

Eficiência de softwares nacionais e internacionais na detecção de similaridade e de plágio em manuscrito

Marcelo Krokosz

Fundação Escola de Comércio Álvares Penteado, São Paulo, SP, Brasil;
marcelok@fecap.br; <https://orcid.org/0000-0002-6869-864X>

Resumo: Este estudo visa identificar a eficiência de quatorze softwares de detecção de similaridades em um texto com fragmentos sobre educação financeira, encontrados em websites da internet, anais de eventos acadêmicos e revistas científicas de acesso aberto e restrito. Os fragmentos foram usados para elaborar um texto coeso, escrito em língua portuguesa, contendo reproduções textuais literais, paráfrases, com trechos traduzidos do idioma inglês, manipulados com inserção de caractere oculto e com substituição de palavras por sinônimos. Os relatórios de similaridade gerados pelos softwares foram analisados de acordo com quatro critérios: 1 identificação de fragmentos citados corretamente; 2 identificação de fragmentos plagiados; 3 identificação de textos manipulados para enganar o software; 4 identificação de fontes originais; os quais foram pontuados em uma escala de 0 a 4 pontos. Os softwares *Turnitin*, *StrikePlagiarism*, *PlagScan* e *Plagium* tiveram performance elevada e *CopySpider* e *Plagium* (complemento do Google) foram os mais ineficientes. Constatou-se que os softwares operam com eficiência variável, o que reforça a convicção de que embora sejam ferramentas úteis para a identificação de plágio, contribuem de forma limitada para aspectos como reconhecimento de reproduções textuais literais que nem sempre correspondem à fonte original. As fontes originais de textos reescritos, manipulados e traduzidos não foram encontradas por nenhum software. O estudo contribui para o aprimoramento da capacidade do usuário na escolha, uso e análise dos relatórios de similaridades gerados por softwares, cuja eficiência pode ser maior no caso da utilização de mais de um software.

Palavras-chave: plágio; similaridade e plágio; detecção de plágio; Software antiplágio

1 Introdução

As estimativas de ocorrência do plágio no meio acadêmico são alarmantes, desde a educação básica até em publicações científicas (AMOS, 2014; KOOCHER; KEITH-SPIEGEL, 2010; NOORDEN, 2012; RIBEIRO; VASCONCELOS, 2018).

Na área de engenharia e computação, são adotadas convenções que estabelecem níveis de classificação da gravidade de ocorrência do plágio. Pode acontecer por falha na identificação das fontes de algumas pequenas partes de texto reproduzidas (Nível 1) ou devido ao uso de paráfrases mal elaboradas sem atribuição correta das fontes utilizadas (Nível 2). Se a transcrição literal de textos de outra fonte corresponde a cerca de 20% do texto que está sendo redigido, já é considerada uma violação moderada (Nível 3). Uma violação significativa ocorre quando a reprodução textual indevida fica entre 20% e 50% do trabalho elaborado (Nível 4). O caso mais extremo, que caracteriza uma violação severa, é quando mais de 50% do texto ou, até mesmo, integralmente, foi reproduzido de uma fonte previamente publicada (Nível 5) (ASSOCIATION FOR COMPUTING MACHINERY (ACM), 2019; INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS (IEEE), 2021).

De forma geral, o advento e o acesso aos recursos tecnológicos têm sido apontados como um dos fatores facilitadores para o aumento da reprodução e apropriação de obras e conteúdos alheios, seja de forma acidental ou intencional. (BARBASTEFANO; SOUZA, 2007; BENNETT, 2005; CHRISTOFE, 1996; FACHINI; DOMINGUES, 2008; MAXWELL; CURTIS; VARDANEGA, 2008; MCCABE; TREVIÑO; BUTTERFIELD, 2001; OLIVEIRA, 2007; PASSOW *et al.*, 2006; PERISSÉ, 2003; STANLICK, 2008; VASCONCELOS *et al.*, 2009; VAZ, 2006).

Entretanto, da mesma forma que a tecnologia amplificou as possibilidades de manejo da informação, têm surgido diversas ferramentas que auxiliam a produtividade do conhecimento, variando desde programas de edição de texto até algoritmos com inteligência artificial que leem e produzem de forma autônoma dissertações completas (GOOGLE FOR EDUCATION, 2022; MICROSOFT, 2022; OPEN AI, 2021).

Entre estes recursos, existe uma miríade de softwares que fazem rastreio de similaridades textuais em fragmentos ou documentos completos, gerando relatórios que permitem constatar reproduções entre fontes distintas. Contudo, a análise de ocorrência de apropriação indevida, ou seja, plágio, depende de análise humana (ABREU, 2011; FOLTÝNEK *et al.*, 2020; FOLTÝNEK;

MEUSCHKE; GIPP, 2019; LUQUINI, 2010; MUSSINI, 2008; OLIVEIRA, 2016; PERTILE, 2011; PERTILE; MOREIRA; ROSSO, 2016).

Esses softwares são classificados em categorias de serviços oferecidos para instituições de ensino ou para uso individual, bem como podem ser encontrados de forma paga, com acesso limitado (*freemium*) ou inteiramente gratuitos. Embora, de forma geral, todos os softwares funcionem de forma semelhante, fazendo rastreio na internet de conteúdos (textos e imagens) idênticos, a eficiência deles depende de fatores como a sofisticação do algoritmo usado, a extensão da base de dados consultada e a qualidade do relatório gerado (FOLTÝNEK; MEUSCHKE; GIPP, 2019).

Foltýnek *et al.* (2020) fizeram um estudo de avaliação de performance de quinze softwares internacionais, analisando aspectos relacionados à cobertura e à usabilidade. Sendo assim, a eficiência dos softwares levou em conta aspectos como tipos de fontes rastreadas, tipos de plágios verificados, idioma dos textos analisados, extensão textual verificável, qualidade dos relatórios de similaridade gerados, integração dos softwares em plataformas de aprendizagem virtual, entre outros fatores. O estudo permitiu identificar os softwares mais adequados e os menos adequados de acordo com tais critérios, o que representa uma contribuição importante na escolha de qual software utilizar. Entretanto, observa-se, como limitações deste estudo, o fato de não ter considerado textos em língua portuguesa, o que pode ter resultados de eficiência diferentes, pois os bancos de dados rastreáveis neste idioma podem ter volume de fontes diferentes comparado às bases internacionais. Além disso, o estudo não verificou a eficiência de dois softwares brasileiros bastante utilizados por estudantes, professores e pesquisadores no país, caso do *CopySpider* e do *Plagius*.

Além desses softwares, mais recentemente, foram lançadas, pela Microsoft e pelo Google, funcionalidades de verificação de similaridades entre textos produzidos nos seus softwares de edição Word e Documentos. Não foi verificada até o momento a publicação de estudos avaliando a eficiência destas ferramentas de forma comparativa. Sendo assim, a proposta deste estudo é replicar de forma adaptada o modelo de estudo realizado por Foltýnek *et al.* (2020), analisando a eficiência de um grupo de softwares de detecção de plágio internacionais e nacionais, representativos das diferentes classificações (pagos X

gratuitos/corporativos X pessoais), considerando como protótipo de verificação um texto padrão com conteúdo sobre educação financeira elaborado exclusivamente com a finalidade de submissão a estes softwares.

Ampliando o estudo europeu que analisou a eficiência de softwares de detecção de plágio, esta pesquisa adota o mesmo modelo de investigação, no entanto, aplica um recorte de pesquisa original, incluindo na análise softwares brasileiros, utilizando como texto preparado para análise um protótipo elaborado no idioma em português, com assunto especificamente relacionado à área de negócios. Sendo assim, apresentamos a seguinte pergunta de pesquisa a ser investigada: **Qual é a eficiência de softwares de detecção de similaridades/plágio, sejam eles nacionais ou internacionais, considerando como objeto de análise um texto com conteúdo da área de negócios escrito em português?**

Este estudo visa alcançar os seguintes objetivos: identificar o nível de eficiência dos softwares de detecção de similaridades textuais, analisar e comparar os relatórios gerados pelos softwares, identificar resultados falsos positivos e falsos negativos e determinar o quanto os softwares são efetivos no reconhecimento de similaridades.

Entre os anos 2009 e 2014, foram realizadas competições entre desenvolvedores de sistemas de detecção de plágio (POTTHAST *et al.*, 2013), bem como, em período análogo (2004-2013), um grupo de estudos liderado pela professora de informática Debora Weber-Wulff realizou estudos de avaliação de eficiência de softwares de detecção de similaridades (JAY, 2022). De forma geral, esses estudos são realizados com textos em inglês. Como os algoritmos de similaridade dependem do idioma do manuscrito bem como da base de dados analisada, a replicação deste estudo em língua portuguesa, com um texto da área de negócios, pode trazer novas conclusões a serem analisadas comparativamente aos estudos prévios. Por exemplo, no estudo realizado por Foltýnek *et al.* (2020), inglês, italiano, espanhol e alemão são os idiomas nos quais os softwares apresentaram melhor desempenho. É interessante observar que, embora a eficiência dos softwares analisados não tenha sido realizada com textos no idioma português, percebe-se o destaque nos resultados com dois idiomas com base latina.

A utilização de ferramentas para a verificação de similaridades textuais é um recurso útil para todos os envolvidos na atividade acadêmica. Alunos, professores, pesquisadores, editores e avaliadores (pareceristas) podem e devem utilizar os softwares de detecção de similaridades textuais para prevenir a ocorrência de plágio indesejável ou acidental, para identificar fraudes textuais cometidas intencionalmente com a finalidade de ludibriar e obter alguma vantagem, seja do ponto de vista de uma nota, aprovação ou publicação de trabalho acadêmico. Nas duas circunstâncias, de forma acidental ou intencional, quando o plágio não é identificado, gera externalidades negativas para todos os envolvidos: para os autores que podem ser prejudicados, para as instituições, que podem ter a sua reputação comprometida e para os professores, avaliadores e editores, que podem ser enganados por agentes de má-fé.

2 Revisão da literatura

Com o advento das novas tecnologias de informação e comunicação, o acesso e o manuseio de conteúdos escritos ficaram bastante facilitados. Para alguns autores, o fenômeno antigo do plágio passou, a partir de então, a ser constatado de forma mais frequente, da mesma forma que a sua visibilidade também se tornou mais evidente (HANSEN, 2003; ISON, 2015; SCANLON; NEUMANN, 2002).

Neste contexto, começaram a surgir, também, ferramentas que permitiam a busca e o rastreio de textos semelhantes com a finalidade de encontrar reproduções não autorizadas. A partir dos anos 1990, surgem os primeiros softwares dedicados exclusivamente a esta tarefa, facilitando a identificação de textos idênticos que podiam ser analisados e caracterizados como cópias indevidas, ou seja, manuscritos plagiados. Com a demanda crescente destas ferramentas especialmente no âmbito acadêmico, surgiram softwares no mundo inteiro oferecendo soluções para identificação de similaridades entre textos publicados na internet (FOLTÝNEK; MEUSCHKE; GIPP, 2019).

De forma geral, estes softwares são classificados de acordo com as suas características operacionais. Alguns são acessíveis por meio da obtenção de licenças de uso pagas, outros são inteiramente gratuitos. Há softwares que podem ser utilizados no próprio site, outros precisam ser instalados. Alguns são

de uso institucional, outros são de uso particular. Para a verificação da similaridade entre textos distintos, em alguns softwares é preciso fazer o *upload* do arquivo com o manuscrito completo e em alguns softwares a verificação é feita apenas em trechos de texto inseridos em uma caixa de formulário no próprio *website* (FOLTÝNEK *et al.*, 2020). Para a verificação da identidade entre textos da área acadêmica, os softwares em geral utilizam como bases de dados de rastreio de similaridades publicações da internet (todo tipo de conteúdo publicado), publicações acadêmicas (revistas científicas de acesso restrito e aberto, repositórios, anais de eventos) e manuscritos já submetidos para verificação de similaridades no próprio software.

Quanto maior a abrangência de busca nessas bases, melhores são os resultados de similaridade. Daí a importância dos algoritmos e a explicação do porquê o mesmo texto, rastreado por softwares diferentes, pode ter similaridades encontradas por um software, mas não por outro.

A qualidade da programação e dos recursos embarcados no software basicamente funcionam de acordo com quatro métodos de detecção de similaridades. (1) Léxico: identificação de similaridades literais entre textos; (2) Sintático: identificação de similaridades com substituição de sinônimos; (3) Semântico: identificação de similaridades de textos traduzidos ou parafraseados; (4) Conceitual: identificação de similaridades de ideias e conceitos. (FOLTÝNEK; MEUSCHKE; GIPP, 2019).

Pertile, Moreira e Rosso (2016) explicam que os métodos de detecção são classificados em três categorias. Os mais utilizados são baseados em conteúdo, ou seja, fazem comparações textuais baseadas em palavras, parágrafos e sentenças do texto suspeito com as prováveis fontes. A técnica mais utilizada neste método é o *n*-gram, isto é, a identificação de uma sequência de *n* palavras consecutivas idênticas entre diferentes textos. A segunda categoria refere-se aos métodos de detecção de similaridades baseados no conteúdo e na estrutura do texto, como por exemplo, seções, títulos, tipologia (fonte e tamanho), pontuação e extensão. A última categoria refere-se aos métodos de detecção baseados nas referências e citações, que podem ser usados como filtros (que verificam se as citações se referem às fontes originais) e também examinando as similaridades

entre as citações e as referências que constam do texto (PERTILE; MOREIRA; ROSSO, 2016).

Os autores investigaram a eficiência desses métodos de verificação de similaridades, utilizando uma base de textos científicos, dando especial atenção aos experimentos relacionados ao uso do método baseado na detecção de citações comparado com os demais. Concluíram que os métodos de detecção baseados em conteúdo ainda são os mais eficientes, mas reconhecem que os demais métodos são promissores devido à necessidade de aprimoramento e qualificação das técnicas de detecção. (PERTILE; MOREIRA; ROSSO, 2016).

Essa evidente limitação dos softwares e a demanda por aperfeiçoamento se devem ao fato de que, em geral, os softwares só conseguem gerar evidências entre textos iguais (análise léxica). Entretanto, há formas de plágio que não são reproduções literais, mas ideias e conteúdos reproduzidos de forma reescrita (parafraseada). Esta alteração semântica no texto ainda não é verificada pelos softwares. Da mesma forma, ainda são poucos e ineficazes os softwares capazes de encontrar similaridades entre textos que são traduzidos de outras fontes. Por exemplo, um texto traduzido do hindi para o inglês dificilmente é reconhecido pelos softwares como sendo idêntico (GUPTA, DEEPA; VANI, 2016; MAURER; KAPPE, 2006).

Atualmente, estão em desenvolvimento algoritmos de inteligência artificial com aprendizagem de máquina (*machine learning*), possibilitando a ampliação das formas de análise textual. Este nível de avanço na programação dos softwares, em parte, resulta das diferentes possibilidades de semelhança textual, bem como das constatações feitas por pesquisadores quanto ao funcionamento de tais softwares em relação à eficiência e contribuição no enfrentamento do plágio acadêmico. De forma geral, os estudos realizados apresentaram as seguintes conclusões: (a) comentam a efetividade da pesquisa com base em outros estudos; (b) apresentam uma classificação preliminar do ponto de vista de recursos, acessibilidade e deficiências; (c) fazem comparações baseadas em características funcionais.

Por exemplo, Chaudhuri (2008) verificou em seus estudos que algumas ferramentas são ineficazes na identificação de similaridades entre fontes de acesso restrito comparadas aos documentos em bases de acesso aberto.

Kakkonen e Mozgovoy (2010) dedicaram-se à análise da eficiência dos softwares comparando os resultados obtidos de documentos de fontes como internet, livros eletrônicos e trabalhos feitos por sites que vendem trabalhos acadêmicos. Foltýnek *et al.* (2020) argumentam que estas conclusões são superficiais e incompletas por não apresentarem resultados comparativos entre os softwares do ponto de vista da experiência do usuário. Neste sentido, aspectos como clareza do relatório de similaridades bem como a eficiência e a eficácia dos algoritmos no alcance de fontes para a cobertura de similaridades em relação ao texto submetido são essenciais. Por este motivo, realizaram o estudo que analisou quinze softwares internacionais do ponto de vista da usabilidade e da cobertura com a finalidade de avaliar a eficiência. Os softwares mais eficientes foram: *PlagScan*, *Ouriginal*, *Turnitin* e *Unicheck*. Entre os softwares menos eficientes, ficaram: *Dupli Checker*, *Intihal*, *Plagiarism Software* e *Copyscape*.

Mesmo considerando a eficiência e a funcionalidade operacional, é importante destacar que os relatórios dos sistemas de detecção de similaridade precisam ser analisados com discernimento e crítica, visto que os resultados encontrados podem ser falsos negativos, isto é, os softwares não identificaram a ocorrência de plágio em uma situação na qual havia, bem como podem ser falsos positivos, ou seja, indicam a ocorrência de similaridade em fragmentos os quais não configuram plágio (WEBER-WULFF *et al.*, 2013).

Apesar das limitações e objeções relacionadas à eficiência dos softwares de detecção de similaridades, estes recursos têm sido utilizados em instituições de ensino superior ao redor do mundo, e são reconhecidos como ferramentas que ajudam no trabalho de escrita científica original, como exemplificado em estudos realizados na Turquia e na Arábia Saudita (KÖSE; ARIKAN, 2011; SANJALAWÉ; ANBAR, 2017).

3 Delineamento metodológico

Para a análise de eficiência de softwares detectores de similaridade e plágio foram escolhidos cinco softwares (*Ouriginal*, *Turnitin*, *StrikePlagiarism*, *PlagScan*, *Unicheck*) com melhor performance de acordo com o estudo de Foltýnek *et al.* (2020). Adicionalmente, foram acrescentados três outros

softwares popularmente conhecidos (*iThenticate*, *Plagiarisma* e *Plagium*) e dois nativamente brasileiros (*CopySpider* e *Plagius*). Também foram testadas duas ferramentas de edição de texto que possuem funcionalidades de verificação de similaridades: *Similarity Checker* (Word/Microsoft) e *Originality Report* (Google Docs). Como o software *Plagium* tem uma versão integrada ao Google Docs, a performance desta versão também foi analisada, bem como foram feitas análises de eficiência de recuperação de similaridades usando a plataforma do Google para buscas na internet.

3.1 Descrição dos softwares analisados

No Quadro 1, são descritas as características e especificidades de cada um dos softwares/ferramentas utilizados neste estudo de acordo com as informações disponibilizadas em seus *websites* ou divulgadas pelas empresas responsáveis pelo serviço.

As informações apresentadas contemplam aspectos como nome comercial, país de origem, ano de lançamento, *websites*, descrição geral, base de dados utilizada, verificação de manipulações textuais, tipo de assinatura realizada, integração com ambientes virtuais de aprendizagem (AVA), em inglês, *Learning Management System (LMS)*, idiomas verificados no rastreamento de similaridades e diferenciais do software.

Quadro 1 - Descrição dos softwares/ferramentas analisados neste estudo

N	SOFTWARE	CARACTERÍSTICAS
1	<i>Ouriginal</i> - Origem: Suécia (SWE), 1999. Acesso online: https://www.ouriginal.com/	Integração com LMS. Verifica alguns tipos de manipulação textual. Afirma reconhecer textos semelhantes, modificados com sinônimos e parafraseados. Promete detectar semelhanças entre os idiomas mais usados. Plataforma intuitiva. Assinatura exclusiva para instituições de ensino.
2	<i>Turnitin Originality</i> - Origem: Estados Unidos (USA), 1998. Acesso online: https://www.turnitin.com/pt	Integração com LMS. Verifica alguns tipos de manipulação textual. Afirma identificar plágio em códigos de programação e verificar fraudes autorais em possíveis trabalhos comprados. Plataforma intuitiva.
3	<i>iThenticate</i> - Origem: Estados Unidos (USA), 2004. Acesso online: https://www.ithenticate.com	Integração com sistemas de gerenciamento de publicações. Rastreia em publicações de editoras como <i>Elsevier</i> , <i>Springer-Nature</i> <i>Taylor&Francis</i> . Afirma comparar o conteúdo com 96% dos 10.000 principais periódicos citados. Permite upload de documentos em lotes. Permite compra de créditos para uso avulso. Assinatura exclusiva para instituições de ensino. Verifica similaridades entre textos do mesmo idioma.

4	StrikePlagiarism - Origem: Polônia (POL), 2022. Acesso online: https://strikeplagiarism.com	Integração com LMS. Verifica alguns tipos de manipulação textual. O sistema tem opção de busca de semelhanças em textos traduzidos. Você pode verificar textos com mais de 100 combinações de idiomas diferentes. Permite compra de créditos para uso avulso. Plataforma intuitiva. Plataforma não está disponível em Português.
5	PlagScan - Origem: Alemanha (DEU), 2009. Acesso online: https://www.plagscan.com	Integração com LMS. Verifica alguns tipos de manipulação textual. Base de verificação com mais de 14 bilhões de páginas da internet, artigos e revistas das mais prestigiadas editoras (BMJ, Gale, a Cengage company, Taylor & Francis, Wiley Blackwell, Springer). Permite compra de créditos para uso avulso. Plataforma intuitiva. Plataforma não está disponível em Português.
6	Unicheck - Origem: Chipre (CYP), 2014. Acesso online: https://unicheck.com/	Integração com LMS. Verifica alguns tipos de manipulação textual. Plataforma intuitiva. Permite a utilização de filtros nos relatórios como seleção de fontes, exclusão de textos citados e referências. Permite compra de créditos para uso avulso. Plataforma intuitiva. Plataforma não está disponível em Português.
7	CopySpider - Origem: Brasil (BRA), 2013. Software instalável. Download: https://copyspider.com.br	O software é um <i>freeware</i> , isto é, pode ser utilizado gratuitamente por usuários individuais, mas que também pode ser adquirido por meio de uma assinatura profissional (mais de 20 instalações). Como diferencial, o proprietário afirma que um arquivo de 20 páginas é verificado entre 2 e 5 minutos. Requer instalação em um hardware. Não identifica textos manipulados.
8	Plagius - Origem: Brasil (BRA), 2009. Software instalável. Download em: https://www.plagius.com/br	Opera localmente no computador e pode ser utilizado de forma individual, com uma assinatura básica, sem limite de uso. O software tem uma assinatura profissional, na qual é possível fazer análises em massa e também analisar textos diretamente de sites de internet. Requer instalação em um hardware. Não identifica textos manipulados.
9	Similarity Checker - Origem: Estados Unidos (USA), 2020. Ferramenta integrada ao editor de textos Word/Microsoft	Durante a elaboração do texto no Word, a ferramenta utiliza, em segundo plano, o Bing, serviço de busca na internet também da Microsoft, para verificar similaridades textuais. Quando isto é verificado, o software indica os trechos similares e sugere a inserção de citações. Disponível apenas para determinado grupo de assinantes do pacote Microsoft 365.
10	Originality Report - Origem: Estados Unidos (USA), 2020. Ferramenta integrada ao Google Docs! para assinantes do Google Workspace	A ferramenta promete comparar os textos entregues com páginas da Web e livros da internet, usando o Google Search como recurso de busca. O relatório destaca os textos similares e apresenta as fontes encontradas. A funcionalidade está disponível para os idiomas: espanhol, finlandês, francês, indonésio, inglês, italiano, japonês, norueguês, português e sueco. Disponível apenas para assinantes do Google Workspace for Education Fundamentals e do Google Workspace for Education Plus.
11	Google Search - Origem: Estados Unidos (USA), 1998. Ferramenta de buscas de informações na internet. Acesso online: https://www.google.com	Sistema de buscas de informações na internet sobre qualquer assunto, destacando-se no mundo inteiro como a ferramenta mais utilizada para encontrar informações. Permite buscas de similaridades apenas em textos com cerca de 300 palavras.
12	Plagiarisma - Origem: Reino Unido (GBR), 2010. Acesso online: http://plagiarisma.net/pt/	O software é um <i>freeware</i> para consulta de arquivos de até 2.000 caracteres. Textos ilimitados para assinantes. Informa ser 5x mais rápido que outros softwares e apresentar melhores resultados. Não identifica textos manipulados.

13	<i>Plagium</i> Complemento - Origem: Estados Unidos (USA), 2020. Integração do software <i>Plagium</i> com o Google Docs!	A ferramenta opera com os recursos e características do software que, nesta versão, pode ser instalado como complemento do Google Docs a partir do Google <i>Workspace Marketplace</i> . Não identifica textos manipulados. Funciona apenas para textos do Google Docs.
14	Plagium - Origem: Estados Unidos (USA), 2006. Acesso online: https://plagium.com/	Trata-se de um <i>freeware</i> que não requer instalação e realiza gratuitamente buscas de até 1.000 caracteres e oferece planos de assinaturas para buscas com mais recursos. Permite compra de créditos para uso avulso e assinaturas mensais. Não identifica textos manipulados.

Fonte: Elaborado pelo autor.

É importante constatar a movimentação desse mercado de softwares de detecção de similaridades. Em setembro de 2020, a *EdTech* sueca conhecida como *Urkund* anunciou a compra da companhia alemã *PlagScan*, passando a ser identificada como *Ouriginal*, ampliando o alcance sobre o mercado europeu, dando à empresa o segundo lugar no mundo no serviço de detecção de similaridades entre textos. Em novembro de 2021, a companhia *Ouriginal* foi adquirida pela americana *Turnitin* depois de um rápido processo de análise e aprovação regulatória, tornando-se líder mundial na oferta de serviços de detecção de similaridades entre textos. A empresa *Turnitin* já havia anunciado a aquisição do software *Unicheck* em junho de 2020, ou seja, em um período curto de cerca de 18 meses, quatro empresas distintas foram incorporadas por uma delas. Até o momento, todos os softwares ainda oferecem seus serviços de forma independente com as características próprias.

3.2 Critérios de análise de resultados

Como se sabe, uma limitação de avaliação da eficácia de sistemas de detecção de plágio se deve à dificuldade de reconhecer se o texto similar (no caso, plagiado) corresponde de fato à fonte original (PERTILE; MOREIRA; ROSSO, 2016). Pode acontecer que a fonte identificada pelo software como idêntica não seja a original, mas outro texto qualquer, até mesmo um texto que tenha sido copiado do manuscrito que o software está analisando, o que caracteriza um resultado falso positivo.

Sendo assim, para afastar a limitação e a dúvida observada, para a obtenção dos resultados produzidos pelos softwares neste estudo, será utilizado um texto de 1.000 caracteres sobre educação financeira, elaborado exclusivamente com fragmentos de dez fontes originais. Foram utilizados

excertos da Wikipédia, de revistas científicas de acesso aberto e de acesso restrito (apenas para assinantes), textos científicos publicados em *sites* de eventos acadêmicos e também foi reproduzido literalmente, na língua portuguesa, um fragmento de texto publicado originalmente no idioma inglês em uma página aberta da internet. Alguns trechos foram reproduzidos no texto prototipado de acordo com a convenção brasileira para elaboração de citações (ABNT NBR 10.520/2002) e referências (ABNT NBR 6023/2018), mas a maioria deles foi reproduzida de forma indevida ou dissimulada, com o intuito de enganar o algoritmo de verificação de similaridades dos softwares. O Quadro 2 apresenta os detalhes das fontes e a forma de utilização.

Quadro 2 - Fragmentos e fontes de textos originais para elaboração de texto protótipo para verificação de similaridade e de plágio por softwares.

Fragmentos e fontes originais	Descrição da fonte e utilização do fragmento
(F1) “Portanto, é visível que a educação financeira é necessária no cotidiano de todo ser humano, principalmente, na realidade que se vivência atualmente.” Fonte: ALVES, L. O.; RODRIGUES, F. G. S.; COSTA FILHO, H. O.; MACIEL, E. T. P.; ALMEIDA, M. G. N. A importância da educação financeira e do orçamento familiar frente a pandemia do covid-19. In: USP INTERNATIONAL CONFERENCE IN ACCOUNTING, 20., 2020, São Paulo. Anais [...]. São Paulo: USP, 2020. Disponível em: https://congressosp.fipecafi.org/anais/20UspInternational/ArtigosDownload/2931.pdf. Acesso em: 26 Mar. 2021.	A fonte está publicada nos Anais de um Congresso com acesso aberto. O fragmento será reproduzido literalmente, sem a citação da fonte. Plágio direto
(F2) “A educação financeira tem por propósito auxiliar os consumidores na administração dos seus rendimentos, nas suas decisões de poupança e investimento, no seu consumo consciente e na prevenção de situações de fraude.” Fonte: WIKIPEDIA. Educação Financeira. 2021. Disponível em: https://pt.wikipedia.org/wiki/Educa%C3%A7%C3%A3o_financeira Acesso em: 26 Mar 2021.	A fonte é uma página da internet de conteúdo público. O fragmento será reproduzido literalmente, por meio de citação direta curta, com a citação da fonte. Não caracteriza plágio.
(F3) “No Brasil, há uma situação preocupante no âmbito da educação financeira, demandando urgência na inserção do tema em todas as esferas, ainda mais considerando a desequilibrada distribuição de renda desse país, onde representativa parte dos recursos produtivos é direcionada ao Estado, tornando imprescindível a excelência na gestão de recursos escassos por parte dos indivíduos e de suas famílias” Fonte: SAVOIA, J. R. F.; SAITO, A. T.; SANTANA, F. A. Paradigmas da educação financeira no Brasil. Revista de Administração Pública, Rio de Janeiro, v. 41, n. 6, p. 1121-1141, 2007.	A fonte é uma revista científica com acesso aberto. O fragmento será reproduzido literalmente, por meio de citação direta longa, com a citação da fonte. Não caracteriza plágio.
(F4) “Uma diferença significativa que a crise provocada pela pandemia do novo coronavírus possui das demais tem relação com a mudança de comportamento das famílias. Mesmo depois de vencida a pandemia, novos hábitos permanecerão e formarão o dito novo normal, o que tem relação direta com o aumento da poupança familiar. Como nunca as famílias, em todo o mundo, estão poupando.” Fonte: AFONSO, J. R.; SOUZA, P. R. V.; ABREU, T. F. Para crescer, poupar ou consumir? Conjuntura econômica, Rio de Janeiro, 2021.	A fonte é uma revista editorial da área de negócios com acesso restrito para assinantes. O fragmento será reproduzido literalmente, sem a citação da fonte Caracteriza plágio direto

(F5) “O presente estudo buscou identificar o impacto que existe da educação financeira e o orçamento familiar para evitar o endividamento frente à pandemia do COVID-19, realizando a pesquisa de forma exploratória, adotando procedimentos bibliográficos.” Fonte: ALVES, L. O.; RODRIGUES, F. G. S.; COSTA FILHO, H. O.; MACIEL, E. T. P.; ALMEIDA, M. G. N. A importância da educação financeira e do orçamento familiar frente a pandemia do covid-19. <i>In</i>: USP INTERNATIONAL CONFERENCE IN ACCOUNTING, 20., 2020, São Paulo. Anais [...]. São Paulo: USP, 2020. Disponível em: https://congressosp.fipecafi.org/anais/20UspInternational/ArtigosDownload/2931.pdf. Acesso em: 26 Mar. 2021.	A fonte está publicada nos Anais de um Congresso com acesso aberto. O fragmento será reproduzido por meio de paráfrase, sem a citação da fonte. Caracteriza plágio de paráfrase.
(F6) “Mesmo que a vacina chegue e se consiga imunizar boa parte da população, ainda persistirá a formação e a manutenção de uma elevada poupança precaucional, visto que a crise desencadeada pela pandemia ainda demandará bastante tempo para ser solucionada.” Fonte: AFONSO, J. R.; SOUZA, P. R. V.; ABREU, T. F. Para crescer, poupar ou consumir? Conjuntura econômica, Rio de Janeiro, 2021.	A fonte é uma revista editorial da área de negócios com acesso restrito para assinantes. O fragmento será reproduzido literalmente, sem a citação da fonte. Caracteriza plágio direto.
(F7) “O grande desafio nesse tema é ensinar a cultura do hábito do controle financeiro. Muitos erram precisamente nesse ponto, pois temem analisar os seus gastos mensais. O controle de gastos é algo muito difícil de se realizar, pois é preciso identificar as questões de cunho emocional que também ofuscam a prática do consumo desordenado. Em pequenas despesas é que as pessoas não se dão conta das despesas no fim do mês, mas valores pequenos alteram definitivamente o orçamento de uma família.” Fonte: WIKIPEDIA. Educação Financeira. 2021. Disponível em: https://pt.wikipedia.org/wiki/Educa%C3%A7%C3%A3o_financeira Acesso em: 26 Mar 2021.	A fonte é uma página da internet de conteúdo público. O fragmento será reproduzido literalmente, com uso de recurso de dissimulação de similaridade (inserção de caractere oculto) Caracteriza plágio direto.
(F8) “Desde o início dos anos 2000, a inclusão financeira passou a ser vista por organismos internacionais e governos como uma frente de combate à pobreza, culminando com o lançamento, em 2010, dos Princípios para Inclusão Financeira Inovadora do Grupo dos 20 (G20).” Fonte: RIBEIRO, C. T. Agenda em políticas públicas: A estratégia de educação financeira no Brasil à luz do modelo de múltiplos. Cadernos EBAPE.BR, Rio de Janeiro, v. 18, n. 3, p. 487-498, 2020.	A fonte é uma revista científica com acesso aberto. O fragmento será reproduzido literalmente, sem a citação da fonte. Caracteriza plágio direto.
(F9) “Pode ser entendida como o conjunto de informações que auxilia as pessoas a lidarem com a sua renda, com a gestão do dinheiro, com gastos e empréstimos monetários, poupança e investimentos de curto e longo prazo. A difusão da educação financeira permite que as pessoas aproveitem as oportunidades de produtos e serviços ofertados de uma forma consciente.” Fonte: WIKIPEDIA. Educação Financeira. 2021. Disponível em: https://pt.wikipedia.org/wiki/Educa%C3%A7%C3%A3o_financeira Acesso em: 26 Mar 2021.	A fonte é uma página da internet de conteúdo público. O fragmento será reproduzido literalmente, com uso de recurso de dissimulação de similaridade (substituição de palavras por sinônimos) Caracteriza plágio direto.
(F10) “So you want to become a financial expert, but don't know where to start? Have no fear, a wealth of information is at your fingertips, and getting started is easy. From a primer on personal finances to advanced securities analysis, anyone interested in learning can get access to the necessary resources.” Fonte: SMITH, L.; HOWARD, E. How to become a self-taught finance expert. Investopedia, New York, 13 jan. 2021. Disponível em: https://www.investopedia.com/articles/basics/11/become-self-taught-finance-expert.asp Acesso em: 26 Mar. 2021.	A fonte é uma página da internet de conteúdo publicado em inglês. O fragmento será reproduzido literalmente, com uso de recurso de tradução para o idioma português, reproduzido literalmente Caracteriza plágio direto.

Fonte: Elaborado pelo autor.

O texto prototipado possui esses 10 (dez) fragmentos de textos, os quais foram extraídos de fontes originalmente publicadas, sendo que 08 trechos de textos foram utilizados de forma indevida, ou seja, foram “plagiados” e dois textos (F2 e F3) foram reproduzidos adequadamente, de acordo com a convenção brasileira (com aspas e/ou recuo).

3.3 Procedimentos de análise

Eficiência refere-se à capacidade de uma técnica obter o melhor rendimento com o mínimo de erros ou perdas relacionadas aos objetivos, padrões, normas ou estimativas estabelecidas. (HOUAISS, 2009). A partir desse pressuposto, admite-se como mais eficiente neste estudo o software que identificar com precisão os trechos textuais copiados e as suas fontes correspondentes, de acordo com os seguintes parâmetros: (a) identificação de fragmentos citados corretamente; (b) identificação de fragmentos plagiados; (c) identificação de textos manipulados para enganar o software e (d) identificação de fontes originais.

A métrica da eficiência será feita com uma escala de 0 a 1, considerando proporcionalmente as seguintes características do texto prototipado, no caso o padrão estabelecido a ser verificado pelos softwares:

- a) **identificação de fragmentos citados corretamente no texto prototipado.** São dois fragmentos (F2 e F3). Caso o software identifique esses dois fragmentos a pontuação será 1.0 ponto, se identificar um fragmento será 0,5 ponto e zero ponto se não identificar nenhum;
- b) **identificação de fragmentos plagiados no texto prototipado.** São sete fragmentos (F1, F4, F5, F6, F8, F9 e F10). Caso o software identifique esses sete fragmentos a pontuação será 1.0 ponto e a pontuação proporcional para as demais quantidades. Se não identificar nenhum fragmento, será marcado zero ponto;
- c) **identificação do fragmento manipulado no texto prototipado para enganar o software.** Trata-se do fragmento F7. Caso o software identifique esse fragmento, a pontuação será 1.0 ponto e se não identificar, será 0 ponto;

- d) **identificação das fontes originais dos 10 fragmentos utilizados.** Caso o software identifique todas, a pontuação será 1.0 ponto.

A pontuação final de eficiência do software será a somatória da pontuação obtida entre estas quatro categorias de análise, resultando em uma escala de variação de 0 (software nada eficiente) a 4 pontos (software muito eficiente).

4 Resultados

A Figura 1 mostra os resultados obtidos de acordo com os critérios metodológicos adotados, comparados à observação dos percentuais de similaridades indicados nos relatórios gerados pelos softwares.

Figura 1 - Resultados da identificação de similaridades, plágio, manipulação e fontes dos dez fragmentos utilizados no texto prototipado.

	Original	Turnitin Originality	iThenticate	StrikePlagiarism	PlagScan	Unicheck	CopySpider	Plagius	Similarity Checker (Microsoft Word)	Originality Report (Google Classroom)	Google Busca	Plagiarisma	Plagium (Complemento Google)	Plagium (software)
Percentuais de SIMILARIDADE indicados	40%	33%	33%	67%	25%	31%	10%	52%	32%	35%	N/A	80%	11%	60%
a) Fragmentos CITADOS identificados (2)	1	1	1	1	1	1	0,5	1	0,5	0,5	1	1	0,5	1
b) Fragmentos PLAGIADOS identificados (7)	0,43	0,43	0,43	0,57	0,43	0,29	0,14	0,57	0,43	0,57	0,57	0,29	0,29	0,71
c) Fragmento MANIPULADO identificado (1)	0	1	0	1	1	0	0	0	0	0	0	0	0	0
d) FONTES ORIGINAIS recuperadas (10)	0,30	0,50	0,50	0,20	0,30	0,20	0,10	0,20	0,20	0,10	0,30	0,3	0,2	0,4
TOTAL	1,73	2,93	1,93	2,77	2,73	1,49	0,74	1,77	1,13	1,17	1,87	1,59	0,99	2,11

Fonte: Elaborado pelo autor.

Inicialmente, pode-se constatar que percentuais altos de similaridade não apresentam correlação direta com o índice de eficiência observado. Esse resultado confirma as observações da literatura e a orientação dada pelos próprios softwares, de que os índices gerados por esses precisam ser analisados, já que não se referem à ocorrência de plágio, mas correspondem apenas ao percentual de similaridades textuais identificadas. De fato, os resultados mostram que a maioria dos softwares identificou como similares os fragmentos (F2 e F3) que são citados corretamente, o que não caracteriza plágio (item “a” na Figura 1).

Entretanto, os resultados dos relatórios em relação aos sete fragmentos plagiados (item “b” na Figura 1) foram bem diversificados. Nenhum software conseguiu identificar os sete fragmentos (F1, F4, F5, F6, F8, F9 e F10). A melhor performance foi a do software *Plagium*, que conseguiu identificar 5 fragmentos (0,71 pontos), seguido pelo *StrikePlagiarism*, *Plagius*, *Originality Report* e Google Busca, que identificaram 4 fragmentos (0,57 pontos). O pior desempenho foi o do *CopySpider*, que identificou apenas um fragmento (0,14 pontos). Três fragmentos (0,43 pontos) foram identificados pelo *Ouriginal*, *Turnitin*, *iThenticate*, *PlagScan* e *Similarity Checker* (Microsoft Word). *Unicheck*, *Plagiarisma* e *Plagium* (Complemento Google) identificaram dois fragmentos (0,29 pontos).

É notável que o software *Plagium* tenha mostrado uma performance muito melhor (0,71) comparada ao seu desempenho como complemento do Google Documentos (0,29). A explicação dessa diferença de performance não está clara, mas tem relação com a integração com o Google, que acaba resultando em uma redução de performance cuja implicação é a geração de um resultado ambíguo, pois boa parte dos fragmentos plagiados não são recuperados igualmente pelo mesmo software em duas situações diferentes de funcionamento.

Outro destaque dos resultados mostrados no item “c” da Figura 1 refere-se ao fato de que o fragmento manipulado para enganar o software (F7) foi identificado apenas pelos softwares *Turnitin*, *StrikePlagiarism* e *PlagScan*. Esses softwares, inclusive, apresentam o recurso de identificação de manipulações como um diferencial de funcionamento. Observa-se que, no fragmento F7, a manipulação realizada foi a inclusão de “caractere em branco”. É importante esclarecer que esta técnica de manipulação textual não é naturalmente visível, pois o caractere é incluído com a cor branca, pelo mecanismo automático de substituição, que, sobre o papel branco, fica imperceptível.

Destaca-se que os softwares *Ouriginal* e *Unicheck* anunciam a disponibilidade do recurso de verificação de manipulação textual, mas não tiveram sucesso no reconhecimento do fragmento prototipado com a inclusão de caractere oculto. Os dois softwares verificam apenas a manipulação no caso de

substituição de caracteres, símbolos de alfabetos ou sistemas de escrita diferentes, mas cujo formato seja semelhante (letras/símbolos homógrafos, como acontece com as letras “a” e “e” dos alfabetos latino e cirílico), mas não conseguiram identificar a manipulação com a utilização de caracteres em branco. Em relação à eficiência na recuperação das fontes originais, utilizadas para elaborar o texto prototipado (item “d” na Figura 1), verifica-se, de acordo com a Figura 1, que os softwares *Turnitin* e *iThenticate* foram os mais eficientes, recuperando cinco fontes originais. O software com pior desempenho foi o *CopySpider*, que recuperou apenas uma fonte original do total de dez.

Na Figura 2, são apresentados os resultados dos fragmentos identificados pelos softwares utilizados nesta pesquisa, do ponto de vista dos tipos de plágio e eficiência na recuperação das fontes originais que têm como característica o fato de terem sido extraídas de documentos de natureza diversa, tais como: (a) conteúdo público (Wikipédia ou websites), (b) anais de congressos da área acadêmica e (c) revistas científicas. A maior parte dessas fontes (oito) são acessíveis de forma aberta (acesso público), mas duas delas são fontes de acesso restrito (apenas para assinantes). Foi atribuído o valor 0,0 (zero) quando a fonte não foi identificada pelo software, 0,5 ponto quando alguma fonte foi identificada, mas não é a fonte original da qual o fragmento foi extraído e 1,0 ponto quando a fonte identificada como similar corresponde ao texto original do qual o fragmento foi extraído.

Figura 2 - Eficiência dos softwares de acordo com os tipos de plágio e fontes utilizadas no texto prototipado.

Fragmento	Tipos de plágio e fontes	Original	Turnitin Originality	iThenticate	StrikePlagiarism	PlagScan	Unicheck	CopySpider	Plagius	Similarity Checker (Microsoft Word)	Originality Report (Google Classroom)	Google Busca	Plagiarisma	Plagium (Complemento Google)	Plagium (Software)
F1	Plágio direto (cópia literal). Fonte de acesso público (Anais de congresso).	0,5	1,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,5	0,5	1,0	0,0	0,0
F2	Não caracteriza plágio. Texto com citação direta curta (uso de aspas). Fonte de acesso público (Wikipedia).	1,0	1,0	1,0	0,5	1,0	1,0	1,0	0,5	1,0	0,5	1,0	1,0	0,0	1,0
F3	Não caracteriza plágio. Texto com citação direta longa (uso de recuo). Fonte de acesso público (Revista científica).	1,0	1,0	1,0	1,0	1,0	1,0	0,0	1,0	0,0	0,0	1,0	1,0	1,0	1,0
F4	Plágio direto (cópia literal). Fonte de acesso restrito (Revista científica).	0,5	0,0	0,0	0,5	0,5	0,5	0,0	0,5	0,5	0,5	0,5	0,5	0,5	0,5
F5	Plágio indireto (paráfrase sem citação). Fonte de acesso público (Anais de congresso).	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
F6	Plágio direto (cópia literal). Fonte de acesso restrito (Revista científica).	0,0	0,0	0,0	0,5	0,0	0,5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,5
F7	Plágio direto (cópia literal com recurso de dissimulação). Fonte de acesso público (Wikipedia).	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
F8	Plágio direto (cópia literal). Fonte de acesso público (Revista científica).	0,0	1,0	1,0	1,0	0,5	0,0	0,0	1,0	1,0	1,0	1,0	0,0	0,0	1,0
F9	Plágio direto (cópia literal com substituição de algumas palavras por sinônimos). Fonte de acesso público (Wikipedia).	1,0	1,0	1,0	0,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0	0,0	1,0	1,0
F10	Plágio direto (cópia literal de texto traduzido do idioma inglês). Fonte de acesso público (website).	0,0	0,0	0,0	0,5	0,0	0,0	0,0	0,5	0,5	0,5	0,5	0,0	0,0	0,5
Total de fontes IDENTIFICADAS		5	5	5	6	5	4	2	5	4	5	6	4	3	7
Total de fontes ORIGINAIS identificadas		3	5	5	2	3	2	1	2	2	1	3	3	2	4
Total de fontes SIMILARES identificadas		2	0	0	4	2	2	1	3	2	4	3	1	1	3

Fonte: Elaborado pelo autor.

Primeiramente, a análise horizontal da Figura 2 permite observar que todos os softwares conseguiram identificar similaridades, no máximo até sete (7) dos dez (10) fragmentos utilizados, caso do *Plagium*. Os softwares com melhor eficiência na identificação de similaridades foram *StrikePlagiarism*, Google Busca e *Plagium* (software), que conseguiram identificar similaridades com seis e sete fontes. O software menos eficiente foi o *Copyspider*, que conseguiu recuperar apenas duas fontes. A fonte mais frequentemente recuperada foi a do fragmento F2, que se trata de um texto copiado literalmente da Wikipédia, cuja fonte original foi a mais encontrada na busca de similaridades dos softwares. Apenas o software *Plagium*, na versão complementar do Google, não identificou o fragmento F2 como similar ao texto publicado na Wikipédia.

Analisando do ponto de vista do total de **fontes originais** identificadas, pode-se constatar que os softwares *Turnitin* e *iThenticate* tiveram a melhor

performance e conseguiram recuperar cinco fontes originais, cada um deles. Os softwares brasileiros *CopySpider* e o *Originality Report* (Google Classroom) tiveram a pior eficiência comparada aos demais softwares, pois conseguiram identificar apenas uma fonte original cada um deles.

Considerando o total de fontes similares e originais identificadas, a melhor eficiência foi a dos softwares *Turnitin* e *iThenticate*, visto que recuperaram as fontes originais de pelo menos cinco do total de dez fragmentos utilizados. Os fragmentos F4, F6 e F10 não tiveram as suas fontes originais identificadas por nenhum software. Os fragmentos F4 e F6 foram extraídos de uma revista científica de acesso restrito e o fragmento F10 foi traduzido de um texto publicado no idioma inglês em uma página eletrônica de acesso aberto. Alguns softwares conseguiram identificar similaridades com outros textos, que não são as fontes originais.

Em relação ao fragmento F4, pode-se verificar, na Figura 2, que apenas os softwares *Turnitin*, *iThenticate* e *Copyspider* não conseguiram recuperar nenhuma fonte similar. Todos os demais softwares encontraram fontes com texto similares, no caso, um dos autores do manuscrito que tinha publicado o respectivo trecho para acesso público em sua página eletrônica pessoal. O fragmento F6 também foi extraído da mesma fonte de acesso restrito, mas apenas os softwares *StrikePlagiarism*, *Unicheck* e *Plagium* conseguiram recuperar uma fonte similar (não original), que também, neste caso, trata-se de uma parte do texto original da revista científica que foi reproduzida na página eletrônica pessoal de acesso público, por um dos autores do manuscrito.

Cabe destacar que o fragmento F10 não teve a fonte original identificada no idioma inglês. Não obstante, alguns softwares identificaram fontes similares em português, caso dos softwares *StrikePlagiarism*, *Plagius*, *Similarity Checker* (Microsoft Word), *Originality Report* (Google Classroom), Google Busca e *Plagium*. Em relação a isto, cabe constatar a imprecisão dos softwares por apresentarem resultado falso positivo, mostrando, como fonte, um texto que não corresponde a isto, mas que pode ser um texto já plagiado da fonte original, no caso o texto publicado anteriormente em inglês. Outra observação é o fato de que a grande maioria dos softwares divulgam que fazem buscas de similaridades em vários idiomas diferentes. No caso deste estudo, o fragmento utilizado era de

um texto em inglês que foi escrito em português e nenhum dos softwares utilizados constatou essa tradução, o que demonstra a ineficiência dos softwares deste estudo neste aspecto.

Os fragmentos F5 e F7 não tiveram quaisquer fontes identificadas (originais ou similares). Em relação ao fragmento F7, que foi manipulado para enganar os softwares, embora tenha sido verificado que cinco softwares afirmam identificar manipulações textuais (Figura 1), nenhum entre todos os softwares conseguiu identificar a fonte original do fragmento. Ademais, o fragmento F5 também não teve a fonte identificada por nenhum software. Trata-se de uma paráfrase da fonte original, que caracteriza plágio indireto ou plágio de paráfrase, pois reproduz as ideias da fonte original utilizando outras palavras, sem atribuir os créditos.

Em suma, considerando-se os resultados apresentados na Figura 2, constata-se que todos os softwares utilizados foram ineficientes na identificação de fontes de acesso restrito e de fontes originalmente escritas no idioma inglês. Da mesma forma, os softwares foram ineficientes no reconhecimento das fontes de textos manipulados e também não são aptos para reconhecerem fontes originais de textos parafraseados, nos quais cabe fazer a citação indireta. Ou seja, caso a fonte não seja citada em uma paráfrase, nenhum software analisado neste estudo é capaz de identificar isto.

Na Figura 2, a eficiência dos softwares também pode ser analisada do ponto de vista do tipo de plágio cometido. Embora haja vários tipos de plágio, foram analisados apenas casos de plágio direto (reprodução literal) e plágio indireto (reprodução com paráfrases). Nos casos dos fragmentos F2 e F3, embora haja reprodução literal, não há caracterização de plágio, pois as fontes estão citadas corretamente.

O principal destaque a ser feito é que nenhum dos softwares deste estudo foi capaz de verificar a ocorrência de plágio indireto, isto é, quando um texto original é reproduzido por meio de paráfrase, sem a citação da fonte original (fragmento F5). Esta evidência demonstra que embora haja estudos e pesquisas com metodologias de reconhecimento de similaridades além dos aspectos lexicais, nos softwares pesquisados, tais tecnologias parecem não estar sendo aplicadas ainda.

A similaridade textual mais encontrada por todos os softwares, exceto o *Plagium*, complemento do Google, foi correspondente ao fragmento F2. Entretanto, trata-se de um texto copiado e citado corretamente, que não caracteriza plágio, mesmo caso do fragmento F3, reproduzido da Wikipédia, com citação direta longa.

O fragmento F1 é uma reprodução literal de um texto apresentado em evento científico, com acesso público, cuja fonte original foi identificada apenas por três softwares (*Turnitin*, *iThenticate* e *Plagiarisma*).

Os fragmentos F4 e F6 foram extraídos de uma revista científica de acesso restrito e reproduzidos literalmente (plágio direto). Nenhum software conseguiu identificar a fonte original, embora sejam textos reproduzidos literalmente, o que denota a limitação dos softwares às publicações com acesso restrito.

O fragmento F7 trata-se de um plágio direto, mas que nenhum software conseguiu identificar a fonte correspondente, embora três softwares (*Turnitin*, *StrikePlagiarism* e *PlagScan*) tenham observado que o texto prototipado foi manipulado, o que pode favorecer a suspeição e análise do usuário.

O fragmento F8 foi extraído de uma revista científica de acesso aberto e foi reproduzido sem a atribuição da fonte. Oito softwares conseguiram identificar a fonte original similar ao texto reproduzido (*Turnitin*, *iThenticate*, *StrikePlagiarism*, *Plagius*, *Similarity Checker* (Microsoft Word), *Originality Report* (Google Classroom), Google Busca e *Plagium* (Software)).

O fragmento F9 teve a fonte original reconhecida por seis softwares, contudo, o texto foi transcrito utilizando um recurso de dissimulação textual, que é o uso de palavras sinônimas dos termos originais. Trata-se de uma reprodução com uma paráfrase malfeita com a finalidade de distrair o algoritmo de verificação de similaridade. Não obstante, os softwares *Ouriginal*, *Turnitin*, *iThenticate*, *PlagScan*, *Plagium* (Complemento Google) e *Plagium* (Software) constataram a reprodução e identificaram a fonte original.

O fragmento F10 caracteriza uma reprodução literal de um texto em inglês cuja fonte nenhum software conseguiu identificar, mas seis softwares conseguiram encontrar um texto similar reproduzido em português, foram eles:

Strike plagiarism, *Plagius*, *Similarity Checker* (Microsoft Word), *Originality Report* (Google Classroom), Google busca e *Plagium* (Software).

Em síntese, considerando todos os aspectos analisados neste estudo, observa-se que os softwares mais eficientes, considerando-se os aspectos (a) identificação de fragmentos citados corretamente; (b) identificação de fragmentos plagiados; (c) identificação de textos manipulados para enganar o software e (d) identificação de fontes originais, são: *Turnitin* (2,93 pontos), *StrikePlagiarism* (2,77 pontos) e *PlagScan* (2,73 pontos). Os softwares mais ineficientes, ou seja, que tiveram pontuações mais próximas de 0 ponto, foram: *CopySpider* (0,74 pontos) e *Plagium/complemento* (0,99 pontos).

No Quadro 3, são apresentados os principais pontos fortes e fracos de cada software com a pontuação obtida classificada em categorias de eficiência, considerando-se a escala de 0 a 4 pontos: baixa (0 e 1 ponto), mediana (entre 1 e 2 pontos) e elevada (2 e 4 pontos).

Quadro 3 - Síntese dos pontos fortes e dos pontos fracos dos softwares analisados.

SOFTWARE	PONTOS FORTES	PONTOS FRACOS
<i>Ouriginal</i> - Eficiência mediana: 1,73.	Foi mais eficiente com fontes de acesso aberto, principalmente Wikipédia.	Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em Inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
<i>Turnitin</i> <i>Originality</i> - Eficiência elevada 2,93.	Melhor performance na recuperação de fontes originais. - Recuperou fontes de revistas científicas e congressos de acesso aberto. Conseguiu identificar que houve manipulação no texto para enganar o software. Foi o software mais eficiente de acordo com os parâmetros deste estudo.	Assinatura exclusiva para instituições de ensino. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em Inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
<i>iThenticate</i> - Eficiência mediana: 1,93.	Melhor performance na recuperação de fontes originais. Recuperou fontes de revistas científicas e congressos de acesso aberto.	Verifica similaridades entre textos do mesmo idioma. Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais - Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em Inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.

StrikePlagiari sm - Eficiência elevada: 2,77.	Recuperou as fontes originais de revistas científicas de acesso aberto. Conseguiu identificar que houve manipulação no texto para enganar o software. Foi o segundo mais eficiente de acordo com os parâmetros deste estudo.	Plataforma não está disponível em português. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em Inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
PlagScan - Eficiência elevada: 2,73.	Conseguiu identificar similaridades em metade dos fragmentos analisados. Conseguiu identificar que houve manipulação no texto para enganar o software. Foi o terceiro mais eficiente de acordo com os parâmetros deste estudo.	Plataforma não está disponível em português. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
Unicheck - Eficiência mediana: 1,49.	Conseguiu identificar 4 similaridades textuais, sendo duas delas fontes originais.	Plataforma não está disponível em português. Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
CopySpider - Eficiência baixa: 0,74.	O software é gratuito.	Requer instalação em um hardware. Não identifica textos manipulados. Foi o software com pior eficiência na identificação de similaridades e plágio (2/10). Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Recuperou apenas 1/10 fontes originais na busca de similaridades. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
Plagius - Eficiência mediana: 1,77.	Identificou 5 similaridades textuais, sendo duas fontes originais (duas revistas científicas de acesso aberto).	Requer instalação em um hardware. Não identifica textos manipulados. Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Recuperou apenas 1/10 fontes originais na busca de similaridades. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
Similarity Checker (Microsoft Word) - Eficiência mediana: 1,13.	Identificou 4 similaridades textuais, sendo duas fontes originais (Wikipédia e revista científica de acesso aberto).	Disponível apenas para determinado grupo de assinantes do pacote Microsoft 365. Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em Inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.

Originality Report (Google Classroom) - Eficiência mediana: 1,17.	Identificou 5 similaridades textuais, mas apenas uma fonte original (revista científica de acesso aberto).	Disponível apenas para assinantes do <i>Google Workspace for Education Fundamentals</i> e do <i>Google Workspace for Education Plus</i> . Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Recuperou apenas 1/10 fontes originais na busca de similaridades. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar a fonte do texto manipulado.
Google Busca - Eficiência mediana: 1,87.	Gratuito. Desempenho semelhante ao <i>Ouriginal</i> . Conseguiu recuperar seis fontes similares, das quais três eram originais. Foi mais eficiente com fontes de acesso aberto, principalmente Wikipédia.	Disponível apenas para assinantes do <i>Google Workspace for Education Fundamentals</i> e do <i>Google Workspace for Education Plus</i> . Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Recuperou apenas 1/10 fontes originais na busca de similaridades. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar a fonte do texto manipulado.
Plagiarisma - Eficiência mediana: 1,59	Gratuito para consulta de arquivos até 2.000 caracteres. Conseguiu identificar quatro similaridades (três fontes originais).	Não tem a funcionalidade para identificar textos manipulados. Não identificou fragmento manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado. Gera resultados falsos positivos e falsos negativos.
Plagium (Complemento Google) - Eficiência baixa: 0,99.	Gratuito para consulta de arquivos até 1.000 caracteres. Ferramenta integrada ao Google Docs.	Não tem a funcionalidade para identificar textos manipulados. Funciona apenas para textos do Google Docs. Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em Inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.
Plagium (Software) - Eficiência elevada: 2,11.	Gratuito para buscas de até 1.000 caracteres. Foi o software com maior número de fragmentos similares identificados (7 do total de 10).	Não tem a funcionalidade para identificar textos manipulados. Não identificou texto manipulado para ludibriar o algoritmo de verificação de similaridades textuais. Não conseguiu acessar a fonte original de acesso restrito. Não conseguiu acessar a fonte original escrita em inglês. Não conseguiu identificar fontes similares de textos parafraseados. Não conseguiu identificar a fonte do texto manipulado.

Fonte: Elaborado pelo autor.

5 Discussão

Embora a literatura observe a existência de diversos softwares para a detecção de similaridades textuais que requerem a análise humana para a verificação de plágio (ABREU, 2011; FOLTÝNEK *et al.*, 2020; FOLTÝNEK; MEUSCHKE; GIPP, 2019; LUQUINI, 2010; MUSSINI, 2008; OLIVEIRA, 2016; PERTILE,

2011; PERTILE; MOREIRA; ROSSO, 2016), independentemente de qualquer subjetividade, os resultados de performance técnica ainda são muito limitados.

Os resultados desta pesquisa evidenciam que apenas alguns softwares conseguem ter performance destacada dos demais e ainda assim alcançam no máximo o terceiro quartil de desempenho (notas abaixo de 3,0 pontos em uma escala de 4,0 pontos).

A limitação observada refere-se a vários aspectos, tais como tipo da fonte original utilizada, maneira como o texto foi reproduzido e condições de acesso à fonte original. Conforme já observado na literatura, a promessa de recuperação de fontes originais em outros idiomas não foi verificável neste estudo, que utilizou um texto original em inglês que foi reproduzido em português (GUPTA, DEEPA; VANI, 2016; MAURER; KAPPE, 2006). Talvez a eficiência dos softwares seja maior no caso de textos que são originalmente escritos em outros idiomas e transcritos para o idioma inglês, mas isto não foi o foco deste estudo.

Um outro aspecto importante de ser analisado nos relatórios de similaridade, que não foram considerados no estudo de Foltýnek *et al.* (2020), refere-se à quantidade de fontes recuperadas e à análise do tipo e qualidade das fontes recuperadas como similares. Por exemplo, de dez fragmentos copiados e colados, seria plenamente eficiente o software que conseguisse encontrar as dez fontes similares. Outra coisa importante é a fidelidade da fonte encontrada com a originalidade dos textos. Isto é, um texto similar da internet pode não ser o texto original, por exemplo, pode ser um texto que copiou o texto que está sendo analisado; pode ser um texto que citou a fonte original; ou pode ser um website que reproduz a fonte original (por exemplo, a Wikipédia).

Os softwares, em geral, procuram fontes na internet que tenham textos similares ao manuscrito que está sendo verificado. Os resultados de correspondências textuais nem sempre são da fonte originalmente utilizada. Por exemplo, o fragmento F3 foi extraído de um texto publicado nos Anais de um evento científico, contudo, o software pode recuperar um texto similar de uma outra fonte, que tenha feito a citação ou a apropriação do fragmento do texto original de forma correta ou incorreta. Deixando de lado a análise de credibilidade desta fonte secundária, cabe observar que a eficiência do software

foi limitada, pois pode influenciar negativamente a análise do usuário, que pode considerar como original uma fonte que não é. Por exemplo, se a fonte secundária utiliza o texto original como se fosse próprio (sem identificar a fonte original) pode ficar parecendo que esta é a fonte original.

Para evitar este equívoco, uma alternativa é olhar para as datas de publicação da fonte, comparando com a data da fonte original, assim, por anterioridade, é possível saber qual é a fonte original. Mas se o software não recupera estas duas fontes, o analista pode considerar como original uma fonte que na realidade é uma cópia. É importante observar que para fins de evidenciação de similaridade e reprodução entre o texto em análise e o texto encontrado como similar, considerando-se que no texto elaborado não há a atribuição de créditos, mesmo que seja para a fonte secundária plagiada, isto é suficiente para sustentar uma prática de plágio, ainda que, neste caso, trate-se de plágio feito a partir de plágio.

Enfim, estas situações referem-se a variações nos resultados encontrados entre os softwares, o que mostra o risco dos resultados falso negativos, ou seja, ao verificar o resultado de um software que não apresenta completamente todos os trechos copiados, pode dar a falsa percepção de que o texto é original, bem como o problema dos resultados falso positivos, que correspondem a similaridades que induzem a uma interpretação de plágio, mas que cujas fontes não sustentam. Foi observado isto em relação aos resultados obtidos com os fragmentos F4 e F6, que eram originalmente de uma fonte de acesso restrito, mas que alguns softwares encontraram como fonte apenas trechos reproduzidos em outras fontes, diferentes da original. Também foram constatados resultados falso positivo no caso do fragmento F10, que originalmente é um texto publicado em inglês, mas que os softwares encontraram como fonte textos reproduzidos em português. Estas distorções dos softwares também já têm sido observadas nos estudos de outros pesquisadores, tais como Weber-Wulff *et al.*, (2013).

Outro aspecto crítico observado a partir dos resultados foi constatado no relatório de similaridade do software *Plagiarisma*. O software identifica vários trechos como similares e apresenta as fontes correspondentes, mas uma análise

detalhada das fontes não permitiu que os excertos destacados no relatório de similaridade do software fossem encontrados nas fontes indicadas. Para cada uma das frases identificadas, foi feita a busca (*Ctrl F*) nas fontes correspondentes encontradas pelo software. Por exemplo, a frase “Entretanto, a realidade pode fazer escola” foi identificada pelo *Plagiarisma* como similar, mas em todas as fontes indicadas pelo *Plagiarisma* esta frase literalmente não foi encontrada. A busca dessas frases foi feita com o Google Busca, inseridas entre aspas para recuperação exata.

Em outro caso, o software apresentou o seguinte excerto como inteiramente original (“*unique*”, de acordo com a notação do relatório do software): “modo de produção capitalista é a habilidade dos indivíduos lidarem com o dinheiro como meio ainda persistirá a formação e a manutenção de uma elevada poupança precaucional, visto que”. Entretanto, este texto não é original, mas uma reprodução, conforme observado no fragmento F6.

Estas evidências demonstram que embora as pesquisas sobre os softwares de detecção de similaridades tenham avançado nos últimos anos, servindo-se do aprimoramento dos métodos e algoritmos de detecção, conforme observado por Pertile, Moreia e Rosso (2016), na aplicação prática demonstrada neste estudo tais avanços não parecem evidentes.

Os resultados deste estudo corroboram as observações feitas por Chaudhuri (2008) no que se refere à natureza das fontes utilizadas: os softwares neste estudo também demonstraram limitações no acesso à fonte de acesso restrito que foi utilizada para os fragmentos F4 e F6.

Outro destaque importante deste estudo refere-se ao fato de ter sido analisada a eficiência de dois softwares brasileiros e a verificação de similaridades em um texto escrito em português. Em relação aos softwares, os resultados demonstraram que o *CopySpider* tem eficiência muito abaixo da média dos demais softwares e, além disso, foi ineficiente na recuperação das fontes similares ao texto prototipado. O software *Plagius* teve performance melhor, mas a quantidade e a qualidade das fontes recuperadas ficaram comparáveis aos softwares com pior desempenho.

Por outro lado, os softwares com melhor desempenho correspondem àqueles já identificados no estudo de Foltýnek *et al.* (2020), o que demonstra

que o fato de o texto estar escrito em português não é um problema para a identificação de similaridades. Trata-se de algo que certamente tem mais relação com a eficiência do algoritmo e da base de alcance utilizada pelo software. Comparativamente, este estudo confirmou que os softwares *StrikePlagiarism*, *PlagScan*, *Turnitin*, *Ouriginal* e *Unicheck* performam bem com textos latinos. (FOLTÝNEK *et al.*, 2020).

6 Considerações finais

De acordo com os quatro critérios de análise adotados neste estudo (1 identificação de fragmentos citados corretamente; 2 identificação de fragmentos plagiados; 3 identificação de textos manipulados para enganar o software e 4 identificação de fontes originais), foi possível identificar diferentes padrões de eficiência entre os softwares analisados, o que impossibilita elaborar um ranking entre eles que permita determinar qual é o melhor, embora tenha sido possível verificar que alguns softwares têm desempenho muito abaixo dos demais.

Esta constatação reforça a recomendação de que seja usado mais de um software simultaneamente para a verificação de similaridades. Este cuidado refere-se à diferença da performance dos softwares na recuperação das similaridades e confiabilidade das fontes encontradas, mas também minimiza o risco de resultados falsos positivos (quando uma similaridade parece plágio, mas não é) e falsos negativos (quando uma similaridade não parece plágio, mas é).

Outra conclusão evidenciada neste estudo em relação à limitação da eficiência dos softwares refere-se ao fato de que ainda eles funcionam basicamente com a recuperação de correspondências literais entre textos. Características como recuperação de fontes em idiomas diferentes, reconhecimento de textos manipulados ou até mesmo verificação de similaridades semânticas, nos casos de textos parafraseados, ainda são recursos desejáveis, mas que não foram satisfatoriamente verificados neste estudo.

Como sugestão para novos estudos, recomenda-se a reprodução da metodologia utilizada, com a prototipação de outro texto, de uma área diferente do conhecimento, extraído de fontes diversificadas e redigido em outro idioma. O estudo pode ser feito com os mesmos softwares ou softwares diferentes.

Financiamento

Este estudo contou com recurso e apoio do Programa Institucional de Iniciação Científica (PRÓ-CIÊNCIA) da Fundação Escola de Comércio Álvares Penteado (FECAP).

Agradecimento

Agradeço à Erika Rocha Barros, bolsista do PRÓ-CIÊNCIA/FECAP, pela colaboração no levantamento bibliográfico que respaldou parte desta pesquisa.

Dados abertos

Acesso aos relatórios de similaridades: <https://bityli.com/jlZLyO>

Referências

ABREU, R. M. **Proposta de arquitetura para um sistema de detecção de plágio multi-algoritmo**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2011.

AMOS, K. A. The ethics of scholarly publishing: exploring differences in plagiarism and duplicate publication across nations. **Journal of the Medical Library Association**, Chicago, v. 102, n. 2, p. 87-91, 2014.

ASSOCIATION FOR COMPUTING MACHINERY (ACM). **Acm policy on plagiarism, misrepresentation, and falsification**. Broadway, 2019. Disponível em: <https://www.acm.org/publications/policies/plagiarism-overview>. Acesso em: 25 fev. 2022.

BARBASTEFANO, R. G.; SOUZA, C. G. Percepção do conceito de plágio acadêmico entre alunos de engenharia de produção e ações para sua redução. **Revista Produção Online**, Florianópolis, v. 7, n. 4, 2007.

BENNETT, R. Factors associated with student plagiarism in a post-1992 university. **Assessment & Evaluation in Higher Education**, Oxon, v. 30, n. 2, p. 137-162, 2005.

CHAUDHURI, J. Deterring digital plagiarism, how effective is the digital detection process? **Webology**, Iran, v. 5, n. 1, p. 1-10, 2008.

CHRISTOFE, L. **Intertextualidade e plágio: questões de linguagem e autoria**. Campinas: Universidade de Campinas, 1996.

FACHINI, G. J.; DOMINGUES, M. J. C. S. Percepção do plágio acadêmico entre alunos de programas de pós-graduação em administração e contabilidade. *In: SEMINÁRIOS EM ADMINISTRAÇÃO FEA-USP*, 11., 2008, São Paulo. **Anais [...]**. São Paulo: USP, 2008. p. 1-14.

FOLTÝNEK, T. *et al.* Testing of support tools for plagiarism detection. **International Journal of Educational Technology in Higher Education**, New York, v. 17, n. 46, p. 31, 2020.

FOLTÝNEK, T; MEUSCHKE, N; GIPP, B. Academic plagiarism detection: a systematic literature review. **ACM Computing Surveys**, New York, v. 52, n. 6, 2019.

GOOGLE FOR EDUCATION. **Use os relatórios de originalidade para procurar com rapidez e facilidade os trechos onde faltam citações das fontes**. 2022. Disponível em: <https://edu.google.com/products/originality>. Acesso em: 25 fev. 2022.

GUPTA, D.; VANI, K. Study on extrinsic text plagiarism detection techniques and tools. **Journal of Engineering Science and Technology Review**, Greece, v. 9, n. 4, p. 150-164, 2016.

HANSEN, B. Combating plagiarism: is the internet causing more students to copy? **The CQ Researcher**, Thousand Oaks, v. 13, n. 32, p. 773-796, 2003.

HOUAISS, A. Eficiência. *In: HOUAISS, A. Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Objetiva, 2009.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS (IEEE). **Plagiarism levels and corrective actions, as taken from section 8.2.4.D of the IEEE publication services and products board operations manual**. Piscataway, 2021. Disponível em: <https://pspb.ieee.org/images/files/files/opsmanual.pdf>. Acesso em: 25 fev. 2022.

ISON, D. C. The influence of the internet on plagiarism among doctoral dissertations: an empirical study. **Journal of Academic Ethics**, Dordrecht, v. 13, n. 2, p. 151-166, 2015.

JAY, R. **Portal plagiat**. Berlin, 2022. Disponível em: <https://plagiat.htw-berlin.de/software-en/>. Acesso em: 22 fev. 2022.

KAKKONEN, T.; MOZGOVOY, M. Hermetic and web plagiarism detection systems for student essays-an evaluation of the state-of-the-art. **Journal of Educational Computing Research**, Thousand Oaks, v. 42, n. 2, p. 135-159, 2010.

KOOCHER, G. P.; KEITH-SPIEGEL, P. Peers nip misconduct in the bud. **Nature**, Berlin, v. 466, n. 7305, p. 438-440, 2010.

- KÖSE, Ö; ARIKAN, A. Reducing plagiarism by using online software: an experimental study. **Contemporary Online Language Education Journal**, Turkey, v. 1, p. 122-129, 2011.
- LUQUINI, E. **Uma proposta para promover a aprendizagem nas disciplinas de programação utilizando-se de redes sociais modeladas por técnicas de detecção de plágio**. São Paulo: Universidade Presbiteriana Mackenzie, 2010.
- MAURER, H; KAPPE, F. Plagiarism: a survey. **Journal of Universal Computer Science**, Austria, v. 12, n. 8, p. 1050-1084, 2006.
- MAXWELL, A.; CURTIS, G. J.; VARDANEGA, L. Does culture influence understanding and perceived seriousness of plagiarism? **International Journal**, London, v. 4, n. 2, p. 25-40, 2008.
- MCCABE, D. L.; TREVIÑO, L. K.; BUTTERFIELD, K. D. Cheating in academic institutions: a decade of research. **Ethics & Behavior**, Oxon, v. 11, n. 3, p. 219-232, 2001.
- MICROSOFT. **Check your document for similarity to online sources**. 2022. Disponível em: <https://support.microsoft.com/en-us/office/check-your-document-for-similarity-to-online-sources-6d942360-b5ca-445f-a84d-6e8c66fc40d2>. Acesso em: 25 fev. 2022.
- MUSSINI, J. A. **Novas arquiteturas para detecção de plágio baseadas em redes P2P**. Curitiba: Pontifícia Universidade Católica do Paraná, 2008.
- NOORDEN, R.V. Science on the move. **Nature**, Berlin, v. 490, p. 326-329, 2012.
- OLIVEIRA, A. M. **Um método de detecção de plágio em códigos-fonte para disciplinas iniciais de programação**. Manaus: Universidade Federal do Amazonas, 2016.
- OLIVEIRA, M. M. **Plágio na constituição de autoria: análise da produção acadêmica de resenhas e resumos publicados na internet**. São Paulo: Universidade Presbiteriana Mackenzie, 2007.
- OPEN AI. **GPT-3 powers the next generation of apps**. 2021. Disponível em: <https://openai.com/blog/gpt-3-apps/>. Acesso em: 25 fev. 2022.
- PASSOW, H. J. *et al.* Factors influencing engineering students' decisions to cheat by type of assessment. **Research in Higher Education**, New York, v. 47, n. 6, p. 643-684, 2006.
- PERISSÉ, G. O Conceito de plágio criativo. **Videtur**, São Paulo, n. 18, p. 9-19, 2003.

PERTILE, S.; MOREIRA, V.; ROSSO, P. Comparing and combining content and citation: based approaches for plagiarism detection. **Journal of the Association for Information Science and Technology**, Hoboken, v. 67, n. 10, p. 2511-2526, 2016.

PERTILE, S. L. **Desenvolvimento e aplicação de um método para detecção de indícios de plágio**. Santa Maria: Universidade Federal de Santa Maria, 2011.

POTTHAST, M. *et al.* Overview of the 5th international competition on plagiarism detection. **CEUR Workshop Proceedings**, Germany, v. 1179, p. 845-876, 2013.

RIBEIRO, M. D.; VASCONCELOS, S. M. R. Retractions covered by Retraction Watch in the 2013-2015 period: prevalence for the most productive countries. **Scientometrics**, Dordrecht, v. 114, p. 719-734, 2018.

SANJALAWA, K.; ANBAR, Y. M. The Plagiarism detection systems for higher education: a case study in Saudi Universities. **International Journal of Software Engineering & Applications**, Australia, v. 8, n. 2, p. 33-49, 2017.

SCANLON, P. M.; NEUMANN, D. R. Internet plagiarism among college students. **Journal of College Student Development**, Baltimore, v. 43, n. 3, p. 374-385, 2002.

STANLICK, N. An open letter to faculty: some thoughts on plagiarism from "Colonel Cheatabuster". **UCF Faculty Focus**, Orlando, v. 7, n. 4, p. 9-10, 2008.

VASCONCELOS, S. *et al.* Discussing plagiarism in Latin American science. Brazilian researchers begin to address an ethical issue. **EMBO reports**, Hoboken, v. 10, n. 7, p. 677-82, 2009.

VAZ, T. R. D. O avesso da ética: a questão do plágio e da cópia no ciberespaço. **Cadernos de Pós-Graduação**, São Paulo, v. 5, n. 1, p. 159-172, 2006.

WEBER-WULFF, D. *et al.* **Plagiarism detection software test**. 2013. Disponível em: <https://plagiat.htw-berlin.de/software-en/test2013/>. Acesso em: 08 ago. 2022.

Eficiencia del software nacional e internacional en la detección de similitud y plagio en manuscritos

Resumen: Este estudio tiene como objetivo identificar el nivel de eficiencia de catorce software para detectar similitudes entre un texto con fragmentos extraídos de contenido en línea sobre educación financiera, encontrados en sitios web de acceso abierto, anales de eventos académicos y revistas científicas de acceso abierto y restringido. Se utilizaron fragmentos extraídos de las fuentes para crear un texto cohesivo, escrito en portugués, que contenía reproducciones textuales literales, paráfrasis, traducción del idioma inglés y manipulación con inserción de caracteres ocultos y reemplazo de palabras por sinónimos. Los informes de similitud generados por los programas se analizaron según cuatro criterios: 1 identificación de fragmentos correctamente citados; 2 identificación de fragmentos plagiados; 3 identificación de textos manipulados para engañar al software; 4 Identificación de fuentes originales, las cuales se puntuaron en una escala de 0 a 4 puntos. Si bien los software Turnitin, Strikeplagiarism y Plagscan obtuvieron los mejores índices de desempeño, se encontró que los programas operan con eficiencia variable y limitada, lo que refuerza la convicción de que, si bien son herramientas útiles para la identificación del plagio, contribuyen significativamente limitadas por aspectos como el reconocimiento principalmente de reproducciones textuales literales que no siempre corresponden a la fuente original. Ningún software encontró las fuentes originales de los textos reescritos, manipulados y traducidos. El estudio contribuye a la mejora de la capacidad del usuario para elegir, utilizar y analizar los informes de similitud generados por el software, cuya eficiencia puede ser mayor en el caso de utilizar más de un programa.

Palabras clave: plagio; similitud y plagio; detección de plagio; Software antiplagio

Efficiency of national and international software in detection of similarity and plagiarism in manuscripts

Abstract: This study aims to identify the efficiency of fourteen software for detecting similarities in a text with fragments on financial education, found on internet websites, proceedings of academic events and open and restricted access scientific journals. The fragments were used to create a cohesive text, written in Portuguese, containing literal textual reproductions, paraphrases, with excerpts translated from the English language, manipulated with hidden character insertion and replacement of words by synonyms. The similarity reports generated by the software were analyzed according to four criteria: 1 identification of correctly cited fragments; 2 identification of plagiarized fragments; 3 identification of texts manipulated to deceive the software; 4 identification of original sources; which were scored on a scale of 0 to 4 points.

The software Turnitin, StrikePlagiarism, PlagScan and Plagium had high performance and CopySpider and Plagium (add-on to Google) were the most inefficient. It was found that the software operates with variable efficiency, which reinforces the conviction that although they are useful tools for the identification of plagiarism, they contribute in a limited way to aspects such as the recognition of literal textual reproductions that do not always correspond to the original source. The original sources of rewritten, manipulated and translated texts were not found by any software. The study contributes to the improvement of the user's ability to choose, use and analyze the similarity reports generated by software, whose efficiency can be greater in the case of using more than one software.

Keywords: plagiarism; similarity and plagiarism; plagiarism detection; anti-plagiarism software

Recebido: 24/03/2022

Aceito: 24/06/2022

Declaração de autoria

Concepção e elaboração do estudo: Marcelo Krokosz.

Coleta de dados: Marcelo Krokosz.

Análise e interpretação de dados: Marcelo Krokosz.

Redação: Marcelo Krokosz.

Revisão crítica do manuscrito: Marcelo Krokosz.

Como citar:

KROKOSZ, Marcelo. Eficiência de softwares nacionais e internacionais na detecção de similaridade e de plágio em manuscrito. **Em Questão**, Porto Alegre, v. 28, n. 4, e-123123, out./dez. 2022. <https://doi.org/10.19132/1808-5245284.123123>

