to investigate their ability to discern whether a given text has been generated with the assistance of artificial intelligence. Despite not having completed my studies, I am already employed in the field. Specifically, I serve as a frontend developer for a reputable company based in Santa Catarina. In this role, I am actively involved in the development of an online platform tailored for the financial sector. Previously, I worked as a backend developer on an API aimed at port control at the Port of Santos. Both jobs were performed remotely. The main programming languages that I work with or have worked with are PHP, JavaScript and TypeScript. I intend to learn new programming languages in the future, starting with Python. My objective is to always acquire more and more knowledge, our area evolves much faster than any other, we cannot accommodate ourselves.
Saída Esperada	Identificar que os textos possuem parte deles elaborada com auxilio da inteligência artificial.

Fonte: Elaborada pelo autor, 2023

3.2 Execução dos testes

Para a coleta e análise dos resultados produzidos por cada ferramenta submetida ao referido estudo, foi utilizada a extensão do Selenium IDE para o

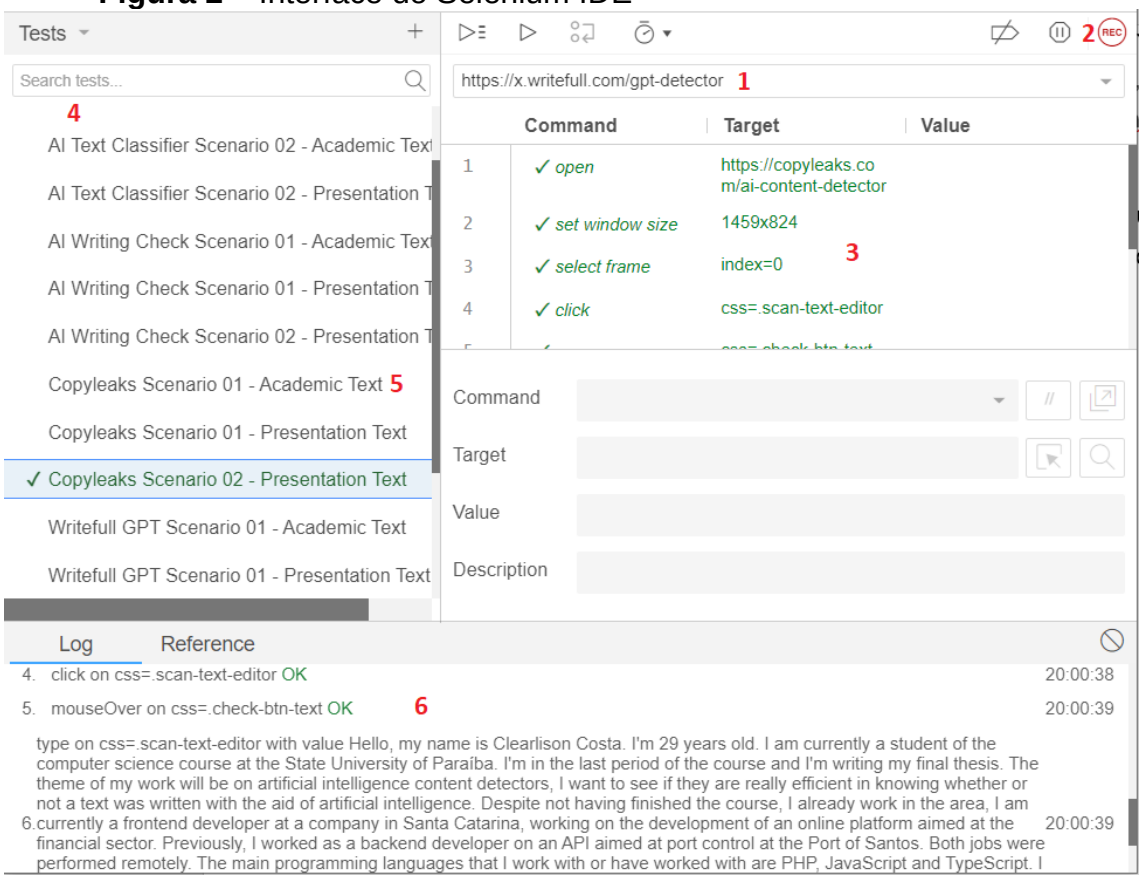
navegador Google Chrome. A automação dos testes realizada dessa maneira permitiu que a execução dos mesmos fosse feita inúmeras vezes sem a necessidade de refazer todos desde o início novamente, otimizando o tempo da coleta dos resultados, além de permitir verificar o tempo decorrido que cada ferramenta levou para trazer a resposta para cada cenário a qual foram expostas.

Todos os detectores de conteúdo de IA tiveram seus testes automatizados da mesma maneira. Foi instalada a extensão do Selenium IDE para Google Chrome, no referido browser. Foi criado um novo projeto dentro do Selenium e para cada cenário de cada ferramenta foi criado um novo teste automatizado.

O processo de automação dos testes foi bastante dinâmico. Foi introduzido a URL da ferramenta a ser testada dentro do campo “Playback Base URL”, logo após isso, clica-se no botão “REC” para iniciar a “gravação” do teste, a partir desse momento, uma nova janela do navegador foi aberta carregando a página da ferramenta. Com a página já carregada, fez-se a interação com a ferramenta, testando o cenário pela primeira vez. Quando o cenário já foi testado e a ferramenta já forneceu a resposta, retornou-se a interface do Selenium IDE e ocorreu mais um clique no botão de gravação, agora para que a mesma seja finalizada. Todo esse processo deixou a interação com a ferramenta mapeada em um caso de teste, o qual posso executar quantas vezes se achar necessário, e todas elas fazendo exatamente o mesmo passo-a-passo que foi feito manualmente.

Para uma melhor compreensão do processo de automatização que foi explanado anteriormente, a Figura 2 traz uma captura de tela que traz a interface do Selenium IDE com alguns pontos destacados de forma ordenada, que ilustram o processo de automação dos testes.

Figura 2 – Interface do Selenium IDE



Fonte: Elaborada pelo autor, 2023.

A seguir, pode-se observar a representação dos pontos relevantes que foram enumerados na imagem acima, com objetivo de fornecer uma explicação um pouco mais detalhada de cada elemento destacado na interface do Selenium:

1. Campo “Playback Base URL”, no qual é passada a URL da ferramenta online que terá seu funcionamento automatizado pelo Selenium;
2. Botão “REC” que tem por finalidade iniciar e finalizar a “gravação” da interação do usuário com a ferramenta;
3. Lista de comandos que foram mapeados na gravação do teste e que serão reproduzidos sempre que o teste seja realizado;
4. Lista de testes que foram gravados para cada ferramenta e cenários de testes que foram abordados no estudo;
5. Nome de um teste automatizado que foi realizado no estudo;
6. Aba que mostra o “Log” do teste, onde é possível visualizar se o teste foi realizado com sucesso e o tempo que levou para o teste ser realizado.

Para que fosse possível identificar qual ferramenta de detecção e a qual cenário estava sendo testado no momento, cada cenário de teste produzido para o estudo seguiu um padrão de nomenclatura – “<nome da ferramenta> scenario <número do cenário> - <tema do texto>”.

A coleta dos dados para cada cenário se deu através da execução de todos os cenários de testes de forma automatizada, clicando no botão de “play” dentro do Selenium IDE e, ao fim da execução, verificando qual foi a resposta que a ferramenta produziu diante do cenário. Para verificar o tempo que cada ferramenta gastou para expor sua resposta, foi observado a diferença entre o horário de início de teste e o horário que o mesmo finalizou com sucesso, essa informação consta na aba de “log” que a interface do Selenium IDE disponibiliza.

3.2.1 Método de classificação dos resultados

Para a obtenção das taxas de acerto e erro nos gráficos apresentados, foi adotado um sistema de pontuação que atribui diferentes pesos a cada resultado. Nesse sistema, um acerto completo é valorizado com peso 1, indicando uma correspondência exata entre a previsão da ferramenta e a realidade. Já um acerto parcial é considerado com peso 0.5, reconhecendo a proximidade da previsão, embora com algumas divergências. Por fim, um erro recebe peso 0, significando que a previsão não corresponde ao resultado esperado.

.

4 RESULTADOS DOS TESTES

Nesta seção, apresentaremos os resultados obtidos por meio da realização dos testes automatizados nas ferramentas de detecção de conteúdo de IA. O objetivo principal deste estudo foi avaliar a eficiência dessas ferramentas ao serem expostas aos cenários previamente delineados. Como mencionado anteriormente, esses cenários foram cuidadosamente selecionados para representar situações relevantes e desafiadoras que essas ferramentas podem encontrar no contexto real.

4.1 Resultados dos testes por cenário

A seguir, pode-se observar os resultados obtidos para cada uma das ferramentas testadas, incluindo a data em que os testes foram realizados e o tempo de resposta de cada ferramenta. As informações contidas nas tabelas 2, 3 e 4 fornecem uma visão geral dos desempenhos individuais das ferramentas diante dos cenários propostos. Os resultados foram divididos em três tabelas com intuito de trazer mais organização e entendimento. Cada tabela detalha os resultados de um cenário.

De modo geral, para o cenário 1, cujo o texto foi elaborado pelo ChatGPT, as ferramentas apresentaram mais acertos quando a temática do texto foi o conteúdo acadêmico. No texto de apresentação, apenas duas ferramentas conseguiram acertos. Quanto ao tempo que os testes levaram para ser executados variou entre 3 e 10 segundos. A Tabela 2 apresenta mais detalhes sobre os resultados que foram obtidos.

Tabela 2 – Resultados do cenário 1 (Texto elaborado por IA)

Ferramenta	Conteúdo	Data do Teste	Tempo	Resposta	Acerto
AI Content Detector (Copyleaks)	Acadêmico	03/07/2023	9s	99.4% probability for AI	Acertou
	Apresentação	03/07/2023	4s	75.0% probability for AI	Acertou
AI Text Classifier	Acadêmico	03/07/2023	4s	unclear if it is AI-generated	Errou
	Apresentação	03/07/2023	5s	unlikely AI-generated	Errou
AI Writing Check	Acadêmico	03/07/2023	3s	Text Written by AI	Acertou
	Apresentação	03/07/2023	3s	Text Written by Human	Errou
Writefull GPT	Acadêmico	03/07/2023	10s	22% likely this comes from AI	Errou
	Apresentação	04/07/2023	7s	1% likely this comes from AI	Errou
ZeroGPT	Acadêmico	04/07/2023	6s	100% AI GPT - Your Text is AI/GPT Generated	Acertou
	Apresentação	04/07/2023	4s	92.43% AI GPT- Your Text is Most Likely AI/GPT generated	Acertou

Fonte: Elaborada pelo autor, 2023.

Seguindo a mesma estrutura da tabela anterior, a Tabela 3 traz os erros e acertos de cada ferramenta quando foram expostas aos textos do cenário 2, os quais representam os textos elaborados por um humano. Este foi o cenário onde as ferramentas obtiveram mais acertos, tendo sido observado apenas um erro no texto de apresentação, quando o mesmo foi exposto a ferramenta “AI Text Classifier”. O tempo gasto para a resposta das ferramentas serem apresentadas variou entre 3 i 8 segundos, bastante semelhante ao primeiro cenário.

Tabela 3 – Resultados do cenário 2 (Texto elaborado por humano)

Ferramenta	Conteúdo	Data do Teste	Tempo	Resposta	Acerto
AI Content Detector (Copyleaks)	Acadêmico	03/07/2023	5s	99.9% probability for Human (This is human text)	Acertou
	Apresentação	03/07/2023	4s	63.7% probability for Human (This is human text)	Acertou
AI Text Classifier	Acadêmico	03/07/2023	4s	very unlikely AI-generated	Acertou
	Apresentação	03/07/2023	4s	unclear if it is AI-generated	Errou
AI Writing Check	Acadêmico	03/07/2023	2s	Text Written by Human	Acertou
	Apresentação	03/07/2023	3s	Text Written by Human	Acertou
Writefull GPT	Acadêmico	04/07/2023	8s	1% likely this comes from AI	Acertou
	Apresentação	04/07/2023	8s	2% likely this comes from AI	Acertou
ZeroGPT	Acadêmico	04/07/2023	4s	51.95% AI GPT - Your Text is Most Likely Human written, may include parts generated by AI/GPT	Acertou Parcialmente
	Apresentação	04/07/2023	5s	0% AI GPT - Your Text is Human written	Acertou

Fonte: Elaborada pelo autor, 2023.

Por fim, a Tabela 4 representa os resultados de cada ferramenta quando foram expostas aos textos que foi elaborado por um ser humano, mas que teve partes deles modificadas pela inteligência artificial, ou seja, textos híbridos. Como pode ser observado na tabela, esse foi o cenário onde houve uma grande taxa de

acertos parciais, uma vez que, apesar das respostas apresentadas pelas ferramentas trazerem na maioria das vezes análises que se aproximam do acerto, elas não são tão conclusivas quanto ao texto ser híbrido. Quanto ao número de erros, obteve a mesma quantidade (4) que o primeiro cenário. Quanto ao tempo gasto na realização desses testes, a variação ficou entre 5 e 8 segundos.

Tabela 4 – Resultados do cenário 3 (Texto híbrido)

Ferramenta	Conteúdo	Data do Teste	Tempo	Resposta	Acerto
AI Content Detector (Copyleaks)	Acadêmico	03/07/2023	5s	47.9% probability for Human (This is human text)	Acertou
	Apresentação	03/07/2023	5s	76.4% probability for Human (This is human text)	Acertou Parcialmente
AI Text Classifier	Acadêmico	03/07/2023	5s	very unlikely AI-generated	Errou
	Apresentação	03/07/2023	6s	unclear if it is AI-generated	Errou
AI Writing Check	Acadêmico	03/07/2023	3s	Text Written by Human	Acertou Parcialmente
	Apresentação	03/07/2023	2s	Text Written by Human	Acertou Parcialmente
Writefull GPT	Acadêmico	04/07/2023	8s	1% likely this comes from AI	Errou
	Apresentação	04/07/2023	6s	5% likely this comes from AI	Errou
ZeroGPT	Acadêmico	04/07/2023	6s	80.98% AI GPT-Your Text is Most Likely AI/GPT generated	Acertou Parcialmente
	Apresentação	04/07/2023	5s	14.44% AI GPT - Your Text is Human written	Acertou Parcialmente

Fonte: Elaborada pelo autor, 2023.

4.2 Análise dos resultados

Nesta seção de análise dos resultados, a avaliação está representada de duas maneiras distintas. Inicialmente, a análise dos resultados foi separada por ferramentas de forma individual. Dessa maneira, foi possível apresentar uma avaliação detalhada de seu desempenho nos diferentes cenários propostos. Essa abordagem proporcionou uma compreensão detalhada e específica sobre o comportamento de cada ferramenta diante dos desafios apresentados, destacando seus pontos fortes e limitações em cenários específicos.

Posteriormente, foi realizada uma análise do espectro geral, considerando os resultados consolidados das cinco ferramentas, que foram testadas no presente estudo, em todos os cenários. Essa avaliação holística permitiu uma visão abrangente do desempenho coletivo das ferramentas de detecção de conteúdo de IA. Desse modo, foi possível identificar padrões e tendências comuns em relação aos acertos e erros. A análise geral também proporcionou uma compreensão mais ampla sobre a eficiência dessas tecnologias em diferentes contextos, subsidiando considerações relevantes para sua aplicabilidade e possíveis direcionamentos futuros.

4.2.1 Análise por ferramenta

Iniciando a análise pela ferramenta ‘AI Content Detector (Copyleaks)’, dentre as cinco ferramentas, foi uma das que apresentou melhor desempenho, conseguiu identificar corretamente tanto os textos do cenário 1 (elaborados por IA) quanto os do cenário 2 (texto elaborado por um humano), não importando a temática do texto. Já no terceiro cenário, a ferramenta acertou quando se tratava de um texto acadêmico, ela conseguiu identificar que o texto foi escrito por um humano, mas que provavelmente também considerou uma boa chance de haver texto produzido por IA. No caso do texto híbrido com a temática de apresentação, foi considerado um acerto parcial, pois a ferramenta considerou o texto como sendo de um texto escrito por um humano, mas com uma probabilidade que foi considerada muito alta (76.5%), uma vez que a parte modificada pela inteligência artificial foi cerca da metade do texto, e que acabou denotando uma dificuldade em perceber as partes que foram melhoradas pela IA.

A segunda ferramenta que foi submetida aos mesmos cenários foi a ‘AI Text Classifier’. Essa apresentou um desempenho muito ruim, tendo acertado apenas em um dos cenários, e ainda assim apenas o texto de apresentação. O cenário onde ocorreu o acerto foi no texto escrito por um humano, nesse caso, a ferramenta identificou que a possibilidade de o texto ter sido produzido por uma ferramenta de IA era praticamente zero. Para os demais textos, a ferramenta trouxe respostas inconclusivas sobre o texto ter sido gerado por uma inteligência artificial e/ou descartou a presença mesmo quando havia trechos ou um texto completamente gerado por IA. Nesse caso, respostas tão ambíguas não podem ser consideradas acertos parciais, pois não trouxe nenhuma ajuda quanto a conclusão sobre o texto que o usuário submeteu a testes.

A ‘AI Writing Check’ foi a terceira ferramenta que foi exposta aos cenários de testes. Quanto aos resultados, para os textos produzidos por IA, a ferramenta conseguiu classificar corretamente quando o conteúdo era o texto com teor acadêmico, já o texto de apresentação, ela classificou incorretamente como se fora um texto escrito sem auxílio de IA. Para o cenário de textos escritos por humano, a ferramenta acertou para ambos os tipos de conteúdo. Já para os textos mesclados, nas duas situações expostas, considerei como acertos parciais, uma vez que, a ferramenta não traz uma resposta mais clara para esse caso, as duas respostas possíveis é se é um texto de IA ou de humano. Nesse caso, a ferramenta considerou para os dois tipos de conteúdo como sendo textos escritos por um humano, o que não deixa de ser verdade, porém também há traços modificados por IA, e sinto a falta da ferramenta ter uma resposta que possa considerar esse ponto.

A quarta ferramenta do estudo foi a ‘Writefull GPT’. Essa ferramenta, comparadas com as demais, teve um desempenho muito ruim quando exposta aos cenários de teste. Para o cenário de texto de IA, a ferramenta deu um índice de 22% do texto seja proveniente de uma inteligência artificial, é uma percentagem muito baixa sabendo que o texto é sim uma produção de uma IA. No caso do texto com conteúdo de apresentação, o resultado foi ainda pior, considerando apenas 1% do conteúdo como sendo produção de uma IA. No caso dos textos escritos por humano, a ferramenta acertou para os dois tipos de conteúdo (acadêmico e apresentação) considerando como apenas 1% e 2% respectivamente vindo de uma IA. Para o terceiro cenário, mais dois erros, considerou apenas 1% do texto

acadêmico e 5% do texto de apresentação provenientes de uma IA, e é uma taxa muito baixa pois uma parte considerável do texto foi alterada por uma IA. Diante desses resultados, até os acertos dessa ferramenta geram dúvidas, pois foram justamente no cenário de texto de humanos. Como em todos os testes a ferramenta deu respostas com taxas muito baixas de que o texto continha conteúdo gerado por IA, os acertos no segundo cenário podem dar a entender que foram mais pela incapacidade dessa ferramenta de identificar conteúdos gerados por IA.

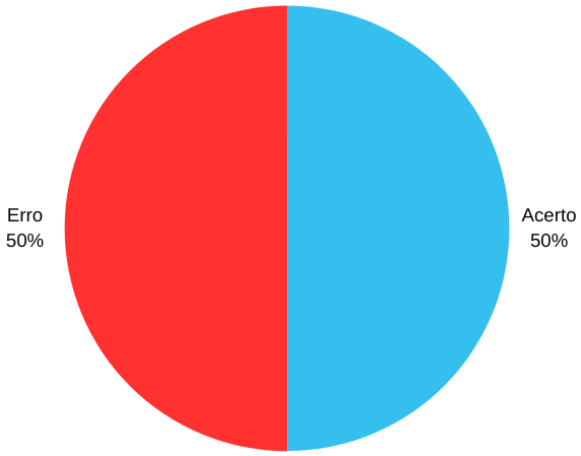
Por fim, a quinta e última ferramenta testada foi a ‘ZeroGPT’. Essa apresentou bons resultados diante dos cenários a qual foi submetida. Para o primeiro cenário, a ferramenta conseguiu identificar sem dificuldades que, tanto o texto acadêmico quanto o de apresentação, eram produções oriundas de uma inteligência artificial, dando como taxas de 100% e 92,43%, respectivamente. Para o segundo cenário, a ferramenta não teve dificuldades em identificar que o texto era escrito por um humano quando o conteúdo era o de apresentação, com uma taxa de 0% de conteúdo de IA. Para o texto acadêmico ocorreu um acerto parcial, a ferramenta considerou ser um texto produzido por um humano, mas que haveria partes de IA, isso se deu muito provavelmente pelo fato de a ferramenta ter textos semelhantes ao trecho utilizado no teste. Para o terceiro cenário, os textos mesclados, a ferramenta acertou de forma parcial para os dois tipos de conteúdo, para o acadêmico, ela considerou que pouco mais de 80% do texto era vindo de uma IA. sabendo que uma parte considerável do texto realmente foi modificado por uma IA, não dá pra considerar como um erro essa resposta, mesmo que a taxa seja um pouco mais alta que o esperado. Para o texto de apresentação houve um comportamento semelhante, porém dando uma porcentagem um pouco abaixo do esperado, mas que não considerei que fosse um erro total.

4.2.2 Análise geral dos resultados

Os gráficos apresentados nas Figuras 2 a 7 mostram os resultados gerais das ferramentas, desse modo foi possível fazer uma análise geral da eficiência das cinco ferramentas que foram escolhidas para o estudo. Desse modo, foi possível identificar em quais cenários as ferramentas foram mais eficientes, em quais tipos de conteúdo dos textos, bem como uma taxa geral de acertos, levando em consideração todos os resultados de todos os testes somados.

Iniciando a análise pelo cenário de textos que foram elaborados por IA, as ferramentas juntas conseguiram uma taxa de acerto de 50%, como pode ser visto na Figura 3. Foi notório que a maioria das ferramentas tiveram problemas para conseguir identificar o texto de apresentação produzido pela IA, como as mesmas se baseiam em dados que estão alocados em sua base, textos com teor mais informal certamente trazem mais dificuldades de classificação, uma vez que, as ferramentas buscam por semelhanças de texto em sua base, como também padrões de escrita e um texto com um teor mais informal pode facilmente burlar essas análises comparativas.

Figura 3 – Taxa de acertos para textos elaborados por inteligência artificial

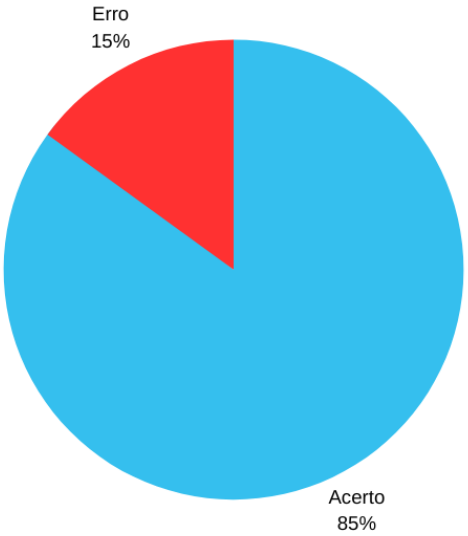


Fonte: Elaborada pelo autor, 2023.

Como pode ser observado na Figura 4, os textos elaborados por humano foi onde as ferramentas conseguiram suas maiores taxas de acerto. Levando em consideração os resultados somados, a taxa de acerto para esse cenário foi de 85%, bastante alta. Inversamente ao cenário 1, os textos com conteúdo mais informal foram os maiores responsáveis por essa taxa de acerto elevada. Uma possibilidade seja justamente o que foi explicado no parágrafo anterior, a dificuldade das ferramentas de conseguirem encontrar semelhanças em suas bases de dados com o texto que fora exposto no teste. Isso acaba dando margem para acreditar que esse nível de acerto seja apenas o fato de as ferramentas não encontrarem nada semelhante em suas bases e assim classificar como sendo textos não produzidos por uma IA. Outra prova disso é o fato de os textos de conteúdo acadêmico terem sido os maiores responsáveis pelos erros nesse cenário, justamente por se tratar de um texto que possui uma melhor estruturação, bem como um assunto que

certamente está contido em textos que estão inseridos nas bases de dados dessas ferramentas.

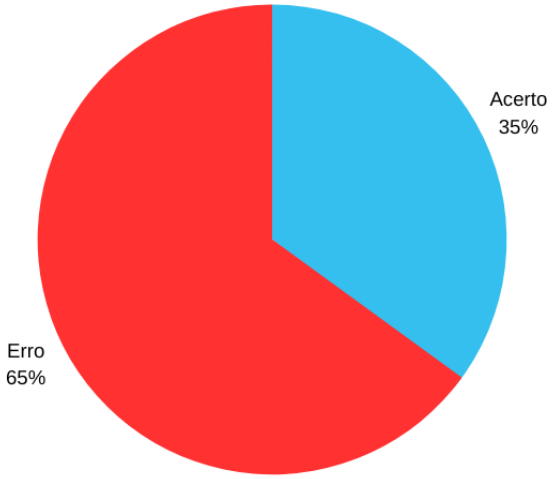
Figura 4 – Taxa de acertos para textos elaborados por humano



Fonte: Elaborada pelo autor, 2023.

Já o terceiro e último cenário, quando se trata de textos híbrido, representado na Figura 5. A taxa de acerto foi a menor de todas. Somados todos os resultados, as ferramentas tiveram uma taxa de apenas 35% de acertos. Isso se deu principalmente pela falta de clareza das respostas que foram obtidas diante desse cenário. O fato de a ferramenta apenas expressar apenas se um texto foi ou não produzido por uma inteligência artificial, para esse cenário, não é o ideal. Uma vez que isso produz mais ambiguidade que uma resposta que realmente auxilie na decisão do usuário final. O melhor caminho seria uma resposta onde a ferramenta argumentasse que o texto é uma produção híbrida.

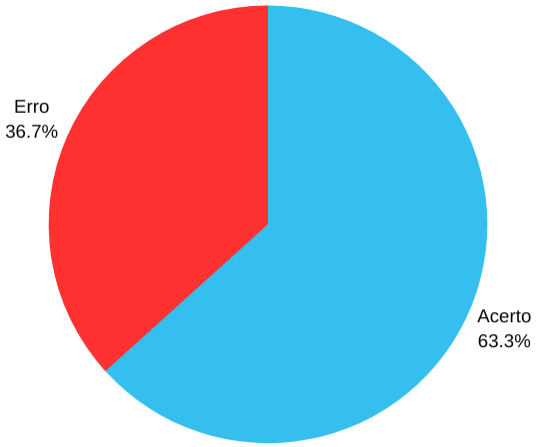
Figura 5 – Taxa de acertos para textos híbridos



Fonte: Elaborada pelo autor, 2023.

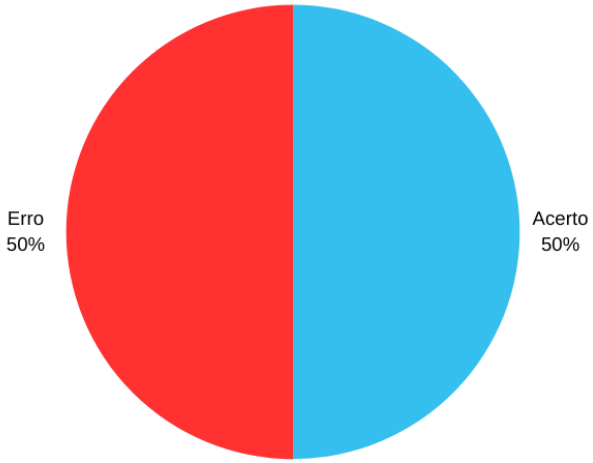
Quando separados por conteúdo dos textos, a taxa de acerto das ferramentas apresentou o resultado de 63,3% para textos acadêmicos e de 50% para textos de apresentação, vide Figuras 6 e 7. Esses resultados corroboram para o que mencionei anteriormente, as ferramentas conseguem ser mais eficientes quando o conteúdo dos textos é mais formal e bem mais estruturado, isso porque elas fazem justamente uma análise dessa estrutura, bem como uma comparação do conteúdo do texto em si com outros textos presentes em suas bases de dados. Desse modo, textos mais informais tendem a fazer com essas ferramentas baixem sensivelmente as suas eficiências.

Figura 6 – Taxa geral de acertos para texto acadêmico



Fonte: Elaborada pelo autor, 2023.

Figura 7 – Taxa de acertos para textos de apresentação



Fonte: Elaborada pelo autor, 2023.

Por fim, a taxa geral de acertos das ferramentas, representada na Figura 8, ficou em 56,7%. Esse resultado demonstra que na maioria dos testes que foram realizados no estudo, as ferramentas conseguiram responder corretamente, seja em sua completude ou de forma parcial. Porém, é possível perceber que ainda há muito espaço para que essas ferramentas possam evoluir em seus métodos de análise e consigam produzir resultados bem mais satisfatórios.

Figura 8 – Taxa geral de acertos

Fonte: Elaborada pelo autor, 2023.

5 CONCLUSÃO

Após a bateria de testes que foi realizada nas ferramentas que foram escolhidas para o presente estudo e da análise minuciosa dos resultados das mesmas para cada cenário a qual foram expostas. Levando em consideração todos os resultados obtidos e explanados com mais detalhes anteriormente no presente trabalho, foi notório que as ferramentas cumpriram seu papel de forma parcial quanto a identificação de tais conteúdos. Levando em consideração que a taxa de sucesso ainda está muito abaixo do ideal, para que essas ferramentas possam ser usadas para tais fins é necessário que elas apresentem uma maior evolução.

Atualmente essas ferramentas estão mais propícias a terem sucesso em seus testes quando expostas à textos com conteúdo um pouco mais acadêmico e temáticas que estão presentes em diversos outros artigos, trabalhos acadêmicos, reportagens e assim sucessivamente. Isso acaba dando impressão de que essas ferramentas são mais voltadas à detecção de plágio do que para a detecção de conteúdo produzido por IA. Obviamente que as inteligências artificiais, quando estão produzindo seus textos, se baseiam em dados previamente apresentados a elas e é isso que ajuda bastante as ferramentas a conseguirem identificar os conteúdos. Mas isso pode abrir brechas para que erros de classificação ocorram, uma vez que, é possível passar um trecho de algum artigo muito conhecido e a ferramenta acabar classificando como sendo produzido por uma inteligência artificial pelo mero fato de essa ferramenta possuir o mesmo artigo em sua base que será usada para fazer a classificação do texto.

Quando o texto produzido pela IA é mais informal, sem temáticas de assuntos amplamente discutidos em outros textos e em outros momentos, as ferramentas acabaram tendo bem mais dificuldades para acertar em suas classificações. Esse é outro ponto que traz precedentes para que ocorram erros de classificação, como foi visto anteriormente. O simples texto onde a IA se passa por alguém que está fazendo uma apresentação trouxe bastante dificuldades e erros de classificação nos testes realizados.

A conclusão alcançada neste estudo, foi que, as ferramentas de detecção serão aliadas para que seja possível discernir entre textos produzidos por humanos e textos produzidos por uma inteligência artificial. Porém, olhando para os dados

aqui detalhados, é nítido que tais ferramentas ainda precisam evoluir muito para que consigam trazer maior confiabilidade em suas respostas. Como sabemos que a tecnologia de produção de conteúdo por inteligência artificial vem se aperfeiçoando cada vez mais e trazendo ainda mais a luz ao interesse pelo assunto, é fundamental que ferramentas com intuito de detectar esses conteúdos acompanhem essa evolução e ganhem mais confiabilidade quanto às suas classificações.

Outro ponto importante a ser destacado é o fato dos cenários de testes aos quais as ferramentas foram expostas tem um escopo limitado. Desse modo, abre-se a possibilidade de que, em cenários mais abrangentes, as ferramentas possam apresentar resultados distintos, e inevitavelmente conclusões igualmente distintas,

REFERÊNCIAS

AI Content Detector. Disponível em: <https://copyleaks.com/ai-content-detector>. Acesso em: 3 jul. 2023.

AI Text Classifier. Disponível em: <https://platform.openai.com/ai-text-classifier>. Acesso em: 3 jul. 2023.

AI Writing Check. Disponível em: <https://aiwritingcheck.org/>. Acesso em: 3 jul. 2023.

ALSULAIMANI, B.; ISLAM, A. **Impact of 4ir technology and its impact on the current deployment**. 2022. Disponível em: http://arxiv.org/abs/2209.01791. Acesso em 24 ago. 2023.

BRANDL, R.; ELLIS, C. **Survey: ChatGPT and AI ContentCan people tell the difference? Tooltester**, 8 mar. 2023. Disponível em: <https://www.tooltester.com/en/blog/chatgpt-survey-can-people-tell-the-difference/>. Acesso em: 24 ago. 2023

CARVALHO, I. **Testes Dinâmicos e Testes Estáticos**. Disponível em: <https://medium.com/@ingrid.carvalho.mo/testes-din%C3%A2micos-e-testes-est%C3%A1ticos-39be46de08ae>. Acesso em: 15 set. 2023.

CONTRI, M. E. **Selenium como uma ferramenta para testes de software**. Disponível em: <https://medium.com/@dudacontri65/selenium-como-uma-ferramenta-para-testes-de-software-22541584b960>. Acesso em: 15 set. 2023.

COUTO, Marlen. **Ataques ao STF, Terra plana e facada em Bolsonaro: veja como o ChatGPT pode potencializar a produção de fake news**. Disponível em: <https://oglobo.globo.com/blogs/sonar-a-escuta-das-redes/post/2023/03/ataques-ao-stf-terra-plana-e-facada-em-bolsonaro-veja-como-o-chatgpt-pode-potencializar-a-producao-de-fake-news.ghtml>. Acesso em: 10 jul. 2023.

Cypress. Disponível em: <http://cypress.io>. Acesso em: 5 out. 2023.

DENG, Jianyang; LIN, Yijia. The benefits and challenges of ChatGPT: An overview. **Frontiers in Computing and Intelligent Systems**, v. 2, n. 2, p. 81-83, 2022.

GARCÍA, B. et al. A survey of the Selenium ecosystem. **Electronics**, v. 9, n. 7, p. 1067, 2020.

HEIKKILÄ, M. Why detecting AI-generated text is so difficult (and what to do about it). **Technology review**, 7 fev. 2023. Disponível em: < https://www.technologyreview.com/2023/02/07/1067928/why-detecting-ai-generated-text-is-so-difficult-and-what-to-do-about-it>. Acesso em: 24 ago. 2023.

HELFSTEIN, C. **7 Detectores de Conteúdo IA Gratuitos: Conheça o Melhor! Niara**, 14 jun. 2023. Disponível em: <https://niara.ai/blog/deteccao-conteudo-ia/>. Acesso em: 5 set. 2023.

KIRCHNER, J.H.; AHMAD, Lama**;** AARONSON, Scott; LEIKE, Jan**. New AI classifier for indicating AI-written text**. Disponível em: <https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text>. Acesso em: 10 jul. 2023.

KRIGER, D. **Revolução tecnológica: importância e impacto na atividade humana**. Disponível em: <https://kenzie.com.br/blog/revolucao-tecnologica/>. Acesso em: 10 jul. 2023.

MCCOY, J. **AI Content Detection: Do you need it? (the truth)**. Disponível em: <https://contenthacker.com/ai-content-detection/>. Acesso em: 15 set. 2023.

OpenAI. Disponível em: <https://openai.com/>. Acesso em: 5 out. 2023.

O'Reilly 2023 tech trends report reveals growing interest in artificial intelligence topics, driven by generative AI advancement. Disponível em: <https://www.oreilly.com/pub/pr/3405>. Acesso em: 24 ago. 2023.

PINHO, C. M. DE A. et al. Identificação de deficiências em textos educacionais com a aplicação de processamento de linguagem natural e aprendizado de máquina. **ETD - Educação Temática Digital**, v. 24, n. 2, p. 350–372, 2022.

PRESSMAN, Roger S.; MAXIM, Bruce R. **Engenharia de Software: uma abordagem profissional**. 8. ed. Porto Alegre: AMGH, 2016.

Ranorex. Disponível em: <https://www.ranorex.com/>. Acesso em: 15 set. 2023.

Selenium. Disponível em: <https://www.selenium.dev/>. Acesso em: 15 set. 2023.

SOMMERVILLE, Ian. **Engenharia de Software**. 9. ed. São Paulo: Pearson Prentice Hall, 2011.

TestComplete. Disponível em: <https://smartbear.com/product/testcomplete/>. Acesso em: 15 set. 2023.

Writefull GPT. Disponível em: <https://x.writefull.com/gpt-detector>. Acesso em: 4 jul. 2023.

ZeroGPT. Disponível em: <https://www.zerogpt.com/>. Acesso em: 4 jul. 2023.