Adaptação do Algoritmo Ferret/WebFerret para Detecção de Plágio em Textos em Português: Desafios com Textos Gerados por IA

Adler Gonçalves Elias[1], **Robson da Silva Lopes**[2]

[1]Graduando em Ciência da Computação/ICET/CUA
Universidade Federal de Mato Grosso (UFMT) - Barra do Garças – MT – Brasil

[2]Docente no curso Ciência da Computação/ICET/CUA
Universidade Federal de Mato Grosso (UFMT) - Barra do Garças – MT – Brasil

adler.elias@sou.ufmt.br, robson.lopes@ufmt.br

Abstract. *This study investigates the adaptation of the Ferret/WebFerret algorithm for plagiarism detection in Portuguese-language texts, with a particular focus on the challenges presented by AI-generated content. The implemented system conducts a phrase-by-phrase analysis using trigrams, leveraging Google's API to search for potential source material. In experiments involving mosaic plagiarism, the algorithm proved effective in identifying instances of copied content. However, it struggled to trace the origins of AI-generated texts. The findings indicate a need for the development of new AI-driven algorithms tailored to this emerging issue.*

Resumo. *Este estudo examina a adaptação do algoritmo Ferret/WebFerret para a detecção de plágio em textos em língua portuguesa, com ênfase nos desafios impostos por textos gerados por Inteligência Artificial (IA). A implementação realizada efetua a análise de similaridade em nível de frases, utilizando trigramas e a API do Google para a busca de fontes. Nos testes realizados com textos classificados como plágio por mosaico, o algoritmo demonstrou eficiência na detecção, porém apresentou limitações na identificação da origem de textos gerados por IA. Os resultados indicam a necessidade de desenvolvimento de novos algoritmos baseados em IA para essa finalidade.*

1. Introdução

Até o final de 2022, sites na internet eram as principais fontes de acesso a textos, vídeos e áudios, democratizando o conhecimento por meio de plataformas como websites e YouTube. Com a facilidade de acesso à informação, o plágio tornou-se uma preocupação em vários âmbitos, principalmente o acadêmico. Em novembro de 2022, foi lançado o ChatGPT [OpenAI 2024], uma inteligência artificial avançada capaz de gerar textos complexos sem citar fontes. Esse lançamento trouxe novos desafios e aumentou a preocupação com o uso ético dessas ferramentas, uma vez que muitas pessoas começaram a utilizá-las para a escrita de textos sem a devida atribuição. Neste contexto, uma pergunta surge: textos gerados por ferramentas de IA, sem citação de fonte, podem ser considerados plágio?

O dicionário Michaelis define plágio como: (1) apresentar como sua uma obra ou ideia de outro autor; (2) usar uma obra como fonte sem mencioná-la; e (3) imitar de forma

servil ou fraudulenta [Niskier 2020]. No Brasil, a Constituição Federal e a Lei de Direitos Autorais reconhecem o plágio como uma violação de direitos autorais, sujeita a sanções legais. Embora a palavra ”plágio”não esteja explicitamente presente no ordenamento jurídico, o ato é juridicamente reconhecido como crime [Presidência da República 1988].

No contexto acadêmico e científico, o plágio é uma prática amplamente condenada, principalmente devido ao impacto que tem na integridade das pesquisas e na originalidade dos trabalhos acadêmicos. [Malcolm and Lane 2008] identificam quatro formas principais de plágio, conhecidas como tradicionais, sendo elas: (i) conluio, quando trabalhos são compartilhados entre colegas; (ii) plágio por meio da internet, onde textos de vários sites são copiados e modificados para disfarçar a cópia; (iii) uso de trabalhos prontos, adquiridos online ou em bancos de dados; e (iv) escrita sob encomenda, quando alguém é contratado para realizar o trabalho.

Embora o plágio tradicional ocorra de diversas formas, ele pode ser detectado com o uso de ferramentas tecnológicas ou comparações de textos. Ferramentas como Plagius, TurnItIn [TurnItIn 2011], iThenticate [iThenticate 2011] e Ferret [Lane et al. 2006] ajudam a detectar o plágio de maneira eficaz, auxiliando na verificação de grandes volumes de texto e garantindo a integridade das produções acadêmicas.

No entanto, estudos recentes [Basic et al. 2023, Lo 2023, Szabo 2023], indicam que esses detectores muitas vezes não conseguem identificar o plágio gerado por IA, pois os textos criados não são diretamente copiados de uma fonte específica, mas construídos a partir de padrões linguísticos e dados processados pela IA.

Diante desse cenário, o presente estudo tem como objetivo reproduzir a ferramenta WebFerret [Malcolm and Lane 2008] para detecção de plágio e avaliar, especificamente para textos em língua portuguesa, se ela apresenta dificuldade em identificar a fonte ou origem de textos gerados por inteligência artificial.

2. Materiais e Métodos

2.1. Ferret e WebFerret

De acordo com [Lane et al. 2006], a ferramenta Ferret calcula a similaridade entre dois textos segundo seus tokens triplos (trigramas) comuns e distintos. Segundo [Lyon et al. 2001], em estudos realizados em grandes corpora, trigramas, em sua maioria, ocorrem uma única vez. Assim, se uma palavra tem baixa probabilidade de ocorrer, a possibilidade de que ela ocorra em conjunto com outras é ainda menor, mesmo que os textos envolvidos tratem basicamente do mesmo assunto. Dessa forma, esses autores propuseram uma metodologia para comparar e classificar dois textos como “independentes” ou “similares”, baseada na ideia de “semelhança” e “contenção” para afirmar se um texto se assemelha a outro e se um texto está contido em outro, respectivamente.

Considerando S(A) e S(B) como conjuntos de trigramas dos textos A e B, respectivamente, o quociente entre o número de trigramas correspondentes dos dois textos e a união desses trigramas determina a semelhança, R(A,B), entre os textos, como mostrado na equação:

$$R(A, B) = \frac{|S(a) \cap S(b)|}{|S(a) \cup S(b)|} \tag{1}$$

denomina coeficiente de Jaccard [Lyon et al. 2001]. A equação 1 gera valores entre 0 e 1, em que valores próximos de 1 indicam alto índice de similaridade e valores próximos de 0 indicam baixo índice de similaridade. Dessa forma, o algoritmo do Ferret converte um texto em trigramas e depois compara-os aos trigramas de outros textos. Caso o valor seja próximo de 1 entre os trigramas dos conjuntos, existe uma alta possibilidade de colusão ou de plágio, mesmo que o texto não seja idêntico.

O WebFerret é uma ferramenta para busca de plágio na web e utiliza em seu processo de identificação o algoritmo Ferret. O algoritmo do WebFerret foi proposto de acordo a Figura 1. Primeiro, o usuário seleciona o texto no qual deseja identificar o plágio. Em seguida, o texto é fragmentado em trigramas, e os trigramas com stop words[1] são excluídos. Com os trigramas restantes, são realizadas consultas na web por meio de motores de busca, um trigrama de cada vez. Para cada trigrama, são armazenados os 10 melhores resultados e, em seguida, os resultados de todos os trigramas são posicionados em ordem de frequência. Por fim, as 10 URLs que apareceram com maior frequência têm seus arquivos baixados e analisados via Ferret.

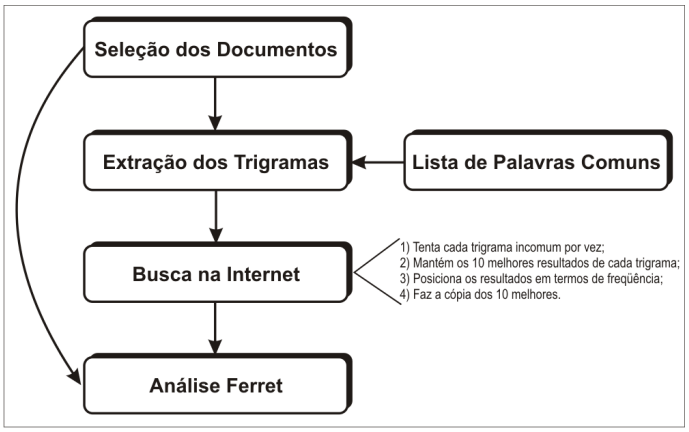

Figura 1. Processo utilizado pelo webferret[Malcolm and Lane 2008]

Mais detalhes sobre os algoritmos do Ferret e WebFerret podem ser encontrados nos trabalhos de [Lane et al. 2006] e [Malcolm and Lane 2008].

2.2. Adaptação do Ferret e Webferret

O presente trabalho foi implementado em linguagem Python e realizou algumas adaptações no algoritmo do trabalho original do Ferret e WebFerret. São elas:

1. O algoritmo original do WebFerret realiza o cálculo do grau de similaridade tomando como base os arquivos inteiros. No presente trabalho, o texto em que se busca identificar o plágio é quebrado em frases, e o grau de similaridade é calculado frase por frase;
2. O algoritmo original do WebFerret utiliza *stop words* da língua inglesa. No presente trabalho, foi utilizada uma base de *stop words* da língua portuguesa.

[1]Stop words: São palavras comuns em um idioma que, em muitas aplicações de processamento de linguagem natural (PLN), são filtradas ou removidas de textos antes de realizar análises mais complexas. O objetivo de remover as *stop words* é reduzir o ”ruído”no processamento, permitindo que mecanismos de busca se concentrem em termos mais relevantes para a tarefa em questão.

Para as buscas na Internet, adotou-se o motor de busca Google, por meio da API *Custom Search Engine*.

A Figura 2 apresenta todo o fluxo de funcionamento do algoritmo desenvolvido neste trabalho. O processo começa com o upload de arquivos de texto, atualmente limitado ao formato .TXT, mas com estrutura para futuras expansões. Em seguida, a ferramenta realiza a extração de frases, usando a biblioteca NLTK, e a extração de trigramas (grupos de três palavras consecutivas) das frases, que são armazenados no banco de dados. Trigramas raros são identificados com base na frequência de ocorrência.

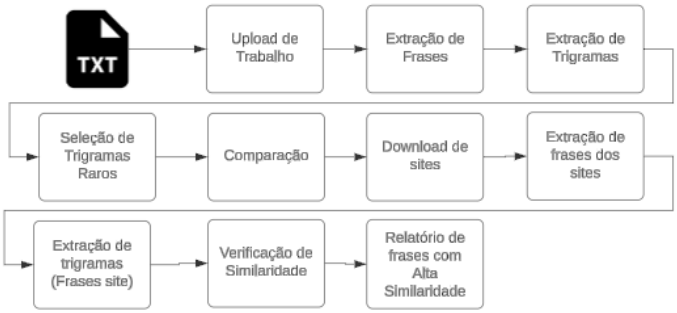

Figura 2. fluxo de funcionamento do algoritmo presente neste trabalho

A fase de busca utiliza a API Custom Search Engine, retornando links relacionados aos termos, que são classificados e armazenados. A partir desses links, o conteúdo dos sites é obtido via web scraping, com o uso das bibliotecas BeautifulSoup e Requests, e processado para a extração de trigramas das frases dos sites. A verificação de similaridade entre os trigramas extraídos do texto e dos sites é realizada com base na equação de interseção e união dos trigramas.

Por fim, é gerado um relatório de frases com alta similaridade, listando frases e suas potenciais fontes e destacando trechos que apresentam maior semelhança, auxiliando na identificação de possíveis plágios.

2.3. Base de textos de teste

Nos testes de detecção de plágio, foram considerados dois tipos de textos, definidos com base em sua origem. O primeiro tipo, denominado plágio por mosaico, refere-se a uma forma mais sofisticada de plágio que envolve a combinação e reorganização de fragmentos extraídos de diversas fontes, resultando em um texto que aparenta ser original, mas que, na realidade, consiste em partes copiadas. O segundo tipo inclui textos gerados por Inteligência Artificial (IA), especificamente utilizando o modelo ChatGPT 3.5, entre os dias 25 de março e 1º de abril de 2024.

2.4. Métrica de similaridade

Para calcular o grau de similaridade dos textos, no trabalho proposto, o grau de similaridade será retornado por frase e será calculado com base na equação 1.

3. Resultados

Para textos do tipo mosaico, foram copiados trechos de diversas fontes da web para criar textos sobre os seguintes assuntos: i) biodiversidade na Amazônia; ii) inteligência artificial; iii) astrofísica; iv) astronomia; v) blockchain; vi) economia comportamental; vii) engenharia genética; viii) psicologia social; ix) dengue; e x) IA generativa.

A Tabela 1 apresenta a média do grau de similaridade das frases dos textos mosaico com os textos encontrados pelo algoritmo aqui proposto.

Tabela 1. Média do grau de similaridade das frases dos textos mosaico

Assunto	i	ii	iii	iv	v	vi	vii	viii	ix	x
Similaridade	0,50	0,78	0,89	0,92	0,63	0,92	0,81	0,54	0,79	0,72

Observa-se que, em sete dos textos analisados, a média de similaridade das frases foi superior a 0,7, evidenciando a eficácia do algoritmo na identificação das fontes utilizadas para compor o texto. Mesmo nos casos em que a média de similaridade foi inferior a 0,7, algumas frases foram localizadas com precisão, atingindo um grau de similaridade de 1, o que indica, de forma clara, a presença de trechos copiados ou plagiados.

Para textos gerados por IA, neste caso o ChatGPT, foi solicitado que gerasse textos sobre os seguintes assuntos: i) mudanças climáticas e agricultura; ii) desenvolvimento de smart cities; iii) a história da música eletrônica; iv) avanços em energia renovável; v) influência da mitologia na cultura pop; vi) inteligência artificial na medicina; vii) o que é Node.js; viii) o que é IA; ix) cidade onde reside o pesquisador; e x) Monteiro Lobato.

A Tabela 2 apresenta a média do grau de similaridade das frases dos textos gerados por IA com os textos encontrados pelo algoritmo aqui proposto.

Tabela 2. Média do grau de similaridade das frases dos textos gerados por IA

Assunto	i	ii	iii	iv	v	vi	vii	viii	ix	x
Similaridade	0,08	0,06	0,07	0,07	0,07	0,08	0,11	0,06	0,14	0,10

Verifica-se que o algoritmo apresentou baixa eficácia na identificação das fontes de textos gerados por IA, resultado que corrobora anteriores [Khalil and Er 2023, Lo 2023, Ventayen 2023]. Esses autores destacam que algoritmos baseados em grau de similaridade textual possuem limitações significativas na detecção de plágio em textos produzidos por sistemas de inteligência artificial, devido à natureza não derivativa desses conteúdos.

3.1. Considerações Finais

Ao analisar os objetivos propostos, verifica-se que foram alcançados, uma vez que a implementação do algoritmo WebFerret foi realizada com sucesso, ajustando-se às particularidades da língua portuguesa, como a utilização de *stop words* específicas. Tal adaptação mostrou-se promissora para a detecção de plágio em textos nessa língua. Ademais, foi proposta uma modificação no algoritmo original, que consiste no cálculo de similaridade por frases, em vez de textos inteiros, reconhecendo que, frequentemente, o plágio ocorre em fragmentos extraídos de diferentes fontes. Essa alteração demonstrou-se eficaz, permitindo a identificação de fontes em textos classificados como plágio por mosaico.

Por fim, conforme previsto nos objetivos e relatado na literatura, constatou-se que textos gerados por IA, como o ChatGPT, apresentam maior dificuldade para a identificação de suas fontes ou para a detecção de plágio. Assim, ferramentas baseadas em análise de similaridade textual mostram-se insuficientes para lidar com conteúdos gerados por IA, evidenciando a necessidade de desenvolver novos algoritmos baseados em IA que sejam capazes de identificar padrões em textos produzidos por essas tecnologias.

Referências

Basic, Z., Banovac, A., Kruzic, I., and Jerkovic, I. (2023). Better by you, better than me, chatgpt3 as writing assistance in students essays. *Humanities and Social Sciences Communications*, 10(1).

iThenticate (2011). Plagiarism detection software — ithenticate. `https://www.ithenticate.com/`. Accessed: 2024-09-19.

Khalil, M. and Er, E. (2023). Will chatgpt get you caught? rethinking of plagiarism detection. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 14040, pages 475–487.

Lane, P. C. R., Lyon, C. M., and Malcolm, J. A. (2006). Demonstration of the ferret plagiarism detector. In *Proceedings of the 2nd International Plagiarism Conference*.

Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410.

Lyon, C., Malcolm, J., and Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 118–125. Citeseer.

Malcolm, J. A. and Lane, P. C. R. (2008). Efficient search for plagiarism on the web. In *Proceedings of The International Conference on Technology Communication Education Kuwait 2008*, pages 206–211.

Niskier, A., editor (2020). *Michaelis Dicionário Brasileiro da Língua Portuguesa*. Editora Melhoramentos, São Paulo, 8ª edição edition.

OpenAI (2024). Chatgpt. Acessado em: 19 de setembro de 2024.

Presidência da República, B. (1988). Constituição da república federativa do brasil de 1988. `http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm`. Acesso em: 28 de outubro de 2024. Art. 5º.

Szabo, P. D. A. (2023). Chatgpt a breakthrough in science and education: Can it fail a test?

TurnItIn (2011). Um resumo da efetividade do turnitin.

Ventayen, R. J. M. (2023). Openai chatgpt generated results: Similarity index of artificial intelligence-based contents. *SSRN Electronic Journal*.