UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

CoGnosis
Revista de Ciencias de la Educación
ISSN 2588-0578

Herramientas para detectar el Plagio a la Inteligencia Artificial: ¿cuán útiles son?

Tools to detect Plagiarism in Artificial Intelligence: how useful are they?

Autores

[1]*Dariel Díaz Arce*

[1]Unidad Educativa Santana, Cuenca, Azuay, Ecuador.

Como citar el artículo:
Díaz Arce, D. (2024). Herramientas para detectar el Plagio a la Inteligencia Artificial: ¿cuán útiles son? Revista Cognosis. ISSN 2588-0578, 9(2). https://doi.org/10.33936/cognosis.v9i2.6195

Enviado: 2023-10-25
Aceptado: 2024-01-15
Publicado: 2024-04-05

Resumen

Las inteligencias artificiales (IA) generativas se abren paso cada vez más en el ámbito educativo, aunque no siempre utilizadas de forma honesta. La preocupación por el uso de estas aplicaciones para crear textos académicos también crece entre docentes de los diferentes niveles educativos y también en el ámbito de las publicaciones científicas. Poder detectar estos trabajos creados por la IA constituye por tanto una prioridad a la que se dio paso en esta investigación. Para ello, se diseñó una investigación cuasiexperimental de evaluación del desempeño diagnóstico de herramientas digitales que se promueven como "antiplagio" para las IA: Copyleaks, AI Text Classifier, Crossplag, Content at scale, Hive moderation. Se conformó un grupo control con trabajos de estudiantes de bachillerato con una antigüedad de 7 a 8 años y otro experimental con documentos generados por IA. Se comparó el rendimiento de las diferentes herramientas mediante los indicadores de pruebas diagnósticas: sensibilidad, especificad, valores predictivos y el índice de validez. Los resultados indican que Copyleaks posee una alta sensibilidad, pero baja especificad, mientras que en las demás ocurre lo contrario. Este trabajo refuerza la necesidad de desarrollar más herramientas o estrategias para detectar este problema entre los estudiantes.

PALABRAS CLAVE: plagio académico; ciberplagio; inteligencia artificial; herramientas antiplagio; educación.

Abstract

The use of generative artificial intelligence (AI) is increasingly common in the educational field, although not always used honestly. This is why concern about the use of these applications to create academic texts is also growing among teachers at different educational levels, which also extends to the field of scientific publications. Detecting the jobs created by AI is therefore a priority that was given way in this research. A quasi-experimental investigation was designed to evaluate the diagnostic performance of digital tools that are promoted as "anti-plagiarism" for AI: Copyleaks, AI Text Classifier, Crossplag, Content at scale, Hive moderation. A control group was organized with works by high school students aged 7 to 8 years and another experimental group was organized with documents generated by AI. The performance of the anti-plagiarism tools was compared using the diagnostic test indicators: sensitivity, specificity, predictive values and the validity index. The results show that Copyleaks has high sensitivity, but low specificity, while the opposite is true for the others. This work reinforces the need to develop more tools or strategies to detect this problem among students.

KEYWORDS: academic plagiarism; cyberplagiarism; artificial intelligence; anti plagiarism tools; education.

CoGnosis
Revista de Ciencias de la Educación

Publicación arbitrada trimestral
Universidad Técnica de Manabí. Ecuador
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.cognosis@utm.edu.ec

INTRODUCCIÓN

En los últimos años el desarrollo vertiginoso de la inteligencia artificial (IA) ha desembocado en un sinnúmero de herramientas disponibles en Internet. De esta forma abundan las aplicaciones para disímiles tareas tales como la generación de textos, códigos, imágenes, videos, audios, presentaciones digitales, hasta la realización de composiciones musicales, entre muchas otras. Por ello, es difícil no solo ver en ellas un futuro prometedor sino también posibles riesgos a un nivel que superan nuestros límites de entendimiento actual.

En el ámbito académico, por ejemplo, el debate sobre los desafíos y utilidades de estas herramientas emerge en un ambiente que resalta la falta de preparación de los docentes para utilizarlas eficientemente, así como el riesgo de deshonestidad académica por parte de los estudiantes (Cotton, Cotton, & Shipway, 2023). En esto último, la IA generativa tendría una suerte o papel de "escritor fantasma" o "ghost writer", donde el trabajo a realizar es encargado a un "tercero" que cede o no puede reclamar sus derechos sobre la creación (Padillah, 2023).

Si bien la incidencia del plagio a la IA aún no está bien estudiada, su mal uso se hace notar ya en estudiantes (Chan & Hu, 2023), así como en el ámbito de los artículos científicos publicados en revistas indexadas (Miller, et al., 2023). A lo anterior se une la preocupación de que los textos creados por estas herramientas son difíciles de detectar y de que los docentes no cuentan con suficientes estrategias para identificar o prevenir este fenómeno. En tal caso, algunos estudios previos ya indican la poca capacidad de programas "antiplagio" comerciales y populares como Turnitin para distinguir la originalidad entre las asignaciones de sus alumnos (Perkins, 2023; Diaz Arce, 2023a).

Con lo anterior, urge la necesidad de contar con aplicaciones que puedan ser útiles y accesibles para la detección del plagio IA. En el mercado existen múltiples de ellas, pero muy pocas investigaciones serias encaminadas a evaluar su funcionalidad en condiciones reales. Por ello, en este trabajo se evalúa la utilidad de cinco de estas herramientas digitales para detectar textos creados por IA a través de parámetros del desempeño de pruebas diagnósticas o de detección.

DESARROLLO

El diseño de la investigación fue quasiexperimental con un enfoque cuantitativo en la evaluación del desempeño de instrumentos de detección de textos generados por IA. Los indicadores empleados fueron: sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo y el índice de validez. El estudio se realizó durante los meses de enero a marzo de 2023. El diseño fue similar a un trabajo previo para evaluar herramientas antiplagio gratuitas (Diaz Arce, 2017).

Se incluyeron 55 trabajos considerados como grupo control conformado por ensayos académicos de estudiantes de bachillerato desde el año 2013 a 2016 elegidos al azar (n=45) y por fragmentos de artículos científicos propios de esos años (n=10). En este periodo las inteligencias artificiales generativas no estaban muy difundidas para uso libre por la sociedad, por lo que se consideran así a los documentos en este grupo sin plagio IA o negativos. No se consideraron otros tipos de deshonestidad que no fuera solamente el plagio IA. Por otro lado, el grupo experimental se incluyeron 45 ensayos creados por diferentes herramientas generadoras de contenido sobre los mismos temas de los ensayos del grupo control.

e-ISNN 2588-0578 | Vol. 9, Núm. 1 (144-150): Abril - Junio 2024 | DOI: 10.33936/cognosis.

La elección de las herramientas de inteligencia artificial generativas de texto se siguieron algunos criterios básicos. Primero se hizo un acercamiento a docentes y estudiantes de bachillerato de la institución en la que se les preguntó sobre las IA que conocen de este tipo. Al mismo tiempo, esto se complementó con una búsqueda libre en Google con las palabras clave "inteligencia artificial" and "generación de textos" así como "inteligencia artificial" and "creación de textos". Se filtraron posteriormente por aquellas que permitían crear al menos tres documentos en su versión libre de pago con una extensión mínima de 250 palabras. Las IA utilizadas al final fueron: ChatGPT 3.5 (https://chat.openai.com/) , Content (https://www.contents.com/), Copy (https://www.copy.ai/), Copymatic (https://copymatic.ai/), Dupla (https://www.dupla.ai/), Escribelo (https://escribelo.ai/), Hypotenuse (https://www.hypotenuse.ai/), Perplexity (https://www.perplexity.ai/), Smodin (https://smodin.io/), Writesonic (https://writesonic.com/), You (https://you.com/). Del total de documentos creados, 29 fueron en español y 16 en inglés.

En cuanto a las herramientas de detección de textos generados con IA, se emplearon cinco de las más mencionadas en internet. Estas se filtraron con los siguientes criterios: acceso gratuito al momento en que se realizó el estudio, soporte de diferentes idiomas, permita analizar tal menos res trabajos por día, reporte de algún indicador que permita evaluar si el texto fue escrito o no por una IA, tener al menos un límite de 500 palabras para analizar. Los trabajos con más de esa cantidad de palabras se analizaron por partes. Las aplicaciones seleccionadas así fueron: Copyleaks (https://copyleaks.com/es/), AI Text Classifier (https://freeaitextclassifier.com/), Crossplag (https://crossplag.com/), Content at Scale (https://contentatscale.ai/ai-content-detector/), Hive Moderation (https://hivemoderation.com/). Adicionalmente, todos los trabajos fueron pasados por Turnitin para estimar el Índice General de Similitud (IGS).

Para el análisis de los resultados se tomaron como positivos todos los documentos que dieran un resultado como probable o posiblemente escrito por una IA, según los datos aportados por cada herramienta. De este modo se construyeron tablas de contingencia 2x2 para analizar el poder de detección de cada aplicación, calculando la sensibilidad como la fracción de documentos con plagio IA real detectada por la herramienta del total experimental. Por su parte la especificidad sería la fracción de documentos del grupo control detectados como libres de plagio. La precisión en la detección o índice de validez se estimó como el porcentaje de trabajos correctamente clasificados del total analizado. Asimismo, se calcularon los valores predictivos positivo y negativo respectivamente como la fracción de los casos de plagio propuestos por la herramienta que realmente lo son, y la fracción de los documentos propuestos como libres de plagio que realmente lo son (Diaz Arce, Beltran, & Cueva Sarmiento, 2018). Se utilizó además la prueba Chi-cuadrado para evaluar la hipótesis de independencia entre lo propuesto y las herramientas de detección y los datos reales. Para comparar la extensión de palabras y el IGS se utilizó el test U de Mann-Whitney. El nivel de significancia fue de 0.05. Para estos cálculos se utilizaron las herramientas SPSS v. 23 y Epidat 3.0.

La extensión media por número de palabras fue similar en ambos grupos: experimental (664) vs. control (694), $p = 0.252$. En cuanto al IGS tampoco hubo diferencias significativas con 38 % vs. 49 % para el grupo experimental y control respectivamente ($p = 0.558$).

El análisis del desempeño de las herramientas para la detección de plagio IA se muestra en la tabla 1. En todos los casos se acepta que existe asociación entre los resultados propuestos por las herramientas para detectar plagio IA y la descripción real de la muestra dado que los valores de p son significativamente menores a 0.05.

En cuanto al desempeño general, Copyleaks presenta una sensibilidad y valor predictivo negativo cercanos al 90 %. Lo anterior significa que con esta herramienta se pueden detectar al menos nueve de cada diez trabajos creados con una IA, al mismo tiempo que en la muestra de estudio, de cada diez casos que propone como libres de plagio, nueve realmente lo son. Por su parte, AI Text Classifier, posee valores predictivos e índice de validez que superan el 70 %. Tales resultados implican que al menos siete de cada diez trabajos de la muestra son correctamente clasificados, al mismo tiempo de más del 70 % de los casos que se predicen con problemas

e-ISNN 2588-0578 | Vol. 9, Núm. 1 (144-150): Abril - Junio 2024 | DOI: 10.33936/cognosis.

CoGnosis
Revista de Ciencias de la Educación

Publicación arbitrada trimestral
Universidad Técnica de Manabí. Ecuador
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.cognosis@utm.edu.ec

de plagio o libres de ello, realmente lo son. Las demás aplicaciones se caracterizan por una baja o muy baja sensibilidad, aunque con una elevada especificidad y valores predictivos positivos.

Tabla 1. Sensibilidad, especificidad e índice de validez de las herramientas empleadas para detectar plagio IA.

Herramienta	Plagio IA	Experimental	Control	S (%)	E (%)	IV (%)	VPP (%)	VPN (%)	p
Copyleaks	Sí	41	20	91.1	63.6	76	67.2	89.7	<0,001
	No	4	35						
AI Text Classifier	Sí	26	7	57.8	87.3	74	78.8	71.6	<0,001
	No	19	48						
Crossplag	Sí	6	0	13.3	100	61	100	58.5	0,007
	No	39	55						
Content at scale	Sí	17	2	48.6	96.4	70	89.5	65.4	<0,001
	No	28	53						
Hive moderation	Sí	18	2	40	96.4	71	90.0	66.3	<0,001
	No	27	53						

Fuente: S, sensibilidad; E, especificidad; IV, índice de validez; VPP, valor predictivo positivo; VPN, valor predictivo negativo.

El desarrollo de la inteligencia artificial en los últimos años ha sido tal que se ha iniciado un debate importante respecto a sus usos en diversas ramas de la ciencia, la tecnología y la sociedad, incluyendo a la educación, donde al mismo tiempo que tiene oportunidades también posee desafíos significativos (Adeshola & Adepoju, 2023; Kung, y otros, 2023). Uno de esos grandes desafíos está en lo referente a la honestidad académica pues el uso indebido de estas herramientas para la generación de contenidos podría no ser aceptable por el riesgo de plagio que representa (Mohammadkarimi, 2023; Cotton, Cotton, & Shipway, 2023; Diaz Arce, 2023a). Por ello, es vital disponer de herramientas y estrategias que permitan identificar con cierto nivel de confianza a las asignaciones realizadas por los estudiantes con ayuda de las IA sin autorización.

El presente trabajo es uno de los primeros en el análisis del desempeño de diferentes herramientas en la detección de trabajos creados enteramente a partir de una IA generativa. En este caso se reporta que Copyleaks parece ser la más adecuada pues detecta la mayoría de los ensayos con dificultades de plagio IA dada su alta sensibilidad. A pesar de ello, para este proceso propone muchos falsos positivos (>30 %), por lo que el índice de clasificación correcta de los trabajos no supera el 80 %.

Los resultados del presente estudio contrastan con los reportados en el sitio web de esta aplicación, quienes aseguran que esta herramienta posee hasta un 99,1 % de precisión para detectar contenido generado por múltiples IAs y en diferentes idiomas, con una tasa de falsos positivos de 0.2 %) (Not All AI Detectors are Created Equal, n.d.). Las diferencias podrían enmarcarse significativamente en las muestras estudiadas, pues

e-ISNN 2588-0578 | Vol. 9, Núm. 1 (144-150): Abril - Junio 2024 | DOI: 10.33936/cognosis.

ellos citan a ChatGPT 4 y Bard, mientras que en el presente trabajo se utilizan otras diez IA generativas que no son mencionadas explícitamente en este sitio. Todo ello denota lo difícil que es la detección del plagio IA entre las tareas o asignaciones de los estudiantes, considerando que podría convertirse en un problema significativo a futuro.

Lo anterior se hace aún más relevante si se considera que la prevalencia de uso de las IA para generar textos académicos no está muy estudiada. En tal caso, un reporte de marzo de 2023 por la propia empresa que diseñó a Copyleaks estimada una frecuencia que oscilaba entre 5.9 % y 10.4 % entre estudiantes de colegio y bachillerato. No obstante, este en este sitio se reconoce que de forma general estos valores pueden estar entre 2.36 % y 35.61 % a nivel mundial (Prevalence of AI-Generated Content in Education , 2023) o ser aún mayores (Chan & Hu, 2023). Incluso en trabajos de investigación ya publicados en revistas científicas arbitradas la incidencia de este problema podría superar el 20 % con una tendencia al aumento desde el año 2020 (Miller et al., 2023).

Por otro lado, hasta el momento de esta investigación no se reportaban datos para América Latina. Un trabajo aislado en jóvenes de una institución educativa de Ecuador indicó que aproximadamente un 4.8 % ha utilizado al menos una vez estas herramientas para crear sus ensayos académicos a inicios del año 2023 (Diaz Arce, 2023b). No obstante, no se descarta que esta frecuencia haya aumentado por la diversidad, popularidad y fácil acceso a las mismas a través de redes sociales y otros entornos virtuales.

En la literatura se encontró un estudio similar que evalúa el desempeño de la herramienta Originality.AI (Collingwood, Ontario, Canadá) para detectar trabajos académicos generados por IA. En dicho reporte, la sensibilidad y especificidad fueron significativamente superiores a las de todas las aplicaciones evaluadas en el presente trabajo, siendo de 100 % y 95 % respectivamente (Miller, et al., 2023). Sin embargo, los autores solo consideraron una única IA para generar contenidos, y solo trabajos en inglés, a diferencia del presente en el que la muestra fue mixta a partir de diferentes aplicaciones, con textos tanto en español como en inglés.

Por su parte, Turnitin habilitó también su herramienta para detectar trabajos realizados con ayuda de las IA, indicando que posee una precisión del 98 % (Turnitin habilita la función de detección de escritura con Inteligencia Artificial, 2023), lo que es muy superior al de todas las herramientas reportadas en este trabajo. A pesar de ello, no se dan muchos datos de cómo se realizó esta evaluación, la muestra analizada, y, por último, solo analiza los trabajos en inglés. De esta forma en otro estudio se observó que esta herramienta solo pudo detectar un 54.8 % de trabajos generados por ChatGPT4 con un IGS similar respecto a los trabajos originales (Perkins, 2023). Esto refuerza la idea previa de que se requieren estrategias y programas más eficaces que ayuden a discernir entre el plagio IA y la originalidad.

Asimismo, hay que recalcar que en el presente estudio ni en ninguno de los reportes consultados sobre el tema se toma en cuenta el uso combinado de estas herramientas con otras de parafraseo, que podrían enmascarar aún más este problema. Asimismo, se debe tener en cuenta que la muestra experimental para el trabajo puede afectar significativamente el desempeño de cualquier herramienta, lo que sugiere la realización de estudios de validación por separado. En este sentido, se debe tomar en cuenta que la prevalencia de este problema podría afectar las estimaciones de los valores predictivos. De este modo, los VPP aumentarían cuando aumenta la frecuencia de casos reales, y lo opuesto sería para el VPN (Diaz Arce, Beltran, & Cueva Sarmiento, 2018).

e-ISNN 2588-0578 | Vol. 9, Núm. 1 (144-150): Abril - Junio 2024 | DOI: 10.33936/cognosis.

CoGnosis
Revista de Ciencias de la Educación

Publicación arbitrada trimestral
Universidad Técnica de Manabí. Ecuador
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.cognosis@utm.edu.ec

CONCLUSIONES

Las herramientas "antiplagio" evaluadas en el presente trabajo no mostraron un nivel aceptable de rendimiento para detectar la escritura por IA generativas. Copyleaks fue la que mejor se desempeño en cuanto a sensibilidad, mientras las demás mostraron valores de especificidad elevados. No obstante, no se puede descartar la utilidad de estas herramientas para identificar el plagio de IA generativas específicas.

Por último, este trabajo deja abiertas otras interrogantes significativas más allá del ¿cuál es el desempeño de otras herramientas digitales para detectar el plagio IA?, ¿qué estrategias utilizan los docentes para detectar los textos generados por IA más allá del uso de herramientas informáticas?, ¿cuán frecuente es el plagio IA y qué herramientas son las que más se utilizan por los estudiantes?, ¿cómo prevenir esta forma de deshonestidad académica?, ¿cómo promover una escritura creativa sin necesidad de las IA?, ¿cómo utilizar realmente las IA en el ámbito de la redacción académica sin que esto pueda considerarse plagio? Estas y otras interrogantes pueden ser parte de investigaciones futuras sobre este tema.

El autor declara que no existen conflictos de intereses que afecten el normal desarrollo de la evaluación del manuscrito.

e-ISNN 2588-0578 | Vol. 9, Núm. 1 (144-150): Abril - Junio 2024 | DOI: 10.33936/cognosis.

REFERENCIAS BIBLIOGRÁFICAS

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. Interactive Learning Environments. doi:10.1080/10494820.2023.2253858

Chan, C. K., & Hu, W. (2023). Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. ArXiv. Retrieved from https://arxiv.org/ftp/arxiv/papers/2305/2305.00290.pdf

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in, 1-12. doi:10.1080/14703297.2023.2190148

Diaz Arce, D. (2017). Evaluación del desempeño de tres herramientas antiplagio gratuitas en la detección de diferentes formas de Copy-Paste procedentes de Internet. Edutec: Revista Electrónica de Tecnología Educativa(59), a354. doi:https://doi.org/10.21556/edutec.2017.59.812

Diaz Arce, D. (2023a). Inteligencia artificial vs. Turnitin: implicaciones para el plagio académico. Cognosis, 8(1), 15-26. Retrieved from https://doi.org/10.33936/cognosis.v8i1.5517

Diaz Arce, D. (2023b). Plagio a la Inteligencia Artificial en estudiantes de bachillerato: un problema real. Revista Innova Educación, 5(2), 108-116. Retrieved from https://www.revistainnovaeducacion.com/index.php/rie/article/view/845

Diaz Arce, D., Beltran, P., & Cueva Sarmiento, J. (2018). ¿Son suficientes los indicadores del rendimiento de una prueba o test diagnóstico para evaluar su desempeño? Revista Cubana de Medicina General Integral, 34(3), 94-109. Retrieved from Revista Cubana de Medicina General Integral: https://www.medigraphic.com/pdfs/revcubmedgenint/cmi-2018/cmi183k.pdf

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De León, L., Elepaño, C., . . . Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Plos Digital Health, 2(2), e0000198. doi:https://doi.org/10.1371/journal.pdig.0000198

Miller, L. E., Bhattacharyya , D., Miller, V. M., & Bhattacharyya , M. (2023). Recent trend in artificial intelligence-assisted biomedical publishing: a quantitative bibliometric analysis. Cureus, 15, e39224. Retrieved from https://doi.org/10.7759/CUREUS.39224.

Mohammadkarimi, I. (2023). Teachers' reflections on academic dishonesty in EFL students' writings in the era of artificial intelligence. Journal of Applied Learning & Teaching, 6(2), 1-9. doi:https://doi.org/10.37074/jalt.2023.6.2.10

Not All AI Detectors are Created Equal. (n.d.). Retrieved from Copyleaks: https://copyleaks.com/ai-content-detector

Padillah, R. (2023). Ghostwriting: a reflection of academic dishonesty in the artificial intelligence era. Journal of Public Health, fdad169. doi:https://doi.org/10.1093/pubmed/fdad169

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic: ChatGPT and beyond. Journal of University Teaching & Learning Practice, 20(2). Retrieved from https://doi.org/10.53761/1.20.02.07

Prevalence of AI-Generated Content in Education . (6 de March de 2023). Retrieved from Copyleaks: https://copyleaks.com/blog/prevalence-of-ai-generated-content-in-education

Turnitin. Contacto de prensa. (4 de Abril de 2023) "Turnitin habilita la función de detección de escritura con Inteligencia Artificial para apoyar a educadores e instituciones académicas. Retrieved from Turnitin: https://www.turnitin.com/es/press/turnitin-habilita-la-funcion-de-deteccion-de-escritura-con-inteligencia-artificial-para-apoyar-a-educadores-e-instituciones-academicas