Text Laundering: Concealing the Use of Generative AI in Text

Aline Lima de Souza[1], Matheus Argôlo[1], Carlos Eduardo Barbosa[1,2], Yuri Oliveira de Lima[1], Herbert dos Santos[1], Alan Lyra[1] and Jano de Souza[1]

[1]Programa de Engenharia de Sistemas e Computação do Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
[2]Centro de Análises de Sistemas Navais, Marinha do Brasil, Rio de Janeiro, Brazil

limaals@cos.ufrj.br
matheusargolo@cos.ufrj.br
eduardo@cos.ufrj.br
yuriodelima@cos.ufrj.br
herbertsds@cos.ufrj.br
alanlyra@cos.ufrj.br
jano@cos.ufrj.br

Abstract: In recent years, we have witnessed a growing increase in applications involving some form of Artificial Intelligence (AI). Conversational AI has gained considerable prominence among the various types of AI applications. Classified as a type of generative AI, chatbot applications such asOpenAI's ChatGPT or Google's Bard are now utilized by multiple authors as tools. Despite its many advantages, the indiscriminate use of this type of generative AI in texts can raise ethical questions about who owns authorship of a particular work, as a human author may have contributed very little to the production of certain content. The concern about authorship is paramount in producing scientific publications, such as articles, dissertations, or theses. To avoid such questioning, some authors may develop strategies to conceal the use of generative AI in their productions. This work aims to introduce the concept of Text Laundering (TL), which we name as such due to its similarity to the practice of money laundering — applying various consecutive alterations to AI-generated text to make its origin indeterminate. This work will also develop the concept by examining examples of similar practices in the literature employing an ad hoc search. A method for applying TL to mask the origin of a text will be proposed, and the results obtained in a simple test will be presented to determine the success of our strategy. We hope this work can contribute to discussions on preventing the fraudulent use of generative AI. Discussing TL methods does not imply endorsing them but rather exploring potential mechanisms that can be maliciously utilized to gain advantages through generative AI, identifying them, and preventing their use. AI-generated texts can harm knowledge management systems, and we highlight the need to include AI-verification tools in such systems. Such tools can enhance the trustworthiness of information in knowledge management systems, improving knowledge externalization and sharing processes. Finally, we expect to promote future research on TL techniques and the proposition of new strategies to combat fraudulent practices in scientific research.

Keywords: Text Laundering, Artificial Intelligence, Generative AI, Conversational AI, AI-generated Text

1. Introduction

Artificial intelligence (AI) has brought forth the capability to generate texts that mimic the style and content of those produced by humans (Kreps, McCain and Brundage, 2022; Barbosa, 2023). Conversational AI has gained considerable prominence in the last few years among the various types of AI applications developed. Chatbot applications such as OpenAI's ChatGPT (Generative Pre-trained Transformer) or Google's Bard focus on creating human-like interactions and responses in a conversation and are now utilized as tools by many authors as they can help facilitate natural and engaging text-based interactions, providing assistance, generating creative content, and supporting various language-related tasks.

While the use of chatbots can be beneficial in empowering humans to develop new products, exploring a wide range of problems and solutions, and consequently increasing innovation performance (Bouschery, Blazevic and Piller, 2023), it is necessary to consider the ethical implications when it comes to their employment even when classified as potential transformative agents in different domains that require textual production (Stokel-Walker and Van Noorden, 2023). In particular, ethical concerns associated with preventing the reproduction of plagiarized texts or the dissemination of erroneous information must be addressed (Pavlik, 2023). An important question regarding the creation of texts through generative AI systems, i.e., systems capable of generating new content, such as text, images, or music, based on patterns learned from large datasets, is to recognize whether a particular text is or is not the result of AI. A possible benefit of concealing generative AI would be that using more humanized AI-generated texts can enhance user interaction, promoting a better understanding and acceptance of conveyed information, as it resembles natural writing (Bozkurt *et al.*, 2021). Nevertheless, the use of this type of technology is not always beneficial.

In the context of higher education, the integration of generative AI within Higher Education Institutions (HEIs) introduces some ethical considerations. While these tools hold promise in enhancing research processes and knowledge dissemination within academia, their implementation raises concerns regarding academic integrity and undergraduates' skills building and evaluation (Chan and Hu, 2023; Michel-Villarreal *et al.*, 2023; Wang, 2023).

Generative AI technologies offer unprecedented capabilities in content creation, aiding researchers and educators in generating diverse materials for study and publication. In an era where the labour market evolves rapidly, HEIs face the imperative to adapt their curricula to meet industry demands (Dos Santos *et al.*, 2023). This pressure often leads to the integration of cutting-edge technologies like generative AI into educational practices. However, their introduction forces HEIs to reassess their educational curricula (Dai, Liu and Lim, 2023), as generative models could potentially overshadow critical thinking and originality in students and researchers work (Spector and Ma, 2019).

The rapid adoption of such tools underscores the importance of fostering skills in prompt engineering among undergraduate students (Dai, Liu and Lim, 2023). As generative AI systems may produce erroneous or misleading outputs (Fui-Hoon Nah *et al.*, 2023), individuals must possess the theoretical knowledge to identify and correct these inaccuracies effectively. Moreover, the rapid advance of generative AI adoption forces a shift in educational focus towards equipping learners with the competencies to engage with AI technologies critically (Dai, Liu and Lim, 2023).

In this rapidly evolving landscape of generative AI adoption, this work aims to introduce the concept of Text Laundering (TL), which we define as such due to its similarity to the practice of money laundering — applying various consecutive alterations to AI-generated text to make its origin indeterminate. This concept not only sheds light on the potential vulnerabilities of AI-generated content detection tools but also underscores the inadequacy of relying solely on such tools to deter the fraudulent use of generative AI by undergraduate students or researchers. We hope to present some preliminary thoughts on this novel concept, develop it, and discuss how to determine whether a text was produced by humans or by AI and subsequently laundered to create ways for discovering AI-generated laundered texts. We expect this work can serve as a foundation for future discussions on preventing the fraudulent use of generative AI in HEIs context, raising the discussion whether identifying fraudulent use of AI should focus on identifying AI-generated texts or on discerning whether the guiding intelligence behind the AI-generated text is human or artificial. Discussing TL methods doesn't mean supporting them. Instead, we explore how the utilization of text laundering methods demonstrates the inefficacy of current AI-generated text identification techniques in detecting the malicious use of the tool for undergraduates and researchers to gaining unfair advantages in their work.

The text laundering of an AI-generated text employs techniques – often automated – to make it more challenging to identify the text's AI origin. How to generate texts more similar to those produced by a human author is already a highly debated and intriguing topic in natural language processing (Camacho-Collados and Pilehvar, 2020). Therefore, there is a need to develop literature for a practice that has been taking place and to assist in improving the detection of the artificial origin as AI systems improve and texts generated by humans and AI tend to look more alike.

Including AI-generated laundered texts in corporate knowledge management systems can have detrimental consequences. For example, it may lead to challenges in accurately externalizing tacit knowledge and documenting lessons learned. Therefore, AI-generated data could result in the loss of nuance or context, leading to less effective knowledge-sharing and decision-making within the organization. Additionally, reliance on AI may introduce biases or errors into the knowledge management system, compromising its effectiveness and reliability.

This work shows how to develop a concealment method for the artificial origin of a text produced by generative AI tools against AI detection tools. We will then discuss ethical questions regarding this process, particularly concerning fraudulent scientific research practices. This topic is of significant educational, academic, and social relevance, as these tools and their growing interest are essential in all these areas. We also hope this work can encourage future research on TL techniques and that researchers can propose new strategies to combat fraudulent practices in scientific research.

2. Theoretical Background

This section explores the theoretical foundations necessary to understand generative AI methodologies used to generate text and Stylometry, which analyze linguistic patterns and determine authorship. Next, we discuss the ethical concerns associated with AI applications, particularly plagiarism and the proliferation of false or misleading information. Finally, we present applications for detecting AI-generated content.

2.1 Generative AI

Generative AI refers to a class of artificial intelligence systems designed to create or generate new content, such as images, text, audio, or even video, that is convincingly similar to the data it has been trained on. Unlike traditional AI models for classification or prediction tasks, generative AI models focus on creating new content. These models typically utilize techniques like deep learning, particularly Generative Adversarial Networks (Creswell *et al.*, 2018) and Variational Autoencoders (Doersch, 2021), to produce output that exhibits characteristics similar to the training data. Generative AI has applications in various fields, including art, design, entertainment, and even scientific research, where it can be used to generate synthetic data for experimentation or to assist in creative endeavors. These systems can also employ techniques such as natural language processing when dealing with text generation (Mou, 2022). The impact of AI on copyright and related rights is also a topic of discussion, particularly regarding the recognition of authorship and responsibility for copyright infringement (Bysaga, Byelov and Zaborovskyi, 2023).

In the academic field, researchers can use generative AI for the automated production of academic texts, such as abstracts, essays, and even dissertations and theses. Additionally, these systems can deliver adaptive learning content by interacting with students through text and voice (Barbosa, 2023). However, generative AI tools are most commonly employed to produce and modify texts. Their ability to generate original and contextually relevant content from a given dataset has significantly advanced natural language processing. Generative text AI tools become widely used as they can produce fluent and contextually relevant texts in just a few seconds, allowing users to generate new content or make their writing process more efficient and agile.

To give some examples of tools, we here will focus on OpenAI's. One of the various generative AI models used to create or edit texts is the GPT-3.5, a transformer-based language model capable of generating coherent and context-relevant texts. GPT-3.5 is trained using a large amount of textual data for translation, text summarization, and dialog generation, among other applications. ChatGPT is a specific implementation of the GPT-3.5 model focused on conversation. It can answer questions, provide information, assist with problem-solving, and engage in natural language dialogues. Another example is OpenAI Codex (OpenAI, 2024). The Codex is a language model trained on programming code and can generate code from natural language descriptions. It can assist developers in writing code by providing suggestions and auto-completing code snippets. Yet another example is TextGPT, a variation of GPT-3.5 specifically designed to work with natural language text that can generate stories, essays, articles, and poems, among others. All these tools can create and modify texts, exemplifying how AI can assist in text creation and editing across various domains.

2.2 Stylometry

A possible method to detect AI-generated content is Stylometry. Stylometry is the field of study that analyzes text's linguistic and stylistic characteristics. Stylometry gained strength by developing computational tools designed explicitly for this purpose (Boto Bravo, 2018), and the possible applications are many: plagiarism detection, countering identity deception, email impersonation detection, authorship detection in SMS messages, identifying speech writers, multi-modal authentication on mobile devices, among others (Neal *et al.*, 2018). Tools such as Stylo R have been used to analyze and classify literary works based on their typological and stylistic characteristics (Boto Bravo, 2018).

Furthermore, computational statistical tools such as ContaWords have also been used to identify stylistic elements in literary texts (Ferreira Barrocal, 2021). Machine learning techniques can also be used for authorship attribution and characterization (Deutsch, 2020). In addition, critical editions of literary works have been produced to examine their plot, attribution, and historical context (Ferreira Barrocal, 2022). In digital humanities, data mining techniques can visualize concepts, author constellations, and indexed citations of digital collections (Rota, 2022). Recurrent neural networks generate obfuscated texts, preserving their original semantics (Franco and Oliveira, 2019). Finally, the automatic detection of fake news was explored (Santos, 2022). These studies

demonstrated applying computational and statistical methods to analyze and comprehend texts. Researchers can use stylometric analysis methods to investigate the possible artificial origin of a text.

There are still challenges in authorship analysis using stylometry techniques, such as a lack of sufficient data or inconsistencies in text samples (Neal *et al.*, 2018). Achieving high accuracy in stylometric analysis, especially with large and diverse datasets, remains a significant challenge. Still, some studies report accuracy rates ranging from 53.8% to 98.312% for tasks like authorship attribution and verification (Neal *et al.*, 2018). The accuracy of stylometric analysis can vary depending on various factors such as the specific task being performed (e.g., authorship attribution, verification, profiling), the size and quality of the dataset, the features and algorithms used, and the complexity of the writing style being analyzed (Neal *et al.*, 2018).

2.3 Ethical Issues

Among the many issues concerning generative AI applications is the issue of plagiarism. If an AI system generates a text that is similar or identical to a pre-existing text, will it be considered a misappropriation of someone else's work? Chatbots frequently do not mention who published it or even from what source (website, book, magazine, etc.) the information came. Still, the model training may use data from specific sources without acknowledging the original authors.

Generative AI also presents ethical challenges regarding its dual modes of operation: autonomously generating content and synthesizing user input. In the former, the AI may inadvertently produce text resembling existing works, raising concerns of plagiarism. In the latter, the AI acts as a facilitator, crafting text based on user input, blurring lines of authorship and authenticity. As such technologies become more widespread, addressing these dilemmas and establishing guidelines for responsible use and content attribution becomes increasingly urgent.

Another issue is the authenticity and authorship of a text. AI-generated texts often present themselves as though a human wrote them, potentially resulting in a lack of authenticity. The authorship of the texts generated by the AI must be appropriately attributed and disclosed to avoid deception or manipulation (Rossetti and Angeluci, 2021).

Regarding the manipulation of information, AI has the potential to create texts similar to authentic and reliable texts. However, they can also increase the proliferation of false or misleading information. This AI capability can lead to the spread of misinformation and undermine public trust in AI-produced texts.

Another important point is responsibility and supervision, as the production of texts by AI also raises questions about who will be responsible for the generated content. Who would be responsible for ensuring the accuracy, integrity, and ethics of texts produced by AI? Careful monitoring and oversight are required to ensure that the texts generated meet ethical and quality standards.

Finally, there is still the possibility of ideological or discriminatory biases in the text. Even today, AIs are trained on large datasets, which may contain built-in biases and biases. These biases can be reflected in texts generated by AI, reproducing or expanding inequalities, discrimination, or societal stereotypes.

2.4 AI-Generated Content Detection

An AI detection tool is software capable of identifying, recognizing, or classifying specific objects, patterns, behaviors, or information in data or images. Some of these tools may use AI to perform stylometric text analysis.

As generative AI tools become more accessible to the public, AI-generated content becomes more common. To be able to determine if AI produced an article or content is something fundamental to content marketers. Fortunately, some methods exist to assess and determine if the content is AI-generated.

One way to identify this AI-generated content is to look for repetitive or unusual patterns in the text. AI text generation is imperfect, i.e., it may not look human in a human or automated analysis. Sometimes, texts tend to repeat sentences or structures. This repetition may indicate that the content was generated by AI rather than a human author. AI content detection tools may also use machine learning and natural language processing techniques to determine whether the content was created by a human or not in an AI versus AI clash. Machine learning techniques can make computer systems more efficient with every text analyzed. AI applications designed to detect the use of AI in texts can also analyze large amounts of information available on the web about a specific content topic in seconds. This analysis can potentially make the detection of AI-generated content more precise. Several free AI content detection tools are available to help identify whether certain content was generated by AI or by a human author; one example is GPTZero.

GPTZero (GPTZero, 2024) is a tool designed to identify text generated by artificial intelligence, specifically by large language models like ChatGPT. It is helpful when understanding the source of a piece of writing is important, such as academic integrity and content authenticity. GPTZero analyzes the statistical properties of text and its Stylometry, such as how random or predictable it is. Based on this analysis, it can tell with some accuracy whether the text was written by a human or generated by a large language model. It's important to note that GPTZero detection is not perfect. As AI language models continue to develop, GPTZero's ability to detect them might require updates.

AI Detector (Corrector App, 2024) is an application designed to scan and analyze texts to identify the use of AI technology. The tool analyzes words like *the*, *it*, and *is,* which tend to be often used by AI text generators. They claim to search for unnatural transitions and errors in the text and the lack of the author's personal opinions since AI-generated texts often present heavy facts in their contents.

3. Methodology

The proposed methodology aims to create techniques that modify an AI-generated text to hide its origin as artificial text. The approach combines four main stages: the setup stage, the text generation stage, the text laundering stage, and the AI-detection & analysis stage.

In the setup stage, we define the tools used in the process. We need to select three types of tools: the generative AI used to produce the text, the tools responsible for altering the generated text, and finally, the tools used to validate if the modified text is detectable as human-made or AI-generated.

In the text generation stage, we use the generative AI (selected in the setup stage) to create a Corpus of text. The AI employed in this stage is supposed to generate coherent and relevant content, but it is also expected to be detected by AI-generated text detection tools.

In the text laundering stage, the text is processed through single or multiple tools (also selected in the setup stage) to resemble a human writing style. We denominate the tools used to alter the AI-generated text to achieve this goal as *text laundering tools*. The idea behind this stage is that sequential laundering processes like translation and paraphrasing will introduce nuances and subtleties often found in human-written texts while removing evidence from known biases derived from AI-generated text. Therefore, we expect to trick the detection tool and statistical analysis.

In the AI-detection & analysis stage, we use the altered text as an input to AI-generated text detection tools (also selected in the setup stage) to assess the success of the laundering. At the end of this stage, we analyze if the results show consistent difficulty in identifying the text as AI-generated. If the results are unsatisfying, we can return to the setup step to modify the tools used in the text laundering process to improve the results. We summarize the described process in Figure 1.

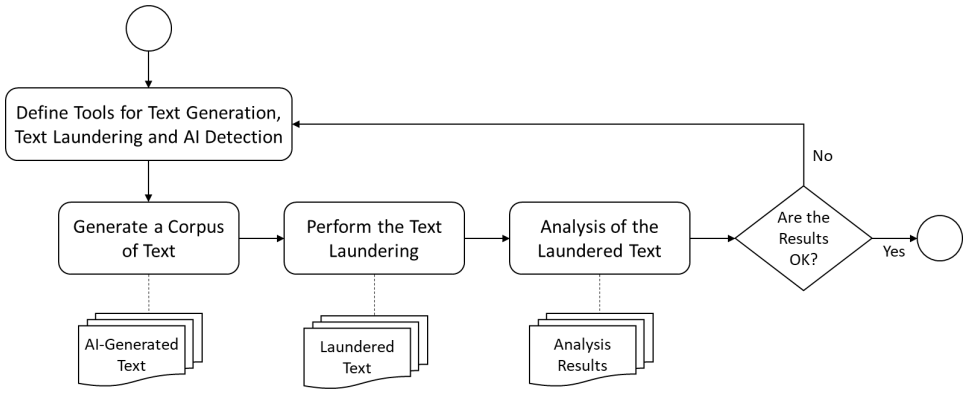

Figure 1: The proposed methodology for the development of Text Laundering techniques.

4. Experiment

In this section, we conduct an experiment on text laundering using the proposed methodology. In the setup stage, we chose ChatGPT as generative AI used to produce the text, Google Translate and Scite as text laundering tools, and GPTZero and AI Detector to validate if the altered text is detectable as human-made or AI-generated. We selected ChatGPT for its popularity and ability to generate quality content in its free version. Google Translate

and Scite were chosen to show that several types of applications can be used for text laundering. Finally, we selected GPTZero and AI Detector because they can detect ChatGPT-produced text.

In the text generation stage, we give ChatGPT the following prompt: "I need you to generate ten topics from different Knowledge Areas and, for each Knowledge Area, a famous and complex topic. I just want the topics; I don't need to generate any text about them." The response was (1) Quantum Physics -Schrödinger's Cat Paradox, (2) Philosophy - JohnRawls's theory of justice, (3) Biology - CharlesDarwin's Theory of Evolution, (4) Mathematics -Fermat's Last Theorem, (5) History - The Cold War and the nuclear arms race, (6) Psychology - SigmundFreud's theory of personality, (7) Sociology - The concept of consumer society, (8) Law - The ethical dilemma of using torture to obtain information, (9) Medicine - The discovery and development of vaccines, and (10) Computer science - The P versus NP hypothesis in computational complexity theory. We then asked ChatGPT to generate a text of approximately 300 words about each topic, one at a time. These texts served as the corpus of text for this experiment.

In the text laundering stage, we use the Google Translate and Scite tools to modify the original text to make it more human-like. In this process, we used Google Translate to perform a double translation (change the text to another language and then back to English). We use Scite to paraphrase and rewrite the text. We end up with three versions of each text: the original version (before the laundering process), the Google Translation laundered text, and the Scite laundered text. We highlight that we can use any tool to create any post-processing workflow to disguise the text, including AI-generating tools (such as ChatGPT) to perform the laundering. We also highlight that we can achieve better results using multiple laundering tools in sequence.

In the AI-detection & analysis stage, we sent each text to the GPTZero and AI Detector to analyze their performance. These tools analyze and identify language patterns generated by AI models, allowing us to confirm the authorship of the texts. The results of the detection experiment are shown in Table 1. We now decide if the results are good *enough* to achieve our goals. If we find the results unsatisfactory, we can revisit the setup stage to adjust the tools utilized in the process or the workflow of laundering tools employed.

Table 1: Results of the experiment.

Text ID	Original Text		Laundered (Google Translate)		Laundered (Scite)	
	GPTZero [AI/Human]	AI Detector [% fake]	GPTZero [AI/Human]	AI Detector [% fake]	GPTZero [AI/Human]	AI Detector [% fake]
Text 1	AI	18.82%	Human	19.94%	AI	8.82%
Text 2	AI	49.97%	AI	57.23%	AI	9.97%
Text 3	AI	49.97%	AI	37.37%	AI	9.97%
Text 4	Human	0.08%	Human	7.38%	Human	0.08%
Text 5	AI	34.67%	AI	65.36%	AI	4.67%
Text 6	AI	1.55%	AI	9.97%	AI	0.55%
Text 7	AI	0%	Human	9.97%	AI	0%
Text 8	AI	0.02%	AI	0.02%	Human	0.02%
Text 9	Human	0.41%	Human	0.03%	Human	0.41%
Text 10	Human	46.70%	Human	41.52%	AI	6.70%

Analyzing the results, we observe that the original AI-generated texts vary in human-likelihood. The original ChatGPT text passes as human in 3 of 10 texts in GPTZero without changes. Similarly, the original text is below *10% fake* in 5 of 10 texts. Regarding the text laundered with Google Translate, we see improvement in the GPTZero assessment, with 5 of 10 texts labeled as human-generated. Meanwhile, in most cases, the AI Detector considered the Google Translate results worse than the original text. Regarding the text laundered with Scite, we see a significant improvement in the AI Detector assessment, with all texts below *10% fake*. However, in the GPTZero results, the Scite laundering kept the 3 of 10 texts labeled as human-generated.

5. Discussion

Our results showed that text laundering can help make texts less artificial and more authentic in the eyes of AI detection systems like the AI Detector. However, we highlight that the effectiveness of this laundering can vary depending on the chosen tool. Our experiment involving the analysis of 10 texts found that four were considered less artificial by the AI Detector after undergoing text laundering. This result indicates that the appropriate choice of laundering tool impacts text detection, making them more genuine and less susceptible to being labeled as artificial.

However, we highlight that the detection methods used by AI Detector systems can vary. Some may be more sensitive to specific laundering techniques, while others may have different rules and algorithms. Therefore, no universally effective method of text laundering works equally well for all cases. Each text and context requires careful analysis to determine the appropriate choice of laundering. We should consider factors such as the type of text, the target audience, the communication objective, and the specificities of the detection system used. For example, in some cases, removing specific language patterns or adding contextual elements may be necessary to decrease the text's authenticity.

We recognize that there is no universally effective approach. Each case requires an individualized analysis, considering specific detection methods and the characteristics of the text in question. The discussion about the ethical implications related to the choice of text laundering tool and its influence on text detection is of utmost importance. Several ethical issues in this context deserve consideration (Bysaga, Byelov and Zaborovskyi, 2023).

One key point is using text laundering to deceive or manipulate artificial intelligence detection systems. Using a text laundering tool to circumvent these systems and provide misleading or false information can produce severe ethical consequences. For example, the spread of misinformation undermines the reliability of information and compromises informed decision-making. Furthermore, the choice of text laundering tool can also raise ethical concerns about privacy and consent. If text laundering involves manipulating or altering other people's texts without their knowledge or consent, it can violate privacy and copyright rights. It is essential to respect individual rights and ensure that any text laundering process is carried out ethically and transparently (Rossetti and Angeluci, 2021). Another important ethical implication is the possibility of bias and discrimination. If specific text laundering tools are more effective in reducing the detection of artificial texts produced by specific groups or communities, it can lead to unfair discrimination. It is crucial to ensure that the choice of laundering tool is unbiased and does not perpetuate or amplify existing prejudices (Temsah *et al.*, 2023). Additionally, we highlight ethical implications related to the reliability and transparency of artificial intelligence detection systems. If the systems cannot identify laundered texts effectively, it can undermine trust in AI technologies and create an environment conducive to spreading false information (Rossetti and Angeluci, 2021).

Although, concerns regarding the detection and prevention of AI-generated content emerge. While these measures can preserve academic integrity, they may inadvertently suppress the creative potential of Generative AI within educational contexts. Rather than solely relying on detecting AI-generated content, HEIs should prioritize educating students on responsible AI usage and encourage the ethical application of Generative AI as a supplemental tool in their learning journey. Therefore, the use of text laundering techniques and tools can serve to demonstrate that alternative methods of detecting fraudulent use of Generative AI should be employed, focusing more on evaluating the ability of undergraduate students to be the guiding minds behind AI-generated texts rather than whether the texts were generated by AI.

Finally, we highlight the necessity of carefully examining the ethical implications of using a text laundering tool. It is necessary to consider the impact on information reliability, privacy, consent, discrimination, and transparency of AI systems. Ensuring that text laundering is conducted ethically and responsibly is essential to promote integrity and trust in AI technologies. The discussion between assistance and complete substitution in the context of AI technologies raises essential questions about acceptable limits and the ethical implications involved. On one hand, AI assistance can bring significant benefits in various areas, such as healthcare, education, and work. The ability of AI technologies to assist humans in complex tasks, provide valuable insights, and enhance efficiency is undeniable. Assistance can involve automating repetitive tasks, analyzing large datasets, and even assisting in decision-making.

6. Conclusions

AI can create texts that closely mimic human writing. However, there's ongoing debate regarding the credibility and impact of AI-generated texts in foreign policy. Platforms like ChatGPT are being investigated for their

potential to produce written content across various fields. However, identifying AI-generated texts and dealing with the ethical considerations they raise pose significant challenges.

This work proposes a methodological approach to develop methods to mask the artificial origin of AI-generated texts, contribute to the development of more effective verification tools, present text laundering techniques, and address associated ethical issues, particularly within the context of HEIs.

Based on the results obtained, we can conclude that the choice of text laundering tool significantly impacts the detection of texts generated by artificial intelligence. Of the ten texts analyzed, we found that four were considered less artificial by the AI Detector after the laundering process. This result indicates that adequate laundering can positively affect detection.

This work contributes significantly to Knowledge Management by developing methods to mask the artificial origin of AI-generated texts, and the approach shows an open issue that must be solved with the development of more effective AI-verification tools, prioritizing the evaluation of the intelligence behind the generated text over the evaluation of AI-generated text. We expect such verification to enhance the trustworthiness and authenticity of information within Knowledge Management Systems and improve the knowledge externalization and sharing processes.

However, we highlight that detection methods vary and require a case-by-case analysis to determine the most appropriate wash. No universally effective method works in all situations. Therefore, it is critical to carefully consider the different laundering methods available and adapt them according to the study's specific needs.

This study has limitations worthy of mentioning. First, we use only a single text-laundering step, which can limit the effectiveness of the process. Furthermore, we analyzed only ten texts, which may not represent all possible variations and results. We also considered only one text generation tool, two laundering tools, and two detection tools, which may restrict the scope of the results.

Future research can explore multiple laundering steps, increase the number of analyzed texts, and include various generation, laundering, and detection tools. This research will help to gain a more complete and accurate understanding of the effects of these processes on detecting AI-generated text. Furthermore, conducting tests on several detection tools would be important to assess their effectiveness in detecting false or misleading information. Automating the laundering process would be crucial to increase efficiency and scalability. Finally, knowledge management systems can integrate AI-detection tools to improve their knowledge externalization and sharing processes.

References

Barbosa, C.R.D.A.C. (2023) 'Transformações no Ensino-Aprendizagem com o Uso da Inteligência Artificial: Revisão Sistemática da Literatura', *RECIMA21 - Revista Científica Multidisciplinar - ISSN 2675-6218*, 4(5), p. e453103. Available at: https://doi.org/10.47820/recima21.v4i5.3103.

Boto Bravo, M.A. (2018) 'Mapa estilométrico de la narrativa de Eduardo Mendoza: aproximación a un análisis estilistico computacional de textos literarios', *Epos : Revista de filología*, (33), p. 99. Available at: https://doi.org/10.5944/epos.33.2017.18281.

Bouschery, S.G., Blazevic, V. and Piller, F.T. (2023) 'Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models', *Journal of Product Innovation Management*, 40(2), pp. 139–153. Available at: https://doi.org/10.1111/jpim.12656.

Bozkurt, A. *et al.* (2021) 'Artificial Intelligence and Reflections from Educational Landscape: A Review of AI Studies in Half a Century', *Sustainability*, 13(2), p. 800. Available at: https://doi.org/10.3390/su13020800.

Bysaga, Y.M., Byelov, D.M. and Zaborovskyi, V.V. (2023) 'Artificial intelligence and copyright and related rights', *Uzhhorod National University Herald. Series: Law*, 2(76), pp. 299–304. Available at: https://doi.org/10.24144/2307-3322.2022.76.2.47.

Camacho-Collados, J. and Pilehvar, M.T. (2020) 'Embeddings in Natural Language Processing', in *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*. *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, Barcelona, Spain (Online): International Committee for Computational Linguistics, pp. 10–15. Available at: https://doi.org/10.18653/v1/2020.coling-tutorials.2.

Chan, C.K.Y. and Hu, W. (2023) 'Students' voices on generative AI: perceptions, benefits, and challenges in higher education', *International Journal of Educational Technology in Higher Education*, 20(1), p. 43. Available at: https://doi.org/10.1186/s41239-023-00411-8.

Corrector App (2024) *AI Checker Free*. Available at: https://corrector.app/ai-content-detector/ (Accessed: 30 March 2024).

Creswell, A. *et al.* (2018) 'Generative Adversarial Networks: An Overview', *IEEE Signal Processing Magazine*, 35(1), pp. 53–65. Available at: https://doi.org/10.1109/MSP.2017.2765202.

Dai, Y., Liu, A. and Lim, C.P. (2023) 'Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education', *Procedia CIRP*, 119, pp. 84–90. Available at: https://doi.org/10.1016/j.procir.2023.05.002.

Deutsch, C. (2020) *Atribuição autoral com uso de classificadores de perfis*. Mestrado em Sistemas de Informação. Universidade de São Paulo. Available at: https://doi.org/10.11606/D.100.2020.tde-04112020-152132.

Doersch, C. (2021) 'Tutorial on Variational Autoencoders'. arXiv. Available at: https://doi.org/10.48550/arXiv.1606.05908.

Dos Santos, H.S. *et al.* (2023) 'A Framework for Assessing Higher Education Courses Employability', *IEEE Access*, pp. 1–1. Available at: https://doi.org/10.1109/ACCESS.2023.3256722.

Ferreira Barrocal, J. (2021) 'Petrarquismo en «Los amores de Marte y Venus», una fábula mitológica del siglo XVI', *Hipogrifo. Revista de literatura y cultura del Siglo de Oro*, 9(2), pp. 823–841. Available at: https://doi.org/10.13035/H.2021.09.02.58.

Ferreira Barrocal, J. (2022) 'Edición crítica de la «Loa con que empezó Escamilla», atribuida a Calderón', *Hipogrifo. Revista de literatura y cultura del Siglo de Oro*, 10(2), pp. 735–757. Available at: https://doi.org/10.13035/H.2022.10.02.45.

Franco, A. and Oliveira, L. (2019) 'Avaliação de Modelos de Redes Neurais Recorrentes para Anonimização de Textos em Português', in *Anais do XIX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg 2019)*. *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, Brasil: Sociedade Brasileira de Computação, pp. 421–426. Available at: https://doi.org/10.5753/sbseg.2019.13992.

Fui-Hoon Nah, F. *et al.* (2023) 'Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration', *Journal of Information Technology Case and Application Research*, 25(3), pp. 277–304. Available at: https://doi.org/10.1080/15228053.2023.2233814.

GPTZero (2024) *GPTZero*, *GPTZero*. Available at: https://gptzero.me/ (Accessed: 30 March 2024).

Kreps, S., McCain, R.M. and Brundage, M. (2022) 'All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation', *Journal of Experimental Political Science*, 9(1), pp. 104–117. Available at: https://doi.org/10.1017/XPS.2020.37.

Michel-Villarreal, R. *et al.* (2023) 'Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT', *Education Sciences*, 13(9), p. 856. Available at: https://doi.org/10.3390/educsci13090856.

Mou, L. (2022) 'Search and learning for unsupervised text generation', *AI Magazine*, 43(4), pp. 344–352. Available at: https://doi.org/10.1002/aaai.12068.

Neal, T. *et al.* (2018) 'Surveying Stylometry Techniques and Applications', *ACM Computing Surveys*, 50(6), pp. 1–36. Available at: https://doi.org/10.1145/3132039.

OpenAI (2024) *OpenAI Codex*. Available at: https://openai.com/blog/openai-codex (Accessed: 30 March 2024).

Pavlik, J.V. (2023) 'Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education', *Journalism & Mass Communication Educator*, 78(1), pp. 84–93. Available at: https://doi.org/10.1177/10776958221149577.

Rossetti, R. and Angeluci, A. (2021) 'Ética Algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação', *Galáxia (São Paulo)*, (46), p. e50301. Available at: https://doi.org/10.1590/1982-2553202150301.

Rota, A.R. (2022) 'El uso de la minería de datos como heurística para la teoría de la historia y la historia de la historiografía', *Amoxtli* [Preprint], (7). Available at: https://doi.org/10.38123/amox7.205.

Santos, R.L.D.S. (2022) *Detecção automática de notícias falsas em português*. Doutorado em Ciências de Computação e Matemática Computacional. Universidade de São Paulo. Available at: https://doi.org/10.11606/T.55.2022.tde-14072022-165613.

Spector, J.M. and Ma, S. (2019) 'Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence', *Smart Learning Environments*, 6(1), pp. 8, s40561-019-0088-z. Available at: https://doi.org/10.1186/s40561-019-0088-z.

Stokel-Walker, C. and Van Noorden, R. (2023) 'What ChatGPT and generative AI mean for science', *Nature*, 614(7947), pp. 214–216. Available at: https://doi.org/10.1038/d41586-023-00340-6.

Temsah, O. *et al.* (2023) 'Overview of Early ChatGPT's Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts', *Cureus* [Preprint]. Available at: https://doi.org/10.7759/cureus.37281.

Wang, T. (2023) 'Navigating Generative AI (ChatGPT) in Higher Education: Opportunities and Challenges', in C. Anutariya et al. (eds) *Smart Learning for A Sustainable Society*. Singapore: Springer Nature Singapore (Lecture Notes in Educational Technology), pp. 215–225. Available at: https://doi.org/10.1007/978-981-99-5961-7_28.