Polo del Conocimiento

Pol. Con. (Edición núm. 83) Vol. 8, No 6
Junio 2023, pp. 1737-1762
ISSN: 2550 - 682X
DOI: 10.23857/pc.v8i6

Modelos de minado de texto para la implementación de sistemas de predicción de plagio de la Universidad Técnica de Manabí

Text mining models for the implementation of plagiarism prediction systems at the Technical University of Manabí

Modelos de mineração de texto para a implementação de sistemas de previsão de plágio na Universidade Técnica de Manabí

Dario Xavier Mieles Macias [I]
dmiles0735@utml.edu.ec
https://orcid.org/0000-0001-8689-8218

Ermenson Ricardo Ordoñez Avila [II]
ermensonrodoñez@gmail.com
https://orcid.org/0000-0003-2583-2076

Correspondencia: dmiles0735@utml.edu.ec

Ciencias Técnica y Aplicadas
Artículo de Investigación

*** Recibido:** 23 de abril de 2023 ***Aceptado:** 12 de mayo de 2023 *** Publicado:** 12 de junio de 2023

I. Estudiante de la carrera de Ingeniería en Sistemas informáticos, Universidad Técnica de Manabí, Ecuador.
II. Magíster en Gestión de Sistemas de Información e Inteligencia de Negocios, Ingeniero en Sistemas Informáticos, Facultad de Ciencias Informáticas, Universidad Técnica de Manabí, Ecuador.

Resumen

El presente estudio tiene como propósito analizar los modelos de minado de texto para la implementación de sistemas de predicción de plagio como herramientas modernas que deben ajustarse a los desafíos complejos de este problema de crecimiento continuo. Para ello se realizó una revisión sistemática de literatura enmarcada en parámetros PRISMA para selección de artículo y reducción de sesgo, identificación de cadenas de búsqueda en bases de datos como ACM, Science direct, IEEE xplore, Scopus considerando criterios de enfoque y contenido para evaluar cada artículo seleccionado. Entre las técnicas de minería de texto fueron más comunes los clasificadores específicamente, las redes neuronales y los árboles de decisión, también se identificaron técnicas de agrupamiento. El sistema de detección de plagio más utilizado es Turnitin, el modelo de minería más utilizado son las redes recurrentes (LSTM) cuya precisión fue del 100%, la recuperación de 97%, exactitud del 99% y una detección de plagio del 94%. En conclusión, las Universidades e institutos se han visto en la necesidad de implementar procesos de detección de plagio a través del uso de sistemas de detección, se ha considerado el empleo de técnicas de minería de texto que facilitan la detección y reconocimiento de elementos, similitudes, coincidencias y semejanzas que aportan en la comprobación de plagio en textos académicos; las redes recurrentes han presentado mejores resultados en diversos escenarios de detección, por ello, se sugieren como modelo de minería de datos de tipo predictivo.

Palabras Clave: Minería de texto; predicción; plagio; software antiplagio; publicaciones académicas.

Abstract

The purpose of this study is to analyze text mining models for the implementation of plagiarism prediction systems as modern tools that must be adjusted to the complex challenges of this continuously growing problem. For this, a systematic review of the literature was carried out framed in PRISMA parameters for article selection and bias reduction, identification of search strings in databases such as ACM, Science direct, IEEE xplore, Scopus considering focus and content criteria to evaluate each study. selected item. Among the text mining techniques, specifically classifiers, neural networks and decision trees were more common, clustering techniques were also identified. The most used plagiarism detection system is Turnitin, the most

used mining model is recurring networks (LSTM) whose accuracy was 100%, recovery 97%, accuracy 99% and plagiarism detection 94%. In conclusion, Universities and institutes have seen the need to implement plagiarism detection processes through the use of detection systems, the use of text mining techniques has been considered that facilitate the detection and recognition of elements, similarities , coincidences and similarities that contribute to the verification of plagiarism in academic texts; recurrent networks have presented better results in various detection scenarios, therefore, they are suggested as a predictive data mining model.

Keywords: Text mining; prediction; plagiarism; anti-plagiarism software; academic publications.

Resumo

O objetivo deste estudo é analisar modelos de mineração de texto para a implementação de sistemas de previsão de plágio como ferramentas modernas que devem ser ajustadas aos complexos desafios desse problema crescente. Para isso, foi realizada uma revisão sistemática da literatura enquadrada nos parâmetros PRISMA para seleção de artigos e redução de viés, identificação de strings de busca em bases de dados como ACM, Science direct, IEEE xplore, Scopus considerando critérios de foco e conteúdo para avaliar cada estudo. item selecionado. Entre as técnicas de mineração de texto, especificamente classificadores, redes neurais e árvores de decisão foram mais comuns, técnicas de agrupamento também foram identificadas. O sistema de detecção de plágio mais utilizado é o Turnitin, o modelo de mineração mais utilizado é redes recorrentes (LSTM) cuja precisão foi de 100%, recuperação 97%, precisão 99% e detecção de plágio 94%. Em conclusão, Universidades e institutos têm visto a necessidade de implementar processos de detecção de plágio através do uso de sistemas de detecção, foi considerado o uso de técnicas de mineração de texto que facilitam a detecção e reconhecimento de elementos, semelhanças, coincidências e semelhanças que contribuem para a verificação de plágio em textos acadêmicos; redes recorrentes têm apresentado melhores resultados em vários cenários de detecção, portanto, são sugeridas como um modelo preditivo de mineração de dados.

Palavras-chave: Mineração de texto; predição; plágio; software antiplágio; publicações acadêmicas.

Introducción

La tecnología ha permitido generar escenarios de información que favorecen los nuevos conocimientos; sin embargo, la gran cantidad de datos que se encuentran en la Web se ha convertido en un arma de dos caras, especialmente en el campo de la investigación académica donde resulta indispensable el buen manejo de la información como una habilidad que contribuya con la localización y uso eficiente de la información (Michán y Álvarez, 2019).

Los formatos digitales y el acceso abierto a gran cantidad de información forman parte de la revolución informática (reconocida como un proceso innovador que ha experimentado con datos científicos) donde los datos constituyen un nuevo recurso valioso que no sólo se genera e impulsa, sino que, además, se comercializa. Por ello, cada vez, existe mayor interés por la creación de enfoques, herramientas, métodos y aplicaciones computacionales innovadores orientados a la caracterización, estudio, sistematización, estructuración, entre otros, para obtener nuevo conocimiento, resolver problemas y tomar decisiones en base al resultado de los procesos informáticos que manejan esos datos (Venkatakrihnan et al., 2016).

En el campo de la investigación, cada vez existe un mayor desafío por parte de las Universidades para aprobar los trabajos investigativos realizados por los estudiantes, pues si bien es un proceso que demanda indagación, pruebas, comprobación y análisis, en la práctica no siempre se cumplen todas esas fases; al contrario, la dinámica actual de los estudiantes con el advenimiento de la era digital, las demandas sociales que exigen cada vez mayor grado de preparación académica a jóvenes profesionales y la deshonestidad académica se ha convertido en una realidad que atenta directamente contra las investigaciones originales y confiables, pues se trata de un problema de principios éticos-morales producto de las nuevas características adquiridas en el plagio académico derivadas de la era digital (Rogerson y McCarthy, 2017).

A este respecto, Llovera (2023) indica que, el "uso de los diferentes recursos e información en formato digital ha conducido al estudiantado a buscar formas más rápidas para realizar sus trabajos académicos y, por ello, incurrir muchas veces en la práctica conocida como ciberplagio" lo cual ocurre de forma consciente (copia y pega de Internet) e inconsciente, esta última cuando se desconoce la debida norma de citación como APA, Vancouver, IEEE, etc., cuya aplicación es fundamental, especialmente cuando se ha parafraseado el texto de referencia.

En este contexto, el plagio académico ha cobrado especial relevancia en el campo de la investigación universitaria, especialmente a partir de casos que han involucrado figuras públicas

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

como congresistas, funcionarios públicos y hasta presidentes (Navarro, 2023); a esto, se suma información como la resultante de la encuesta del Programa Universitario de Bioética realizado por la Universidad Autónoma de México que revela cómo un 52% de académicos de dicha casa de estudios que ha sido testigo de plagio académico por parte de sus colegas en procesos de investigación para titulación de pregrado, postgrado y hasta doctorado (Cruz, 2023).

Frente a este creciente problema del ciberplagio, el mismo que ocurre en el contexto académico en investigaciones realizadas por estudiantes de educación superior, se ha incrementado el uso de programas y sistemas informáticos por parte de las universidades y revistas científicas para la detección de coincidencias y patrones que puedan evidenciar plagio en el material que se presente ante las autoridades universitarias en virtud de evitar investigaciones fraudulentas y generar las respectivas sanciones o correctivos necesarios para mantener la confianza y validez de los trabajos que se aprueben para su futura publicación.

La minería de textos forma parte de esas soluciones informáticas que se han perfeccionado con el paso de los años y las innovaciones tecnológicas que han ocurrido, pues se trata de un subconjunto de la minería de datos útil para extraer información de datos no estructurados y, a su vez, detectar grupos, tendencias, asociaciones y derivaciones de patrones a partir de técnicas basadas en el procesamiento de textos como la "lingüística computacional y la recuperación de información" las cuales se aplican tanto en la fase de pre-procesamiento, donde los textos se transforman en un tipo de representación semiestructurada, previo a la fase de descubrimiento, donde se detectan agrupamientos, asociaciones, desviaciones o tendencias (Gil, 2021).

Este proceso de descubrimiento se realiza mediante el uso de métodos de aprendizaje automático, estadísticos, matemáticos o artificiales para explorar en grandes bases de datos (Mancilla et al., 2020) que, de otra forma, no se podrían analizar. Cuando se hace referencia a esta técnica, es preciso entender que la minería de datos puede ser descriptiva o predictiva; en el primer caso, se trata de aquella que encuentra patrones y relaciones en los datos utilizando técnicas de asociación y agrupamiento, mientras que, en el segundo caso se trata de aquellas que predicen el valor particular de un atributo a partir de otros atributos enfocadas en algoritmos de clasificación y regresión (Santamaria, 2015).

En el contexto universitario, en donde se desenvuelven los estudiantes de la Universidad Técnica de Manabí, es necesario explorar las alternativas tecnológicas que permitan y garanticen una

adecuada revisión de las publicaciones de sus estudiantes, reduciendo el plagio, y a su vez, alcanzando niveles adecuados de calidad.

Es por ello que, describiendo las características técnicas y metodológicas de las herramientas utilizadas para el diseño e implementación de soluciones orientadas a la predicción del plagio, se obtendrían mejores márgenes de confiabilidad en los procesos de revisión de las producciones intelectuales elaboradas en el seno de esta casa de estudio.

Esta revisión sistemática de literatura tiene como objetivo explorar los modelos de minado de texto utilizados en sistemas de predicción de plagio en instituciones de educación superior. Para alcanzar este propósito, se formularon las siguientes preguntas de investigación:

RQ1. ¿Qué técnicas de minería de texto se han utilizado para predecir el plagio en publicaciones académicas?, RQ2. ¿Cuáles son los sistemas de predicción de plagio utilizados en instituciones de educación superior?, RQ3. ¿Cuáles son los modelos de minería de datos, con mejores indicadores de rendimiento, implementados en sistemas de predicción de plagio en universidades?

Finalmente, es preciso indicar que este trabajo de investigación presenta el orden que sugiere el modelo PRISMA, para revisiones sistemáticas de literatura: introducción, método, resultados, discusión y conclusiones.

Método

El presente artículo de revisión sistemática parte de la aplicación de los parámetros PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses por sus siglas en inglés) para este tipo de investigaciones en el que se considera fundamental utilizar la lista de verificación al momento de seleccionar los artículos y publicaciones que conformarán la sistematización, así como la determinación de una estrategia de búsqueda que responda a dichos parámetros con la finalidad de reducir los sesgos informativos.

En este caso, la estrategia de búsqueda utilizada inició con la determinación de los criterios de búsquedas, entre los cuales destacan como criterios de inclusión: publicaciones, artículos de revisión sistemática, artículos originales y artículos de revisión bibliográfica; asimismo, se consideraron las publicaciones realizadas en revistas de alto impacto, redactadas en inglés o español, cuyo contenido sea completo y disponible, publicaciones realizadas en un período de 2015-2023, que compartan más de una palabra clave.

Por su parte, los criterios excluyentes se enmarcaron en: publicaciones incompletas, tesis doctorales, monografías o libros, investigaciones publicadas en revistas no indexadas o de bajo impacto, publicaciones realizadas en otro idioma distinto al inglés o español, que no comparten variables o palabras claves, publicadas antes del 2015.

Posterior a la determinación de los criterios de búsqueda, se especificaron las palabras claves a utilizar en base a las principales variables de investigación las cuales fueron: “Plagio”, “minería de texto”, “aprendizaje automático” “técnicas de minería de datos”, “predicción de plagio”, “algoritmos de predicción”, “educación superior” en español y, “Plagiarism”, “text mining”, “machine learning”, “data mining techniques”, “plagiarism prediction”, “prediction algorithms”, “higher education” en inglés. Estas palabras junto a los criterios de búsqueda orientaron la indagación a través de los buscadores de alto impacto tales como: ACM, Science direct, IEEE Xplore, Scopus y Google Academy. Con estos términos clave, se diseñó la cadena de búsqueda ideal (Tabla 1).

Tabla 1.- *Cadena de búsqueda por cada buscador*

Base de Datos	**Cadena de Búsqueda**
ACM	[All: plagiarism] AND [[All: prediction] OR [All: detection]] AND [All: "text mining"] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]
Science direct	plagiarism AND (prediction OR detection) AND "text mining"
Google Academy	plagiarism + (prediction OR detection) + "text mining"
IEEE xplore	((plagiarism AND(prediction OR detection) AND "text mining"))
Scopus	**1** (plagiarism AND (prediction OR detection) AND "text mining") AND PUBYEAR > 2014 AND PUBYEAR < 2024 AND PUBYEAR > 2014 AND PUBYEAR < 2024

Para el proceso de revisión y selección de los artículos que conforman la sistematización, se utilizó la lista de verificación para resúmenes estructurados de PRISMA, haciendo énfasis en los ítems de: título, resumen, objetivos, métodos y resultados, los cuales permitieron llevar a cabo la búsqueda y selección en sus diferentes fases de identificación, cribado, evaluación e inclusión.

Para la evaluación de los artículos primarios se valoraron dos aspectos principales: enfoque y contenido, en tres niveles de acuerdo a los percentiles indicados donde moderado corresponde entre 0 y 40 de aportación, aceptable entre 41 y 80, y finalmente, óptimo entre 81 y 100 (Tabla 2).

Tabla 2.- *Valoración de cada nivel de aporte de los artículos revisados*

Nivel de aporte	**Percentil**
Moderado	0-40
Aceptable	41-80
Óptimo	81-100

En el primer caso, se evaluaron las referencias de sistemas de predicción de plagio y, en el segundo caso, se evaluaron las referencias relacionadas con la minería de datos, cada una de ellas con un conjunto de criterios (Tabla 3) que se ponderaron de acuerdo al aporte que tuvo cada uno de ellos a las variables: Nada (0), Algo (0.5), Cumple Totalmente (1).

Tabla 3.- *Criterios a evaluar por cada aspecto de investigación*

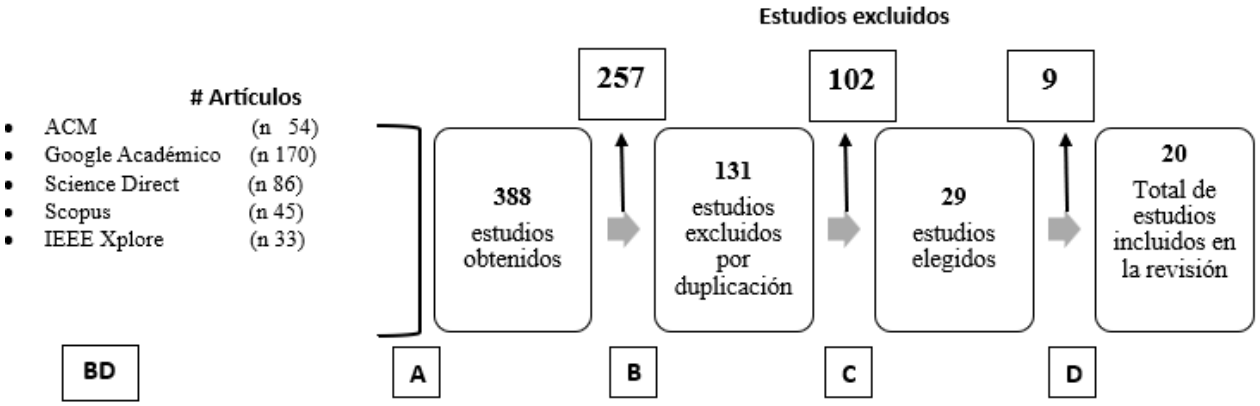
Criterios	**Aspectos**	
	Enfoque	**Contenido**
1	E1-Menciona criterios de predicción de plagio	C1-Metodología de minería de datos utilizadas
2	E2-Presenta indicadores de rendimiento de los algoritmos utilizados	C2-Secciones de limitaciones
3	E3-Describe las técnicas de minerías de textos utilizadas	C3-Propuestas o referencias para el desarrollo de investigaciones futuras

Resultados

Los principales hallazgos de este estudio evidencian que los artículos seleccionados y revisados cumplieron en un 100% con los parámetros de verificación de resúmenes estructurados PRISMA (Figura 1), los cuales se realizaron en su mayoría en el año 2020 en países asiáticos, seguidos de países latinoamericanos y, finalmente, los de Europa, los cuales cumplieron cada uno con las

respectivas palabras claves en relación a: detección de plagio, machine learning, minería de texto, minería de datos, plagio académico, etc., (Tabla 4)

Figura 1. El flujo de búsqueda y selección de artículos.



BD: Base de Datos. A: Extracciones de artículos de la Base de Datos. B: Remover artículos duplicados. C: Leer títulos, resúmenes, y palabras claves. D: Revisión de textos completos,

Tabla 4

Identificación, evaluación y selección de artículos según PRISMA

P	Autor	Año	Título	Lugar del estudio	Tipo de estudio	Palabras Claves Asociadas
P1	Sindhu y Idicula	2017	Plagiarism detection in Malayalam language text using a composition of similarity measures	Singapur	Revisión	Detección de plagio
P2	Duracick et al.	2020	Método optimizado basado en algoritmo K-means como herramienta en la detección de plagio en código fuente	Colombia	Artículo Original	Plagio, algoritmos, código fuente
P3	Qiubo et al.	2019	Research on code plagiarism detection model based on Random	Hong Kong	Artículo Original	Detección de plagio, árbol de decisión

			Forest and Gradient Boosting Decision Tree			
P4	Xylogiannopoulos, et al.	2020	Text mining for plagiarism detection: multivariate pattern detection for recognition of text similarities	España	Artículo Original	Minería de texto y detección de plagio
P5	Viuginov et al.	2020	A Machine Learning based plagiarism detection in source code	China	Artículo Original	Machine learning
P6	Alí et al.	2018	Detection of plagiarism in URDU text documents	Pakistán	Artículo Original	Plagio, algoritmos de clasificación
P7	Mansoor y Al Tamimi	2022	Plagiarism detection system in scientific publication using LSTM networks		Artículo Original	Detección de plagio, minería de texto
P8	Massagram et al.	2018	A novel technique for Thai document plagiarism detection using syntactic parse trees	Tailandia	Revisión	Minería de texto y detección de plagio
P9	Chakrabarty y Roy	2018	An efficient context-aware agglomerative fuzzy clustering framework for plagiarism detection	India	Artículo Original	Minería de texto y detección de plagio
P10	El-Rashidy et al.	2022	reliable plagiarism detection system based on deep learning approaches	Egipto	Artículo Original	Minería de texto y detección de plagio

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

P11	Priya et al.	2019	Plagiarism detection in source code using machine learning	India	Artículo Original	Minería de texto, minería de datos, machine learning y detección de plagio
P12	Perilla, M.	2020	Detección de plagio en código fuente java mediante tokenización y aprendizaje de máquina	Colombia	Artículo Original	Plagio, código fuente, tokenización
P13	Reducindo et al.	2017	Integración de plataformas LMS y algoritmo de código abierto para detección y prevención de plagio en Educación Superior	México	Artículo Original	Plagio académico, algoritmo de detección
P14	Santamaría, W.	2015	Técnicas de minería de datos aplicadas en la detección de fraude: Estado del arte	Colombia	Artículo Original	Detección de fraude, minería de datos
P15	Hany y Gomaa	2022	A hybrid approach to paraphrase detection based on text	Egipto	Artículo Original	Detección de fraude, minería de datos
P16	Huang et al.	2020	Code plagiarism detection method based on code similarity and student behavior characteristics	China	Artículo Original	Detección de plagio, minería de datos
P17	Nennuri, et al.,	2021	Plagiarism detection through data mining techniques	Suiza	Artículo Original	Detección de plagio, minería de datos

P18	Kulkarni et al.	2021	Analysis of Plagiarism Detection Tools and Methods		Revisión Sistemática	Detección de plagio
P19	Shakeel, et al.	2020	A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts	Pakistan	Artículo Original	Detección, minería
P20	Awale et al.,	2020	Plagiarism Detection in Programming Assignments using Machine Learning	Nepal	Artículo Original	Detección de plagio, Minería de textos

Los artículos primarios seleccionados se valoraron por criterio en cada uno de los aspectos evaluados: enfoque y contenido ponderados de acuerdo a su nivel de aportación en cada una de las variables de est6udio (Tabla 5). En este particular, se evidencia un mayor aporte aceptable y óptimo en enfoque, mientras que, en el contenido, el mayor aporte es el moderado. En cuanto a los criterios de enfoque, la presentación de indicadores de rendimiento de los algoritmos utilizados (E2) fue el mejor ponderado; por su parte, el criterio de contenido mejor ponderado fue el de metodología de minería de datos utilizadas (C1) (Figura 2).

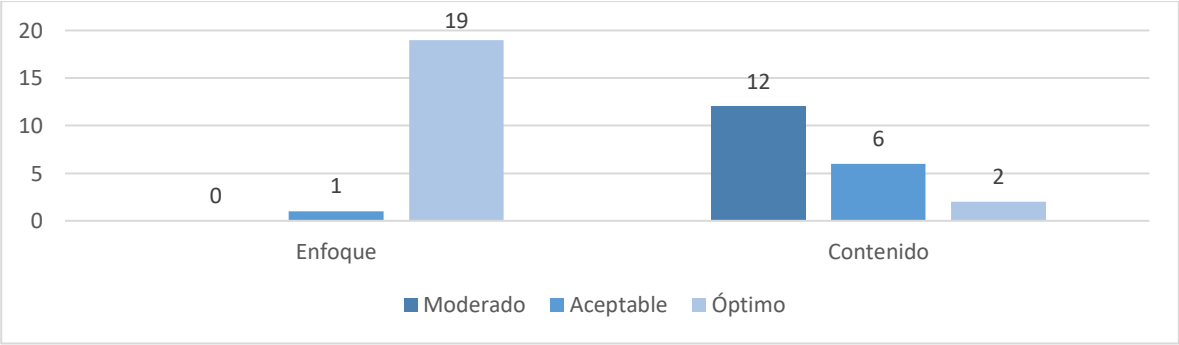
Tabla 5.- *Tabla de valoración según aporte (enfoque-contenido)*

P	**Enfoque**			**%**	**Aporte**	**Contenido**			**%**	**Aporte**
	E1	**E2**	**E3**			**C1**	**C2**	**C3**		
P1	1	1	1	100%	Óptimo	1	0	1	67%	Aceptable
P2	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P3	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P4	1	1	0.5	83%	Óptimo	0.5	0	0.5	33%	Moderado
P5	1	1	0.5	83%	Óptimo	0.5	0	1	50%	Aceptable

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

P6	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P7	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P8	1	1	1	100%	Óptimo	1	1	1	100%	Óptimo
P9	1	1	1	100%	Óptimo	1	0	1	67%	Aceptable
P10	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P11	1	0.5	1	83%	Óptimo	1	0	0	33%	Moderado
P12	1	1	1	100%	Óptimo	1	1	1	100%	Óptimo
P13	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P14	0.5	1	1	83%	Óptimo	1	0	1	67%	Aceptable
P15	1	1	1	100%	Óptimo	1	0	0	33%	Aceptable
P16	1	1	1	100%	Óptimo	1	1	0	67%	Aceptable
P17	1	1	1	100%	Óptimo	1	0	0	33%	Moderado
P18	1	0.5	0.5	67%	Aceptable	0.5	0	0	17%	Moderado
P19	1	0.5	1	83%	Óptimo	1	0	0	33%	Moderado
P20	1	1	1	100%	Óptimo	1	0	0	33%	Moderado

Figura 2.- *Nivel de aporte de los artículos primarios (enfoque-contenido)*



Entre las principales técnicas de minería de texto utilizadas para predecir plagio en las publicaciones académicas destacan los clasificadores de tipo predictivo a través de redes neuronales, árboles de decisiones, redes bayesianas y otros como datos etiquetados de Machine Learning; por su parte, el agrupamiento de tipo descriptivo se presentó a partir del uso de

agrupamiento difuso. En cuanto a los sistemas de predicción de plagio utilizados en las instituciones de educación superior que fueron analizadas, destaca Turnitin como el principal y más común sistema de detección a pesar de que se enuncian otros como Plagscam, Chamilo, Jplag. En cuanto a las características funcionales de las soluciones informáticas que emplean la minería de datos para la predicción de plagio se utilizaron procesos como el algoritmo K-Means, Naïve Bayes, KDD, K-NN, C4.5, clasificación binaria, máquina de soporte vectorial en algunos casos aplicados en WEKA. En la mayoría de los casos los procedimientos realizados se enmarcaron en el análisis, consenso, patrones de comportamiento, tokenización de código fuente, limpieza, extracción, recuperación, agrupamiento, validación y localización de conjuntos.

Por su parte, las soluciones funcionales que se presentaron en los estudios revisados contemplaron la creación de nuevos algoritmos para la obtención de correlaciones entre conjuntos de intemsets relevantes para reducir redundancias (Díaz y García, 2018), detección de plagio de código fuente, mapeo de uso Weka, descubrimiento de conocimiento usando KDD, identificación de datos para caracterizar fenómenos, identificación de diversos tipos de plagio, determinación de variables asociadas, selección, limpieza, transformación y proyección de datos, comparación de niveles de uso de texto, detección de patrones investigativos, detección de fraude, predicción de fracaso escolar, incremento de la eficiencia en la detección de plagio y detección de plagio semántico (Tabla 6).

Tabla 6.- *Principales resultados en técnicas y modelo de minería utilizada*

P	TMD utilizada	TM	Solución funcional	Proceso	Procedimiento
P1	Red Neuronal (clasificador)	Predictiva	Predicción de plagio rápido y con óptima clasificación	algoritmo NLP	Combinación de puntuaciones de similitud
P2	Agrupación	Descriptiva	Detección de plagio de código fuente	K-Means	Herramienta de clasificación previa de vectores

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

P3	Árboles de decisión (clasificador)	Predictiva	Mejor rendimiento para determinar nivel de sospecha del código	Algoritmos Random Forest y Gradient Boosting Decision Tree	Combinación de algoritmos para determinar rango de grado de similitud
P4	Agrupación	Descriptiva	Detección de plagio en bibliotecas digitales de big data, detección de patrones comunes entre documentos bajo inspección y bibliotecas de referencia y detección eficiente de diferentes tipos de plagio	Algoritmo LERP-RSA y ARPaD	Combinado multivariante que mejora la estructura de datos para la detección de patrones
P5	Árbol de análisis comprimido (Clasificadores)	Predictiva	Canalización para clasificar códigos fuente de pares de soluciones para problemas de ACM	AST (Árbol de Sintaxis Abstracta)	Producción de árbol estructurado con diferentes tipos de nodos
P6	Redes Bayesianas (Clasificadores)	Predictiva	Identificar diferentes tipos de plagio, como el reordenamiento de oraciones, la similitud intertextual inerte/borrada y la similitud de copia cercana	Support Vector Machine y Naïve Bayes	Método de concenso

P7	Redes Neuronales (Clasificadores)	Predictiva	Detectar plagios internos y externos, amplía la memoria para aprender de sus experiencias recordando sus entradas.	Algoritmo LSTM (Long - Short Term Memory)	Extensión de redes neuronales recurrentes
P8	Árboles de análisis sintáctico (clasificador)	Predictiva	Identificación de clases semánticas de las oraciones. Mejora la precisión de la detección de plagio	SRL (Semantic Role Labeling)	Etiquetado jerárquico-no secuencial
P9	Agrupamiento difuso (Fuzzy clustering)	Descriptiva	Mejorar solidez y consistencia de resultados para agrupar artículos multidisciplinarios	Enfoque aglomerativo	Construir jerarquía de grupos
P10	Redes Neuronales convolucionales (Clasificador)	Predictiva	Extrae automáticamente características que se utilizarán para la clasificación de objetos	RNN/CNN/ Modelo LSTM	Clasificar y predecir
P11	Datos etiquetados Mahine Learning (Clasificadores)	Predictiva	Determinar presencia o ausencia de plagio, estimar función de densidad de las predictoras, reducir sesgo y	LSTM	Combinación de algoritmos clasificadores para optimizar precisión de resultados

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

			varianza en el contexto de aprendizaje supervisado		
P12	Clasificadores	Predictiva	Detección de plagio de código fuente	SMO usado en WEKA	Tokenización de código fuente
P13	Agrupamiento	Descriptivo	Detección de plagio de código fuente	AAPD	Extracción-recuperación
P14	Agrupamiento, árboles de decisión y redes neuronales	Descriptiva y predictiva	Detección de fraude	K-Means, CART, MLP	Descubrimiento y extracción de conocimiento
P15	Red Neuronal (clasificador)	Predictiva	Predicción de plagio rápido y con óptima clasificación	Algoritmo NLP	Combinación de técnicas de similitud (semántica, de cadena y de incrustración
P16	Árboles de decisión (clasificadores)	Predictivo	Detección de plagio de código basado en similitud del código	Clasificación binaria utilizando SCD (concentración de similitud de código)	Identificar distribución de similitud entre todos los códigos
P17	Redes Neuronales (Clasificadores)	Predictivo	Incrementar la eficiencia en la detección de plagio	Enfoque k-NN	Localización de conjuntos de datos copiados

P18	Redes Neuronales (Clasificadores)	Predictivo	Detección de plagio semántico	Enfoque K-NN	Localización de conjuntos de datos copiados
P19	Redes Neuronales (Clasificadores)	Predictivo	Mejorar el rendimiento de los modelos de aprendizaje profundo y analizar el impacto de varios pasos de aumento de datos	CNN y LSTM	Detección de paráfrasis en textos breves
P20	Árboles de decisión (Clasificador)	Predictivo	Incrementar precisión en el modelo de detección	Algoritmo xgBoost	Predecir pares de código fuentes plagiados

Finalmente, los indicadores de rendimiento mejor valorados en los modelos de minería de datos revisados en los artículos son: la precisión (f-measure) con un 100% en modelos como el enfoque aglomerativo, 99% en la clasificación binaria y 98% en las LSTM; en cuanto a la recuperación, el marco SPT y SRL reportó un 100%, el algoritmo xgBoost 97% y las LSTM un 97%; por su parte, la exactitud tuvo mejor valoración en modelos como las LSTM (99%), xgBoost (94%) y Gradient Boosting Decision Tree (95%) (Tabla 7).

Asimismo, se analizó el nivel de detección de plagio que reportó mejor valoración en las LSTM con un 94% y la especificidad de 98% fue generada utilizando Gradient Boosting Decision Tree; el mejor tiempo fue de 1.64 segundos y lo reportó el SMO.

Tabla 7.- *Indicadores de rendimiento por técnica y modelo de minería*

	Modelo		**Medida de rendimiento**

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

Publicación		**Técnica de minería**	**Precisión**	**Recuperación**	**Exactitud**	**Det. Plagio**	**Tiempo**	**Esp.Sen.**
P1	NLP	Red Neuronal (PNN)	0.93	0.95	**	**	**	****
P2	K-means	Clustering	**	**	**	**	5.2 "	*****
P3	Random Forest y Gradient Boosting Decision Tree	Árbol de decisión	0.202 (RF) 0.929 (GBDT)	**	0.202 (RF) 0.959 (GBDT)	**	**	**1 (RF) 0.864 (GBDT)**
P4	Algoritmo LERP-RSA y ARPaD	Clustering	**	**	**	1	**	****
P6	Support Vector Machine y Naïve Bayes	Redes Bayesianas	0.73 (SVM) 0.71(NB)	0.83 (SVM) 0.80 (NB)	**	**	**	****
P7	Algoritmo LSTM (Long - Short Term Memory)	Redes neuronales	0.98	0.97	0.99	**	**	****
P8	Marco SPT y SRL	Árboles de análisis sintáctico	0.33 (SRL)	1	**	**	**	****

			0.79 (SPT)					
P9	Enfoque aglomerativo	Agrupamiento difuso	1	0.95	**	**	**	**
P10	Modelo LSTM	Redes Neuronales convolucionales (Clasificador)	0.95	0.92	**	0.94	**	**
P11	LSTM	Datos etiquetados Mahine Learning	0.89	0.887	**	0.887	**	**
P12	SMO usado en WEKA	Clasificadores	**	**	**	**	1.64 "	**
P15	NLP	Red Neuronal (PNN)	0.76	**	**	**	**	**
P16	Clasificación binaria	Árbol de decisión	0.99	**	**	**	**	**
P17	Enfoque k-NN	Redes neuronales	Alta	**	**	**	**	**
P18	Enfoque K-NN	Redes neuronales	Alta	**	**	**	**	**
			Alta	**	**	**	**	**
P19	CNN y LSTM	Redes neuronales	0.70	0.80	0.90	**	**	**

Pol. Con. (Edición núm. 83) Vol. 8, No 6, Junio 2023, pp. 1737-1762, ISSN: 2550 - 682X

P20	**Algoritmo xgBoost**	**Árbol de decisión**	**0.95 (NP) 0.89 (P)**	**0.97 (NP) 0.82 (P)**	**0.94 (NP) 0.94 (P)**	**	**	**

** Valores no disponibles

Discusión

Los hallazgos de este estudio muestran que el problema del plagio académico, especialmente, a nivel universitario, es común en todos los continentes y en diversos idiomas, lo que supone un reto para la minería de datos en la creación de algoritmos y programas de detección de plagio que superen las barreras del idioma en virtud de incrementar la precisión en este tipo de recursos informáticos.

En atención a la pregunta de investigación RQ1: ¿Qué técnicas de minería de texto se han utilizado para predecir el plagio en publicaciones académicas? En esta investigación destacaron entre las técnicas de minería de textos utilizadas para predecir plagio en publicaciones académicas los clasificadores de redes neuronales tal como exponen autores como: Sindhu e Idicula (2017); Mansoor y Al Tamimi (2022); El-Rashidy (2022); Perilla (2019); Hany (2022); Nennuri et al., (2021); Kullkarni et al., (2021); Shakeel et al., (2020), los árboles de decisiones referidos por: Awale et al., (2020); Huang et al., (2020); Massagram et al., (2018); Viuginov (2020), Qiubo, (2019); Santamaría (2015) y las redes bayesianas (Alí et al., 2018). En este contexto, los clasificadores son idóneos para identificar coincidencias y generar métricas de similitud, especialmente utilizadas en la detección de plagio porque, específicamente en la identificación de paráfrasis, permiten inferir el contexto adecuado sobre una oración debido a su corta longitud (Hunt et al., 2019).

Asimismo, los árboles de decisión constituyen una potente herramienta de clasificación porque soportan los posibles problemas de clasificación y regresión que puedan surgir en el proceso al tiempo que son más fáciles de comprender; en el caso de las predicciones, permite seleccionar el mejor punto de corte para hacerlas y repetir el proceso hasta alcanzar la profundidad fija deseada (Espinoza, 2018).

En el caso de las redes bayesianas, permiten observar el comportamiento dinámico de un patrón a partir de una aproximación en función de los valores que toman el resto de las variables; en este

sentido, se genera un modelo empírico, inductivo que permite reconstruir un modelo de información real a partir de la propagación de las influencias por esa red bayesiana (Sarmiento y Ocampo, 2023).

Por otro lado, los hallazgos mostraron una incidencia significativa en el uso de técnicas de agrupamiento o clustering, especialmente del agrupamiento difuso, también resulta útil para la detección de plagio al tener la capacidad de pertenecer a más de un grupo, lo cual permite acortar el tiempo de análisis, pues cada uno de los grupos al que pertenece se asocia a un conjunto de niveles de pertenencia que indican la fuerza de asociación entre un dato específico y uno o varios grupos (Villanza et al., 2012).

En cuanto a la pregunta de investigación RQ2: ¿Cuáles son los sistemas de predicción de plagio utilizados en instituciones de educación superior? En relación a los sistemas de predicción de plagio utilizadas en las instituciones de educación superior en todo el mundo, Turnitin es la herramienta antiplagio más común que apoya al docente y a los estudiantes, especialmente cuando se consultan fuentes electrónicas (Moreno, 2018). Este sistema realiza sus búsquedas de similitud entre más de un billón de páginas y sitios de Internet, siendo útil en la reducción de porcentaje de similitud y mejoramiento de los trabajos de investigación académica debido a que permite realizar retroalimentación por parte del docente (Díaz, 2015).

Finalmente, en relación a RQ3: ¿Cuáles son los modelos de minería de datos, con mejores indicadores de rendimiento, implementados en sistemas de predicción de plagio en universidades? En este caso, las redes recurrentes de LSTM (Long Short Term Memory por sus siglas en inglés) fueron las más usadas y mejor valoradas, pues además de presentar elevados niveles de precisión, recuperación, exactitud y detección de plagio, tal como mencionan El-Rashidy et al., (2022); Mansoor y Al Tamimi, (2022); Priya et al., (2019) Shakeel et al., (2020); (Reducindo et al., 2017) son altamente efectivas para tal fin por su capacidad de aprender y recordar secuencias por largos períodos de tiempo debido a la elevada sensibilidad que tienen a los datos de entrada (Sánchez, et al., 2020).

Otro modelo de minería de datos que fue valorado con el 100% de precisión fue el enfoque aglomerativo que se utilizó para mejorar la solidez y consistencia de los resultados en virtud de poder realizar una mejor agrupación de artículos multidisciplinarios para dar respuesta a la integración de características semánticas y alcanzar una mejor y optimizada función (Chakrabarty y Roy, 2018)

El algoritmo xgBoost, también es uno de los modelos mejor valorados con una recuperación del 97% y una exactitud del 94%; en este caso, se utilizó como parte del aprendizaje automático bajo el marco de Gradient Boosting optimizada y distribuida que brinda una elevada eficiencia en la resolución de problemas manejando grandes cantidades de datos con mayor rapidez (Awale et al., 2020).

Limitaciones

Si bien el presente estudio se realizó atendiendo a los parámetros de revisiones sistemáticas con estándares internacionales, el acceso limitado a plataformas con mayor número de artículos con textos completos disponibles dificultó el hallazgo de estudio óptimo, especialmente en relación al contenido. Además, sólo se atendieron estudios en inglés y español, lo cual es otra limitante entendiendo que existen estudios en otros idiomas que se realizan en países desarrollados y más avanzados en esta materia que no fueron revisados y podrían generar información idónea para profundizar en este estudio.

Conclusiones

El plagio académico se ha convertido en uno de los problemas más graves, desde el punto de vista ético, a los que se enfrentan las universidades frente al uso desmedido, poco ético e irresponsable de las publicaciones e información que se encuentran en Internet por parte de los estudiantes que presentan investigaciones que atentan contra el derecho de autor de quienes sí se han tomado la tarea de analizar, profundizar y crear textos científicos de calidad. Ante esta compleja situación, las Universidades e institutos se han visto en la necesidad de implementar procesos de detección de plagio a través del uso de sistemas de detección como es el caso de Turnitin o Urkund; no obstante, el elevado costo de las licencias que autorizan su uso y el incremento en los falsos positivos de estos han contribuido a la necesidad de replantear los sistemas y usos implementados para tal fin.

En consecuencia, se ha considerado el empleo de técnicas de minería de texto que facilitan la detección y reconocimiento de elementos, similitudes, coincidencias y semejanzas que aportan en la comprobación de plagio en textos académicos en estudios universitarios, pues permiten atender a este problema que cada vez crece y se vuelve más complejo de detectar.

Por ello, utilizar modelos que tengan elevados niveles de precisión, exactitud y recuperación constituye una premisa al analizar la idoneidad de estas herramientas para la detección de plagio académico, siendo las redes recurrentes (LSTM) las que han presentado mejores resultados en diversos escenarios de detección, por ello, se sugieren como modelo de minería de datos de tipo predictivo.

Referencias

1. Alí, W., Ahmed, T., Rehman, Z., Rehman, A., Slaman, M. (22 de noviembre de 2018). Detection of plagiarism in URDU text documents. Conferencia internacional sobre tecnologías emergentes (ICET) de 2018, Islambad, Pakistán. DOI: 10.1109/ICET.2018.8603616.

2. Awale, N., Pandey, M., dulal, A., Timsiná, B. (2020). Plagiarism Detection in Programming Assignments using Machine Learning. Journal or artificial intelligence and capsule networks, 2(3), 177-184. DOI: 10.36548/jaicn.2020.3.005

3. Chakrabarty, A., Roy, S. (2018). An efficient context-aware agglomerative fuzzy clustering framework for plagiarism detection. International journal of data mining modelling and management, 10(2), 188. DOI: 10.1504/IJDMMM.2018.092533

4. Cruz, E. (30 de enero 2023). Desde 2013 encuesta de UNAM reveló que 52% de académicos atestiguaron algún plagio de tesis. La Hoguera. https://lahoguera.mx/desde-2013-encuesta-de-unam-revelo-que-52-de-academicos-atestiguaron-algun-plagio-de-tesis/

5. Díaz, D. (2015). El uso de Turnitin con retroalimentación mejora la propiedad académica de estudiantes de bachillerato. Ciencia, docencia y tecnología, 26(51), 197-216. https://dialnet.unirioja.es/servlet/articulo?codigo=5265867

6. Díaz, A., García, L. (2018). FP-MAXFLOW: Un algoritmo para la minería de patrones relevantes de longitud máxima. Computación y Sistemas, 22(2), 563-583. DOI: 10.13053/cys-22-2-2498

7. Duracik, M., Callejas, M., Mikusova, M. (2020). Método optimizado basado en algoritmo K-Means como herramienta en la detección de plagio de código fuente. RISTI, (e29),620-

632. https://www.proquest.com/openview/fb8bfe36673b48be7b95c99d83529f32/1?pq-origsite=gscholar&cbl=1006393

8. El-Rashidy, M., Mohamed, R., El-Fishawy, N., Shouman, M. (2022). Reliable plagiarism detection system based on deep learning approaches. Neural Computing and Applications, 34, 18837-18858. https://doi.org/10.1007/s00521-022-07486-w

9. Espinoza, M. (2018). Weka, áreas de aplicación y sus algoritmos: una revisión sistemática de literatura. Revista Científica Ecociencia, 5(Edición Especial), 1-26. DOI: https://doi.org/10.21855/ecociencia.50.153

10. Gil, J. (2021). Minería de texto con R: Aplicaciones y técnicas estadísticas de apoyo. UNED.

11. Hany, M., Gomaa, W. (09 de mayo de 2022). A hybrid approach to paraphrase detection based on text similarities and machine learning classifiers. 2nd International Mobile, Intelligent and Ubiquitous computing conference, El Cairo, Egipto. DOI: 10.1109/MIUCC55081.2022.9781678.

12. Huang, Q., Song, X., Fang, G. (01 de junio de 2020). Code plagiarism detection method based on code similarity and student behavior characteristics. IEEE International Conference on Artificia Intelligence and Computer Applications, Dalian, China. DOI: 10.1109/ICAICA50127.2020.9182389.

13. Hunt, E., Janamsetty, R., Kinares, C., Koh, C., Sánchez, A., Zhan, F., Özdemir, M., Wasim, S., Yolcu, O., Dahal, B., Zhan, J., Geali, L., Oh, P. (2019). Modelos de aprendizaje automático para la identificación de paráfrasis y sus aplicaciones en la detección de plagio. Conferencia Internacional IEEE sobre Gran conocimiento.

14. Kulkarni, S., Govilkar, S., Amin, D. (7 de mayo de 2021). Analysis of Plagiarism Detection Tools and Methods. Proceedings of the 4th international conference on advances in science & technology. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3869091

15. Llovera, Y., Aragón, Y., Cano, P. (2023). Ciberplagio académico entre el estudiantado universitario: un acercamiento al estado actual de la temática (2017-2020). Revista Colombiana de Educación, (87), 207-226. DOI: https://doi.org/10.17227/rce.num87-13143

16. Mancilla, G., Leal, P., Sánchez, A., Vidal, C. (2020). Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: un análisis en minería de datos. Formación Universitaria, 13(6), 23-36. DOI: http://dx.doi.org/10.4067/S0718-50062020000600023

17. Mansoor, M., Al Tamimi, M. (2022). Plagiarism detection system in scientific publication using LSTM networks. Internacional Journal Technical and physical problems of engineering, 4(4), 17-24. http://www.iotpe.com/IJTPE/IJTPE-2022/IJTPE-Issue53-Vol14-No4-Dec2022/3-IJTPE-Issue53-Vol14-No4-Dec2022-pp17-24.pdf

18. Massagram, W., Prapanitisatian, S., Kerson, K. (2018). A novel technique for Thai document plagiarism detection using syntactic parse trees. Engineering & Applied Science Research, 45(4), 290-311. DOI: 10.14456/easr.2018.39

19. Michán, L., Álvarez, E. (2019). Tendencias actuales en el manejo de datos de investigación. BIOCIT, 12(45), 869-880. https://dialnet.unirioja.es/servlet/articulo?codigo=6971157

20. Moreno, J. (2018). Plagio en universidades: estudio de Turnitin y Compilatorio. Sego-Bit (7), 16-23. https://www.researchgate.net/publication/329151488_Plagio_en_universidades_estudio_de_Turnitin_y_Compilatio

21. Navarro, M. (07 de febrero de 2023). Denuncian ante la CNMS la “cara oculta” de las publicaciones científicas universitarias. El cierre digital. https://elcierredigital.com/investigacion/945608780/llevan-juzgado-cara-oculta-negocio-publicaciones-cientificas-universitarias.html

22. Nennuri, R., Geetha, M., Samhitha, M., Sandeep, S., Rochini, G. (26 de mayo2021). Plagiarism detection through data mining techniques. Journal of physiscs: conference series, International Conference on Recent Trends in Computing, San Francisco, EE.UU. DOI: 10.1088/1742-6596/1979/1/012070

23. Perilla, M. (2019). Detección de plagio en código fuente java mediante tokenización y aprendizaje de máquina. Educación, ciencia y tecnologías emergentes para la generación del siglo 21, 79-100. https://www.researchgate.net/publication/344755167

24. Priya, S., Dixit, A., Das, K., Harish, R. (2019). Plagiarism detection in source code using Machine Learning. International journal of engineering and advanced technology, 8,898-900. https://www.ijeat.org/wp-content/uploads/papers/v8i4/D6359048419.pdf

25. Qiubo, H., Jingdong, T., Guozheng, F. (28 de abril de 2019). Research on code plagiarism detection model based on Random Forest and Gradient Boosting Decision Tree. Conferencia internacional de 2019 sobre minería de datos y aprendizaje automático, Hong Kong. DOI: 10.1145/3335656.3335692

26. Reducindo, I., Rivera, L., Rivera, J., Olvera, M. (2017). Integración de plataforma LMS y algoritmo de código abierto para detección y prevención de plagio en Educación Superior. Revista general de información y documentación, 27(2), 299-315. DOI: https://doi.org/10.5209/RGID.58205

27. Rogerson, A., McCarthy, G. (2017). Using internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? International Journal for Educational Integrity, 13(2), 1-15. DOI: 10.1007/s40979-016-0013-y

28. Sánchez, D., González, H., Hernández, Y. (2020). Revisión de algoritmos de detección y seguimiento de objetos con redes profundas para videovigilancia inteligente. Revista Cubana de Ciencias Informáticas, 14(3), 165-197. https://www.redalyc.org/journal/3783/378365834009/html/

29. Santamaría, W. (2015). Técnicas de minería de datos aplicadas en la detección de fraude: Estado del Arte. Universidad Nacional de Colombia. https://www.researchgate.net/publication/240724702_Tecnicas_de_Mineria_de_Datos_Aplicadas_en_la_Deteccion_de_FraudeEstado_del_Arte

30. Sarmiento, J., Ocampo, C. (2023). Enfoques frecuentistas y bayesiano en el estudio del plagio académico. Una propuesta innovadora en investigación educativa. REICE, 21(1), 139-158. DOI: https://doi.org/10.15366/reice2023.21.1.007

31. Shakeel, M., Karim, A. Khan, I. (2020). A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. Information processing & management, 57(3), 102204. DOI: https://doi.org/10.1016/j.ipm.2020.102204

32. Sindhu, L., Idicula, S. (24 de febrero de 2017). Plagiarism detection in Malayalam language text using a composition of similarity measures. Conferencia internacional sobre aprendizaje automático y computación, Singapur. DOI: https://doi.org/10.1145/3055635.3056655

33. Venkatakrishnan, S., Mohan, K., Beattie, J., Correa, E., Dart, J., Deslippe, A., Hexemer, H., Krishnan, A., MacDowell, S., Marchesini, S., Patton, T., Perciano, J., Sethian, R., Stromsness, B., Tierney, C., Tull, D., Ushizima, D., Parkinson, D. (2016). Making advanced scientific algorithms and big scientific data management more accesible. Electronic Imaging, (19),1-7. DOI: 10.2352/ISSN.2470-1173.2016.19.COIMG-155

34. Villanaza S., Arteaga, F., Seijas, c., Rodríguez, O. (2012). Estudio comparativo entre algoritmos de agrupamiento basado en SVM y C-medios difuso aplicados a señales electrocardiográficas arrítmicas. Revista Ingeniería UC, 19(1), 16-24. https://www.redalyc.org/articulo.oa?id=70732261003

35. Viuginov, N., Grachev, P., filchenkov, A. (26 de diciembre de 2020). A Machine Learning based plagiarism detection in source code. 3ra Conferencia Internacional sobre algoritmos, computación e Inteligencia Artificial. Sanya, China. DOI: 10.1145/3446132

36. Xylogiannopoulos, K., Karampelas, P., Alhajj, R. (31 de agosto de 2018). Text mining for plagiarism detection: Multivariate pattern detection for recognition of text similarities. Conferencia Internacional IEEE/ACM 2018 sobre avances en análisis y minería de redes sociales, Barcelona, España. DOI: 10.1109/ASONAM.2018.8508265.